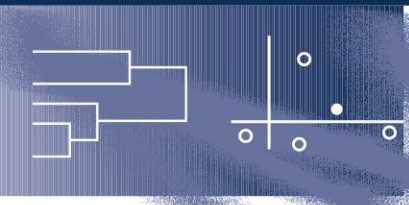
STUDIES IN CLASSIFICATION, DATA ANALYSIS, AND KNOWLEDGE ORGANIZATION

C. Weihs
W. Gaul
Editors

Classification – the Ubiquitous Challenge



Proceedings of the 28th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Dortmund, March 9–11, 2004

Studies in Classification, Data Analysis, and Knowledge Organization

Titles in the Series

H.-H. Bock, W. Lenski, and M.M. Richter (Eds.)
Information Systems and Data Analysis. 1994 (out of print)

E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, and B. Burtschy (Eds.)
New Approaches in Classification and Data Analysis. 1994 (out of print)

W. Gaul and D. Pfeifer (Eds.)
From Data to Knowledge. 1995

H.-H. Bock and W. Polasek (Eds.)
Data Analysis and Information Systems. 1996

E. Diday, Y. Lechevallier, and O. Opitz (Eds.)
Ordinal and Symbolic Data Analysis. 1996

R. Klar and O. Opitz (Eds.)
Classification and Knowledge Organization. 1997

C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock, and Y. Baba (Eds.)
Data Science, Classification, and Related Methods. 1998

I. Balderjahn, R. Mathar, and M. Schader (Eds.)
Classification, Data Analysis, and Data Highways. 1998

A. Rizzi, M. Vichi, and H.-H. Bock (Eds.)
Advances in Data Science and Classification. 1998

M. Vichi and O. Opitz (Eds.)
Classification and Data Analysis. 1999

W. Gaul and H. Locarek-Junge (Eds.)
Classification in the Information Age. 1999

H.-H. Bock and E. Diday (Eds.)
Analysis of Symbolic Data. 2000

H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, and M. Schader (Eds.)
Data Analysis, Classification, and Related Methods. 2000

W. Gaul, O. Opitz, and M. Schader (Eds.)
Data Analysis. 2000

R. Decker and W. Gaul (Eds.)
Classification and Information Processing at the Turn of the Millenium. 2000

S. Borra, R. Rocci, M. Vichi, and M. Schader (Eds.)
Advances in Classification and Data Analysis. 2001

W. Gaul and G. Ritter (Eds.)
Classification, Automation, and New Media. 2002

K. Jajuga, A. Sokołowski, and H.-H. Bock (Eds.)
Classification, Clustering and Data Analysis. 2002

M. Schwaiger and O. Opitz (Eds.)
Exploratory Data Analysis in Empirical Research. 2003

M. Schader, W. Gaul, and M. Vichi (Eds.)
Between Data Science and Applied Data Analysis. 2003

H.-H. Bock, M. Chiodi, and A. Mineo (Eds.)
Advances in Multivariate Data Analysis. 2004

D. Banks, L. House, F.R. McMorris, P. Arabie, and W. Gaul (Eds.)
Classification, Clustering, and Data Mining Applications. 2004

D. Baier and K.-D. Wernecke (Eds.)
Innovations in Classification, Data Science, and Information Systems. 2005

M. Vichi, P. Monari, S. Mignani and A. Montanari (Eds.)
New Developments in Classification and Data Analysis. 2005

D. Baier, R. Decker, and L. Schmidt-Thieme (Eds.)
Data Analysis and Decision Support. 2005

Claus Weihs · Wolfgang Gaul
Editors

Classification – the Ubiquitous Challenge

Proceedings of the 28th Annual Conference
of the Gesellschaft für Klassifikation e.V.
University of Dortmund, March 9–11, 2004

With 181 Figures and 108 Tables

Springer

Professor Dr. Claus Weihs
Universität Dortmund
Fachbereich Statistik
44221 Dortmund
weihs@statistik.uni-dortmund.de

Professor Dr. Wolfgang Gaul
Universität Karlsruhe (TH)
Institut für Entscheidungstheorie
und Unternehmensforschung
76128 Karlsruhe
wolfgang.gaul@wiwi.uni-karlsruhe.de

ISSN 1431-8814
ISBN 3-540-25677-6 Springer-Verlag Berlin Heidelberg New York

Library of Congress Control Number: 2005927145

Springer · Part of Springer Science+Business Media

springeronline.com

Printed in Germany

Softcover-Design: Erich Kirchner, Heidelberg

SPIN 11419365 43/3153 – 5 4 3 2 1 0 – Printed on acid-free paper

Preface

This volume contains revised versions of selected papers presented during the 28th Annual Conference of the Gesellschaft für Klassifikation (GfKl), the German Classification Society. The conference was held at the Universität Dortmund in Dortmund, Germany, in March 2004. Wolfgang Gaul chaired the program committee, Claus Weihs and Ernst-Erich Doberkat were the local organizers. Patrick Groenen, Iven van Mechelen, and their colleagues of the Vereniging voor Ordinatie en Classificatie (VOC), the Dutch-Flemish Classification Society, organized special VOC sessions.

The program committee recruited 17 notable and internationally renowned invited speakers for plenary and semi-plenary talks on their current research work regarding classification and data analysis methods as well as applications. In addition, 172 invited and contributed papers by authors from 18 countries were presented at the conference in 52 parallel sessions representing the whole field addressed by the title of the conference "Classification: The Ubiquitous Challenge". Among these 52 sessions the VOC organized sessions on Mixture Modelling, Optimal Scaling, Multiway Methods, and Psychometrics with 18 papers. Overall, the conference, which is traditionally designed as an interdisciplinary event, again provided an attractive forum for discussions and mutual exchange of knowledge.

Besides the results obtained in the fundamental subjects Classification and Data Analysis, the talks in the applied areas focused on various application topics. Moreover, along with the conference a competition on "Social Milieus in Dortmund", co-organized by the city of Dortmund, took place. Hence the presentation of the papers in this volume is arranged in the following parts:

I. (Semi-)Plenary Presentations
II. Classification and Data Analysis
III. Applications, and
IV. Contest: Social Milieus in Dortmund.

The part on applications has sub-chapters according to the different application fields Archaeology, Astronomy, Bio-Sciences, Electronic Data and Web, Finance and Insurance, Library Science and Linguistics, Macro-Economics, Marketing, Music Science, and Quality Assurance. Within (sub-)parts papers are mainly arranged in alphabetical order with respect to (first) author's names.

I.

Plenary and semi-plenary lectures enclose both conceptual and applied papers. Among the conceptual papers Erosheva and Fienberg present a fully

Bayesian approach to soft clustering and classification within a general framework of mixed membership, Friendly introduces the Milestones Project on documentation and illustration of historical developments in statistical graphics, Hornik discusses consensus partitions particularly when applied to analyze the structure of cluster ensembles, Kiers gives an overview of procedures for constructing bootstrap confidence intervals for the solutions of three-way component analysis techniques, Pahl argues that a classification framework can organize knowledge about software components' characteristics, and Uter and Gefeller define partial attributable risk as a unique solution for allocating shares of attributable risk to risk factors. Within the applied papers Beran presents preprocessing of musical data utilizing prior knowledge from musicology, Fischer et al. introduce a method for the prediction of spatial properties of molecules from the sequence of amino acids incorporating biological background knowledge, Grzybek et al. discuss how far word length may contribute to quantitative typology of texts, and Snoek and Worring present the Time Interval Multimedia Event framework as a robust approach for classification of semantic events in multimodal soccer video.

II.

The second part of this volume is concerned with methodological progress in classification and data analysis and methods presented cover a variety of different aspects.

In the **Classification** part, more precise confidence intervals for the parameters of latent class models using the bootstrap method are proposed (Dias), as well as a method of feature selection for ensembles that significantly reduces the dimensionality of subspaces (Gatnar), and a sensitive two-stage classification system for the detection of events in spite of a noisy background in the processing of thousands of images in a few seconds (Hader and Hamprecht). Variants of bagging and boosting are discussed, which make use of an ordinal response structure (Hechenbichler and Tutz), a methodology for exploring two quality aspects of cluster analyses, namely separation and homogeneity of clusters (Hennig), and a comparison of Adaboost to Arc-x(h) for different values of h in the subsampling of binary classification data is carried out (Khanchel and Limam). The method of distance-based discriminant analysis (DDA) is introduced finding a linear transformation that optimizes an asymmetric data separability criterion via iterative majorization and the necessary number of discriminative dimensions (Kosinov et al.), an efficient hybrid methodology to obtain CHAID tree segments based on multiple dependent variables of possibly different scale types is proposed (Magidson and Vermunt), and possibilities of defining the expectation of p-dimensional intervals (Nordhoff) are described. Design of experiments is introduced into variable selection in classification (Pumplün et al.), as well as the KMC/EDAM method for classification and visualization as an alternative to Kohonen Self-Organizing Maps (Raabe et al.). A clustering of variables approach extended

to situations with missing data based on different imputation methods (Sahmer et al.), a method for binary online-classification incorporating temporal distributed information (Schäfer et al.), and a concept of characteristic regions and a new method, called DiSCo, to simultaneously classify and visualize data (Szepannek and Luebke) are described. The part concludes with two papers discussing multivariate Pareto Density Estimation (PDE), based on information optimality, for data sets containing clusters (Ultsch) and an extension of standard latent class or mixture models that can be used for the analysis of multilevel and repeated measures data (Vermunt and Madgison).

The part on **Data Analysis** starts with papers proposing a robust procedure for estimating a covariance matrix under conditional independence restrictions in graphical modelling (Becker) and a new approach to find principal curves through a multidimensional, possibly branched, data cloud (Einbeck et al.). A three–way multidimensional scaling approach developed to account for individual differences in the judgments about objects, persons or brands (Krolak-Schwerdt), and the Time Series Knowledge Mining (TSKM) framework to discover temporal structures in multivariate time series based on the Unification-based Temporal Grammar (UTG) (Mörchen and Ultsch) are introduced. A framework for the comparison of the information in continuous and categorical data (Nishisato) and an external analysis of two-mode three-way asymmetric multidimensional scaling for the disclosure of asymmetry (Okada and Imaizumi) are presented. Finally, nonparametric regression with the Relevance Vector Machine under inclusion of covariate measurement error (Rummel) is described.

III.

In the third part of this volume all contributions are also related to applications of classification and data analysis methods but structured by their application field.

Two papers deal with applications in **Archaeology**. The first is a historical overview (Ihm) over early publications about formal methods on seriation of archaeological finds, in the second article some cluster analysis models including different data transformations in order to differentiate between brickyards of different areas on the basis of chemical analysis are investigated (Mucha et al.).

Another two papers (both by Bailer-Jones) discuss applications in **Astronomy**. A brief overview of the upcoming Gaia astronomical survey mission, a major European project to map and classify over a billion stars in our Galaxy, and an outline of the challenges are given in the first paper while in the second a novel method based on evolutionary algorithms for designing filter systems for astronomical surveys in order to provide optimal data on stars and to determine their physical parameters is introduced.

The articles with applications in the **Bio-Sciences** all deal with enzyme, DNA, microarray, or protein data, except the presentation of results of a sys-

tematic and quantitative comparison of pattern recognition methods in the analysis of clinical magnetic resonance spectra applied to the detection of brain tumor (Menze et al.). The Generative Topographic Mapping approach as an alternative to SOM for the analysis of microarray data (Grimmenstein et al.) and a finite conservative test for detecting a change point in a binary sequence with Markov dependence and applications in DNA analysis (Krauth) are proposed as well as a new algorithm for finding similar substructures in enzyme active sites with the use of emergent self-organizing neural networks (Kupas and Ultsch). How the feature selection procedure "Significance Analysis of Microarrays" (SAM) and the classification method "Prediction Analysis of Microarrays" (PAM) can be applied to "Single Nucleotide Polymorphism" (SNP) data is explained (Schwender) as well as that using relative differences (RelDiff) instead of LogRatios for cDNA microarray analysis solves several problems like unlimited ranges, numerical instability and rounding errors (Ultsch). Finally, a novel method, PhyNav, to reconstruct the evolutionary relationship from really large DNA and protein datasets is introduced applying the maximum likelihood principle (Vinh et al.).

Among the contributions on applications to **Electronic Data and Web** one paper discusses the application of clustering with restricted random walks on library usage histories in large document sets containing millions of objects (Franke and Thede). In the other four papers different aspects of web-mining are tackled. A tool is described assisting users of online news web-sites in order to reduce information overload (Bomhardt and Gaul), benchmarks are offered with respect to competition and visibility indices as predictors for traffic in web-sites (Schmidt-Mänz and Gaul), an algorithm is introduced for fuzzy two-mode clustering that outperforms collaborative filtering (Schlecht and Gaul), and visualizations of online search queries are compared to improve understanding of searching, viewing, and buying behavior of online shoppers and to further improve the generation of recommendations (Thoma and Gaul).

Two of the articles on **Finance and Insurance** deal with insurance problems: A strategy based on a combination of support vector regression and kernel logistic regression to detect and to model high-dimensional dependency structures in car insurance data sets is proposed (Christmann) and support vector machines are compared to traditional statistical classification procedures in a life insurance environment (Steel and Hechter). Applications in Finance deal with evaluation of global and local statistical models for complex data sets of credit risks with respect to practical constraints and asymmetric cost functions (Schwarz and Arminger), show how linear support vector machines select informative patterns from a credit scoring data pool serving as inputs for traditional methods more familiar to practitioners (Stecking and Schebesch), analyze the question of risk budgeting in continuous time (Straßberger), and formulate a one-factor model for the correlation between probabilities of default across industry branches, comparing it

to more traditional methods on the basis of insolvency rates for Germany (Weißbach and Rosenow).

Besides one contribution on **Library Science** where it is argued that the history of classification is intensively linked to the history of library science (Lorenz) the volume encloses five papers on applications in **Linguistics**. It is shown that one meta-linguistic relation suffices to model the concept structure of the lexicon making use of intensional logic (Bagheri), that improvements of the morphological segmentation of words using classical distributional methods are possible (Benden), and that in Russian texts (letters and poems by three different authors) word length is a characteristic of genre, rather than of authorship (Kelih et al.). A validation method of cluster analysis methods concerning the number and stability of clusters is described with the help of an application in linguistics (Mucha and Haimerl), clustering of word contexts is used in a large collection of texts for word sense induction, i.e. automatic discovery of the possible senses for a given ambiguous word (Rapp), and formal graphs that structure a document-related information space by using a natural language processing chain and a wrapping procedure are proposed (Rist).

There are three papers with applications in **Macro-Economics**, two of them dealing with the comparison of economic structures of different countries. The sensitivity of economic rankings of countries based on indicator variables is discussed (Berrer et al.), structural variables of the 25 member European Union are analyzed and patterns are found to be quite different between the 15 current and the 10 new members (Sell), while the question whether methods measuring (relative) importance of variables in the context of classification allow interpretation of individual effects of highly correlated economic predictors for the German business cycle (Enache and Weihs) is tackled in a more methods-based contribution.

Within the **Marketing** applications one article shows by means of an intercultural survey (Bauer et al.) that the cyber community is not a homogeneous group since online consumers can be classified into the three clusters: "risk avers doubters", "open minded online-shoppers" and "reserved information seekers". Two papers deal with reservation prices. A novel estimation procedure of reservation prices combining adaptive conjoint analysis with a choice task using individually adapted price scales is proposed (Breidert et al.), and an explicit evaluation of variants of conjoint analysis together with two types of data collection is described for the detection of reservation prices of product bundles applied to a seat system offered by a German car manufacturer (Stauß and Gaul).

Music Science is an application field that is present at GfKl conferences for the first time. In this volume one paper deals with time series analysis, the other five papers apply classification methods. A new algorithm structure is introduced for feature extraction from time series, its efficiency is proofed, and illustrated by different classification tasks for audio data (Mierswa). Classifi-

cation methods are used to show that the more the musical sound is unstable in time domain the more pitch bending is admitted to the musician expressing emotions by music (Fricke). Classification rules for quality classes of "sight reading" (SR) are derived (Kopiez et al.) based on indicators of piano practice, mental speed, working memory, inner hearing etc. as well as the total SR performance of 52 piano students. Classification rules are also found for digitized sounds played by different instruments based on the Hough-transform (Röver et al.). Finally, classifications of possibly overlapping drum sounds by linear support vector machines (Van Steelant et al.) and of singers and instruments into high or low musical registers only by means of timbre, i.e. after elimination of pitch information, are proposed (Weihs et al.).

Applications in **Quality Assurance** include one methodological paper (Jessenberger and Weihs) which proposes the use of the expected value of the so-called desirability function to assess the capability of a process. The other papers discuss different statistical aspects of a deep hole drilling process in machine building. The Lyapunov exponent is used for the discrimination between well-predictable and not-well-predictable time series with applications in quality control (Busse). Two multivariate control charts to monitor the drilling process in order to prevent chatter vibrations and to secure production with high quality are proposed (Messaoud et al.) as well as a procedure to assess the changing amplitudes of relevant frequencies over time based on the distribution of periodogram ordinates (Theis and Weihs).

IV.

The fourth part of this volume starts with an introduction to the competition on "Social Milieus in Dortmund" (Sommerer and Weihs). Moreover, the best three papers of the competition by Scheid, by Schäfer and Lemm, and by Röver and Szepannek appear in this volume. We would like to thank the head of the "dortmund-project", Udo Mager, and the head of the Fachbereich "Statistik und Wahlen" of the City of Dortmund, Ernst-Otto Sommerer, for their kind support.

The conference owed much to its sponsors (in alphabetical order)

- Deutsche Forschungsgemeinschaft (DFG), Bonn,
- dortmund-project, Dortmund,
- Fachbereich Statistik, Universität Dortmund, Dortmund,
- Landesbeauftragter für die Beziehungen zwischen den Hochschulen in NRW und den Beneluxstaaten,
- Novartis, Basel, Switzerland,
- Roche Diagnostics, Penzberg,
- sas Deutschland, Heidelberg,
- Sonderforschungsbereich 475, Dortmund,
- Springer-Verlag, Heidelberg,
- Universität Dortmund, and
- John Wiley and Sons, Chicester, UK.

who helped in many ways. Their generous support is gratefully acknowledged.

Additionally, we wish to express our gratitude to the authors of the papers in the present volume, not only for their contributions, but also for their diligence and timely production of the final versions of their papers. Furthermore, we thank the reviewers for their careful reviews of the originally submitted papers, and in this way, for their support in selecting the best papers for this publication.

We would like to emphasize the outstanding work of Uwe Ligges and Nils Raabe who did an excellent job in organizing the program of the conference and the refereeing process as well as in preparing the abstract booklet and this volume, respectively. We also wish to thank our colleague Prof. Dr. Ernst-Erich Doberkat, Fachbereich Informatik, University Dortmund, for co-organizing the conference, and the Fachbereich Statistik of the University Dortmund for all the support, in particular Anne Christmann, Dr. Daniel Enache, Isabelle Grimmenstein, Dr. Sonja Kuhnt, Edelgard Kürbis, Karsten Luebke, Dr. Constanze Pumplün, Oliver Sailer, Roland Schultze, Sibylle Sturtz, Dr. Winfried Theis, Magdalena Thöne, and Dr. Heike Trautmann as well as other members and students of the Fachbereich for helping to organize the conference and making it a big success, and Alla Stankjawitschene and Dr. Stefan Dißmann from the Fachbereich Informatik for all they did in organizing all financial affairs.

Finally, we want to thank Christiane Beisel and Dr. Martina Bihn of Springer-Verlag, Heidelberg, for their support and dedication to the production of this volume.

Dortmund and Karlsruhe, *Claus Weihs, Wolfgang Gaul*
April 2005

Contents

Part II. Classification and Data Analysis

Classification

Data Analysis

Part III. Applications

Archaeology

Astronomy

Bio-Sciences

Electronic Data and Web

Finance and Insurance

Library Science and Linguistics

Macro-Economics

Marketing

Music Science

Quality Assurance

Part IV. Contest: Social Milieus in Dortmund

Part I

(Semi-) Plenary Presentations

Classification and Data Mining in Musicology

Jan Beran

Department of Mathematics and Statistics,
University of Konstanz, 78457 Konstanz, Germany

Abstract. Data in music are complex and highly structured. In this talk a number of descriptive and model-based methods are discussed that can be used as pre-processing devices before standard methods of classification, clustering etc. can be applied. The purpose of pre-processing is to incorporate prior knowledge in musicology and hence to filter out information that is relevant from the point of view of music theory. This is illustrated by a number of examples from classical music, including the analysis of scores and of musical performance.

1 Introduction

Mathematical considerations in music have a long tradition. The most obvious connection between mathematics and music is through physics. For instance, in ancient Greece, the Pythagoreans discovered the musical significance of simple frequency ratios such as 2/1 (octave), 3/2 (pure fifth), 4/3 (pure fourth) etc., and their relation to the length of a string. There are, however, deeper connections between mathematical and musical structures that go far beyond acoustics. Many of these can be discovered using techniques from data mining, together with a priori knowledge from music theory. The results can be used, for instance, to solve classification problems. This is illustrated in the following sections by three types of examples.

2 Music, $1/f$-noise, fractal and chaos

In their celebrated – but also controversial – paper, Voss and Clarke (1975) postulated that recorded music is essentially $1/f$-noise (in the spectral domain), after high frequencies have been eliminated. (The term $1/f$-noise is generally used for random processes whose power spectrum is dominated by low frequencies f such that its value is proportional to $1/f$.) Can we verify this statement? At first, the following question needs to be asked: Which aspects of a composition does recorded music represent? Sound waves are determined not only by the selection of notes, but also by the instrumental sound itself. It turns out, however, that the sound wave of a musical instrument often resembles $1/f$-noise (see e.g. Beran (2003)). Thus, if recorded music looks like $1/f$-noise, this may be due to the instrument rather than a particular composition. To separate instrumental sounds from composed music, we therefore consider the score itself, in terms of pitch and onset

time. The problem of superposition of notes in polyphonic music is solved by replacing chords by arpeggio chords, replacing a chord by the sequence of notes in the chord starting with the lowest note. In order to eliminate high frequencies and to simplify the spectral density, data are aggregated by taking averages over disjoint blocks of $k = 7$ notes (see Beran and Ocker (2001) and Tsai and Chan (2004) for a theoretical justification). Subsequently, a semiparametric fractional model with nonparametric trend function, the so-called SEMIFAR-model (Beran and Feng (2002), also see Beran (1994)), is fitted to the aggregated series. In a SEMIFAR-model, the stochastic part has a generalized spectral density behaving at the origin like $1/f^\alpha$ (where f is the frequency) with $\alpha = 2d$ for some $-\frac{1}{2} < d$. Thus, $1/f$-noise corresponds to $d = 1/2$. Figure 1 shows smoothed histograms of α for four different time periods. The results are based on 60 compositions ranging from the 13th to the 20th century. Apparently a value around $\alpha = 1$ is favored in classical music up to the early romantic period (first three distributions, from above). However, this preference is less clear in the late 19th and the 20th century. Similar investigations can be made for other characteristics of a composition. For instance, we may consider onset time gaps between the occurence of a particular note. Figure 2 displays typical log-log-periodograms and fitted spectra, for gap series referring to the most frequent note (modulo 12). Note that, near zero, each fitted log-log-curve essentially behaves like a straight line with estimated slope $\hat{\alpha}$.

In summary, we may say that $1/f^\alpha$-behaviour with $\alpha > 0$ appears to be common for many musical parameters. The fractal parameter $\alpha = 2d$ may be interpreted as a summary statistic of the degree of variation and memory. From the examples here it is clear, however, that $1/f$-noise is not the only, though perhaps the most frequent, type of variation.

3 Music and entropy

The fractal parameter d (or $\alpha = 2d$) is a measure of randomness and coherence (memory) in the sense mentioned above. Another, in some sense more direct, measure of randomness is entropy. Consider, for instance, the distribution of notes modulo 12 and its entropy. We calculate the entropy for 148 compositions by the following composers: Anonymus (dates of birth between 1200 and 1500), Halle (1240-1287), Ockeghem (1425-1495), Arcadelt (1505-1568), Palestrina (1525-1594), Byrd (1543-1623), Dowland (1562-1626), Hassler (1564-1612), Schein (1586-1630), Purcell (1659-1695), D. Scarlatti (1660-1725), F. Couperin (1668-1733), Croft (1678-1727), Rameau (1683-1764), J.S. Bach (1685-1750), Campion (1686-1748), Haydn (1732-1809), Clementi (1752-1832), W.A. Mozart (1756-1791), Beethoven (1770-1827), Chopin (1810-1849), Schumann (1810-1856), Wagner (1813-1883), Brahms (1833-1897), Faure (1845-1924), Debussy (1862-1918), Scriabin (1872-1915), Rachmaninoff (1873-1943), Schoenberg (1874-1951), Bartok (1881-1945), Webern

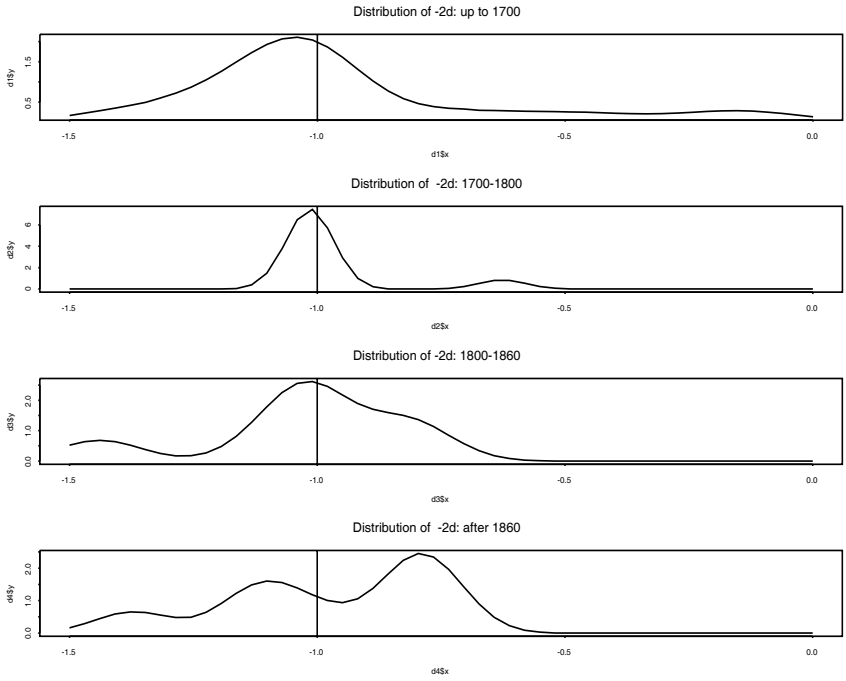


Fig. 1. Distribution of $-\alpha = -2d$ for four different time periods.

(1883-1945), Prokoffieff (1891-1953), Messiaen (1908-1992), Takemitsu (1930-1996) and Beran (*1959). For a detailed description how the entropy is calculated see Beran (2003). A plot of entropy against the date of birth of the composer (figure 3) reveals a positive dependence, in particular after 1400. Why that is so can be seen, at least partially, from star plots of the distributions. Figure 4 shows a random selection of star plots ranging from the 15th to the 20th century. In order to reveal more structure, the 12 note categories are ordered according to the ascending circle of fourths. The most striking feature is that for compositions that may be classified as purely tonal in a traditional sense, there is a neighborhood of 7 to 8 adjacent notes where beams are very long, and for the rest of the categories not much can be seen. The plausible reason is that in tonal music, the circle of fourths is a dominating feature that determines a lot of the structure. This is much less the case for classical music of the 20th century. With respect to entropy it means that for newer music, the (marginal) distribution of notes is much less predictable than in earlier music (see figure 3 where composers born after 1881 are marked as "20th century", namely Prokoffieff, Messiaen, Takemitsu, Webern and Beran). Note, however, that there are also a few outliers in figure 3. Thus, the rule is not universal, and entropy may depend on the individual composer or

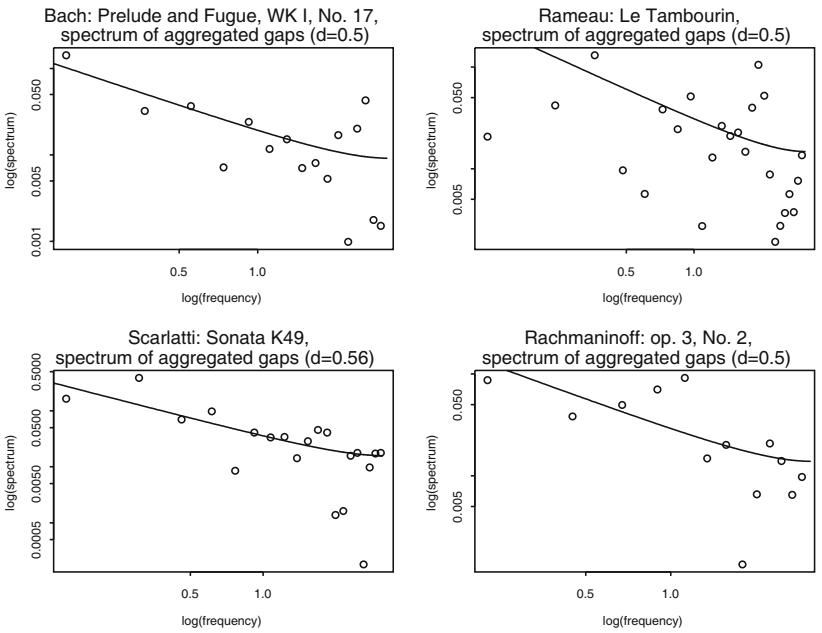


Fig. 2. Log-log-periodograms and fitted spectra for gap time series.

even the composition. In the last millennium, music moved gradually from rather strict rules to increasing variety. It is therefore not surprising that variability increases throughout the centuries - composers simply have more choice. On the other hand, a comparison of Schumann's entropies (which were not included in figure 3) with those by Bach points in the opposite direction (figure 5). As a cautionary remark it should also be noted that this data set is a very small, and partially unbalanced, sample from the huge number of existing compositions. For instance, Prokoffieff is included 15 times whereas many other composers of the 20th century are missing. A more systematic empirical investigation will need to be carried out to obtain more conclusive results.

4 Score information and performance

Due to advances in music technology, performance theory is a very active area of research where statistical analysis plays an essential role. In contrast to some other branches of musicology, repeated observations and controlled experiments can be carried out. With respect to music where a score exists, the following question is essential: Which information is there in a score, and how can it be quantified? Beran and Mazzola (1999a) (also see Mazzola (2002) and Beran (2003)) propose to encode structural information of a

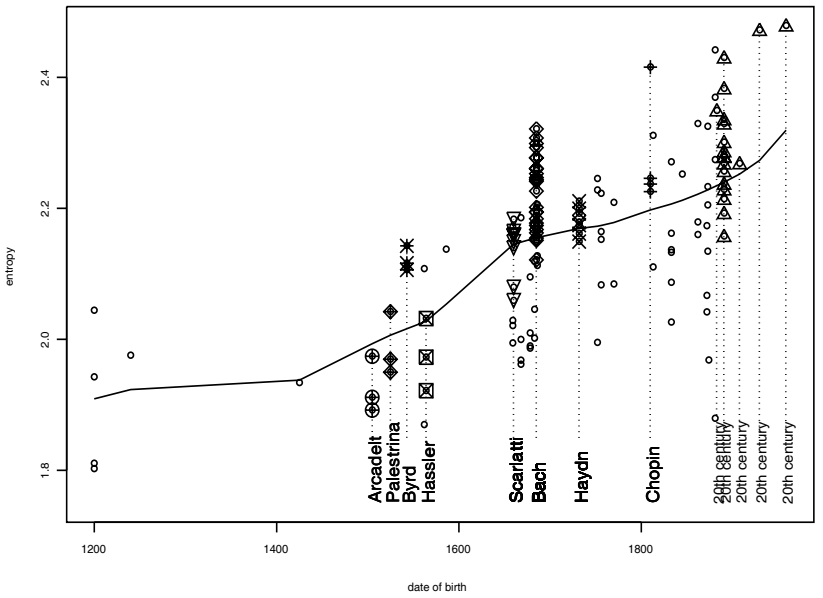


Fig. 3. Entropy of notes in $\mathbb{Z}_{12}$ versus date of birth.

score by so-called metric, harmonic and melodic weights or indicators. These curves quantify the metric, harmonic and melodic importance of a note respectively. A modified motivic indicator based on a priori knowledge about motifs in the score is defined in Beran (2003). Figure 6 shows some indicator functions corresponding to eight different motifs in Schumann's Träumerei. These curves can be related to observed performance data by various statistical methods (see e.g. Beran (2003), Beran and Mazzola (1999b, 2000, 2001)). For instance, figure 7 displays tempo curves of different pianists after applying data sharpening with the indicator function of motif 2. Sharpening was done by considering only those onset times where the indicator curve of motif 2 is above its 90th percentile. This leads to simplified tempo curves where differences and communalities are more visible. Also, sharpened tempo curves can be used as input for other statistical techniques, such as classification. A typical example is given in figure 8, where clustering is based the motif-2-sharpened tempo curves in figure 7.

Acknowledgements

I would like to thank B. Repp for providing us with the tempo measurements.

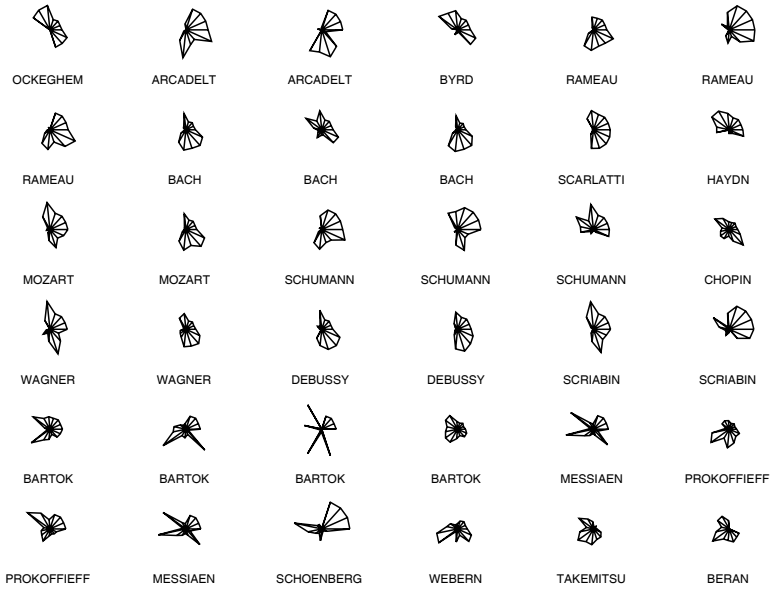


Fig. 4. Star plots of $\mathbb{Z}_{12}$-distribution, ordered according to the circle of fourths.

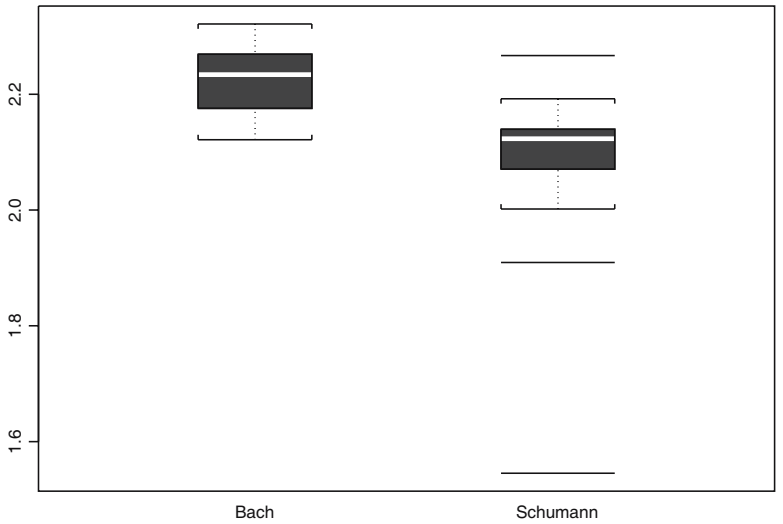


Fig. 5. Boxplots of entropies for Bach (left) and Schumann (right), based on note distribution in $\mathbb{Z}_{12}$.

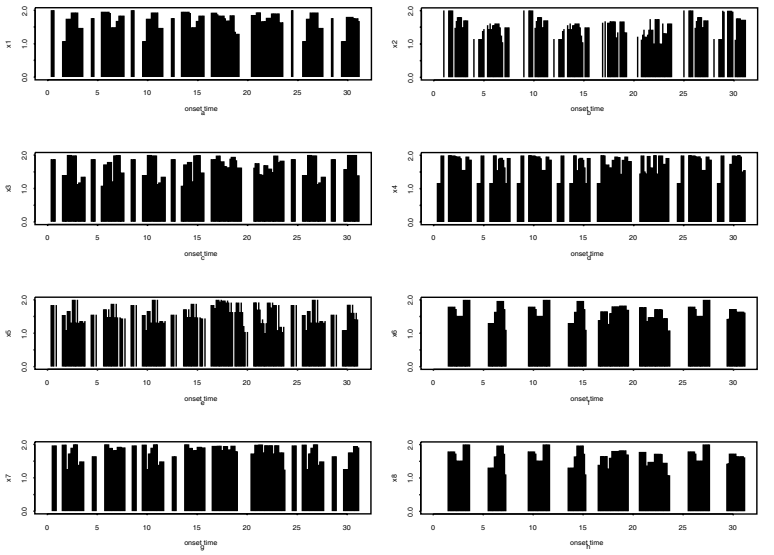


Fig. 6. Motivic indicators for Schumann's Träumerei.

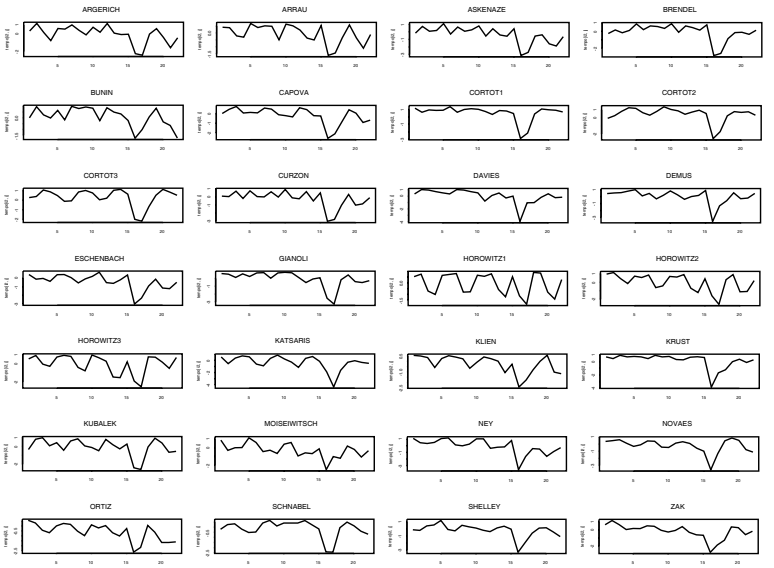


Fig. 7. Schumann's Träumerei: Tempo curves sharpened by 90th percentile of motif-curve 2.

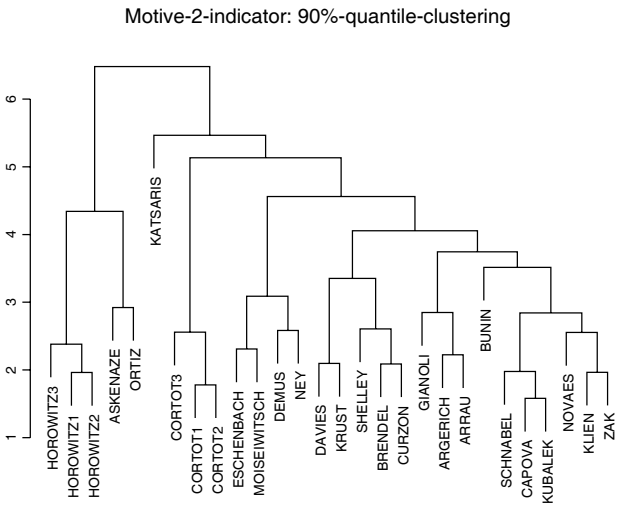


Fig. 8. Schumann's Träumerei: Tempo clusters based on sharpened tempo.

References

BERAN, J. (2003): *Statistics in Musicology*. Chapman & Hall, CRC Press, Boca Raton.

BERAN, J. (1994): *Statistics for long-memory processes*. Chapman & Hall, London.

BERAN, J. and FENG, Y. (2002): SEMIFAR models – a semiparametric framework for modeling trends, long-range dependence and nonstationarity. *Computational Statistics and Data Analysis, 40(2), 690–713*.

BERAN, J. and MAZZOLA, G. (1999a): Analyzing musical structure and performance - a statistical approach. *Statistical Science, 14(1), 47–79*.

BERAN, J. and MAZZOLA, G. (1999b): Visualizing the relationship between two time series by hierarchical smoothing. *J. Computational and Graphical Statistics, 8(2), 213–238*.

BERAN, J. and MAZZOLA, G. (2000): Timing Microstructure in Schumann's Träumerei as an Expression of Harmony, Rhythm, and Motivic Structure in Music Performance. *Computers Mathematics Appl., 39(5-6), 99–130*.

BERAN, J. and MAZZOLA, G. (2001): Musical composition and performance - statistical decomposition and interpretation. *Student, 4(1), 13–42*.

BERAN, J. and OCKER, D. (2000): *Temporal Aggregation of Stationary and Nonstationary FARIMA(p,d,0) Models*. CoFE Discussion Paper, No. 00/22. University of Konstanz.

MAZZOLA, G. (2002): *The topos of music*. Birkhäuser, Basel.

TSAI, H. and CHAN, K.S. (2004): *Temporal Aggregation of Stationary and Nonstationary Discrete-Time Processes*. Technical Report, No. 330, University of Iowa, Statistics and Actuarial Science.

VOSS, R.F. and CLARKE, J. (1975): 1/f noise in music and speech. *Nature, 258, 317–318*.

Bayesian Mixed Membership Models for Soft Clustering and Classification

Elena A. Erosheva[1] and Stephen E. Fienberg[2]

[1] Department of Statistics,
School of Social Work,
Center for Statistics and the Social Sciences,
University of Washington, Seattle, WA 98195, U.S.A.
[2] Department of Statistics,
Center for Automated Learning and Discovery,
Center for Computer and Communications Security
Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.

Abstract. The paper describes and applies a fully Bayesian approach to soft clustering and classification using mixed membership models. Our model structure has assumptions on four levels: population, subject, latent variable, and sampling scheme. Population level assumptions describe the general structure of the population that is common to all subjects. Subject level assumptions specify the distribution of observable responses given individual membership scores. Membership scores are usually unknown and hence we can also view them as latent variables, treating them as either fixed or random in the model. Finally, the last level of assumptions specifies the number of distinct observed characteristics and the number of replications for each characteristic. We illustrate the flexibility and utility of the general model through two applications using data from: (i) the National Long Term Care Survey where we explore types of disability; (ii) abstracts and bibliographies from articles published in *The Proceedings of the National Academy of Sciences*. In the first application we use a Monte Carlo Markov chain implementation for sampling from the posterior distribution. In the second application, because of the size and complexity of the data base, we use a variational approximation to the posterior. We also include a guide to other applications of mixed membership modeling.

1 Introduction

The canonical clustering problem has traditionally had the following form: for N units or objects measured on J variables, organize the units into G groups, where the nature, size, and often the number of the groups is unspecified in advance. The classification problem has a similar form except that the nature and the number of groups are either known theoretically or inferred from units in a training data set with known group assignments. In machine learning, methods for clustering and classification are referred to as involving "unsupervised" and "supervised learning" respectively. Most of these methods assume that every unit belongs to exactly one group. In this paper, we will primarily focus on clustering, although methods described can be used for both clustering and classification problems.

Some of the most commonly used clustering methods are based on hierarchical or agglomerative algorithms and do not employ distributional assumptions. Model-based clustering lets $\mathbf{x} = (x_1, x_2, \ldots, x_J)$ be a sample of J characteristics from some underlying joint distribution, $Pr(\mathbf{x}|\theta)$. Assuming each sample is coming from one of G groups, we estimate $Pr(\mathbf{x}|\theta)$ indicating presence of groups or lack thereof. We represent the distribution of the gth group by $Pr_g(\mathbf{x}|\theta)$ and then model the observed data using the mixture distribution:

$$Pr(\mathbf{x}|\theta) = \sum_{g=1}^{G} \pi_g Pr_g(\mathbf{x}|\theta), \tag{1}$$

with parameters $\{\theta, \pi_g\}$, and G.

The assumption that each object belongs exclusively to one of the G groups or latent classes may not hold, e.g., when characteristics sampled are individual genotypes, individual responses in an attitude survey, or words in a scientific article. In such cases, we say that objects or individuals have mixed membership and the problem involves *soft clustering* when the nature of groups is unknown or *soft classification* when the nature of groups is known through distributions $Pr_g(\mathbf{x}|\theta)$, $g = 1, \ldots, G$, specified in advance.

Mixed membership models have been proposed for applications in several diverse areas. We describe six of these here:

1. *NLTCS Disability Data.* The National Long Term Care Survey assesses disability in U.S. elderly population. We have been working with a 2^{16} contingency table on functional disability drawing on combined data from the 1982, 1984, 1989, and 1994 waves of the survey. The dimensions of the table correspond to 6 Activities of Daily Living (ADLs)–e.g., getting in/out of bed and using a toilet–and 10 Instrumental Activities of Daily Living (IADLs)–e.g., managing money and taking medicine. In Section 3, we describe some of our results for the combined NLTCS data. We note that further model extensions are possible to account for the longitudinal nature of the study, e.g., via employing a powerful conditional independence assumption to accommodate a longitudinal data structure as suggested by Manton et al. (1994).
2. *DSM-III-R Psychiatric Classifications.* One of the earliest proposals for mixed membership models was by Woodbury et al. (1978), in the context of disease classification. Their model became known as the *Grade of Memebership* or GoM model, and was later used by Nurnberg et al. (1999) to study the DSM-III-R typology for psychiatric patients. Their analysis involved $N = 110$ outpatients and used the $J = 112$ DSM-III-R diagnostic criteria for clustering in order to reassess the appropriateness of the "official" 12 personality disorders. One could also approach this problem as a classical classification problem but with $J > N$.

3. *Peanut Butter Market Segmentation.* Seetharaman et al. (2001) describe data on peanut butter purchases drawn from A.C. Nielsen's scanner database. They work with data from 488 households over 4715 purchase occasions (chosen such that there are at least 5 per household) for 8 top brands of peanut butter. For each choice occasion we have: (a) shelf price, (b) information on display/feature promotion, and a set of household characteristics used to define "market segments" or groupings of households. Market segmentation has traditionally been thought of as a standard clustering problem but Varki et al. (2000) proposed a mixed-membership model for this purpose which is a variant on the GOM model.
4. *Matching Words and Pictures.* Blei and Jordan (2003) and Barnard et al. (2003) have been doing mixed-membership modeling in machine learning combining different sources of information in text documents, i.e., main text, photographic images, and image annotations. They estimate the joint distribution of these characteristics via employing hierarchical versions of a model known as the Latent Dirichlet Allocation in machine learning. This allows them to perform such tasks as automatic image annotations (recognizing image regions that portray, for example, clouds, water, and flowers) and text-based image retrieval (finding unannotated images that correspond to a given text query) with remarkably good performance.
5. *Race and Population Genetic Structure.* In a study of human population structure Rosenberg et al. (2002) used genotypes at 377 autosomal microsatellite loci in 1056 individuals from 52 populations and part of their analysis focuses on the soft clustering of individuals in groups. One of the remarkable results of their study which uses the mixed membership methods of Pritchard et al. (2002), is a typology structure that is very close to the "traditional" 5 main racial groups, a notion much maligned in the recent social science and biological literatures.
6. *Classifying Scientific Publications.* Erosheva, Fienberg et al. (2004) and Griffiths and Styvers (2004) have used mixed membership models to analyse related data bases involving abstracts, text, and references of articles drawn from the Proceedings of the National Academy of Sciences U S A (PNAS). Their mutual goal was to understand the organization of scientific publications in PNAS and we explore the similarities and differences between their approaches and results later in Section 4.

What these examples have in common is the mixed membership structure. In the following sections, we first introduce our general framework for mixed membership models and then we illustrate its application in two of the examples, using the PNAS and NLTCS data sets.

2 Mixed membership models

The general mixed membership model relies on four levels of assumptions: population, subject, latent variable, and sampling scheme. At the population level, we describe the general structure of the population that is common to all subjects, while at the subject level we specify the distribution of observable responses given individual membership scores. At the latent variable level, we declare whether the membership scores are considered fixed or random with some distribution. Finally, at the last level, we specify the number of distinct observed characteristics and the number of replications for each characteristic. Following the exposition in Erosheva (2002) and Erosheva et al. (2004), we describe the assumptions at the four levels in turn.

Population level. We assume that there are K basis subpopulations (extreme or pure types) in the population of interest. For each subpopulation k, we denote by $f(x_j|\theta_{kj})$ the probability distribution for response variable j, where θ_{kj} is a vector of parameters. Moreover, we assume that, within a subpopulation, responses for the observed variables are independent.

Subject level. For each subject, membership vector $\lambda = (\lambda_1, \ldots, \lambda_K)$ represents the degrees of a subject's membership in each of the subpopulations or the consonance of the subject with each of the pure types. The form of the conditional probability, $Pr(x_j|\lambda) = \sum_k \lambda_k f(x_j|\theta_{kj})$, combined with the assumption that the response variables x_j are independent conditional on membership scores, fully defines the distribution of observed responses x_j for each subject. In addition, given the membership scores, we take the observed responses from different subjects to be independent.

Latent variable level. We can either assume that the latent variables are fixed unknown constants or that they are random realizations from some underlying distribution.

1. If the membership scores λ are fixed but unknown, then

$$Pr(x_j|\lambda;\boldsymbol{\theta}) = \sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj}) \tag{2}$$

   is the conditional probability of observing x_j, given the membership scores λ and parameters $\boldsymbol{\theta}$.
2. If the membership scores λ are realizations of latent variables from some distribution D_α, parameterized by α, then

$$Pr(x_j|\alpha,\boldsymbol{\theta}) = \int \left(\sum_{k=1}^{K} \lambda_k f(x_j|\theta_{kj}) \right) dD_\alpha(\lambda) \tag{3}$$

   is the marginal probability of observing x_j, given the parameters.

Sampling scheme. Suppose we observe R independent replications of J distinct characteristics for one subject, $\{x_1^{(r)}, \ldots, x_J^{(r)}\}_{r=1}^R$. If the membership scores are realizations from the distribution D_α, the conditional probability is

$$Pr\left(\{x_1^{(r)}, \ldots, x_J^{(r)}\}_{r=1}^R | \alpha, \boldsymbol{\theta}\right) = \int \left(\prod_{j=1}^{J} \prod_{r=1}^{R} \sum_{k=1}^{K} \lambda_k f(x_j^{(r)} | \theta_{kj})\right) dD_\alpha(\lambda). \quad (4)$$

If we treat the latent variables as unknown constants, we get an analogous representation for the conditional probability of observing R replications of J variables. In general, the number of observed characteristics J need not be the same across subjects, and the number of replications R need not be the same across observed characteristics.

This mixed membership model framework unifies several specialized models that have been developed independently in the social sciences, in genetics, and in machine learning. Each corresponds to different choices of J and R, and different latent variable assumptions. For example, the standard GoM model of Woodbury and Clive (1974) and Manton et al. (1994) assumes that we observe responses to J survey questions without replications, i.e., $R = 1$, and treats the membership scores as fixed unknown constants (fixed-effects). Examples of the "fixed-effect" GoM analyses include but are not limited to: an analysis mentioned earlier of DSM-III psychiatric classifications in Nurnberg et al. (1999), a study of data on remote sensing (Talbot (1996)), an analysis of business opportunities (Talbot et al. (2002)), and a classification of individual tree crowns into species groups from aerial photographs (Brandtberg (2002)).

Another class of mixed membership models is based directly on the standard GoM model but places a distribution on the membership scores. Thus, Potthoff et al. (2000) treat the membership scores as realizations of Dirichlet random variables and are able to use marginal maximum likelihood estimation in a series of classification examples when the number of items J is small. Erosheva (2002) provides a Markov chain Monte Carlo estimation scheme for the GoM model also assuming the Dirichlet distribution on the membership scores. Varki et al. (2000) employ a mixture of point and Dirichlet distributions as the generating distribution for the membership scores in their work.

Independently from the GoM developments, in genetics Pritchard et al. (2000) use a *clustering model with admixture*. For diploid individuals the clustering model assumes that $R = 2$ replications (genotypes) are observed at J distinct locations (loci) and that the membership scores are random Dirichlet realizations. Again, J and N vary in this and related applications. In the Introduction, we briefly described an example of findings obtained via this model in the study on race and population genetic structure by Rosenberg et al. (2002).

A standard assumption in machine learning of text and other objects is that a single characteristic is observed multiple times. For example, for a

text document of length L only one distinct characteristic, a word, is observed with $R = L$ realizations. In this set-up, the work of Hofmann (2001) on *probabilistic latent semantic analysis* treated membership scores as fixed unknown constants and that of Blei et al. (2003) adopted a Dirichlet generating distribution for the membership scores. More recently, this line of modeling has moved from considering a single characteristic (e.g., words in a document) to working with a combination of distinct characteristics. An example that we discussed in this area is by Barnard et al. (2003) who modeled a combination of words and segmented images via a mixed membership structure.

Given this multiplicity of unrelated mixed membership model developments, we should not be surprised by the variety of estimation methods adopted. Broadly speaking, estimation methods are of two types: those that treat membership scores as fixed and those that treat them as random. The first group includes the numerical methods introduced by Hofmann (2003) and by Kovtun et al. (2004b), and joint maximum likelihood type methods described in Manton et al. (1994) and Varki and Cooil (2003) where fixed effects for the membership scores are estimated in addition to the population parameter estimates. The statistical properties of the estimates in these approaches, such as consistency, identifiability, and uniqueness of solutions, are suspect. The second group includes variational estimation methods used by Blei et al. (2003), expectation-propagation methods developed by Minka and Lafferty (2002), joint maximum likelihood approaches of Potthoff et al. (2000) and Varki et al. (2000), and Bayesian MCMC simulations (Pritchard et al. (2002), Erosheva (2002, 2003a)). These methods solve some of the statistical and computational problems, but many other challenges and open questions still remain as we illustrate below.

3 Disability types among older adults

3.1 National Long Term Care Survey

The National Long-Term Care Survey (NLTCS), conducted in 1982, 1984, 1989, 1994, and 1999, was designed to assess chronic disability in the U.S. elderly Medicare-enrolled population (65 years of age or older). Beginning with a screening sample in 1982, individuals were followed in later waves and additional samples were subsequently added maintaining the sample at about 20,000 Medicare enrollees in each wave. The survey aims to provide data on the extent and patterns of functional limitations (as measured by activities of daily living (ADL) and instrumental activities of daily living (IADL), availability and details of informal caregiving, use of institutional care facilities, and death. NLTCS public use data can be obtained from the Center for Demographic Studies, Duke University.

Erosheva (2002) considered the mixed membership model with up to $K = 5$ subpopulations or extreme profiles for the 16 ADL/IADL measures, pooled

across four survey waves of NLTCS, 1982, 1984, 1989, and 1994. For each ADL/IADL measure, individuals can be either disabled or healthy. Thus the data form a 2^{16} contingency table. The table has 65,536 cells, only 3,152 of which are non-zero and there are a total of $N = 21,574$ observations. This is a large sparse contingency table that is not easily analyzed using classical statistical methods such as those associated with log-linear models.

3.2 Applying the mixed membership model

Following the GoM structure for dichotomous variables, we have $J = 16$ dichotomous characteristics observed for each individual and the number of replications R is 1. For each extreme profile k, the probability distribution for characteristic j, $f(x_j|\theta_{kj})$ is binomial parameterized by the probability of the positive response μ_{kj}.

We assume that the membership scores follow a Dirichlet distribution D_α and employ Monte Carlo Markov chain estimation for the latent class representation of the GoM model (Erosheva (2003a)). We obtain posterior means for the response probabilities of the extreme profiles and posterior means of the membership scores conditional on observed response patterns. Estimated response probabilities of the extreme profiles provide a qualitative description of the extreme categories of disability as tapped by the 16 ADL/IAD measures while the estimated parameters α of the Dirichlet distribution describe the distribution of the mixed membership scores in the population.

Although the Deviance Information Criteria (Spiegelhalter et al. (2002)) indicates an improvement in fit for K increasing from 2 to 5 with the largest improvement for K going from 2 to 3, other considerations point out that a $K = 4$ solution might be appropriate for this data set (Erosheva (2002)). In Table 1, we provide posterior means and standard deviation estimates for the parameters of the GoM model with four extreme profiles. The estimates of ξ_i and α_0 reported in Table 1 and their product gives the vector of Dirichlet distribution parameters. The estimated distribution of the membership scores is bathtub shaped, indicating that the majority of individual profiles are close to estimated extreme profiles.

One of the most significant findings in this analysis is based on examining interpretations of the extreme profiles for the mixed membership models for $K = 4, 5$ which rejects the hypothesis of a unidimensional disability structure, i.e., the extreme profiles are qualitatively different and can not be ordered by severity. In particular, individuals at two of the estimated extreme profiles can be described as mostly cognitive and mostly mobility impaired individuals. For more details on the analysis and substantive findings see Erosheva (2002).

Table 1. Posterior mean (standard deviation) estimates for $K = 4$ extreme profiles. The ADL items are: (1) eating, (2) getting in/out of bed, (3) getting around inside, (4) dressing, (5) bathing, (6) using toilet. The IADL items are: (7) doing heavy house work, (8) doing light house work, (9) doing laundry, (10) cooking, (11) grocery shopping, (12) getting about outside, (13) traveling, (14) managing money, (15) taking medicine, (16) telephoning.

k	1	2	3	4
$\mu_{k,1}$	0.000 (3e-04)	0.002 (2e-03)	0.001 (6e-04)	0.517 (1e-02)
$\mu_{k,2}$	0.000 (3e-04)	0.413 (1e-02)	0.001 (5e-04)	0.909 (7e-03)
$\mu_{k,3}$	0.001 (5e-04)	0.884 (1e-02)	0.018 (8e-03)	0.969 (5e-03)
$\mu_{k,4}$	0.007 (2e-03)	0.101 (6e-03)	0.016 (4e-03)	0.866 (8e-03)
$\mu_{k,5}$	0.064 (4e-03)	0.605 (9e-03)	0.304 (9e-03)	0.998 (2e-03)
$\mu_{k,6}$	0.005 (2e-03)	0.316 (9e-03)	0.018 (4e-03)	0.828 (8e-03)
$\mu_{k,7}$	0.230 (7e-03)	0.846 (7e-03)	0.871 (7e-03)	1.000 (3e-04)
$\mu_{k,8}$	0.000 (2e-04)	0.024 (4e-03)	0.099 (7e-03)	0.924 (7e-03)
$\mu_{k,9}$	0.000 (3e-04)	0.253 (9e-03)	0.388 (1e-02)	0.999 (1e-03)
$\mu_{k,10}$	0.000 (2e-04)	0.029 (5e-03)	0.208 (1e-02)	0.987 (4e-03)
$\mu_{k,11}$	0.000 (3e-04)	0.523 (1e-02)	0.726 (1e-02)	0.998 (2e-03)
$\mu_{k,12}$	0.085 (5e-03)	0.997 (2e-03)	0.458 (1e-02)	0.950 (4e-03)
$\mu_{k,13}$	0.021 (4e-03)	0.585 (1e-02)	0.748 (1e-02)	0.902 (5e-03)
$\mu_{k,14}$	0.001 (7e-04)	0.050 (5e-03)	0.308 (1e-02)	0.713 (8e-03)
$\mu_{k,15}$	0.013 (2e-03)	0.039 (4e-03)	0.185 (8e-03)	0.750 (8e-03)
$\mu_{k,16}$	0.014 (2e-03)	0.005 (2e-03)	0.134 (7e-03)	0.530 (9e-03)
ξ_k	0.216 (2e-02)	0.247 (2e-02)	0.265 (2e-02)	0.272 (2e-02)
α_0	0.197 (5e-03)			

4 Classifying publications by topic

4.1 Proceedings of the National Academy of Sciences

The *Proceedings of the National Academy of Sciences* (PNAS) is the world's most cited multidisciplinary scientific journal. Historically, when submitting a research paper to the Proceedings, authors have to select a major category from Physical, Biological, or Social Sciences, and a minor category from the list of topics. PNAS permits dual classifications between major categories and, in exceptional cases, within a major category. The lists of topics change over time in part to reflect changes in the National Academy sections. Since in the nineties the vast majority of the PNAS research papers was in the Biological Sciences, our analysis focuses on this subset of publications. Another reason for limiting ourselves to one major category is that we expect papers from different major categories to have a limited overlap.

In the Biological Sciences there are 19 topics. Table 2 gives the percentages of published papers for 1997-2001 (Volumes 94-98) by topic and and numbers of dual classification papers in each topic.

Table 2. Biological Sciences publications in PNAS volumes 94–98, by subtopic, and numbers of papers with dual classifications. The numbers in the final column represent projections based on our model.

	Topic	Number		% Dual	% Dual	More Dual?
1	Biochemistry	2578	21.517	33	18.436	338
2	Medical Sciences	1547	12.912	13	.263	84
3	Neurobiology	1343	11.209	9	5.028	128
4	Cell Biology	1231	10.275	10	5.587	111
5	Genetics	980	8.180	14	7.821	131
6	Immunology	865	7.220	9	5.028	39
7	Biophysics	636	5.308	40	22.346	62
8	Evolution	510	4.257	12	6.704	167
9	Microbiology	498	4.157	11	6.145	42
10	Plant Biology	488	4.073	4	2.235	54
11	Developmental Biology	366	3.055	2	1.117	43
12	Physiology	340	2.838	1	0.559	34
13	Pharmacology	188	1.569	2	1.117	34
14	Ecology	133	1.110	5	2.793	16
15	Applied Biological Sciences	94	0.785	6	3.352	7
16	Psychology	88	0.734	1	0.559	22
17	Agricultural Sciences	43	0.359	2	1.117	8
18	Population Biology	43	0.359	5	2.793	4
19	Anthropology	10	0.083	0	0	2
	Total	11981	100	179	100	1319

4.2 Applying the mixed membership model

The topic labels provide an author-designated classification structure for published materials. Notice that the vast majority of the articles are members of only a single topic. We represent each article by collections of words in the abstract and references in the bibliography. For our mixed membership model, we assume that there is a fixed number of extreme categories or aspects, each of which is characterized by multinomial distributions over words (in abstracts) and references (in bibliographies). A distribution of words and references in each article is given by the convex combination of the aspects' multinomials weighted by proportions of the article's content coming from each category. These proportions, or membership scores, determine soft clustering of articles with respect to the internal categories.

Choosing a suitable value for the number of internal categories or aspects, K, in this type of setting is difficult. We have focused largely on two versions of the model, one with eight aspects and the other with ten. The set of parameters in our model is given by multinomial word and reference probabilities for each aspect, and by the parameters of Dirichlet distribution, which is a generating distribution for membership scores. There are 39,616 unique words and 77,115 unique references in our data, hence adding an aspect corresponds

to having $39{,}615 + 77{,}114 + 1 = 116{,}730$ additional parameters. Because of the large numbers of parameters involved, it is difficult to assess the extent to which the added pair of aspects actually improve the fit of the model to the data. In a set of preliminary comparisons we found little to choose between them in terms of fit and greater ease of interpretation for the eight aspect model. In Erosheva et al. (2004) we report on the details of the analysis of the $K = 8$ aspect model and its interpretation and we retain that focus here.

From our analysis of high probability words and references, the 8 aspects of our model have the following interpretation:

1. Intracellular signal transaction, neurobiology.
2. Evolution, molecular evolution.
3. Plant molecular biology.
4. Developmental biology; brain development.
5. Biochemistry, molecular biology; protein structural biology.
6. Genetics, molecular biology; DNA repair, mutagenesis, cell cycle.
7. Tumor immunology; HIV infection.
8. Endocrinology, reporting of experimental results; molecular mechanisms of obesity.

Based on the interpretations, it is difficult to see whether the majority of aspects correspond to a single topic from the official PNAS classifications. To investigate a correspondence between the estimated aspects and the given topics further, we examine aspect "loadings" for each paper. Given estimated parameters of the model, the distribution of each article's "loadings" can be obtained via Bayes' theorem. The variational and expectation-propagation procedures give Dirichlet approximations to the posterior distribution $p(\boldsymbol{\lambda}(d), \boldsymbol{\theta})$ for each document d. We employ the mean of this Dirichlet as an estimate of the "weight" of the document on each aspect.

We can gauge the sparsity of the loadings by the parameters of the Dirichlet distribution, which for the $K = 8$ model we estimate as $\alpha_1 = 0.0195$, $\alpha_2 = 0.0203$, $\alpha_3 = 0.0569$, $\alpha_4 = 0.0346$, $\alpha_5 = 0.0317$, $\alpha_6 = 0.0363$, $\alpha_7 = 0.0411$, $\alpha_8 = 0.0255$. This estimated Dirichlet, which is the generative distribution of membership scores, is "bathtub shaped" on the simplex; as a result, articles will tend to have relatively high membership scores in only a few aspects.

To summarize the aspect distributions for each topic, we provide a graphical representation of these values for $K = 8$ and $K = 10$ in Figure 1 and Figure 2, respectively. Examining the rows of Figure 1, we see that, with the exception of Evolution and Immunology, the subtopics in Biological Sciences are concentrated on more than one internal category. The column decomposition, in turn, can assist us in interpreting the aspects. Aspect 8, for example, which from the high probability words seems to be associated with the reporting of experimental results, is the aspect of origin for a combined 37% of Physiology, 30% of Pharmacology, and 25% of Medical Sciences papers, according to the mixed membership model.

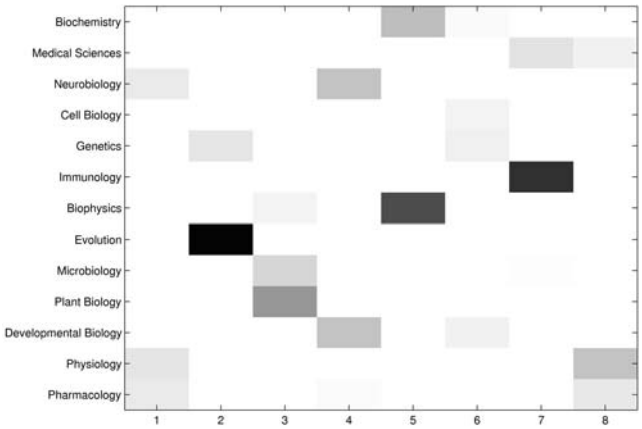


Fig. 1. Graphical representation of mean decompositions of aspect membership scores for $K = 8$. Source: Erosheva et al.(2004).

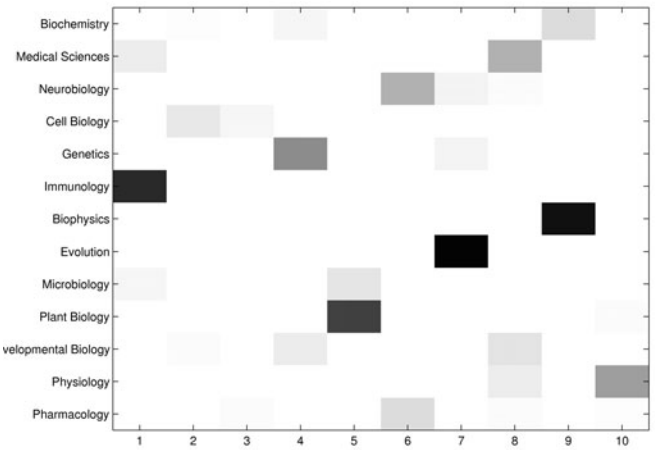


Fig. 2. Graphical representation of mean decompositions of aspect membership scores for $K = 10$.

Finally, we compare the loadings (posterior means of the membership scores) of dual-classified articles to those that are singly classified. We consider two articles as having similar membership vectors if their loadings are equal for the first significant digit for all aspects. One might consider singly classified articles that have membership vectors similar to those of dual-classified articles as interdisciplinary, i.e., the articles that should have had dual classification but did not. We find that, for 11 percent of the singly classified articles, there is at least one dual-classified article that has similar membership scores. For example, three biophysics dual-classified articles with

loadings 0.9 for the second and 0.1 for the third aspect turned out to have similar loading to 86 singly classified articles from biophysics, biochemistry, cell biology, developmental biology, evolution, genetics, immunology, medical sciences, and microbiology. In the last column of Table 2, we give the numbers of projected additional dual classification papers by PNAS topic.

4.3 An alternative approach with related data

Griffiths and Steyvers (2004) use a related version of the mixed membership model on the words in PNAS abstracts for the years 1991-2001, involving 28,154 abstracts. Their corpus involves 20,551 words that occur in at least five abstracts, and are not on the "stop list". Their version of the model does not involve the full hierarchical probability structure. In particular, they employ Dirichlet(α) distribution for membership scores λ, but they fix α at $50/K$, and a Dirichlet(β) distribution for aspect word probabilities, but they fix β at 0.1. These choices lead to considerable computational simplification that allows using a Gibbs sampler for the Monte Carlo computation of marginal components of the posterior distribution.

In Griffiths and Steyvers (2004) they report on estimates of $Pr(data|K)$ for K= 50, 100, 200, 300, 400, 500, 600, 1000, integrating out the latent variable values. They then pick K to maximize this probability. This is referred to in the literature as a maximum *a posteriori* (MAP) estimate (e.g., see Denison et al. (2002)), and it produces a value of K approximately equal to 300, more than an order of magnitude greater than our value of $K = 8$.

There are many obvious and some more subtle differences between our data and those analyzed by Griffiths and Steyvers as well as between our approaches. Their approach differs from ours because of the use of a words-only model, as well as through the simplification involving the fixing of the Dirichlet parameters and through a more formal selection of dimensionality. While we can not claim that a rigorous model selection procedure would estimate the number of internal categories close to 8, we believe that a high number such as $K = 300$ is at least in part an artifact of the data and analytic choices made by Griffiths and Steyvers. For example, we expect that using the class of Dirichlet distributions with parameters $50/K$ when $K > 50$ for membership scores biases the results towards favoring many more categories than there are in the data due to increasingly strong preferences towards extreme membership scores with increasing K. Moreover, the use of the MAP estimate of K has buried within it an untenable assumption, namely that $Pr(K)$ constant *a priori*, and pays no penalty for an excessively large number of aspects.

4.4 Choosing K to describe PNAS topics

Although the analyses in the two preceding subsections share the same general goal, i.e., detecting the underlying structure of PNAS research publica-

tions, they emphasize two different levels of focus. For the analysis of words and references in Erosheva et al. (2004), we aimed to provide a succinct high-level summary of the population of research articles. This led us to narrow our focus to research reports in biology and to keep the numbers of topics within the range of the current classification scheme. We found the results for $K = 8$ aspects were more easily interpretable than those for $K = 10$ but because of time and computational expense we did not explore more fully the choice of K.

For their word-only model, Griffiths and Steyvers (2004) selected the model based on $K = 300$ which seems to be aimed more at the level of prediction, e.g., obtaining the most detailed description for each paper as possible. They worked with a database of all PNAS publications for given years and considered no penalty for using a large number of aspects such as that which would be associated with the Bayesian Information Criterion applied to the marginal distributions integrating out the latent variables.

Organizing aspects hierarchically, with sub-aspects having mixed membership in aspects, might allow us to reconcile our higher level topic choices with their more fine-grained approach.

5 Summary and concluding remarks

In this paper we have described a Bayesian approach to a general mixed membership model that allows for:

- Identification of internal clustering categories (unsupervised learning).
- Soft or mixed clustering and classifications.
- Combination of types of characteristics, e.g., numerical values and categories, words and references for documents, features from images.

The ideas behind the general model are simple but they allow us to view seemingly disparate developments in soft clustering or classification problems in diverse fields of application within the same broad framework. This unification has at least two saluatory implications:

- Developments and computational methods from one domain can be imported to or shared with another.
- New applications can build on the diverse developments and utilize the general framework instead of beginning from scratch.

When the GoM model was first developed, there were a variety of impediments to its implementation with large datasets, but the most notable were technical issues of model identifiability and consistency of estimation, since the number of parameters in the model for even a modest number of groups (facets) is typically greater than the number of observations, as well as possible multi-modal likelihood functions even when the model was properly identified. These technical issues led to to practical computational problems

and concerns about the convergence of algorithms. The Bayesian hierarchical formulation described here allows for solutions to a number of these difficulties, even in high dimensions, as long as we are willing to make some simplifying assumptions and approximations. Many challenges remain, both statistical and computational. These include computational approaches to full posterior calculations; model selection (i.e., choosing K), and the development of extensions of the model to allow for both hierarchically structured latent categories and dependencies associated with longitudinal structure.

Acknowledgments

We are indebted to John Lafferty for his collaboration on the analysis of the PNAS data which we report here, and to Adrian Raftery and Christian Robert for helpful discussions on selecting K. Erosheva's work was supported by NIH grants 1 RO1 AG023141-01 and R01 CA94212-01, Fienberg's work was supported by NIH grant 1 RO1 AG023141-01 and by the Centre de Recherche en Economie et Statistique of the Institut National de la Statistique et des Études Économiques, Paris, France.

References

BARNARD, K., DUYGULU, P., FORSYTH, D., de FREITAS, N., BLEI, D. M. and JORDAN, M. I. (2003): Matching words and pictures. *Journal of Machine Learning Research, 3, 1107–1135.*

BLEI, D. M. and JORDAN, M. I. (2003a): Modeling annotated data. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, 127–134.

BLEI, D. M., JORDAN, M. I. and NG, A. Y. (2003b): Latent Dirichlet models for application in information retrieval. In J. Bernardo, et al. eds., *Bayesian Statistics 7. Proceedings of the Seventh Valencia International Meeting*, Oxford University Press, Oxford, 25–44.

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003c): Latent Dirichlet allocation. *Journal of Machine Learning Research, 3, 993–1002.*

BRANDTBERG, T. (2002): Individual tree-based species classification in high spatial resolution aerial images of forests using fuzzy sets. *Fuzzy Sets and Systems, 132, 371–387.*

COHN, D. and HOFMANN, T. (2001): The missing link: A probabilistic model of document content and hypertext connectivity. *Neural Information Processing Systems (NIPS*13)*, MIT Press .

COOIL, B. and VARKI, S. (2003): Using the conditional Grade-of-Membership model to assess judgment accuracy. *Psychometrika, 68, 453–471.*

DENISON, D.G.T., HOLMES, C.C., MALLICK, B.K., and SMITH, A.F.M. (2002: *Bayesian Methods for Nonlinear Classification and Regression.* Wiley, New York.

EROSHEVA, E. A. (2002): *Grade of Membership and Latent Structure Models With Applicsation to Disability Survey Data.* Ph.D. Dissertation, Department of Statistics, Carnegie Mellon University. PhD thesis, Carnegie Mellon University.

EROSHEVA, E. A. (2003a): Bayesian estimation of the Grade of Membership Model. In J. Bernardo et al. (Eds.): *Bayesian Statistics 7. Proceedings of the Seventh Valencia International Meeting*, Oxford University Press, Oxford, 501–510.

EROSHEVA, E. A. (2003b): Partial Membership Models With Application to Disability Survey Data In H. Bozdogan (Ed.): *New Frontiers of Statistical Data Mining, Knowledge Discovery, and E-Business*, CRC Press, Boca Raton, 117–134.

EROSHEVA, E.A., FIENBERG, S.E. and LAFFERTY, J. (2004): Mixed Membership Models of Scientific Publications. *Proceedings of the National Academy of Sciences*, in press.

GRIFFITHS, T. L. and STEYVERS, M. (2004): Finding scientific topics. *Proceedings of the National Academy of Sciences*, in press.

HOFMANN, T. (2001): Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42, 177–196.*

KOVTUN, M., AKUSHEVICH, I., MANTON, K.G. and TOLLEY, H.D. (2004a): Grade of membership analysis: Newest development with application to National Long Term Care Survey. Unpublished paper presented at Annual Meeting of Population Association of America (dated March 18, 2004).

KOVTUN, M., AKUSHEVICH, I., MANTON, K.G. and TOLLEY, H.D. (2004b): Grade of membership analysis: One possible approach to foundations. Unpublished manuscript.

MANTON, K. G., WOODBURY, M. A. and TOLLEY, H. D. (1994): *Statistical Applications Using Fuzzy Sets*. Wiley, New York.

MINKA, T. P. and LAFFERTY, J., (2002): Expectation-propagation for the generative aspect model. *Uncertainty in Artificial Intelligence: Proceedings of the Eighteenth Conference (UAI-2002)*, Morgan Kaufmann, San Francisco, 352–359.

NURMBERG, H.G., WOODBURY, M.A. and BOGENSCHUTZ, M.P. (1999): A mathematical typology analysis of DSM-III-R personality disorder classification: grade of membership technique. *Compr Psychiatry, 40, 61–71.*

POTTHOFF, R. F., MANTON, K. G. and WOODBURY, M. A., (2000): Dirichlet generalizations of latent-class models. *Journal of Classification, 17, 315–353.*

PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P., (2000): Inference of population structure using multilocus genotype data. *Genetics, 155, 945–959.*

ROSENBERG, N. A., PRITCHARD, J. K., WEBER, J. L., CANN, H. M., KIDD, K. K., ZHIVOTOVSKY, L. A. and FELDMAN, M. W. (2002): Genetic structure of human populations. *Science, 298, 2381–2385.*

SEETHARAMAN, P.B., FEINBERG, F.M. and CHINTGUNTA, P.K. (2001): Product line management as dynamic, attribute-level competition. Unpublished manuscript.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, Methodological, 64, 1–34.*

TALBOT, B.G., WHITEHEAD, B.B. and TALBOT, L.M. (2002): Metric Estimation via a Fuzzy Grade-of-Membership Model Applied to Analysis of Business Opportunities. *14th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2002*, 431–437.

TALBOT, L.M. (1996): A Statistical Fuzzy Grade-of-Membership Approach to Unsupervised Data Clustering with Application to Remote Sensing. Unpublished Ph.D. dissertation, Department of Electrical and Computer Engineering, Brigham Young University.

VARKI, S. and CHINTAGUNTA, K. (2003): The augmented latent class model: Incorporating additional heterogeneity in the latent class model for panel data. *Journal of Marketing Research, forthcoming.*

VARKI, S., COOIL, B. and RUST, R.T. (2000): Modeling Fuzzy Data in Qualitative Marketing Research. *Journal of Marketing Research, XXXVII, 480–489.*

WOODBURY, M. A. and CLIVE, J. (1974): Clinical pure types as a fuzzy partition. *Journal of Cybernetics, 4, 111–121.*

WOODBURY, M. A., CLIVE, J. and GARSON, A. (1978): Mathematical typology: A Grade of Membership technique for obtaining disease definition. *Computers and Biomedical Research, 11, 277–298.*

Predicting Protein Secondary Structure with Markov Models

Paul Fischer, Simon Larsen, and Claus Thomsen

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Kongens Lyngby

Abstract. The *primary structure* of a protein is the sequence of its amino acids. The *secondary structure* describes structural properties of the molecule such as which parts of it form sheets, helices or coils. Spacial and other properties are described by the higher order structures. The classification task we are considering here, is to predict the secondary structure from the primary one. To this end we train a Markov model on training data and then use it to classify parts of unknown protein sequences as sheets, helices or coils. We show how to exploit the directional information contained in the Markov model for this task. Classifications that are purely based on statistical models might not always be biologically meaningful. We present combinatorial methods to incorporate biological background knowledge to enhance the prediction performance.

1 Introduction

The primary structure of a DNA-sequence is given by the sequence of its amino-acids. The secondary structure is a classification of contiguous stretches of a DNA-molecule according to their conformation. We use a threefold classification, namely the conformation *helices*, *sheets*, and *coils*. Most databases contain a finer classification into 6 or more classes. We use the mapping from Garnier et al. (1996) and Kloczkowski et al. (2002) to reduce to the three aforementioned classes.

The task is to determine the secondary structure from the primary one. We use a supervised learning approach for this purpose. From a database one collects a number of DNA-sequences for which the classifications are known. On these a (statistical) model is trained and then used to assign classifications to new, unclassified protein sequences. There is a number of such classifiers which are based on different learning concepts. Some use statistical methods like, e.g., the GOR algorithm, Garnier et al. (1996) and Kloczkowski et al. (2002). GOR are the initials of the authors of the first version of this method: Garnier, Osguthorpe, and Robson. Other algorithms rely on neural networks like PHD, Rost and Sander (1993) and (1994). The acronym means "Profile network from HD", where HD is the number plate code for Heidelberg, Germany, where the authors worked. Most of them incorporate biological background knowledge at some stage. For example a first classification given

by a statistical model is then checked for biological plausibility and, if necessary, corrected.

We use a first order Markov model as classifier. This type of classifier has been successfully used before in a related setting, Brunnert et al. (2002). There, the order and length of the helix and sheet subsequences was given (but no information on the intermediate coil parts was known). Here, we investigate how this classifier performs without the additional information on order and length and how its performance can be improved. The aim is to push the basic statistical method to its limits before combining it with other techniques. We use the GOR algorithms as references. They have been re-implemented without the incorporation of background knowledge.

2 The method

Let Σ_a denote the alphabet for the 20 amino acids, and let $\Sigma_c = \{H, E, C\}$ denote the classification alphabet, where H denotes helix, E sheet, and C coil. In the following let $\mathbf{x} = x_1, \ldots, x_n$ be a protein sequence, where $x_i \in \Sigma_a$. Let $\|\mathbf{x}\|$ denote its length. Let $C = c_1, \ldots, c_n$ be the corresponding classification sequence, $c_i \in \Sigma_c$.

We shall use a first order Markov model for the prediction. The model uses a parameter ℓ, the *window size*. Such a model assigns probabilities p to subsequences of $\mathbf{x}$ of length l as follows:

$$p(x_i, \ldots, x_{i+\ell-1}) = p(x_i)\, p(x_{i+1} \mid x_i) \cdots p(x_{i+\ell-1} \mid x_{i+\ell-2}) \tag{1}$$

For the threefold classification task we have in mind, three such models are used, one for each of the classes $\{H, E, C\}$. The probability functions of the respective models are denoted by p_H, p_E, and p_C. The three models are trained by estimating their parameters of the kinds $p_X(\cdot)$ and $p_X(\cdot \mid \cdot)$ $X \in \{H, E, C\}$. Then they can be used for classification of new sequences as follows: One evaluates all three models and then assigns that class which corresponds to the model with highest probability: $\arg\max\{p_H, p_E, p_C\}$. The obvious problem with this approach is, that a Markov model assign probabilities to subsequences (windows) and not to individual amino acids. This might lead to conflicting predictions. If, for example, $E = \arg\max_X\{p_X(x_1, \ldots, x_{\ell-1})\}$ and $H = \arg\max_X\{p_X(x_2, \ldots, x_\ell)\}$, it is not clear which of the two classifications x_2 should get. We choose to assign the classification of a window to the first amino acid in that window. The estimation of the model parameters is then performed to support this choice. We denote this by using the term $p(i)$ for this, i.e.,

$$p_X(i) = p_X(x_i)\, p_X(x_{i+1} \mid x_i) \cdots p_X(x_{i+\ell-1} \mid x_{i+\ell-2}) \tag{2}$$

We investigated several modifications of the Markov model, some of which also differ in the training process. The basic training is conducted as follows.

The training data consists of N DNA-sequences $\mathbf{x}^{(j)}$ and the corresponding classification sequences $\mathbf{c}^{(j)}$, $j = 1, \ldots, N$. Now, three sets of subsequences are constructed, one for each of the three classes. Each DNA-sequence $\mathbf{x}^{(j)}$ is divided into maximal substrings according to the classification $\mathbf{c}^{(j)}$: Let $x_k^{(j)} x_{k+1}^{(j)} \cdots x_{k+\ell-1}^{(j)}$ be such a subsequence. Then $c_k^{(j)} = c_{k+1}^{(j)} = \cdots = c_{k+\ell-1}^{(j)}$ and either $k = 1$ or $c_{k-1}^{(j)} \neq c_k^{(j)}$ and either $k + \ell - 1 = \|\mathbf{x}\|$ or $c_{k+\ell-1}^{(j)} \neq c_{k+\ell}^{(j)}$. When we use the term *subsequence* in the following we mean such a maximal subsequence. We denote the three collections of subsequences by $\mathcal{S}_H$, $\mathcal{S}_E$, and $\mathcal{S}_C$. On each of these sets a Markov model is trained by estimating its parameters. Let $\mathcal{M}_H$, $\mathcal{M}_E$, and $\mathcal{M}_C$ be the respective models. The estimations are the relative frequencies of (pairs of) residues in the training data. To avoid zero empirical probabilities, we introduce a pseudocount value $c \geq 0$, where $c = 0$ is the estimation without pseudocounts. Let $X \in \{H, E, C\}$ be the class and let $\mathbf{y}^{(j)}$ denote the maximal subsequences. Then the estimations are

$$p_X(a) := \frac{c + \left|\{j \mid \mathbf{y}^{(j)} \in \mathcal{S}_X \wedge y_1^{(j)} = a\}\right|}{|\Sigma_a|\, c + |\mathcal{S}_X|} \tag{3}$$

$$p_X(b|a) := \frac{c + \left|\left\{(i,j) \mid \mathbf{y}^{(j)} \in \mathcal{S}_X \wedge 1 < i \leq \left\|\mathbf{y}^{(j)}\right\| \wedge y_i^{(j)} = a \wedge y_{i-1}^{(j)} = b\right\}\right|}{\left|\left\{(i,j) \mid \mathbf{y}^{(j)} \in \mathcal{S}_X \wedge 1 < i \leq \left\|\mathbf{y}^{(j)}\right\| \wedge y_i^{(j)} = a\right\}\right| + |\Sigma_a|\, c} \tag{4}$$

3 Improvements

The basic classification method described above has been analysed and modified in order to detect the importance of the various parameters and to improve its performance. The tests were carried out while maintaining the statistical nature of the approach. No biological background knowledge was incorporated. Also, the method was not combined with other techniques. The aim was to push the performance of the basic method as far as possible before applying other techniques. In the following we describe the modifications and their influence on the performance.

The results shown here come from tests performed in Larsen and Thomsen (2004) on the GOR data set (Garnier et al. (1996)), which consists of 267 protein sequences. It was evaluated using a leave-one-out cross-validation. We also used the benchmark data set of 513 protein sequences. The results on the latter set showed no relevant difference to those on the GOR data set. Due to the structure of the Markov model with window size ℓ, the last $\ell - 1$ residuals of a sequence cannot be classified. The percent figures thus are the ratios of correctly classified residuals and all classified residuals.

Pseudocount and window size: These two parameters have been varied independently. The window size parameter ℓ is the number of terms used in the Markov expansion (1). The range for the window size was 1 through 10.

One would expect that a very small window size results in bad performance, because too few information is used in the classification process. Also very large window sizes should decrease the performance because the local information is blurred by far off data. The pseudocount parameter c was varied from 0 through 1000. The effect of this parameter depends on the size of the training set. In our case the set was so large, that no zero empirical probabilities occurred. Nevertheless, the performance of the classifier was improved when using small positive pseudocount values. We believe that this is due to the fact, that statistical fluctuation in small (unprecisely estimated) empirical probabilities are leveled by this.

The optimal choice of the parameters was a window size of 5 and a pseudocount value of 5. These settings were used in all following results. We also varied the window size and pseudocount constant in combination with other modifications but the aforementioned values stayed optimal. Figure 1 shows a plot of the test results. With this choice, the basic model has a classification rate (number of correctly classified residuals) of 51.0%. The naive classification – constantly predicting the most frequent residual (coil) – would give 43%.

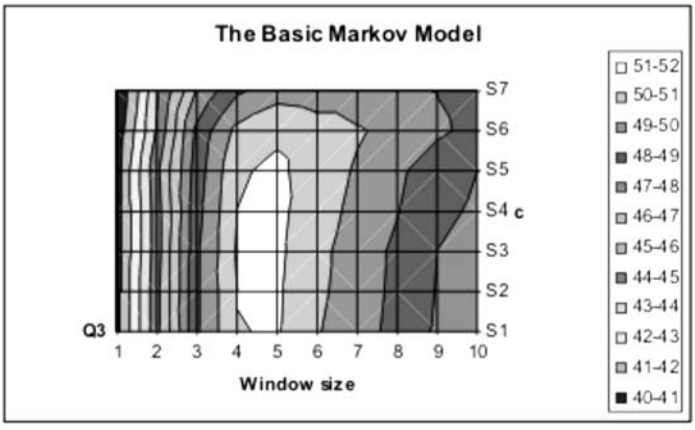


Fig. 1. A contour plot of the prediction performance of the Markov model as a function of the window size and the pseudocount constant. The vertical axis is from $S1 = 0$ to $S7 = 15$

Estimation of $p_X(a)$: In Equation (3) the parameter $p_X(a)$ was estimated as the empirical frequency of a as a first letter of a maximal subsequence with classification X. This definition stems from the application in Brunnert et al. (2002) where additional knowledge on the structure of the

subsequences (length/order) was available. We changed the estimation (3) to

$$p'_X(a) := \frac{c + \left|\{(i,j) \mid \mathbf{y}^{(j)} \in \mathcal{S}_X \wedge y_i^{(j)} = a\}\right|}{|\Sigma_a|\, c + \sum_j \left|\mathbf{y}^{(j)}\right|}\,, \tag{5}$$

the frequency of the letter a in all subsequences with classification X. Using this definition improved the classification performance by 1.4 percentage points. The increase was expected, because the information on residuals in the middle of the subsequence is increased.

Estimation of $p(a|b)$: Instead of using Equation (4) to estimate the conditional probabilities, we also considered the reversed sequence. That is we computed $p^{forw}(a|b)$ as in Equation (4) and $p^{rev}(a|b)$ as in Equation (4) but on the reversed sequence. Then we set $p(a|b)$ to the sum of $p^{fore}(a|b)$ and $p^{rev}(a|b)$ and normalize to get a probability distribution. Using this definition of $p(a|b)$ improved the prediction performance by 2 percentage points.

Direction: Markov models exploit directional information. We therefore tried another modification, namely to reverse the sequences in the training and the classification process. We did not expect a significant increase from this. To our surprise the classification performance was increased by 1.5 percentage points when using reversed sequences. This indicates that the sequence data is more informative in one direction than in the other one.

Momentum: This variation of the basic method tries to achieve a more "stable" classification as the classification window moves along the DNA-sequence. To this end we consider the discounted values of previous classifications. The new classification value, denoted by $p'_X(i)$, replaces the original values $p_X(i)$ from (2) and is defined by

$$\begin{aligned} p'_X(1) &= p_X(1) \\ p'_X(i) &= w \cdot p'_X(i-1) + (1-w) \cdot p_X(i) \end{aligned}$$

To determine a good value for the discount constant w, different settings of $w \in [0,1]$ were tested. The choice of $w = 0.5$ showed the best results with an increase of 4.3 percentage points in the prediction performance.

One can say that the use of a momentum term does model some biological knowledge. It is known that helix, coil, or sheet subsequences usually consist of a number of amino acids, not just a single one. The momentum method eliminates a number of subsequences of length one from the prediction. This often replaces the old prediction by the correct one, which results in the better performance.

Combination of methods: A number of combinations of the above methods were tested. Combining the definition given in Equation (5) for $p(a)$, the momentum and the modified definition of $p(a|b)$ proved to be the most successful one. It resulted in the considerable increase of the prediction performance of 6.3 percentage points resulting in 57.3%.

Our implementations of the GOR algorithm versions I, III and IV, all without the incorporation of background knowledge and with window size 17, gave classification rates of 60.7%, 59.6%, and 63.4%. It is not surprising that the GOR-algorithms outperform the Markov approach, as it uses a statistic of all pairs in the window. It is however surprising, how close one can come to versions I and III of GOR.

4 Ongoing research

We are currently considering "peaks" of the probabilities. The idea of using the concept of a peak is motivated by the shapes of the graphs of the three probability functions $p_E(i)$, $p_H(i)$, and $p_C(i)$. Often the function p_X has a peak at the first residuum of a X-subsequence. See Figure 2 for an example. The peaks are more prominent when using the original definition (3) of the term $p(a)$ than that given in (5). A peak could be used as indicator of the start of a new subsequence. Then the corresponding classification is maintained until a peak of another probability function is found.

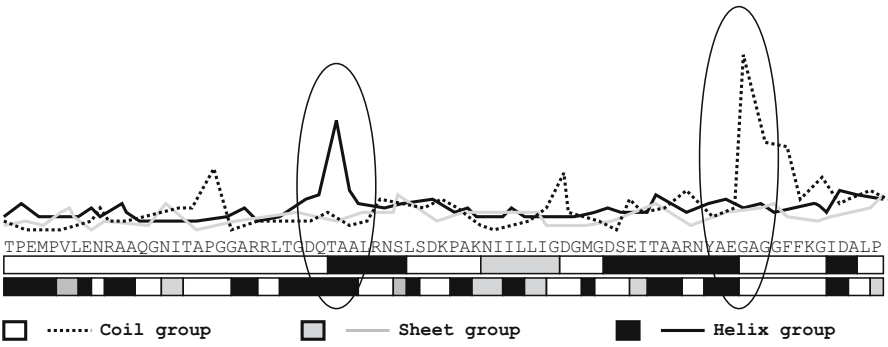


Fig. 2. The picture shows the probability functions for the three classes. Two peaks at the left start points of subsequences are marked by ovals. Below the graphs is the protein sequence with the correct classifications shown by colors. The colors in the line below show the predictions of the Markov model.

The problem here is to find an appropriate combinatorial definition of the term "peak". The absolute value of the functions p_X cannot be used due to their strong variation. Also, a peak of one function, say p_E, does not necessarily exceed the values of the two other functions. On the other hand, a peak value of p_E should not be ridiculously small relative to the two other functions.

First tests with a simple definition of a peak show that using this concept as a start indicator only gives an improvement of 4 percentage points over the naive classification leading to 47%. The plan is to incorporate peak indicators into the Markov method (or other prediction methods). One way of

doing this is to compare the peak locations with a prediction given by some other method. Then the alignment of a peak with the start of a predicted subsequence would raise our confidence in the prediction. If a predicted subsequence does not coincide with a peak, then the prediction at this location should be checked.

5 Summary

We have significantly improved a simple statistical prediction method by a thorough analysis of the influence of its different components. Now, the next step is to incorporate biological background knowledge into the classification process and to combine the Markov predictor with other classifiers. The investigations also exposed the "peak" concept as a promising alternative for using the statistical information.

References

BRUNNERT, M., FISCHER, P. and URFER, W. (2002): Sequence-structure alignment using a statistical analysis of core models and dynamic programming. Technical report, SFB 475, Universität Dortmund.

GARNIER, J., GIBRAT, J.-F. and ROBSON, B. (1996): GOR method for predicting protein secondary structure from amino acid sequence. *Methods in Enzymology, 266, 540–553.*

KLOCZKOWSKI, A., TING, K.L., JERNIGAN, R.L. and GARNIER, J (2002): Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from aminoacid sequence. *Proteins, 49, 154–166.*

LARSEN, S. and THOMSEN, C. (2004): Classification of protein sequences using Markov models, Masters thesis, Informatics and Mathematical Modelling. Technical University of Denmark.

ROST, B. and SANDER, C. (1993): Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol., 232, 584–599.*

ROST, B. and SANDER, C. (1994): Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins, 19, 55–72.*

Milestones in the History of Data Visualization: A Case Study in Statistical Historiography

Michael Friendly

Psychology Department
York University, Toronto, Canada
friendly@yorku.ca

Abstract. The Milestones Project is a comprehensive attempt to collect, document, illustrate, and interpret the historical developments leading to modern data visualization and visual thinking. This paper provides an overview and brief tour of the milestones content, with a few illustrations of significant contributions to the history of data visualization. This forms one basis for exploring interesting questions and problems in the use of statistical and graphical methods to explore this history, a topic that can be called "statistical historiography."

1 Introduction

The only new thing in the world is the history you don't know.—Harry S Truman

The graphic portrayal of quantitative information has deep roots. These roots reach into the histories of the earliest map-making and visual depiction, and later into thematic cartography, statistics and statistical graphics, medicine, and other fields, which are intertwined with each other. They also connect with the rise of statistical thinking and widespread data collection for planning and commerce up through the 19th century. Along the way, a variety of advancements contributed to the widespread use of data visualization today. These include technologies for drawing and reproducing images, advances in mathematics and statistics, and new developments in data collection, empirical observation and recording.

From above ground, we can see the current fruit; we must look below to understand their germination. Yet the great variety of roots and nutrients across these domains, that gave rise to the many branches we see today, are often not well known, and have never been assembled in a single garden, to be studied or admired.

The Milestones Project is designed to provide a broadly comprehensive and representative catalog of important developments in *all* fields related to the history of data visualization. Toward this end, a large collection of images, bibliographical references, cross-references and web links to commentaries on these innovations has been assembled.

This is a useful contribution in its own right, but is a step towards larger goals as well. First, we see this not as a static collection, but rather a dynamic database that will grow over time as additional sources and historical contributions are uncovered or suggested to us. Second, we envisage this project as providing a tool to enable researchers to work with or study this history, finding themes, antecedents, influences, patterns, trends, and so forth. Finally, as implied by our title, work on this project has suggested several interesting questions subsumed under the self-referential term "statistical historiography."

1.1 The *Milestones Project*

The past only exists insofar as it is present in the records of today. And what those records are is determined by what questions we ask.—(Wheeler (1982), p. 24)

There are many historical accounts of developments within the fields of probability (Hald (1990)), statistics (Pearson (1978), Porter (1986), Stigler (1986)), astronomy (Riddell (1980)), cartography (Wallis and Robinson (1987)), which relate to, *inter alia*, some of the important developments contributing to modern data visualization. There are other, more specialized accounts, which focus on the early history of graphic recording (Hoff and Geddes (1959), Hoff and Geddes (1962)), statistical graphs (Funkhouser (1936), Funkhouser (1937), Royston (1970), Tilling (1975)), fitting equations to empirical data (Farebrother (1999)), cartography (Friis (1974), Kruskal (1977)) and thematic mapping (Palsky (1996), Robinson (1982)), and so forth; (Robinson (1982, Ch. 2)) presents an excellent overview of some of the important scientific, intellectual, and technical developments of the 15^{th}–18^{th} centuries leading to thematic cartography and statistical thinking.

But there are no accounts that span the entire development of visual thinking and the visual representation of data, and which collate the contributions of disparate disciplines. In as much as their histories are intertwined, so too should be any telling of the development of data visualization. Another reason for interweaving these accounts is that practitioners in these fields today tend to be highly specialized, often unaware of related developments in areas outside their domain, much less their history. Extending (Wheeler (1982)), the records of history also exist insofar as they are collected, illustrated, and made coherent.

The initial step in portraying the history of data visualization was a simple chronological listing of milestone items with capsule descriptions, bibliographic references, markers for date, person, place, and links to portraits, images, related sources or more detailed commentaries. Its current public and visible form is that of hyper-linked, interactive documents available on the web and in PDF form (`http://www.math.yorku.ca/SCS/Gallery/milestone/`). We started with the developments listed by (Beniger and Robyn (1978)) and incorporated additional listings from Hankins (1999)), Tufte

(1983), Tufte (1990), Tufte (1997)), (Heiser (2000)), and others. With assistance from *Les Chevaliers*, many other contributions, original sources, and images have been added. As explained below, our current goal is to turn this into a true multi-media database, which can be searched in flexible ways and can be treated as data for analysis.

2 Milestones tour

In organizing this material, it proved useful to divide history into epochs, each of which turned out to be describable by coherent themes and labels. In the larger picture— recounting the history of data visualization— each milestone item has a story to be told: What motivated this development? What was the communication goal? How does it relate to other developments? What were the pre-cursors? What makes it a milestone? To illustrate, we present just a few exemplars from a few of these periods. For brevity, we exclude the earliest period (pre-17^{th} century) and the most recent period (1975–present) in this description.

2.1 1600-1699: Measurement and theory

Among the most important problems of the 17th century were those concerned with physical measurement— of time, distance, and space— for astronomy, surveying, map making, navigation and territorial expansion. This century also saw great new growth in theory and the dawn of practice— the rise of analytic geometry, theories of errors of measurement and estimation, the birth of probability theory, and the beginnings of demographic statistics and "political arithmetic."

As an example, Figure 1 shows a 1644 graphic by Michael Florent van Langren, a Flemish astronomer to the court of Spain, believed to be the first visual representation of statistical data (Tufte (1997, p. 15)). At that time, lack of a reliable means to determine longitude at sea hindered navigation and exploration.[1] This 1D line graph shows all 12 known estimates of the difference in longitude between Toledo and Rome, and the name of the astronomer (Mercator, Tycho Brahe, Ptolemy, etc.) who provided each observation.

What is notable is that van Langren could have presented this information in various tables— ordered by author to show provenance, by date to show priority, or by distance. However, only a graph shows the wide variation in the estimates; note that the range of values covers nearly half the length of the scale. Van Langren took as his overall summary the center of the range, where there happened to be a large enough gap for him to inscribe "ROMA." Unfortunately, all of the estimates were biased upwards; the true distance ($16°30'$) is shown by the arrow. Van Langren's graph is also a milestone

[1] For navigation, latitude could be fixed from star inclinations, but longitude required accurate measurement of time at sea, an unsolved problem until 1765.

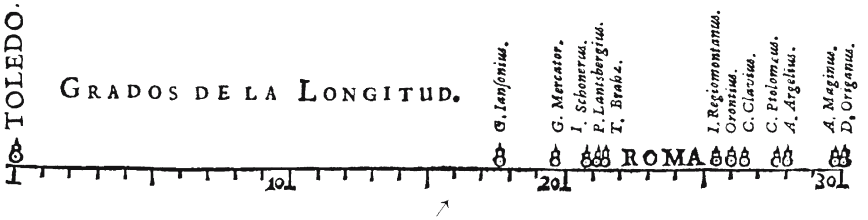


Fig. 1. Langren's 1644 graph of determinations of the distance, in longitude, from Toledo to Rome. The correct distance is 16°30′. Source: Tufte (1997, p.15.)

as the earliest-known exemplar of the principle of "effect ordering for data display" (Friendly and Kwan (2002)).

2.2 1700-1799: New graphic forms

The 18th century witnessed, and participated in, the initial germination of the seeds of visualization that had been planted earlier. Map-makers began to try to show more than just geographical position on a map. As a result, new graphic forms (isolines and contours) were invented, and thematic mapping of physical quantities took root. Towards the end of this century, we see the first attempts at the thematic mapping of geologic, economic, and medical data.

Abstract graphs, and graphs of functions were introduced, along with the early beginnings of statistical theory (measurement error) and systematic collection of empirical data. As other (economic and political) data began to be collected, some novel visual forms were invented to portray them, so the data could "speak to the eyes."

As well, several technological innovations provided necessary nutrients. These facilitated the reproduction of data images (color printing, lithography), while other developments eased the task of creating them. Yet, most of these new graphic forms appeared in publications with limited circulation, unlikely to attract wide attention.

William Playfair (1759–1823) is widely considered the inventor of most of the graphical forms widely used today— first the line graph and bar chart (Playfair (1786)), later the pie chart and circle graph (Playfair (1801)). A somewhat later graph (Playfair (1821)), shown in Figure 2, exemplifies the best that Playfair had to offer with these graphic forms. Playfair used three parallel time series to show the price of wheat, weekly wages, and reigning monarch over a ∼250 year span from 1565 to 1820, and used this graph to argue that workers had become better off in the most recent years.

2.3 1800-1850: Beginnings of modern graphics

With the fertilization provided by the previous innovations of design and technique, the first half of the 19th century witnessed explosive growth in

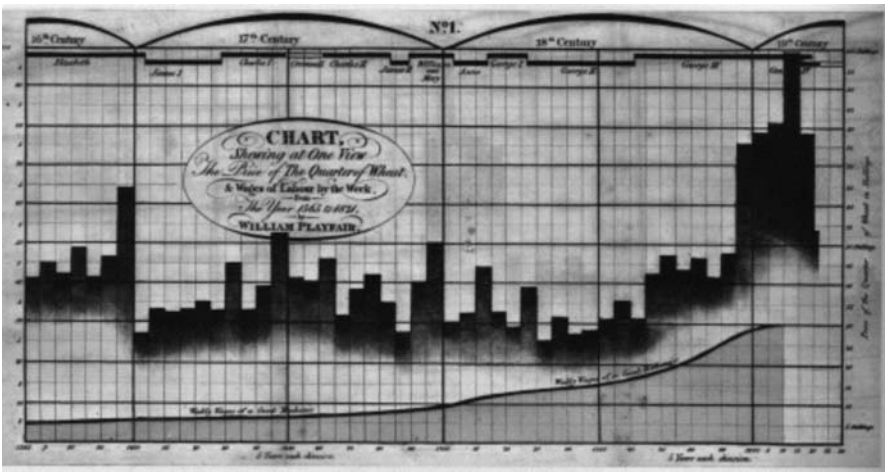


Fig. 2. William Playfair's 1821 time series graph of prices, wages, and ruling monarch over a 250 year period. *Source*: Playfair (1821), image from Tufte (1983, p. 34)

statistical graphics and thematic mapping, at a rate which would not be equalled until modern times.

In statistical graphics, all of the modern forms of data display were invented: bar and pie charts, histograms, line graphs and time-series plots, contour plots, scatterplots, and so forth. In thematic cartography, mapping progressed from single maps to comprehensive atlases, depicting data on a wide variety of topics (economic, social, moral, medical, physical, etc.), and introduced a wide range of novel forms of symbolism.

To illustrate this period, we choose an 1844 "tableau-graphique" (Figure 3) by Charles Joseph Minard, an early progenitor of the modern mosaic plot (Friendly (1994)). On the surface, mosaic plots descend from bar charts, but Minard introduced two simultaneous innovations: the use of divided and proportional-width bars so that area had a concrete visual interpretation. The graph shows the transportation of commercial goods along one canal route in France by variable-width, divided bars (Minard (1844)). In this display the width of each vertical bar shows distance along this route; the divided bar segments have height $\sim$ amount of goods of various types (shown by shading), so the area of each rectangular segment is proportional to cost of transport. Minard, a true visual engineer (Friendly (2000)), developed such diagrams to argue visually for setting differential price rates for partial vs. complete runs. Playfair had tried to make data "speak to the eyes," but Minard wished to make them "calculer par l'oeil" as well.

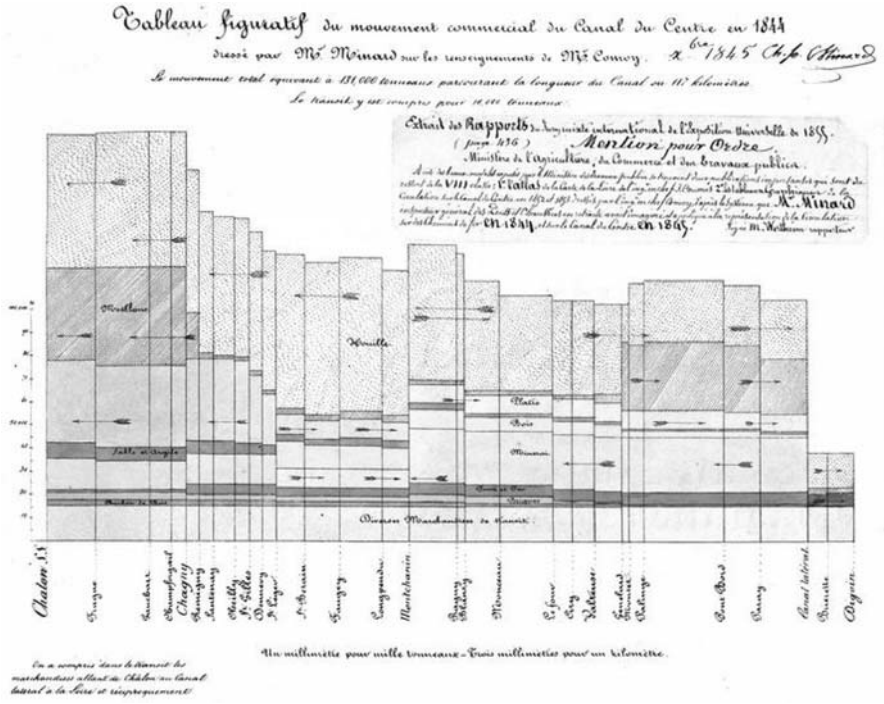


Fig. 3. Minard's Tableau Graphique, showing the transportation of commercial goods along the Canal du Centre (Chalon–Dijon). Intermediate stops are spaced by distance, and each bar is divided by type of goods, so the area of each tile represents the cost of transport. Arrows show the direction of transport. *Source:* ENPC:5860/C351 (Col. et cliché ENPC; used by permission)

2.4 1850-1900: The Golden Age of statistical graphics

By the mid-1800s, all the conditions for the rapid growth of visualization had been established. Official state statistical offices were established throughout Europe, in recognition of the growing importance of numerical information for social planning, industrialization, commerce, and transportation. Statistical theory, initiated by Gauss and Laplace, and extended to the social realm by Quetelet, provided the means to make sense of large bodies of data.

What started as the *Age of Enthusiasm* (Palsky (1996)) for graphics may also be called the *Golden Age*, with unparalleled beauty and many innovations in graphics and thematic cartography.

2.5 1900-1950: The modern dark ages

If the late 1800s were the "golden age" of statistical graphics and thematic cartography, the early 1900s could be called the "modern dark ages" of visualization (Friendly and Denis (2000)).

There were few graphical innovations, and, by the mid-1930s, the enthusiasm for visualization which characterized the late 1800s had been supplanted by the rise of quantification and formal, often statistical, models in the social sciences. Numbers, parameter estimates, and, especially, standard errors were precise. Pictures were— well, just pictures: pretty or evocative, perhaps, but incapable of stating a "fact" to three or more decimals. Or so it seemed to statisticians.

But it is equally fair to view this as a time of necessary dormancy, application, and popularization, rather than one of innovation. In this period statistical graphics became main stream. It entered textbooks, the curriculum, and standard use in government, commerce and science. In particular, perhaps for the first time, graphical methods proved crucial in a number of scientific discoveries (e.g. the discovery of atomic number by Henry Mosely, lawful clusterings of stars based on brightness and color in the Hertzprung-Russell diagrams; see Friendly and Denis (2004) for details.)

2.6 1950-1975: Re-birth of data visualization

Still under the influence of the formal and numerical zeitgeist from the mid-1930s on, data visualization began to rise from dormancy in the mid 1960s, spurred largely by three significant developments:

(a) In the USA, John W. Tukey began the invention of a wide variety of new, simple, and effective graphic displays, under the rubric of "Exploratory Data Analysis." (b) In France, Jacques Bertin published the monumental *Sémiologie Graphique* (Bertin (1967), Bertin (1983)). To some, this appeared to do for graphics what Mendeleev had done for the organization of the chemical elements, that is, to organize the visual and perceptual elements of graphics according to the features and relations in data. (c) Finally, computer processing of data had begun, and offered the possibility to construct old and new graphic forms by computer programs. True high-resolution graphics were developed, but would take a while to enter common use.

By the end of this period significant intersections and collaborations would begin: (a) computer science research (software tools, C language, UNIX, etc.) at Bell Laboratories (Becker (1994)) and elsewhere would combine forces with (b) developments in data analysis (EDA, psychometrics, etc.) and (c) display and input technology (pen plotters, graphic terminals, digitizer tablets, the mouse, etc.). These developments would provide new paradigms, languages and software packages for expressing statistical ideas and implementing data graphics. In turn, they would lead to an explosive growth in new visualization methods and techniques.

Other themes began to emerge, mostly as initial suggestions: (a) various visual representations of multivariate data (Andrews' plots, Chernoff faces, clustering and tree representations); (b) animations of a statistical process; and (c) perceptually-based theory (or just informed ideas) related to how

graphic attributes and relations might be rendered to better convey the data visually.

3 Problems and methods in statistical historiography

As we worked on assembling the Milestones collection, it became clear that there were several interesting questions and problems related to conducting historical research along these lines.

3.1 What counts as a Milestone?

In order to catalog the contributions to be considered as "milestones" in the history of data visualization, it is necessary to have some criteria for inclusion: for form, content, and substantive domain, as well as for "what counts" as a milestone in this context. We deal only with the last aspect here.

We have adopted the following scheme. First, we decided to consider several types of contributions as candidates: true innovations, important precursors and developments or extensions. Second, we have classified these contributions according to several themes, categories and rubrics for inclusion. Attributions without reference here are listed in the Milestones Project web documents.

- ***Contributions to the development and use of graphic forms***. In statistical graphics, inventions of the bar chart, pie chart, line plot (all attributed to Playfair), the scatterplot (attributed to J.F.W. Herschel; see Friendly and Denis (2004)), 3D plots (Luigi Perozzo), boxplot (J. W. Tukey), and mosaic plot (Hartigan & Kleiner) provided new ways of representing statistical data. In thematic cartography, isolines (Edmund Halley), choropleths (Charles Dupin) and flow maps (Henry Harness; C. J. Minard) considerably extended the use of a map-based display to show more than simple geographical positions and features.
- ***Graphic content: data collection and recording***. Visual displays of information cannot be done without empirical data, so we must also include contributions to measurement (geodesy), recording devices, collection and dissemination of statistical data (e.g., vital statistics, census, social, economic data).
- ***Technology and enablement***. It is evident that many developments had technological prerequisites, and conversely that new technology allowed new advances that could not have been achieved before. These include advances in (a) reproduction of printed materials (printing press, lithography), (b) imaging (photography, motion pictures), and (c) rendering (computing, video display).
- ***Theory and practice***. Under this heading we include theoretical advances in the treatment and analysis such as (a) probability theory and

notions of errors of measurement, (b) data summarization (estimation and modelling), (c) data exposure (e.g., EDA), as well as (d) awareness and use of these ideas and methods.

- ***Theory and data on perception of visual displays***. Graphic displays are designed to convey information to the human viewer, but how people use and understand this form of communication was not systematically studied until recent times. As well, proposals for graphical standards, and theoretical accounts of graphic elements and graphic forms provided a basis for thinking of and designing visual displays.
- ***Implementation and dissemination***. New techniques become *available* when they are introduced, but additional steps are needed to make them widely *accessible* and useable. We are thinking here mainly of implementations of graphical methods in software, but other contributions fall under this heading as well.

3.2 Who gets credit?

All of the Milestones items are attributed to specific individuals where we have reason to believe that names can be reasonably attached. Yet, *Stigler's Law of Eponomy* (Stigler (1980)) reminds us that standard attributions are often not those of priority. The Law in fact makes a stronger claim: "No scientific discovery is named after its original discoverer." As *prima facie* evidence, Stigler attributes the origin of this law to Merton (1973).

As illustrations, Stigler (1980) states that Laplace first discovered the Fourier transform, Poisson first discovered the Cauchy distribution, and both de Moivre and Laplace have prior claims to the Gaussian distribution. He concludes that epononyms are conveyed by the community of scholars, not by historians.

Thus, although all of the events listed are correctly attributed to their developers, it cannot be claimed with certainty that we are always identifying the first instance, nor that we give credit to all who have, perhaps independently, given rise to a new idea or method. Similarly, in recent times there may be some difficulty distinguishing credit among developers of (a) an underlying method or initial demonstration, (b) a corresponding algorithm, or (c) an available software implementation.

3.3 Dating milestones

In a similar way, there is some unavoidable uncertainty in the dates attached to milestone items, in a degree which generally increases as we go back in time. For example, in the 18^{th} and 19^{th} centuries, many papers were first read at scientific meetings, but recorded in print some years later; William Smith's geological map of England was first drawn in 1801, but only finished and published in 1815; some pre-1600 dates are only known approximately.

In textual accounts of history this does not present any problem— one can simply describe the circumstances and range of events, dated specifically or approximately, contributing to some development.

It does matter, however, if we wish to treat item dates as data, either for retrieval or analysis/display. For retrieval, we clearly want any date within a specified range to match; for analysis or display, the end points will sometimes be important, but sometimes it will suffice to use a middle value.

3.4 What is milestones "data"

The *Milestones Project* represents ongoing work. We continually update the web and pdf versions as we add items and images, many of which have been contributed by *Les Chevaliers*. To make this work, we rely on software tools to generate different versions from a single set of document sources, so that all versions can be updated automatically. For this, we chose to use LaTeX and BibTeX.

More recently, we have developed tools to translate this material to other forms (e.g., XML or CSV) in order to be able to work with it as "data." In doing so, it seemed natural to view the information as coming from three distinct sources, that we think of as a relational database, linked by unique keys in each, as shown in Figure 4.

3.5 Analyzing milestones "data"

Once the milestones data has been re-cast as a database, statistical analysis becomes possible. The simplest case is to look at trends over time. Figure 5 shows a density estimate for the distribution of milestones items from 1500 to the present, keyed to the labels for the periods in history. The bumps, peaks and troughs all seem interpretable: note particularly the steady rise up to $\sim$ 1880, followed by a decline through the "modern dark ages" to $\sim$ 1945, then the steep rise up to the present.

If we classify the items by place of development (Europe vs. North America), other interesting trends appear (Figure 6). The decline in Europe following the Golden Age was accompanied by an initial rise in North America, largely due to popularization (e.g., text books) and significant applications of graphical methods, then a steep decline as mathematical statistics held sway.

3.6 What was he thinking?: Understanding through reproduction

Historical graphs were created using available data, methods, technology, and understanding current at the time. We can often come to a better understanding of intellectual, scientific, and graphical questions by attempting a re-analysis from a modern perspective.

Earlier, we showed Playfair's time-series graph (Figure 2) of wages and prices, and noted that Playfair wished to show that workers were better off at

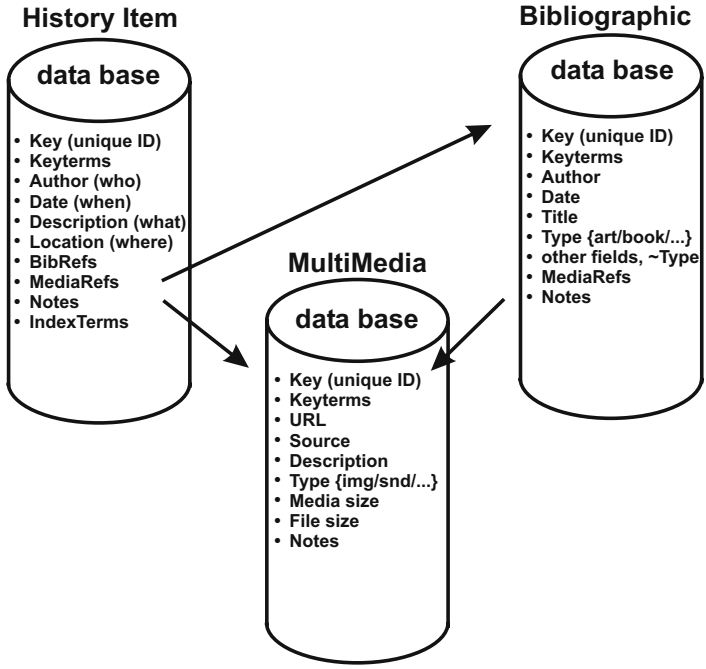


Fig. 4. Milestones data as a relational database composed of history-item, bibliographic, and multimedia databases

the end of the period shown than at any earlier time. Presumably he wished to draw the reader's eye to the narrowing of the gap between the bars for prices and the line graph for wages. Is this what you see?

What this graph shows directly is quite different than Playfair's intension. It appears that wages remained relatively stable, while the price of wheat varied greatly. The inference that wages increased relative to prices is indirect and not visually compelling.

We cannot resist the temptation to give Playfair a helping hand here—by graphing the ratio of wages to prices (labor cost of wheat), as shown in Figure 7. But this would not have occurred to Playfair, because the idea of relating one time series to another by ratios (index numbers) would not occur for another half-century (Jevons). See Friendly and Denis (2004) for further discussion of Playfair's thinking.

3.7 What kinds of tools are needed?

We have also wondered how other advances in statistics and data visualization could be imported to a historical realm. Among other topics, there has recently been a good deal of work in document analysis and classification that

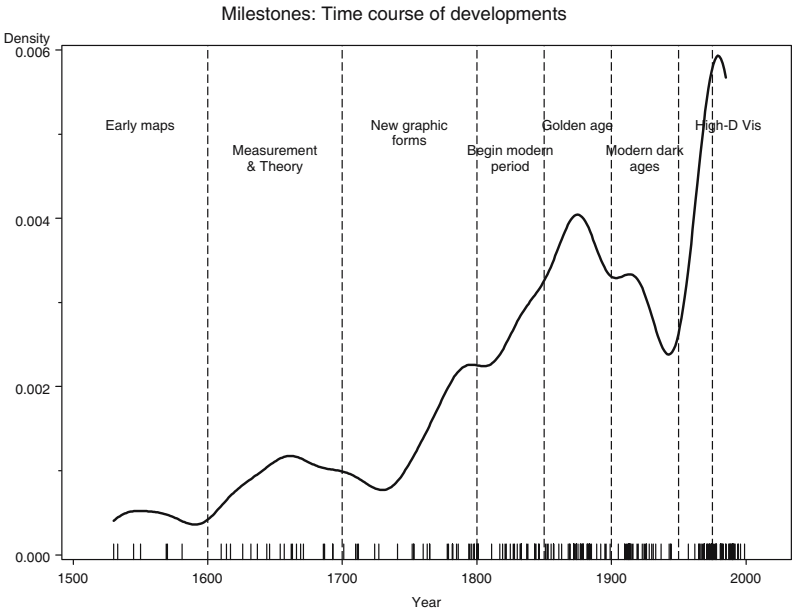


Fig. 5. The distribution of milestone items over time, shown by a rug plot and density estimate.

suggests an analog of EDA we might call Exploratory Bibliographic Analysis (EBA).

It turns out that there are several instances of software systems that provide some basic tools for this purpose. An example is RefViz (`http://www.refviz.com`), shown in Figure **??**. This software links to common bibliographic software (EndNote, ProCite, Reference Manager, etc.), codes references using key terms from the title and abstract and calculates an index of similarity between pairs of references based on frequencies of co-occurrence. Associations between documents can be shown in a color-coded matrix view, as in Figure **??**, or a galaxy view (combining cluster analysis and MDS), and each view offers zoom/unzoom, sorting by several criteria, and querying individual documents or collections.

4 How to visualize a history?

A timeline is obvious, but has severe limitations. We record a history of over 8000 years, but only the last 300-400 have substantial contributions. As well, a linear representation entails problems of display, resolution and access to detailed information, with little possibility to show either content or context. We explore a few ways to escape these constraints below.

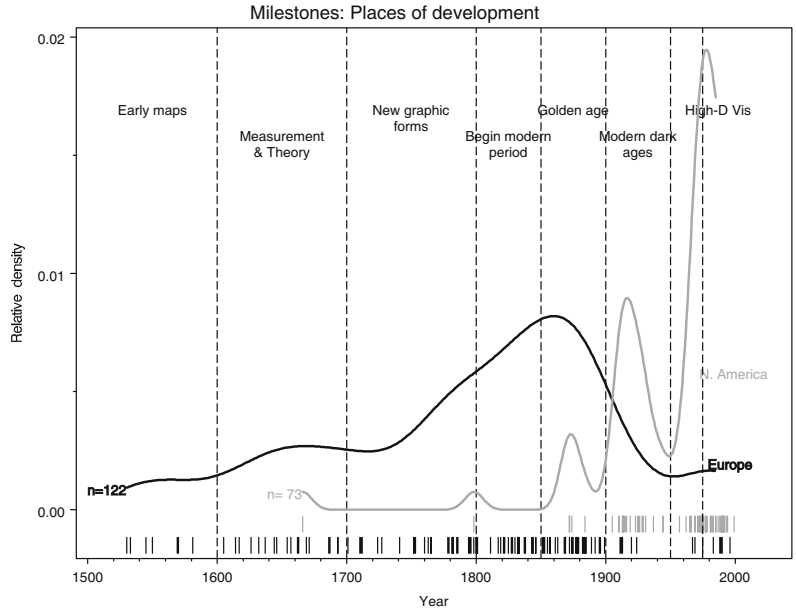


Fig. 6. The distribution of milestone items over time, comparing trends in Europe and North America.

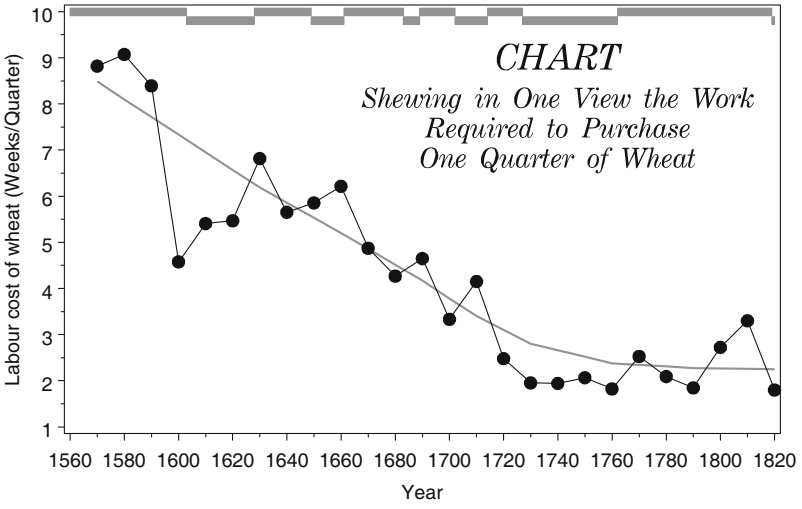


Fig. 7. Redrawn version of Playfair's time series graph showing the ratio of price of wheat to wages, together with a loess smoothed curve.

4.1 Lessons from the past

In the milestones collection, we have three examples of attempts to display a history visually. It is of interest that all three used essentially the same

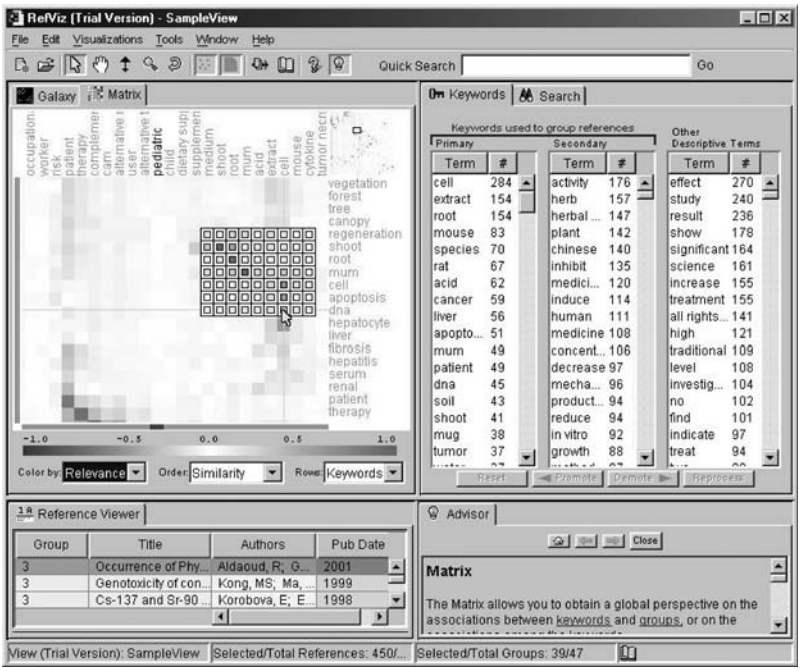


Fig. 8. RefViz similarity matrix view of a bibliographic database. The popup grid is a zoomed display of the region surrounding a selected cell.

format: a horizontal, linear scale for time, with different content or context stacked vertically, as separate horizontal bands.

We illustrate with Joseph Priestley's *Chart of Biography* (Priestley (1765)), showing the lifespans of famous people from 1200 BC to 1750 (Figure 9). Priestley divided people into two groups: 30 "men of learning" and 29 "statesmen," showing each lifespan as a horizontal line. He invented the convention of using dots to indicate uncertainty about exact date of birth or death.

4.2 Lessons from the present

In modern times, a variety of popular publications, mostly in poster form, have attempted to portray graphically various aspects of the history of civilization, geographic regions, or of culture and science.

For example, Hammond's *Graphic History of Mankind* (Figure 10) shows the emergence of new cultures and the rise and fall of various empires, nations and ethnic groups from the late Stone Age to the present in a vertical format. It uses a varying-resolution time scale, quite coarse in early history, getting progressively finer up to recent times. It portrays these using flow lines of different colors, whose width indicates the influence of that culture, and with shading or stripes to show conquest or outside influence.

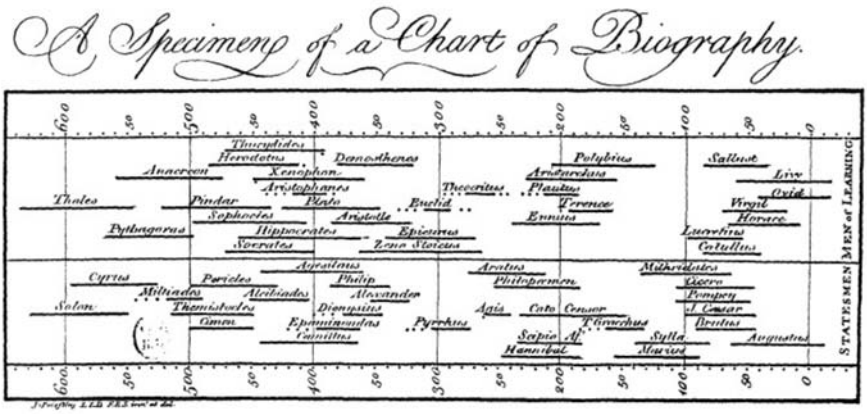


Fig. 9. Priestley's *Chart of Biography. Source*: Priestley (1765)

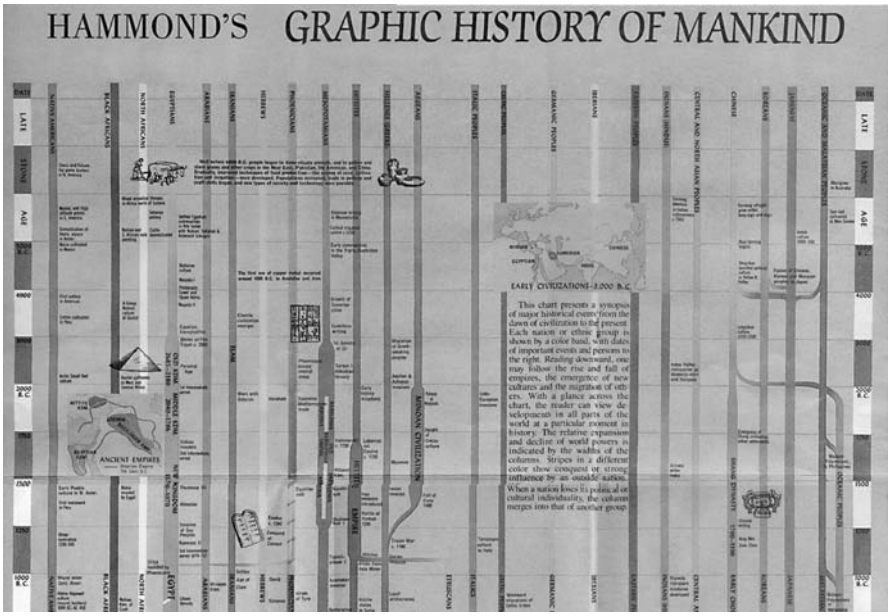


Fig. 10. Hammond's *Graphic History of Mankind* (first of 5 panels)

4.3 Lessons from the web

A large component of the milestones collection is the catalog of graphic images and portraits associated with the milestones items. At present, they are stored as image files of fixed resolution and size, and presented as hyper-links in the public versions. How can we do better, to make this material more easily accessible?

There are now a number of comprehensive image libraries available on the web that provide facilities to search for images by various criteria and in some cases to view these at varying resolutions. Among these, David Rumsey's Map Collection (`http://www.davidrumsey.com`) is notable. It provides access to a collection of over 8800 historical maps (mostly 18^{th}-19^{th} century, of North/South America, with some European content) online, extensively indexed so they may be searched by author, category, country or region, and a large number of other data fields. The maps are stored using Mr. Sid technology (`http://www.lizardtech.com`), which means that they can be zoomed and panned in real time. Rumsey provides several different browsers, including a highly interactive Java client.

4.4 Lessons from the data visualization

Modern data visualization also provides a number of different ideas and approaches to multivariate complexity, time and space we may adapt (in a self-referential way) to the history of data visualization itself.

Interactive viewers provide one simple solution to the trade-off between detail and scope of a data view through zoom and unzoom, but in the most basic implementation, any given view is a linear scaling of the section of the timeline that will fit within the given window.

We can do better by varying resolution *continuously* as a non-linear, decreasing function of distance from the viewer's point of focus. For example, Figure 11 shows a fisheye view (Furnas (1986)) of central Washington, D.C., using a hyperbolic scale, so that resolution is greatest at the center and decreases as 1/distance. The map is dynamic, so that moving the cursor changes the focal point of highest resolution, This has the property that it allows the viewer to see the context surrounding the point of focus, yet navigate smoothly throughout the entire space. Similar ideas have been applied to tables in the Table Lens (`http://www.tablelens.com`) and hierarchies (Lamping et al. (1995)) such as web sites and file systems, and can easily be used for a 1D timeline.

Acknowledgments

This paper is based on joint work with Dan Denis, now at the University of Montana. It was supported by Grant 8150 from the National Science and Research Council of Canada and was prepared with the assistance of *Les Chevaliers des Album de Statistique Graphique*, particularly Antoine de Falguerolles and Rüdiger Ostermann, to whom we are grateful. We also thank several reviewers for helping to make this presentation more cogent.

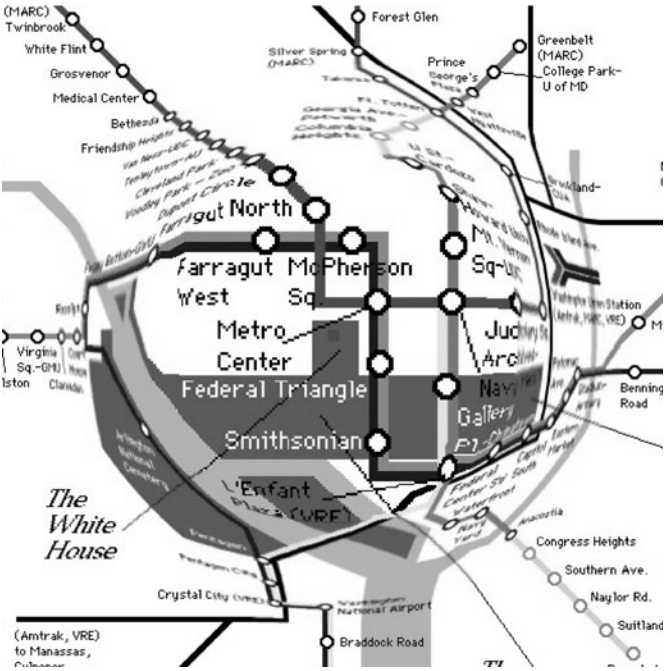


Fig. 11. Fisheye view of central Washington, D.C., illustrating a hyperbolic view

References

BECKER, R. A. (1994): A brief history of S. In: P. Dirschedl and R. Ostermann (Eds.): *Computational Statistics*. Physica, Heidelberg, 81–110.

BENIGER, J. R. and ROBYN, D. L. (1978): Quantitative graphics in statistics: A brief history. *The American Statistician, 32, 1–11.*

BERTIN, J. (1967): *Sémiologie Graphique: Les diagrammes, les réseaux, les cartes*. Gauthier-Villars, Paris.

BERTIN, J. (1983): *Semiology of Graphics*. University of Wisconsin Press, Madison, WI. (trans. W. Berg).

FAREBROTHER, R. W. (1999): *Fitting Linear Relationships: A History of the Calculus of Observations*. Springer, New York, 1750–1900.

FRIENDLY, M. (1994): Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association, 89, 190–200.*

FRIENDLY, M. (2000): Re-Visions of Minard. *Statistical Computing & Statistical Graphics Newsletter, 11/1, 1, 13–19.*

FRIENDLY, M. and DENIS, D. (2000): The roots and branches of statistical graphics. *Journal de la Société Française de Statistique, 141/4, 51–60.* (published in 2001).

FRIENDLY, M. and DENIS, D. (2004): The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences.* (In press, accepted 7/09/04).

FRIENDLY, M. and KWAN, E. (2003): Effect ordering for data displays. *Computational Statistics and Data Analysis, 43/4, 509–539.*

FRIIS, H. R. (1974): Statistical cartography in the United States prior to 1870 and the role of Joseph C. G. Kennedy and the U.S. Census Office. *American Cartographer, 1, 131–157.*

FUNKHOUSER, H. G. (1936): A note on a tenth century graph. *Osiris, 1, 260–262.*

FUNKHOUSER, H. G. (1937): Historical development of the graphical representation of statistical data. *Osiris, 3/1, 269–405.* Reprinted St. Catherine Press, Brugge, 1937.

FURNAS, G. W. (1986): Generalized fisheye views. In: *Proceedings of the ACM CHI '86 Conference on Human Factors in Computing Systems.* ACM, Boston, MA, 16–23.

HALD, A. (1990): *A History of Probability and Statistics and their Application before 1750.* John Wiley & Sons, New York.

HANKINS, T. L. (1999): Blood, dirt, and nomograms: A particular history of graphs. *Isis, 90, 50–80.*

HEISER, W. J. (2000): Early roots of statistical modelling. In: J. Blasius, J. Hox, E. DE Leeuw, and P. Schmidt (Eds.): *Social Science Methodology in the New Millenium: Proceedings of the Fifth International Conference on Logic and Methodology.* TT-Publikaties, Amsterdam.

HOFF, H. E. and GEDDES, L. A. (1959): Graphic recording before Carl Ludwig: An historical summary. *Archives Internationales d'Histoire des Sciences, 12, 3–25.*

HOFF, H. E. and GEDDES, L. A. (1962): The beginnings of graphic recording. *Isis, 53, 287–324.* Pt. 3.

KRUSKAL, W. (1977): Visions of maps and graphs. In: *Proceedings of the International Symposium on Computer- Assisted Cartography, Auto-Carto II.* 27–36.

LAMPING, J., RAO, R. and PIROLLI, P. (1995): A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In: *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 401–408.

MERTON, R. K. (1973): *Sociology of Science: Theoretical and Empirical Investigations.* University of Chicago Press, Chicago, IL.

MINARD, C. J. (1844): Tableaux figuratifs de la circulation de quelques chemins de fer. lith. (n.s.). ENPC: 5860/C351, 5299/C307.

PALSKY, G. (1996): *Des Chiffres et des Cartes: Naissance et développement de la Cartographie Quantitative Français au XIX*^{e} *siècle.* CHTS, Paris.

PEARSON, E. S., ed. (1978): *The History of Statistics in the 17th and 18th Centuries Against the Changing Background of Intellectual, Scientific and Religeous Thought.* Griffin & Co. Ltd, London. Lectures by Karl Pearson given at University College London during the academic sessions 1921–1933.

PLAYFAIR, W. (1786): *Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century.* Corry, London. 3rd edition, Stockdale, London, 1801; French edition, Tableaux d'arithmétique linéaire, du commerce, des finances, et de la dette nationale de l'Angleterre (Chez Barrois l'Aîné, Paris, 1789).

PLAYFAIR, W. (1801): *Statistical Breviary; Shewing, on a Principle Entirely New, the Resources of Every State and Kingdom in Europe.* Wallis, London.

PLAYFAIR, W. (1821): Letter on our agricultural distresses, their causes and remedies; accompanied with tables and copperplate charts shewing and comparing the prices of wheat, bread and labour, from 1565 to 1821.

PORTER, T. M. (1986): *The Rise of Statistical Thinking 1820–1900*. Princeton University Press, Princeton, NJ.

PRIESTLEY, J. (1765): *A chart of biography*. London.

RIDDELL, R. C. (1980): Parameter disposition in pre-Newtonain planetary theories. *Archives Hist. Exact Sci., 23, 87–157*.

ROBINSON, A. H. (1982): *Early Thematic Mapping in the History of Cartography*. University of Chicago Press, Chicago.

ROYSTON, E. (1970): Studies in the history of probability and statistics, III. a note on the history of the graphical presentation of data. *Biometrika, 241–247*. 43, Pts. 3 and 4 (December 1956); reprinted In: E. S. Pearson and M. G. Kendall (Eds.): *Studies in the History Of Statistics and Probability Theory*. Griffin, London.

STIGLER, S. M. (1980): Stigler's law of eponomy. *Transactions of the New York Academy of Sciences, 39, 147–157*.

STIGLER, S. M. (1986): *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press, Cambridge, MA.

TILLING, L. (1975): Early experimental graphs. *British Journal for the History of Science, 8, 193–213*.

TUFTE, E. R. (1983): *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.

TUFTE, E. R. (1990): *Envisioning Information*. Graphics Press, Cheshire, CT.

TUFTE, E. R. (1997): *Visual Explanations*. Graphics Press, Cheshire, CT.

WALLIS, H. M. and ROBINSON, A. H. (1987): *Cartographical Innovations: An International Handbook of Mapping Terms to 1900*. Map Collector Publications, Tring, Herts.

WHEELER, J. A. (1982): Bohr, Einstein, and the strange lesson of the quantum. In: R. Q. Elvee (Ed.): *Mind in Nature*. Harper and Row, San Francisco.

Quantitative Text Typology: The Impact of Word Length

Peter Grzybek[1], Ernst Stadlober[2], Emmerich Kelih[1], and Gordana Antić[2]

[1] Department for Slavic Studies, University Graz, A-8010 Graz, Austria
[2] Department for Statistics, Graz University of Technology, A-8010 Graz, Austria

Abstract. The present study aims at the quantitative classification of texts and text types. By way of a case study, 398 Slovenian texts from different genres and authors are analyzed as to their word length. It is shown that word length is an important factor in the synergetic self-regulation of texts and text types, and that word length may significantly contribute to a new typology of discourse types.[2]

1 Introduction: Structuring the universe of texts

Theoretically speaking, we assume that there is a universe of texts representing an open (or closed) system, i.e. an infinite (or finite) number of textual objects. The structure of this universe can be described by two processes: identification of its objects, based on a definition of 'text', and classification of these objects, resulting in the identification and description of hierarchically ordered sub-systems. To pursue the astronomic metaphor, the textual universe will be divided into particular galaxies, serving as attractors of individual objects. Finally, within such galaxies, particular sub-systems of lower levels will be identified, comparable to, e.g., stellar or solar systems. The two processes of identification and classification cannot be realized without recourse to theoretical assumptions as to the obligatory and/or facultative characteristics of the objects under study: neither quantitative nor qualitative characteristics are immanent to the objects; rather, they are the result of analytical cognitive processes.

1.1 Classification and quantification

To one degree or another, any kind of classification involves quantification: Even in seemingly qualitative approaches, quantitative arguments come into play, albeit possibly only claiming – implicitly or explicitly – that some objects are 'more' or 'less' similar or close to each other, or to some alleged norm or prototype. The degree of quantification is governed by the traits incorporated into the meta-language. Hence it is of relevance on which analytical level the process of classification is started. Note that each level has its own problems as to the definition of sub-systems and their boundaries.

[2] This study is related to research project #15485 (Word Length Frequencies in Slavic Texts), supported by the Austrian Research Fund (FWF).

In any case, a classification of the textual universe cannot be achieved without empirical research. Here, it is important to note that the understanding of empirical work is quite different in different disciplines, be they concerned with linguistic objects or not. Also, the proportion of theory and practice, the weighting of qualitative and quantitative arguments, may significantly differ. Disciplines traditionally concentrating on language tend to favor theoretical and qualitative approaches; aside from these approaches, corpus linguistics as a specific linguistic sub-discipline has a predominant empirical component. Defining itself as "data-oriented", the basic assumption of corpus linguistics is that a maximization of the data basis will result in an increasingly appropriate ("representative") language description. Ultimately, none of these disciplines – be they of predominantly theoretical or empirical orientation – can work without quantitative methods.

Here, quantitative linguistics comes into play as an important discipline in its own right: as opposed to the approaches described above, quantitative linguistics strives for the detection of regularities and connections in the language system, aiming at an empirically based theory of language. The transformation of observed linguistic data into quantities (i.e., variables and constants), is understood as a standardized approach to observation. Specific hypotheses are statistically tested and, ideally speaking, the final interpretation of the results obtained is integrated into a theoretical framework.

1.2 Quantitative text analysis: From a definition of the basics towards data homogeneity

The present attempt follows these lines, striving for a quantitative text typology. As compared to corpus linguistics, this approach – which may be termed *quantitative text analysis* – is characterized by two major lines of thinking: apart from the predominantly theoretical orientation, the assumption of quantitative text analysis is that 'text' is the relevant analytical unit at the basis of the present analysis. Since corpus linguistics aims at the construction, or re-construction, of particular norms, of "representative" standards, of (a given) language, corpus-oriented analyses are usually based on a mixture of heterogeneous texts, of a "quasi text", in a way (Orlov (1982)). On contrast, quantitative text analysis focuses on texts as homogeneous entities. The basic assumption is that a (complete) text is a self-regulating system, ruled by particular regularities. These regularities need not necessarily be present in text segments, and they are likely to intermingle in any kind of text combination. Quite logically, the question remains, what a 'text' is: is it a complete novel, composed of books?, or the complete book of a novel, consisting of several chapters?, or each individual chapter of a given book?, or perhaps even a paragraph, or a dialogical or narrative sequence within it? Ultimately, there is no clear definition in text scholarship, and questions whether we need a "new" definition of text, regularly re-occur in relevant discussions. Of course,

this theoretical question goes beyond the scope of this paper. From a statistical point of view, we are faced with two major problems: the problem of data homogeneity, and the problem of the basic analytical units. Thus, particular decisions have to be made as to the boundary conditions of our study:

- ▷ We consider a *'text'* to be the result of a homogeneous process of text generation. Therefore, we concentrate on letters, or newspaper comments, or on chapters of novels, as individual texts. Assuming that such a 'text' is governed by synergetic processes, these processes can and must be quantitatively described. The descriptive models obtained for each 'text' can be compared to each other, possibly resulting in one or more general model(s); thus, a quantitative typology of texts can be obtained.
- ▷ But even with a particular definition of 'text', it has to be decided which of their traits are to be submitted to quantitative analyses. Here, we concentrate on *word length*, as one particular linguistic trait of a text.

1.3 Word length in a synergetic context

Word length is, of course, only one linguistic trait of texts, among others, and one would not expect a coherent text typology, based on word length only. However, the criterion of word length is not an arbitrarily chosen factor (cf. Grzybek (2004)). First, experience has shown that genre is a crucial factor influencing word length (Grzybek and Kelih (2004); Kelih et al., this volume); this observation may as well turned into the question to what degree word length studies may contribute to a quantitative typology of texts. And second, word length is an important factor in a synergetic approach to language and text. We cannot discuss the synergetics of language in detail, here (cf. Köhler (1986)); yet, it should be made clear that word length is no isolated linguistic phenomenon: given one accepts the distinction of linguistic levels, as (1) phoneme/grapheme, (2) syllable/morpheme, (3) word/lexeme, (4) clause, and (5) sentence, at least the first three levels are concerned with recurrent units. Consequently, on each of these levels, the re-occurrence of units results in particular frequencies, which may be modelled with recourse to specific frequency distribution models. Both the units and their frequencies are closely related to each other. The units of all five levels are characterized by length, again mutually influencing each other, resulting in specific frequency length distributions. Table 1 demonstrates the interrelations.

Finally, in addition to the decisions made, it remains to be decided which shall be the analytical units, that is not only what a 'word' is (a graphemic, phonetic, phonological, intonational, etc. unit), but also in which units word length is supposed to be measured (number of letters, of graphemes, of phonemes, syllables, morphemes, etc.).

- ▷ In the present analysis, we concentrate on word as an orthographic-phonemic category (cf. Antić et al. (2004)), measuring word length as the number of syllables per word.

Table 1. Word length in a synergetic circuit

	SENTENCE	Length	Frequency
	↕		
	CLAUSE	Length	Frequency
↱		↰ ↕	
Frequency	WORD / LEXEME	Length	Frequency
↕ ↱		↰ ↕	
Frequency	SYLLABLE / MORPHEME	Length	Frequency
↕ ↱		↰ ↕	
Frequency	PHONEME / GRAPHEME	Length	Frequency

1.4 Qualitative and quantitative classifications: A priori and a posteriori

Given these definitions, we can now pursue our basic question as to a quantitative text typology. As mentioned above, the quantitative aspect of classification is often neglected or even ignored in qualitative approaches. As opposed to this, qualitative categories play an overtly accepted role in quantitative approaches, though the direction of analysis may be different:

1. One may favor a *"tabula rasa" principle* not attributing any qualitative characteristics in advance; the universe of texts is structured according to word length only, e.g. by clustering methods, by analyzing the parameters of frequency distributions, etc.;
2. One may prefer an *a priori* ↔ *a posteriori principle*: in this case, a particular qualitative characteristic is attributed to each text, and then, e.g. by discriminant analysis, one tests whether these categorizations correspond to the quantitative results obtained.

Applying qualitative categories, the problem of data heterogeneity once again comes into play, now depending on the meta-language chosen. In order to understand the problem, let us suppose, we want to attribute a category such as 'text type' to each text. In a qualitative approach, the text universe is structured with regard to external (pragmatic) factors – "with reference to the world". The categories usually are based either on general communicative functions of language (resulting in particular *functional styles*) or on specific situational functions (resulting in specific *text sorts*).

(a) The concept of functional style, successfully applied in previous quantitative research (cf. Mistrík (1966)), has been mainly developed in Russian and Czechoslovak stylistics, understanding style as serving particular socio-communicative functions. A functional style thus relates to particular discourse spheres, such as everyday, official-administrative, scientific, journalistic, or artistic communication. Such a coarse categorization with about only half a dozen of categories necessarily results in an extreme heterogeneity of the texts included in the individual categories.

(b) Contemporary text sort research (cf. Adamczik (1995), 255ff.) distinguishes ca. 4,000 categories. In this case, the categories are less broad and general, the material included tends to be more homogeneous, but the number of categories can hardly be handled in empirical research.

In order to profit from the advantages of both approaches, it seems reasonable to combine these two principles (cf. Grzybek and Kelih (2004)): each text sort thus tentatively is attributed to a functional style (cf. Figure 1), the attribution being understood as a more or less subjective a priori classification. Thus, in the subsequent quantitative analysis, both bottom-up (text → text sort → functional style) and top-down analyses are possible in a vertical perspective, as well as first order and second order cross-comparisons, in a horizontal perspective (i.e., between different functional styles or text sorts). Our basic assumption is that the highest level – the entities of which are

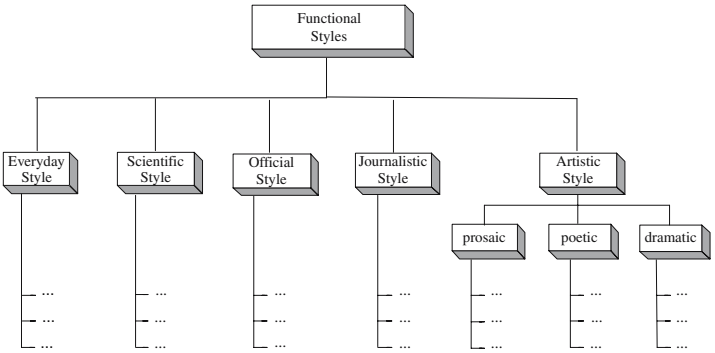


Fig. 1. Functional styles and text sorts

comparable to 'text galaxies' (see above) – should not primarily considered to be defined by socio-communicative functions, but regarded as linguistic phenomena: It seems reasonable to assume that different text sorts (analogous to our "stellar systems"), which serve particular functions as well, should be characterized by similar linguistic or stylistic traits. As opposed to merely qualitative text typologies, the attribution of text sorts to functional styles is to be understood as an a priori hypothesis, to be submitted to empirical tests. As a result, it is likely that either the a priori attributions have to be modified, or that other categories have to be defined at the top level, e.g. specific *discourse types*, instead of functional styles.

2 A case study: Classifying 398 Slovenian texts

The present case study is an attempt to arrive at a classification of 398 Slovenian texts, belonging to various sorts, largely representing the spectrum of functional styles; the sample is characterized in Table 2. The emphasis

Table 2. 398 Slovenian texts

FUNCTIONAL STYLE	AUTHOR(S)	TEXT TYPE(S)	no.
⊡ **Everyday**	Cankar, Jurčič	Private Letters	61
⊡ **Public**	various	Open Letters	29
⊡ **Journalistic**	various	Readers' Letters, Comments	65
⊡ **Artistic**			
⊙ *Prose*	Cankar	Individual Chapters from Short Novels (*povest*)	68
	Švigelj-Mérat / Kolšek	Letters from an Epistolary Novel	93
⊙ *Poetry*	Gregorčič	Versified Poems	40
⊙ *Drama*	Jančar	Individual Acts from Dramas	42

on different types of letters is motivated by the fact that 'letter' as a genre often is regarded to be prototypical of (a given) language in general, since a 'letter' is assumed to be located between oral and written communication, and considered as the result of a unified, homogeneous process of text generation. This assumption is problematic, however, if one takes into account the fact that contemporary text sort research (cf. Adamczik (1995), 255ff.) distinguishes several dozens of different letter types. Consequently, it would be of utmost importance (i) to compare how the genre of letters as a whole relates to other genres, and (ii) to see how different letter types relate to each other – in fact, any difference would weaken the argument of the letter's prototypicality.

In our analyses, each text is analyzed with regard to word length, the mean (m_1) being only one variable characterizing a given frequency distribution. In fact, there is a pool of ca. 30 variables at our disposal, including the four central moments, variance and standard deviation, coefficient of variation, dispersion index, entropy, repeat rate, etc. These variables are derived from the word length frequencies of a given text; Figure 2 examplarily represents the relative frequencies of x-syllable words for two arbitrarily chosen texts. In this case, there are significant differences between almost all length classes.

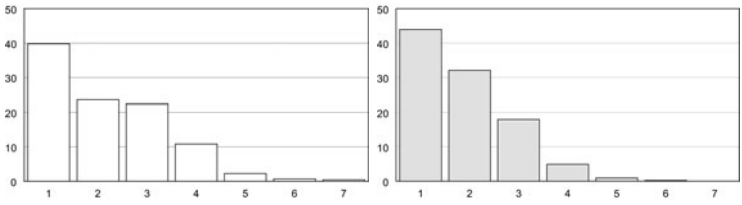


Fig. 2. Word length frequencies (in %) of two different texts (Left: Comment (#324). Right: Private letter (#1))

2.1 Post hoc analysis of mean word length

By way of a first approximation, it seems reasonable to calculate a post-hoc-analysis of the mean values. As a result of this procedure, groups without significant differences form homogeneous subgroups, whereas differing groups are placed in different groups. As can be seen from Table 3, which is based on mean word length (m_1) only, homogeneous subgroups do in fact exist; even more importantly, however, all four letter types fall into different categories. This observation gives rise to doubt the assumption, that 'letter' as a category can serve as a prototype of language without further distinction.

Table 3. Post hoc analyses (m_1)

		Subgroup for $\alpha = .05$				
Text sort	n	1	2	3	4	5
Poems	40	1.7127				
Short stories	68		1.8258			
Private letters	61		1.8798			
Drama	42		1.8973			
Epistolary novel	93			2.0026		
Readers' letters	30				2.2622	
Comments	35				2.2883	
Open Letters	29					2.4268

2.2 Discriminant analyses: The whole corpus

In linear discriminant analyses, specific variables are submitted to linear transformations in order to arrive at an optimal discrimination of the individual cases. At first glance, many variables of our pool may be important for discrimination, where the individual texts are attributed to groups, on the basis of these variables. However, most of the variables are redundant due to their correlation structure. The stepwise procedures in our analyses resulted in at most four relevant predictor variables for the discriminant functions.
Figure 3 shows the results of the discriminant analysis for all eight text sorts, based on four variables: mean word length (m_1), variance (m_2), coefficient of variation ($v = s/m_1$), and relative frequency of one-syllable words (p_1). With only 56.30% of all texts being correctly discriminated, some general tendencies can be observed: (1) although some text sorts are located in clearly defined areas, there are many overlappings; (2) poems seem to be a separate category, as well as readers' letters, open letters, and comments, on the other end; (3) drama, short story, private letters and the letters from the epistolary novel seem to represent some vaguely defined common area.

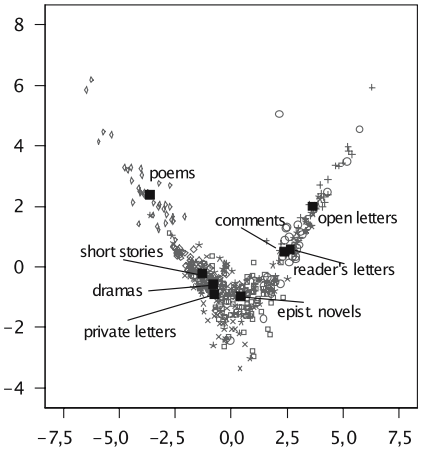


Fig. 3. Discriminant analysis: Eight text sorts

2.3 From four to two letter types

In a first approach to explore the underlying structure of the textual universe, we concentrate on the four letter types, only, since they were all attributed to different classes in the post hoc analyses. Treating all of them – i.e., private letters (PL), open letters (OL), readers' letters (RL), and letters from an epistolary novel (EN) –, as separate classes, a percentage of 70.40% correctly discriminated texts is obtained, with only two relevant variables: m_1 and v. There is an obvious tendency that private letters (PL) and the letters from

Table 4. Discriminant analysis: Four letter types ($n = 213$)

	Predicted group				
Letter Type	PL	OL	RL	EN	Total
PL	37	0	2	22	61
OL	0	22	3	4	29
RL	1	9	10	10	30
EN	10	0	3	80	93

the epistolary novel (EN) represent a common category, whereas open letters (OL) and readers's letters (RL) display this tendency to a lesser degree, if at all. Combining private letters and the letters from the epistolary novel in one group, thus discriminating three classes of letters, yields a percentage of 86.90% correctly discriminated texts, with only two variables: m_1 and p_2 (i.e., the percentage of two-syllable words). Table 5 shows the results in detail: 98% of the combined group are correctly discriminated. This is a strong argument

Table 5. Discriminant analysis: Three letter types ($n = 213$)

	Predicted group			
Group	1	2	3	Total
1	151	0	3	154
2	2	20	6	28
3	12	5	14	31
1=$\{PL, EN\}$		2=OL		3=RL

in favor of the assumption that we are concerned with some common group of private letters, be they literary or not. This result sheds serious doubt on the possibility to distinguish fictional literary letters: obviously, they reproduce or "imitate" the linguistic style of private letters, what generally calls into question the functional style of prosaic literature. Given this observation, it seems reasonable to combine readers' letters (RL) and open letters (OL) in one common group, too, and to juxtapose this group of public letters to the group of private letters. In fact, this results in a high percentage of 92.00%, with m_1 and p_2 being the relevant variables.

2.4 Towards a new typology

On the basis of these findings, the question arises if the two major groups – private letters (PL/EN) and public letters (OL/RL) – are a special case of more general categories, such as, e.g., 'private/everyday style' and 'public/official style'. If this assumption should be confirmed, the re-introduction of previously eliminated text sorts should yield positive results.

The re-introduction of journalistic comments (CO) to the group of public texts does not, in fact, result in a decrease of the good discrimination result: as Table 6 shows, 91.10% of the 248 texts are correctly discriminated (again, with m_1 and p_2, only). Obviously, some distinction along the line of public/official vs. private/everyday texts seems to be relevant.

Table 6. Discriminant analysis: Five text sorts in two categories: Public/Official vs. Private/Everyday ($n = 248$)

	Predicted group		
Group	1	2	Total
1	148	6	154
2	16	78	94
1=$\{PL, EN\}$		2=$\{OL, RL, CO\}$	

The re-introduction of the dramatic texts (DR), as well, seems to be a logical consequence, regarding them as the literary pendant of everyday dialogue. We thus have 290 texts, originating from six different text sorts, and grouped in two major classes; as Table 7 shows, 92.40% of the texts are correctly discriminated. One might object, now, that the consideration of only two classes is likely to be effective. Yet, it is a remarkable result that the addition of two non-letter text sorts does not result in a decrease of the previous result.

Table 7. Discriminant analysis: Six text sorts in two categories: Public/Official vs. Private/Everyday ($n = 290$)

	Predicted group		
Group	1	2	Total
1	190	6	196
2	16	78	94

1=$\{PL, EN, DR\}$ 2=$\{OL, RL, CO\}$

The re-introduction of the poetic texts (PO) as a category in its own right, results in three text classes. Interestingly enough, under these circumstances, too, the result is not worse: rather, a percentage of 91.20% correct discriminations is obtained on the basis of only three variables: m_1, p_2, v. The results are represented in detail, in Table 8.

Table 8. Discriminant analysis: Seven text sorts in three categories: Public/Official vs. Private/Everyday vs. Poetry ($n = 330$)

	Predicted group			
Group	1	2	3	Total
1	191	3	2	196
2	19	75	0	94
3	5	0	35	40

1=$\{PL, EN, DR\}$ 2=$\{OL, RL, CO\}$ 3=$\{PO\}$

It can clearly be seen that the poetic texts represent a separate category and imply almost no mis-classifications. At this point, the obvious question arises if a new typology might be the result of our quantitative classification. With this perspective in mind, it should be noticed that seven of our eight text sorts are analyzed in Table 8.

The re-introduction of the literary prose texts (LP) is the last step, thus again arriving at the initial number of eight text sorts. As can be seen from Table 9, the percentage of correctly discriminated texts now decreases to 79.90%.

Table 9. Discriminant analysis: Eight text sorts in four categories ($n = 398$)

	Predicted group				
Group	1	2	3	4	Total
1	183	3	9	1	196
2	19	75	0	0	94
3	42	0	26	0	68
4	1	0	5	34	40

1=$\{PL, EN, DR\}$ 2=$\{OL, RL, CO\}$
3=$\{LP\}$ 4=$\{PO\}$

A closer analysis shows that the most mis-classifications appear between literary texts and private letters. Interestingly enough, many of these texts are from one and the same author (Ivan Cankar). One might therefore suspect authorship to be an important factor; however, Kelih et al. (this volume) have good arguments (and convincing empirical evidence) that word length is less dependent on authorship, than it is on genre. As an alternative interpretation, the reason may well be a specific for the analyzed material because in case of the literary texts, we are concerned with short stories which aim at the imitation of orality, and include dialogues to varying degree.

Therefore, including the literary prose texts (LP) in the group of inofficial/oral texts, and separating them from the official/written group, on the one hand, and the poetry group, on the other, results in a percentage of 92.70% correctly discriminated texts, as can be seen from Table 10. The final outcome of our classification is represented in Figure 4.

Table 10. Discriminant analysis: Eight text sorts in three categories: Inofficial / Oral vs. Official / Written vs. Poetry ($n = 398$)

	Predicted group			
Group	1	2	3	Total
1	260	3	1	264
2	19	75	0	94
3	6	0	34	40

1=$\{PL, EN, DR, LP\}$ 2=$\{OL, RL, CO\}$
3=$\{PO\}$

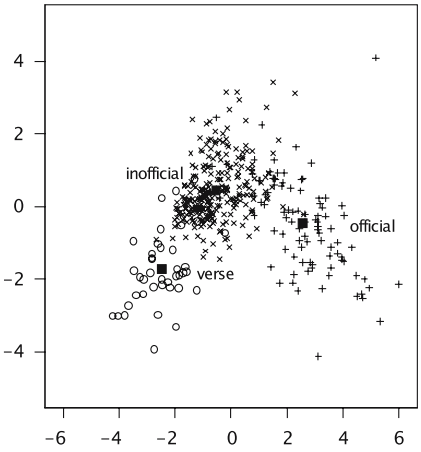


Fig. 4. Discriminant analysis: Final results and new categorization

2.5 Conclusion

The results suggest the existence of specific discourse types, which do not coincide with traditional functional styles. Future research must concentrate on possible additional discourse types and their relation to text sorts.

References

ADAMCZIK, Kirsten (1995): *Textsorten – Texttypologie. Eine kommentierte Bibliographie.* Nodus, Münster.

ANTIĆ, G., KELIH, E., and GRZYBEK, P. (2004): Zero-syllable Words in Determining Word Length. In: P. Grzybek (Ed.): *Contributions to the Science of Language. Word Length Studies and Related Issues.* [In print]

GRZYBEK, P. (2004): History and Methodology of Word Length Studies: The State of the Art. In: P. Grzybek (Ed.): *Contributions to the Science of Language: Word Length Studies and Related Issues.* [In print]

GRZYBEK, P. and KELIH, E. (2004): Texttypologie in/aus empirischer Sicht. In: J. Bernard, P. Grzybek and Ju. Fikfak (Eds.): *Text and Reality.* Ljubljana. [In print].

GRZYBEK, P. and STADLOBER, E. (2003): Zur Prosa Karel Čapeks – Einige quantitative Bemerkungen. In: S. Kempgen, U. Schweier and T. Berger (Eds.), *Rusistika – Slavistika – Lingvistika. Festschrift für Werner Lehfeldt zum 60. Geburtstag.* Sagner, München, 474–488.

KELIH, E., ANTIĆ, G., GRZYBEK, P., and STADLOBER, E. (2004): Classification of Author and/or Genre? [Cf. this volume]

KÖHLER, R. (1986): Zur synergetischen Linguistik: Struktur und Dynamik der Lexik. Brockmeyer, Bochum.

ORLOV, Ju.K. (1982): Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? (Die Antinomie "Sprache–Rede" in der statistischen Linguistik). In: Ju.K. Orlov and M.G. Boroda, I.Š. Nadarešvili: *Sprache, Text, Kunst. Quantitative Analysen.* Brockmeyer, Bochum.

Cluster Ensembles

Kurt Hornik

Institut für Statistik und Mathematik,
Wirtschaftsuniversität Wien,
Augasse 2–6, A-1090 Wien, Austria

Abstract. Cluster ensembles are collections of individual solutions to a given clustering problem which are useful or necessary to consider in a wide range of applications. Aggregating these to a "common" solution amounts to finding a consensus clustering, which can be characterized in a general optimization framework. We discuss recent conceptual and computational advances in this area, and indicate how these can be used for analyzing the structure in cluster ensembles by clustering its elements.

1 Introduction

Ensemble methods create solutions to learning problems by constructing a set of individual (different) solutions ("base learners"), and subsequently suitably aggregating these, e.g., by weighted averaging of the predictions in regression, or by taking a weighted vote on the predictions in classification. Such methods, which include Bayesian model averaging (Hoeting et al. (1999)), bagging (Breiman (1996)) and boosting (Friedman et al. (2000)) have already become very popular for supervised learning problems (Dietterich (2002)).

In general, aggregation yields algorithms with "low variance" in the statistical learning sense so that the results obtained by aggregation are more "structurally stable". Based on the success and popularity of ensemble methods, the statistical and machine learning communities have recently also become interested in employing these in unsupervised learning tasks, such as clustering. (Note that in these communities, the term "classification" is used for discriminant analysis. To avoid ambiguities, we will use "supervised classification" to refer to these learning problems.) For example, a promising idea is to obtain more stable partitions of a given data set using bagging (Bootstrap Aggregating), i.e., by training the same base clusterer on bootstrap samples from the data set and then finding a "majority decision" from the labelings thus obtained. But obviously, aggregation is not as straightforward as in the supervised classification framework, as these labelings are only unique up to permutations and therefore not necessarily matched. In the classification community, such aggregation problems have been studied for quite some time now. A special issue of the Journal of Classification was devoted to "Comparison and Consensus of Classifications" (Day (1986)) almost two decades ago. By building on the readily available optimization framework for obtaining

consensus clusterings it is possible to exploit the full potential of the cluster ensemble approach.

Employing cluster ensembles can be attractive or even necessary for several reasons, the main ones being as follows (see e.g. Strehl and Ghosh (2002)):

- To improve quality and robustness of the results. Bagging is one approach to reduce variability via resampling or reweighting of the data, and is used in Leisch (1999) and Dudoit and Fridlyand (2002). In addition, many clustering algorithms are sensitive to random initializations, choice of hyper-parameters, or the order of data presentation in on-line learning scenarios. An obvious idea for possibly eliminating such *algorithmic* variability is to construct an ensemble with (randomly) varied characteristics of the base algorithm. This idea of "sampling from the algorithm" is used in Dimitriadou et al. (2001, 2002). Aggregation can also *leverage* performance in the sense of turning weak into strong learners; both Leisch (1999) and Dimitriadou et al. (2002) illustrate how e.g. suitable aggregation of base k-means results can reveal underlying non-convex structure which cannot be found by the base algorithm. Other possible strategies include varying the "features" used for clustering (e.g., using various preprocessing schemes), and constructing "meta-clusterers" which combine the results of the application of *different* base algorithms as an attempt to reduce dependency of results on specific methods, and take advantage of today's overwhelming method pluralism.
- To aggregate results over conditioning/grouping variables in situations where repeated measurements of features on objects are available for several levels of a grouping variable, such as the 3-way layout in Vichi (1999) where the grouping levels correspond to different time points at which observations are made.
- To reuse existing knowledge. In applications, it may be desired to reuse legacy clusterings in order to improve or combine these. Typically, in such situations only the cluster labels are available, but not the original features or algorithms.
- To accommodate the needs of distributed computing. In many applications, it is not possible to use all data simultaneously. Data may not necessarily be available in a single location, or computational resources may be insufficient to use a base clusterer on the whole data set. More generally, clusterers can have access to either a subset of the objects ("object-distributed clustering") or the features ("feature-distributed clustering"), or both.

In all these situations, aggregating (subsets of) the cluster ensemble by finding "good" consensus clusterings is fundamental. In Section 2, we consider a general optimization framework for finding consensus partitions. Extensions are discussed in Section 3.

2 Consensus partitions

There are three main approaches to obtaining consensus clusterings (Gordon and Vichi (2001)): in the *constructive* approach, a way of constructing a consensus clustering is specified: for example, a strict consensus clustering is defined to be one such that objects can only be in the same group in the consensus partition if they were in the same group in all base partitions. In the *axiomatic* approach, emphasis is on the investigation of existence and uniqueness of consensus clusterings characterized axiomatically. The *optimization* approach formalizes the natural idea of describing consensus clusterings as the ones which "optimally represent the ensemble" by providing a criterion to be optimized over a suitable set $\mathcal{C}$ of possible consensus clusterings. Given a function d which measures dissimilarity (or distance) between two clusterings, one can e.g. look for clusterings which minimize average dissimilarity, i.e., which solve

$$C^* = \text{argmin}_{C \in \mathcal{C}} \sum_{b=1}^{B} d(C, C_b)$$

over $\mathcal{C}$. Analogously, given a measure of similarity (or agreement), one can look for clusterings maximizing average similarity. Following Gordon and Vichi (1998), one could refer to the above C^* as the *median* or *medoid* clustering if the optimum is sought over the set of all possible base clusterings, or the set $\{C_1, \ldots, C_B\}$ of the base clusterings, respectively.

When finding consensus *partitions*, it seems natural to look for optimal *soft* partitions which make it possible to assign objects to several groups with varying degrees of "membership" (Gordon and Vichi (2001), Dimitriadou et al. (2002)). One can then assess the amount of belongingness of objects to groups via standard impurity measures, or the so-called classification margin (the difference between the two largest memberships). Note that "soft" partitioning includes fuzzy partitioning methods such as the popular fuzzy c-means algorithm (Bezdek (1974)) as well as probabilistic methods such as the model-based approach of Fraley and Raftery (2002). In addition, one can compute global measures Φ of the softness of partitions, and use these to extend the above optimization problem to minimizing

$$\sum_{b=1}^{B} \omega_b d(C, C_b) + \lambda \Phi(C)$$

over all soft partitions, where the ω indicate the importance of the base clusterings (e.g., by assigning importance according to softness of the base partitions), and λ controls the amount of "regularization". This extension also allows for a soft-constrained approach to the "simple" problem of optimizing over all hard partitions. Of course, one could consider criterion functions resulting in yet more robust consensus solutions, such as the median or trimmed mean of the distances $d(C, C_b)$.

One should note that the above optimization problems are typically computationally very hard. Finding an optimal hard partition with K labels in

general makes it necessary to search all possible hard partitions (the number of which is of the order $(K+1)^n$ (Jain and Dubes (1988)) for the optimum. Such exhaustive search is clearly impossible for most applications. Local strategies, e.g. by repeating random reassigning until no further improvement is obtained, or Boltzmann-machine type extensions (Strehl and Ghosh (2002)) are still expensive and not guaranteed to find the global optimum.

Perhaps the most popular similarity measure for partitions of the same data set is the Rand index (Rand (1971)) used in e.g. Gordon and Vichi (1998), or the Rand index corrected for agreement by chance (Hubert and Arabie (1985)) employed by Krieger and Green (1999). Finding (hard) consensus partitions by maximizing average similarity is NP-hard in both cases. Hence, Krieger and Green (1999) propose an algorithm (SEGWAY) based on the combination of local search by relabeling single objects together with "smart" initialization using random assignment, latent class analysis (LCA), multiple correspondence analysis (MCA), or a greedy heuristic. Note also that using (dis)similarity measures adjusted for agreement by chance works best if the partitions are stochastically independent, which is not necessarily the case in all cluster ensemble frameworks described in Section 1.

In what follows, the following terminology will be useful. Given a data set $\mathcal{X}$ with the measurements of the same features (variables) on n objects, a *K-clustering* of $\mathcal{X}$ assigns to each x_i in $\mathcal{X}$ a (sub-)probability K-vector $C(x_i) = (\mu_{i1}, \ldots, \mu_{iK})$ (the "membership vector" of the object) with $\mu_{i1}, \ldots, \mu_{iK} \geq 0$, $\sum_k \mu_{ik} \leq 1$. Formally,

$$C : \mathcal{X} \to M \in \mathcal{M}_K; \qquad \mathcal{M}_K = \{M \in \mathbb{R}^{n \times K} : M \geq 0, M1_K \leq 1_K\},$$

where 1_K is a lenght K column vector of ones, and $M1_K$ is the matrix product of M and 1_K. This framework includes hard partitions (where each $C(x_i)$ is a unit Cartesian unit vector) and soft ones, as well as incomplete (e.g., completely missing, for example if a sample from $\mathcal{X}$ was used) results where $\sum_k \mu_{ik} < 1$. Permuting the labels (which correspond to the columns of the membership matrix M) amounts to replacing M by $M\Pi$, where Π is a suitable permutation matrix.

The dissimilarity measure used in Models I and II of Gordon and Vichi (2001) and in Dimitriadou et al. (2002) use the Euclidean dissimilarity of the membership matrices, adjusted for optimal matching of the labels. If both partitions use the same number of labels, this is given by

$$d_F(M, \tilde{M}) = \min_\Pi \|M - \tilde{M}\Pi\|^2$$

where the minimum is taken over all permutation matrices Π and $\|\cdot\|$ is the Frobenius norm (so that $\|Y\|^2 = \mathrm{tr}(Y'Y)$, where $'$ denotes transposition). As $\|M - \tilde{M}\Pi\|^2 = \mathrm{tr}(M'M) - 2\,\mathrm{tr}(M'\tilde{M}\Pi) + \mathrm{tr}(\Pi'\tilde{M}'\tilde{M}\Pi) = \mathrm{tr}(M'M) - 2\,\mathrm{tr}(M'\tilde{M}\Pi) + \mathrm{tr}(\tilde{M}'\tilde{M})$, we see that minimizing $\|M - \tilde{M}\Pi\|^2$ is equivalent to maximizing $\mathrm{tr}(M'\tilde{M}\Pi) = \sum_{i,k} \mu_{ik}\tilde{\mu}_{i,\pi(k)}$, which for hard partitions is

the number of objects with the same label in the partitions given by M and $\tilde{M}\Pi$. Finding the optimal Π is thus recognized as an instance of the assignment problem (or weighted bipartite graph matching problem), which can be solved by a linear program using the so-called Hungarian method in time $O(K^3)$ (e.g., Papadimitriou and Steiglitz (1982)). If the partitions have different numbers of labels, matching also includes suitably collapsing the labels of the finer partition, see Gordon and Vichi (2001) for details.

Finding the consensus K-clustering of given base K-clusterings with membership matrices $M_1, \ldots, M_B$ amounts to minimizing $\sum_{b=1}^{B} d_F(M, M_b)$ over $\mathcal{M}_K$, and is equivalent to minimizing $\sum_{b=1}^{B} \|M - M_b\Pi_b\|^2$ over $M \in \mathcal{M}_K$ and all permutation matrices $\Pi_1, \ldots, \Pi_K$. Dimitriadou et al. (2002) show that the optimal M is of the form

$$M = \frac{1}{B} \sum\nolimits_{b=1}^{B} M_b \Pi_b$$

for suitable permutation matrices $\Pi_1, \ldots, \Pi_B$. A hard partition obtained from this consensus partition by assigning objects to the label with maximal membership thus performs simple majority voting after relabeling, which motivates the name "voting" for the proposed framework. The $\Pi_1, \ldots, \Pi_B$ in the above representation are obtained by simultaneously maximizing the profile criterion function

$$\sum\nolimits_{1 \le \beta, b \le B} \mathrm{tr}(\Pi_\beta' M_\beta' M_b \Pi_b)$$

over all possible permutation matrices (of course, one of these can be taken as the identity matrix). This is a special case (but not an instance) of the multiple assignment problem, which is known to be NP-complete, and can e.g. be approached using randomized parallel algorithms (Oliveira and Pardalos (2004)). However, we note that unlike in the general case, the above criterion function only contains second-order interaction terms of the permutations. Whether the determination of the optimal permutations and hence of the consensus clustering is possible in time polynomial in both B and K is currently not known.

Based on the characterization of the consensus solution, Dimitriadou et al. (2002) suggest a greedy forward aggregation strategy for determining approximate solutions. One starts with $\tilde{M}_0 = M_1$ and then, for all b from 1 to B, first determines a locally optimal relabeling $\tilde{\Pi}_b$ of M_b to $\tilde{M}_{b-1}$ (i.e., solves the assignment problem $\mathrm{argmin}_\Pi \|\tilde{M}_{b-1} - M_b\Pi\|^2$ using the Hungarian method), and determines the optimal $M = \tilde{M}_b = (1/b) \sum_{\beta=1}^{b} \tilde{M}_\beta \tilde{\Pi}_\beta$ for fixed $\tilde{\Pi}_1, \ldots, \tilde{\Pi}_b$ by on-line averaging as $\tilde{M}_b = (1 - 1/b)\tilde{M}_{b-1} + (1/b)M_b\tilde{\Pi}_b$. The final $\tilde{M}_B$ is then taken as the approximate consensus clustering. One could extend this approach into a fixed-point algorithm which repeats the forward aggregation, with the order of membership matrices possibly changed, until convergence. Gordon and Vichi (2001) propose a different approach

which iterates between simultaneously determining the optimal relabelings $\Pi_1, \ldots, \Pi_B$ for fixed M by solving the corresponding assignment problems, and then optimizing for M for fixed $\Pi_1, \ldots, \Pi_B$ by computing the average $(1/B)\sum_{b=1}^{B} M_b \Pi_b$.

In the aggregation strategy Bag1 of Dudoit and Fridlyand (2002), the same base clusterer is applied to both the original data set and B bootstrap samples thereof, giving membership matrices M_{ref} and $M_1, \ldots, M_B$. Optimal relabelings Π_b are obtained by matching the M_b to M_{ref}, and (a hard version of) the consensus partition is then obtained by averaging the $M_b \Pi_b$. There seems to be no optimization criterion underlying this constructive approach.

According to Messatfa (1992), historically the first index of agreement between partitions is due to Katz and Powell (1953), and based on the Pearson product moment correlation coefficient of the off-diagonal entries of the co-incidence matrices MM' of the partitions. (Note that the (i, j)-th element of MM' is given by $\sum_{k=1}^{K} \mu_{ik}\mu_{jk}$, which in the case of hard partitions is one if objects i and j are in the same group, and zero otherwise, and that relabeling does not change MM'.) A related dissimilarity measure (using covariance rather than correlation) is

$$d_C(M, \tilde{M}) = \|MM' - \tilde{M}\tilde{M}'\|^2$$

The corresponding consensus problem is the minimization of $\sum_b \|MM' - M_b M_b'\|^2$, or equivalently of

$$\left\| MM' - \frac{1}{B} \sum\nolimits_{b=1}^{B} M_b M_b' \right\|^2$$

over $\mathcal{M}_K$. This is Model III of Gordon and Vichi (2001), who suggest to use a sequential quadratic programming algorithm (which can only be guaranteed to find local minima) for obtaining the optimal $M \in \mathcal{M}_K$. The average co-incidence matrix $(1/B)\sum_{b=1}^{B} M_b M_b'$ also forms the basis of the constructive consensus approaches in Fred and Jain (2002) and Strehl and Ghosh (2002).

3 Extensions

The optimization approach to finding consensus clusterings is also applicable to the case of hierarchical clusterings (Vichi (1999)). If these are represented by the corresponding ultra-metric matrices $U_1, \ldots, U_B$, a consensus clustering can be obtained e.g. by minimizing $\sum_b \|U - U_b\|^2$ over all possible ultra-metric matrices U.

In many applications of cluster ensembles, interest is not primarily in obtaining a global consensus clustering, but to analyze (dis)similarity patterns in the base clusterings in more detail—i.e., to cluster the clusterings. Gordon and Vichi (1998) present a framework in which all clusterings considered

are hard partitions. Obviously, the underlying concept of "clustering clusterings", based on suitable (dis)similarity measures between clusterings, such as the ones discussed in detail in Section 2, is much more general. In particular, it is straightforward to look for hard prototype-based partitions of a cluster ensemble characterized by the minimization of

$$\sum_{k=1}^{K} \sum_{C(M_b)=e_k} d(M_b, P_k),$$

where e_k is the k-th Cartesian unit vector over all possible hard assignments C of membership matrices to labels and all suitable prototypes $P_1, \ldots, P_K$. If the usual algorithm which alternates between finding optimal prototypes for fixed assignments and reassigning the M_b to their least dissimilar prototype is employed, we see that finding the prototypes amounts to finding the appropriate consensus partitions in the groups. Similarly, soft partitions can be characterized as the minima of the fuzzy c-means style criterion function $\sum_{k,b} u_{kb}^q d(M_b, P_k)$.

References

BEZDEK, J. C. (1974): Numerical taxonomy with fuzzy sets. *Journal of Mathematical Biology, 1, 57–71.*

BREIMAN, L. (1996): Bagging predictors. *Machine Learning, 24(2), 123–140.*

DAY, W. H. E. (1986): Foreword: Comparison and consensus of classifications. *Journal of Classification, 3, 183–185.*

DIETTERICH, T. G. (2002): Ensemble learning. In: M. A. Arbib (Ed.): *The Handbook of Brain Theory and Neural Networks.* The MIT Press, Cambridge, MA, 405–408.

DIMITRIADOU, E., WEINGESSEL, A. and HORNIK, K. (2001): Voting-merging: An ensemble method for clustering. In: G. Dorffner, H. Bischof and K. Hornik (Eds.): *Artificial Neural Networks – ICANN 2001*, volume 2130 of *LNCS.* Springer Verlag, 217–224.

DIMITRIADOU, E., WEINGESSEL, A. and HORNIK, K. (2002): A combination scheme for fuzzy clustering. *International Journal of Pattern Recognition and Artificial Intelligence, 16(7), 901–912.*

DUDOIT, S. and FRIDLYAND, J. (2002): A prediction-based resampling method to estimate the number of clusters in a dataset. *Genome Biology, 3(7), 0036.1–0036.21.*

FRALEY, C. and RAFTERY, A. E. (2002): Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97, 611–631.* URL `http://www.stat.washington.edu/mclust`.

FRED, A. L. N. and JAIN, A. K. (2002): Data clustering using evidence accumulation. In: *Proceedings of the 16th International Conference on Pattern Recognition (ICPR 2002)*, 276–280.

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000): Additive logistic regression: A statistical view of boosting. *The Annals of Statistics, 28(2), 337–407.*

GORDON, A. D. and VICHI, M. (1998): Partitions of partitions. *Journal of Classification, 15, 265–285*.

GORDON, A. D. and VICHI, M. (2001): Fuzzy partition models for fitting a set of partitions. *Psychometrika, 66(2), 229–248*.

HOETING, J., MADIGAN, D., RAFTERY, A. and VOLINSKY, C. (1999): Bayesian model averaging: A tutorial. *Statistical Science, 14, 382–401*.

HUBERT, L. and ARABIE, P. (1985): Comparing partitions. *Journal of Classification, 2, 193–218*.

JAIN, A. K. and DUBES, R. C. (1988): *Algorithms for Clustering Data*. Prentice Hall, New Jersey.

KATZ, L. and POWELL, J. H. (1953): A proposed index of the conformity of one sociometric measurement to another. *Psychometrika, 18, 149–256*.

KRIEGER, A. M. and GREEN, P. E. (1999): A generalized Rand-index method for consensus clustering of separate partitions of the same data base. *Journal of Classification, 16, 63–89*.

LEISCH, F. (1999): *Bagged clustering*. Working Paper 51, SFB "Adaptive Information Systems and Modeling in Economics and Management Science". URL `http://www.ci.tuwien.ac.at/~leisch/papers/wp51.ps`.

MESSATFA, H. (1992): An algorithm to maximize the agreement between partitions. *Journal of Classification, 9, 5–15*.

OLIVEIRA, C. A. S. and PARDALOS, P. M. (2004): Randomized parallel algorithms for the multidimensional assignment problem. *Applied Numerical Mathematics, 49(1), 117–133*.

PAPADIMITRIOU, C. and STEIGLITZ, K. (1982): *Combinatorial Optimization: Algorithms and Complexity*. Prentice Hall, Englewood Cliffs.

RAND, W. M. (1971): Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association, 66(336)*.

STREHL, A. and GHOSH, J. (2002): Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research, 3, 583–617*.

VICHI, M. (1999): One-mode classification of a three-way data matrix. *Journal of Classification, 16, 27–44*.

Bootstrap Confidence Intervals for Three-way Component Methods

Henk A.L. Kiers*

Heymans Institute,
University of Groningen,
Grote Kruisstraat 2/1,
9712TS Groningen, The Netherlands

Abstract. The two most common component methods for the analysis of three-way data, CANDECOMP/PARAFAC (CP) and Tucker3 analysis, are used to summarize a three-mode three-way data set by means of a number of component matrices, and, in case of Tucker3, a core array. Until recently, no procedures for computing confidence intervals for the results from such analyses were available. Recently, such procedures have come available by Riu and Bro (2003) for CP using the jack-knife procedure, and by Kiers (2004) for CP and Tucker3 analysis using the bootstrap procedure. The present paper reviews the latter procedures, discusses their performance as reported by Kiers (2004), and illustrates them on an example data set.

1 Introduction

For the analysis of three-way data sets (e.g., data with scores of a number of subjects, on a number of variables, under a number of conditions) various exploratory three-way methods are available. The two most common methods for the analysis of three-way data are CANDECOMP/PARAFAC (Carroll and Chang (1970), Harshman (1970)) and Tucker3 analysis (Tucker (1966), Kroonenberg and De Leeuw (1980)). Both methods summarize the data by components for all three modes, and for the entities pertaining to each mode they yield component loadings; in the case of Tucker3 analysis, in addition, a so-called core array is given, which relates the components for all three modes to each other.
If we denote our $I \times J \times K$ three-way data array (which has usually been preprocessed by centering and/or scaling procedures) by $\underline{\mathbf{X}}$, then the two methods can be described as fitting the model

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk}, \tag{1}$$

where a_{ip}, b_{jq}, and c_{kr} denote elements of the component matrices $\mathbf{A}$ (for the first mode, e.g., the subjects), $\mathbf{B}$ (for the second mode, e.g., the variables),

* e-mail: : h.a.l.kiers@ppsw.rug.nl

and $\mathbf{C}$ (for the third mode, e.g., the conditions), of orders $I \times P$, $J \times Q$, and $K \times R$, respectively; g_{pqr} denotes the element (p, q, r) of the $P \times Q \times R$ core array $\underline{\mathbf{G}}$, and e_{ijk} denotes the error term for element x_{ijk}; P, Q, and R denote the numbers of components for the three respective modes. Once the solution has been obtained, component matrices and/or the core are usually rotated to simplify the interpretation (see, Kiers, 1998), without loss of fit.
CANDECOMP/PARAFAC (CP) differs from Tucker3 analysis in that in CP the core is set equal to a superidentity array (i.e., with $g_{pqr} = 1$ if $p = q = r$, and $g_{pqr} = 0$ otherwise). As a consequence, in the case of CP, for all modes we have the same number of components, and (1) actually reduces to

$$x_{ijk} = \sum_{r=1}^{R} a_{ir} b_{jr} c_{kr} + e_{ijk}, \tag{2}$$

Because the CP model is unique up to scaling and permutation, no rotations can be used to simplify the interpretation.
Both models are fitted to sample data by minimizing the sum of squared errors. Usually, the model that fits optimally to the sample is assumed to be, at least to some extent, also valid for the population from which the sample was drawn. However, until recently, only global measures for indicating the reliability of such generalizations from sample to population were available, for instance by using cross-validation (e.g, see Kiers and van Mechelen (2001)). Recently, however, resampling procedures (see, e.g., Efron and Tibshirani (1993)) have been proposed for obtaining confidence intervals for all generalizable individual parameters resulting from a three-way component analysis. For this purpose, Riu and Bro (2003) proposed a jack-knife procedure for CP, and Kiers (2004) proposed various bootstrap procedures for CP and Tucker3 analysis. Both procedures can be used when the entities in one of the three modes can be considered a random sample from a population. According to Efron and Tibshirani, in general, the bootstrap can be expected to be more efficient than the jack-knife, so here we will focus on the bootstrap rather than the jackknife.
Bootstrap analysis can be applied straightforwardly when solutions are uniquely determined. However, the Tucker3 solution is by no means uniquely determined. Kiers (2004) described various procedures for handling this nonuniqueness in case of Tucker3. He also studied their performance in terms of coverage of the resulting confidence intervals, and in terms of computational efficiency by means of a simulation study. The main purpose of the present paper is to review these procedures briefly, and to describe how such a procedure works in practice in case of the analysis of an empirical data set.

2 The bootstrap for fully determined solutions

The basic idea of the bootstrap is to mimic the sampling process that generated our actual data sample, as follows. We suppose that the entities in

the first mode (e.g., the individuals) are a random sample from a population. Then, with the bootstrap procedure we assess what could happen if we would consider our sample as a population, and if we randomly (re)sample with replacement from this 'pseudo-population'. In fact, we consider the distribution of score profiles in our actual sample as a proxy of the distribution of such profiles in our population; by randomly resampling from our sample, we mimic the construction of a sampling distribution, on the basis of which we intend to make inferences about actual population characteristics. In practice, if our three-way data set of order $I \times J \times K$ is denoted by $\underline{\mathbf{X}}$, and the score profiles for individual i are denoted by X_i, which is a matrix of order $J \times K$, then we randomly draw (with replacement) I matrices X_i from the set of matrices $\{X_1, \ldots, X_I\}$. This creates one bootstrap sample (in which some matrices may occur repeatedly, and others not at all), which is then reorganized into a three-way array. This procedure is to be repeated in order to obtain, for instance, 500 such bootstrap three-way arrays. (In the sequel, 500 is taken as the number of bootstrap samples, but this is just meant as an example; a higher number will always be better, although the improvement may be little).

Now to each bootstrap three-way array, we apply a three-way component method in *exactly the same way* as we applied it originally to our sample (hence including the pre- and postprocessing procedures we used in the analysis of the sample data), and we compute the statistics we are interested in. The statistics of main interest are the loadings and the core values. Let these be collected in a single vector θ. The vector of outcomes for our original sample is denoted as $\theta^{\mathbf{s}}$, whereas those for the bootstrap samples are denoted as $\theta^{\mathbf{b}}$, $b = 1, \ldots, 500$. Now, the variation in the 500 bootstrap sample outcome vectors indicates how and how much the outcome vectors vary if we randomly resample from our pseudo-population. This is used as an estimate of how much real samples from our real population can be expected to vary if we would sample repeatedly from our actual population.

A simple way to describe the variation across the bootstrap sample outcomes is to give, for each parameter separately, a percentile interval (e.g., a 95%percentile interval) which describes the range in which we find the middle 95% values (out of the total of 500 values) of the parameter at hand. In this way, for each loading and each core value we get a 95%percentile interval. Such percentile intervals can be interpreted as approximate confidence intervals.

The above procedure was based on computing the loadings and core values in exactly the same way for each bootstrap three-way array. However, this requires that the models are completely identified, in some way or another. Identification of CP or Tucker3 solutions can be done as follows. One of the key features of the CP model is that it is 'essentially' uniquely identified (see Carroll and Chang (1970), Harshman (1970)). By this it is meant that the component matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ resulting from a CP analysis are, under mild assumptions unique up to a joint permutation of the columns of the

three matrices, and up to scaling of the columns of the three matrices. Hence a simple procedure to further identify this solution is to scale components such that the component matrices for two modes (e.g., the first two modes) have unit column sums of squares. A procedure to fix the order of the components is by ordering them such that the column sums of squares decrease. Then it still remains to identify the sign of the component matrices. This can be done in various ways, that are, however, all rather arbitrary (e.g., ensure that the column sums in the component matrices $\mathbf{A}$ and $\mathbf{B}$ are positive).

The Tucker3 model is not at all uniquely identified: As already shown by Tucker (1966), the fit is not affected by arbitrarily multiplying each of the component matrices by a nonsingular square matrix, provided that the core is multiplied appropriately by the inverse of these transformations. Specifically, postmultiplication of the component matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ by nonsingular matrices $\mathbf{S}$, $\mathbf{T}$, and $\mathbf{U}$, respectively, does not affect the model estimates if the core array is multiplied in the appropriate way by $\mathbf{S^{-1}}$, $\mathbf{T^{-1}}$, and $\mathbf{U^{-1}}$, respectively.

To identify the Tucker3 solution, a first commonly used step is to require the component matrices to be columnwise orthonormal, which reduces the transformational nonuniqueness to *rotational* nonuniqueness. A requirement to further identify the solution is to rotate the component matrices to what Kiers (2004) called the "principal axes" orientation of the Tucker3 solution. This identifies the rotation of all component matrices, as well as their permutation.

The principal axes solution has nice theoretical properties, but usually is not easy to interpret. Alternatively, identification can be obtained by some simple structure rotation of the core and component matrices (Kiers (1998)). Such rotations identify the Tucker3 solution up to permutation and scaling. We thus end up in the same situation as with CP, and can hence use the same procedure to obtain full identification (see above).

Above it has been shown how the CP solution and the Tucker3 solution can be identified completely. If we would use exactly the same identification procedure for all bootstrap solutions, then we can compare bootstrap solutions, and sensibly compute percentile intervals, and use these as estimates of confidence intervals for our parameters. However, in doing so, we imply that in our actual data analysis, we consider as our solution only the one that we get from exactly the same identification procedure. As a consequence, if we would have two samples from the same population, and we analyze both in exactly the same way, and it so happens that the solutions are almost identical but have a different ordering of the columns or (in case of Tucker3 analysis) a different rotation of the component matrix at hand, then we would not recognize this near identity of the solutions (as is illustrated by Kiers (2004)). To avoid overlooking such near similarities, we should not take the identifications used for obtaining the bootstrap solutions too seriously, and we should use

procedures that consider bootstrap solutions similar when they *only* differ by permutations or (in case of Tucker3) other nonsingular transformations.

3 Smaller bootstrap intervals using transformations

When, in computing percentile intervals, we wish to consider bootstrap solutions as similar when they differ only by a permutation and scaling, this can be taken into account as follows. Before comparing bootstrap solutions, the components are all reflected and permuted (as far is possible without affecting the fit) in such a way that they optimally resemble the sample solution. In case of Tucker3, also the core should be appropriately reflected and rescaled. For details, the reader is referred to Riu and Bro (2003) or Kiers (2004), who offer two slightly different procedures for achieving this. As an obvious consequence, then the bootstrap solutions will also resemble each other well. For each loading and core value, the associated 500 values in the resulting permuted and reflected bootstrap solutions can then be used to set up a 95%percentile interval. Typically, these intervals will be smaller than the ones based on fully identified solutions. If orderings and scalings are not to be taken seriously (as is often the case in practice), this is indeed desirable, because then the fully identified solutions would lead to artificially wide intervals, as bootstrap solutions that differ mainly in irrelevant ways (i.e., by permutations and/or scalings) would nevertheless be considered as strongly different.

As mentioned above, fully identified Tucker3 bootstrap solutions may differ considerably even when fit preserving transformations exist that make them almost equal. For example, simple structure rotation may lead to very different solutions for two bootstrap samples, when for the original sample two rather different rotations will yield almost the same degree of simplicity; in such cases solutions for some bootstrap samples may, after rotation, resemble one rotated sample solution, while others may resemble the other rotated sample solution, even when, *before rotation*, both would resemble the original *unrotated* sample solution very much. Often, the optimal simplicity of a solution, as such, is not taken seriously (similarly as the actual ordering with respect to sums of squares is usually not taken seriously). Then, it is appropriate to consider as similar all bootstrap solutions that are similar after an optimal transformation towards each other, or to a reference solution. This idea has repeatedly been used in bootstrap or jack-knife procedures for two-way analysis techniques (Meulman and Heiser (1983), Krzanowski (1987), Markus (1994), Milan and Whittaker (1995), Groenen, Commandeur and Meulman (1998)). These two-way techniques cannot as such be used in the three-way situation. For Tucker3, Kiers (2004) proposed two procedures to make three-way bootstrap solutions optimally similar to the sample solution. The first uses 'only' rotational freedom, leaving intact the columnwise orthonormality of the component matrices; the other uses the full transfor-

mational freedom in the Tucker3 model. The basic idea is as follows.
Let a Tucker3 solution be given by $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\underline{\mathbf{G}}$, and bootstrap solutions be indicated by $\mathbf{A^b}$, $\mathbf{B^b}$, $\mathbf{C^b}$ and $\underline{\mathbf{G}}^b$. As in the usual solutions, the component matrices are columnwise orthonormal. Now we want to transform $\mathbf{B^b}$, $\mathbf{C^b}$ and $\underline{\mathbf{G}}^b$ such that they become optimally similar to $\mathbf{B}$, $\mathbf{C}$ and $\underline{\mathbf{G}}$, respectively. Thus, we first search (possibly orthonormal) transformation matrices $\mathbf{T}$ and $\mathbf{U}$, such that $\mathbf{B^bT}$, and $\mathbf{C^bU}$ become optimally similar to $\mathbf{B}$ and, $\mathbf{C}$, respectively. For this purpose, we minimize $f_1(\mathbf{T}) = \| \mathbf{B^bT} - \mathbf{B} \|^2$ and $f_2(\mathbf{U}) = \| \mathbf{C^bU} - \mathbf{C} \|^2$. The inverses of the optimal transformations $\mathbf{T}$ and $\mathbf{U}$ are then appropriately applied to the core array $\underline{\mathbf{G}}^b$. Next, transformational freedom for the first mode (associated with component matrix $\mathbf{A^b}$ and the first mode of the core array) is exploited by transforming the current bootstrap core array across the first mode such that it optimally resembles the sample core array $\underline{\mathbf{G}}$ in the least squares sense. See Kiers (2004) for technical details.
For each loading and core value, the associated 500 values in the resulting transformed bootstrap solutions can be used to set up a 95%percentile interval. These intervals will typically be smaller than both the ones based on fully identified solutions, and the ones based on only permuting and scaling bootstrap solutions, because now similarity across all possible transformations is taken into account.

4 Performance of bootstrap confidence intervals

The above described bootstrap percentile intervals are considered estimates of confidence intervals. This means that, if we have a 95%percentile interval, we would like to conclude from this that with 9% certainty it covers the true population parameter. If we would work with fully identified solutions, then it is clear what the actual population parameters refer to. When transformational freedom is used, the coverage property of our intervals should be that, in 95% of all possible samples from our population, after optimal transformation of the population component matrices and core towards their sample counterparts, the population parameters fall in the confidence intervals we set up. Obviously, for transformation we should read "permutation and scaling", or "orthonormal rotation", if this is the kind of transformation actually used.
By means of a simulation study, Kiers (2004) assessed the coverage properties of the above described bootstrap procedures, both for CP and Tucker3. Specifically, first, large population data sets were constructed according to the three-way model at hand, to which varying amounts of noise were added. The numbers of variables were 4 or 8, and the numbers of conditions were 6 or 20; the numbers of components used varied between 2 and 4. The appropriate three-way solution was computed for the population. Next, samples (of sizes 20, 50, and 100) were drawn from this population, the three-way method at

hand was applied to each sample, and this was followed by a bootstrap procedure set up in one of the ways described above. Finally, for each parameter it was assessed whether the population parameter, after optimal transformation of the population solution towards the sample solution, was covered by the 95%bootstrap confidence interval estimated for it. Across all samples and populations, the percentage of coverages should be 95%, and it was verified whether the actual coverages came near to this percentage.
It was found in the simulation studies that the overall average coverages per method and per type of parameter (variable loading, situation loading, or core), ranged from 92% to 95%, except for Tucker3 using the principal axes solution and making bootstrap solutions comparable to it by only using permutations and reflections (here the worst average coverage was found for the elements of the **B** matrices: 85%). Such overall average coverages are optimistic, because they may have resulted from over- and undercoverages cancelling each other. Therefore, coverage percentages for individual conditions (which are less reliable, because they pertain to smaller numbers of cases) were also inspected. It turned out that these range from roughly 84% to 98%, disregarding the troublesome case mentioned above. The lowest coverage percentages were found for the smallest sample size (20). It can be concluded that, when interpreting the bootstrap percentile intervals as confidence intervals, it should be taken into account that they tend to be too small, especially in case of sample sizes as small as 20.
For a single full bootstrap analysis, 500 three-way component analyses have to be carried out, so it is important to know whether this can be done in reasonable time. Kiers (2004) reported that, for the largest sizes in his study, the Tucker3 bootstrap analyses cost about 30 seconds, which seems acceptable. For CP, however, even for sample sizes of 50, computation time was about 5 minutes. Fortunately, a procedure using the sample solution as start for the bootstrap analyses, can help to decrease this computation time considerably, while not affecting the coverage performance.

5 An application: Bootstrap confidence intervals for results from a Tucker3 Analysis

Kiers and van Mechelen (2001) reported the Tucker3 analysis of the scores of 140 subjects on 14 five-point scales measuring the degree of experiencing various anxiety related phenomena in 11 different stressful situations. The data have been collected by Maes, Vandereycken, and Sutren at the University of Leuven, Belgium. Here we have reanalyzed their data, using the very same options as they used, and now computed bootstrap confidence intervals for the outcomes. Here we used the procedure where the bootstrap component loadings for the anxiety scales and for the situations, and the core array are matched by means of optimal orthogonal rotations to the corresponding sample component matrices and core.

The bootstrap confidence intervals for the loadings of the anxiety scales are given in Table 1, and those for the situations in Table 2. In Table 3, the core values are reported, and, to keep the results insightful, only for the values that play a role in the interpretation, confidence intervals are reported.

Table 1. Confidence intervals for loadings of anxiety scales on components (the latter interpreted as by Kiers and van Mechelen, 2001).

component\ anxiety scale	"Approach-avoidance"		"Autonomic physiology"		"Sickness"		"Excretory need"	
Heart beats faster	-0.12	-0.00	**0.44**	**0.64**	-0.19	0.07	-0.26	-0.08
"Uneasy feeling"	-0.34	-0.19	0.15	0.38	-0.05	0.23	-0.18	0.02
Emotions disrupt	-0.25	-0.09	0.11	0.34	0.10	0.41	-0.15	0.11
Feel exhilarated	**0.41**	**0.52**	0.04	0.20	-0.03	0.17	-0.00	0.15
Not want to avoid	**0.29**	**0.45**	-0.22	-0.03	-0.12	0.14	-0.09	0.10
Perspire	-0.13	-0.02	**0.41**	**0.58**	-0.11	0.11	-0.11	0.05
Need to urinate	-0.04	0.15	0.02	0.35	-0.24	0.23	**0.34**	**0.61**
Enjoy challenge	**0.43**	**0.53**	0.02	0.17	-0.01	0.18	-0.05	0.08
Mouth gets dry	-0.06	0.16	**0.12**	**0.51**	-0.24	0.25	**0.18**	**0.49**
Feel paralyzed	-0.14	0.02	0.01	0.33	0.03	0.44	0.07	0.36
Full stomach	-0.09	0.05	-0.10	0.13	**0.54**	**0.85**	-0.20	0.05
Seek experiences	**0.42**	**0.54**	0.03	0.23	-0.05	0.22	-0.12	0.06
Need to defecate	-0.15	0.03	-0.27	0.10	-0.30	0.25	**0.48**	**0.81**
Feel nausea	-0.19	-0.08	-0.24	-0.03	**0.28**	**0.54**	0.12	0.35

Note: intervals for high loadings used in the original interpretation are set in bold.

The confidence intervals for the anxiety scale loadings vary somewhat in width, but are usually rather small for the highest loadings (which are set in bold face). These values are the ones on which Kiers and van Mechelen (2001) based their interpretation of the components, hence it is comforting to see that these intervals are usually not too wide. The main exceptions are the intervals for the loadings of "Mouth gets dry" on "Autonomic physiology" and "Excretory need", which are both wide, and which indicates that it is not at all clear with which component this anxiety scale is related strongest (which is remarkable, because this refers clearly and solely to an autonomic physiological reaction). Without confidence intervals, this unclarity had gone unnoticed.

The confidence intervals for the situation loadings vary somewhat more in width. Again the intervals for the highest loadings are set in bold face. They are small for "Performance judged by others", but for "Inanimate danger" especially the "Sail boat on rough sea" situation has a wide interval, and both highest loadings on the "Alone in woods at night" component have wide

intervals as well. Clearly, the judgement component is better determined than the other two.

Table 2. Confidence intervals for component values of situations on components (the latter interpreted as by Kiers and van Mechelen, 2001).

component\ situation	Performance judged by others		Inanimate danger		Alone in wood at night	
Auto trip	0.02	0.22	0.01	0.26	-0.29	0.20
New date	0.14	0.31	0.03	0.24	**-0.40**	**-0.03**
Psychological experiment	-0.19	0.23	-0.12	0.42	-0.30	0.72
Ledge high on mountain side	-0.05	0.19	**0.56**	**0.87**	-0.18	0.30
Speech before large group	**0.36**	**0.55**	-0.25	0.02	-0.24	0.12
Consult counseling bureau	0.10	0.45	-0.27	0.12	-0.25	0.53
Sail boat on rough sea	0.02	0.29	**0.27**	**0.70**	-0.34	0.26
Match in front of audience	**0.22**	**0.46**	-0.02	0.26	-0.25	0.24
Alone in woods at night	0.06	0.33	-0.09	0.17	**0.25**	**0.93**
Job-interview	**0.36**	**0.51**	-0.18	0.01	-0.13	0.21
Final exam	**0.38**	**0.53**	-0.29	-0.03	0.02	0.33

Note: intervals for high loadings used in the original interpretation are set in bold.

The core values are used to interpret the components for the individuals indirectly, through the interpretation of the components for the anxiety scales and the situations, see Kiers and van Mechelen (2001). The confidence intervals for the highest core values are usually relatively small. This is even the case for the values just higher than 10 (in absolute sense). The confidence intervals suggest even these smaller values can be taken rather seriously.

6 Discussion

In the present paper, procedures have been described for determining bootstrap percentile intervals for all parameters resulting from a Tucker3 or CP three-way analysis. These can be used as such, that is, as intervals indicating the stability of solutions across resampling from the same data, and hence give an important primary indication of their reliability. However, it was found that the 95%percentile intervals also turn out to be fairly good approximations to 95%confidence intervals in most cases. Thus, they can at least tentatively be used as confidence intervals as well. Some improvement of these intervals, however, still remains to be desired.

Note: core values higher than 10 (in absolute sense) are set in bold.

Table 3. Core array with confidence intervals, in brackets, only for the high core values. Labels A1,. . . ,A6 refer to the 6 components for summarizing the subjects.

Performance judged by others

	Appr.Avoid.	Auto.phys.	Sickness	Excr.need
A1	**36.5** [28, 45]	-1.0	-0.5	-0.2
A2	0.8	1.6	-0.3	2.2
A3	-0.2	0.7	-0.1	**36.9** [29, 43]
A4	-0.9	**40.0** [31, 48]	1.2	1.2
A5	0.5	-0.1	1.2	0.9
A6	-0.3	1.0	**34.9** [28, 43]	0.2

Inanimate Danger

	Appr.Avoid.	Auto.phys.	Sickness	Excr.need
A1	1.6	3.4	2.0	1.0
A2	**30.3** [25, 34]	**-11.0** [−15, −7]	**-11.8** [−15, −8]	-9.0
A3	2.8	3.5	2.4	**15.2** [9, 20]
A4	2.7	**11.2** [5, 16]	0.6	-0.6
A5	0.4	-2.6	1.9	1.9
A6	-0.4	-4.0	6.5	-4.7

Alone in woods at night

	Appr.Avoid.	Auto.phys.	Sickness	Excr.need
A1	2.5	4.3	1.7	-2.2
A2	0.4	-0.4	0.8	2.4
A3	1.6	-0.5	3.9	**12.4** [7, 16]
A4	1.2	5.0	-4.8	-7.0
A5	**26.4** [19, 30]	**-18.4** [−22, −11]	-8.3	-6.6
A6	3.0	1.7	9.8	2.2

Different procedures have been proposed, depending on which transformations one allows for the bootstrap solutions. The choice between these should be made on theoretical grounds, not on empirical grounds. That is, this depends on whether or not the ordering of components in terms of column sums of squares, and the optimal simplicity of solutions in terms of varimax is taken seriously or not. If such characteristics are not taken seriously, indeed one should use all the rotational freedom that is available in setting up bootstrap intervals.

The approximate confidence intervals given here pertain to each individual output parameter. However, obviously, the output parameters are not independent from each other. For instance, already the unit column sums of squares constraints on the component matrices ensure that elements within columns of such matrices depend on each other. Moreover, the optimality of

a solution does not depend on each parameter individually, but on the complete configuration of all output parameters. Thus, one may expect that, if a percentile interval for a particular element of **B**, say, does not contain the population parameter value, then it is rather likely that percentile intervals for other elements of **B** will not cover their population counterparts either. Such dependence even holds for elements from different matrices: Consider that an 'extreme' solution for **B** is found (such that the associated percentile intervals miss most of the population parameters), then this will most likely also affect the solution for the core $\underline{\mathbf{G}}$ (and hence lead to misfitting percentile intervals for many elements of $\underline{\mathbf{G}}$). Clearly, further research is needed to deal with the dependence of output parameters. For now, it suffices to remark that the confidence intervals are each taken as if they 'were on their own', and in interpreting the confidence intervals their dependence should not be overlooked, in particular when they are to be used to make probability statements on sets of parameters jointly.

The bootstrap method is sometimes called a computer intensive method. When we apply it to three-way analysis, indeed, this intensity becomes apparent, especially when using CP. Computation times for moderately sized problems are nonnegligible, although not prohibitive. Some speed improvement was obtained, and further speed improvement may be possible. All in all, however, it can be concluded that the bootstrap now is a viable procedure for estimating confidence intervals for the results from exploratory three-way methods.

References

CARROLL, J.D. and CHANG, J-J. (1970): Analysis of individual differences in multidimensional scaling via an N way generalization of "Eckart Young" decomposition. *Psychometrika, 35, 283–319.*

EFRON, B. and TIBSHIRANI, R.J. (1993): *An introduction to the bootstrap.* New York, Chapman & Hall.

GROENEN, P.J.F., COMMANDEUR, J.J.F. and MEULMAN, J.J. (1998): Distance analysis of large data sets of categorical variables using object weights. *British Journal of Mathematical and Statistical Psychology, 51, 217–232.*

HARSHMAN, R.A. (1970): Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multi-mode factor analysis. *UCLA Working Papers in Phonetics, 16, 1–84.*

KIERS, H.A.L. (1998): Joint orthomax rotation of the core and component matrices resulting from three-mode principal components analysis. *Journal of Classification, 15, 245–263.*

KIERS, H.A.L. (2004): Bootstrap confidence intervals for three-way methods. *Journal of Chemometrics, 18, 22–36.*

KIERS, H.A.L. and VAN MECHELEN, I. (2001): Three-way component analysis: Principles and illustrative application. *Psychological Methods, 6, 84–110.*

KROONENBERG, P.M. and DE LEEUW, J. (1980): Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika, 45, 69–97.*

KRZANOWSKI, W.J. (1987): Cross-validation in principal component analysis. *Biometrics, 43, 575–584.*

MARKUS, M.T. (1994):*Bootstrap confidence regions in nonlinear multivariate analysis*. Leiden, DSWO Press.

MEULMAN, J.J. and HEISER, W.J. (1983): *The display of bootstrap solutions in multidimensional scaling*. Unpublished technical report, University of Leiden, The Netherlands.

Organising the Knowledge Space for Software Components

Claus Pahl

School of Computing,
Dublin City University, Dublin 9, Ireland

Abstract. Software development has become a distributed, collaborative process based on the assembly of off-the-shelf and purpose-built components. The selection of software components from component repositories and the development of components for these repositories requires an accessible information infrastructure that allows the description and comparison of these components.

General knowledge relating to software development is equally important in this context as knowledge concerning the application domain of the software. Both form two pillars on which the structural and behavioural properties of software components can be expressed. Form, effect, and intention are the essential aspects of process-based knowledge representation with behaviour as a primary property.

We investigate how this information space for software components can be organised in order to facilitate the required taxonomy, thesaurus, conceptual model, and logical framework functions. Focal point is an axiomatised ontology that, in addition to the usual static view on knowledge, also intrinsically addresses the dynamics, i.e. the behaviour of software. Modal logics are central here – providing a bridge between classical (static) knowledge representation approaches and behaviour and process description and classification.

We relate our discussion to the Web context, looking at Web services as components and the Semantic Web as the knowledge representation framework.

1 Introduction

The style of software development has changed dramatically over the past decades. Software development has become a distributed, collaborative process based on the assembly of off-the-shelf and purpose-built software components – an evolutionary process that in the last years has been strongly influenced by the Web as a software development and deployment platform.

This change in the development style has an impact on information and knowledge infrastructures surrounding these software components. The selection of components from component repositories and the development of components for these repositories requires an accessible information infrastructure that allows component description, classification, and comparison. Organising the space of knowledge that captures the description of properties and the classification of software components based on these descriptions is central. Discovery and composition of software components based on these

descriptions and classifications have become central activities in the software development process (Crnkovic and Larsson (2002)). In a distributed environment where providers and users of software components meet in electronic marketplaces, knowledge about these components and their properties is essential; a shared knowledge representation language is a prerequisite (Horrocks et al. (2003)). Describing software behaviour, i.e. the effect of the execution of services that a component might offer, is required.

We will introduce an ontological framework for the description and classification of software components that supports the discovery and composition of these components and their services – based on a formal, logical coverage of this topic in (Pahl (2003)). Terminology and logic are the cornerstones of our framework. Our objective is here twofold:

- We will illustrate an ontology based on description logics (a logic underlying various ontology languages), i.e. a logic-based terminological classification framework based on (Pahl (2003)). We exploit a connection to modal logics to address behavioural aspects, in particular the safety and liveness of software systems.
- Since the World-Wide Web has the potential of becoming central in future software development approaches, we investigate whether the Web can provide a suitable environment for software development and what the requirements for knowledge-related aspects are. In particular Semantic Web technologies are important for this context.

We approach the topic here from a general knowledge representation and organisation view, rather than from a more formal, logical perspective.

In Section 2 we describe the software development process in distributed environments in more detail. In Section 3, we relate knowledge representation to the software development context. We define an ontological framework for software component description, supporting discovery and composition, in Section 4. We end with some conclusions in Section 5.

2 The software development process

The World-Wide Web is currently undergoing a change from a document- to a services-oriented environment. The vision behind the Web Services Framework is to provide an infrastructure of languages, protocols, and tools to enable the development of services-oriented software architectures on and for the Web (W3C (2004)). Service examples range from simple information providers, such as weather or stock market information, to data storage support and complex components supporting e-commerce or online banking systems. An example for the latter is an account management component offering balance and transfer services. Service providers advertise their services; users (potential clients of the provider) can browse repository-based marketplaces to find suitable services, see Fig. 1. The prerequisite is a common

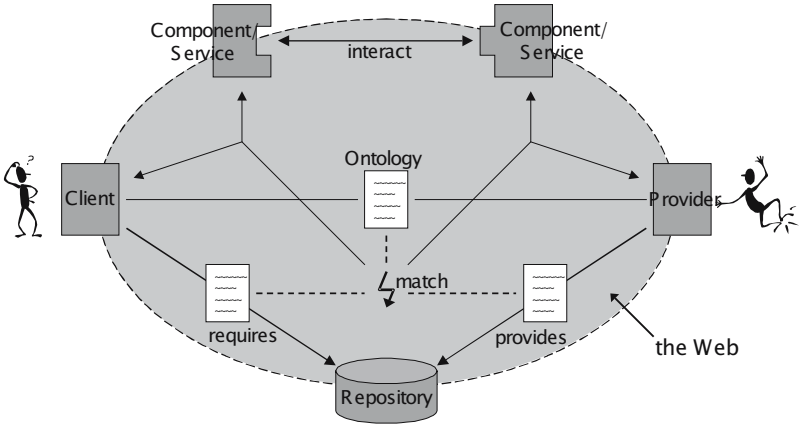


Fig. 1. A Service Component Development Scenario.

language to express properties of these Web-based services and a classification approach to organise these. The more knowledge is available about these services, the better can a potential client determine the suitability of an offer.

Services and components are related concepts. Web services can be provided by software components; we will talk about service components in this case. If services exhibit component character, i.e. are self-contained with clearly defined interfaces that allows them to be reused and composed, then their composition to larger software system architectures is possible. Pluggable and reusable software components are one of the approaches to software developments that promises risk minimisation and cost reduction. Composition can be physical, i.e. a more complex artefact is created through assembly, or logical, i.e. a complex system is created by allowing physically distributed components to interact. Even though our main focus are components in general, we will discuss them here in the context of the Web Services platform.

The ontological description of component properties is our central concern (Fig. 1). We will look at how these descriptions are used in the software development process. Two activities are most important:

- Discovery of provided components (lower half of Fig. 1) in structured repositories. Finding suitable, reusable components for a given development based on abstract descriptions is the problem.
- Composition of discovered components in complex service-based component architectures through interaction (upper half of Fig. 1). Techniques are needed to compose the components in a consistent way based on their descriptions.

For a software developer, the Web architecture means that most software development and deployment activities will take place outside the boundaries of her/his own organisation. Component descriptions can be found in external

repositories. These components might even reside as provided services outside the own organisation. Shared knowledge and knowledge formats become consequently essential.

3 A knowledge space for software development

The Web as a software platform is characterised by different actors, different locations, different organisations, and different systems participating in the development and deployment of software. As a consequence of this heterogeneous architecture and the development paradigm as represented in Fig. 1, shared and structured knowledge about components plays a central role. A common understanding and agreement between the different actors in the development process are necessary.

A shared, organised *knowledge space* for software components in service-oriented architectures is needed. The question how to organise this knowledge space is the central question of this paper. In order to organise the knowledge space through an ontological framework (which we understand essentially as basic notions, a language, and reasoning techniques for sharable knowledge representation), we address three *facets of the knowledge space*: firstly, types of knowledge that is concerned, secondly, functions of the knowledge space, and, finally, the representation of knowledge (Sowa (2000)).

Three *types of knowledge* can be represented in three layers:

- The *application domain* as the basic layer.
- Static and dynamic *component properties* as the central layer.
- Meta-level *activity-related knowledge* about discovery and composition.

We distinguish four *knowledge space functions* (Daconta et al. (2003)) that characterise how knowledge is used to support the development activities:

- *Taxonomy* – terminology and classification; supporting structuring and search.
- *Thesaurus* – terms and their relationships; supporting a shared, controlled vocabulary.
- *Conceptual model* – a formal model of concepts and their relationships; here of the application domain and the software technology context.
- *Logical theory* – logic-supported inference and proof; here applied to behavioural properties.

The third facet deals with how knowledge is represented. In general, *knowledge representation* (Sowa (2000)) is concerned with the description of entities in order to define and classify these. Entities can be distinguished into objects (static entities) and processes (dynamic entities). *Processes* are often described in three *aspects* or *tiers*:

- *Form* – algorithms and implementation – the 'how' of process description

- *Effect* – abstract behaviour and results – the 'what' of process description
- *Intention* – goal and purpose – the 'why' of process description

We have related the aspects form, effect, and intention to software characteristics such as algorithms and abstract behaviour. The service components are software entities that have process character, i.e. we will use this three-tiered approach for their description.

The three facets of the knowledge space outline its structure. They serve as requirements for concrete description and classification techniques, which we will investigate in the remainder.

4 Organising the knowledge space

4.1 Ontologies

Ontologies are means of knowledge representation, defining so-called shared conceptualisations. Ontology languages provide a notation for terminological definitions that can be used to organise and classify concepts in a domain. Combined with a symbolic logic, we obtain a framework for specification, classification, and reasoning in an application domain. Terminological logics such as description logics (Baader et al. (2003)) are an example of the latter.

The Semantic Web is an initiative for the Web that builds up on ontology technology (Berners-Lee et al. (2001)). XML – the eXtensible Markup Language – is the syntactical format. RDF – the Resource Description Framework – is a triple-based formalism (subject, property, object) to describe entities. OWL – the Web Ontology Language – provides additional logic-based reasoning based on RDF.

We use Semantic Web-based ontology concepts to formalise and axiomatise processes, i.e. to make statements about processes and to reason about them. Description logic, which is used to define OWL, is based on concept and role descriptions (Baader et al. (2003)). *Concepts* represent classes of objects; *roles* represent relationships between concepts; and *individuals* are named objects. Concept descriptions are based on primitive logical combinators (negation, conjunction) and hybrid combinators (universal and existential quantification). Expressions of a description logic are interpreted through sets (concepts) and relations (roles).

We use a connection between *description logic* and *dynamic logic* (Sattler et al. (2003), Chapter 4.2.2). A dynamic logic is a modal logic for the description of programs and processes based on operators to express necessity and possibility (Kozen and Tiuryn (1990)). This connection allows us to address safety (necessity of behaviour) and liveness (possibility of behaviour) aspects of service component behaviour by mapping the two modal operators 'box' (or 'always', for safety) and 'diamond' (or 'eventually', for liveness) to the description logic universal and existential quantification, respectively. The central idea behind this connection is that roles can be interpreted as

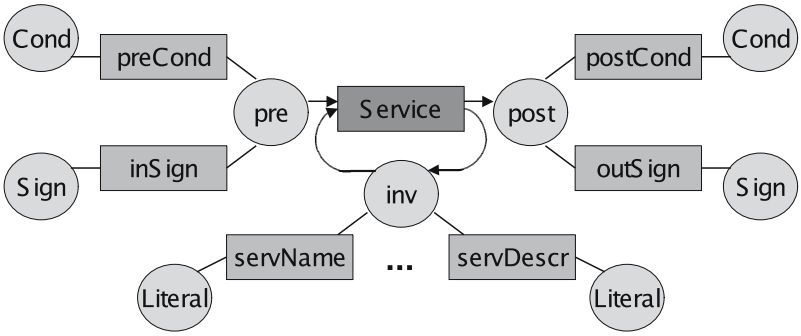


Fig. 2. A Service Component Ontology.

accessibility relations between states, which are central concepts of process-oriented software systems. The correspondence between description logics and a multi-modal dynamic logic is investigated in detail in (Schild (1991)).

4.2 A discovery and composition ontology

An intuitive approach to represent software behaviour in an ontological form would most likely be to consider components or services as the central concepts (DAML-S Coalition (2002)). We, however, propose a different approach. Our objectve is to represent software systems. These systems are based on inherent notions of state and state transition. Both notions are central in our approach. Fig. 2 illustrates the central ideas. Service executions lead from old (pre)states to new (post)states, i.e. the service is represented as a role (a rectangle in the diagram), indicated through arrows. The modal specifications characterise in which state executions might (using the possibility operator to express liveness properties) or should (using the necessity operator to express safety properties) end. For instance, we could specify that a customer may (possibly) check his/her account balance, or, that a transfer of money must (necessarily) result in a reduction of the source account balance. Transitional roles such as *Service* in Fig. 2 are complemented by more static, descriptional roles such as *preCond* or *inSign*, which are associated through non-directed connections. For instance, *preCond* associates a precondition to a prestate; *inSign* associates the type signatures of possible service parameters. Some properties, such as the service name *servName*, will remain invariant with respect to state change.

Central to our approach is the intrinsic specification of process behaviour in the ontology language itself. Behaviour specifications based on the descriptions of necessity and possibility are directly accessible to logic-based methods. This makes reasoning about behaviour of components possible.

We propose a *two-layered ontology* for discovery and composition. The *upper ontology layer* supports *discovery*, i.e. addresses description, search,

discovery, and selection. The *lower ontology layer* supports *composition*, i.e. addresses the assembly of components and the choreography of their interactions. We assume that execution-related aspects are an issue of the provider – shareable knowledge is therefore not required.

Table 1 summarises development activities and knowledge space aspects. It relates the activities discovery, composition, and execution on services (with the corresponding ontologies) to the three knowledge space facets.

Table 1. Development Activities and Knowledge Space Facets.

	Knowledge Aspect	Knowledge Type	Function
Discovery (upper ontology)	intention (terminology)	domain	taxonomy thesaurus
Composition (lower ontology)	effect (behaviour)	component component activities	conceptual model logical theory
Execution	form (implementation)	component	conceptual model

4.3 Description of components

Knowledge describing software components is represented in three layers. We use two ontological layers here to support the abstract properties.

- The *intention* is expressed through assumptions and goals of services in the context of the application domain.
- The *effect* is a contract-based specification of system invariants, pre- and postconditions describing the obligations of users and providers.
- The *form* defines the implementation of service components, usually in a non-ontological, hidden format.

We focus on effect descriptions here. Effect descriptions are based on modal operators. These allow us to describe process behaviour and composition based on the choreography of component interactions. The notion of composition shall be clarified now. Composition in Web- and other service-oriented environments is achieved in a logical form. Components are provided in form of services that will reside in their provider location. Larger systems are created by allowing components to interact through remote operation invocation. Components are considered as independent concurrent processes that can interact (communicate) with each other. Central in the composition are the abstract effect of individual services and the interaction patterns of components as a whole.

We introduce role expressions based on the role constructors sequential composition $R;S$, iteration $!R$, and choice $R+S$ into a basic ontology language to describe interaction processes (Pahl (2003)). We often use $R \circ S$

instead of $R;S$ if R and S are functional roles, i.e. are interpreted by functions – this notation will become clearer when we introduce names and service parameters. Using this language, we can express ordering constraints for parameterised service components. These process expressions constrain the possible interaction of a service component with a client.

For instance, $Login;!(Balance + Transfer)$ is a role expression describing an interaction process of an online banking user starting with a login, then repeatedly executing balance enquiry or money transfer.

An effect specification[1] focussing on safety is for a given system state

$$\forall preCond.positive(Balance(no)) \quad \text{and}$$
$$\forall Transfer.\forall postCond.reduced(Balance(no))$$

saying that if the account balance for account *no* is positive, then money can be transfered, resulting (necessarily) in a reduced balance. *Transfer* is a service; $positive(Balance(no))$ and $reduced(Balance(no))$ are pre- and postcondition, respectively. These conditions are concept expressions. The specification above is formed by navigating along the links created by roles between the concepts in Fig. 2 – *Transfer* replaces *Service* in the diagram.

In Fig. 3, we have illustrated two sample component descriptions – one representing the requirements of a (potential) client, the other representing a provided bank account component. Each component lists a number of individual services (operations) such as *Login* or *Balance*. We have used pseudocode for signatures (parameter names and types) and pre-/postconditions – a formulation in proper description logic will be discussed later on. We have limited the specification in terms of pre- and postconditions to one service, *Transfer*.

The requirements specification forms a query as a request, see Fig. 1. The ontology language is the query language. The composition ontology provides the vocabulary for the query. A query should result ideally in the identification of a suitable (i.e. matching) description of a provided component. In our example, the names correspond – this, however, is in general not a matching prerequisite. Behaviour is the only definitive criterion.

4.4 Discovery and composition of components

Component-based development is concerned with discovery and composition. In the Web context, both activities are supported by Semantic Web and Web Services techniques. They support semantical descriptions of components, marketplaces for the discovery of components based on *intention* descriptions as the search criteria, and composition support based on semantic *effect* descriptions. The deployment of components is based on the *form* description.

[1] This safety specification serves to illustrate effect specification. We will improve this currently insufficient specification (negative account balances are possible, but might not be desired) in the next section when we introduce names and parameters.

Component `AccountRequirements`
signatures and pre-/postconditions
`Login`
inSign `no:int,user:string`
outSign `void`
`Balance`
inSign `no:int`
outSign `real`
`Transfer`
inSign `no:int,dest:int,sum:real`
outSign `void`
preCond `Balance(no) ≥ sum`
postCond `Balance(no) = Balance(no)@pre - sum`
`Logout`
inSign `no:int`
outSign `void`
interaction process
`Login;!Balance;Logout`

Component `BankAccount`
signatures and pre-/postconditions
`Login(no:int,user:string)`
inSign `no:int,user:string`
outSign `void`
`Balance(no:int):real`
inSign `no:int`
outSign `real`
`Transfer(no:int,dest:int,sum:real)`
inSign `no:int,dest:int,sum:real`
outSign `void`
preCond `true`
postCond `Balance(no) = Balance(no)pre - sum`
`Logout(no:int)`
inSign `no:int`
outSign `void`
interaction process
`Login;!(Balance+Transfer);Logout`

Fig. 3. Bank Account Component Service.

Query and Discovery. The aim of the discovery support is to find suitable provided components in a first step that match based on the application domain related goals and that, in a second step, match based on the more technical effect descriptions. Essentially, the ontology language provides a query language. The client specifies the requirements in a repository query in terms of the ontology, which have to be matched by a description of a provided component.

Matching requires technical support, in particular for the formal effect descriptions. Matching can be based on techniques widely used in software development, such as refinement (which is for instance formalised as the consequence notion in dynamic logic). We will focus on the description of effects, i.e. the lower ontology layer (cf. Fig. 2):

- Service component-based software systems are based on a central state concept; additional concepts for auxilliary aspects such as the pre- and poststate-related descriptions are available.
- Service components are behaviourally characterised by transitional roles (for state changes between prestate and poststate) and descriptional roles (auxilliary state descriptions).

Matching and composition. In order to support matching and composition of components through ontology technology, we need to extend the (already process-oriented) ontology language we presented above (Pahl and Casey (2003)). We can make statements about component interaction processes, but we cannot refer to the data elements processed by services. The role expression sublanguage needs to be extended by names (representing data elements) and parameters (which are names passed on to services for processing):

- Names: a name is a role $n[Name]$ defined by the identity relation on the interpretation of an individual n.
- Parameters: a parameterised role is a transitional role R applied to a name $n[Name]$, i.e. $R \circ n[Name]$.

We can make our *Transfer* service description more precise by using a data variable (sum) in pre- and postconditions and as a parameter:

$\forall preCond.(Balance(no) \geq sum)$ and
$\forall Transfer \circ sum[Name].\forall postCond.(Balance(no) = Balance(no)@pre - sum)$

This specification requires *Transfer* to decrease the pre-execution balance by sum.

Matching needs to be supported by a comparison construct. We already mentioned a refinement notion as a suitable solution. This definition, however, needs to be based on the support available in description logics. Subsumption is the central inference technique. Subsumption is the subclass relationship on concept and role interpretations. We define two types of *matching*:

- For *individual services*, we define a *refinement* notion based on weaker preconditions (allowing a service to be invoked in more states) and stronger postconditions (improving the results of a service execution). For example *true* as the precondition and $Balance(no) = Balance(no)@pre - sum$ as the postcondition for $Transfer \circ sum[Name]$ matches, i.e. refines the requirements specification with $Balance(no) >= sum$ as the precondition

and $Balance(no) = Balance(no)@pre - sum$ as the postcondition since it allows the balance to become negative (i.e. allows more flexibility for an account holder).

- For *service processes*, we define a *simulation* notion based on sequential process behaviour. A process matches another process if it can simulate the other's behaviour. For example the expression $Login; !(Balance + Transfer); Logout$ matches, i.e. simulates $Login; !Balance; Logout$, since the transfer service can be omitted.

Both forms of matching are sufficient criteria for subsumption. Matching of effect descriptions is the prerequisite for the composition of services. Matching guarantees the proper interaction between composed service components.

5 Conclusions

Knowledge plays an important role in the context of component- and service-oriented software development. The emergence of the Web as a development and deployment platform for software emphasises this aspect.

We have structured a knowledge space for software components in service-oriented architectures. Processes and their behavioural properties were the primary aspects. We have developed a process-oriented ontological model based on the facets form, effect, and intention. The discovery and the composition of process-oriented service components are the central activities. This knowledge space is based on an ontological framework formulated in a description logic. The defined knowledge space supports a number of different functions – taxonomy, thesaurus, conceptual model, and logical theory. These functions support a software development and deployment style suitable for the Web and Internet environment.

Explicit, machine-processable knowledge is the key to future automation of software development activities. In particular, Web ontologies have the potential to become an accepted format that supports such an automation endeavour.

Acknowledgements

The author is greatly indebted to the anonymous reviewers for their helpful comments and suggestions.

References

BAADER, F., MCGUINESS, D., NARDI, D. and SCHNEIDER, P. (Eds.) (2003): *The Description Logic Handbook*. Cambridge University Press.

BERNERS-LEE, T., HENDLER, J. and LASSILA, O. (2001): The Semantic Web. *Scientific American, 284(5)*.

CRNKOVIC, I. and LARSSON, M. (Eds.) (2002): *Building Reliable Component-based Software Systems*. Artech House Publishers.

DACONTA, M.C., OBRST, L.J. and SMITH, K.T. (2003): *The Semantic Web – A Guide to the Future of XML, Web Services, and Knowledge Management*. Wiley & Sons.

DAML-S COALITION (2002): DAML-S: Web Services Description for the Semantic Web. In: I. Horrocks and J. Hendler (Eds.): *Proc. First International Semantic Web Conference ISWC 2002*. Springer-Verlag, Berlin, 279–291.

HORROCKS, I., MCGUINESS, D. and WELTY, C. (2003): Digital Libraries and Web-based Information Systems. F. Baader, D. McGuiness, D. Nardi and P. Schneider (Eds.): *The Description Logic Handbook*. Cambridge University Press.

KOZEN, D. and TIURYN, J. (1990): Logics of programs. In: J. van Leeuwen (Ed.): *Handbook of Theoretical Computer Science, Vol. B*. Elsevier Science Publishers, 789–840.

PAHL, C. (2003): An Ontology for Software Component Matching. In: *Proc. Fundamental Approaches to Software Engineering FASE'2003*. Springer-Verlag, Berlin, 208–216.

PAHL, C. and CASEY, M. (2003): Ontology Support for Web Service Processes. In: *Proc. European Software Engineering Conference / Foundations of Software Engineering ESEC/FSE'03*. ACM Press.

SATTLER, U., CALVANESE, D. and MOLITOR, R. (2003): Description Logic - Relationships with other Formalisms. In: F. Baader, D. McGuiness, D. Nardi and P. Schneider (Eds.): *The Description Logic Handbook*. Cambridge University Press.

SOWA, J.F. (2000): *Knowledge Representation – Logical, Philosophical, and Computational Foundations*. Brooks/Cole.

SCHILD, K. (1991): A Correspondence Theory for Terminological Logics: Preliminary Report. In *Proc. 12th Int. Joint Conference on Artificial Intelligence*.

W3C – WORLD WIDE WEB CONSORTIUM (2004): *Web Services Framework*. http://www.w3.org/2002/ws.

Multimedia Pattern Recognition in Soccer Video Using Time Intervals

Cees G.M. Snoek* and Marcel Worring

Intelligent Sensory Information Systems, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

Abstract. In this paper we propose the Time Interval Multimedia Event (TIME) framework as a robust approach for recognition of multimedia patterns, e.g. highlight events, in soccer video. The representation used in TIME extends the Allen temporal interval relations and allows for proper inclusion of context and synchronization of the heterogeneous information sources involved in multimedia pattern recognition. For automatic classification of highlights in soccer video, we compare three different machine learning techniques, i.c. C4.5 decision tree, Maximum Entropy, and Support Vector Machine. It was found that by using the TIME framework the amount of video a user has to watch in order to see almost all highlights can be reduced considerably, especially in combination with a Support Vector Machine.

1 Introduction

The vast amount of sport video that is broadcasted on a daily basis, is even for sports enthusiasts too much to handle. To manage the video content, annotation is required. However, manual annotation of video material is cumbersome and tedious. This fact has already been acknowledged by the multimedia research community more than a decade ago, and has resulted in numerous methodologies for automatic analysis and indexing of video documents, see Snoek and Worring (2005).

However, automatic indexing methods suffer from the *semantic gap* or the lack of coincidence between the extracted information and its interpretation by a user, as recognized for image indexing in Smeulders et al. (2000). Video indexing has the advantage that it can profit from combined analysis of visual, auditory, and textual information sources. For this multimodal indexing, two problems have to be unravelled. Firstly, when integrating analysis results of different information channels, difficulties arise with respect to synchronization. The synchronization problem is typically solved by converting all modalities to a common layout scheme, e.g. camera shots, hereby ignoring the layout of the other modalities. This introduces the second problem, namely the difficulty to properly model context, i.e. how to include clues that do not occur at the exact moment of the highlight event of interest? When synchronization and context have been solved, multimodal video indexing might be able to bridge the semantic gap to some extent.

* This research is sponsored by the ICES/KIS MIA project and TNO-TPD.

Existing methods for multimedia pattern recognition, or multimodal video indexing, can be grouped into knowledge based approaches (Babaguchi et al. (2002), Fischer et al. (1995)) and statistical approaches (Assfalg et al. (2002), Han et al. (2002), Lin and Hauptmann (2002), Naphade and Huang(2001)). The former approaches typically combine the output of different multimodal detectors into a rule based classifier. To limit model dependency, and improve the robustness, a statistical approach seems more promising. Various statistical frameworks can be exploited for multimodal video indexing. Recently there has been a wide interest in applying the Dynamic Bayesian Network (DBN) framework for multimedia pattern recognition (Assfalg et al. (2002), Naphade and Huang(2002)). Other statistical frameworks that were proposed include Maximum Entropy (Han et al. (2002)), and Support Vector Machines (Lin and Hauptmann (2002)). However, all of these frameworks suffer from the problems of synchronization and context, identified above. Furthermore, they lack satisfactory inclusion of the textual modality.

To tackle the problems of proper synchronization and inclusion of contextual clues for multimedia pattern recogntion, we propose the Time Interval Multimedia Event (TIME) framework. Moreover, as it is based on statistics, TIME yields a robust approach for multimedia pattern recognition. To demonstrate the viability of our approach we provide a systematic evaluation of three statistical classifiers, using TIME, on the domain of soccer and discuss their performance. The soccer domain was chosen because contextual clues like replays and distinguishing camera movement don't appear at the exact moment of the highlight event. Hence, their synchronization should be taken into account. We improve upon existing work related to soccer video indexing, e.g. Assfalg et al. (2002) and Ekin et al. (2003), by exploiting multimodal, instead of unimodal, information sources, and by using a classifier based on statistics instead of heuristics that is capable to handle both synchronization and context.

The rest of this paper is organized as follows. First we introduce the TIME framework, discussing both representation and classification. Then we discuss the multimodal detectors used for classification of various highlight events in soccer video in section 3. Experimental results are presented in section 4.

2 Multimedia event classification framework

We view a video document from the perspective of its author (Snoek and Worring (2005)). Based on a predefined semantic intention, an author combines certain multimedia layout and content elements to express his message. For analysis purposes this authoring process should be reversed. Hence, we start with reconstruction of layout and content elements. To that end, discrete detectors, indicating the presence or absence of specific layout and content elements, are often the most convenient means to describe the layout and content. This has the added advantage that detectors can be developed

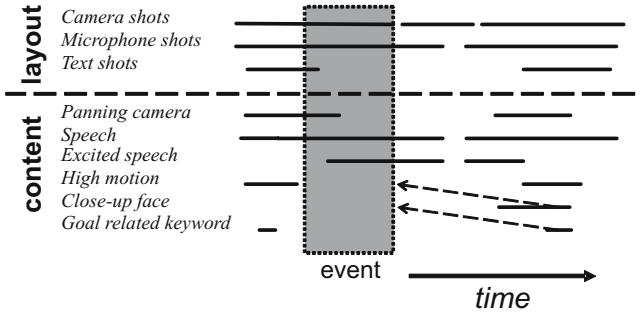


Fig. 1. Detector based segmentation of a multimodal soccer video document into its layout and content elements with a goal event (box) and contextual relations (dashed arrows).

independently of one another. To combine the resulting detector segmentations into a common framework, some means of synchronization is required. To illustrate, consider Fig. 1. In this example a soccer video document is represented by various time dependent detector segmentations, defined on different asynchronous layout and content elements. At a certain moment a goal occurs. Clues for the occurrence of this event are found in the detector segmentations that have a value within a specific position of the time-window of the event, e.g. excited speech of the commentator. But also in contextual detector segmentations that have a value before, e.g. a camera panning towards the goal area, or after the actual occurrence of the event, e.g. the occurrence of the keyword *score* in the time stamped closed caption. Clearly, in terms of the theoretical framework, it doesn't matter exactly what the detector segmentations are. What is important is that we need means to express the different visual, auditory, and textual detector segmentations into one fixed representation without loss of their original layout scheme.

Hence, for automatic classification of a semantic event, ω, we need to grasp a video document into a common pattern representation. In this section we first consider how to represent such a pattern, x, using multimodal detector segmentations and their relations, then we proceed with statistical pattern recognition techniques that exploit this representation for classification using varying complexity.

2.1 Pattern representation

Applying layout and content detectors to a video document results in various segmentations, we define:

Definition 1 (TIME Segmentation) *Decomposition of a video document into one or more series of time intervals, τ, based on a set of multimodal detectors.*

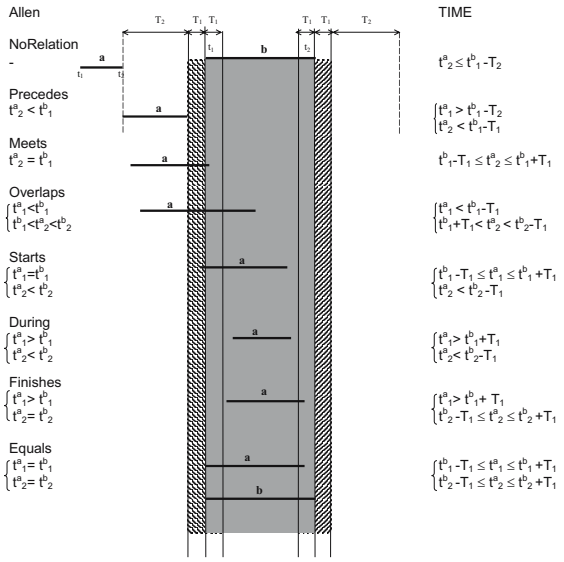


Fig. 2. Overview of the differences between exact Allen relations and TIME relations, extended from Aiello et al. (2002).

To model synchronization and context, we need means to express relations between these time intervals. Allen showed that thirteen relationships are sufficient to model the relationship between any two intervals. To be specific, the relations are: *precedes*, *meets*, *overlaps*, *starts*, *during*, *finishes*, *equals*, and their inverses, identified by adding *_i* to the relation name (Allen (1983)). For practical application of the Allen time intervals two problems occur. First, in video analysis exact alignment of start- or endpoints seldom occurs due to noise. Second, two time intervals will always have a relation even if they are far apart in time. To solve the first problem a fuzzy interpretation was proposed by Aiello et al. (2002) . The authors introduce a margin, T_1, to account for imprecise boundary segmentations, explaining the fuzzy nature. The second problem only occurs for the relations *precedes* and *precedes_i*, as for these the two time intervals are disjunct. Thus, we introduce a range parameter, T_2, which assigns to two intervals the type *NoRelation* if they are too far apart in time. Hence, we define:

Definition 2 (TIME Relations) *The set of fourteen fuzzy relations that can hold between any two elements from two segmentations, τ_1 and τ_2, based on the margin T_1 and the range parameter T_2.*

Obviously the new relations still assure that between two intervals one and only one type of relation exists. The difference between standard Allen relations and TIME relations is visualized in Fig. 2.

Since TIME relations depend on two intervals, we choose one interval as a reference interval and compare this interval with all other intervals. Contin-

uing the example, when we choose a camera shot as a reference interval, the goal can be modelled by a swift camera pan that *starts* the current camera shot, excited speech that *overlaps_i* the camera shot, and a keyword in the closed caption that *precedes_i* the camera shot within a range of 6 seconds. This can be explained because of the time lag between actual occurrence of the event and its mentioning in the closed caption. By using TIME segmentations and TIME relations it now becomes possible to represent events, context, and synchronization in one common framework:

Definition 3 (TIME Representation) *Model of a multimedia pattern x based on the reference interval τ_{ref}, and represented as a set of n TIME relations, with d TIME segmentations.*

In theory, the number of TIME relations, n, is bounded by the number of TIME segmentations, d. Since, every TIME segmentation can be expressed as a maximum of fourteen TIME relations with the fixed reference interval, the maximum number of TIME relations in any TIME representation is equal to $14(d-1)$. In practice, however, a subset can be chosen, either by feature selection techniques (Jain et al. (2000)), experiments, or domain knowledge.

With the TIME representation we are able to combine layout and content elements into a common framework. Moreover, it allows for proper modelling of synchronization and inclusion of context as they can both be expressed as time intervals.

2.2 Pattern classification

To learn the relation between a semantic event ω, and corresponding pattern x, we exploit the powerful properties of statistical classifiers. In standard pattern recognition, a pattern is represented by features. In the TIME framework a pattern is represented by related detector segmentations. In literature a varied gamut of statistical classifiers is proposed, see Jain et al. (2000). We will discuss three classifiers with varying complexity. We start with the C4.5 decision tree (Quinlan (1993)), then we proceed with the Maximum Entropy framework (Jaynes (1957), Berger et al. (1996)), and finally we discuss classification using a Support Vector Machine (Vapnik (2000)).

C4.5 Decision tree The C4.5 decision tree learns from a training set the individual importance of each TIME relation by computing the gain ratio (Quinlan (1993)). Based on this ratio a binary tree is constructed where a leaf indicates a class, and a decision node chooses between two subtrees based on the presence of some TIME relation. The more important a TIME relation is for the classification task at hand, the closer it is located near the root of the tree. Because the relation selection algorithm continues until the entire training set is completely covered, some pruning is necessary to prevent overtraining. Decision trees are considered suboptimal for most applications

(Jain et al. (2000)). However, they form a nice benchmark for comparison with more complex classifiers and have the added advantage that they are easy to interpret.

Maximum Entropy Whereas a decision tree exploits individual TIME relations in a hierarchical manner, the Maximum Entropy (MaxEnt) framework exploits the TIME relations simultaneously. In MaxEnt, first a model of the training set is created, by computing the expected value, E_{train}, of each TIME relation using the observed probabilities $\tilde{p}(x, \omega)$ of pattern and event pairs, (Berger et al. (1996)). To use this model for classification of unseen patterns, we require that the constraints for the training set are in accordance with the constraints of the test set. Hence, we also need the expected value of the TIME relations in the test set, E_{test}. The complete model of training and test set is visualized in Fig. 3. We are left with the problem of finding the optimal reconstructed model, p^*, that finds the most likely event ω given an input pattern x, and that adheres to the imposed constraints. From all those possible models, the maximum entropy philosophy dictates that we select the one with the maximum entropy. It is shown by Berger et al. (1996) that there is always a unique model $p^*(\omega|x)$ with maximum entropy, and that $p^*(\omega|x)$ has a form equivalent to:

$$p^*(\omega|x) = \frac{1}{Z} \prod_{j=1}^{n} \alpha_j^{\tau_j(x,\omega)} \tag{1}$$

where α_j is the weight for TIME relation τ_j and Z is a normalizing constant, used to ensure that a probability distribution results. The values for α_j are computed by the *Generalized Iterative Scaling* (GIS) algorithm (Darroch and Ratcliff (1972)). Since GIS relies on both E_{train} and E_{test} for calculation of α_j, an approximation proposed by Lau et al. (1993) is used that relies only on E_{train}. This allows to construct a classifier that depends completely on the training set. The automatic weight computation is an interesting property of the MaxEnt classifier, since it is very difficult to accurately weigh the importance of individual detectors and TIME relations beforehand.

Support Vector Machine The Support Vector Machine (SVM) classifier follows another approach. Each pattern x is represented in a n-dimensional space, spanned by the TIME relations. Within this relation space an optimal hyperplane is searched that separates the relation space into two different categories, ω, where the categories are represented by $+1$ and -1 respectively. The hyperplane has the following form: $\omega|(\mathbf{w} \cdot x + b)| \geq 1$, where $\mathbf{w}$ is a weight vector, and b is a threshold. A hyperplane is considered optimal when the distance to the closest training examples is maximum for both categories. This distance is called the margin. Consider the example in Fig. 3. Here a two-dimensional relation space consisting of two categories is visualized. The

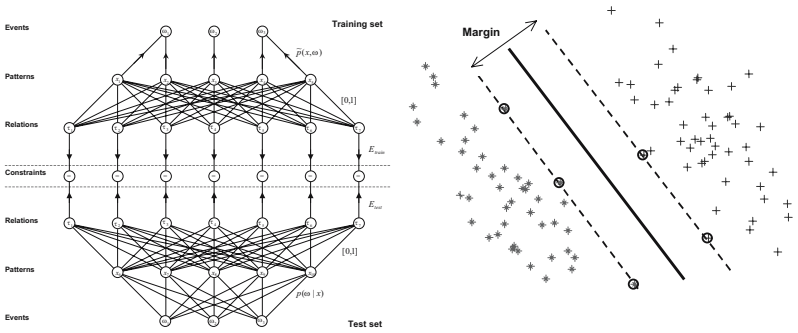


Fig. 3. *(a) Simplified visual representation of the maximum entropy framework. (b) Visual representation of Support Vector Machine framework in two dimensions. The optimal hyperplane is indicated as a thick solid line.*

solid bold line is chosen as optimal hyperplane because of the largest possible margin. The circled data points closest to the optimal hyperplane are called the support vectors. The problem of finding the optimal hyperplane is a quadratic programming problem of the following form (Vapnik (2000)):

$$\min_{\mathbf{w},\xi} \left\{ \frac{1}{2}\mathbf{w} \cdot \mathbf{w} + C\Big(\sum_{i=1}^{l} \xi_i \Big) \right\} \tag{2}$$

Under the following constraints:

$$\omega|(\mathbf{w} \cdot x_i + b)| \geq 1 - \xi_i, \quad i = 1, 2, \ldots, l \tag{3}$$

Where C is a parameter that allows to balance training error and model complexity, l is the number of patterns in the training set, and ξ_i are slack variables that are introduced when the data is not perfectly separable. These slack variables are useful when analyzing multimedia, since results of individual detectors typically include a number of false positives and negatives.

3 Highlight event classification in soccer broadcasts

Important events in a soccer game are scarce and occur more or less random. Examples of such events are goals, penalties, yellow cards, red cards, and substitutions. We define those events as follows:

- *Goal*: the entire camera shot showing the actual goal;
- *Penalty*: beginning of the camera shot showing the foul until the end of the camera shot showing the penalty;
- *Yellow card*: beginning of the camera shot showing the foul until the end of the camera shot that shows the referee with the yellow card;

Table 1. *TIME representation for soccer analysis. T_2 indicates the contextual range used by the precedes and precedes_i relations.*

TIME segmentation	*TIME relations*	T_2 *(s)*
Camera work	*during*	
Person	*during*	
Close-up	*precedes_i*	0 - 40
Goal keyword	*precedes_i*	0 - 6
Card keyword	*precedes_i*	0 - 6
Substitution keyword	*precedes_i*	0 - 6
Excitement	*All relations*	0 - 1
Info block statistics	*precedes_i*	20 - 80
Person block statistics	*precedes_i*	20 - 50
Referee block statistics	*precedes_i*	20 - 50
Coach block statistics	*precedes_i*	20 - 50
Goal block statistics	*precedes_i*	20 - 50
Card block statistics	*precedes_i*	20 - 50
Substitution block statistics	*during*	
Shot length	*during*	

- *Red card*: beginning of the camera shot showing the foul until the end of the camera shot that shows the referee with the red card;
- *Substitution*: beginning of the camera shot showing the player who goes out, until the end of the camera shot showing the player who comes in;

Those events are important for the game and therefore the author adds contextual clues to make the viewer aware of the events. For accurate detection of events, this context should be included in the analysis.

Some of the detectors, used for the segmentation, are soccer specific. Other detectors were chosen based on reported robustness and training experiments. The parameters for individual detectors were found by experimentation using the training set. Combining all TIME segmentations with all TIME relations results in an exhaustive use of relations, we therefore use a subset, tuned on the training set, to prevent a combinatory explosion. For all events, all mentioned TIME segmentations and TIME relations are used, i.e. we used the same TIME representation for all events from the same domain.

The teletext (European closed caption) provides a textual description of what is said by the commentator during a match. This information source was analyzed for presence of informative keywords, like *yellow, red, card, goal, 1-0, 1-2*, and so on. In total 30 informative stemmed keywords were defined for the various events. On the visual modality we applied several detectors. The type of camera work (Baan et al. (2001)) was computed for each camera shot, together with the shot length. A face detector by Rowley et al. (1998) was applied for detection of persons. The same detector formed the basis for a close-up detector. Close-ups are detected by relating the size of detected faces to the total frame size. Often, an author shows a close-up of a player after an event of importance. One of the most informative pieces of information in a soccer broadcast are the visual overlay blocks that give information about the game. We subdivided each detected overlay block as either info, person, referee, coach, goal, card, or substitution block (Snoek

and Worring (2003)), and added some additional statistics. For example the duration of visibility of the overlay block, as we observed that substitution and info blocks are displayed longer on average. Note that all detector results are transformed into binary output before they are included in the analysis. From the auditory modality the excitement of the commentator is a valuable resource. For the proper functioning of an excitement detector, we require that it is insensitive to crowd cheer. This can be achieved by using a high threshold on the average energy of a fixed window, and by requiring that an excited segment has a minimum duration of 4 seconds.

We take the result of automatic shot segmentation as a reference interval. An overview of the TIME representation for the soccer domain is summarized in Table 1. In the next section we will evaluate the automatic indexing of events in soccer video, based on the presented pattern representation.

4 Evaluation

For the evaluation of the TIME framework we recorded 8 live soccer broadcasts, about 12 hours in total. We used a representative training set of 3 hours and a test set of 9 hours. In this section we will first present the evaluation criteria used for evaluating the TIME framework, then we present and discuss the classification results obtained.

4.1 Evaluation criteria

The standard measure for performance of a statistical classifier is the error rate. However, this is unsuitable in our case, since the amount of relevant events are outnumbered by irrelevant pieces of footage. An alternative is to use the precision and recall measure adapted from information retrieval. This measure gives an indication of correctly classified highlight events, falsely classified highlight events, and missed highlight events. However, since highlight events in a soccer match can cross camera shot boundaries, we merge adjacent camera shots with similar labels. As a consequence, we loose our arithmetic unit. Therefore, precision and recall can no longer be computed. As an alternative for precision we relate the total duration of the segments that are retrieved to the total duration of the relevant segments. Moreover, since it is unacceptable from a users perspective that scarce soccer events are missed, we strive to find as many events as possible in favor of an increase in false positives. Finally, because it is difficult to exactly define the start and end of an event in soccer video, we introduce a tolerance value T_3 (in seconds) with respect to the boundaries of detection results. We used a T_3 of 7 s. for all soccer events. A merged segment is considered relevant if one of its boundaries plus or minus T_3 crosses that of a labelled segment in the ground truth.

Table 2. *Evaluation results of the different classifiers for soccer events, where duration is the total duration of all segments that are retrieved.*

	Ground truth		*C4.5*		*MaxEnt*		*SVM*	
	Total	Duration	Relevant	Duration	Relevant	Duration	Relevant	Duration
Goal	12	3^m07^s	2	2^m40^s	10	10^m14^s	11	11^m52^s
Yellow Card	24	10^m35^s	13	14^m28^s	22	26^m12^s	22	12^m31^s
Substitution	29	8^m09^s	25	15^m27^s	25	7^m36^s	25	7^m23^s
$\sum$	65	21^m51^s	40	32^m35^s	57	44^m02^s	58	31^m46^s

4.2 Classification results

For evaluation of TIME on the soccer domain, we manually labelled all the camera shots as either belonging to one of four categories: yellow card, goal, substitution, or unknown. Red card and penalty were excluded from analysis since there was only one instance of each in the data set. For all three remaining events a C4.5, MaxEnt, and SVM classifier was trained. Results on the test set are visualized in Table 2.

When analyzing the results, we clearly see that the C4.5 classifier performs worst. Although it does a good job on detection of substitutions, it is significantly worse for both yellow cards and goals when compared to the more complex MaxEnt and SVM classifiers. When we compare results of MaxEnt and SVM, we observe that almost all events are found independent of the classifier used. The amount of video data that a user has to watch before finding those events is about two times longer when a MaxEnt classifier is used, and about one and a half times longer when a SVM is used, compared to the best case scenario. This is a considerable reduction of watching time when compared to the total duration, 9 hours, of all video documents in the test set. With the SVM we were able to detect one extra goal, compared to MaxEnt. Analysis of retrieved segments learned that results of Maximum Entropy and SVM are almost similar. Except for goal events, where nine events were retrieved by both, the remaining classified goals were different for each classifier.

When we take a closer look to the individual results of the different classifiers, it is striking that C4.5 can achieve a good result on some events, e.g. substitution, while performing bad on others, e.g. goal. This can, however, be explained by the fact that the events where C4.5 scores well, can be detected based on a limited set of TIME relations. For substitution events in soccer an overlay during the event is a very strong indicator. When an event is composed of several complex TIME relations, like goal, the relatively simple C4.5 classifier performs worse than both complex MaxEnt and SVM classifiers.

To gain insight in the meaning of complex relations in the soccer domains, we consider the GIS algorithm from section 2.2, which allows to compute the importance or relative weight of the different relations used. The weights computed by GIS indicate that for the soccer events goal and yellow card specific keywords in the closed captions, excitement with during and overlaps

Fig. 4. *The Goalgle soccer video search engine.*

relations, a close-up afterwards, and the presence of an overlay nearby are important relations.

Overall, the SVM classifier achieves comparable or better results than MaxEnt. When we analyze false positives for both classifiers, we observe that those are caused because some of the important relations are shared between different events. This mostly occurs when another event is indeed happening in the video, e.g. a hard foul or a scoring chance. False negatives are mainly caused by the fact that a detector failed. By increasing the number of detectors and relations in our model we might be able to reduce those false positives and false negatives.

5 Conclusion

To bridge the semantic gap for multimedia event classification, a new framework is required that allows for proper modelling of context and synchronization of the heterogeneous information sources involved. We have presented the Time Interval Multimedia Event (TIME) framework that accommodates those issues, by means of a time interval based pattern representation. Moreover, the framework facilitates robust classification using various statistical classifiers.

To demonstrate the effectiveness of TIME it was evaluated on the domain of soccer. We have compared three different statistical classifiers, with varying complexity, and show that there exists a clear relation between narrowness of the semantic gap and the needed complexity of a classifier. When there exists a simple mapping between a limited set of relations and the semantic concept we are looking for, a simple decision tree will give comparable results as a more complex SVM. When the semantic gap is wider, detection will profit from combined use of multimodal detector relations and a more complex classifier, like the SVM. Results show that a considerable reduction of watching time can be achieved. The indexed events were used to build the *Goalgle* soccer video search engine, see Fig. 4.

References

AIELLO, M. et al. (2002): Document understanding for a broad class of documents. *Int'l Journal on Document Analysis and Recognition, 5/1, 1–16.*

ALLEN. J.F. (1983): Maintaining knowledge about temporal intervals. *Communications of the ACM 26/11, 832–843.*

ASSFALG, J. et al. (2002): Soccer highlights detection and recognition using HMMs. In: *IEEE Int'l Conf. on Multimedia & Expo*, Lausanne, Switzerland.

BAAN, J. et al. (2001): Lazy users and automatic video retrieval tools in (the) lowlands. In: *Proc. of the 10th Text REtrieval Conf.* Gaithersburg, USA.

BABAGUCHI, N. et al. (2002): Event based indexing of broadcasted sports video by intermodal collaboration. *IEEE Trans. on Multimedia, 4/1, 68–75.*

BERGER, A. et al. (1996): A maximum entropy approach to natural language processing. *Computational Linguistics, 22/1, 39–71.*

DARROCH, J.N. and RATCLIFF, D. (1972): Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics, 43/5, 1470–1480.*

EKIN, A. et al. (2003): Automatic soccer video analysis and summarization. *IEEE Trans. on Image Processing, 12/7, 796–807.*

FISCHER, S. et al. (1995): Automatic recognition of film genres. In: *ACM Multimedia*, San Francisco, USA, 295–304.

HAN, M. et al. (2002): An integrated baseball digest system using maximum entropy method. In: *ACM Multimedia*, Juan-les-Pins, France.

JAIN, A.K. et al. (2000): Statistical pattern recognition: A review. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 22/1, 4–37.*

JAYNES, E.T. (1957): Information theory and statistical mechanics. *The Physical Review, 106/4, 620–630.*

LAU, R. et al. (1993): Adaptive language modelling using the maximum entropy approach. In: *ARPA Human Language Technologies Workshop* Princeton, USA, 81–86.

LIN, W.-H. and HAUPTMANN, A.G. (2002): News video classification using SVM-based multimodal classifiers and combination strategies. In: *ACM Multimedia*, Juan-les-Pins, France.

NAPHADE, M.R. and HUANG, T.S. (2001): A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Trans. on Multimedia, 3/1, 141–151.*

QUINLAN, J.R. (1993): *C4.5: Programs for Machine Learning.* Morgan Kaufmann.

ROWLEY, H.A. et al. (1998): Neural network-based face detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 20/1, 23–38.*

SMEULDERS, A.W.M. et al. (2000): Content based image retrieval at the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 22/12, 1349–1380.*

SNOEK, C.G.M. and WORRING, M. (2003): Time interval maximum entropy based event indexing in soccer video. In *IEEE Int'l Conf. on Multimedia & Expo*, volume 3, Baltimore, USA, 481–484.

SNOEK, C.G.M. and WORRING, M. (2005): Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications.* In press.

VAPNIK, V.N. (2000): *The Nature of Statistical Learning Theory*, 2th ed. Springer, New York.

Quantitative Assessment of the Responsibility for the Disease Load in a Population

Wolfgang Uter and Olaf Gefeller

Department of Medical Informatics, Biometry and Epidemiology,
University of Erlangen Nuremberg, Germany

Abstract. The concept of attributable risk (AR), introduced more than 50 years ago, quantifies the proportion of cases diseased due to a certain exposure (risk) factor. While valid approaches to the estimation of crude or adjusted AR exist, a problem remains concerning the attribution of AR to each of a set of several exposure factors. Inspired by mathematical game theory, namely, the axioms of fairness and the Shapley value, introduced by Shapley in 1953, the concept of partial AR has been developed. The partial AR offers a unique solution for allocating shares of AR to a number of exposure factors of interest, as illustrated by data from the German Göttingen Risk, Incidence, and Prevalence Study (G.R.I.P.S.).

1 Introduction

Analytical epidemiological studies aim at providing quantitative information on the association between a certain exposure, or several exposures, and some disease outcome of interest. Usually, the disease etiology under study is multifactorial, so that several exposure factors have to be considered simultaneously. The effect of a particular exposure factor on the dichotomous disease variable is quantified by some measure of association, including the relative risk (RR) or the odds ratio (OR), which will be explained in the next section.

While these measures indicate by which factor the disease risk increases if a certain exposure factor is present in an individual, the concept of attributable risk (AR) addresses the impact of an exposure on the overall disease load in the population. This paper focusses on the AR, which can be informally introduced as the answer to the question, "what proportion of the observed cases of disease in the study population suffers from the disease due to the exposure of interest?". In providing this information the AR places the concept of RR commonly used in epidemiology in a public health perspective, namely by providing an answer also to the reciprocal question, "what proportion of cases of disease could — theoretically — be prevented if the exposure factor could be entirely removed by some adequate preventive action?".

Since its introduction in 1953 (Levin (1953)), the concept of AR is increasingly being used by epidemiological researchers. However, while the methodology of this invaluable epidemiological measure has constantly been extended to cover a variety of epidemiological situations, its practical use has not followed these advances satisfactorily (reviewed by Uter and Pfahlberg (1999)).

One of the difficulties in applying the concept of AR is the question of how to adequately estimate the AR associated with several exposure factors of interest, and not just one single exposure factor. The present paper briefly introduces the concept of sequential attributable risk (SAR) and then focusses on the partial attributable risk (PAR), following an axiomatic approach founded on game theory. For illustrative purposes, data from a German cohort study on risk factors for myocardial infarction are used.

2 Basic definitions of attributable risk

Suppose a population can be divided into an exposed subpopulation ($E = 1$) and an unexposed one ($E = 0$), as well as a diseased part ($D = 1$) and a non-diseased one ($D = 1$). Denote $P(A)$ the probability that a randomly chosen subject from this population belongs to subpopulation A, and $P(A|B)$ the corresponding conditional probability of A given B.
The definition of the relative risk (RR) is then as follows:

$$RR := \frac{P(D = 1|E = 1)}{P(D = 1|E = 0)}. \tag{1}$$

Another, well-established measure of individual risk is the odds ratio (OR), which compares the odds of being diseased instead of the risk of being diseased between the exposed ($E = 1$) and the unexposed ($E = 0$) subpopulation:

$$OR := \frac{P(D = 1|E = 1)/P(D = 0|E = 1)}{P(D = 1|E = 0)/P(D = 0|E = 0)}. \tag{2}$$

The definition of attributable risk, in contrast, is as follows (for more formal details see Eide and Heuch (2001)) :

$$AR := \frac{P(D = 1) - P(D = 1|E = 0)}{P(D = 1)}. \tag{3}$$

Alternatively, the AR can be expressed in algebraically equivalent forms, as originally introduced by Levin (1953) as

$$AR = \frac{P(E = 1) * [RR - 1]}{P(E = 1) * [RR - 1] + 1}. \tag{4}$$

or, as defined by Miettinen (1974),

$$AR = P(E = 1|D = 1) * \frac{RR - 1}{RR}. \tag{5}$$

As can be seen from these definitions, the AR depends both on the individual risk (RR) and on the exposure prevalence ($P(E = 1)$): the larger the RR, the larger the AR will be, given a fixed $P(E = 1)$, and the higher the exposure prevalence, the larger the AR will be, given a certain RR. Moreover, a

certain AR may result from different scenarios — a rare exposure associated with a high individual (relative) risk, or a common, but weak risk factor. Knowledge of the underlying scenario, and not only of the AR alone, is important for public health decisions regarding intervention strategies: in the first case, a targeted approach aiming at the small, identifiable subgroup at high risk would be appropriate, while in the latter case, a "population strategy" offering intervention for nearly the whole population would be more advisable.

3 Crude and adjusted attributable risk

The maximum likelihood estimator of AR can be easily obtained in 2×2 tables from cohort and cross-sectional studies by substituting sample proportions for the respective probabilities in (1) leading to what has been termed *crude estimators of the AR* (Walter (1976)). Some additional approximations (i.e., replacing RR by OR in (3)) is necessary for the case-control design (Whittemore (1982), Benichou (1991)).

However, often we face a multifactorial etiology of disease, some of these factors being potential confounders for the impact of one certain factor of interest. In this situation, crude estimates of the AR derived from a $D \times E$ contingency table will be biased. If only one exposure factor is of interest in terms of AR estimation, confounding of this estimate can be overcome by calculating an adjusted AR:

$$AR_{adj} := \frac{P(D=1) - \sum_i P(D=1|E=0, C=i)}{P(D=1)}. \tag{6}$$

where C denotes the stratum variable formed by the combination of all other L exposures considered as nuisance factors. This AR adjusted for the total effect of all L nuisance factors may be interpreted as the proportion of the diseased population that is potentially preventable if the risks of disease in the exposed sub-populations were changed to the risks of the unexposed ($E = 0$) population in all C strata of the adjusting variables. Estimation of the adjusted AR based on stratification methods (Gefeller (1992)) or on a logistic regression approach (Bruzzi et al. (1985)) have been investigated in detail some time ago already.

Such an approach, however, is only reasonable whenever the specific aim of the study is to evaluate the role of only one particular exposure factor. Otherwise, the implicit hierarchy imposed on the variables involved in the calculation is not justified and another approach to AR estimation is required (Gefeller and Eide (1993)).

4 Sequential attributable risk

As a first step to overcome the limited usefulness of adjusted AR estimates when dealing with the problem of quantifying several ARs of several exposure factors of interest, the sequential AR (SAR) has been suggested. The idea behind the SAR is to consider sequences of exposure variables of interest and quantify the additional effect of one exposure on disease risk after the preceding variables have already been taken into account. For didactic purposes, the approach is outlined below in its basic form ignoring for the moment any hierarchy or grouping of exposure variables as well as additional nuisance variables used exclusively for the purpose of controlling confounding (Eide and Gefeller (1995)).

Suppose a total of the $K+1$ exposure classes are generated by L exposure factors each with K_l+1 exposure categories, i.e., $K+1=\prod_{l=1}^{L}(K_l+1)$. Our interest lies in the potential reduction of cases when preventing the L exposures, one at a time, in a given sequence, for instance starting with exposure no. 1, then exposure no. 2, and so on, until all L exposures are eliminated in the population. A reasonable way to accomplish this will be first to calculate the adjusted AR as shown in the previous section with all exposure factors but the first one included among the adjusting variables. This results in an adjusted AR denoted by $AR_{adj}^{(1)}$ derived from a situation with $\prod_{l=2}^{L}(K_l+1)$ strata and K_1+1 exposure classes. Thereafter, define $AR_{adj}^{(2)}$ as the adjusted attributable risk calculated for the combined effect of first and second exposure variable (creating $(K_1+1)*(K_2+1)$ exposure classes), and the remaining exposures, including the adjustment variables (confounders) forming the $\prod_{l=3}^{L}(K_l+1)$ strata. This stepwise procedure of calculating adjusted AR for different sets of exposure variables can be continued until all L exposures are incorporated among the exposure classes generating variables. The last term of this sequence $AR_{adj}^{(L)}$ corresponds to the total population impact of all L exposures.

Any difference $AR_{adj}^{(r)}-AR_{adj}^{(p)},\ p<r,\ p,r\in\mathcal{N}$, describes the additional effect of considering the $(p+1)st, (p+2)nd, ..., r-th$ exposure after having previously taken into account the effect of the first p exposures in the specified sequence. These differences may be called sequential attributable risks (SAR). Notice that the SAR of a specific exposure factor may differ even for the same set of L exposures according to the sequence of exposure variables considered during the stepwise process of calculation. Hence, the SAR depends on the ordering within the sequence and is not unique for an exposure (for an illustrative example of this property see Gefeller et al. (1998)). Thus, the problem of an unambiguous quantification of the contribution of one exposure to the disease load on a population in a multifactorial situation under the assumption of quasi equal-ranking of factors remains, but in situations where, e.g., a specific sequence of exposure factors targeted in a prevention campaign is given the SAR can be of intrinsic interest (Rowe et al. (2004)).

5 Partial attributable risk

As a solution of the problem of ambiguity, the partial AR (PAR) has been suggested. Originally, the idea has been proposed in a preliminary form by Cox (1987). The PAR is estimated in two steps:

1. by deriving the joint attributable risk for all exposures $E^1, ..., E^L$ under consideration, i.e., the AR for at least one of these exposure factors, and then
2. additively partitioning this quantity into shares for each exposure E^i using an appropriate allocation rule

These resulting shares for E^i are referred to as "partial attributable risk for E^i" $PAR(E^i)$. The development of an appropriate allocation rule has been inspired by game theory. A classical problem in game theory is the following: how can the (momentary) profit that several players have gained by cooperative action in varying coalitions be fairly divided among them? In 1953, Shapley developed a set of axioms for profit division which leads to a unique solution, the Shapley value. The Shapley value averages the contributions of single players to all possible coalitions and is still one of the most common methods of payoff allocation in game theory based on the following assumptions:

- **Efficiency**: Entire value of each coalition must be paid out to members
- **Symmetry**: Payoffs must be independent from order of players
- **Additivity**: Payoff of the sum of two games must be the sum of two separate payoffs
- **Null player**: Any player, whose value added to any coalition is 0, receives 0 payoff in any coalition

Table 1. Comparison of game-theoretic and epidemiological setting

Game theory	Epidemiology
Player P^i	Exposure E^i
Varying coalitions among $P^1, ..., P^L$	Combinations of exposures among $E^1, ..., E^L$
Profit	Risk
Fair division of profit among all players	Allocation of AR to each exposure factor
Shapley value: average of the contribution of a single player P^i to all coalitions	Partial AR: average of all $L!$ sequential ARs of an exposure factor E^i

While the problem of fair profit division of players, and of "fair" allocation of shares of AR to certain exposure factors bears striking similarity (Table 1),

the axioms have to be reformulated. In particular, the axiom of additivity with its clear meaning in typical game-theoretic applications with successive games has no meaningful counterpart in epidemiological applications. Therefore, an algebraically equivalent set of axioms has been derived to be applicable in the epidemiological context including the following "properties of fairness" (Land and Gefeller (1997)):

Marginal rationality ensures a consistent comparison of the population impact of one exposure factor E^i with respect to separate (sub-)populations (denoted by superscripts I and II, respectively). If the AR associated with this risk factor E^i is higher in subpopulation I than in subpopulation II for all combinations with other exposure factors under study, then the attributable share allocated to E^i in subpopulation I should be larger than in subpopulation II. More formally:

$$AR^I(S \cup E^i) - AR^I(S) \geq AR^{II}(S \cup E^i) - AR^{II}(S), \forall S \subset \{E^1, ..., E^L\} \backslash \{E^i\}$$
$$\Rightarrow PAR^I(E^i) \geq PAR^{II}(E^i)$$

Internal marginal rationality ensures a consistent comparison of different exposure factors concerning their respective impact on the disease load in one population, i.e., if the AR associated with a certain risk factor E^i is larger than the AR associated with another risk factor E^k in all corresponding combinations with other exposure factors under study in a given population, then $PAR(E^i)$ should also be larger than $PAR(E^k)$. More formally:

$$AR^I(S \cup E^i) \geq AR^I(S \cup E^k), \forall S \subset \{E^1, ..., E^L\} \backslash \{E^i, E^k\}$$
$$\Rightarrow PAR(E^i) \geq PAR(E^k)$$

Symmetry ensures that the method used for dividing up the joint AR among the L different exposure factors is not influenced by any ordering among the variables. While $SARs$ are not symmetrical, as pointed out above, the PAR is symmetrical by virtue of the averaging process.

Finally, there is exactly one way of partitioning the joint attributable risk for L exposure factors into L single components, which then sum up to the joint AR, which satisfies both marginal rationality and internal marginal rationality as well as symmetry (Land and Gefeller (1997)). Recently, extensions of the concept of PAR have been introduced to address the situation of grouped (hierarchical) exposure variables. In this situation, a "top down" approach of first deriving the $PARs$ for the group variables and then further subdividing these into shares for each single exposure factors must be followed (Land et al. (2001)).

6 Illustrative example: The G.R.I.P.S. Study

Data of the G.R.I.P.S. study (Göttingen Risk, Incidence, and Prevalence Study), a cohort study with 6029 male industrial workers aged 40 to 60, designed to analyze the influence of potential risk factors for myocardial infarction (Cremer et al. (1991)), are used to illustrate the different concepts of

attributable risk. For our purposes, we focus on the effect of the three lipoprotein fractions (LDL-, VLDL- and HDL-cholesterol) and cigarette smoking as exposures of interest, while controlling for age, familiar disposition, alcohol consumption, blood pressure and glucose level as potential confounders. All analyzes were performed with the SAS software package. Table 2 shows estimates for crude AR (derived from the corresponding 2×2 tables), adjusted AR (based on a logistic regression analysis) and partial AR. Note that estimates of precision are omitted. From the comparison of crude and adjusted values it is evident that some confounding of the relationship between lipoprotein exposure variables and the outcome is present, while this is not the case for cigarette smoking. Moreover, in the situation of the G.R.I.P.S. study the PAR for each exposure variable is generally much lower than the corresponding crude and adjusted AR. Due to their construction the $PARs$ given in table 2 reveal an additive property, i.e., the sum of all PAR values equals the total effect of all four exposures measured by the corresponding adjusted AR, according to expression (4) (adjusted for the set of five other, confounding variables quoted above). Consequently, in all situations the sum of PAR values cannot exceed the natural limit of one, which must be regarded as a strong advantage with respect to the interpretation of this measure in a multifactorial setting.

Table 2. Crude, adjusted and partial AR for exposure factors in G.R.I.P.S.

Exposure factor	Definition of "unexposed"	Crude AR	Adjusted AR	Partial AR
LDL-cholesterol	$<$ 160mg/dl	0.612	0.577	0.396
HDL-cholesterol	$>$ 35mg/dl	0.204	0.172	0.072
VLDL-cholesterol	$<$ 30mg/dl	0.217	0.167	0.067
Smoking	nonsmoker	0.371	0.370	0.234
Total effect of all 4 factors (joint AR)		0.803	0.769	0.769

7 Conclusion

The estimation of attributable risks from epidemiological data forms an integral part of modern analytical approaches quantifying the relationship between some binary disease variable and a set of exposure factors. Whereas the relative risk quantifies the impact of exposure factors on an individual level, the AR addresses the impact on a population level. The multifactorial situation usually encountered in epidemiological studies should be reflected in the definition of these risk parameters. The definition of partial attributable risks incorporates the multifactorial nature of the attribution problem and

offers a solution to the task of assigning shares for several exposure factors. Further methodological research will address interval estimation of the PAR to promote its utilization in practical epidemiological studies.

Acknowledgment

This work has been supported by a grant of the Deutsche Forschungsgemeinschaft (grant no. Ge 637/4-1).

References

BENICHOU, J. (1991): Methods of adjustment for estimating the attributable risk in case-control studies: a review. *Statistics in Medicine, 10, 1753–1773.*

BRUZZI, P., GREEN, S.B., BYAR, D.P., BRINTON, L.A. and SCHAIRER, D. (1985): Estimating the population attributable risk for multiple risk factors using case-control data. *American Journal of Epidemiology, 122, 904–914.*

COX, L.A. (1987): A new measure of attributable risk for public health applications. *Management Science, 31, 800–813.*

CREMER, P., NAGEL, D., LABROT, B., MUCHE, R., ELSTER, H., MANN, H. and SEIDEL, D. (1991): *Göttinger Risiko-, Inzidenz- und Prävalenzstudie (GRIPS).* Springer, Berlin.

EIDE, G.E. and GEFELLER, O. (1995): Sequential and average attributable fractions as aids in the selection of preventive strategies. *Journal of Clinical Epidemiology, 48, 645–655.*

EIDE, G.E. and HEUCH, I. (2001): Attributable fractions: fundamental concepts and their visualization. *Statistical Methods in Medical Research, 10, 159–193.*

GEFELLER, O. (1992): Comparison of adjusted attributable risk estimators. *Statistics in Medicine, 11, 2083–2091.*

GEFELLER, O. and EIDE, G.E. (1993): Adjusting attributable risk versus partitioning attributable risk. *Statistics in Medicine, 12, 91–94.*

GEFELLER, O., LAND, M. and EIDE, G.E. (1998): Averaging attributable fractions in the multifactorial situation: assumptions and interpretation. *Journal of Clinical Epidemiology, 51, 437–441.*

LAND, M. and GEFELLER, O. (1997): A game-theoretic approach to partitioning attributable risks in epidemiology. *Biometrical Journal, 39, 777–792.*

LAND, M., VOGEL, C. and GEFELLER, O. (2001): A multifactorial variant of the attributable risk for groups of exposure variables. *Biometrical Journal, 43, 461–481.*

LEVIN, M.L. (1953): The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum, 9, 531–541.*

MIETTINEN, O.S. (1974): Proportion of disease caused or prevented by a given exposure, trait or intervention. *American Journal of Epidemiology, 99, 325–332.*

ROWE, A.K., POWELL, K.E. and FLANDERS, D. (2004): Why Population Attributable Fractions Can Sum to More Than One. *American Journal of Preventive Medicine, 26, 243–249.*

SHAPLEY, L.S. (1953): A value for n-person games. In: H. Kuhn and A. Tucker (Eds.): *Contributions to the theory of games II. Ann. Math. Studies 28.* 307–317.

UTER, W. and PFAHLBERG, A. (1999): The concept of attributable risk in epidemiological practice. *Biometrical Journal, 41, 1–9.*

WALTER, S.D. (1976): The estimation and interpretation of attributable risk in health research *Biometrics, 32, 829–849.*

WHITTEMORE, A.S. (1982): Statistical methods for estimating attributable risk from retrospective data. *Statistics in Medicine, 1, 229–24.*

Part II

Classification and Data Analysis

Bootstrapping Latent Class Models

José G. Dias*

Department of Quantitative Methods,
Instituto Superior de Ciências do Trabalho e da Empresa - ISCTE,
Av. das Forças Armadas, Lisboa 1649–026, Portugal

Abstract. This paper deals with improved measures of statistical accuracy for parameter estimates of latent class models. It introduces more precise confidence intervals for the parameters of this model, based on parametric and nonparametric bootstrap. Moreover, the label-switching problem is discussed and a solution to handle it introduced. The results are illustrated using a well-known dataset.

1 Introduction

The finite mixture model is formulated as follows. Let $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$ denote a sample of size n, with $\mathbf{y}_i$ a J-dimensional vector. Each data point is assumed to be a realization of the random variable $\mathbf{Y}$ with S-component mixture probability density function (p.d.f.) $f(\mathbf{y}_i; \varphi) = \sum_{s=1}^{S} \pi_s f_s(\mathbf{y}_i; \theta_s)$, where the mixing proportions π_s are nonnegative and sum to one, θ_s denotes the parameters of the conditional distribution of component s defined by $f_s(\mathbf{y}_i; \theta_s)$, and $\varphi = \{\pi_1, ..., \pi_{S-1}, \theta_1, ..., \theta_S\}$. In this paper, we focus on the case where S is fixed. Note that $\pi_S = 1 - \sum_{s=1}^{S-1} \pi_s$. The log-likelihood function of the finite mixture model (i.i.d. observations) is $\ell(\varphi; \mathbf{y}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i; \varphi)$, which is straightforward to maximize by the EM algorithm (Dempster et al. (1977)).

This paper focuses on the following question: how accurate is the ML estimator of φ? A natural methodology to answering this question is the bootstrap technique. Bootstrap analysis has been applied in finite mixture modeling mainly to compute uncertainty of parameters by bootstrap standard errors (*e.g.*, de Menezes (1999)). As a result of the difficulties of using likelihood ratio tests for testing the number of components of finite mixtures, another application is the boostrapping of the likelihood ratio statistic (McLachlan and Peel (2000)). A full computation of bootstrap confidence intervals for finite mixture models has not been reported in the literature. In this paper we focus on the latent class model.

The paper is organized as follows: Section 2 gives a short review of the bootstrap technique; Section 3 discusses specific aspects of bootstrap when

* His research was supported by Fundação para a Ciência e Tecnologia Grant no. SFRH/BD/890/2000 (Portugal) and conducted at the University of Groningen (Population Research Centre and Faculty of Economics), The Netherlands. I would like to thank Jeroen Vermunt and one referee for their helpful comments on a previous draft of the manuscript.

applied to finite mixture models; Section 4 illustrates the contributions for the latent class model (finite mixture of conditionally independent multinomial distributions). Section 5 summarizes main results and needs for further research.

2 Bootstrap analysis

The bootstrap is a computer intensive resampling technique introduced by Efron (1979) for assessing among other things standard errors, biases, and confidence intervals, in situations where theoretical statistics are difficult to obtain. The bootstrap technique is easily stated. Suppose we have a random sample $\mathcal{D}$ from an unknown probability distribution F, and we want to estimate the parameter $\varphi = t(F)$. Let $S(\mathcal{D}, F)$ be a statistic. In order to infer, the underlying sampling distribution of $S(\mathcal{D}, F)$ has to be known. The bootstrap method estimates F by some estimate $\hat{F}$ based on $\mathcal{D}$, giving a sampling distribution based on $S(\mathcal{D}^*, \hat{F})$, where the bootstrap sample $\mathcal{D}^* = (\mathbf{y}_1^*, \mathbf{y}_2^*, ..., \mathbf{y}_n^*)$ is a random sample of size n drawn from $\hat{F}$, and $\hat{\varphi}^* = S(\mathcal{D}^*, \hat{F})$ is a bootstrap replication of $\hat{\varphi}$. The bootstrap uses a Monte Carlo evaluation of the properties of $\hat{\varphi}$, repeating sampling, say B times, from $\hat{F}$ to approximate the sampling distribution of $\hat{\varphi}$. The B samples are obtained using the following cycle:

1. Draw a bootstrap sample $\mathcal{D}^{(*b)} = \{\mathbf{y}_i^{(*b)},\ i = 1, \ldots, n\}$, $\mathbf{y}_i^{(*b)} \overset{i.i.d.}{\sim} \hat{F}$;
2. Estimate $\hat{\varphi}^{(*b)} = S(\mathcal{D}^{(*b)}, \hat{F})$ by the plug-in principle.

For an overview of bootstrap methodology, we refer to Efron and Tibshirani (1993). The quality of the approximation depends on the value of B and how close $\hat{F}$ is to distribution F. Efron and Tibshirani (1993, 13) suggest that typical values of B for computing standard errors are in the range from 50 to 200. For example, Albanese and Knott (1994) used 100 replications. For confidence intervals, typical values of B are ≥ 1000 (Efron (1987)). Because we wish to compute more precise confidence intervals for finite mixture models, there is no prior indication on the appropriate number of bootstrap samples. We set $B = 5000$. The application of the bootstrap technique depends on the estimation of F ($\hat{F}$) that can be parametric or nonparametric. Parametric bootstrap assumes a parametric form for F ($\hat{F}_{par}$) and estimates the unknown parameters by their sample quantities. That is, one draws B samples of size n from the parametric estimate of the function F. Nonparametric bootstrap estimates F ($\hat{F}_{nonpar}$) by its nonparametric maximum likelihood estimate, the empirical distribution function which puts equal mass $1/n$ at each observation. Then, sampling from $\hat{F}$ means sampling with replacement from $\mathcal{D}$. In our analyses, we compare results from the nonparametric (NP) and parametric (PAR) versions of the bootstrap.

After generating B bootstrap samples and computing bootstrap (ML) estimates $\{\hat{\varphi}^{(*b)},\ b = 1, ..., B\}$, standard errors, bias, and confidence intervals for φ can easily be computed (Efron and Tibshirani (1986)). Bias and standard error of φ_r are given by $bias^*(\varphi_r)=\hat{\varphi}_r^{(*)} - \hat{\varphi}_r$ and $\sigma^*(\varphi_r)=\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\varphi}_r^{(*b)} - \hat{\varphi}_r^{(*)}\right)^2}$ respectively, with $\hat{\varphi}_r^{(*)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\varphi}_r^{(*b)}$, where r indexes the elements of vector φ.

Standard errors are a crude but useful measure of statistical accuracy, and are frequently used to give approximate confidence intervals for an unknown parameter φ_r given by $\hat{\varphi}_r \pm \hat{\sigma}_r z^{(\alpha)}$, where $\hat{\varphi}_r$ and $\hat{\sigma}_r$ are the ML estimate and the estimated standard error of φ_r respectively, and $z^{(\alpha)}$ is the $100 \times \alpha$ percentile point of a standard normal variate. In our analyses, we compare this approximation using the asymptotic standard error (ASE), the nonparametric bootstrap standard error (BSE_NP), and the parametric bootstrap standard error (BSE_PAR).

The percentile method takes a direct $(1 - \alpha)100\%$ bootstrap confidence interval using the empirical $\alpha/2-$ and $(1 - \alpha/2)-$quantiles of the bootstrap replicates. The BC_a confidence interval improves precision by correcting for bias and nonconstant variance (skewness) and is especially important for asymmetric distributions (Efron (1987); Efron and Tibshirani (1993)). The confidence interval for φ_r with endpoints $\alpha/2$ and $(1 - \alpha/2)$ is $\left((\varphi_r)^*_{BC_a}(\alpha/2), (\varphi_r)^*_{BC_a}(1 - \alpha/2)\right)$, with

$$(\varphi_r)^*_{BC_a}(\alpha) = \hat{G}^{-1}\left(\Phi\left\{z_0 + \frac{z_0 + z^{(\alpha)}}{1 - \hat{a}(z_0 + z^{(\alpha)})}\right\}\right),$$

where $\hat{G}$ is the bootstrap cumulative density function (c.d.f.), Φ is the c.d.f. of the normal distribution, and $\hat{z}_0 = \Phi^{-1}\{\hat{G}(\hat{\varphi}_r)\}$ roughly measures the median bias of $\hat{\varphi}_r$, *i.e.*, the discrepancy between the median of the bootstrap distribution of φ_r and $\hat{\varphi}_r$ (Efron and Tibshirani (1993), 185). The acceleration value $\hat{a}$ (for nonconstant variance) can be estimated using jackknife values (for details, see Efron and Tibshirani (1993), 186). Setting $\hat{a} = 0$, yields the BC confidence interval that corrects bias. Confidence intervals by the percentile method are simpler to compute from the bootstrap distribution by $\left(\hat{G}^{-1}(\alpha/2), \hat{G}^{-1}(1 - \alpha/2)\right)$, but may be less precise, since they assume $\hat{z}_0 = \hat{a} = 0$.

3 Bootstrap analysis in finite mixture models

The complex likelihood function of finite mixture models adds extra difficulties in implementing the bootstrap method due to local optima and nonidentifiability.

For estimating the parameters of the finite mixture model $(\hat{\varphi}^{(*b)})$, one needs to use an iterative process. The EM algorithm is an elegant alternative,

but its success depends on different issues such as starting values. Because the original sample $\mathcal{D}$ and the replicated sample $\mathcal{D}^{(*b)}$ may not differ too much, McLachlan and Peel (2000) suggest the use of the maximum likelihood estimate of φ from $\mathcal{D}$ as a starting value. As stopping rule we set an absolute difference of two successive log-likelihood values smaller than 10^{-6}.

The likelihood function of the finite mixture model is invariant under permutations of the S components, *i.e.*, rearrangement of the component indices will not change the likelihood (label-switching). In bootstrap analysis as well as Bayesian analysis by Markov chain Monte Carlo (MCMC) techniques, a permutation of the components may occur, resulting in the distortion of the distribution of quantities of interest. One way of eliminating this non-identifiability is to define a natural order for each bootstrap sample, based, for example, on $\pi_1^{(*b)} \leq \pi_2^{(*b)} \leq ... \leq \pi_S^{(*b)}$, $b = 1, ..., B$, commonly utilized in Bayesian analysis. We refer to this procedure as the Order strategy. However, it has been shown for Bayesian analysis that also this method can distort the results. Stephens (2000) suggests relabeling or reordering the classes based on the minimization of a function of the posterior probabilities $\alpha_{is}^{(*b)} = \pi_s^{(*b)} f_s(\mathbf{y}_i; \theta_s^{(*b)}) \left[\sum_{h=1}^{S} \pi_h^{(*b)} f_h(\mathbf{y}_i; \theta_h^{(*b)})\right]^{-1}$. It has been shown that this method performs well in comparison with other label-switching methods in the Bayesian setting (Dias and Wedel (2004)). Let $\upsilon_{(*b)}(\varphi^{(*b)})$ define a permutation of the parameters for the bootstrap sample b, and $\mathbf{Q}^{(b-1)} = (q_{is}^{(b-1)})$ be the bootstrap estimation of $\alpha = (\alpha_{is})$, based on the previous $b-1$ bootstrap samples. The algorithm is initialized with a small number of runs, say B^*: $\mathbf{Q}^{(0)} = \left(\frac{1}{B^*} \sum_{m=1}^{B^*} \hat{\alpha}_{is}^{(m)}\right)$. Then, for the bth bootstrap sample, choose $\upsilon_{(*b)}$ to minimize the Kullback-Leibler (KL) divergence between the posterior probabilities $\hat{\alpha}_{is}$ $\left\{\upsilon_{(*b)}(\hat{\varphi}^{(*b)})\right\}$, and the estimate of the posterior probabilities $\mathbf{Q}^{(b-1)}$, and compute $\mathbf{Q}^{(b)}$. For computational details, we refer to Stephens (2000). In MCMC, as a result of the underlying Markov chain, label switching happens less often than in independent situations such as the bootstrap resampling. Therefore, an initial estimate using a small number of B^* bootstrap estimates (without taking into account label switching) may not be appropriate, and a better solution is to take $\mathbf{Q}^{(0)}$ as the MLE solution. This strategy of dealing with the label switching is referred to as KL.

4 An application to the latent class model

The finite mixture of conditionally independent multinomial distributions, also known as a latent class (LC) model, has become a popular technique for clustering and subsequent classification of discrete data (Vermunt and Magidson (2003)). For binary data, let Y_j have 2 categories, *i.e.*, $y_{ij} \in \{0, 1\}$. The latent class model with S latent classes for $\mathbf{y}_i$ is defined by the density $f_s(\mathbf{y}_i; \theta_s) = \prod_{j=1}^{J} \theta_{sj}^{y_{ij}} (1 - \theta_{sj})^{1-y_{ij}}$, where θ_{sj} denotes the parameters of the conditional distribution of component s, $\theta_{sj} = P(Y_{ij} = 1 \mid Z_i = s)$, *i.e.*,

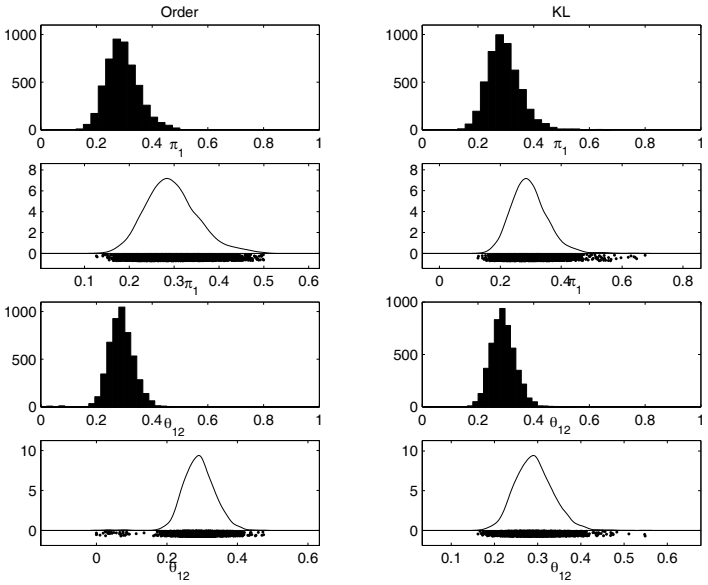


Fig. 1. Effect of the label-switching strategy.

the probability that observation i belonging to component s has category 1 (success) in variable j. This definition assumes conditional independence of the J manifest variables given the latent variable. The estimation of the LC model is straightforward by the EM algorithm.

This application uses the well-known Stouffer-Toby dataset, which has been also used by others (*e.g.*, Albanese and Knott (1994)). It corresponds to 216 observations with respect to whether they tend toward particularistic or universalistic values when confronted by each of four different role conflict situations. We set universalist values as the reference category, and reported conditional probabilities (θ_{sj}) refer to particularistic values. We set $S = 2$ (identified model).

We started the EM algorithm 10 times with random values of the parameters φ from the uniform distribution on [0, 1] for each bootstrap sample. Comparing to starting the EM algorithm from the MLE solution, we concluded that differences are very small, and starting with the MLE solution works well for parametric and nonparametric bootstrap.

Figure 1 depicts the histogram and kernel density estimation of the nonparametric bootstrap distribution of π_1 and θ_{12} for order and KL strategies. For π_1, the order strategy truncates the distribution at 0.5, forcing the relabeling of the components, whereas the KL strategy relabels the components respecting the geometry of the distribution, allowing values above 0.5. In this application, the effect of the order strategy is small, since only 0.64% of the bootstrap estimates of π_1 are above 0.5, and the bootstrap estimates of π_1 by

both procedures have similar values. However, even for a very small number of bootstrap samples, the effect on other parameter estimates can be serious. For example, Figure 1 shows the distribution of the bootstrap estimates of θ_{12}. As can be seen, the order strategy in which π_1 is truncated at 0.5 creates multimodality in the distribution of θ_{12}. This leads to a serious overestimation of the standard error and the confidence interval for θ_{12}. Results presented below are, therefore, based on KL relabeling.

Table 1 reports ML estimates (MLE), the bootstrap mean (BMean), bootstrap median (BMedian), and bootstrap bias (Bias) for nonparametric and parametric estimates. Though most of the parameter estimates present some bias, it is somewhat larger for $\hat{\theta}_{23}$. Bootstrap means and medians tend to be similar, which may indicate similar symmetry of the bootstrap distributions. From the comparison of parametric and nonparametric estimates, we conclude that differences are small.

Table 1. ML estimates, bootstrap mean and median, and bias

	MLE	BMean		BMedian		Bias	
		NP	PAR	NP	PAR	NP	PAR
Class 1							
π_1	0.279	0.295	0.289	0.289	0.285	0.015	0.009
θ_{11}	0.007	0.015	0.015	0.006	0.005	0.008	0.008
θ_{12}	0.074	0.082	0.079	0.076	0.076	0.009	0.006
θ_{13}	0.060	0.070	0.068	0.062	0.062	0.010	0.008
θ_{14}	0.231	0.233	0.228	0.237	0.233	0.002	−0.003
Class 2							
π_2	0.721	0.706	0.711	0.711	0.715	−0.015	−0.009
θ_{21}	0.286	0.291	0.288	0.289	0.287	0.005	0.002
θ_{22}	0.646	0.654	0.652	0.652	0.652	0.007	0.006
θ_{23}	0.646	0.679	0.676	0.677	0.675	0.033	0.030
θ_{24}	0.868	0.876	0.875	0.876	0.874	0.008	0.007

Table 2 presents standard errors and respective 95% confidence intervals. Note that π_2 is not a free parameter, and so asymptotic results are not defined for it. We concluded that standard errors are, in general, similar, however, with slight differences. The relation between them is difficult to generalize. We observe that for θ_{11}, θ_{12}, and θ_{13} the normal approximation does not give accurate results as a consequence of the symmetry of the interval close to the boundary of the parameter space. Another approximation results from applying the logit transformation to the probability parameters defined on $[0, 1]$, *i.e.* $\log[\varphi_r/(1-\varphi_r)] = \psi_r$, $\psi_r \in (-\infty, \infty)$. As an attempt to improve this approximation when bootstrap is applied, one may transform every bootstrap estimate and ML estimates to the logit scale, compute the confidence interval on the logit scale, and finally apply the inverse transformation. We concluded that the logit scale gives poor results. For example, for π_1 the 95%

confidence interval is (0.180, 0.700) and (0.201, 0.671) for nonparametric and parametric bootstrap respectively.

Table 2. Standard errors and 95% confidence intervals

	Standard error			Normal approximation		
	Asymp.	NP	PAR	Asymptotic	BSE (NP)	BSE (PAR)
Class 1						
π_1	0.056	0.061	0.046	(0.169, 0.389)	(0.160, 0.398)	(0.190, 0.369)
θ_{11}	0.025	0.021	0.021	(−0.043, 0.057)	(−0.034, 0.047)	(−0.034, 0.047)
θ_{12}	0.064	0.064	0.058	(−0.052, 0.199)	(−0.052, 0.199)	(−0.040, 0.187)
θ_{13}	0.065	0.060	0.056	(−0.067, 0.187)	(−0.057, 0.177)	(−0.050, 0.171)
θ_{14}	0.093	0.100	0.094	(0.049, 0.413)	(0.036, 0.426)	(0.046, 0.416)
Class 2						
π_2	–	0.061	0.046	–	(0.602, 0.840)	(0.631, 0.810)
θ_{21}	0.039	0.044	0.040	(0.209, 0.363)	(0.201, 0.372)	(0.208, 0.365)
θ_{22}	0.048	0.049	0.049	(0.552, 0.740)	(0.550, 0.742)	(0.551, 0.741)
θ_{23}	0.049	0.052	0.049	(0.550, 0.742)	(0.544, 0.748)	(0.550, 0.742)
θ_{24}	0.038	0.038	0.037	(0.793, 0.942)	(0.793, 0.942)	(0.796, 0.939)

The percentile method and BC_a do not impose the symmetry condition of the previous approximations and respect the parameter space (Table 3). The value of $\hat{a}$ (not shown) is relatively larger (absolute value) for θ_{11}, θ_{12}, θ_{13}, and θ_{24}. The larger skewness of the bootstrap distributions of θ_{11}, θ_{12}, θ_{13}, and θ_{24} leads to a larger correction undertaken by the BC_a confidence interval for these parameters.

Table 3. Bootstrap 95% confidence intervals

	Percentile method		BC_a	
	NP	PAR	NP	PAR
Class 1				
π_1	(0.192, 0.430)	(0.210, 0.389)	(0.177, 0.393)	(0.199, 0.371)
θ_{11}	(0.000, 0.071)	(0.000, 0.070)	(0.000, 0.082)	(0.000, 0.083)
θ_{12}	(0.000, 0.225)	(0.000, 0.205)	(0.000, 0.240)	(0.000, 0.211)
θ_{13}	(0.000, 0.207)	(0.000, 0.191)	(0.000, 0.217)	(0.000, 0.201)
θ_{14}	(0.008, 0.426)	(0.015, 0.401)	(0.001, 0.420)	(0.021, 0.403)
Class 2				
π_2	(0.570, 0.808)	(0.611, 0.790)	(0.607, 0.823)	(0.629, 0.801)
θ_{21}	(0.213, 0.383)	(0.212, 0.371)	(0.210, 0.379)	(0.212, 0.371)
θ_{22}	(0.562, 0.755)	(0.561, 0.750)	(0.544, 0.734)	(0.547, 0.735)
θ_{23}	(0.584, 0.787)	(0.582, 0.774)	(0.572, 0.765)	(0.570, 0.761)
θ_{24}	(0.803, 0.951)	(0.804, 0.948)	(0.773, 0.928)	(0.784, 0.930)

5 Conclusion

This paper proposed and described improved measures for estimation of the statistical accuracy of finite mixture model parameters. To our knowledge, for the first time more precise confidence intervals for the latent class model were computed, avoiding approximations with asymptotic standard errors, or using bootstrap standard errors coupled with normal approximations. Our comparison shows the improvement provided by full bootstrap confidence intervals, namely the BC_a confidence interval. We observed in the application similar results for the parametric and nonparametric bootstrap.

Furthermore, we showed that label-switching strategies are needed to handle the non-identifiability of component labels of finite mixture models. We introduced an adaptation of the Stephens method to the bootstrap methodology that alleviates the effect of hard constraints and respects the geometry of the bootstrap distributions.

Future research could extend our findings to other finite mixture models such as finite mixture of generalized linear models.

References

ALBANESE, M.T. and KNOTT, M. (1994): Bootstrapping Latent Variable Models for Binary Response. *British Journal of Mathematical and Statistical Psychology, 47, 235–246.*

DE MENEZES, L.M. (1999): On Fitting Latent Class Models for Binary Data: The Estimation of Standard Errors. *British Journal of Mathematical and Statistical Psychology, 52, 149–168.*

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society B, 39(1), 1–38.*

DIAS, J.G. and WEDEL, M. (2004): An Empirical Comparison of EM, SEM and MCMC Performance for Problematic Gaussian Mixture Likelihoods. *Statistics & Computing*, 14(4), 323–332.

EFRON, B. (1979): Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics, 7(1), 1–26.*

EFRON, B. (1987): Better Bootstrap Confidence Intervals (with discussion). *Journal of the American Statistical Association, 82(397), 171–200.*

EFRON, B. and TIBSHIRANI, R.J. (1986): Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy (with discussion). *Statistical Science, 1(1), 54–96.*

EFRON, B. and TIBSHIRANI, R.J. (1993): *An Introduction to the Bootstrap.* Chapman & Hall, London.

MCLACHLAN, G.J. and PEEL, D. (2000): *Finite Mixture Models.* John Wiley & Sons, New York.

STEPHENS, M. (2000): Dealing with Label Switching in Mixture Models. *Journal of the Royal Statistical Society B, 62(4), 795–809.*

VERMUNT, J.K. and MAGIDSON, J. (2003): Latent Class Models for Classification. *Computational Statistics & Data Analysis, 41(3-4), 531–537.*

Dimensionality of Random Subspaces

Eugeniusz Gatnar

Institute of Statistics,
Katowice University of Economics, ul. Bogucicka 14, 40-226 Katowice, Poland

Abstract. Significant improvement of classification accuracy can be obtained by aggregation of multiple models. Proposed methods in this field are mostly based on sampling cases from the training set, or changing weights for cases. Reduction of classification error can also be achieved by random selection of variables to the training subsamples or directly to the model. In this paper we propose a method of feature selection for ensembles that significantly reduces the dimensionality of the subspaces.

1 Introduction

Combining classifiers into an ensemble is one of the most interesting recent achievements in statistics aiming at improving accuracy of classification. Multiple models are built on the basis of training subsets (selected from the training set) and combined into an ensemble or a committee. Then the component models determine the predicted class.

Combined classifiers work well if the component models are "weak" and diverse. The term "weak" refers to poorly performing classifiers, that have high variance and low complexity. The diversity of base classifiers is obtained by using different training subsets, assigning different weights to instances or selecting different subsets of features.

Examples of the component classifiers are: classification trees, nearest neighbours, and neural nets.

A number of aggregation methods have been developed so far. Some are based on sampling cases from the training set while others use systems of weights for cases and combined models, or choosing variables randomly to the training samples or directly to the model.

Selecting variables for the training subsamples is the projection of cases into the space of lower dimensionality to the original space. Therefore, reduction of the number of dimensions of the subspaces is an important problem in statistics.

In sections 1-4 of this paper we give a review of model aggregation and feature selection methods for ensembles. Then in section 5 we propose a new method of feature selection for combined models that significantly reduces the classification error. Section 6 contains a brief description of related work in correlation-based feature selection. Results of our experiments are presented in section 7. The last section contains a short summary.

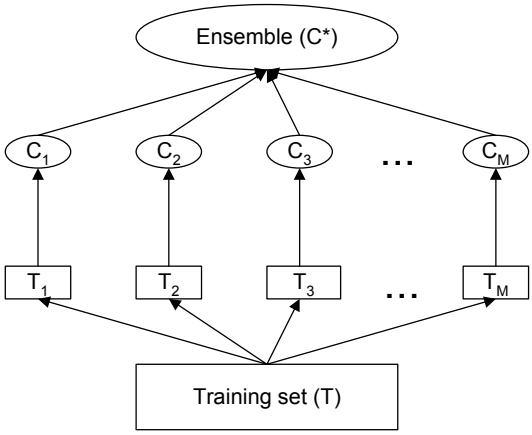


Fig. 1. The aggregation of classification models.

2 Model aggregation

Given a set of training instances:

$$T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}, \tag{1}$$

we form a set of training subsets: $T_1, T_2, \ldots, T_M$ and a classifier is fitted to each of them, resulting in a set of base models: $C_1, C_2, \ldots, C_M$. Then they are combined in some way to produce the ensemble C^*. When component models are tree-based models the ensemble is called a forest. Figure 1 shows process of combining classification models.

Several variants of aggregation methods have been developed that differ in two aspects. The first one is the way that the training subsets are formed on the basis of the original training set. Generally three approaches are used:

- Manipulating training examples: *Windowing* (Quinlan (1993)); *Bagging* (Breiman (1996)); *Wagging* (Bauer and Kohavi (1999)); *Boosting* (Freund and Shapire (1997)) and *Arcing* (Breiman (1998)).
- Manipulating output values: *Adaptive bagging* (Breiman (1999)); *Error-correcting output coding* (Dietterich and Bakiri (1995)).
- Manipulating features (predictors): *Random subspaces* (Ho (1998)); *Random split selection* (Amit and Geman (1997)), (Dietterich (2000)); *Random forests* (Breiman (2001)).

The second aspect is the way that the outputs of base models are combined for the aggregate $C^*(\mathbf{x})$. There are three methods:

- *Majority voting* (Breiman (1996)), when the component classifiers vote for the most frequent class as the predicted class:

$$\hat{C}^*(\mathbf{x}) = argmax_{y \in Y} \left\{ \sum_{m=1}^{M} I(\hat{C}_m(\mathbf{x}) = y) \right\}. \tag{2}$$

- *Weighted voting* (Freund and Schapire (1997)), where predictions of base classifiers are weighted. For example in boosting the classifiers with lower error rate are given higher weights:

$$\hat{C}^*(\mathbf{x}) = argmax_{y \in Y} \left\{ \sum_{m=1}^{M} a_m I(\hat{C}_m(\mathbf{x}) = y) \right\}. \tag{3}$$

  where: $a_m = log\left(\frac{1-e_m}{e_m}\right)$, and e_m is the error rate of the classifier C_m.
- *Stacked generalisation* (Wolpert (1992)) that also uses a system of weights for component models:

$$\hat{C}^*(\mathbf{x}) = \sum_{m=1}^{M} \hat{w}_m \hat{C}_m(\mathbf{x}), \tag{4}$$

  where: $\hat{w}_m = argmin_w \sum_{i=1}^{N} \left\{ y_i - \sum_{m=1}^{M} w_m \hat{C}_m^{-i}(\mathbf{x}_i) \right\}^2$. The models $\hat{C}_m^{-i}(\mathbf{x})$ are fitted to training samples U_m^{-i} obtained by leave-one-out cross-validation (i.e. with i-th observation removed).

3 Random Subspace Method

Ho (1998) introduced a simple aggregation method for classifiers called "Random Subspace Method" (RSM). Each component model in the ensemble is fitted to the training subsample containing all cases from the training set but with randomly selected features. Varying the feature subsets used to fit the component classifiers results in their necessary diversity.

This method is very useful, especially when data are highly dimensional, or some features are redundant, or the training set is small compared to the data dimensionality. Similarly, when the base classifiers suffer from the "curse of dimensionality".

The RSM uses a parallel classification algorithm in contrast to boosting or adaptive bagging that are sequential. It does not require specialised software or any modification of the source code of the existing ones.

A disadvantage of the RSM is the problem of finding the optimal number of dimensions for random subspaces. Ho (1998) proposed to choose half of the available features while Breiman (2001) - the square root of the number of features, or twice the root.

Figure 2 shows the classification error for the committee of trees built for the Satellite dataset (Blake et al. (1998)). The error has been estimated on the appropriate test set. Note that the error starts to rise up after quick decrease, while the number of dimensions of Random Subspaces increases.

We propose to reduce the dimensionality of the subspaces by applying a feature selection to the initial number of variables chosen at random.

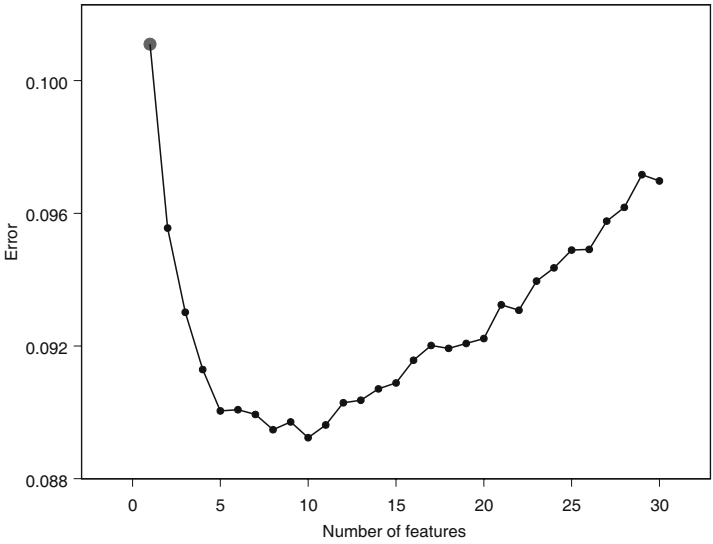


Fig. 2. Effect of the number of features on classification error rate.

4 Feature selection for ensembles

The aim of feature selection is to find the best subset of variables. There are three approaches to feature selection for ensembles:

- *filter methods* that filter undesirable features out of the data before classification,
- *wrapper methods* that use the classification algorithm itself to evaluate the usefulness of feature subsets,
- *ranking methods* that score individual features.

Filter methods are the most common used methods for feature selection in statistics. We will focus on them in the next two sections.

The wrapper methods generate sets of features. Then they run the classification algorithm using features in each set and evaluate resulting models using 10-fold cross-validation. Kohavi and John (1997) proposed a stepwise wrapper algorithm that starts with an empty set of features and adds single features that improve the accuracy of the resulted classifier. Unfortunately, this method is only useful for data sets with relatively small number of features and very fast classification algorithms (e.g. trees). In general, the wrapper methods are computationally expensive and very slow.

The RELIEF algorithm (Kira and Rendell (1992)) is an interesting example of ranking methods for feature selection. It draws instances at random, finds their nearest neighbors, and gives higher weights to features that discriminate the instance from neighbors of different classes. Then those features with weights that exceed a user-specified threshold are selected.

5 Proposed method

We propose to reduce the dimensionality of random subspaces using a filter method based on Hellwig heuristic. The method is a correlation-based feature selection and consists of two steps:

1. Iterate m=1 to M:
   - Choose at random half of the data set features ($L/2$) to the training subset T_m.
   - Determine the best subset F_m of features in T_m according to the Hellwig's method.
   - Grow and prune the tree using the subset F_m.
2. Finally combine the component trees using majority voting.

The heuristic proposed by Hellwig (1969) takes into account both class-feature correlation and correlation between pairs of variables. The best subset of features is selected from among all possible subsets $F_1, F_2, \ldots, F_K$ ($K = 2^L - 1$) that maximises the so-called "integral capacity of information":

$$H(F_m) = \sum_{j=1}^{L_m} h_{mj}, \tag{5}$$

where L_m is the number of features in the subset F_m and h_{mj} is the capacity of information of a single feature x_j in the subset F_m:

$$h_{mj} = \frac{r_{cj}^2}{1 + \sum_{\substack{i=1 \\ i \neq j}}^{L_m} |r_{ij}|}. \tag{6}$$

In the equation (6) r_{cj} is a class-feature correlation, and r_{ij} is a feature-feature correlation.

The correlations r_{ij} are computed using the formula of symmetrical uncertainty coefficient (Press et al. (1988)) based on the entropy function $E(x)$:

$$r_{ij} = 2\left[\frac{E(x_i) + E(x_j) - E(x_i, x_j)}{E(x_i) + E(x_j)}\right]. \tag{7}$$

The measure (7) lies between 0 and 1. If the two variables are independent, then it equals zero, and if they are dependent, it equals unity.

Continuous features have been discretised using the contextual technique of Fayyad and Irani (1993).

6 Related work

Several correlation-based methods of feature selection for ensembles have been developed so far. We can assign them into one of the following groups:

simple correlation-based selection, advanced correlation-based selection, and contextual merit-based methods.

Oza and Tumar (1999) proposed a simple method that belongs to the first group. It ranks the features by their correlations with the class. Then the L features of highest correlation are selected to the model. This approach is not effective if there is a strong feature interaction (multicollinearity).

The correlation feature selection (CFS) method developed by Hall (2000) is advanced because it also takes into account correlations between pairs of features. The set of features F_m is selected to the model that maximizes the value:

$$CFS(F_m) = \frac{L_m \left|\bar{r}_c\right|}{\sqrt{L_m + L_m(L_m - 1)\left|\bar{r}_{ij}\right|}}. \tag{8}$$

where $\bar{r}_c$ is the average feature-class correlation and $\bar{r}_{ij}$ – the average feature-feature correlation.

Hong (1997) developed a method that assigns a merit value to the feature x_i that is the degree to which the other features are capable of classifying the same instances as x_i. The distance between the examples $\mathbf{x_i}$ and $\mathbf{x_j}$ in the set of features F_m is defined as:

$$D_{ij} = \sum_{k=1}^{L_m} d_{ij}^{(k)}, \tag{9}$$

For the categorical feature x_k the component distance is:

$$d_{ij}^{(k)} = \begin{cases} 0 \; if \; x_{k_i} = x_{k_j} \\ 1 \; if \; x_{k_i} \neq x_{k_j} \end{cases} \tag{10}$$

and for a continuous one it is:

$$d_{ij}^{(k)} = min \left\{ \frac{\left|x_{k_i} - x_{k_j}\right|}{t_k}, 1 \right\} \tag{11}$$

where t_k is a feature-dependent threshold (i.e. the half of the range).

The contextual merit of the feature x_k is:

$$CM(x_k) = \sum_{i=1}^{N} \sum_{j \in C(i)} w_{ij} d_{ij}^{(k)}, \tag{12}$$

where $w_{ij} = 1/D_{ij}^2$ if $\mathbf{x_j}$ is one of the K-nearest counter examples to $\mathbf{x_i}$ and $w_{ij} = 0$ otherwise. $C(i)$ in equation (12) is the set of counter examples to $\mathbf{x_i}$ (all instances not in the set of $\mathbf{x_i}$).

7 Experiments

To compare prediction accuracy of ensembles for different feature selection methods we used 9 benchmark datasets from the Machine Learning Repository at the UCI (Blake et al. (1998)).

Results of the comparisons are presented in Table 1. For each dataset an aggregated model has been built containing M=100 component trees[1]. Classification errors have been estimated for the appropriate test sets.

Table 1. Classification errors and dimensionality of random subspaces.

Data set	Single tree (Rpart)	CFS	New method	Average number of features (new method)
DNA	6.40%	5.20%	4.51%	12.3
Letter	14.00%	10.83%	5.84%	4.4
Satellite	13.80%	14.87%	10.32%	8.2
Soybean	8.00%	9.34%	6.98%	7.2
German credit	29.60%	27.33%	26.92%	5.2
Segmentation	3.70%	3.37%	2.27%	3.4
Sick	1.30%	2.51%	2.14%	6.7
Anneal	1.40%	1.22%	1.20%	5.8
Australian credit	14.90%	14.53%	14.10%	4.2

8 Summary

In this paper we have proposed a new correlation-based feature selection method for classifier ensembles that is contextual (uses feature intercorrelations) and based on the Hellwig heuristic. It gives more accurate aggregated models than those built with the CFS correlation-based feature selection method. The differences in classification error are statistically significant at the $\alpha = 0.05$ level (two-tailed t-test).

The presented method also considerably reduces the dimensionality of random subspaces.

References

AMIT, Y. and GEMAN, G. (2001): Multiple Randomized Classifiers: MRCL. *Technical Report*, Department of Statistics, University of Chicago, Chicago.

BAUER, E. and KOHAVI R. (1999): An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. *Machine Learning, 36, 105–142*.

BLAKE, C., KEOGH, E. and MERZ, C. J. (1998): *UCI Repository of Machine Learning Databases*. Department of Information and Computer Science, University of California, Irvine.

[1] In order to grow trees we have used the Rpart procedure written by Therneau and Atkinson (1997) for the S-PLUS and R environment.

BREIMAN, L. (1996): Bagging predictors. *Machine Learning, 24, 123–140.*

BREIMAN, L. (1998): Arcing classifiers. *Annals of Statistics, 26, 801–849.*

BREIMAN, L. (1999): Using adaptive bagging to debias regressions. *Technical Report 547*, Department of Statistics, University of California, Berkeley.

BREIMAN, L. (2001): Random Forests. *Machine Learning 45, 5–32.*

DIETTERICH, T. (2000): An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization. *Machine Learning, 40, 139–158.*

DIETTERICH, T. and BAKIRI, G. (1995): Solving multiclass learning problem via error-correcting output codes. *Journal of Artificial Intelligence Research, 2, 263–286.*

FAYYAD, U.M. and IRANI, K.B. (1993): Multi-interval discretisation of continuous-valued attributes. In: *Proceedings of the XIII International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Francisco, 1022–1027.

FREUND, Y. and SCHAPIRE, R.E. (1997): A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences, 55, 119–139.*

HALL, M. (2000): Correlation-based feature selection for discrete and numeric class machine learning. In: *Proceedings of the 17th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco.

HELLWIG, Z. (1969): On the problem of optimal selection of predictors. *Statistical Revue, 3–4* (in Polish).

HO, T.K. (1998): The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 832–844.*

HONG, S.J. (1997): Use of contextual information for feature ranking and discretization. *IEEE Transactions on Knowledge and Data Engineering, 9, 718–730.*

KIRA, A. and RENDELL, L. (1992): A practical approach to feature selection. In: D. Sleeman and P. Edwards (Eds.): *Proceedings of the 9th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, 249–256.

KOHAVI, R. and JOHN, G.H. (1997): Wrappers for feature subset selection. *Artificial Intelligence, 97, 273–324.*

OZA, N.C. and TUMAR, K. (1999): Dimensionality reduction thrrough classifier ensembles. *Technical Report NASA-ARC-IC-1999-126*, Computational Sciences Division, NASA Ames Research Center.

PRESS, W.H., FLANNERY, B.P., TEUKOLSKY, S.A., VETTERLING, W.T. (1989): *Numerical recipes in Pascal.* Cambridge University Press, Cambridge.

QUINLAN, J.R. (1993): *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo.

THERNEAU, T.M. and ATKINSON, E.J. (1997): *An introduction to recursive partitioning using the RPART routines*, Mayo Foundation, Rochester.

WOLPERT, D. (1992): Stacked generalization. *Neural Networks, 5, 241–259.*

Two-stage Classification with Automatic Feature Selection for an Industrial Application

Sören Hader[1] and Fred A. Hamprecht[2]

[1] Robert Bosch GmbH, FV/PLF2, Postfach 30 02 40
D-70442 Stuttgart, Germany
[2] Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR),
Universität Heidelberg, D-69120 Heidelberg, Germany

Abstract. We address a current problem in industrial quality control, the detection of defects in a laser welding process. The process is observed by means of a high-speed camera, and the task is complicated by the fact that very high sensitivity is required in spite of a highly dynamic / noisy background and that large amounts of data need to be processed online. In a first stage, individual images are rated and these results are then aggregated in a second stage to come to an overall decision concerning the entire sequence. Classification of individual images is by means of a polynomial classifier, and both its parameters and the optimal subset of features extracted from the images are optimized jointly in the framework of a wrapper optimization. The search for an optimal subset of features is performed using a range of different sequential and parallel search strategies including genetic algorithms.

1 Introduction

Techniques from data mining have gained much importance in industrial applications in recent years. The reasons are increasing requirements of quality, speed and cost minimization and the automation of high-level tasks previously performed by human operators, especially in image processing. Since the data streams acquired by modern sensors grow at least as fast as the processing power of computers, more efficient algorithms are required in spite of Moor's law.

The industrial application introduced here is an automated supervision of a laser welding process. A HDRC (High-Dynamic-Range-CMOS) sensor records a welding process on an injection valve. It acquires over 1000 frames with a resolution of 64×64 pixels per second. The aim is to detect welding processes which are characterized by *sputter*, i.e. the ejection of metal particles from the keyhole, see Fig. 1. These events are rare and occur at most once in a batch of 1000 valves. Potential follow-up costs of a missed detection are high and thus a detection with high sensitivity is imperative, while a specificity below 100% is tolerable.

The online handling and processing of the large amounts of raw data is particularly difficult; an analysis becomes possible if appropriate features are extracted which can represent the process. Of the large set of all conceivable

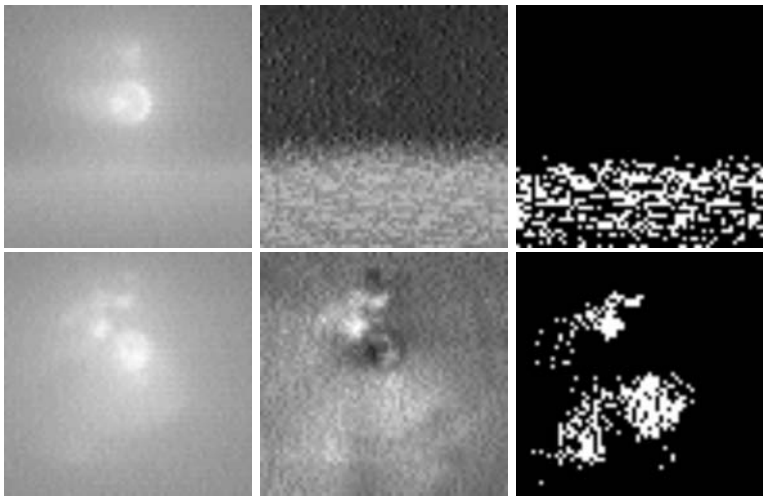


Fig. 1. Top row, left: original frame (64 × 64) from laser welding process, showing a harmless perturbation which should not be detected. Middle: image of the estimated pixel-wise standard deviations, illustrating in which areas the keyhole is most dynamic. Right: pixels which exceed the expected deviation from the mean are marked. Large aggregations of marked pixels are merged to an "object hypothesis". Bottom row: as above, but for original image showing a few sputter that should be detected.

features, we should choose the ones that maximize the classification performance on an entire sequence of images. An exhaustive evaluation of all possible combinations of both features and classifiers is usually too expensive. On the other hand, the recognition performance using a manually chosen feature set is not sufficient in most cases. An intermediate strategy is desired and proposed here: section 2 introduces a two-stage classification system which is optimized using the wrapper approach (section 3) while experimental results are given in section 4.

2 Two-stage classification

2.1 Motivation

While the task is to evaluate the entire sequence of images, we have implemented a divide-and-conquer strategy which focuses on individual images first. In particular, we use a very conservative classifier on individual images: even if there is only a weak indication of an abnormality, the presumed sputter is segmented from the background and stored as an *object hypothesis*. Evidence for a sputter is substantiated only if several such hypotheses appear in consecutive frames.

The advantage of a simple classification in the first stage is the fast evaluation and adaptation of the classifier. The second stage aggregates classifica-

tions derived from individual images into an overall decision with increased reliability.

2.2 First stage – object classification

In the first stage, object hypotheses from single images are extracted and classified.

In particular, an image of pixel-wise means and an image of pixel-wise standard deviations are computed from the entire sequence. Deviations from the mean, which are larger than a constant (e.g. $\in [2.0, 4.0]$) times the standard deviation at that pixel are marked as suspicious (Brocke (2002), Hader (2003)). Sufficiently large agglomerations of suspicious pixels then become an object hypothesis $O_{t,i}$ with indices for time t and object number i. Next, features such as area, eccentricity, intensity, etc. (Teague (1980)) are computed for all object hypotheses.[1] Based on these features, we compute (see section 2.4) an index $d(O_{t,i}) \in [0, 1]$ for membership of object hypothesis $O_{t,i}$ in class "sputter".

2.3 Second stage – image sequence classification

The first stage leaves us with a number of object hypotheses and their class membership indices. Sputters appear in more than one consecutive frame, whereas random fluctuations have less temporal correlation. The second stage exploits this temporal information by aggregating the membership indices into a single decision for the entire sequence as follows: for each frame, we retain only the highest membership index: $d_t := \max_i d(O_{t,i})$. If there is no hypothesis in a frame, the value is set to 0. The d_t can be aggregated using a variety of functions. We use a sliding window located at time t, and apply the $\sum, \prod, \min$ operators to the indices $d_t, \ldots, d_{t+w-1}$ to obtain aggregating functions $a_w(t)$. The length of the time window w is arbitrary, but should be no longer than the shortest sputter event in the training database. The largest value of the aggregate function then gives the decision index for the entire sequence,

$$d_{\text{sequence}} = \max_{t \in T} a_w(t) \tag{1}$$

If d_{sequence} exceeds a threshold Θ, the entire sequence is classified as defective, otherwise as faultless. The optimum value for the threshold Θ depends on the loss function, see section 3.

2.4 Polynomial classifier

The choice of the classifier used in the first stage is arbitrary. We use the polynomial classifier (PC, Schürmann (1996)) which offers a high degree of

[1] This list of features is arbitrarily expandable and previous knowledge on which (subset of) features are useful is not necessary, see section 3.1.

flexibility if sufficiently high degrees are used. Since it performs a least-squares minimization, the optimization problem is convex and its solution unique. Training is by solving a linear system of equations and is faster than that of classifiers like multilayer perceptrons or support vector machines (LeCun et al. (1995)), which is important in case the training is performed repeatedly such as a wrapper optimization (section 3.1). PCs have essentially only one free parameter, the polynomial degree.

In the development stage, a tedious manual labeling of image sequences is required to assemble a training set. Based on an initial training set and the resultant classifier, further sequences can be investigated. The variance of predictions for single object hypotheses can be estimated and those for which a large variance is found can be assumed to be different from the ones already in the training set and added to it. In particular, under a number of assumptions (uncorrelated residuals with zero mean and variance σ^2) the variance of a prediction can be estimated by $\sigma^2 x^T (X^T X)^{-1} x$ where X is the matrix of all explanatory variables (features and monomials formed from these) for all observations in the training set, and x is the new observation (Seeber and Lee (2003)).

3 System optimization

As stated above, sensitivity is of utmost importance in our application, while an imperfect specificity can be afforded. These requirements are met by optimizing the detection threshold Θ such that the overall cost is minimized. The losses incurred by missed detections or false positives are given by $L_{NIO,IO}$ and $L_{IO,NIO}$, respectively, with the former much larger than the latter.

It is customary to arrange the loss function in a matrix as shown below:

$$L = \begin{vmatrix} L_{IO,IO} & L_{IO,NIO} \\ L_{NIO,IO} & L_{NIO,NIO} \end{vmatrix}, \; L_{IO,IO} = L_{NIO,NIO} = 0, \; L_{IO,NIO} \ll L_{NIO,IO}$$

The first index gives the true class, the second one the estimated class, with IO faultless, and NIO defective. The aim is to find a decision function which minimizes the Bayes risk $r = \mathbb{E}\{L\}$. A missed NIO part makes for a large contribution to the risk $\hat{r}$.

The generalization error of a given feature subset and classifier is estimated from the bins that are held out in a k-fold cross-validation (CV). 5- or 10-fold CV is computationally faster than leave-one-out and is a viable choice in the framework of a wrapper algorithm; moreover, these have performed well in a study by Breiman and Spector (1992).

3.1 Wrapper approach

We see great potential in the testing of different feature subsets. In earlier applications the filter approach (which eliminates highly correlated variables

or selects those that correlate with the response) was the first step in finding the relevant features. The filter approach attempts to assess the importance of features from the data alone. In contrast, the *wrapper approach* selects features using the induction algorithm as a black box without knowledge of feature context (Kohavi and John (1997)). The evaluation of a large number of different subsets of features with a classifier is possible only with computationally efficient procedures such as the PC. We use the wrapper approach to simultaneously choose the feature subset, the polynomial degree G, the operator in the aggregation function a, the window width w and the threshold Θ. Evaluating a range of polynomial degrees $1, \ldots, G$ is expensive; in section 3.3 we show how PCs with degree $< G$ can be evaluated at little extra cost.

3.2 Search strategies in feature subsets

The evaluation of all 2^n combinations of n individual features is usually prohibitive. We need smart strategies to get as close as possible to the global optimum without an exhaustive search. Greedy sequential search strategies are among the simplest methods, with two principal approaches, sequential forward selection (SFS) and sequential backward elimination (SBE). SFS starts with an empty set and iteratively selects from the remaining features the one which leads to the greatest increase in performance. Conversely, SBE begins with the complete feature set and iteratively eliminates the feature that leads to the greatest improvement or smallest loss in performance. Both SFS and SBE have a reduced complexity of $\mathcal{O}(n^2)$. Both heuristics can miss the global optimum because once a feature is selected/eliminated, it is never replaced again.

A less greedy strategy is required to reach the global optimum. In particular, locally suboptimal steps can increase the search range. We use a modified BEAM algorithm (Aha and Bankert (1995)) in which not only the best, but the q best local steps are stored in a queue and explored systematically. Deviating from the original BEAM algorithm, we allow either the adding of an unused feature or the exchange of a selected with an unused feature.

Another global optimization method are genetic algorithms (GAs), which represent each feature subset as member of a population. Individuals can mutate (add or lose a feature) and mate with others (partly copy each other's feature subsets), where the probability of mating increases with the predictive performance of the individuals / subsets involved. It is thus possible to find solutions beyond the paths of a greedy sequential search. A disadvantage is the large number of parameters that need to be adjusted and the suboptimal performance that can result if the choice is poor.

3.3 Efficiency

The analysis of the runtime is important to understand the potential of the PC for speed-up. A naive measure of the computational effort is the total

count of multiplications. Although it is just a "quick and dirty" method ignoring memory traffic and other overheads, it provides good predictions.

The coefficient matrix A for the PC is obtained by solving

$$\mathbb{E}\{xx^T\} \cdot A = \mathbb{E}\{xy^T\} \tag{2}$$

with x a column vector specifying the basis functions (i.e. the monomials built from the original features) of an individual observation and y a vector which is $[1\ 0]^T$ for one class and $[0\ 1]^T$ for the other. The expectation values are also called moment matrices.

The computational effort mainly consists of two steps: estimation of the moment matrix $\mathbb{E}\{xx^T\}$ and its inversion. The former requires D^2N multiplications, with N the number of observations and $D = \binom{F+G}{G}$ the dimension of x, that is the feature space obtained by using all F original features as well as all monomials thereof up to degree G.

In CV, the data is partitioned into k bins; accordingly, the $N \times D$ design matrix X can be partitioned into $N_i \times D$ matrices X_i, with $\sum_{i=1}^{k} N_i = N$. The moment matrices are estimated for each bin separately by $X_i^T X_i$. For the jth training in the course of a k-fold CV, the required correlation matrix is obtained from

$$X_{-j}^T X_{-j} = \sum_{i \neq j} X_i^T X_i \tag{3}$$

that is, $D \times D$ matrices are added only.

In summary, while the correlation matrices need to be inverted in each of the k runs in a k-fold CV (requiring a total of $k\frac{2}{3}D^3$ multiplications for a Gauss-Jordan elimination), they are recomputed at the cost of a few additions or subtractions only once the correlation matrices for individual bins have been built (requiring a total of D^2N multiplications).

In addition, once the correlation matrix for a full feature set F and polynomial degree G has been estimated, all moment matrices for $F' \subseteq F$ and $G' \leq G$ are obtained by a mere elimination of appropriate rows and columns.

4 Experimental results

The system has been tested on a dataset of 633 IO and 150 NIO image sequences which comprise a total of 5294 object hypotheses that have been labeled by a human expert. A large part of the IO sequences selected for training were "difficult" cases with sputter look-alikes. The loss function used was $L_{IO,NIO} = 1$ and $L_{NIO,IO} = 100$ and generalization performance was estimated using a single 10-fold CV. A total of 19 features were computed for each object hypothesis. The four subset selection strategies described were tested. For the modified BEAM algorithm, the parameter $q = 5$ and 20 generations were used. The GA ran for 50 generations with 60 individuals each.

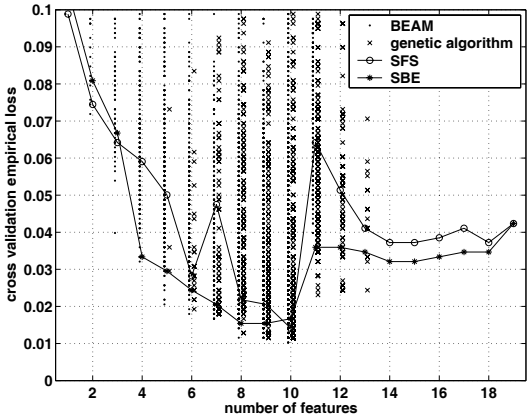


Fig. 2. Each point gives the generalization performance, as estimated by CV, for a particular subset of features and an optimized classifier. For a given subset, all classifier parameters such as aggregation function operator and its window width, degree of polynomial, and threshold Θ, were optimized using a grid search.

Results are shown in Fig. 2. SBE works better than SFS on average, though their best results are similar ($\hat{r} = 0.015$ and 0.014). BEAM and GA offer minor improvements ($\hat{r} = 0.010$ and 0.012) only.

Surprisingly, the final optimized system recognizes individual object hypotheses with a low accuracy: $\hat{r} = 0.341$ with $L_{Non-Sputter,Sputter} = 10$ and $L_{Sputter,Non-Sputter} = 1$. The high performance obtained in the end is entirely due to the temporal aggregation of evidence from individual frames.

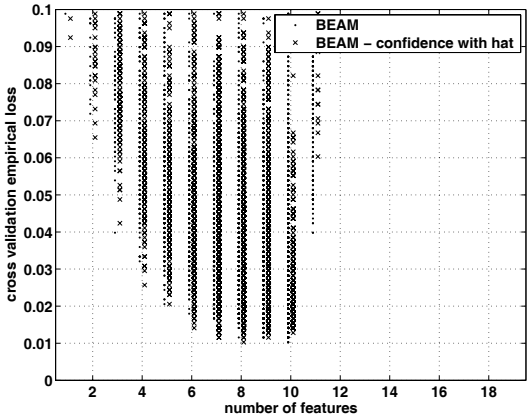


Fig. 3. Results obtained when replacing object estimates of class membership in individual images with the lower bound of interval estimates, see section 2.4.

Figure 3 shows the results obtained when the membership index $d(O_{t,i})$ is not given by the object estimate obtained from the PC, but by the lower

bound of an interval estimate to reflect the strongly asymmetric loss function. Overall classification accuracy is not improved, but the magnitude of the interval can help identifying sequences that ought to be labeled manually and should be included in future training sets.

5 Conclusion and outlook

Since the number of objects, N, is typically much larger than the number of basis functions, D, the most expensive part in training a PC is the computation of the correlation matrix and not its inversion. Recomputations of the former can be avoided in the framework of cross-validation, as illustrated in section 4 For our particular data set, advanced subset selection strategies did not lead to a much improved performance.

Even though all features computed on object hypotheses were chosen with the aim of describing the phenomenon well, the generalization performance varies greatly with the particular subset that is chosen in a specific classifier. A systematic search for the optimal subset is thus well worth while, and is made possible by the low computational cost of the PC which allows for a systematic joint optimization of parameters and feature subset.

References

AHA, D.W. and BANKERT, R.L. (1995): A comparative evaluation of sequential feature selection algorithms. In: D. Fischer and H. Lenz (Eds.): *Fifth International Workshop on Artificial Intelligence and Statistics*. 1–7.

BREIMAN, L. and SPECTOR, P. (1992): Submodel selection and evaluation in regression: The X-random case. *International Statistical Review, 60, 291–319*.

BROCKE, M. (2002): Statistical Image Sequence Processing for Temporal Change Detection. In: L. Van Gool (Ed.): *DAGM 2002, Pattern Recognition*. Springer, Zurich, 215–223.

HADER, S. (2003): System Concept for Image Sequence Classification in Laser Welding. In: B. Michaelis and G. Krell (Eds.): *DAGM 2003, Pattern Recognition*. Springer, Magdeburg, 212–219.

KOHAVI, R. and JOHN, G.H. (1997): Wrappers for Feature Subset Selection. *Artificial Intelligence, 97(1-2), 273–324*.

LECUN, Y., Jackel, J.D., Bottou, L., Brunot, A., Cortes, C., Denker, J.S., Drucker, H., Guyon, I., Muller, U.A., Sackinger, E., Simard, P. and Vapnik, V.(1995): Comparison of learning algorithms for handwritten digit recognition. In: F. Fogelman and P. Gallinari (Eds.): *International Conference on Artificial Neural Networks*. 53–60.

SCHÜRMANN, J. (1996): *Pattern Classification*. John Wiley and Sons, Inc., New York.

SEEBER, G. and Lee, A. (2003): *Linear Regression Analysis*. Wiley-Interscience.

TEAGUE, M.R. (1980): Image Analysis via the General Theory of Moments. *Opt. Soc. of America, 70, 920–930*.

Bagging, Boosting and Ordinal Classification

Klaus Hechenbichler and Gerhard Tutz

Institut für Statistik,
Ludwig-Maximilians-Universität München,
80539 München, Germany

Abstract. Since the introduction of bagging and boosting many new techniques have been developed within the field of classification via aggregation methods. Most of them have in common that the class indicator is treated as a nominal response without any structure. Since in many practical situations the class must be considered as an ordered categorical variable, it seems worthwhile to take this additional information into account. We propose several variants of bagging and boosting, which make use of the ordinal structure and it is shown how the predictive power might be improved. Comparisons are based not only on misclassification rates but also on general distance measures, which reflect the difference between true and predicted class.

1 Introduction

In statistical classification covariates are used to predict the value of an unobserved class variable. Various methods have been proposed and are nicely summarized e.g. in Hastie et al. (2001).

In recent years especially the introduction of aggregation methods like bagging (Breiman (1996)) and boosting (Freund (1995), Freund and Schapire (1996)) led to spectacular improvements of standard techniques. In all these methods a basic discrimination rule is used not only once but in different (weighted or unweighted) bootstrap versions of the data set.

A special problem is how to treat categorical ordered response variables. This additional information should be used to improve the accuracy of a classification technique. The purpose of our work is to combine aggregation methods with ordered class problems. Therefore aggregated classifiers for ordered response categories are developed and compared considering empirical data sets.

2 Aggregating classifiers

In a classification problem each object is assumed to come from one out of k classes. Let $L = \{(y_i, x_i), i = 1, \ldots, n_L\}$ denote the learning or training set of observed data, where $y_i \in \{1, \ldots, k\}$ denotes the class and $x_i' = (x_{i1}, \ldots, x_{ip})$ are associated covariates. Based on these p characteristics a classifier of the

form

$$C(.,L) : X \longrightarrow \{1,\ldots,k\}$$
$$x \longrightarrow C(x,L)$$

is built, where $C(x,L)$ is the predicted class for observation x. In the following three variants of aggregated classifiers, that are used as building blocks later, are shortly sketched.

Bagging (bootstrap aggregating) uses perturbed versions L_m of the learning set and aggregates the corresponding predictors by plurality voting, where the winning class is the one being predicted by the largest number of predictors

$$\text{argmax}_j \left(\sum_{m=1}^{M} I(C(x,L_m)=j) \right) \quad .$$

The perturbed learning sets of size n_L are formed by drawing at random from the learning set L. The predictor $C(.,L_m)$ is built from the m-th bootstrap sample.

In *boosting* the data are resampled adaptively and the predictors are aggregated by weighted voting. The *Discrete AdaBoost* procedure starts with weights $w_1 = \cdots = w_{n_L} = 1/n_L$ which form the resampling probabilities. Based on these probabilities the learning set L_m is sampled from L with replacement and the classifier $C(.,L_m)$ is built. The learning set is run through this classifier yielding error indicators $\epsilon_i = 1$ if the i-th observation is classified incorrectly and $\epsilon_i = 0$ otherwise. With

$$e_m = \sum_{i=1}^{n_L} w_i \epsilon_i \quad \text{and} \quad c_m = \log \frac{1-e_m}{e_m}$$

the resampling weights are updated for the next step by

$$w_{i,new} = \frac{w_i \exp(c_m \epsilon_i)}{\sum_{j=1}^{n_L} w_j \exp(c_m \epsilon_j)} \quad .$$

After M steps the aggregated voting for an observation is obtained by

$$\text{argmax}_j \left(\sum_{m=1}^{M} c_m I(C(x,L_m)=j) \right) \quad .$$

Real AdaBoost (Friedman et al. (2000)) uses real valued classifier functions $f(x,L)$ instead of $C(x,L)$. Since the original algorithm only works for two class problems, we present a variant that can be used for more than two classes in the following: Again it starts with the weights $w_1 = \cdots = w_{n_L} = 1/n_L$ which form the resampling probabilities. The learning set is run through a classifier that yields class probabilities

$$p_j(x_i) = \hat{P}(y_i = j | x_i) \quad .$$

Based on these probabilities real valued scores $f_j(x_i, L_m)$ are built by

$$f_j(x_i, L_m) = 0.5 \cdot \log \frac{p_j(x_i)}{(\prod_{l \neq j}^{k} p_l(x_i))^{\frac{1}{k-1}}}$$

and the weights are updated for the next step by

$$w_{i,new} = \frac{w_i \exp(-f_{y_i}(x_i, L_m))}{\sum_{j=1}^{n_L} w_j \exp(-f_{y_j}(x_j, L_m))} \quad .$$

After M steps the aggregated voting for observation x is obtained by

$$\text{argmax}_j \left(\sum_{m=1}^{M} f_j(x, L_m) \right) \quad .$$

Both AdaBoost algorithms are based on weighted resampling. In alternative versions of boosting observations are not resampled but the classifiers are computed by weighting the original observations by weights $w_1, \ldots, w_{n_L}$ that are updated iteratively. Then $C(., L_m)$ should be read as the classifier based on the current weights in the m-th step.

3 Ordinal prediction

In the following it is assumed that the classes in $y \in \{1, \ldots, k\}$ are ordered. In *Fixed split boosting* the classification procedure is divided into two stages: First aggregation is done by splitting the response categories. Then the resulting binary classifiers are combined. It works by defining

$$y^{(r)} = \begin{cases} 1\,, & y \in \{1, \ldots, r\} \\ 2\,, & y \in \{r+1, \ldots, k\} \end{cases}$$

for $r = 1, \ldots, k-1$.

Let $C^{(r)}(., L)$ denote the classifier for the binary class problem defined by $y^{(r)}$. For fixed r, by using any form of aggregate classifier one obtains the predicted class for observation x by computing

$$C_{agg}^{(r)}(x) = \text{argmax}_j \left(\sum_{m=1}^{M} c_m^{(r)} I(C^{(r)}(x, L_m^{(r)}) = j) \right) \quad .$$

These first stage aggregate classifiers $C_{agg}^{(r)}(.)$ have been designed for fixed split at r. The combination of $C_{agg}^{(1)}(.), \ldots, C_{agg}^{(k-1)}(.)$ is based on the second stage aggregation, now by exploiting the ordering of the categories. Thereby let the result of the classifier be transformed into the sequence $\hat{y}_1^{(r)}, \ldots, \hat{y}_k^{(r)}$ of binary variables.

For $C^{(r)}(x) = 1$ corresponding to $\hat{y}(x) \in \{1, \ldots, r\}$ one has

$$\hat{y}_1^{(r)}(x) = \cdots = \hat{y}_r^{(r)}(x) = \frac{1}{r}, \quad \hat{y}_{r+1}^{(r)}(x) = \cdots = \hat{y}_k^{(r)}(x) = 0 \quad .$$

For $C^{(r)}(x) = 2$ corresponding to $\hat{y}(x) \in \{r+1, \ldots, k\}$ one has

$$\hat{y}_1^{(r)}(x) = \cdots = \hat{y}_r^{(r)}(x) = 0, \quad \hat{y}_{r+1}^{(r)}(x) = \cdots = \hat{y}_k^{(r)}(x) = \frac{1}{k-r} \quad .$$

Thus the classifier $C_{agg}^{(r)}(.)$ yields the binary sequence

$$\frac{1}{r} \cdot (1, 1, \ldots, 1, 0, 0, \ldots, 0) \quad or \quad \frac{1}{k-r} \cdot (0, 0, \ldots, 0, 1, 1, \ldots, 1)$$

where the change from 1 to 0 or 0 to 1 is after the r-th component. We divide these sequences by r or $k-r$ respectively to take into account the different number of categories within each dichotomization. The final classifier is given by the second stage aggregation

$$C_{agg}(x) = \text{argmax}_j \left(\sum_{r=1}^{k-1} \hat{y}_j^{(r)} \right) \quad .$$

In Fixed split boosting the ordinal structure of the response is not used in the reweighting scheme. Only in the final combining step it is exploited that the response is ordinal. Therefore in the following an alternative algorithm (called *Ordinal Discrete AdaBoost*) is suggested which connects the weights in Discrete AdaBoost to the ordered performance of the classifier.

Again we start with weights $w_1, \ldots, w_{n_L}$ which form the resampling probabilities. Based on these probabilities the learning set L_m is sampled from L with replacement. Based on L_m the classifiers $C^{(r)}(., L_m)$ are built for all dichotomous splits of the ordinal class variable at value r. The learning set is run through each classifier $C^{(r)}(., L_m)$ yielding the information if the i-th observation is predicted into a class higher or lower than r. The results of the classifiers for different split values r are combined by majority vote into the aggregated classifier $C(., L_m)$.

Let the error indicators now be given by

$$\epsilon_i = \frac{|C(x_i, L_m) - y_i|}{k-1} \quad .$$

Therefore with $e_m = \sum_{i=1}^{n_L} w_i \epsilon_i$ and $c_m = \log((1-e_m)/e_m)$ the weights are updated by

$$w_{i,new} = \frac{w_i \exp(c_m \epsilon_i)}{\sum_{j=1}^{n_L} w_j \exp(c_m \epsilon_j)} \quad .$$

After M steps the aggregated voting for observation x is obtained by

$$\text{argmax}_j \left(\sum_{m=1}^{M} c_m I(C(x, L_m) = j) \right) \quad .$$

In a similar way an ordinal version of Real AdaBoost (called *Ordinal Real AdaBoost*) can be developed: For each dichotomization probabilities $p^{(r)}(x) = \hat{P}(y \leq r|x)$ are provided by a dichotomous classifier. From these probabilities one obtains a sequence of scores $\hat{y}^{(r)}(x)$ for each class by

$$\hat{y}_1^{(r)}(x) = \cdots = \hat{y}_r^{(r)}(x) = p^{(r)}(x), \quad \hat{y}_{r+1}^{(r)}(x) = \cdots = \hat{y}_k^{(r)}(x) = 1 - p^{(r)}(x) \quad .$$

For the aggregation across splits one considers the value

$$\hat{y}_j(x) = \sum_{r=1}^{k-1} \hat{y}_j^{(r)}(x)$$

which reflects the strength of prediction in class j. Then an ordinal algorithm, that still shows a close relationship to Real AdaBoost, works as follows: Based on weights $w_1, \ldots, w_{n_L}$ a classifier $\hat{y}_j(.)$ for ordered classes $1, \ldots, k$ is built and the learning set is run through it yielding scores $\hat{y}_j(x_i)$, $j = 1, \ldots, k$. Based on these scores real valued terms $f_j(x_i, L_m)$ are built by

$$f_j(x_i, L_m) = 0.5 \cdot \log \frac{\hat{y}_j(x_i)}{\frac{1}{\sum_{l \neq j}^{k} d_{il}} \sum_{l \neq j}^{k} d_{il} \hat{y}_l(x_i)}$$

where d_{ij} is the distance between true class and class j for observation i. So class probabilities are weighted according to the distance between current and true class. The weights are updated for the next step by

$$w_{i,new} = \frac{w_i \exp(-f_{y_i}(x_i, L_m))}{\sum_{j=1}^{n_L} w_j \exp(-f_{y_j}(x_j, L_m))} \quad .$$

After M steps the aggregated voting for observation x is obtained by

$$\text{argmax}_j \left(\sum_{m=1}^{M} f_j(x, L_m) \right) \quad .$$

In Friedman et al. (2000) a variant for Real AdaBoost, called Gentle AdaBoost, is suggested, which uses a different update function and seems to work more stable. Without a detailed description of this algorithm, the results of the ordinal adaption just in the same way as for Real AdaBoost are presented in the empirical part.

Finally a boosting variant is presented, that originally was developed to predict binary classes or real valued variables, but is easily transformed to cope with ordinal classes: L_2-*Boost* (Bühlmann and Yu (2002)), a special case of the more general gradient descent boosting algorithm, works without any kind of weighting. In the first step a real-valued initial learner $\hat{F}_0(x) = \hat{f}(x)$ is computed by means of least squares $\min \sum_{i=1}^{n_L} (y_i - \hat{f}(x_i))^2$. Then the iteration starts with $m = 0$. The negative gradient vector

$$u_i = y_i - \hat{F}_m(x_i)$$

is computed and the real-valued learner $\hat{f}_{m+1}(x)$ is now fit to these current residuals, again by means of least squares $\min \sum_{i=1}^{n_L}(u_i - \hat{f}_{m+1}(x_i))^2$. Finally the prediction $\hat{F}_{m+1}(x)$ is updated by

$$\hat{F}_{m+1}(x) = \hat{F}_m(x) + \hat{f}_{m+1}(x)$$

and the iteration index m is increased by one.

As the mean squared error is a sensitive indicator for ordinal distances, this algorithm can be used for ordinal classification with only one little adjustment: The values of $\hat{F}_{m+1}(x)$ are rounded to the nearest class label in order to follow the allowed domain of y.

4 Empirical studies

The scapula data are part of a dissertation (Feistl and Penning (2001)) written at the *Institut für Rechtsmedizin der LMU München*. The aim was to predict the age of dead bodies only by means of the scapula. Therefore a lot of measures, implying angles, lengths, descriptions of the surface, etc. were provided. We preselected 15 important covariates to predict age, which was splitted into 8 distinct ordinal classes. Each class covers ten years. The data set consists of 153 complete observations.

In the following we compare the different ordinal approaches to simpler alternative methods, that either do not use the ordinal information within the data or do not use aggregation techniques. The first one is a simple classification tree, called nominal CART. Here a tree is built by means of the deviance criterion and grown up to maximal depth. Afterwards it is pruned on the basis of resubstitution misclassification rates until a fixed tree size (given by the user and dependent on the data set) is reached. An alternative approach which uses the ordering of the classes, but still without bagging or boosting, is to build a tree for every dichotomization in r separately and aggregate them according to $\text{argmax}_j \sum_{r=1}^{k-1} I(C^{(r)}(x, L) = j)$. This method is called ordinal CART. In the same way two bagging variants are considered: A nominal approach, where every tree predicts the multi class target variable and the final result is obtained by a majority vote of these predictions, and ordinal bagging, in which bagging is applied to fixed splits. The results are aggregated over dichotomizations according to $\text{argmax}_j \sum_{r=1}^{k-1} I(C^{(r)}(x, L) = j)$ and over the bagging cycles. The nominal methods are considered as baseline for possible improvements by ordinal bagging or ordinal boosting.

In addition we consider ten different boosting versions: The simple nominal Discrete, Real and Gentle AdaBoost, which do not use the ordinal structure within the data, are used for comparison only. The new methods are the ordinal boosting techniques: On the one hand we distinguish between real, discrete and gentle methods, on the other hand between Fixed split boosting and Ordinal AdaBoost. Finally L_2-Boost results are shown.

The evaluation of the methods is based on several measures of accuracy. As a raw criterion for the accuracy of the prediction we take the misclassification error rate $\frac{1}{n}\sum_{i=1}^{n} 1_{\{y_i \neq \hat{y}_i\}}$. But in the case of ordinal class structure measures should take into account that a larger distance is a more severe error than a wrong classification into a neighbour class. Therefore we use the mean absolute value (here called *mean abs*) of the differences $\frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$ and the mean squared difference (here called *mean squ*) $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ which penalizes larger differences even harder.

When measuring accuracy one has to distinguish between resubstitution error and validation (or test) error. Resubstitution error is used to examine how fast the misclassification error can be lowered by the different techniques, but because of its bias it is not an appropriate measure for prediction accuracy. Therefore we divide the data set at random into two parts consisting of one respectively two thirds of the observations. From the larger (learning) data set the classification model is built and the observations of the smaller (test) data set are predicted. We use 50 different random splits into learning and test set and give the mean over these splits. As testing for differences between the performance of various techniques is quite difficult because of the statistical dependence between the different test sets, we omit it in the framework of this study.

When aggregating classifiers one has to choose the number of cycles, which means the number of different classifiers that are combined in one bagging or boosting run. As standard we use a number of 50 cycles in this study. The last parameter that has to be chosen is the (fixed) tree size, that is defined as the number of terminal nodes of each tree. Here we use a number of 15 terminal nodes in the nominal approach, which seems necessary for a problem with 8 classes and after all 15 covariates. All ordinal approaches are performed with trees of size 5, because as far as trees are concerned only two class problems are treated.

The interpretation of Table 1 leads to the following conclusions: As far as the misclassification error is concerned there are only slight differences between the classifiers. However, the more important measures for problems with ordinal classes are the distance measures. Here the results of CART are improved by all aggregation methods. Especially Fixed split boosting, but also the other ordinal boosting methods and ordinal bagging perform very well. For example the mean squared distance 2.365 of the classification tree is reduced to 1.215 by Discrete Fixed split boosting.

5 Concluding remarks

The concept to combine aggregating classifiers with techniques for ordinal data structure led to new methods that can be compared with common classification techniques. In further studies (Tutz and Hechenbichler (2003)) we

Table 1. Test error for scapula data

method	misclass	mean abs	mean squ
nominal CART	0.676	1.085	2.365
ordinal CART	0.652	0.995	2.112
nominal bagging	0.663	0.982	1.925
ordinal bagging	0.628	0.828	1.375
Discrete AdaBoost	0.649	0.932	1.747
Real AdaBoost	0.646	0.904	1.619
Gentle AdaBoost	0.643	0.876	1.502
Discrete Fixed split boosting	0.629	0.799	1.215
Real Fixed split boosting	0.646	0.818	1.244
Gentle Fixed split boosting	0.638	0.808	1.216
Ordinal Discrete AdaBoost	0.611	0.841	1.473
Ordinal Real AdaBoost	0.691	0.878	1.330
Ordinal Gentle AdaBoost	0.644	0.806	1.210
L_2-Boost	0.652	0.851	1.316

found promising results for other empirical data sets as well. Ordinal techniques definitely improve the performance of a simple CART tree.

All in all there seems to be no dominating method as for different data sets the best results occur by different methods. Although ordinal bagging shows satisfying results for all data sets, it often is outperformed by at least one of the ordinal boosting techniques. These findings suggest that further research seems to be a worthwhile task.

References

BREIMAN, L. (1996): Bagging Predictors. *Machine Learning, 24, 123–140.*

BREIMAN, L. (1998): Arcing Classifiers. *Annals of Statistics, 26, 801–849.*

BÜHLMANN, P. and YU, B. (2002): Boosting with the L_2-Loss: Regression and Classification. To appear in *Journal of the American Statistical Association.*

FREUND, Y (1995): Boosting a Weak Learning Algorithm by Majority. *Information And Computation, 121, 256–285.*

FREUND, Y. and SCHAPIRE, R.E. (1996): Experiments with a New Boosting Algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference.*

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000): Additive Logistic Regression: A Statistical View of Boosting. *Annals of Statistics, 28, 337–374.*

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): The Elements of Statistical Learning. New York, Spinger.

SCHAPIRE, R.E. (2002): The Boosting Approach to Machine Learning: An Overview. *MSRI Workshop on Nonlinear Estimation and Classification.*

TUTZ, G. and HECHENBICHLER, K. (2003): Aggregating Classifiers with Ordinal Response Structure. *Discussion Paper 359, SFB 386 der Ludwig-Maximilians-Universität München.*

A Method for Visual Cluster Validation

Christian Hennig

Fachbereich Mathematik - SPST,
Universität Hamburg, 20146 Hamburg, Germany

Abstract. Cluster validation is necessary because the clusters resulting from cluster analysis algorithms are not in general meaningful patterns. I propose a methodology to explore two aspects of a cluster found by any cluster analysis method: the cluster should be separated from the rest of the data, and the points of the cluster should not split up into further separated subclasses. Both aspects can be visually assessed by linear projections of the data onto the two-dimensional Euclidean space. Optimal separation of the cluster in such a projection can be attained by asymmetric weighted coordinates (Hennig (2002)). Heterogeneity can be explored by the use of projection pursuit indexes as defined in Cook, Buja and Cabrera (1993). The projection methods can be combined with splitting up the data set into clustering data and validation data. A data example is given.

1 Introduction

Cluster validation is the assessment of the quality and the meaningfulness of the outcome of a cluster analysis (CA). Most CA methods generate a clustering in all data sets, whether there is a meaningful structure or not. Furthermore, most CA methods partition the data set into subsets of a more or less similar shape, and this may be adequate only for parts of the data, but not for all. Often, different CA methods generate different clusterings on the same data and it has to be decided which one is the best, if any. Therefore, if an interpretation of a cluster as a meaningful pattern is desired, the cluster should be validated by information other than the output of the CA. A lot of more or less formal methods for cluster validation are proposed in the literature, many of which are discussed, e.g., in Gordon (1999, Section 7.2) and Halkidi et al. (2002). Six basic principles for cluster validation can be distinguished:

Use of external information External information is information that has not been used to generate the clustering. Such information can stem from additional data or from background knowledge. However, such information is often not available.

Significance tests for structure Significance tests against null models formalizing "no clustering structure at all" are often used to justify the interpretation of a clustering. While the rejection of homogeneity is a reasonable minimum requirement for a clustering, such tests cannot validate the concrete structure found by the CA algorithm.

Comparison of different clusterings on the same data Often, the agreement of clusterings based on different methods is taken as a confirmation of clusters. This is only meaningful if sufficiently different CA methods have been chosen, and in the case of disagreement it could be argued that not all of them are adequate for the data at hand.

Validation indexes In some sense, the use of validation indexes is similar to that of different clusterings, because many CA methods optimize indexes that could otherwise be used for validation.

Stability assessment The stability of clusters can be assessed by techniques such as bootstrap, cross-validation, point deletion, and addition of contamination.

Visual inspection Recently (see, e.g., Ng and Huang (2002)), it has been recognized that all formal approaches of cluster validation have limitations due to the complexity of the CA problem and the intuitive nature of what is called a "cluster". Such a task calls for a more subjective and visual approach. To my knowledge, the approach of Ng and Huang (2002) is the first visual technique which is specifically developed for the validation of a clustering.

Note that these principles address different aspects of the validation problem. A clustering that is well interpretable in the light of external information will not necessarily be reproduced by a different clustering method. Structural aspects such as homogeneity of the single clusters and heterogeneity between different clusters as indicated by validation indexes or visual inspection are not necessarily properties of clusters which are stable under resampling. However, these aspects are not "orthogonal". A well chosen clustering method should tend to reproduce well separated homogeneous clusters even if the data set is modified.

In the present paper, a new method for visual cluster validation is proposed. As opposed to the approach of Ng and Huang (2002), the aim of the present method is to assess every cluster individually. The underlying idea is that a valid cluster should have two properties:

- separation from the rest of the data, so that it should not be joined with other parts of the data,
- homogeneity, so that the points of the cluster can be said to "belong together".

In Section 2, asymmetric weighted coordinates (AWCs) are introduced. AWCs provide a linear projection of the data in order to separate the cluster under study optimally from the rest of the data. In Section 3, I propose the application of some projection pursuit indexes to the points of the cluster to explore its heterogeneity. Additionally, if there is enough data to split the data set into a "training sample" and a "validation sample", the projections obtained from clustering and visual validation on the training sample can be applied also to the points of the validation sample to see if the found patterns can be reproduced. Throughout the paper, the data is assumed to come

from the p-dimensional Euclidean space. The methods can also be applied to distance data after carrying out an appropriate multidimensional scaling method. Euclidean data is only needed for the validation; the clustering can be done on the original distances. In Section 4, the method is applied to a real data set.

2 Optimal projection for separation

The most widespread linear projection technique to separate classes goes back to Rao (1952) and is often called "discriminant coordinates" (DCs). The first DC is defined by maximizing the ratio

$$F(\mathbf{c}_1) = \frac{\mathbf{c}_1'\mathbf{B}\mathbf{c}_1}{\mathbf{c}_1'\mathbf{W}\mathbf{c}_1}, \text{ where}$$

$$\mathbf{W} = \frac{1}{n-s}\sum_{i=1}^{s}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\mathbf{m}_i)(\mathbf{x}_{ij}-\mathbf{m}_i)',$$

$$\mathbf{B} = \frac{1}{n(s-1)}\sum_{i=1}^{s} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})'.$$

n denotes the number of points, n_i is the number of points of class i, s denotes the number of classes, $\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i}$ are the p-dimensional points of class i, $\mathbf{m}_i$ is the mean vector of class i and $\mathbf{m}$ is the overall mean. The further DCs maximize F under the constraint of orthogonality to the previous DCs w.r.t. $\mathbf{W}$. $\mathbf{B}$ is a covariance matrix for the class means and $\mathbf{W}$ is a pooled within-class covariance matrix. Thus, F gets large for projections that separate the means of the classes as far as possible from each other while keeping the projected within-class variation small. Some disadvantages limit the use of DCs for cluster validation. Firstly, separation is formalized only in terms of the class means, and points of different classes far from their class means need not to be well separated (note that the method of Ng and Huang (2002) also aims at separating the cluster centroids). Secondly, $s-1$ dimensions are needed to display all information about the separation of s classes, and therefore there is no guarantee that the best separation of a particular cluster shows up in the first two dimensions in case of $s > 3$. This could in principle be handled by declaring the particular cluster to be validated as class 1 and the union of all other clusters as class 2 (this will be called the "asymmetry principle" below). But thirdly, DCs assume that the covariance matrices of the classes are equal, because otherwise $\mathbf{W}$ would not be an adequate covariance matrix estimator for a single class. If the asymmetry principle is applied to a clustering with $s > 2$, the covariance matrices of these classes cannot be expected to be equal, not even if they would be equal for the s single clusters.

A better linear projection technique for cluster validation is the application of asymmetric linear dimension reduction (Hennig (2002)) to the two

classes obtained by the asymmetry principle. Asymmetry means that the two classes to be projected are not treated equally. Asymmetric discriminant coordinates maximize the separation between class 1 and class 2 while keeping the projected variation of class 1 small. Class 2, i.e., the union of all other data points, may appear as heterogeneously as necessary. Four asymmetric projection methods are proposed in Hennig (2002), of which asymmetric weighted coordinates (AWCs) are the most suitable for cluster validation. The first AWC is defined by maximizing

$$F^*(\mathbf{c}_1) = \frac{\mathbf{c}_1'\mathbf{B}^*\mathbf{c}_1}{\mathbf{c}_1'\mathbf{S}_1\mathbf{c}_1}, \text{ where}$$

$$\mathbf{S}_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(\mathbf{x}_{1j} - \mathbf{m}_1),$$

$$\mathbf{B}^* = \sum_{i,j} w_j(\mathbf{x}_{1i} - \mathbf{x}_{2j})(\mathbf{x}_{1i} - \mathbf{x}_{2j})',$$

$$w_j = \min\left(1, \frac{d}{(\mathbf{x}_{2j} - \mathbf{m}_1)'\mathbf{S}_1^{-1}(\mathbf{x}_{2j} - \mathbf{m}_1)}\right),$$

$d > 0$ being some constant, for example the 0.99-quantile of the χ^2_p-distribution. The second AWC $\mathbf{c}_2$ maximizes F^* subject to $\mathbf{c}_1'\mathbf{S}_1^{-1}\mathbf{c}_2 = 0$ and so on. $\mathbf{c}_1'\mathbf{B}^*\mathbf{c}_1$ gets large if the projected differences between points of class 1 and class 2 are large. The weights w_j downweight differences from points of class 2 that are very far away (in Mahalanobis distance) from class 1. Otherwise, $\mathbf{c}_1'\mathbf{B}^*\mathbf{c}_1$ would be governed mainly by such points, and class 1 would appear separated mainly from the furthest points in class 2, while it might be mixed up more than necessary with closer points of class 2. The weights result in a projection that separates class 1 also from the closest points as well as possible. More motivation and background is given in Hennig (2002). As for DCs, the computation of AWCs is very easily be done by an Eigenvector decomposition of $\mathbf{S}_1^{-1}\mathbf{B}^*$. Note that AWCs can only be applied if $n_1 > p$, because otherwise class 1 could be projected onto a single point, thus $\mathbf{c}_1'\mathbf{S}_1^{-1}\mathbf{c}_1 = 0$. If n_1 is not much larger than p, $\mathbf{c}_1'\mathbf{S}_1^{-1}\mathbf{c}_1$ can be very small, and some experience (e.g., with simulated data sets from unstructured data) is necessary to judge if a seemingly strong separation is really meaningful.

3 Optimal projection for heterogeneity

Unfortunately, AWCs cannot be used to assess the homogeneity of a cluster. The reason is that along projection directions that do not carry any information regarding the cluster, the cluster usually does not look separated, but often more or less homogeneous. Thus, to assess separation, the projected separation has to be maximized, which is done by AWCs. But to assess homogeneity, it is advantageous to maximize the projected *heterogeneity* of the cluster.

Projection pursuit is the generic term for linear projection methods that aim for finding "interesting", i.e., heterogeneous projections of the data (Huber (1985)). The idea is to project only the points of the cluster to be validated in order to find a most heterogeneous visualization. There are lots of projection pursuit indexes. Some of them are implemented in the data visualization software XGOBI (Buja et al. (1996)). A main problem with projection pursuit is that the indexes can only be optimized locally. XGOBI visualizes the optimization process dynamically, and after a local optimum has been found, the data can be rotated toward new configurations to start another optimization run.

Two very simple and useful indexes have been introduced by Cook et al. (1993) and are implemented in XGOBI. The first one is the so-called "holes index", which is defined by minimizing

$$F^{**}(\mathbf{C}) = \sum_{i=1}^{n_1} \varphi_2(\mathbf{C}'\mathbf{x}_{1i}),$$

over orthogonal $p \times 2$-projection matrices $\mathbf{C}$, where φ_2 denotes the density of the two-dimensional Normal distribution and the points $\mathbf{x}_{1i}$ are assumed to be centered and scaled. F^{**} becomes minimal if as few points as possible are in the center of the projection, in other words, if there is a "hole". Often, such a projection shows a possible division of the cluster points into subgroups.

It is also useful to maximize F^{**}, which is called "central mass index" in XGOBI. This index attempts to project as many points as possible into the center, which can be used to find outliers in the cluster. But it can also be useful to try out further indexes, as discussed in Cook et al. (1993).

4 Example

As an example, two CA methods have been applied to the "quakes" data set, which is part of the base package of the free statistical software R (to obtain from `www.R-project.org`). The data consist of 1000 seismic events on Fiji, for which five variables have been recorded, namely geographical longitude and latitude, depth, Richter magnitude and number of stations reporting. Because of the favorable relation of n to p, I divided the data set into 500 points that have been used for clustering and 500 points for validation.

The first clustering has been generated by MCLUST (Fraley and Raftery (2003)), a software for the estimation of a Normal mixture model including noise, i.e., points that do not belong to any cluster. The Bayesian information criterion has been used to decide about the number of clusters and the complexity of the model for the cluster's covariance matrices. It resulted in four clusters with unrestricted covariance matrices plus noise. As a comparison, I have also performed a 5-means clustering on sphered data.

Generally, the validity of the clusters of the MCLUST-solution can be confirmed. In Figure 1, the AWC plot is shown for the second cluster (points

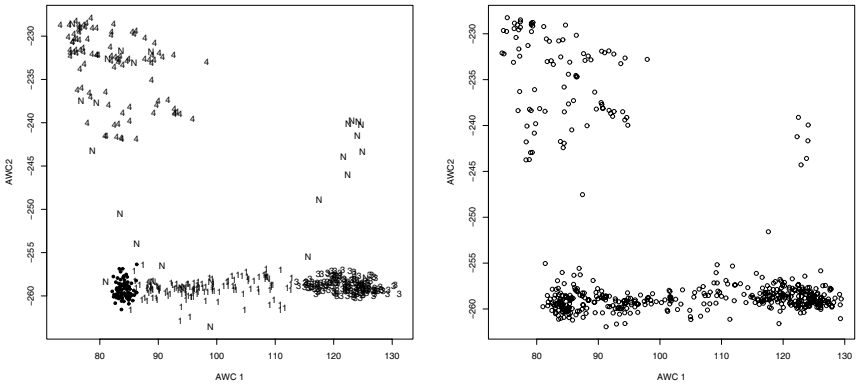


Fig. 1. Left: AWCs of cluster 2 (black points) of the MCLUST solution. Right: validation data set projected onto the AWCs.

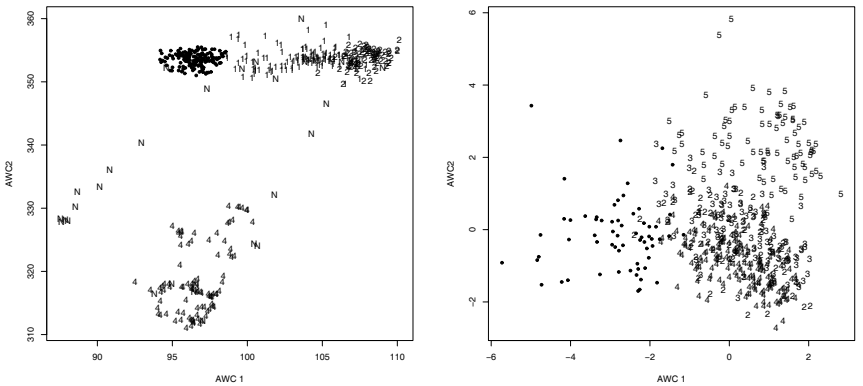


Fig. 2. Left: AWCs of cluster 3 (black points) of the MCLUST solution. Right: AWCs of cluster 1 (black points) of the 5-means solution.

of other clusters are always indicated with the cluster numbers). These points do neither appear separated in any scatterplot of two variables nor in the principal components (not shown), but they are fairly well separated in the AWC plot, and the projection of the validation points on the AWCs (right side) confirms that there is a meaningful pattern. Other clusters are even better separated, e.g., cluster 3 on the left side of Figure 2. Some of the clusters of the 5-means solution have a lower quality. For example, the AWC-plot of cluster 1 (right side of Figure 2) shows the separation as dominated by the variation within this cluster.

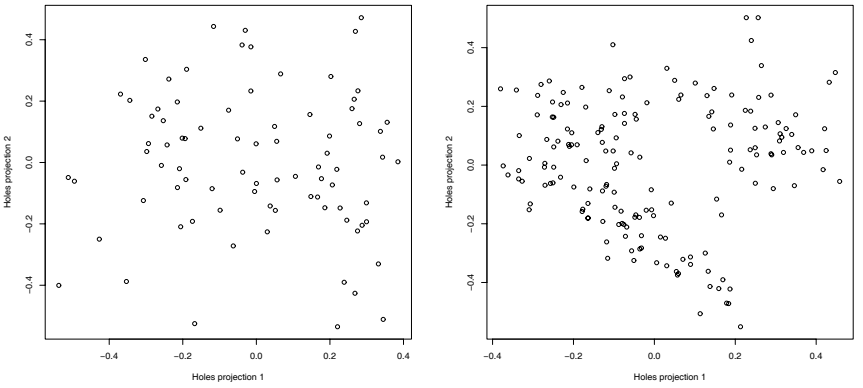


Fig. 3. Left: "holes" projection of cluster 2 of the MCLUST solution. Right: "holes" projection of cluster 3 of the MCLUST solution.

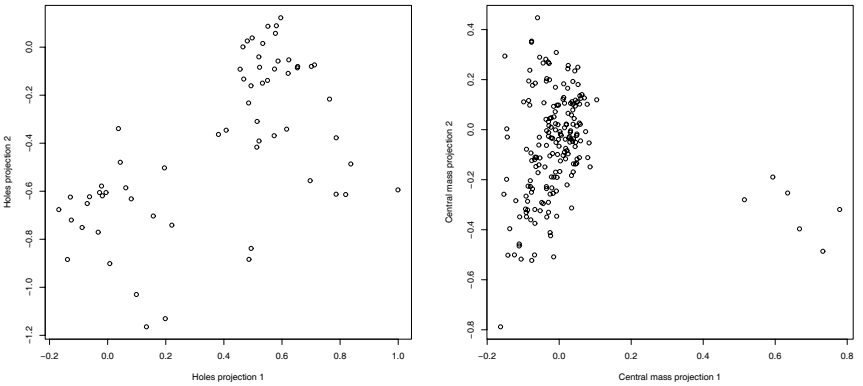


Fig. 4. Left: "holes" projection of cluster 1 of the 5-means solution. Right: "central mass" projection of cluster 4 of the 5-means solution.

Optimization of the holes index did not reveal any heterogeneity in MCLUST-cluster 2, see the left side of Figure 3, while in cluster 3 (right side) two subpopulations could roughly be recognized. Sometimes, when applying MCLUST to other 500-point subsamples of the data, the corresponding pattern is indeed divided into two clusters (it must be noted that there is a non-negligible variation in the resulting clustering structures from MCLUST, including the estimated number of clusters, on different subsamples). Some of the 5-means clusters show a much clearer heterogeneity. The holes index reveals some subclasses of cluster 1 (right side of Figure 4), while the central mass index highlights six outliers in cluster 4 (right side).

5 Conclusion

A combination of two plots for visual cluster validation of every single cluster has been proposed. AWCs optimize the separation of the cluster from the rest of the data while the cluster is kept homogeneous. Projection pursuit is suggested to explore the heterogeneity of a cluster.

Note that for large p compared to n, the variety of possible projections is large. Plots in which the cluster looks more or less separated or heterogeneous are found easily. Thus, it is advisable to compare the resulting plots with the corresponding plots from analogous cluster analyses applied to data with the same n and p generated from "null models" such as a normal or uniform distribution to assess if the cluster to be validated yields a stronger pattern. This may generally be useful to judge the validity of visual displays.

The proposed plots are static. This has the advantage that they are reproducible (there may be a non-uniqueness problem with projection pursuit) and they are optimal with respect to the discussed criteria. However, a further dynamical visual inspection of the data by, e.g., the grand tour as implemented in XGOBI (Buja et al. (1996)), can also be useful to assess the stability of separation and heterogeneity as revealed by the static plots.

AWCs are implemented in the add-on package FPC for the statistical software package R, available under `www.R-project.org`.

References

BUJA, A., COOK, D. and SWAYNE, D. (1996): Interactive High-Dimensional Data Visualization. *Journal of Computational and Graphical Statistics, 5, 78–99.*

COOK, D., BUJA, A. and CABRERA, J. (1993): Projection Pursuit Indexes Based on Orthonormal Function Expansions. *Journal of Computational and Graphical Statistics, 2, 225–250.*

FRALEY, C. and RAFTERY, A. E. (2003): Enhanced Model-Based Clustering, Density Estimation and Discriminant Analysis Software: MCLUST. *Journal of Classification, 20, 263–293.*

GORDON, A.D. (1999): *Classification*, 2nd Ed. Chapman & Hall/CRC, Boca Raton.

HALKIDI, M., BATISTAKIS, Y. and VAZIRGIANNIS, M. (2002): Cluster Validity Methods: Part I. *SIGMOD Record, 31, 40–45.*

HENNIG, C. (2002): Symmetric, asymmetric and robust linear dimension reduction for classification. To appear in *Journal of Computational and Graphical Statistics*, `ftp://ftp.stat.math.ethz.ch/Research-Reports/108.html`.

HUBER, P. J. (1985): Projection pursuit (with discussion). *Annals of Statistics, 13, 435–475.*

NG, M. and HUANG, J. (2002): M-FastMap: A Modified FastMap Algorithm for Visual Cluster Validation in Data Mining. In: M.-S. Chen, P. S. Yu and B. Liu (Eds.): *Advances in Knowledge Discovery and Data Mining. Proceedings of PAKDD 2002, Taipei, Taiwan.* Springer, Heidelberg, 224–236.

RAO, C. R. (1952): *Advanced Statistical Methods in Biometric Research*, Wiley, New York.

Empirical Comparison of Boosting Algorithms

Riadh Khanchel and Mohamed Limam

Institut superieur de gestion
41, Rue da la liberte, Le Bardo 2000 Tunis, Tunisia

Abstract. Boosting algorithms combine moderately accurate classifiers in order to produce highly accurate ones. The most important boosting algorithms are Adaboost and Arc-x(j). While belonging to the same algorithms family, they differ in the way of combining classifiers. Adaboost uses weighted majority vote while Arc-x(j) combines them through simple majority vote. Breiman (1998) obtains the best results for Arc-x(j) with $j = 4$ but higher values were not tested. Two other values for j, $j = 8$ and $j = 12$ are tested and compared to the previous one and to Adaboost. Based on several real binary databases, empirical comparison shows that Arc-x4 outperforms all other algorithms.

1 Introduction

Boosting algorithms are one of the most recent developments in classification methodology. They repeatedly apply a classification algorithm as a subroutine and combine moderately accurate classifiers in order to produce highly accurate ones. The first boosting algorithm, developed by Schapire(1990), converts a weak learning algorithm into a strong one. A strong learning algorithm achieves low error with high confidence while a weak learning algorithm drops the requirement of high accuracy. Freund (1995) presents another boosting algorithm, boost-by-majority, which outperforms the previous one.

Freund and Schapire (1997) present another boosting algorithm, Adaboost. It is the first adaptive boosting algorithm because its strategy depends on the advantages of obtained classifiers, called hypotheses. For binary classification, the advantage of a hypothesis measures the difference between its performance and random guessing. The only requirement of Adaboost is to obtain hypotheses with positive advantage. Furthermore, the final hypothesis is a weighted majority vote of the generated hypotheses where the weight of each hypothesis depends on its performance. Due to its adaptive characteristic, Adaboost has received more attention than its predecessors. Experimental results (Freund and Schapire (1996), Bauer and Kohavi (1999)) show that Adaboost decreases the error of the final hypothesis.

Breiman (1998) introduces the ARCING algorithm's family: **A**daptively **R**esampling and **C**ombining which Adaboost belongs to. In order to better understand the behavior of Adaboost, Breiman (1998) develops a simpler boosting algorithm denoted by Arc-x(j). This algorithm uses a different

weight updating rule and combines hypotheses using simple majority vote. The best results of Arc-x(j) are obtained for $j = 4$. When compared to Adaboost, Breiman's results show that both algorithms perform equally well. Breiman (1998) argues that the success of Adaboost is not due to its way of combining hypotheses but on its adaptive property. He argues also that since higher values for j were not tested further improvement is possible.

In this paper, two other values for the parameter j of Arc-x(j) algorithm, $j = 8$ and $j = 12$, are tested and their performance compared to Adaboost and Arc-x4 in the subsampling framework using a one node decision tree algorithm.

In section two, the different boosting algorithms used are briefly introduced. In section three, the empirical study is described and the results are presented. Finally, section four provides a conclusion to this article.

2 Arcing algorithms

Adaboost was the first adaptive boosting algorithm. First, the general framework of boosting algorithms is introduced, then Adaboost and some of its characteristics are reviewed. Finally, arcing algorithm's family is discussed.

Given a labeled training set $(x_1, y_1), \ldots, (x_n, y_n)$, where each x_i belongs to the instance space X, and each label y_i to the label set Y. Here only the binary case is considered where $Y = \{-1, 1\}$. Adaboost applies repeatedly, in a series of iterations $t = 1, \ldots, T$, the given learning algorithm to a reweighted training set. It maintains a weight distribution over the training set. Starting with equal weight assigned to all instances, $D(x_i) = 1/n$, weights are updated after each iteration such that the weight of misclassified instances is increased. Weights represent instance importance. Increasing instance's weight will give it more importance and thus forcing the learning algorithm to focus on it in the next iteration. The learning algorithm outputs in each iteration a hypothesis that predicts the label of each instances $h_t(x_i)$. For a given iteration, the learning algorithm tends to minimize the error:

$$\epsilon_t = Pr[h_t(x_i) \neq y_i], \tag{1}$$

where $Pr[.]$ denote empirical probability on the training sample.

2.1 Adaboost

Adaboost requires that the learning algorithm outputs hypotheses with error less than 0.5. A parameter α_t is used to measure the importance assigned to each hypothesis. This parameter depends on hypothesis' performance. For the binary case this parameter is set to:

$$\alpha_t = \frac{1}{2} \ln(\frac{1 - \epsilon_t}{\epsilon_t}). \tag{2}$$

The weight distribution is updated using α_t (see Figure 2.1). This parameter is positive because Adaboost requires that the learning algorithm output hypotheses with error less than 0.5. At the end of the process, a final hypothesis is obtained by combining all hypotheses from previous iterations using weighted majority vote. The parameter α_t represents the weight of the hypothesis h_t generated in iteration t. The pseudocode of Adaboost for binary classification is presented in Figure 2.1.

Adaboost requires that the base learner performs better than random guessing. The error can be written as follows:

$$\epsilon_t = 1/2 - \gamma_t, \tag{3}$$

where γ_t is a positive parameter that represents the advantage of the hypothesis over random guessing. The training error of the final hypothesis is bounded by:

$$\prod_t 2\sqrt{\epsilon_t(1-\epsilon_t)} \tag{4}$$

Given: $(x_1, y_1), \ldots, (x_n, y_n)$ where $x_i \in X, y_i \in Y = \{-1; +1\}$
1-Initialize $D_1(i) = 1/n$
2-For t = 1 to T:

- Train the weak classifier using D_t and get a hypothesis $h_t : X \mapsto \{-1; +1\}$
- Compute $\epsilon_t = \sum_{i:h_t(x_i)\neq y_i} D_t(x_i)$
- If $\epsilon_t \geq 0.5$ stop.
- Choose: $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$
- Update: $D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$

  where Z_t is a normalization factor

3-output the final hypothesis: $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$

Figure 2.1: Adaboost algorithm

This bound can be expressed in term of the advantage sequence γ_t:

$$\prod_t 2\sqrt{\epsilon_t(1-\epsilon_t)} \leq \exp(-2\sum_t \gamma_t^2). \tag{5}$$

Thus, if each hypothesis is slightly better than random guessing, that is $\gamma_t > \gamma$ for $\gamma > 0$, the training error will drop exponentially fast.

The bound of generalization error, or the error of the final hypothesis over the whole instance space X, depends on the training error, the size of the sample n, the Vapnik-Chervonenkis (VC, Vapnik 1998) dimension d of the weak hypothesis space and the number of boosting iterations T. The generalization error is at most:

$$Pr[H(x) \neq y] + \hat{O}\left(\sqrt{\frac{T.d}{n}}\right). \tag{6}$$

This bound depends on the number of iterations T and we would think that it will overfit as T becomes large but experimental results (Freund and Schapire (1996)) show that Adaboost continue to drop down generalization error as T becomes large.

2.2 Arcing family

Breiman (1998) used the ARCING term to describe the family of algorithms that Adaptively Resample data and Combine the outputted hypotheses. Adaboost was the first example of an arcing algorithm.

In order to study the behavior of Adaboost, Breiman developed an ad-hoc algorithm, Arc-x(j). This algorithm is similar to Adaboost but differs in the following:

- it uses a simpler weight updating rule:

$$D_{t+1}(i) = \frac{1 + m(i)^j}{\sum(1 + m(i)^j)}, \tag{7}$$

  where $m(i)$ is the number of misclassifications of instance i by classifiers $1, \ldots, t$ and j is an integer.
- classifiers are combined using simple majority vote.

Since the development of arcing family, Adaboost and Arc-x4 were compared in different framework and using different collections of datasets. Breiman (1998) and Bauer and Kohavi (1999) show that Arc-x4 has an accuracy comparable to Adaboost without using the weighting scheme to construct the final classifier. Breiman (1998) argues that higher values of j were not tested so improvement is possible.

In this empirical study, two other values of the parameter j, $j = 8$ and $j = 12$, are tested in the subsampling framework and compared to Adaboost and Arc-x4.

3 Empirical study

First, the base classifier and the performance measure used in the experiments are introduced then we the experimental results of each algorithm are presented. Finally, the performance of all algorithms are compared.

3.1 Base classifier and performance measure

Boosting algorithms require a base classifier as a subroutine that performs slightly better than random guessing. In our experiments, we use a simple algorithm, developed by Iba and Langley (1992), that induces a one node decision tree from a set of preclassified training instances.

In order to compare different boosting algorithms, we use a collection of binary data sets from UCI Machine learning Repository (Blakes et al. (1998)). Details of these data sets are presented in Table 2.

For each data set, we repeat the experiment 50 times. Each time, the data set is randomly partitioned into two equally sized sets. Each set is used once as a training set and once as a testing set. We run each algorithm for $T = 25$ and 75 iterations and report the average test error.

Bauer and Kohavi (1999) measures of performance are used. For a fixed number of iterations, the performance of each algorithm is evaluated using test error averaged over all data sets. To measure improvement produced by a boosting algorithm, absolute test error reduction and relative test error reduction are used.

3.2 Results

Results are reported in Table 1 and interpreted as follows: for a fixed number of iterations, we evaluate the performance of each algorithm on the collection of data sets and on each data set. Then all algorithms are compared for 25 and 75 iterations using test error averaged over all data sets.

Table 1. Average test error for each algorithm for 25 and 75 iterations on each data set.

	base	Adaboost		Arc-x4		Arc-x8		Arc-x12	
Data	Classifier	25	75	25	75	25	75	25	75
Liv.	41.81 %	29.78%	29.35%	29.96%	28.94%	31.94%	29.24%	34.28%	29.60%
Hea.	28.96%	19.58%	20.38%	18.99%	18.93%	20.25%	19.41%	21.79%	20.07%
Ion.	18.93%	12.38%	11.32%	12.27%	11.83%	12.22%	11.21%	12.59%	11.01%
Bre.	8.32%	4.56%	4.62%	3.88%	3.87%	4.22%	4.14%	4.40%	4.26%
Tic.	34.66%	28.80%	28.68%	29.59%	28.43%	31.38%	28.82%	30.11%	29.36%
mean	26.54%	19.02%	18.87%	18.94%	18.40%	20.00%	18.56%	20.63%	18.86%

Adaboost results: Adaboost decreases the average test error by 7.52% for 25 iterations and by 7.67% for 75 iterations. All data sets have relative test error reduction higher than 15%. The results for 75 iterations are better than those obtained for 25 iterations except for breast cancer data and heart data.

Table 2. Data sets used in the experimental study

Data set	number of instances	number of attributes
Liver disorders(Liv)	345	7
Heart (Hea)	270	13
Ionosphere (Ion)	351	34
Breast cancer (Bre)	699	10
Tic tac toe (Tic)	958	9

Arc-x(j) results: All Arc-x(j) algorithms decrease the test error. The relative test error reduction is higher than 15% for all datasets except when Arc-x(j) algorithms are applied for 25 iterations on the tic tac toe dataset. Results produced for 75 iterations are better than those obtained for 25 iterations.

Comparing algorithms: When comparing the results of the different boosting algorithms for 25 and 75 iterations, we notice that:

- For 25 iterations, the lowest average test error is produced by Arc-x4 algorithm.
- The relative average error reduction between Arc-x4 and Adaboost is 0.43% which is not significant.
- The average error of Arc-x4 is better than the average error of Arc-x8 by 5.62% and by 8.96% for Arc-x12 which are significant at 5% level.
- Arc-x4 and Adaboost produce the lowest error on 2 databases, Arc-x8 outperforms the other algorithms on 1 data set.
- For 75 iterations, Arc-x4 outperforms all other algorithms.
- Adaboost and Arc-x12 performs equally well and less accurately than Adaboost and Arc-x8.
- Arc-x4 produces the lowest error on 4 data sets and Arc-x12 on 1 data set.
- The relative average error reduction between the lowest and the highest error is 2.55% which is not significant.

4 Conclusion

This empirical study is an extension to Breiman's (1998) study on the family of boosting algorithms, the ARCING family. Two extensions of arcing weight updating rules are tested and compared to the one used by Breiman (1998) and to Adaboost in the subsampling framework.

Our empirical study shows that, based on these empirical results, increasing the factor j of Arc-x(j) algorithm does not improve the performance of

Arcing algorithms. The absolute test error reduction is higher for the first 25 iterations than for the last 50 iterations. It is interesting to look for another way of combining classifiers which gives more weight to the first ones and thus producing lower test error.

References

BAUER, E. KOHAVI, R. (1999): An empirical comparison of voting classification algorithms: Bagging, boosting and variants.*Machine learning, 36(1), 105–142.*

BLAKES, C. KEOGH and E. MERZ, C.J. (1998): UCI repository of machine learning databases. *http://www.ics.uci.edu/ mlearn/MLRepository.html*

BREIMAN, L. (1998): Arcing classifiers.*The Annals of Statistics, 26(3), 801–849.*

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000): Additive logistic regression: A statistical view of boosting. *The Annals of Statistics, 28(2), 337-407.*

FREUND, Y. (1995): Boosting a weak learning algorithm by majority. *Information and Computation, 121(2), 256–285.*

FREUND, Y. and SCHAPIRE, R.E. (1996): Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference, 148–156.*

FREUND, Y. and SCHAPIRE, R.E. (1997): A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences, 55(1), 119–139.*

IBA, W. and LANGLEY, P. (1992): Induction of one-level decision trees. *Proceedings the ninth international conference on machine learning*, 233–240.

SCHAPIRE, R.E. (1990): The strength of weak learnability. *Machine Learning, 5(2), 197–227.*

VAPNIK, V. (1998): *Statistical learning theory.* John Wiley & Sons INC., New York. A Wiley-Interscience Publication.

Iterative Majorization Approach to the Distance-based Discriminant Analysis

Serhiy Kosinov, Stéphane Marchand-Maillet, and Thierry Pun

Computer Vision and Multimedia Lab, University of Geneva,
24 Rue du General-Dufour, CH-1211, Geneva 4, Switzerland

Abstract. This paper proposes a method of finding a discriminative linear transformation that enhances the data's degree of conformance to the compactness hypothesis and its inverse. The problem formulation relies on inter-observation distances only, which is shown to improve non-parametric and non-linear classifier performance on benchmark and real-world data sets. The proposed approach is suitable for both binary and multiple-category classification problems, and can be applied as a dimensionality reduction technique. In the latter case, the number of necessary discriminative dimensions can be determined exactly. The sought transformation is found as a solution to an optimization problem using iterative majorization.

1 Introduction

Efficient algorithms, developed originally in the field of multidimensional scaling (MDS), quickly gained popularity and paved their way into discriminant analysis. Koontz and Fukunaga (1972), as well as Cox and Ferry (1993) proposed to include class membership information in the MDS procedure and recover a discriminative transformation by fitting *a posteriori* a linear or quadratic model to the obtained reduced-dimensionality configuration. The wide-spread use of guaranteed-convergence optimization techniques in MDS sparked the development of more advanced discriminant analysis methods, such as one put forward by Webb (1995), that integrated the two stages of scaling and model fitting, and determined the sought transformation as a part of the MDS optimization. These methods, however, focused mostly on deriving the transformation without adapting it to the specific properties of the classifier that is subsequently applied to the observations in the transformed space. In addition to that, these techniques do not explicitly answer the question of how many dimensions are needed to distinguish among a given set of classes.

In order to address these issues, we propose a method that relies on an efficient optimization technique developed in the field of MDS and focuses on finding a discriminative transformation based on the compactness hypothesis (see Arkadev and Braverman (1966)). The proposed method differs from the above work in that it specifically aims at improving the accuracy of the non-parametric type of classifiers, such as nearest neighbor (NN), Fix and Hodges (1951), and can determine exactly the number of necessary discriminative

dimensions, since feature selection is embedded in the process of deriving the sought transformation.

The remainder of this paper is structured as follows. In Section 2, we formulate the task of deriving a discriminant transformation as a problem of minimizing a criterion based on the compactness hypothesis. Then, in Section 3, we demonstrate how the method of iterative majorization (IM) can be used to find a solution that optimizes the chosen criterion. Section 4 describes the extensions of the proposed approach for dimensionality reduction and multiple class discriminant analysis, whereas the details of our experiments are provided in Section 5.

2 Problem formulation

Suppose that we seek to distinguish between two classes represented by matrices X and Y having N_X and N_Y rows of m-dimensional observations, respectively. For this purpose, we are looking for a transformation matrix $T \in \mathbb{R}^{m \times k}$, $k \ll m$, that eventuates in compactness within members of one class, and separation within members of different classes.

While the above preamble may fit just about any class-separating transformation method profile (e.g., Duda and Hart (1973)), we must emphasize several important assertions that distinguish the presented method and naturally lead to the problem formulation that follows. First of all, we must reiterate that our primary goal is to improve the NN performance on the task of discriminant analysis. Therefore, the sought problem formulation must relate only to the factors that directly influence the decisions made by the NN classifier, namely - the distances among observations. Secondly, in order to benefit as much as possible from the non-parametric nature of the NN, the sought formulation must not rely on the traditional class separability and scatter measures that use class means, weighted centroids or their variants which, in general, connote quite strong distributional assumptions. Finally, an asymmetric product form should be more preferable, justified as consistent with the properties of the data encountered in the target application area of multimedia retrieval and categorization, Zhou and Huang (2001). More formally, these requirements can be accommodated by an optimization criterion expressed in terms of distances among the observations from the two datasets as follows:

$$J(T) = \frac{\left(\prod_{i<j}^{N_X} \Psi\left(d_{ij}^W(T)\right) \right)^{\frac{2}{N_X(N_X-1)}}}{\left(\prod_{i=1}^{N_X} \prod_{j=1}^{N_Y} d_{ij}^B(T) \right)^{\frac{1}{N_X N_Y}}}, \tag{1}$$

where the numerator and denominator of (1) represent the geometric means of the within- and between-class distances defined as $\sqrt{(x_i - x_j)TT^T(x_i - x_j)^T}$

and $\sqrt{(x_i - y_j)TT^T(x_i - y_j)^T}$, respectively, and $\Psi(\cdot)$ denotes a Huber robust estimation function, Huber (1964), parametrized by a positive constant c and defined as:

$$\Psi(d_{ij}^W) = \begin{cases} \frac{1}{2}\left(d_{ij}^W\right)^2 \text{ if } d_{ij}^W \leq c; \\ cd_{ij}^W - \frac{1}{2}c^2 \text{ if } d_{ij}^W > c. \end{cases} \quad (2)$$

The choice of Huber function in (1) is motivated by the fact that at c the function switches from quadratic to linear penalty allowing to mitigate the consequences of an implicit unimodality assumption that the formulation of the numerator of (1) may lead to. In the logarithmic form, criterion (1) is written as:

$$\begin{aligned} \log J(T) &= \frac{2}{N_X(N_X-1)} \sum_{i<j}^{N_X} \log \Psi\left(d_{ij}^W(T)\right) - \frac{1}{N_X N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \log d_{ij}^B(T) \quad (3) \\ &= \alpha S_W(T) - \beta S_B(T). \end{aligned}$$

Our preliminary studies, Kosinov (2003), have shown that neither straightforward gradient descent nor some of the state-of-the-art optimization routines are suitable for solving the above optimization problem mostly due to susceptibility to local minima, adverse dependence on the initial value, and difficulties related to the discontinuities of the derivative of (3). However, by deriving some approximations of $S_W(T)$ and $S_B(T)$ one can make the task of minimizing $\log J(T)$ criterion amenable to a simple iterative procedure based on the majorization method (Borg and Groenen (1997), de Leeuw (1977), Heiser (1995)), which we discuss in the following section.

3 Iterative majorization

It can be verified that majorization remains valid under additive decomposition. Therefore, a possible strategy for majorizing (3) is to deal with $S_W(T)$ and $-S_B(T)$ separately and subsequently recombine their respective majorizing expressions. We begin by noting that both the logarithm and Huber function are majorizable by linear and quadratic functions, respectively, Heiser (1995). This fact makes it possible to derive a majorizing function of $S_W(T)$ as follows:

$$S_W(T) = \sum_{i<j}^{N_X} \log \Psi\left(d_{ij}^W(T)\right) \leq \sum_{i<j}^{N_X} \frac{\bar{w}_{ij} \cdot \left(d_{ij}^W(T)\right)^2}{2\Psi\left(d_{ij}^W(\bar{T})\right)} + K_1 = \mu_{S_W}(T, \bar{T}), \quad (4)$$

where $T, \bar{T} \in \mathbb{R}^{m \times m}$, $\bar{T}$ is a supporting point for T, $\bar{w}_{ij}$ is a weight of the Huber function majorizer, that in this case is equal to 1 if $\Psi(d_{ij}^W(\bar{T})) < c$ or $c/\Psi(d_{ij}^W(\bar{T}))$ otherwise, and K_1 is a constant term with respect to T.

Switching to matrix notation and defining a square symmetric design matrix B dependent on $\bar{T}$:

$$b_{ij} = \begin{cases} -\dfrac{\bar{w}_{ij}}{\Psi\left(d_{ij}^{W}(\bar{T})\right)} & \text{if } i \neq j; \\ \displaystyle\sum_{k=1,k\neq i}^{N_X} b_{ik} & \text{if } i = j; \end{cases} \tag{5}$$

leads to the majorizing expression of $S_W(T)$ in its final form:

$$\mu_{S_W}(T, \bar{T}) = \frac{1}{2}\mathbf{tr}\left(T^T X^T B X T\right) + K_1. \tag{6}$$

An attempt to majorize $-S_B(T)$ directly runs into problems due to the difficulties of finding a proper quadratic majorizing function of the negative logarithm. As a practical solution, we replace the neg-logarithm with its piece-wise linear approximation (see Figure 1, left panel), which, in turn, can

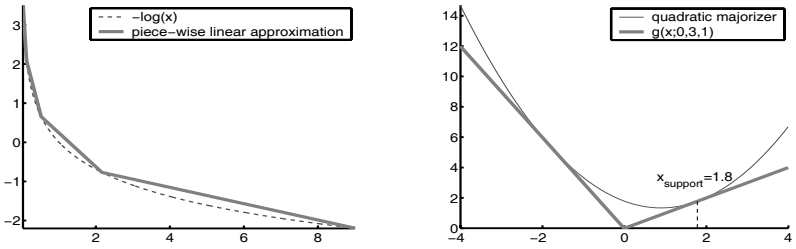


Fig. 1. Majorization of piecewise-linear approximation of $-log(x)$

be represented as a sum of the functions defined as:

$$g(x; x_0, l, r) = \begin{cases} r(x - x_0) & \text{if } x \geq x_0, \\ -l(x - x_0) & \text{if } x < x_0; \end{cases} \tag{7}$$

where $l + r > 0$, to ensure convexity. It is easy to see that the family of functions defined in (7) is one of the many possible generalizations of the absolute value function $|x|$, the former being equivalent to the latter whenever $x_0 = 0$ and $l = r = 1$. Similarly to $|x|$, $g(x; x_0, l, r)$ can be majorized by a quadratic $ax^2 + bx + c$ with coefficients $a > 0$, b and c determined from the majorization requirements (see an example in Figure 1, right panel). Finally, $-S_B(T)$ expressed in terms of the above quadratics can be majorized by the following function, written in matrix notation as:

$$\mu_{-S_B}(T, \bar{T}) = \mathbf{tr}(T^T Z^T G Z T) - \mathbf{tr}(T^T Z^T C Z \bar{T}) + K_2, \tag{8}$$

where Z is the matrix obtained by joining X and Y together, row-wise, and G, C are design matrices dependent on $\bar{T}$, whose non-zero elements m_{ij} are:

$$m_{ij} = \begin{cases} p_{ij} & \text{for } i \in [1; N_X] \text{ and } j \in [N_X + 1; N], \\ p_{ij} & \text{for } i \in [N_X + 1; N] \text{ and } j \in [1; N_X], \\ \sum\limits_{k=1, k \neq i}^{N_X + N_Y} m_{ik} & \text{for } i = j, \end{cases} \tag{9}$$

where p_{ij} is equal to -1 and $-1/\left(d_{ij}^B(\bar{T})\right)^2$ for C and G, respectively (see Kosinov (2003) for derivation details and a description of an alternative faster method based on Taylor series expansion).

Finally, combining results (6) and (8), we obtain a majorizing function of the $\log J(T)$ optimization criterion:

$$\begin{aligned} \mu_{\log J}(T, \bar{T}) &= \alpha \mu_{S_W} + \beta \mu_{-S_B} \\ &= \frac{\alpha}{2} \mathbf{tr}\left(T^T X^T B X T\right) + \beta \mathbf{tr}(T^T Z^T G Z T) \\ &\quad - \beta \mathbf{tr}(T^T Z^T C Z \bar{T}) + K_3, \end{aligned} \tag{10}$$

that is used to find an optimal transformation T minimizing $\log J(T)$ criterion via the iterative procedure described in Heiser (1995), and, thus, constitutes the core of the proposed distance-based discriminant analysis (DDA) method.

While at every iteration it is possible to minimize (10) by solving a system of linear equations, it is often recommended, Krogh and Hertz (1992), that a length-constrained solution be found, especially in the case of classifiers capable of achieving zero training error, to prevent overfitting. By incorporating the constraint into the Lagrangian, we obtain a standard trust-region subproblem, for which efficient solution methods exist, Rojas et al. (2000), Hager (2001).

4 Dimensionality reduction and multiple-class setting

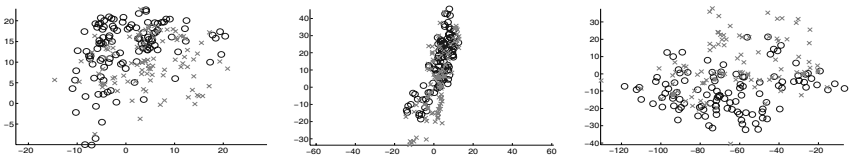
For any $T \in \mathbb{R}^{m \times k}$, $k < m$, the proposed method has an additional advantage of being a dimensionality reduction technique. Moreover, the value of k, i.e., the exact number of dimensions the data can be reduced to without loss of discriminatory power with respect to (3), is precisely determined by the number of non-zero singular values of T. Indeed, the distances between the transformed observations may be viewed as distances between the original observations in a different metric TT^T, that can be expressed as $TT^T = USV^TVSU^T = U_k S_k^2 U_k^T$ using the singular value decomposition of T. The obtained expression reveals that the effect of the full-dimensional transformation T is captured by the first k left-singular vectors of T scaled by the corresponding non-zero singular values, whose number gives an answer to the question of how many dimensions are needed in the transformed space.

While the above discussion is concentrated mostly on the two-class configuration, it is straightforward to generalize the presented formulation to a multiple-class discriminant analysis setting, for the number of classes $K \geq 2$:

$$\log J_K(T) = \sum_{i=1}^{K-1} \left(\alpha^{(i)} S_W(T)^{(i)} - \beta^{(i)} S_B(T)^{(i)} \right). \tag{11}$$

5 Experimental results

Our empirical analysis was based on data sets from the UCI Machine Learning Repository, Blake and Merz (1998). First of all, we verified that the solutions



(a) $\log J = -0.17$ (b) $\log J = -0.22$ (c) $\log J = -0.19$

Fig. 2. Two-dimensional discriminative projections of the Sonar data set: inferior solutions found by the gradient descent method

of the optimization problem formulated in Section 2 found by the proposed method were of better quality compared to those of generic techniques, confirming the results reported by Van Deun and Groenen (2003), and Webb (1995). Indeed, numerous random initializations of the gradient search led to inferior as well as unstable results reflected in higher values of $\log J$ (see Figure 2), while the IM-based method proved nearly insensitive to the choice of the initial supporting point and regularly reached far lower criterion values maintaining convergence property at all times. We also validated the pro-

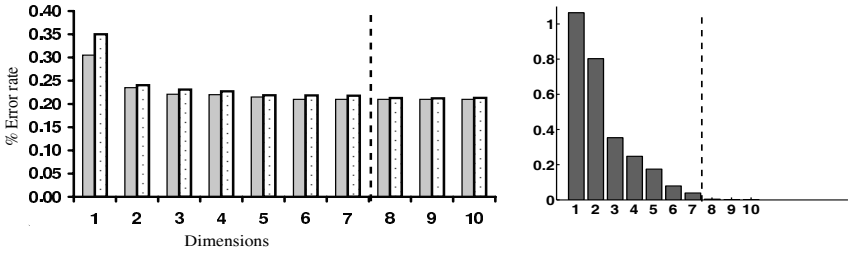


Fig. 3. Dimensionality reduction experiments: classification performance results (left) and singular values of $T \in \mathbb{R}^{m \times m}$ (right). The dashed lines mark the boundary that determines the dimensionality of the transformed space.

posed dimensionality reduction technique by analysing how the classification performance varied with respect to k, the dimensionality of the transformed space, and how it was related to the number of non-zero singular values of the full-dimensional transformation, an example of which for the Sonar data set is depicted in Figure 3. The right pane plots 10 largest out of 60 singular values of the full-dimensional transformation, in descending order, while the left diagram shows the results of 10-fold cross-validation experiments with respect to the transformed space dimensionality. Dot-filled bars denote performance achieved by fixing k *a priori*, while shaded bars show results obtained from a k-truncated SVD of the full-dimensional transformation. It is easy to see that the singular values beyond the 7^{th} are virtually zero. And as the diagram on the left confirms, adding dimensions beyond 7 no longer improves the classification performance (confirmed by Chow test at 99% confidence).

The experiments with the rest of the UCI data sets compared 10-fold cross-validation classification performance of the nearest neighbor classifier in the original feature space (denoted as NN) and that achieved in the transformed space derived by the proposed distance-based discriminant analysis method (denoted henceforth as DDA+NN). Hence, the goal of this analysis was to assess the effect of applying a DDA transformation on the accuracy of the NN classifier. The error rates of NN and DDA+NN data classification experiments are presented in Table 1, showing a consistent improvement

Table 1. Classification results for UCI data sets

Data set	Classes	% Error of NN	% Error of DDA+NN
Hepatitis	2	29.57	0.00
Ionosphere	2	13.56	7.14
Diabetes	2	30.39	27.11
Heart	2	40.74	21.11
Monk's P1	2	14.58	0.69
Balance	3	21.45	3.06
Iris	3	4.00	3.33
DNA	3	23.86	6.07
Vehicle	4	35.58	24.70

in performance. A separate set of experiments (see Kosinov (2003) for details) using the ETH80 database also revealed the importance of the length constraint, introduced in Section 3 to avoid overfitting. The results of these tests demonstrated up to 20% better classification accuracy for the length-constrained version of the method. Additionally, the results of our more recent experiments reveal that the DDA combined with an SVM classifier, Cristianini and Shawe-Taylor (2000), produces a smaller number of support vectors leading to better classification accuracy.

References

ARKADEV, A. and BRAVERMAN, E. (1966): *Computers and Patter Recognition*. Thompson, Washington, D.C.

BLAKE, C. and MERZ, C. (1998), UCI Repository of machine learning databases.

BORG, I. and GROENEN, P. J. F. (1997): *Modern Multidimensional Scaling*. New York, Springer.

COX, T., FERRY, G. (1993): Discriminant analysis using nonmetric multidimensional scaling. *Pattern Recognition, 26(1), 145–153*.

CRISTIANINI, N. and SHAWE-TAYLOR, J. (2000): *An introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, Cambridge.

DE LEEUW, J. (1977): Applications of convex analysis to multidimensional scaling. *Recent Developments in Statistics*, 133–145.

DUDA, R. O. and HART, P. E. (1973): *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York.

FISHER, R. A. (1936): The Use of Multiple Measures in Taxonomic Problems. *Ann. Eugenics, 7, 179–188*.

FIX, E. and HODGES, J. (1951): Discriminatory analysis: Nonparametric discrimination: Consistency properties. Tech. Rep. 4, USAF School of Aviation Medicine.

HAGER, W. (2001): Minimizing quadratic over a sphere. *SIAM Journal on Optimization, 12(1), 188–208*.

HEISER, W. (1995): Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. *Recent advances in descriptive multivariate analysis*, 157–189.

HUBER, P. (1964): Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35, 73–101*.

KOONTZ, W. and FUKUNAGA, K. (1972): A nonlinear feature extraction algorithm using distance information. *IEEE Trans. Comput., 21(1), 56–63*.

KOSINOV, S. (2003): Visual object recognition using distance-based discriminant analysis. Tech. Rep. 03.07, Computer Vision and Multimedia Laboratory, Computing Centre, University of Geneva, Rue Général Dufour, 24, CH-1211, Geneva, Switzerland.

KROGH, A. and HERTZ, J. A. (1992): A Simple Weight Decay Can Improve Generalization. In: J. E. Moody, S. J. Hanson and R. P. Lippmann (Eds.), *Advances in Neural Information Processing Systems*, vol. 4, 950–957. Morgan Kaufmann Publishers, Inc.

ROJAS, M., SANTOS, S. and Sorensen, D. (2000): A New Matrix-Free Algorithm for the Large-Scale Trust-Region Subproblem. *SIAM Journal on Optimization, 11(3), 611–646*.

VAN DEUN, K. and GROENEN, P. J. F. (2003): Majorization Algorithms for Inspecting Circles, Ellipses, Squares, Rectangles, and Rhombi. Tech. rep., Econometric Institute Report EI 2003-35.

WEBB, A. (1995): Multidimensional scaling by iterative majorization using radial basis functions. *Pattern Recognition, 28(5), 753–759*.

ZHOU, X. and HUANG, T. (2001): Small Sample Learning during Multimedia Retrieval using BiasMap. In: *IEEE Computer Vision and Pattern Recognition (CVPR'01)*, Hawaii.

An Extension of the CHAID Tree-based Segmentation Algorithm to Multiple Dependent Variables

Jay Magidson[1] and Jeroen K. Vermunt[2]

[1] Statistical Innovations Inc., 375 Concord Avenue, Belmont, MA 02478, USA
[2] Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, Netherlands

Abstract. The CHAID algorithm has proven to be an effective approach for obtaining a quick but meaningful segmentation where segments are defined in terms of demographic or other variables that are predictive of a *single* categorical criterion (dependent) variable. However, response data may contain ratings or purchase history on *several* products, or, in discrete choice experiments, preferences among alternatives in each of *several* choice sets. We propose an efficient hybrid methodology combining features of CHAID and latent class modeling (LCM) to build a classification tree that is predictive of *multiple* criteria. The resulting method provides an alternative to the standard method of profiling latent classes in LCM through the inclusion of (active) covariates.

1 Background and summary of approach

The CHAID (Chi-Squared Automatic Interaction Detection) tree-based segmentation technique has been found to be an effective approach for obtaining meaningful segments that are predictive of a K-category (nominal or ordinal) criterion variable. For example, the dependent variable might be response to a mailing (responders vs. non-responders). Each of the resulting segments, depicted as a terminal node in a tree diagram, is defined as a combination of directly observable categorical predictors such as AGE = 18-24 & INCOME = \$80,000+. Descriptive entries in each tree node consist of the sample size and the corresponding observed distribution on the dependent variable (e.g., associated response rate).

Latent class (LC) models are useful in identifying segments that underlie *multiple* response variables. While the resulting latent classes can be either ordered (ordinal latent variable) or unordered (nominal latent variable), they are not actionable like CHAID segments, because by definition they are unobservable (latent).

In this paper we propose a hybrid methodology that combines strengths of both approaches. After decomposing a set of M response variables into K underlying latent class segments, a modified CHAID algorithm is used with the K latent classes serving as the K-category nominal (ordinal) criterion variable. The resulting CHAID segments, derived from selected demographic

or other exogenous variables that are predictive of the classes, should also tend to be predictive of the M criterion variables.

The hybrid method also provides an alternative to the use of covariates in LCM to profile the classes. In practice, one or more demographic or other exogenous variables are included in an LCM to describe/predict the latent classes using a multinomial logit model. The proposed CHAID-based alternative is especially advantageous when the number of covariates is large, when covariate effects are non-linear, or when there are complicated higher-order interactions.

In the next section we provide brief introductions to the standard CHAID algorithm and the standard LC (cluster and factor) models. We then provide the technical details of the hybrid approach, followed by an empirical example from a pre-post survey (Burns et al. (2001)). We conclude with some final remarks.

2 The CHAID algorithm

The original CHAID algorithm was introduced by Kass (1980) for nominal dependent variables. CHAID is a recursive partitioning method useful in exploratory analyses that relate a potentially large number of categorical predictor variables to a single categorical nominal dependent variable. It was extended to ordinal dependent variables by Magidson (1993) who illustrated how this extension could be used to take advantage of fixed scores such as profitability, for each category of the dependent variable when such scores are known, as well as how to estimate meaningful scores when category scores are unknown. Chi-squared goodness of fit tests are used to identify significant predictors, and to merge predictor categories that do not differ in their prediction of the dependent variable.

Predictor categories are eligible to be merged according to specified scale types. Any categories of Nominal ("free") predictors can be merged, while only adjacent categories of ordinal or grouped continuous ("monotonic") predictors are allowed to merge. A final scale type ("float") may be used to specify that the variable is to be treated as monotonic except for the final category, often corresponding to a 'don't know' or 'missing' response, which is free to merge with any of the other categories. Technical settings include significance levels associated with merging and splitting and a stopping rule. A case weight and a frequency variable may also be included in the analysis.

As an example, Figure 1 illustrates a CHAID analysis based on data from a post-election survey on 1,051 persons who voted for either Bush or Gore in the 2000 U.S. election. The dependent variable (VOTE) is the candidate voted for and the predictors are 5 demographic variables: 1) MARSTAT (1=married, 2=widowed, 3= separated/divorced, 4= never married, 5= other – "Free"), AGEr (1=18-24, 2=25-34, 3=35-44, 4= 45-54, 5= 55-64, 6=65+, '.' = refused – "Float"), GENDER (1 = male, 2 = female), EDUCATION

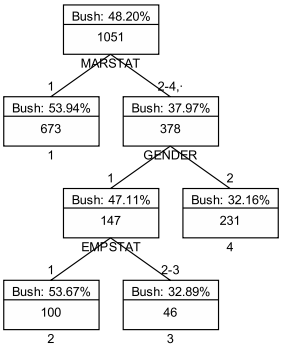


Fig. 1. CHAID tree for VOTE.

(1 = less than HS, 2= HS grad, 3= some college, 4=college grad, 5= post grad, 5-refused – "Float"), and EMPLOYED (1 = Yes, 2 = No, 3 = retired – "Free").

Overall, 48.2% voted for Bush. This is displayed in the top (root node) of the tree. Among the 5 demographic predictors included in this analysis, only 2 were significant at the root node – MARSTAT ($p<.00001$), and GENDER ($p<.01$). The CHAID analysis resulted in 4 segments. The best segments for Bush are S1, consisting of the 673 married voters (53.94% for Bush) and S2 consisting of 100 unmarried employed males (53.67% for Bush). The remaining segments – S3 (unmarried unemployed males) and S4 (unmarried females) – voted more than 2:1 in favor of Gore over Bush.

One limitation of CHAID is that segments are defined based on a single criterion variable. Given situations where multiple criteria exist, it is not clear how one should go about obtaining a single common segmentation. Using one dependent variable as the criterion may result in one set of segments, while use of an alternative dependent variable will likely yield a different set of segments. Moreover, the categories of a predictor may merge in different ways depending upon which dependent variable is used, again leading to different segments.

In addition, when multiple dependent variables do exist, they may be of different scale types (nominal, ordinal, continuous, count, etc.). Using a 3-category response variable as an example Magidson (1993) showed that CHAID segments resulting from treating the dependent variable as ordinal (using profitability scores for the categories) differed substantially from segments derived from the nominal algorithm which ignored the scores. The hybrid approach resolves the need to chose between different segmentations because indicators with differing scale types can be used in extended LCMs, yielding a single LC solution. An important advantage of this hybrid approach over approaches based on specific measures for node homogeneity rather than a model (e.g., Kim and Lee (2003)) is that the LC model used here can handle dependent variables of different scale types.

3 Latent class modeling

A LC model postulates a nominal K-category latent (unobservable) variable X to explain the associations/correlations between the observed response variables (multiple criteria; Lazarsfeld and Henry (1968); Goodman (1974)). Each category of X is called a latent class. Let Y_m denote one of M nominal response variables, $m = 1, 2..., M$; j_m is a particular response category and J_m the number of categories of variable Y_m. Notation $\mathbf{Y}$ and $\mathbf{j}$ is used to refer to a full response vector and a full set of response categories. The LC model for M response variables is defined as

$$P(\mathbf{Y} = \mathbf{j}) = \sum_{k=1}^{K} P(X = k, \mathbf{Y} = \mathbf{j}) = \sum_{k=1}^{K} P(X = k) P(\mathbf{Y} = \mathbf{j} | X = k)$$
$$= \sum_{k=1}^{K} P(X = k) \prod_{m=1}^{M} P(Y_m = j_m | X = k), \quad (1)$$

where $P(X = k)$ denotes the probability of being in latent class k, $k = 1, 2, ..., K$, and $P(Y_m = j_m | X = k)$ denotes the conditional probability of obtaining the j_mth response to item Y_m, from members of class k, $j_m = 1, 2, ..., J_m$.

Cases with response pattern $\mathbf{j}$ are typically classified into the latent class for which the posterior membership probability $P(X = k|\mathbf{Y} = \mathbf{j})$ is highest. Estimates for the posterior membership probabilities – for $k = 1, 2, ..., K$ – can be obtained using Bayes theorem as follows:

$$P(X = k|\mathbf{Y} = \mathbf{j}) = \frac{P(X = k, \mathbf{Y} = \mathbf{j})}{P(\mathbf{Y} = \mathbf{j})}. \quad (2)$$

The numerator and denominator were defined in equation (1).

Recent advances allow for dependent variables (indicators) of varying scale types to be used – including mixing categorical, continuous, and count variables – by specifying the appropriate probability densities $P(Y_m = j_m|X = k)$ (Vermunt and Magidson (2002)). By expressiong the mean of these densities in terms of a generalized linear model (GLM), one can include direct effects between 2 or more indicators, multiple categorical latent variables, continuous latent factors and/or other terms into the model (see Magidson and Vermunt (2001); Vermunt and Magidson, (2005)).

It is also possible to include one or more exogenous variables called covariates in a LCM, allowing one to explore the relationship between exogenous variables and the latent classes and assess the significance of such relationships in a formal way. However, the covariates included in LCM influence the estimates of the parameters in the original measurement model. If the covariate part of the model holds true, inclusion of the covariates improves the efficiency of the estimates. However, if it is misspecified, the estimates

may become somewhat biased. In addition, profiling latent classes in terms of many covariates may cause the solution to become unstable. As an alternative, Magidson and Vermunt (2001) allow covariates to be treated in an *inactive* manner – providing appropriate cross-tabulations but not influencing the original measurement model. But this approach comes at the expense of no longer being able to assess statistical significance.

In the next section, we show how the hybrid algorithm provides an alternative treatment to the use of both active and inactive covariates in LC models. The new approach provides an assessment of statistical significance for *selected* covariates included within the LCM framework, whether the covariate is specified as active or inactive. Those covariates specified as inactive do not alter the estimates obtained from the LCM.

4 The hybrid CHAID algorithm

Our hybrid CHAID algorithm involves 3 steps.

1. Perform an LC cluster analysis on M response variables to obtain K latent classes.
2. Perform a CHAID analysis using the K classes as a nominal dependent variable.
3. Obtain predictions for each of M response variables based on the resulting CHAID segments and/or on any preliminary set of CHAID segments.

Step 1 yields class-specific predicted probabilities for each category of the m-th dependent variable[1], as well as posterior membership probabilities for each case.

Step 2 yields a set of CHAID segments that differ with respect to their average posterior membership probabilities for each class. We use the posterior membership probabilities defined in equation (2) as fixed case weights as opposed to the modal assignment into one of the K classes. This weighting eliminates bias due to the misclassification error that occurs if cases were equated (with probability one) to that segment having the highest posterior probability. Specifically, each case contributes K records to the data, the kth record of which contains the value k for the dependent variable, and contains a case weight of $P(X = k|\mathbf{Y} = \mathbf{j})$, the posterior membership probability associated with that case. Thus, as opposed to the original algorithm where chi-square is calculated on observed 2-way tables, in the hybrid algorithm, the chi-squared statistic is computed on 2-way tables of *weighted* cell counts.[2]

If as an alternative to performing a standard LC analysis, one performs an LC factor analysis in step 1, in step 2 the CHAID *ordinal* algorithm can

[1] When one or more of the dependent variables are quantitative, for each class this step also yields predicted means for the quantitative dependent variables.

[2] The new algorithm also incorporates sampling weights, if present, using an efficient ML algorithm proposed by Vermunt and Magidson (2001).

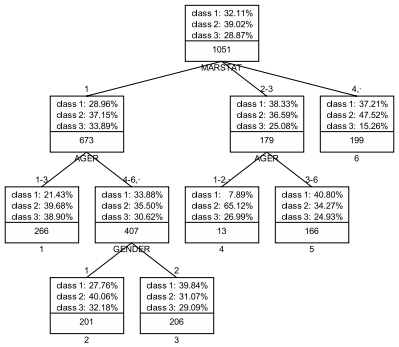


Fig. 2. Hybrid CHAID tree for 11 dependent variables.

be used to obtain segments based on the use of any of the LC factors as the ordinal dependent variable, or a single segmentation can be obtained using the nominal algorithm to identify segments based on the single joint latent variable defined as a combination of two or more identified LC factors.

Step 3 involves obtaining predictions for any or all of the M dependent variables for each of the I CHAID segments by cross-tabulating the resulting CHAID segments by the desired dependent variable(s). An alternative is to obtain predictions as follows

$$P(Y_m = j|i) = \sum_{k=1}^{K} P(Y_m = j|X = k)P(X = k|i).$$

As can be seen, we compute a weighted average of the class-specific distributions for dependent variable Y_m obtained in step 1 $[P(Y_m = j|X = k)]$, with the average posterior membership probabilities obtained in step 2 for segment i being used as the weights $[P(X = k|i)]$.

5 Empirical example

Among other questions, the pre-election survey solicited ratings for each candidate on 5 attributes – leadership, caring, knowledge, honesty and morality. A LCM was fit to these data, using VOTE as an active covariate, and the 5 demographics as inactive covariates. This model may be viewed as a kind of unsupervised regression with 11 dependent variables – VOTE, plus the 10 attribute ratings. This LCM yielded 3 segments. The first segment (32%) favored Gore, the second (39%) was neutral and the third favored Bush with respect to the attribute ratings and in their votes. These percentages are displayed in the root node of the hybrid CHAID tree in Figure 2.

The hybrid CHAID used the 3-category latent variable (segments) as the dependent variable and again utilized the 5 demographics as the predictors.

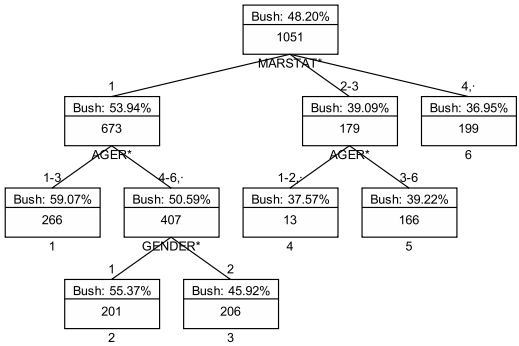


Fig. 3. Hybrid CHAID tree for VOTE.

At the root node 3 of the 5 predictors were found to be significant – MARSTAT ($p < .00002$), AGEr ($p<.001$), and GENDER ($p=.01$). Compared to our earlier CHAID, age is more important than when VOTE was the only dependent variable. The hybrid CHAID analysis resulted in 6 segments (Figure 2). Since the attributes are now included as additional dependent variables (the latent classes are a proxy for these dependent variables) we might expect that the resulting segments might predict any single dependent variable less well than CHAID based on only that dependent variable.

Figure 3 shows how the 6 hybrid segments predict VOTE. To compare this to the predictions based on our original segments (Figure 1) we first compare those segments favorable to Bush. Our previous analysis identified segments S1 and S2 as favorable to Bush. The hybrid CHAID (Figure 3) identifies 3 segments most likely to vote for Bush – segments 1, 2 and 3. Note that these 3 segments combined, are equivalent to the original segment S1. Since the hybrid CHAID fails to yield any additional segments that prefer Bush such as S2, it appears that the hybrid segmentation predicts VOTE less well than the original CHAID. Similarly, focusing on segments most favorable to Gore, our previous CHAID identified S3 and S4 (n= 277 cases) as favoring Gore by more than 2:1. The hybrid CHAID finds segments 4, 5 and 6 as favoring Gore, but not by as much as 2:1 over Bush.

6 Final comments

In this paper, we introduced a hybrid CHAID algorithm[3] as an extension of CHAID to multiple dependent variables of possibly differing scale types. Alternatively, this hybrid algorithm could be described as an alternative to the standard treatment of active and/or inactive covariates in LCM. The

[3] The extended CHAID algorithm has been implemented in a commercially available computer program called SI-CHAID 4.0, and works in conjunction with the latent class programs Latent GOLD 4.0 and Latent GOLD Choice 4.0.

CHAID-type output can simplify the process of examining the relationship between the demographics and/or other exogenous variables and the latent segments by 1) ranking the covariates from most to least significant and 2) for each covariate, merging categories that are not significantly different. This new output is especially valuable when the number of covariates is large.

We illustrated the hybrid algorithm here with dependent variables consisting of favorability ratings of Bush and Gore on 5 attributes plus the actual vote among 1,051 voters in the 2000 U.S. election. We showed how the hybrid CHAID provides a unique segmentation. We showed how it compares with a segmentation obtained using the traditional CHAID algorithm for a single dependent variable – VOTE. The results suggest that the segments resulting from the hybrid CHAID may fall somewhat short of predictability of any single dependent variable in comparison to the original algorithm, but makes up for this by providing a single unique set of segments that are predictive of all dependent variables.

References

BURNS, N., KINDER, D.R., ROSENSTONE, S.J., SAPIRO, V., and the National Election Studies (2001): National Election Studies, 2000: Pre-/Post- Election Study [dataset id:2000.T]. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer and distributor].

GOODMAN, L.A. (1974): Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61, 215–231.*

KASS, G. (1980): An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29, 119–127.*

KIM, S.J. and Lee, K.B. (2003): Constructing decision trees with multiple response variables. *International Journal of Management and Decision Making, 4, 289 – 311.*

LAZARSFELD, P. F. and HENRY, N.W. (1968): *Latent structure analysis.* Houghton Mifflin, Boston.

MAGIDSON, J. (1993): The use of the new ordinal algorithm in CHAID to target profitable segments. *The Journal of Database Marketing, 1, 29–48.*

MAGIDSON, J. and VERMUNT, J.K. (2001): Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology, 31, 223–264.*

VERMUNT, J.K. and MAGIDSON, J. (2001): *Latent class analysis with sampling weights.* Paper presented at the 6th annual meeting of the Methodology Section of the American Sociological Association, University of Minnesota, May 4-5, 2001.

VERMUNT, J.K. and MAGIDSON, J. (2002): Latent class cluster analysis. In: J.A. Hagenaars and A.L. McCutcheon (Eds.): *Applied latent class analysis.* Cambridge University Press, Cambridge, 89–106.

VERMUNT, J. K. and MAGIDSON, J. (2005): *Latent GOLD 4.0 User Manual.* Statistical Innovations Inc, Belmont MA.

Expectation of Random Sets and the 'Mean Values' of Interval Data

Ole Nordhoff

Institut für Statistik und Wirtschaftsmathematik,
RWTH Aachen, 52056 Aachen, Germany
nordhoff@stochastik.rwth-aachen.de

Abstract. Several possibilities of defining the expectation of random p-dimensional intervals are proposed. After defining the expectation via reducing intervals to their extremal points p-dimensional intervals (rectangles) are treated as Random Closed Sets (RCSs). In this framework Random Closed Rectangles (RCRs) are defined and the properties of different definitions for expectations of RCSs, applied on RCRs are studied. In addition known mean values of interval data are integrated in this generalized approach.

1 Introduction

Clustering methods often use class representatives or prototypes to describe data clusters. Prototypes are involved in many clustering criteria, where the dissimilarity between a data point and a cluster representative is considered. Moreover, the properties of a cluster are often characterised briefly by one single data point, e.g. the class centroid.
There are several clustering methods preparing p-dimensional interval data $x_1, \ldots, x_n$ with

$$\begin{aligned} x_i &= [a_i, b_i],\ a_i \leq b_i \in \mathbb{R}^p \quad (\text{that means } a_{i,j} \leq b_{i,j}\ \forall i, j), \\ &:= [a_{i,1}, b_{i,1}] \times \cdots \times [a_{i,p}, b_{i,p}], \quad i = 1, \ldots, n. \end{aligned}$$

For instance these data could be daily meteoroligical data (atmospheric pressure, temperature, air humidity, etc.) of a certain city, or medical data (blood pressure, temperature,...) of different patients. These data consist of many different measuremnts which are contained in an interval.
When preparing p-dimensional interval data with a certain clustering method one is correspondingly searching for the mean of intervals as the representative of all intervall data in a class. This leads to the question how the mean of some p-dimensional intervals is defined. We will introduce two different ways of defining a mean of (p-dimensional) intervals, which also can be taken by hyper-rectangles. First we reduce an interval to its minimum and its maximum value to shift the problem to the case of real-valued data, where the definition of expectation and mean is well known. In the second approach we treat an interval as a a special form of closed (convex) sets and use the theory of Random Closed Sets (RCSs) to define the mean via some different definitions of expectation.

2 Reduction to characteristic points

In this chapter a very obvious possibility to average a set of p-dimensional rectangles is studied. Like circles can be characterised only by their midpoint and their radius rectangles can also be reduced to some few points. Hence, we want to use the lower left vertex and the upper right vertex to characterise a subset of $\mathbb{R}^p$ by only two p-dimensional real-valued vectors. To put this approach in a more formal framework we use the transformations t, t^{-1} to switch between rectangles and pairs of p-dimensional vectors. On $\mathcal{Q} := \{Q \subset \mathbb{R}^p |\ Q = [a,b], a \leq b \in \mathbb{R}^p\}$ we define

$$\begin{aligned} t : \mathcal{Q} &\longrightarrow \mathbb{R}^p \times \mathbb{R}^p, && (1)\\ t(Q) = t([a,b]) &:= (a,b) \quad \forall Q \in \mathcal{Q}, \\ t^{-1} &: \{(a,b) \in \mathbb{R}^p \times \mathbb{R}^p |\ a \leq b\} \longrightarrow \mathcal{Q}, && (2)\\ t^{-1}((a,b)) &:= [a,b] \quad \forall (a,b) \in \{(a,b) \in \mathbb{R}^p \times \mathbb{R}^p |\ a \leq b\}. \end{aligned}$$

Definition 1. Let $(\Omega, \mathfrak{A}, P)$ be a probability space and $X = (A,B) : (\Omega, \mathfrak{A}) \to (\mathbb{R}^p \times \mathbb{R}^p, \mathfrak{B}^{2p})$ a random variable which satisfies

$$A \leq B \quad a.s. \tag{3}$$

Then we call X a *Random Point Rectangle (RPR).*

Definition 2. Let $X = (A,B)$ a Random Point Rectangle with the property that A, B are integrable. Then the *expectation* of X is defined as

$$\hat{\mathrm{E}}[X] := (\mathrm{E}[A], \mathrm{E}[B]). \tag{4}$$

Remark 1. Treating a p-dimensional interval as a pair of points gives us the ability to obtain a definition for the mean of a finite set of rectangles.
Let $\{Q_1, \ldots, Q_n\}$ be a set of p-dimensional intervals and $\mathrm{M}(\cdot)$ the empirical mean (corresponding to Definition 2 of expectation) of a finite set of p-dimensional RPRs. Then we obtain

$$\hat{\mathrm{M}} := t^{-1}(\mathrm{M}(t(Q_1), \ldots, t(Q_n))) \tag{5}$$

as a mean of p-dimensional intervals. This result coincides with the intuitive way to built the mean of finitely many rectangles and this identicalness leads from the also intuitive construction of $\hat{\mathrm{E}}$. Later in Section 4 we will get the same mean from a more general construction of expectation.

3 Several expectations of Random Closed Sets

Bearing in mind that the source of the treated rectangles can be values which shall be represented by their complete spectrum and not only by their extremal points now we treat the rectangles as 'real' subsets of $\mathbb{R}^p$.

According to this proceeding one considers p-dimensional intervals as realisations of set-valued random variables. We will introduce briefly to the more general theory of Random Closed Sets (RCSs) based on Matheron (Matheron (1975), Stoyan and Mecke (1983)) to provide a basis for several definitions of expectations.
Let $\mathcal{F}$ be the system of all closed subsets in $\mathbb{R}^p$ and $\mathcal{K}$ the system of all compact subsets. Then we consider the σ-algebra $\mathfrak{F}$ on F which contains for all $K \in \mathcal{K}$:

$$\mathcal{F}_K := \{F \in \mathcal{F} |\ F \cap K \neq \emptyset\}. \tag{6}$$

Definition 3. A *Random Closed Set (RCS)* is a random variable X with values in $(\mathcal{F}, \mathfrak{F})$.
X is called *convex*, if X is convex almost surely.
X is called a *Random Closed Rectangle (RCR)*, if X is a closed rectangle almost surely.

Remark 2. The distribution P^X of a RCS X is determined by knowing $P^X(\mathcal{F}_K)$ for all $K \in \mathcal{K}$.

3.1 The Aumann expectation

Definition 4. Let $(\Omega, \mathfrak{A}, P)$ be a probability space and $X : (\Omega, \mathfrak{A}) \to (\mathcal{F}, \mathfrak{F})$ a RCS. A random point $\phi : (\Omega, \mathfrak{A}) \to (\mathbb{R}^p, \mathfrak{B}^p)$ is called *selection* of X, if

$$\phi(\omega) \in X(\omega) \quad a.s. \tag{7}$$

Definition 5. Let $(\Omega, \mathfrak{A}, P)$ be a probability space, $X : (\Omega, \mathfrak{A}) \to (\mathcal{F}, \mathfrak{F})$ a RCS and Φ_X the set of all selections of X. Then

$$\mathrm{E}_A[X] := \{\mathrm{E}[\phi] |\ \phi \in \Phi_X\} \tag{8}$$

is called the *Aumann expectation* of X.

3.2 The Frechet expectation

Definition 6. For closed $A, B \subseteq \mathbb{R}^p$ and with $d_e : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$ the euclidean distance we can define the *Hausdorff-distance* between A and B as:

$$d(A, B) := \max\{\max_{x \in A} \min_{y \in B} d_e(x, y), \max_{y \in B} \min_{x \in A} d_e(x, y)\}. \tag{9}$$

Let be X a Random Closed Set. Then the solution K_0 of

$$\mathrm{E}[d(X, K_0)^2] = \min_{K \in \mathcal{K}} \mathrm{E}[d(X, K)^2] \tag{10}$$

is called the *Frechet expectation* $\mathrm{E}_F[X]$ of X.

Remark 3. Although the Frechet expectation is in general very hard to specify, because it is a solution (existence, uniqueness?) of a hard optimisation problem it can be generalised to a whole class of different expectations, if the type of distance is changed. We can consider the Aumann expectation as a special case (see Molchanov (1997)) and moreover, we are able to choose special distances for rectangles such that the expectation of RCRs has a closed solution (see Section 4).
The second advantage of this definition of expectation is the fact that we obtain a formulation of a corresponding variance, too. If

$$\mathrm{V}_F := \min_{K \in \mathcal{K}} \mathrm{E}[d(X,K)^2] \tag{11}$$

is considered as a *Frechet-Variance*, it is possible to show several properties of this object the variance of a real-valued random variable has.

3.3 The Doss expectation

As the kinds of expectation introduced above the Doss expectation is also defined via a distance measure.

Definition 7. Let $(\Omega, \mathfrak{A}, P)$ be a probability space and $X : (\Omega, \mathfrak{A}) \to (\mathcal{F}, \mathfrak{F})$ a RCS. Consider for an arbitrary $x \in \mathbb{R}^p$

$$M_x := \{y \in \mathbb{R}^p |\ d_e(x,y) \leq \mathrm{E}[d(x,X)]\}, \tag{12}$$

then

$$\mathrm{E}_D[X] := \bigcap_{x \in \mathbb{R}^p} M_x = \{y \in \mathbb{R}^p |\ d_e(x,y) \leq \mathrm{E}[d(x,X)]\ \forall x \in \mathbb{R}^p\} \tag{13}$$

is called the *Doss expectation* of X.

Remark 4. In analogy to the Frechet expectation it is possible to vary the distance measure. A whole family of different expectations of RCSs can be obtained that way, but there are even more possibilities. Instead of using the expectation $\mathrm{E}[d(x,X)]$ in (12) it would be possible to use an arbitrary functional on $\mathbb{R}^p$.

Theorem 1. *The Doss expectation of a Random Closed Set is convex.*

Proof. see Nordhoff (2003).

3.4 The Vorob'ev expectation

In contrast to the other presented expectations the Vorob'ev expextation is defined via the volume of a Random Closed Set. As the volume of the boundary is zero, the boundary of the RCS will not play that important role

as in the other definitions of expectation.
The Vorob'ev expectation uses the characteristic function of a RCS, viz the function $\mathbb{1}_X : \mathbb{R}^p \times \Omega \to \{0,1\}$ with

$$\mathbb{1}_{X(\omega)}(x) = \begin{cases} 1, & \text{if } x \in X(\omega), \\ 0, & \text{else.} \end{cases} \tag{14}$$

Definition 8. Let $(\Omega, \mathfrak{A}, P)$ be a probability space and $X : (\Omega, \mathfrak{A}) \to (\mathcal{F}, \mathfrak{F})$ a RCS. Then the function $p_X : \mathbb{R}^p \to [0,1]$ with

$$p_X(x) := \mathrm{E}[\mathbb{1}_X(x)] = \int_\Omega \mathbb{1}_{X(\omega)}(x)\, dP(\omega) = P(x \in X) \tag{15}$$

is called *cover function* of X.

Definition 9. Let $(\Omega, \mathfrak{A}, P)$ be a probability space, $X : (\Omega, \mathfrak{A}) \to (\mathcal{F}, \mathfrak{F})$ a RCS and $p_X : \mathbb{R}^p \to [0,1]$ the cover function of X. For an arbitrary $q \in [0,1]$ and the Lebesgue-measure λ^p in $\mathbb{R}^p$ one considers

$$Lq(X) := \{x \in \mathbb{R}^p |\ p_X(x) \geq q\} \quad \text{and} \tag{16}$$

$$q_0 := \inf\{q \in [0,1] |\ \lambda^p(L_q(X)) \leq \mathrm{E}[\lambda^p(X)]\}. \tag{17}$$

Then the *Vorob'ev expectation* $\mathrm{E}_V[X]$ of X is defined by

$$\mathrm{E}_V[X] := L_{q_0}. \tag{18}$$

Remark 5. The set $L_{\frac{1}{2}}(X)$ of a compact RCS X is often named *Median* (see Stoyan and Stoyan (1994)), because it minimises the expected volume of the symmetric difference between X and a Borel-set.

4 Expectations of Random Closed Rectangles

After introducing several definitions of epectations for RCSs one is interested in the behaviour of theses expectations, if the underlying sets are Random Closed Rectangles.

4.1 The Aumann expectation

A very pleasant property of the Aumann expectation and the resulting mean for a finite number of fixed closed rectangles is the easy evaluation. More preciseley they coincide with the expectation/mean of Definition 2 and Remark 1.

Theorem 2. *Let $(\Omega, \mathfrak{A}, P)$ be a probability space, $X : (\Omega, \mathfrak{A}) \to (\mathcal{F}, \mathfrak{F})$ a RCR and the functions t, t^{-1} defined like in (2) und (3). If the expectation $\hat{E}[t(X)]$ exists, it is applied to the Aumann- Expectation:*

$$E_A[X] = t^{-1}(\hat{E}[t(X)]). \tag{19}$$

Proof. Let $(\Omega, \mathfrak{A}, P)$ be a probability space, $\mathcal{F}$ the system of closed sets in $\mathbb{R}^p$ and let for the RCR $X : (\Omega, \mathfrak{A}) \to (\mathcal{F}, \mathfrak{F})$

$$\begin{aligned} X(\omega) &= [A(\omega), B(\omega)] \\ &= [A_1(\omega), B_1(\omega)] \times [A_2(\omega), B_2(\omega)] \times \cdots \times [A_p(\omega), B_p(\omega)] \end{aligned}$$

with $A(\omega) \leq B(\omega) \in \mathbb{R}^p$ *a.s.* be valid.
To raise survey we abandon the transformations t, t^{-1}. That means, we write $\hat{E}[X]$ instead of $t^{-1}(\hat{E}[t(X)])$ in this proof, although the expectation $\hat{E}[\cdot]$ is defined on the set of Random Point Rectangles.
Like in Chapter 2 it is

$$\hat{E}[X] := [E[A], E[B]] = [E[A_1], E[B_1]] \times \cdots \times [E[A_p], E[B_p]]. \tag{20}$$

We want to show: $E_A[X] = \hat{E}[X]$.

$\subseteq$: We consider $x \in E_A[X]$, then there is a selection $\phi : (\Omega, \mathfrak{A}) \to (\mathbb{R}^p, \mathfrak{B}^p)$ of X with $x = E[\phi]$ and $\phi(\omega) \in X(\omega)$ *a.s.* Thus

$$\begin{aligned} E[A] = \int_\Omega A(\omega)\ dP(\omega) \leq \int_\Omega \phi(\omega)\ dP(\omega) = x \\ \leq \int_\Omega B(\omega)\ dP(\omega) = E[B], \end{aligned} \tag{21}$$

because $A(\omega) \leq \phi(\omega) \leq B(\omega)$ *a.s.* It follows from (21) that

$$x \in \hat{E}[X]. \tag{22}$$

$\supseteq$: Consider now $x \in \hat{E}[X]$, then there are real-valued $0 \leq t_1, \ldots, t_p \leq 1$ satisfying

$$x = \begin{pmatrix} t_1 E[A_1] + (1-t_1)E[B_1] \\ \vdots \\ t_p E[A_p] + (1-t_p)E[B_p] \end{pmatrix} = E\left[\begin{pmatrix} t_1 A_1 \\ \vdots \\ t_p A_p \end{pmatrix}\right] + E\left[\begin{pmatrix} (1-t_1)B_1 \\ \vdots \\ (1-t_p)B_p \end{pmatrix}\right].$$

We choose now $\phi : (\Omega, \mathfrak{A}) \to (\mathbb{R}^p, \mathfrak{B}^p)$ with

$$\phi(\omega) = \begin{pmatrix} t_1 A_1(\omega) \\ \vdots \\ t_p A_p(\omega) \end{pmatrix} + \begin{pmatrix} (1-t_1)B_1(\omega) \\ \vdots \\ (1-t_p)B_p(\omega) \end{pmatrix} \quad \forall \omega \in \Omega.$$

Then ϕ is a selection of X, due to $\phi(\omega) \in X(\omega)\ \forall \omega \in \Omega$ and we obtain $x = E[\phi]$. So we can conclude that $x \in E_A[X]$.

With the aid of Theorem 2 the canonical definition of *Aumann mean* can be replaced by a simple representation, if the underlying objects are rectangles.

Proposition 1. *Let $Q_1, \ldots, Q_n$ be fixed closed p-dimensional intervals. To get a definition for 'mean' via Aumann expectation we construct a finite probability space $(\Omega, \mathfrak{A}, P)$ with $\Omega = \{\omega_1, \ldots, \omega_n\}$, $\mathfrak{A} = \mathfrak{P}(\Omega)$ and P as the uniform distribution on Ω. By defining a Random Closed Rectangle X with $X(\omega_i) := Q_i, \quad i = 1, \ldots, n$, we get the* Aumann mean $M_A(Q_1, \ldots, Q_n) := E_A[X]$, *and taking note of Theorem 2 we obtain*

$$M_A(Q_1, \ldots, Q_n) = \hat{M}(Q_1, \ldots, Q_n). \tag{23}$$

4.2 The Frechet expectation

As proposed in Remark 3 it is possible to define various kinds of expectation and mean, if we consider several distance measures between sets. In the case of p-dimensional intervals there are several distance measures between intervals (for example see Chavent (2000)).

Example 1. Now, we take a look at the distance $\tilde{d} : \mathcal{Q} \times \mathcal{Q} \to \mathbb{R}+$, satisfying

$$\tilde{d}([a,b],[a',b']) := \sqrt{\sum_{i=1}^{p} (a_i - a_i')^2 + (b_i - b_i')^2} \quad a \leq b, a' \leq b' \in \mathbb{R}^p. \tag{24}$$

Looking for a closed form of the Frechet expectation of RCRs with respect to this special kind of distance-measure one can show easily (see Nordhoff (2003)) that this expectation coincides with the expectation which is defined in Definition 2.

In the case of a finite number of rectangles the optimisation problem which is connected to the Frechet expectation is treated in Chavent and Lechevallier (2002) for a special form of Hausdorff-distance.

4.3 The Doss expectation

Like the Aumann expectation and the version of Frechet expectation in Example 1 the Doss-Expectaion of Random Closed Rectangles coincides under the assumption of uniform boundedness with the expectation which is defined in Def. 2 (for details see Nordhoff (2003). Therefore the *Doss mean* of a finite number of closed rectangles which is defined in a canonical way like the Aumann mean (via constructing a uniformly bounded RCR) is concordant with the empirical mean of the corresponding RPRs.

4.4 The Vorob'ev expectation

The Vorob'ev expectation depends on the volume of the RCR and therefore does not conserve the shape as the following simple example shows.

Example 2. In the case of $p = 2$ let be X the (convex) RCR with

$$X(\omega) = \begin{cases} Q_1 := [1,2] \times [1,2], & \text{with probability } \frac{1}{2}, \\ Q_2 := [3,6] \times [1,8], & \text{with probability } \frac{1}{2}. \end{cases} \tag{25}$$

Then the Vorob'ev expectation of X is $\mathrm{E}_V[X] = Q_1 \cup Q_2$, in particular it is no rectangle and not convex.

Additionally in case of simple probability spaces , the 'approximation of the expected volume' often fails. So this kind of expectation is not suited to built a mean of random rectangles.

5 Discussion

We have considered the problem of building a mean of p-dimensional rectangles in a more general framework. With the aid of Random Closed sets the intuitive way of building the mean can be embedded as a special case. This fact legitimates the intuitive approach and in some cases it specifies a closed form for expectations of Random Closed Sets.
There are more imaginable approaches to define an expectation for RCSs and thus a mean of p-dimensional intervals, but it has to be analysed if the resulting mean has reasonable properties. The different kinds of expectations considered in this paper have shown that only those expectations of RCSs seem useful for RCRs which take the shape of the sets into account.
Furthermore, in this paper we always use the empirical mean as a standard estimator for the expectation of RCSs. But taking other statistical models (Stoyan and Mecke (1983)) into account one could use an estimated distribution of RCRs to build the mean of intervals.

References

CHAVENT, M. (2000): Criterion-Based Divisive Clustering for Symbolic Data. In: Bock, H.-H. with E. Diday (Eds.): *Analysis of Symbolic Data*. Springer, Berlin, 299–311.

CHAVENT, M. and LECHEVALLIER, Y. (2002): Dynamical Clustering of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff Distance. In: K. Jajuga, A. Sokołowski and H.H. Bock (Eds.): *Classification, Clustering, and Analysis*. Springer, Berlin, 203–210.

MATHERON, G. (1975): *Random Sets and Integral Geometry*. Wiley, New York.

MOLCHANOV, I. (1997): Statistical Problems for Random Sets. In: J. Goutsias (Ed.): *Random Sets: Theory and Applications*. Springer, Berlin, 27–45.

NORDHOFF, O. (2003): *Erwartungswerte zufälliger Quader*. Diplomarbeit, RWTH Aachen.

STOYAN, D. and MECKE, J. (1983): *Stochastische Geometrie*. Akademie-Verlag, Berlin.

STOYAN, D. and STOYAN, H. (1994): *Fractals, random shapes and point fields*. Wiley, New York.

Experimental Design for Variable Selection in Data Bases

Constanze Pumplün, Claus Weihs, and Andrea Preusser

University of Dortmund*
Department of Statistics
44221 Dortmund, Germany

Abstract. This paper analyses the influence of 13 stylized facts of the German economy on the West German business cycles from 1955 to 1994. The method used in this investigation is Statistical Experimental Design with orthogonal factors. We are looking for all existing Plackett-Burman designs realizable by coded observations of these data. The plans are then analysed by regression with forward selection and various classification methods to extract the relevant variables for separating upswing and downswing of the cycles. The results are compared with already existing studies on this topic.

1 Introduction

In the following, existing data are analysed using the method of statistical experimental design. The aim of experimental design is to estimate factor effects with the highest accuracy possible. Usually, an experimental design with fixed factor levels is taken and the response of the experiment is used to find factors of high influence with as few experiments as possible. Thus the optimal factors determining the response are found faster and with less expense than by carrying out all experiments with all possible factor level combinations. In order to detect the variables which do influence the up- and downswing phases of the economy, we use a special type of screening plans, namely Plackett-Burman plans. Contrary to the method of full factorial designs, which investigate main effects and all possible interactions, these plans are employed to find only the main effects in the model.

The original data used here are highly correlated. In order to eliminate these correlations, the data are coded by -1 and +1 only and then special observations are selected building Plackett-Burman plans. The main advantage of this method is that it selects the most important factors not disturbed by correlations in the data. By this procedure, on the one hand, the data are reduced by the discrete coding by -1 and +1 and on the other hand by choosing special observations only. In order to at least partially compensate this, we are analysing all existing Plackett-Burman plans with respect to the data

* This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475. We also thank Uwe Ligges and Karsten Luebke for their support.

and finally choose those variables which are, what we call, uniquely correlated to the up- and down phases of the economy. The following investigations are based on 13 stylized facts of the West German economy (cf. Heilemann and Münch (1999)) which have been selected by Heilemann and Münch to explain the German business cycle. There exists already a number of papers which analyse and interpret these data based on, e.g. classification methods like linear discriminant analysis and time series analysis (cp. Heilemann and Münch (1999), Weihs, Röhl and Theis (1999), Weihs and Garczarek (2002)).

In this paper in the first step, we code the data to -1 and +1 and in the second step we look for all Plackett-Burman plans in the coded data. All these plans are analysed by stepwise regression with forward selection, by unpruned classification trees, by trees consisting only of the tree stump and by stepwise linear discriminant analysis (cp. Röver (2003)). All this is based on an a priori classification of the response in the phases 'up', and 'down' in the years under investigation, based on Heilemann and Münch (1999). Finally, the variables which have turned out to be important are compared with the results of existing studies.

2 Data

The predictor data set consists of 13 variables which have been measured quarterly (157 quarters) in the years 1955/4 to 1994/4 (price index base is 1991) (cf. Heilemann and Münch (1999)). The variables (and their abbreviations) are real-gross-national product-gr (BSP91JW), real-private-consumption-gr (CP91JW), government-deficit-rate (DEFRATE), wage-and salary-earners-gr (EWAJW), net-export-rate (EXIMRATE), money-supply-M1-gr (GM1JW), real-investment-in-equipment-gr (IAU91JW), real-investment-in-construction-gr (IB91JW), unit-labour-cost-gr (LSTKJW), GNP-price-deflator-gr (PBSPJW), consumer-price-index-gr (PCPJW), nominal-short-term-interest-rate (ZINSK), real-long-term-interest-rate (ZINSLR). The letters 'gr' are an abbreviation of 'growth rates relative to last years corresponding quarter'.

3 Plackett-Burman designs

Heilemann and Münch (1999) distinguish 4 phases of the business cycle: 'upswing', 'upper turning point', 'downswing' and 'lower turning point'. Each quarter has been assigned one of these phases which we assume to be the correct one. Here only the phases 'up-', and 'downswing' are considered. Therefore, the phases 'upper turning point' and 'lower turning point' are split in the middle, i.e. if, e.g., the 'upper turning point' phase lasts for k quarters, $k \in \mathbb{N}, [k/2]$ quarters will be added to the 'upswing' phase and $k - [k/2]$ quarters will be added to the succeeding 'downswing' phase, where $[x]$ denotes the so called Gauß brackets, i.e. the largest integer less or equal

to x, $x \in \mathbb{N}$. An analogous convention holds for the 'lower turning point' phase. These two phases 'upswing' and 'downswing' are coded by 0 and 1, respectively. Note that two phase consideration is standard in business cycle analysis. Thus, it is the natural starting point for our studies. Extensions to 4 classes are planned.

Plackett-Burman plans only exist if the number of experiments n is a multiple of four and the number of variables is n-1 (cf. Plackett and Burman (1946), Weihs and Jessenberger (1999)). The Plackett-Burman plan for $n = 8$ is shown in Table 1.

Table 1. Plackett-Burman plan with 8 experiments.

	x1	x2	x3	x4	x5	x6	x7
1	-	-	-	-	-	-	-
2	-	-	+	-	+	+	+
3	+	-	-	+	-	+	+
4	+	+	-	-	+	-	+
5	+	+	+	-	-	+	-
6	-	+	+	+	-	-	+
7	+	-	+	+	+	-	-
8	-	+	-	+	+	+	-

The second row is called generating row, as it generates the rows 3–8 of the matrix by being shifted one position to the right at each step. Plackett-Burman plans are orthogonal arrays in the sense of (Hedayat et al. (1999)), they are of the form $OA(4\lambda, 4\lambda - 1, 2, 2)$, $\lambda \in \mathbb{N}$, ($\lambda = 2$ in Table 1), i.e. each factor has only two levels -1 and +1, the sum of each column is 0 and columns are pairwise orthogonal. If an 8th column consisting only of +1's is added to the matrix, one gets a unique Hadamard matrix of order 8 (cp. Hedayat et al. (1999)). Therefore it is necessary to code the existing data in +1 and -1, in order to look for Plackett-Burman designs. For each variable, all values less than its median are taken as -1 and all values greater than or equal to its median are taken as +1. As there are 13 variables, one looks for Plackett-Burman plans with $n = 8$ or $n = 12$ in the coded data. 113 different plans were found for $n = 8$ and none for $n = 12$.

The algorithm for finding these plans is first to look for all rows which contain at least seven times the number -1. The corresponding columns are then searched for the generating row. After this has been found, the search continues for the generating row shifted one position to the right, etc. This process has to be carried out for all possible permutations of the original seven columns. A much faster algorithm has been suggested by S.Haustein (private communication), where one looks for the base row $u0 = (- - - - - - - -)$ and then searches for a row v in the corresponding columns with Hamming distance 4 to $u0$. After this has been found, one looks for a row $v1$ with Hamming distance 4 to $u0$ and v. This process is continued until eight rows have been found which are equidistant with Hamming distance 4. These eight

rows form a Plackett-Burman plan for $n = 8$, because the Plackett-Burman plan for $n = 8$ is an orthogonal array of the form $OA(8, 8-1, 2, 2)$ and this class has only one isomorphism class. Here two arrays are said to be isomorphic (cf. Hedayat et al. (1999)), if one can be obtained from the other by permutations of rows, columns or factor levels.

In the following investigations, a linear screening model is used, $y = X\beta + \epsilon$, where $X = (\mathbf{1}, A)$ is an $(n \times n)$ matrix with $\mathbf{1} = (1, 1, 1, ...)^t$ and A the Plackett-Burman matrix. β is the vector of unknown coefficients, y the result vector with the coded business cycle phases and ε the error vector.

4 Results

4.1 Stepwise regression by forward selection

113 different Plackett-Burman plans were found by the method described in 3. When evaluating these plans by stepwise regression with forward selection with respect to y (cp. Weihs and Jessenberger (1999)), we used the F-test at level 0.2. Figures 1, and 2 show the absolute and the relative frequency of the selected variables (dark bars). The light bars show how often each variable appears in all 113 Plackett-Burman plans. Figure 1 thus shows that each variable is at least once in a plan (light bars). The variables which turn out to be most important by this method are 'DEFRATE', 'EXIMRATE', 'LSTKJW', 'IAU91JW' and 'ZINSK'(cp. Figure 2). If one uses the F-test with level 0.05 one gets the same variables except 'EXIMRATE'. It is also interesting that in almost half of all cases none of the variables turns out to be important. Furthermore it strikes that for all variables the dark bars are rather small, compared to the light ones. That means that although a variable appears often in the plans it is chosen only a few times as important concerning the up- and down of the economy.

4.2 Classification methods

In the 113 plans, variables are selected also by different classification methods, i.e.unpruned classification trees (TreeAllNodes), classification trees with only the tree stump (TreeStump) and stepwise linear discriminant analysis (cp. Röver (2003)). Figures 3, and 4 again show the absolute and the relative frequency of selected variables by the different methods. The number in brackets following the variable name indicates how often the variable appears in a Plackett-Burman plan. Classification by unpruned trees yields as important variables 'BSP91JW', 'CP91JW', 'DEFRATE' and 'EXIMRATE'. Using only the tree stump yields the same variables without 'CP91JW' as important. This is the same result one gets by stepwise linear discriminant analysis. On the whole, these three classification methods yield similar results but on different levels.

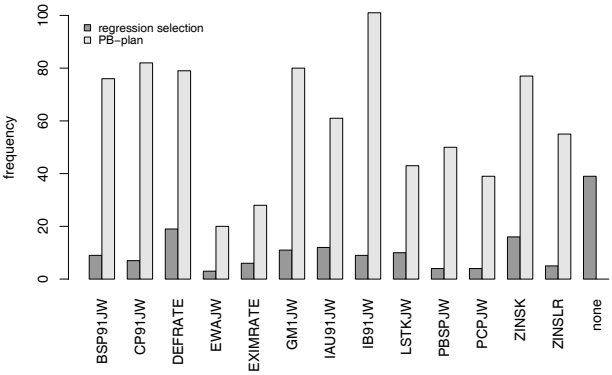


Fig. 1. Absolute frequency of variable selected by stepwise regression with forward selection.

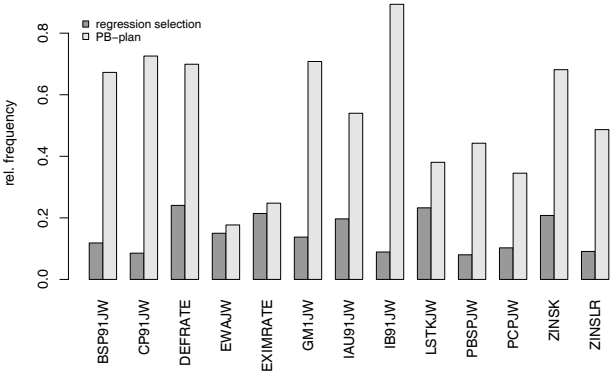


Fig. 2. Relative frequency of variable selected by stepwise regression with forward selection.

For all used classification methods as well as for stepwise regression with forward selection it is important to know how the rows which build the Plackett-Burman plans are distributed. This is illustrated in Figure 5 which shows how often each row is contained in a plan. Note that the outstanding row number 72 refers to the 4th quarter of 1972 and row number 145 to the 1st quarter of 1991. These years are special years from an economic point of view, as in 1972 the German economy suffered from the oil price shock. The German unification influences the post 1990 data, an effect shown in the first quarter of 1991.

4.3 Variable assessment

If one wants to decide which of the above variables plays a dominant role with respect to the business cycle, it is important to assess their correlation

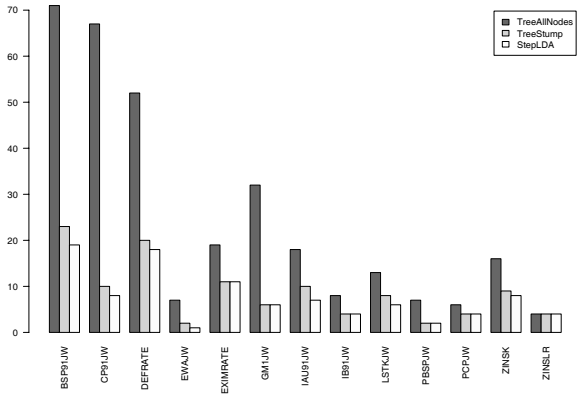


Fig. 3. Absolute frequency of variables selected in Plackett-Burman design.

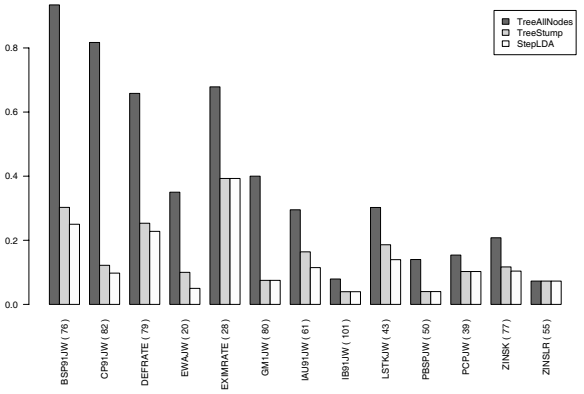


Fig. 4. Relative frequency of variables selected in Plackett-Burman design.

in all those Plackett-Burman plans where the corresponding variable was included. It turns out (see Table 2) that unit labour costs('LSTKJW') is clearly positively correlated to y (84% of all cases) and the government deficit ('DEFRATE') can still be considered as positive correlated (78% of all cases), taking into account a possible error margin. No variable is clearly negatively correlated to y. Hence, one may finally consider those variables as important which on the one hand are chosen most often, both by regression and by classification, and which on the other hand possess a distinct positive or negative correlation to y. Using this decision criterion, one gets 'unit labour costs' ('LSTKJW') and 'government deficit' ('DEFRATE') as variables which clearly determine the West German business cycles.

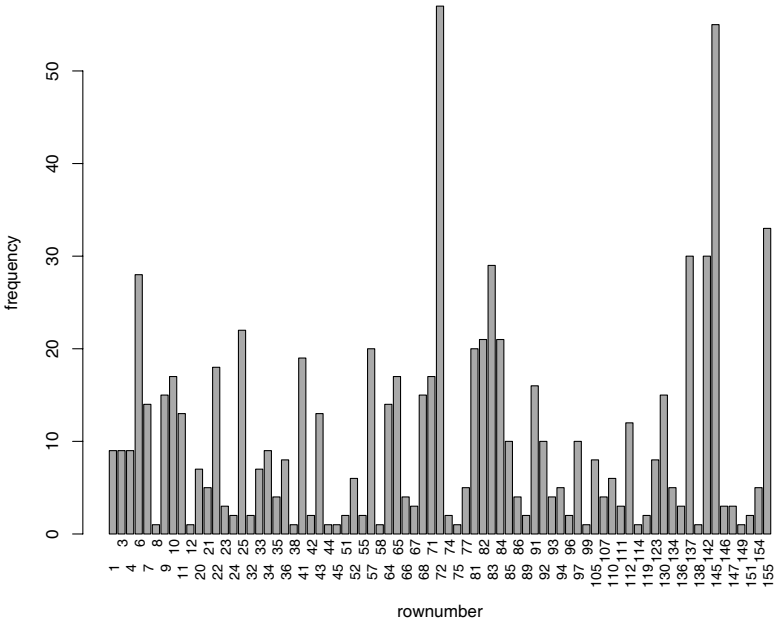


Fig. 5. Absolute frequency of rows in all Plackett-Burman plans.

In previous studies of this topic (cp. e.g. Weihs and Garczarek (2002), Weihs et al. (1999)) the variables most influential for the West German business cycle in the 4 phase case were 'wage and salary earners' ('EWAJW') and 'unit labour costs' ('LSTKJW'). Moreover, if one compares the above method to stepwise regression by forward selection on the whole data set,

Table 2. Correlation with respect to y.

	Positive	**Negative**	**No Cor.**	**% positive**
BSP91JW	18	41	17	24
CP91JW	32	24	26	39
DEFRATE	62	3	14	78
EWAJW	13	2	5	65
EXIMRATE	19	5	4	68
GM1JW	8	53	19	10
IAU91JW	5	40	16	8
IB91JW	21	51	29	21
LSTKJW	36	1	6	84
PBSPJW	29	6	15	58
PCPJW	24	6	9	49
ZINSK	50	14	13	65
ZINSLR	25	18	12	45

again taking level 0.2 in the F-test, the model 'LSTKJW' + 'IAU91JW' + 'DEFRATE' + 'ZINSK' + 'CP91JW' + 'BSP91JW' is chosen. This strongly indicates the importance of 'LSTKJW' and 'DEFRATE'. Also stepwise linear discriminant analysis, classification by unpruned trees and classification trees using only the tree stump were applied on the whole data set. The application of unpruned classification trees shows 'IAU91JW' to be the most important variable, as does classification trees using only the tree stump. Stepwise linear discriminant analysis shows that besides 'IAU91JW', also two other variables are important, 'LSTKJW' and 'PCPJW'.

5 Conclusion

'Unit labour costs' ('LSTKJW') has been detected as an important variable by this method as well as by previous methods (cp. 4.3). This strongly indicates that this variable has a great influence on the West German business cycle. The question why the 'government deficit' ('DEFRATE') turns out to be important here, but does not so in previous studies, requires a thorough analysis of the influence of the methods applied here on the results. The advantage of using Plackett-Burman plans lies in the clean and easy selection of variables in determining the important variables. This is only a first step in this direction. Right now, we are investigating only the correlations of those variables with the business cycle, which have turned out to be important in the above described investigations. A next step could be to investigate a similar procedure with full factorial designs or fractional factorial designs. These plans also respect orthogonality, but in addition permit interactions between the factors.

References

HEDAYAT, A., SLOANE, N. and STUFKEN, J. (1999): *Orthogonal Arrays.* Springer Verlag, New York, Berlin, Heidelberg.

HEILEMANN, U. and MÜNCH, J. (1999): Classification of West German Business Cycles. *Technical Report 11, SFB 475 Universität Dortmund.*

PLACKETT, R. L. and BURMAN, J. P. (1946): The design of optimum multifactorial experiments. *Biometrica, 33, 305–325.*

RÖVER, C. (2003): Musikinstrumentenerkennung mit Hilfe der Hough - Transformation. *Diplomarbeit, Fachbereich Statistik, Universität Dortmund.*

WEIHS, C. and JESSENBERGER, J. (1999): *Statistische Methoden zur Qualitätssicherung und -optimierung in der Industrie.* Wiley-Vch, Weinheim.

WEIHS, C., RÖHL, M.C. and THEIS, W. (1999): Multivariate Classification of Business Phases. *Technical Report 26, SFB 475, Universität Dortmund.*

WEIHS, C. and GARCZAREK, U. (2002): Stability of multivariate representation of business cycles over time. *Technical Report 20, SFB 475, Universität Dortmund.*

KMC/EDAM: A New Approach for the Visualization of K-Means Clustering Results

Nils Raabe, Karsten Luebke, and Claus Weihs

University of Dortmund*
Department of Statistics
44221 Dortmund, Germany

Abstract. In this work we introduce a method for classification and visualization. In contrast to simultaneous methods like e.g. Kohonen SOM this new approach, called KMC/EDAM, runs through two stages. In the first stage the data is clustered by classical methods like K-means clustering. In the second stage the centroids of the obtained clusters are visualized in a fixed target space which is directly comparable to that of SOM.

1 Introduction

In many applications a classification of the examined objects in both inter-heterogeneous and intra-homogeneous groups (clusters) is desired. Many methods have been developed to solve this problem and are subsumed under the term classification-methods as well as clustering-methods.

In the context of clustered objects another problem often occurs. This problem consists of the graphical representation - called visualization - of the objects resp. classes which are often represented by high-dimensional data vectors in a space of lower dimension. The requirement for such representations is topology preservation, i.e. objects which are comparatively close in the original space should also be close together in the representation space and, corresponding by, pairs of distant objects should have high distances in the visualization.

One method, which can be interpreted both as a visualization and a classification method, is the so called Kohonen Self-Organizing-Map (SOM) (Kohonen (1990)). SOM performs classification and visualization simultaneously. Many alternatives to SOM have been proposed in the past. One example is another simultaneous method suggested by Bock (1997). Bezdek and Pal (1995) compare the methods principal component analysis (PCA) and the Sammon algorithm to SOM concerning topology preservation. They try to avoid the problem of different solution spaces - with SOM in contrast to the latter methods only a subset of the objects is visualized - by assigning to each object an image in the neighborhood of the nearest visualized object.

* This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

Since this is done by randomly jittering it is questionable if the corresponding results can still be seen as the results generated by SOM. Hence the results of Bezdek and Pal – PCA and Sammon are superior to SOM – are to be interpreted cautiously.

Being aware of the aforementioned comparability-problems we introduce a new approach of carrying out classification and visualization one after the other. This approach consists of a combination of classical classification methods (mainly K-Means-Clustering, KMC) and a new approach for the visualization of the corresponding centroids. This approach is called Eight-Directions-Arranged-Map (EDAM) and has a fixed representation space. This solution space can be chosen in SOM as well. Under these conditions criteria for classification and topology preservation can be defined and compared between the two methods.

This paper starts with a description of the methods in section 2. Then section 3 gives a view on a few examples. The paper concludes in a summary given in section 4.

2 Methods

2.1 Preliminaries

All following methods refer to a data matrix $X \in \mathbb{R}^{n \times k}$. Its rows $x_{1\cdot}, ..., x_{n\cdot} \in \mathbb{R}^k$ represent the data vectors of n corresponding objects and its columns $x_{\cdot 1}, ..., x_{\cdot k} \in \mathbb{R}^n$ represent the measurement vectors of k corresponding variables. Distances between two data vectors $x_{i\cdot}$ and $x_{j\cdot}$ are denoted by $d(x_{i\cdot}, x_{j\cdot})$. We use the ordinary euclidean distance in this paper.

A classification of X is a set of c clusters, where each object belongs to exactly one cluster. A classification is denoted by a vector $\kappa \in \{1, ..., c\}^n$, where the ith element κ_i of κ gives the cluster-number of the ith object. A common representative of cluster i is the so called centroid $\mu_i \in \mathbb{R}^k$, which is defined as:

$$\begin{gathered} \mu_i = (\mu_{i1}, ..., \mu_{ik})' \quad \text{with} \quad \mu_{ih} = \frac{1}{n_i} \sum_{j:\kappa_j = i} x_{jh}, \quad h = 1, ..., k, \\ n_i = \#\{j : \kappa_j = i\}, \quad i = 1, ..., c. \end{gathered} \tag{1}$$

All centroids are compiled in the centroid matrix $M = (\mu_{ij})_{\substack{1 \le i \le c \\ 1 \le j \le k}}$.

A visualization of X is a function $f : \{x_{1\cdot}, ...x_{n\cdot}\} \to \mathcal{Z} \subseteq \mathbb{R}^{n \times m}, m < k$, which assigns an image $z^i = (z^i_1, .., z^i_m)' = f(x_{i\cdot})$ to each row of X. $\mathcal{Z}$ is called the image-space.

With $Z = (z^i_j)_{\substack{1 \le i \le n \\ 1 \le j \le m}}$ the visualization f may be written as $f(X) = Z$. In the following we only consider the case of $m = 2$.

2.2 Basic idea

Our approach to visualize high-dimensional data in a plane is based on the idea of considering the plane as a topographical map. When the images are visualized as the vertices of a rectangular grid, each object has eight direct neighbors, one in each direction of the compass (by taking NE, SE, SW and NW into account, compare figure 1). We try to obtain topology preservation by re-ordering the objects on each of these eight directions corresponding to the distances of their data vectors in the original space $\mathbb{R}^k$. Considering the example of the vector pointing from z^{20} to west in figure 1 this means, that with $x_{i\cdot} = f^{-1}(z^i)$ after re-ordering, i.e. interchanging the values of $x_{21\cdot}$ to $x_{24\cdot}$, the relation $d(x_{20\cdot}, x_{21\cdot}) \leq d(x_{20\cdot}, x_{22\cdot}) \leq d(x_{20\cdot}, x_{23\cdot}) \leq d(x_{20\cdot}, x_{24\cdot})$ holds.

The method EDAM visualizes by repeating this "star-shaped" re-ordering step successively for all objects up to either convergence or to another stopping criterion. The following subsection gives a formal definition of the method.

2.3 KMC/EDAM

The classification of X into a set of $c < n$ clusters, c given, by the method KMC/EDAM is performed by a combination of a K-Means-algorithm and a hierarchical method. First $g > c$ clusters are constructed by applying the K-Means-algorithm suggested by Forgy (see Anderberg (1973)). Then the agglomerative hierarchical Centroid-method (see Kaufmann and Pape (1996)) is applied to these clusters. After $(g - c)$ steps of this method the final classification κ of the n objects into c clusters is obtained.

In the next stage of KMC/EDAM the centroids $\{\mu_1, ..., \mu_c\}$ of κ are visualized. Therefore first the image space is fixed to the points of intersections of b_1 vertical and b_2 horizontal lines of a two-dimensional, equally spaced grid, with $c = b_1 \cdot b_2$. By labelling the images by their integer Euclidean coordinates and enumerating them from the lower left corner by rows the image-space can be written as:

$$\mathcal{Z} = \left\{ z^1, \cdots, z^c \right\} \quad \text{with} \quad z^i = \begin{pmatrix} z_1^i \\ z_2^i \end{pmatrix} = \begin{pmatrix} i - \lfloor \frac{i-1}{b_1} \rfloor \cdot b_1 \\ \lceil \frac{i}{b_1} \rceil \end{pmatrix}. \tag{2}$$

The problem of visualizing the centroids in $\mathcal{Z}$ by a visualization f is to find a permutation π of $\{1, ..., c\}$, such that $f(\mu_{\pi(i)}) = z^i, i = 1, ..., c$, preserves topology as well as possible (concerning to a predefined criterion).

The main idea of our method is to consider[1] each centroid $\mu_{\pi_{t-1}(i)}$ as a "reference point" for the centroids whose images are lying on the vectors

[1] The consideration of one centroid defines one step denoted by index t; the index i defining the actual centroid computes to $i = t - \lfloor \frac{t-1}{c} \rfloor \cdot c$, i.e. each time t exceeds a multiple of c, i is switched back to 1.

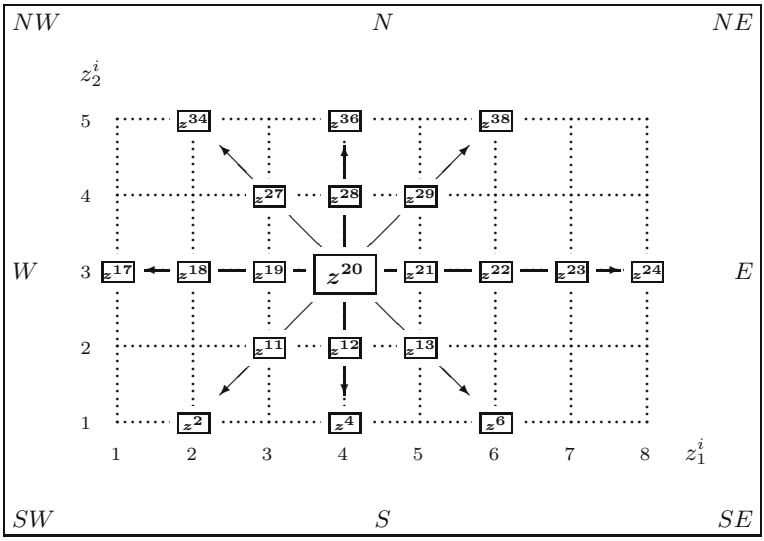


Fig. 1. $\mathcal{Z}$ as topographical map

pointing from z^i to each direction $D \in \{N, NE, ..., NW\}$, where π_0 is a randomly chosen initial permutation. First, for each direction D, the indices $j_q^D, q = 1, ..., n_D$ of these images are determined. Table 1 gives an overview of how these indices are calculated for all directions.

Table 1. Calculation of indices

D	j_q^D	n_D	D	j_q^D	n_D
N	$i + qb_1$	$b_2 - z_2^i$	NE	$i + qb_1 + q$	$min(n_N, n_E)$
E	$i + q$	$b_1 - z_1^i$	SE	$i - qb_1 + q$	$min(n_S, n_E)$
S	$i - qb_1$	$z_2^i - 1$	SW	$i - qb_1 - q$	$min(n_S, n_W)$
W	$i - q$	$z_1^i - 1$	NW	$i + qb_1 - q$	$min(n_N, n_W)$

Let now φ_D be the permutation of $\{\pi_{t-1}(j_1^D), ..., \pi_{t-1}(j_{n_D}^D)\}$ so that

$$d(\mu_{\pi_{t-1}(i)}, \mu_{\varphi_D[\pi_{t-1}(j_1^D)]}) \leq d(\mu_{\pi_{t-1}(i)}, \mu_{\varphi_D[\pi_{t-1}(j_2^D)]}) \leq ... \leq d(\mu_{\pi_{t-1}(i)}, \mu_{\varphi_D[\pi_{t-1}(j_{n_D}^D)]})$$

for each direction D. Now, set $\pi_t := \pi_{t-1}$. Next, the following substeps are repeated for all directions D:

1. $\pi_t^D := \pi_t$

2. $\{\pi_t^D(j_1^D), ..., \pi_t^D(j_{n_D}^D)\} := \{\varphi_D([\pi_t^D(j_1^D)], ..., \varphi_D([\pi_t^D(j_{n_D}^D)]\}$
3. $\pi_t := \begin{cases} \pi_t^D & , \quad if \quad S(\pi_t^D) < S(\pi_t) \\ \pi_t & , \quad else \end{cases}$.

The function S is a predefined criterion for visualizations with lower values indicating better visualizations. Repeating the described procedure for all centroids – i.e. a set of c steps – builds one iteration. In our investigations we choose S as the STRESS known from MDS (see Hamerle and Pape (1996, p. 769)).

Each time, when no more improvement can be obtained after a complete iteration (or alternatively if a given maximum number of iterations is reached), the area, in which re-ordering is possible is decreased by changing the values of n_D in table 1 to $min(n_D, max[b_1, b_2] - r)$, where r runs successively from 1 to $max[b_1, b_2] - 2$. A set of iterations with the same value of r is called iteration cycle.

The final visualization result $f(\mu_{\pi(i)}) = z^i, i = 1, ..., c$, of KMC/EDAM is obtained by setting $\pi := \pi_t$ where t is the number of the last step.

3 Examples

First the introduced method is applied to the synthetic Chainlink data, which consist of two three-dimensional interlocking ring-shaped classes as seen in figure 2. In our example each class contains 1000 data points.

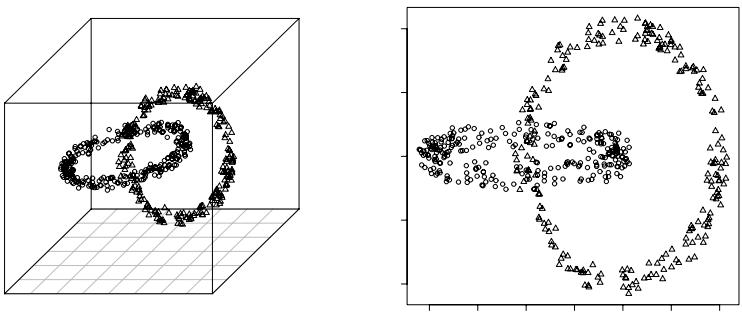


Fig. 2. The Chainlink data and their MDS visualization

On the right side of figure 2 the two-dimensional result of the method MDS for this example is depicted. The STRESS of this result is 0.246. Note that in the original space the two classes have exactly the same relation to each other, i.e. they have the same shape, the rings have the same radius

and the center of each ring lies on the other one. But looking at the MDS visualization one gets the impression that there are differences between the shapes of the classes.

For the computation of KMC/EDAM for the Chainlink data the following settings were used: g=750, c=500,$b_1 = 20$, $b_2 = 25$, maximum number of iterations per cycle: 10. The result is shown in U-Matrix-representation on the left side of figure 3. The U-matrix is a well-known tool developed for the representation of Self-Organizing Maps (compare Ultsch (2003)). Since the image space of KMC/EDAM is restricted to a rectangular grid the U-Matrix can easily be applied to the results of this method as well. For comparison purposes the right side of figure 3 shows the U-matrix of a SOM of the same size applied to the same data. For the computation of the SOM the package `som` available for the statistical software `R` (R Development Core Team (2004)) with its default settings was used. The size of the symbols in both pictures corresponds to the number of objects assigned to each cluster.

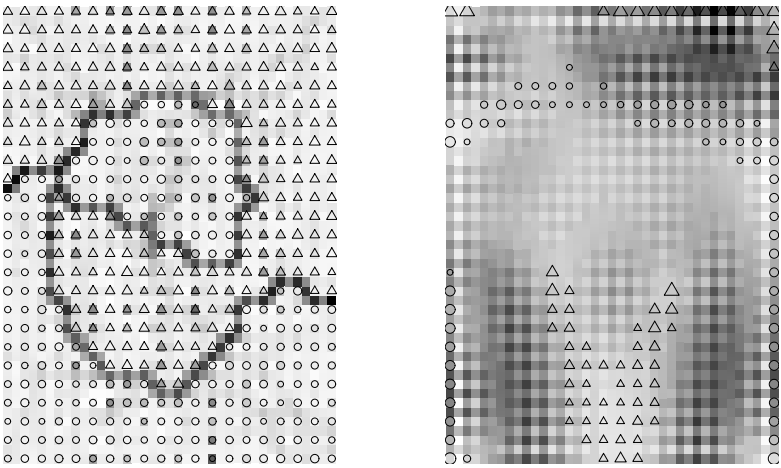


Fig. 3. KMC/EDAM and SOM visualization of the Chainlink data

The STRESS of the KMC/EDAM solution is 0.209, that of SOM is 0.252, so KMC/EDAM seems to be better. Beyond this superiority of KMC/EDAM to the MDS and SOM the result of KMC/EDAM gives an evidently better mirror of the fact, that the Chainlink classes are equally placed relatively to each other. This is not the case for SOM, since the class depicted by circles seems to surround parts of the classes depicted by triangles. At first glance

the separation of the classes seems better for MDS resp. SOM, since the latter leave gaps between the classes. But in the U-Matrix of the KMC/EDAM result a dark line is visible which corresponds to relatively high distances between objects along the line. This line runs between the two classes like a boundary. The brightness of the rest of the map is well-adjusted, which suggests that the topology within the classes is well preserved. Another advantage of the KMC/EDAM result compared to that of SOM is, that it maintains the connection of the classes, i.e. there are now exclaves. In the SOM result there are apparently a few objects of the "triangle class" separated from the rest by the "circle class".

The next example we consider is the well-known iris data set introduced by Fisher (1936), which contain setal and petal lengths and widths of three species of iris for 150 flowers. Figure 4 shows a plot of the MDS result and the U-matrices of KMC/EDAM and SOM results for this example. The settings of KMC/EDAM were: g=50, c=35,$b_1 = 5$, $b_2 = 7$, maximum number of iterations per cycle: 10.

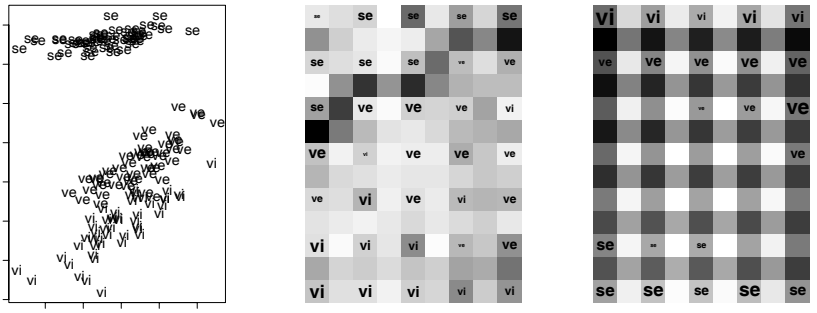


Fig. 4. MDS, KMC/EDAM and SOM visualizations of the iris data

The STRESS of KMC/EDAM is 0.351 in this case, while that of SOM is 0.252. MDS performs even better with a STRESS of 0.04. Similarly to the previous example SOM leaves a gap between the well-separated classes versicolor and setosa. Again this separation is visible as a dark line in the U-matrix of the KMC/EDAM result. The separation of the classes virginica and versicolor seems slightly better in the SOM result, since one can notice the darkest squares between these classes than at other regions of the map.

4 Conclusion

With KMC/EDAM a method is introduced which allows to visualize the results of classical clustering methods. Because of its specially chosen target space the results of this method are directly comparable to those of SOM. The method is applied to two popular examples, the artificial Chainlink data and Fisher's iris data. In the critical Chainlink example KMC/EDAM leads to better results than MDS and SOM.

In the iris example KMC/EDAM has the highest STRESS. But the relative positions of classes are the same with KMC/EDAM. Furthermore the lacking separation – which is probably the reason for the higher STRESS – becomes visible as well by representing the result in an U-matrix.

Modifications for the improvement of EDAM are conceivable. Such modifications may concern the optimization of the initial ordering of the centroids. On the other hand a method like Simulated Annealing could be integrated into the algorithm to avoid local optima. First attempts in this direction led to promising results.

References

ANDERBERG, M.R. (1973): *Cluster Analysis for Applications*. Academic Press Inc., New York.

BEZDEK, J.C., PAL, N.R. (1995): An index of topological preservation for feature extraction. *Pattern Recognition, 28/3, 381–391.*

BOCK, H.H. (1997): Simultaneous visualization and clustering methods as an alternative to Kohonen maps. In: G. Della Riccia, R. Kruse and H.-J. Lenz(Eds.): *Learnings, networks and statistics*. CISM Courses and Lectures, 382, Springer, New York, 67–85.

FISHER, R.A. (1936): The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7/2, 179–188.*

HAMERLE, A., PAPE, H. (1996): Grundlagen der mehrdimensionalen Skalierung. In: L. Fahrmeir, A. Hamerle, and G. Tutz (Eds.): *Multivariate statistische Verfahren*. De Gruyter, Berlin 765–792.

KAUFMANN, H., PAPE, H. (1996): Clusteranalyse. In: L. Fahrmeir, A. Hamerle, and G. Tutz (Eds.): *Multivariate statistische Verfahren*. De Gruyter, Berlin 437–536.

KOHONEN, T. (1990): The Self-Organizing Map. *Proceedings of the IEEE, 78/9, 1464–1480.*

R DEVELOPMENT CORE TEAM (2004): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna,Austria. `http://www.R-project.org`

ULTSCH, A. (2003): Maps for the visualization of high-dimensional data spaces. *Proc. Workshop on Self organizing Maps, 225-230.*

Clustering of Variables with Missing Data: Application to Preference Studies

Karin Sahmer[1], Evelyne Vigneau[1], El Mostafa Qannari[1], and Joachim Kunert[2]

[1] Laboratoire de sensométrie et de chimiométrie,
ENITIAA / INRA, rue de la Géraudière, BP 82 225,
F-44322 Nantes Cedex 03, France
e-mail: sahmer@enitiaa-nantes.fr
[2] Fachbereich Statistik,
Universität Dortmund, D-44221 Dortmund, Germany

Abstract. Clustering of variables around latent components is a means of organizing multivariate data into meaningful subgroups. We extend the approach to situations with missing data. A straightforward method is to replace the missing values by some estimates and cluster the completed data set. This basic imputation method is improved by more sophisticated procedures which update the imputations within each group after an initial clustering of the variables. We compare the performance of the different imputation methods with the help of a simulation study.

1 Introduction

The problem of missing values occurs very often in practice. In this paper we propose methods to deal with the problem of missing values when we want to cluster variables. A method of clustering variables around latent components (CAVALC) was proposed by Vigneau and Qannari (2003). This procedure bears some similarity to VARCLUS which is implemented in SAS (SAS/STAT (1990)). However, it is based on a simple algorithm and may be extended in various ways as discussed by Vigneau and Qannari (2002). This method is briefly described in section 2.

An important application of this technique is given by the clustering of consumers who give their scores of preference to different products. Usually these preference scores are analysed by means of a preference mapping technique which mainly consists in performing a principal components analysis on the data set whose rows are the products and columns are the scores given by the various consumers (Greenhoff and MacFie (1994)).

However, it is not always possible to present all the products to each consumer, especially when we have saturating products such as beers. Therefore each consumer evaluates a subset of the products. The resulting data set is incomplete.

In section 3 we propose some imputation methods for the clustering in this situation. The real data set under study is briefly described in section

4. We compare these imputation methods on the basis of a simulation study (section 5).

2 Clustering of variables around latent components

We consider p variables $x_1, x_2, \ldots, x_p$ measured on a sample of n observations. The procedure of clustering discussed herein consists in representing each cluster of variables by a latent component.
More precisely, the strategy consists in simultaneously determining K (supposed to be fixed) clusters of variables and K latent components $c_1, c_2, \ldots, c_K$ such that

$$S = \sum_k \sum_j \delta_{jk} \mathrm{cov}(x_j, c_k)$$

is maximized under the constraint $c_k' c_k = 1$. In this criterion S, the parameter δ_{jk} is equal to 1 if variable x_j belongs to cluster k and 0, otherwise, and cov stands for the covariance.

S is maximized by a partitioning algorithm in the course of which latent components and group memberships are iteratively updated. The initialization of the partitioning algorithm is based on an agglomerative hierarchical procedure (Vigneau and Qannari (2003)).

For a given cluster G_k, the latent variable c_k is collinear with the mean $\bar{x}_k$ of the variables belonging to cluster G_k.

3 Imputation methods

3.1 Direct imputation methods

An intuitive method for dealing with missing values consists in replacing each missing value by the mean of the observed values for the variable under consideration. We will refer to this method as the vertical imputation.

In the special case of preference data it is also possible to replace each missing value by the mean of the observed values for the observation (product) under consideration. In that way, the mean score observed on the whole panel is attributed to all missing data for a product. We will refer to this imputation method as the horizontal imputation.

After replacing the missing values by these estimates, we cluster the completed data set.

3.2 Imputation within each cluster

We can improve the results by updating the imputations after clustering. Each missing value is replaced by an estimate that is based on the values observed on the variables of the same group.

In preference studies the consumers use the given scale in different manners. More precisely, consumers may score at different levels of the scale, as it will be illustrated in section 4, or may differ in the range of scoring. To make the variables comparable, it is necessary to firstly standardize them as follows:

$$\tilde{x}_{ij} = \frac{x_{ij} - \hat{\mu}_j}{\hat{\sigma}_j}.$$

The needed estimations $\hat{\mu}_j$ and $\hat{\sigma}_j$ of the mean and standard-deviation are calculated using the observed values of the variable. In each group G_k, the mean m_{ik} of each observation i $(i = 1, \ldots, n)$ is calculated using the standardized observed values.

If the value of observation i for variable j in group G_k is missing, $\hat{z}_{ij} = m_{ik}$ is used as an estimate. Finally, the imputed data are 'destandardized' as follows: $\hat{x}_{ij} = \hat{z}_{ij}\hat{\sigma}_j + \hat{\mu}_j$. The calculation of the latent components is finally updated using the observed values and the new imputations $\hat{x}_{ij}$ of the unobserved values.

3.3 Method based on a cross-partition

This method is based on the two different initial imputations outlied in section 3.1 (horizontal and vertical imputations). Generally, the clustering of the two completed data sets provides two different partitions into K groups. The analysis of their cross-partition may improve the grouping. We proceed as follows (Sahmer (2003)):

We calculate the cross-partition which consists of K^2 groups, called 'stable groups' ('groupements stables', Lebart et al. (2000)). As an illustration, consider the clustering of six variables which we denote 1, 2, 3, 4, 5, 6. Suppose that the first clustering leads to the partition (1, 2, 3) and (4, 5, 6) and the second clustering leads to the partition (1, 2, 6) and (3, 4, 5). The four stable groups are obtained by considering the intersections of each group from the first partition with each group of the second partition. This leads to the partition (1, 2), (3), (6) and (4, 5). The imputations are then updated in the stable groups according to the procedure described in section 3.2. However, it should be noted that the stable groups may contain only very few variables. So it is possible that in group G_k there is no data for an observation i. In this case, the mean of the observed values of the variable (vertical imputation) is used.

The K stable groups with the largest numbers of variables are determined. In the case of ties, we may randomly select which of the tied groups to retain.

The latent components $c_k = \bar{x}_k / \sqrt{\bar{x}_k' \bar{x}_k}$ of these K largest stable groups are calculated. The covariances of all the variables not belonging to these K groups with each of the K latent components are determined. Each variable is assigned to a group, considering its largest covariance with the group latent components.

Finally, in each of the K groups, the imputations and the latent component are again updated as described in section 3.2.

4 Illustration: data set 'jam'

The methods are compared on the basis of a real data set from a preference study. These data were collected by students at ENITIAA (Nantes, France) during a training period. These students manufactured seven varieties of jam with different percentages of apple and pear and added vanilla or cinnamon flavour. Hedonic ratings were given by a panel of consumers. In addition, the sensory properties of the jams were evaluated. Therefore, we have three kinds of variables concerning the jams: the compositional data, sensory variables and the hedonic scores.

In this paper we focus on the analysis of the hedonic data. 56 consumers gave their scores of preference. They scored each product on a non structured 10 cm scale according to their liking.

Table 1. Preference data of two consumers

	Observed data		Standardized data	
	Consumer 41	Consumer 53	Consumer 41	Consumer 53
Jam 1	5.0	1.2	-0.9	-0.2
Jam 2	8.3	3.6	1.6	1.9
Jam 3	6.0	0.4	-0.1	-0.9
Jam 4	4.0	0	-1.7	-1.2
Jam 5	7.0	2.4	0.6	0.9
Jam 6	7.0	1.1	0.6	-0.3
Jam 7	6.0	1.2	-0.1	-0.2
Average	6.2	1.4	0	0
Standard deviation	1.3	1.1	1	1

As an illustration, the left side of Table 1 shows the scores given by consumers 41 and 53 to the seven varieties of jam. This example makes it possible to outline the different use of the scale by the consumers. Obviously, consumer 41 gives higher scores than consumer 53. Nevertheless, their patterns of liking are very similar as reflected by the standardized data.

We used this data set together with simulated data in order to compare the different imputation methods.

5 Simulation study

5.1 Jam data set

We clustered the complete data set and determined two groups of consumers. In a subsequent stage, we simulated one to four missing values per consumer. We clustered the incomplete data sets with each of the imputation methods and compared the results to the clusters obtained from the complete data set. For each number of missing values, we repeated this procedure 100 times.

5.2 Simulated data

We simulated data that were designed to reflect the different ways consumers use the given scale. For each group k, we simulated an average score m_{ik} for each product i. The score of each consumer j is based on this average. More precisely, the score of a consumer j in cluster k is simulated by multiplying m_{ik} by a scaling factor d_{jk} which reflects the different ranges of scale used by the consumers. Thereafter, a translation t_{jk} and a normally distributed noise ϵ_{ijk} are added. The score of consumer j in group k for product i is given by (Callier (1996)):

$$x_{ijk} = t_{jk} + \bar{m}_k + d_{jk}(m_{ik} - \bar{m}_k) + d_{jk}\epsilon_{ijk}, \qquad \epsilon_{ijk} \;\; i.i.d. \sim N(0, \sigma^2).$$

We simulated data sets with two and three groups. The standard deviation of the noise was set to $\sigma = 0.5$ and to $\sigma = 2$. For each combination of these two parameters we simulated 100 data sets with eight products and 56 consumers each. Then, in each data set we simulated one to five missing values per consumer. We clustered the incomplete data sets with the different imputation methods. We compared the resulting groups to the simulated groups.

5.3 Criterion for comparison

The comparison of the methods was based on different criteria. We show herein the results regarding the criterion ACALC (Average of the Correlations of the Associated Latent Components). It was proposed by Callier (1996) under the acronym of MCGA (Moyenne des Corrélations entre Groupes Associés). This criterion indicates the extent to which the latent components of the incomplete data are related to those of the complete data. However, the labelling of the groups is arbitrary. Group 1 of the complete data set does not necessarily correspond to group 1 of the incomplete data set, etc. Consequently, we first have to determine the association between the groups and the latent components in the two partitions (from the incomplete and complete data sets). In practice, we consider each possible combination and calculate the average correlation between associated latent components

for each of them. The criterion ACALC corresponds to the maximum average correlation.

The calculation is illustrated for the case of $K = 2$ as follows. The average correlations of the possible combinations are given by

$$mcor_1 = \frac{1}{2}(cor(c_1, \tilde{c}_1) + cor(c_2, \tilde{c}_2))$$

and

$$mcor_2 = \frac{1}{2}(cor(c_1, \tilde{c}_2) + cor(c_2, \tilde{c}_1)),$$

where c_k are the latent components of the complete data and $\tilde{c}_k$ the latent components of the incomplete data.

We define

$$\text{ACALC} = max(mcor_1, mcor_2).$$

5.4 Results

It was observed that the two direct-imputation methods were almost always improved by an update of the imputations after clustering.

In the following, we consider the vertical imputation with an update of the imputations, the horizontal imputation with an update of the imputations and the method based on a cross-partition. Their results are shown in Figure 1. For the simulated data sets we only give the results for two groups and $\sigma = 0.5$ and for three groups and $\sigma = 2$. The results for the other combinations of parameters lay between these two extremes.

The figure shows the average of the criterion ACALC over the simulations as a function of the number of missing values per consumer. As it can be expected, the quality of the clustering decreases when the number of missing values increases.

For the simulated data with two groups and $\sigma = 0.5$ the three methods are almost equivalent. In this situation (small noise) we obtain good results when up to five out of eight values (60 %) are missing.

For the case of three groups and $\sigma = 2$ and also for the data 'jam', the method which is based on a cross-partition and the horizontal imputation with an update of the imputations perform best. We observe a sharp decrease in criterion ACALC when the number of missing values increases. If we set up a limit at 0.9 regarding the average value of criterion ACALC, it turns out that the two best methods give good results when up to three out of eight values (40 %) are missing (simulated data with three groups and $\sigma = 2$) or when up to two out of seven values (30 %) are missing (data 'jam').

However, the results of the different simulations showed great variations. Therefore, it is interesting to consider not only the average value of ACALC but the complete distribution. Boxplots of the values of criterion ACALC for the data with three groups and $\sigma = 2$ are given in Figure 2, for three missing

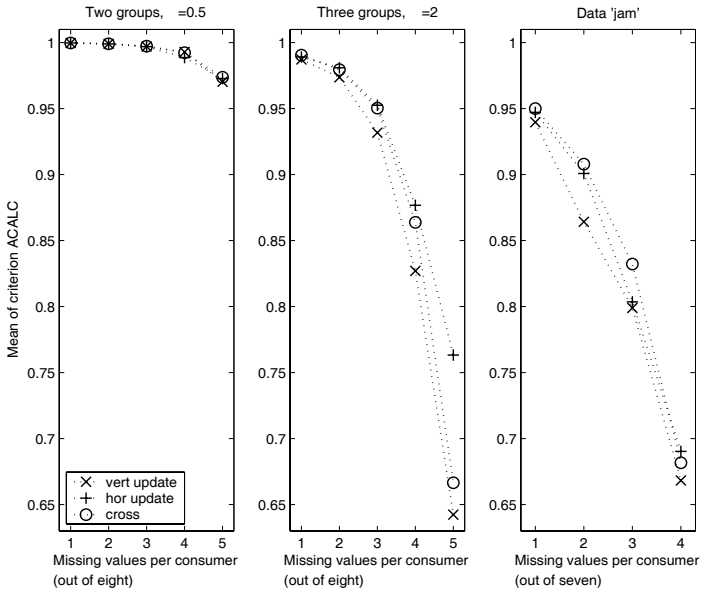


Fig. 1. Mean of criterion ACALC for the vertical imputation with an update ('vert update'), horizontal imputation with an update ('hor update') and the method based on a cross-partition ('cross')

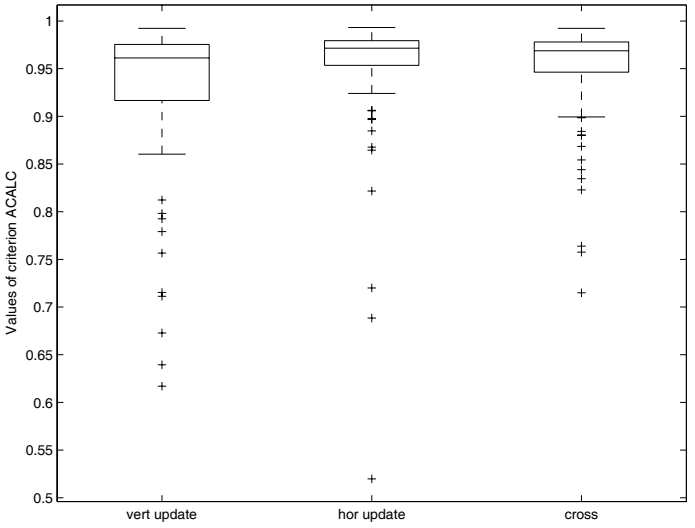


Fig. 2. Boxplots for three groups, $\sigma = 2$ and three missing values

values per consumer. It appears that the method based on a cross-partition safeguards against very poor results.

6 Conclusion

We have compared several imputation methods with regard to their performance within a clustering of variables framework. Two methods performed very well when applied in preference studies context. The first one replaces each missing value by the mean of the observed values for the observation under consideration and then clusters the completed data set. After clustering, the imputations are updated within each cluster. The second one is based on the clustering of two completed data sets and the cross-partition of the two partitions thus obtained.

In an ideal situation with small noise these methods perform very well even when more than half of the values are missing. In more realistic situations, they can perform well if less than one third of the values are missing.

Acknowledgment

The first author wishes to thank the Martin-Schmeisser-Stiftung who supported part of her work.

References

CALLIER, P. (1996): *La cartographie des préférences. Son application en milieu industriel et son extension aux plans incomplets.* Université Montpellier II (Doctorat en biostatistique).

GREENHOFF, K. and Mac FIE, H.J.H. (1994): Preference mapping in practice. In: H.J.H. Mac Fie and D.M.H. Thomson (Eds.): *Measurement of food preferences*, Blackie academic & professional, 137–166.

LEBART, L., MORINEAU, A. and PIRON, M. (2000): *Statistique exploratoire multidimensionnelle*, 3ième édition. Dunod, Paris.

SAHMER, K. (2003): *Classification des variables en présence de données manquantes : Application aux données de préférence.* Diplomarbeit, Fachbereich Statistik, Universität Dortmund.

SAS/STAT (1999): *User's guide, Version 8,* SAS Institute Inc., Cary, North Carolina.

VIGNEAU, E. and QANNARI, E.M. (2002): Segmentation of consumers taking account of external data. A clustering of variables approach. *Food Quality and Preference, 13, 515–521.*

VIGNEAU, E. and QANNARI, E.M. (2003): Clustering of variables around latent components. *Communications in Statistics – Simulation and Computation, 32, 1131–1150.*

Binary On-line Classification Based on Temporally Integrated Information

Christin Schäfer[1], Steven Lemm[1,2], and Gabriel Curio[2]

[1] Fraunhofer Institute FIRST, Intelligent Data Analysis Group
Kekuléstr.7, 12489 Berlin, Germany
[2] Dept. of Neurology, Campus Benjamin Franklin, Charité,
University Medicine Berlin, Hindenburgdamm 30, 12200 Berlin, Germany

Abstract. We present a method for on-line classification of triggered but temporally blurred events that are embedded in noisy time series. This means that the time point at which an event is initiated or a dynamical system is perturbed is known, e.g., the moment an injection of a therapeutic agent is given to a patient. From the ongoing monitoring of the system one has to derive a classification of the event or the induced change of the state of the system, e.g., whether the state of health improves or degrades. For simplification we assume that the reactions form two classes of interest. In particular the goal of the binary classification problem is to obtain the decision on-line, as fast and as reliable as possible.

To provide a probabilistic decision at every time-point t the presented method gathers information across time by incorporating decisions from prior time-points using an appropriate weighting scheme. For this specific weighting we utilize the Bayes error to gain insight into the discriminative power between the instantaneous class distributions.

The effectiveness of this procedure is verified by its successful application in the context of a Brain Computer Interface, especially to the binary discrimination task of left against right imaginary hand-movements from ongoing raw EEG data.

1 General framework

In this paper we present an approach how to improve on-line classification of sequential data by combining information across time. Accordingly the proposed method has the following underlying assumptions: First we consider only binary classification problems, which can directly be interpreted as the detection of two distinguishable states of a dynamical system, embedded in a high-dimensional noisy environment. Second we assume that the event onset or the time the system is perturbed is known. However the development of the event or the change of the systems state might be fuzzy, in the sense that relevant informations are blurred or spread over time. Examples of such kind of problem often occur in biomedical investigations, e.g., monitoring the vital state of health of a patient after an injection (Morik et al. (2000)). Also in the more general context of control and feedback control systems, this is a commonly used framework.

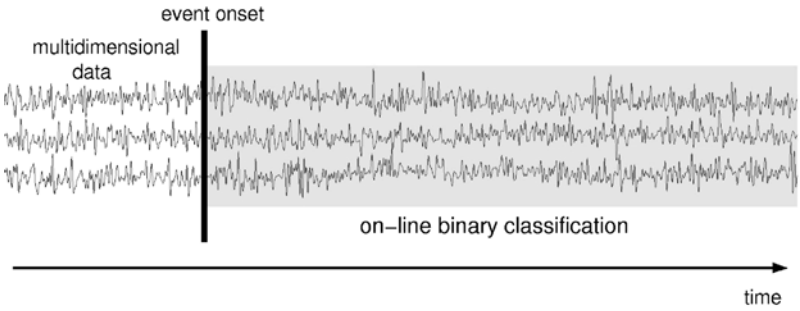


Fig. 1. Single epoch: containing a pre-event time interval and the event embedded in a high-dimensional time series. The task is to classify on-line the change of the system during the time window of interest after the event onset (gray area).

The task is to provide an on-line classification of the systems state at every time point t in a time window of interest after the event onset. The decision must be derived from the ongoing observations of the high-dimensional noisy process.

1.1 Data format

Relative to the given event onsets we cut the time-series into epochs where each epoch contains a single event in the corresponding time window of interest. Additionally, the epochs can also include data from pre-event time intervals, that can be used for calibration or baseline correction (see Fig. 1).

Let $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}$ denote the training sample, where $\mathbf{x}_k(t)$ is the high-dimensional observation of the k-th epoch, $k = 1, \ldots, N$, and $t = T_a, \ldots, T_e$ the time index. The event onset takes place at t_0, $T_a \leq t_0 \leq T_e$. The corresponding class labels $y_1, \ldots, y_N \in \mathcal{Y} = \{-1, 1\}$ of the training sample are given. The on-line classification can now be formulated as a collection of mappings $f_t : \mathcal{X} \to \mathcal{Y}$. These functions f_t can be estimated on the basis of $\mathcal{Z}_t = \{(\mathbf{x}_k(t_i), y_k),\ k = 1, \ldots, N,\ t_0 \leq t_i \leq t\}$. Utilizing these estimated functions the on-line classification of unlabeled epochs can be derived.

1.2 On-line classification

Regardless of the used classification algorithm one can distinguish between two opposite on-line classification approaches: 'instantaneous' and 'batch' classification. The instantaneous classification gives a decision d_{t_i} for each time point t_i based only on the observation at this time $x_k(t_i)$. On the opposite batch classification derives the decision D_{t_i} based on all previous observations $\{x_k(t), t \leq t_i\}$. Both approaches have advantages and drawbacks. The series of instantaneous classifications d_t can be very unsteady, whereas in contrast the series of D_t is more stable. However this stability comes at the cost

of an increased model complexity, also the amount of used data can become intractable due to memory restrictions. The complexity of the instantaneous classification problem is lower and keeps constant over the time.

Similar to batch classification one can also combine all preceding instantaneous decisions d_t, $t \leq t_i$, to a single decision D_{t_i}. This combination can be done in several ways. One can use:

1) expectation $D_{t_i} = \frac{1}{t_i - t_0} \sum_{t=t_0}^{t_i} d_t$,
2) majority vote $D_{t_i} = \text{argmax}_{y \in \mathcal{Y}} \{\sum_{t=t_0}^{t_i} I(d_t = y)\}$,
3) product $D_{t_i} = \prod_{t=t_0}^{t_i} d_t$.

Applying anyone of these combination methods one implicit assumes that each time point t contains the same 'amount' of information. In many practical applications this assumption is questionable. Furthermore by using the product based combination scheme one assumes independence of the decisions, that does not hold in general for the class of problems we are addressing.

Instead of using these combination schemes we suggest to derive a decision D_{t_i} at each time point t_i as a weighted combination of all previous, instantaneous decisions $d_t, t \leq t_i$, where the weights are chosen proportional to an estimate of the amount of discriminative information that can be derived from this specific time point.

The paper is organized as follows: The introduction of the method in section 2 is followed by an experimental section, in which we present results of a case study, illustrating the benefit of the proposed approach.

2 Integration of information across time

In order to finally derive the on-line classification at a certain time t_i, we incorporate knowledge from all preceding time points $t_0 \leq t \leq t_i$, leading to an evidence accumulation over time about the binary decision process d_t. The temporal combination is realized by taking the expectation of the class probability with respect to the discriminative power of each time instance.

More formally, a decision at time t_i is given through a weighted linear combination of the previous decision process d_t:

$$D_{t_i} = \sum_{t=t_0}^{t_i} g_t d_t. \tag{1}$$

The weights g_t represent the discriminative power. The question arises how to measure this discriminative power. In our situation we estimate two class distributions $P(x(t)|y)$, $y \in \{-1, 1\}$, on the training data. Applying Bayes decision rule, one decides for $y = 1$ if $P(y = 1|\mathbf{x}(t)) > P(y = -1|\mathbf{x}(t))$, otherwise for $y = -1$. The Bayes error of this decision rule is given through: $P(\text{error}|\mathbf{x}(t)) = \min[P(y = 1|\mathbf{x}(t)), P(y = -1|\mathbf{x}(t))]$. A small Bayes error

indicates a good separability of the two class distributions while on the other hand, if the Bayes error is high the data contain less discriminative information. Consequently, we use the Bayes error as a measure of the discriminative power at time point t. Thus the weights are defined as

$$g_t = 0.5 - P(\text{error}|\mathbf{x}(t)). \tag{2}$$

Since the estimation of the Bayes error is usually intractable, we exploit the Chernoff bound (Duda et al. (2001)), which upper bounds the Bayes error. The Chernoff bound is defined as the minimum over all $\beta \in [0, 1]$ of the right hand side of the following inequality:

$$P(\text{error}) \leq P(y=1)^{\beta} P(y=-1)^{1-\beta} \int p(\mathbf{x}|y=1)^{\beta} p(\mathbf{x}|y=-1)^{1-\beta} d\mathbf{x}. \tag{3}$$

Notice that, if $p(\mathbf{x}|y=1)$ and $p(\mathbf{x}|y=-1)$ are normal, the Chernoff bound can be evaluated analytically by finding β that minimizes

$$\int p(\mathbf{x}|y=1)^{\beta} p(\mathbf{x}|y=-1)^{1-\beta} d\mathbf{x} = e^{-k(\beta)}, \tag{4}$$

where

$$2k(\beta) = \beta(1-\beta)(\mu_+ - \mu_-)'[\beta\Sigma_- + (1-\beta)\Sigma_+]^{-1}(\mu_+ - \mu_-) + \ln \frac{|\beta\Sigma_- + (1-\beta)\Sigma_+|}{|\Sigma_-|^{\beta}|\Sigma_+|^{1-\beta}}.$$

3 Application

As an application of the proposed method we choose the binary classification problem of distinguishing between left and right imagined hand movements from recordings of electroencephalogram (EEG). This is a common task in the newly emerging field of Brain Computer Interface (BCI)(Krausz et al. (2003), Dornhege et al. (2004)).

The data we use in this study, are taken from the 2003 BCI-competition (Blankertz et al. (2003)). In particular we apply our method to data set III - "imagined hand movement", provided by the Dept. of Med. Informatics, Inst. for Biomed. Eng. at the Univ. of Techn. Graz. The EEG from three channels (C3, Cz, C4) was acquired with band filter settings of 0.5 to 30 Hz and sampled at 128 Hz. The data consist of 140 labeled and 140 unlabeled trials of imaginary hand movements, with an equal number of left and right hand trials. Each trial has a duration of 9 s: after a 3 s preparation period a visual cue (arrow) is presented pointing either to the left or the right. This is followed by another 6 s for performing the imagination task (for further details see Blankertz et al. (2003)). The specific competition task is to provide an on-line discrimination between left and right movements for each of the 140 unlabeled single trials (STs). In particular, at every time instance in the

interval from 3 to 9 seconds a decision and its confidence must be supplied. The objective of the competition was to detect the respective motor intention as early and as reliable as possible and therefore perfectly meets the settings of the proposed method.

3.1 Neurophysiology

The human perirolandic sensorimotor cortices show rhythmic macroscopic EEG oscillations (μ-rhythm) (Hari and Salmelin (1997)), with spectral peak energies around 10 Hz (localized predominantly over the postcentral somatosensory cortex) and 20 Hz (over the precentral motor cortex). Modulations of the μ-rhythm have been reported for different physiological manipulations, e.g., by motor activity, both actual and imagined (Jasper and Penfield (1949), Pfurtscheller and Arabibar (1979), Schnitzler et al. (1997)). Standard trial averages of μ-rhythm power show a sequence of attenuation, termed event-related desynchronization (ERD) (Pfurtscheller and Arabibar (1979)), followed by a rebound (event-related synchronization: ERS) which often overshoots the pre-event baseline level. Imaginary movements modulate the μ-rhythm on the hemisphere contralateral to the respective event more than ipsilateral (Pfurtscheller and Arabibar (1979), Schnitzler et al. (1997), Nikouline et al. (2000)), e.g. left imaginary movement causes stronger perturbations on the right motor cortex (C4) and vice versa (see Figure 2).

3.2 Model

In order to distinguish between STs of left and right hand imaginary movements, we utilize the accompanying EEG μ-rhythm perturbation. Similar approaches were pursued in (Pfurtscheller et al. (1997), Neuper et al. (1999)). Since we assume that the mid-line channel Cz contains little discriminative information, we exclude it and restrict the analysis to C3 and C4. To extract the modulations in the two relevant frequency bands, we map the EEG to the time-frequency domain by means of Morlet wavelets (Torrence and Compo (1998)). Furthermore, we assume the existence of two distinguishable prototypical behaviors of modulation for the *absolute* amplitude of the μ-rhythm caused by either imaginary left or, respectively, right hand movements. Based on these physiological concepts we estimate two probabilistic models, one for each class of imaginary movement. For each class and at any time instance $t \in [0-9]\,\mathrm{s}$ we assume a 4-dimensional Gaussian distribution of the feature vectors $\mathbf{a}(t)$ (the amplitudes of the two relevant frequency bands at the two electrodes), i.e.,

$$p(\mathbf{a}(t)|y) = N(\mu^y(t), \Sigma^y(t))\,, \tag{5}$$

where $\mu^y(t)$ and $\Sigma^y(t)$ are the individual means and the covariance matrices of the two classes $y \in \{\mathrm{L}, \mathrm{R}\}$ that have been estimated in a robust manner.

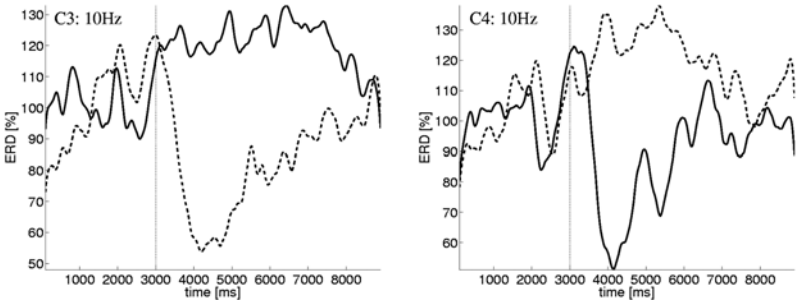


Fig. 2. The panels show the averaged event-related desynchronization (ERD) of the μ-rhythm at 10 Hz for the imagination of left (solid line) and right (dashed line) hand movement. The vertical line indicates the begin of the imagination period. The μ-rhythm amplitude is attenuated in relation to the preceding baseline during the motor intention. This attenuation is prominent contralateral to the intended movement, i.e., for right hand movement over the left hemisphere (C3) and over the right hemisphere (C4) for the left hand.

The instantaneous classification at a single time point is given by

$$p\left(y|\mathbf{a}(t)\right) = \frac{p\left(\mathbf{a}(t)|y\right)}{p\left(\mathbf{a}(t)|\mathrm{L}\right) + p\left(\mathbf{a}(t)|\mathrm{R}\right)}. \tag{6}$$

The temporal combination according to eq.(1) is realized by taking the expectation of the class probabilities from eq.(6) with respect to the discriminative power g_t at each time point:

$$p\left(y|\mathbf{a}(t_0), \ldots, \mathbf{a}(t_i)\right) = \frac{\sum_{t_0 \le t \le t_i} g_t p(y|\mathbf{a}(t))}{\sum_{t_0 \le t \le t_i} g_t}. \tag{7}$$

As described above (section 2) we measure the discriminative power through the Bayes error, which itself is approximated from above by the Chernoff bound eq.(3) and finally define g_t using equal class prior probabilities by

$$2g_t := 1 - \min_{0 \le \beta_t \le 1} \int p(\mathbf{a}(t)|\mathrm{L})^{\beta_t} p(\mathbf{a}(t)|\mathrm{R})^{1-\beta_t} d\mathbf{a}(t). \tag{8}$$

The distributions of the feature vectors $\mathbf{a}(t)$ are normal, therefore the minimization can be obtained easily. Fig. 3 shows the estimated Chernoff bound, given the labeled training data. Note that the most discriminative information occurs around 4.5 s, as indicated by the minimum of the error bound that corresponds to the maximum weight in the integration process.

Due to the submission requirements of the competition the final decision at this time point is

$$d_{t_i} = 0.5 - p(L|\mathbf{a}_{t_0}, \ldots, \mathbf{a}_{t_i}), \tag{9}$$

where a positive or negative sign refers to right or left movements, while the magnitude indicates the confidence in the decision.

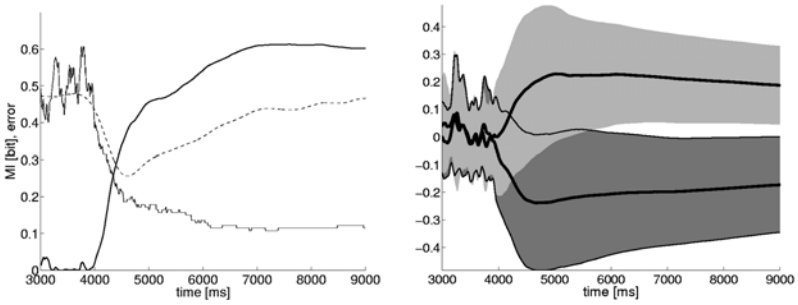


Fig. 3. Left panel shows the time course of the classification error (thin solid), the Chernoff bound on the Bayes error (dashed) and the mutual information (thick solid). Right panel displays time course of the mean and standard deviations of the decision according eq. (9) for right (light grey) and left (dark grey) imaginary movements on the test data.

3.3 Results

After all model parameters have been estimated by means of a leave-one-out cross-validation optimization on the labeled training data, we applied the estimated model to the feature vectors of the unlabeled STs of the test data. The resulting time courses for both the model error of the binary classification and the mutual information (MI) (Schlögl et al. (2003)) on the previously unlabeled data are presented in Fig. 3. During the first four seconds the classification is rather by chance, after four seconds a steep ascent in the classification accuracy can be observed in both the raising MI and the decreasing classification error. Although the Bayes error bound starts to gradually increase again after 4.5 s, indicating fading separability, the full model still gains information due to the integration process, so that at 6.8 s an overall minimum error of 10.7% is achieved. The MI maximum of 0.61 Bit occurs at 7.6 s indicating a peak decision confidence at this time. Demonstrating the time courses of the class means and the standard deviations of the decision the right panel of Fig. 3 emphazises the high discriminative ability of the proposed procedure: around 6 s there is no overlap between the class standard deviation tubes, reflecting the high confidence of the decisions. A comprehensive comparison of all submitted techniques to solve the specific task for the data set III of the BCI competition is provided in (`http://ida.first.fraunhofer.de/projects/bci/competition/...`
`...results/index.html#graz`). Basically this evaluation reveals that the proposed algorithm outperforms all competing approaches, including traditional adaptive AR-parameter based methods.

Acknowledgement The authors want to thank the IDA group at Fraunhofer FIRST for valuable discussions and suggestions. The studies were supported in part by BMBF Grants FKZ 011BB02A,B and DFG SFB 618-B4.

References

BLANKERTZ, B., MÜLLER, K.-R., VAUGHAN, T.M., CURIO, G., SCHALK, G., WOLPAW, J.R., SCHLÖGL, A., NEUPER, C., PFURTSCHELLER, G., HINTERBERGER, T., SCHRÖDERAND, M. and BIRBAUMER, N. (2003): The BCI competition 2003. *IEEE Trans. Biomed. Eng.*

DORNHEGE, G., BLANKERTZ, B., CURIO, G. and MÜLLER, K.-R. (2004): Boosting bit rates in non-invasive EEG single-trial classification by feature combination and multi-class paradigms. *IEEE Trans. on Biomed. Eng.*, in press.

DUDA, R.O., HART, P.E. and STORK, D.G. (2001): *Pattern classification.* 2nd ed. John Wiley & Sons, New York.

HARI, R. and SALMELIN, R. (1997): Human cortical oscillations: a neuromagnetic view through the skull. *Trends in Neuroscience, 20, 44–49.*

JASPER, H. and PENFIELD, W. (1949): Electrocorticograms in man: Effect of voluntary movement upon the electrical activity of the precentral gyrus. *Arch. Psychiatrie Zeitschrift Neurol., 183, 163–174.*

KRAUSZ, G., SCHERER, R., KORISEK, G. and PFURTSCHELLER, G. (2003): Graz-BCI: state of the art and clinical applications. *IEEE Trans Neural Syst Rehabil Eng., 11(2), 177–180.*

MORIK, K., IMHOFF, M., BROCKHAUSEN, P., JOACHIMS, T. and GATHER, U. (2000): Knowledge discovery and knowledge validation in intensive care. *Technical Report 14*, SFB 475, University Dortmund.

NEUPER, C., SCHLÖGL, A. and PFURTSCHELLER, G. (1999): Enhancement of left-right sensorimotor EEG differences during feedback-regulated motor imagery. *Journal Clin. Neurophysiol., 16, 373–382.*

NIKOULINE, V., LINKENKAER-HANSEN, K., WIKSTRÖM, H., KESÄNIEMI, M., ANTONOVA, E., ILMONIEMI, R. and HUTTUNEN, J. (2000): Dynamics of mu-rhythm suppression caused by median nerve stimulation: a magnetoencephalographic study in human subjects. *Neuroscience Letters*, 294.

PFURTSCHELLER, G. and ARABIBAR, A. (1979): Evaluation of event-related desynchronization preceding and following voluntary self-paced movement. *Electroenceph. clin. Neurophysiol, 46, 138–146.*

PFURTSCHELLER, G., NEUPER, C., FLOTZINGER, D. and PREGENZER, M. (1997): EEG-based discrimination between imagination of right and left hand movement. *Electroenceph. clin. Neurophysiol., 103, 642–651.*

SCHLÖGL, A., SCHERER, R., KEINRATH, C. and PFURTSCHELLER, G. (2003): Information transfer of an EEG-based brain-computer interface. In: *Proc. First Int. IEEE EMBS Conference on Neural Engineering*, 641–644.

SCHNITZLER, A., SALENIUS, S., SALMELIN, R., JOUSMÄKI, V. and HARI, R. (1997): Involvement of primary motor cortex in motor imagery: a neuromagnetic study. *Neuroimage, 6,201–208.*

TORRENCE, C. and COMPO, G.P. (1998): A practical guide to wavelet analysis. *Bull. Am. Meterol., 79,61–78.*

Different Subspace Classification

Gero Szepannek * and Karsten Luebke

University of Dortmund**, Department of Statistics, 44221 Dortmund, Germany

Abstract. We introduce the idea of **Characteristic Regions** to solve a classification problem. By identifying regions in which classes are dense (i.e. many observations) and also relevant (for discrimination) we can characterize the different classes. These Characteristic Regions are used to generate a classification rule. The result can be visualized so the user is provided with an insight into data for an easy interpretation.

1 Introduction

Supervised Classification or Discrimination often involves two goals: the first is allocation or prediction, i.e. assigning class labels to new observations. The second goal, which can be even more important, is descriptive and involves the disclosure of the underlying differences between the classes. The new Different Subspace Classification (DiSCo) method is a method to simultaneously visualize and classify multi-class-problems in high dimensional spaces and therefore is designed to attain both predictive and descriptive goals.

The problem of classification or pattern recognition is given in the following way: N objects $x_n, n = 1, \ldots, N$, are observed, each object belonging to one and only one class $k_n, k_n \in \{1, \ldots, K\}, n = 1, \ldots, N$. The class membership is known to the user. N_k objects are observed from class k. This set of objects is called training data. For each object D variables $x^d, d = 1, \ldots, D$, are observed. Every object x_n can be considered as a D-dimensional realization of a random vector X_n following an unknown distribution that depends on its class k_n.

The first goal is to be able to determine the correct (unknown) class for objects x_{new} that will be observed in future. The second goal is to find out the characteristics of the different classes by analyzing the training data. The higher the dimension of the data the more challenging is the understanding of the data. So if there are many observed variables, methods of variable selection are often used to reduce the dimension of the data. These methods identify and retain those of the variables that separate the classes best. Then following this procedure a classification method is (re-)applied to the resulting subspace of variables. A problem may be that in general the variables do not contain equal separating-information for all classes. So a variable can contain

** This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

* e-mail: szepannek@statistik.uni-dortmund.de

information for separating class i from the rest but no information for the separation of class $j \neq i$.

In DiSCo variable selection is intrinsic to the classification method. The resulting subsets of variables which are used for discrimination of the classes can differ between the classes.
A focus is also laid on the visualization of the class-characteristics. The proposed method does not make any assumptions about the underlying distribution of the data. The only weak assumption is that objects of the same class are similar in some of their predictor values.

In the following chapter the principle of Characteristic Regions is defined and a classification rule developed. Chapter 3 explains the visualization of the results. Chapter 4 briefly summarizes the choice of parameters for the implementation of the method while chapter 5 contains a simulation study with comparison to Classification trees and Discriminant analysis.

2 Notation and method

The idea of the new method is to search for Characteristic Regions, i.e. sets of values in some variables that indicate the class-membership. To build up these Characteristic Regions two steps are needed. The first step is to search for intervals of the realizations of the random variables that contain a large probability mass of the classes. The resulting "regions" are called Dense Regions. The second step, which is independent of the first, identifies regions that discriminate at least one class from the others because of a relatively high density. These regions are called Relevant Regions. Regions that are both dense and relevant are then called Characteristic Regions.

2.1 Characteristic regions

Definition 1. S being the set of all possible predictor values of an object x_n, for all d let $\{R_m^d \ : 0 \leq m \leq M^d + 1\}$ be a contiguous segmentation of an interval covering $S \cap X^d$ following

1. $\bigcup_{m=0}^{M^d+1} R_m^d \supseteq S \cap X^d$
   (All possible values of X^d are covered by the union of all its regions.)
2. $\forall\, x_1, x_2 \in R_m^d \; and \; \alpha \in [0,1] \; : \; \alpha x_1 + (1-\alpha) x_2 \in R_m^d$
   (The regions of every variable are contiguous.)
3. $\forall\, x_1 \in R_{m_1}^d, x_2 \in R_{m_2}^d, \; m_1 < m_2 \; : \; x_1 < x_2$
   (In every variable the regions are disjoint and also ordered.)

R_m^d are called ***regions*** of variable X^d.

By restriction 2 all the objects that fall into one region can be considered to be similar.

Definition 2. Let x_n^d be the value taken by object n in variable X^d and let k_n be the corresponding, known index of its class. Then

$$n_m^d(k) := \sum_{n=1}^{N} I_{[R_m^d]}(x_n^d)\, I_{[k]}(k_n) \tag{1}$$

with $I_{[\cdot]}$ as the indicator function is called the ***corresponding frequency*** of class k in Region m of variable d.

As the $n_m^d(k)$ should represent the density of the data it is assumed for simplicity of comparisons that for any fixed d and all $1 \leq m \leq M^d$: $\sup_{x \in R_m^d} - \inf_{x \in R_m^d} \equiv const.$, so the regions of a variable have equal width. By this the corresponding frequencies are proportional to heights of histogram bars of the classes if the bandwidths are given by the regions.

Let ***Dense Regions*** be those regions which contain most of the classes' probability masses. Let $S_{DR} > 0$ be a threshold to construct classwise Dense Regions. Then Dense Regions are regions $R_{m_0}^d(k)$ with

$$n_{m_0}^d(k) \geq S_{DR} \frac{\sum_{m=0}^{M^d+1} n_m^d(k)}{M^d} \tag{2}$$

This proceeding corresponds to comparing the observed corresponding frequency to the mean over all regions.

Relevant Regions should be the regions where the density of one class k is high compared to those of the other classes and so a new observed object lying in this region strongly indicates its membership to class k. Let $S_{RR} > 0$ be a threshold to construct classwise Relevant Regions. Then Relevant Regions are regions $R_m^d(k_0)$ with:

$$\frac{n_m^d(k_0)}{N_{k_0}} \geq S_{RR} \frac{\sum_{k=1}^{K} \frac{n_m^d(k)}{N_k}}{K} \tag{3}$$

To be able to compare the regions' densities of different classes by corresponding frequencies they have to be weighted by their observed absolute frequencies. Finally, ***Characteristic Regions*** are regions that are both dense and relevant.

2.2 Classification rule

Let $w_m^d(k) \geq 0$ be the ***class wise weight of a region*** of class k connected to region R_m^d.

The Characteristic Regions are used to build up the classification rule by summing the weights over all variables. Then the assignment of the class is obtained by

$$\hat{k}(x_{new}) = \arg\max_k \sum_{d=1}^{D} \sum_{m=0}^{M^d+1} I_{[R_m^d]}(x_{new}) w_m^d(k) \tag{4}$$

where the weights of the Characteristic Regions are defined by

$$w_m^d(k_0) := \begin{cases} 0 & \text{if (2) or (3) do not hold} \\ \frac{n_m^d(k_0)\frac{p(k_0)N}{N_{k_0}}}{\sum_{k=1}^{K} n_m^d(k)\frac{p(k)N}{N_k}} & \text{if } R_m^d \text{ is characteristic for class } k_0 \end{cases} \tag{5}$$

$\frac{p(k)N}{N_k}$ is a correction term for the absolute frequency of the classes in the data with the prior probabilities of the classes – if it differs from the observed frequency. The weights are motivated by the marginal probability of $k_{new} = k$ given $x_{new}^d \in R_m^d$, if R_m^d is "characteristic" for class k.

As only Characteristic Regions are used for the classification rule the cutpoints of the regions may disregard information. So to keep more of the classes' probability masses we propose another smoothed classification rule where the weights $w_m^d(k)$ are as before but additionally the adjoining regions are included in the model. Then:

$$\hat{k}(x_{new}) = \arg\max_k \sum_{d=1}^{D} \sum_{m=0}^{M^d+1} I_{[R_m^d]}(x_{new})(\frac{1}{2}w_{m-1}^d(k) + w_m^d(k) + \frac{1}{2}w_{m+1}^d(k)) \tag{6}$$

with $w_m^d(k) = 0$ for $m = -1, M^d + 2$.

3 Visualization

The weights $w_m^d(k)$ described above mimic marginal conditional probability of the different classes. As only Characteristic Regions will be shown in our visualization only robust information relevant for classification is given. So plotting these class wise weights of the regions (see equation 5) provides a visualization of the class characteristics and an interpretation may be simplified.

As example we illustrate the method in Figure 1 on the well known Iris data set introduced by Fisher. The values of the variables are shown on the x-axes while the different colours of the bars symbolize the different true classes (black = "Setosa", light grey = "Virginica" and dark grey = "Versicolor"). The heights are the weights of the Characteristic Regions. It can be seen that the variable "Sepal length" only serves to indicate membership of one of the classes "Virginica" or "Setosa" but not for "Versicolor", while the variable "Sepal width" just serves to characterize a plant of class "Setosa" or "Versicolor".

The "Petal" variables seem to separate all three classes with the lowest values for class "Setosa".The upper extreme values indicate the class "Virginica". As the plots of these two variables are of the same structure one can suppose a correlation between these variables.

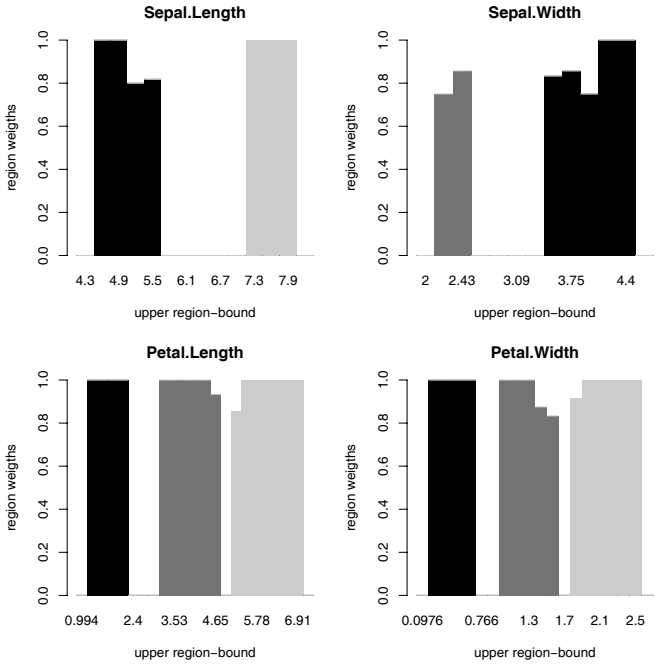


Fig. 1. Example: Visualization of a result for Iris data

4 Parameter choice for DiSCo

For the implementation one has to find the Characteristic Regions. So the problem is how to form the regions and how to choose the thresholds.

4.1 Building the regions

As mentioned earlier the corresponding frequencies should be proportional to heights of histogram bars for convenience so we can refer to the theory of nonparametric density estimation to build the regions. In histogram density estimation the problem consists in smoothing but not over-smoothing the empirical distribution of the data. Thus the bandwidth of a histogram should be chosen neither too small nor too large. Freedman and Diaconis (1981) suggest a choice of

$$bw = \frac{2}{\sqrt[3]{N}} IQR \tag{7}$$

as bandwidth where IQR is the interquartile range. Under weak assumptions this histogram is L^2-convergent for density estimation (Freedman and Diaconis (1981)). As the distribution may be different in the classes this

must be done for every class – and every variable. The number of classwise bins is then $M^d(k) = \lfloor \frac{x^d_{(N_k)} - x^d_{(1_k)}}{bw(k,d)} \rfloor$ with $x^d_{(N_k)}$ and $x^d_{(1_k)}$ being the classwise maximum respective minimum and $\lfloor . \rfloor$ being the rounding operator. With $IV^d := [x^d_{(1)}, x^d_{(N)}]$ and $IV^d_k := [x^d_{(1_k)}, x^d_{(N_k)}]$ let:

$$M^d := \left\lfloor \frac{\int_{IV^d} 1dt}{\int_{\cup_k IV^d_k} 1dt} \left\{ \sum\nolimits_k \left(M^d(k) \int_{IV^d_k} \{ \sum\nolimits_k I_{[IV^d_k]}(s) \}^{-1} ds \right) \right\} \right\rfloor \tag{8}$$

This means that the classwise number of bins is interpolated resp. averaged for intervals covered by none, one or more than one class. So the regions of variable d are IV^d divided into M^d equal parts. R^d_0 and $R^d_{M^d+1}$ cover the upper and lower rest.

4.2 Optimizing the thresholds

There remains the question how to choose the thresholds in equation 2 and equation 3. So far no theoretical background is known for an optimal choice of both S_{DR} (Dense Regions) and S_{RR} (Relevant Regions).

The optimal parameters are found by a contracting 2-dimensional grid-search algorithm. As the criterion for optimization the cross validated error rate is used. It should be noticed that since the number of observations is finite small changes of the two thresholds will not change the resulting model. In order to check the parameters one can consider that a rather small threshold S_{DR} eliminates outliers but keeps a large probability mass in the remaining regions. A S_{RR} rather large keeps only regions in the model that strongly indicate one class.

5 Simulation study

5.1 Data generation

In order to obtain more general results an experimental design is used in data generation to be able to compare the effects of possibly influencing factors in the data on the classification result of DiSCo and of two well-established other methods: Classification Trees (CART) and Linear Discriminant Analysis (LDA).

With the factor levels described below, data of 8 or 12 variables are first drawn from independent normal distributions with variance 1 but different expectations in 3 classes. These data are transformed to possess different kurtosis and skewness and to be deflected.

Below we give a brief description of the seven investigated factors:

- The class priors may be equal or not.

- We investigated two different class mean settings in the first 6 variables: either only one class mean separated from the others or all three class means are different (one in the middle between the others). For 3 variables the doubled 0.95 quantile is chosen, for the 3 other variables the doubled 0.9 quantile of the standard normal distribution is chosen for the tallest differences in the location of the class means.
- 2 or 6 irrelevant independent variables are attached to the data that are $N(0, 1)$ distributed for all classes, i.e. either a quarter or half of the variables do not contain any separating information.
- All variables are transformed to have high or low kurtosis and skewness following the Johnson-System (see Johnson (1949)) to generate a wide range of values in the kurtosis-skewness-plane.
- The probability of an object to be deflected is fixed to be 0.1 or 0.4, where deflection means that an object is moved into one of two directions: half the distance towards its class mean or half the way away from it into the direction of its nearest wrong class.

The factors and levels included in the experimental design of the simulation study are summarized in Table 1. A Plackett-Burman design (Plackett and Burman (1946)) for these factors was repeated 20 times.

Table 1. Effects and levels on the simulated data sets

Effect	Low level	High level
Class priors	$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	$(\frac{2}{3}, \frac{1}{6}, \frac{1}{6})$
Number of different class means	3	2
Added irrelevant variables	2	6
Kurtosis	2.7	5
Skewness	0.1^2	1.15^2
Probability to be deflected	0.1	0.4
Direction of deflection	towards class mean	away from class mean

5.2 Results

Compared are both proposed classification rules for the DiSCo method including (labelled (1)) and not including (2) the adjoining regions, CART (Breiman et al. (1984)) and LDA. Table 1 shows the mean error rates on the test data and the estimated effects of the main factors (coded to $-1/+1$) used in the design (cp. table 1) on $log(odds(hitrate))$. These effects can be estimated independently by a regression on the coded influencing factors:

DiSCo seems to outperform the Classification trees and is almost as good as LDA. One can also see that there are only small differences between both

Table 2. Results: Overall mean error and estimated effects on $log(odds(hitrate))$

	DiSCo (1)	DiSCo (2)	CART	LDA
Overall mean error	0.085	0.079	0.127	0.075
Class Priors	0.41	0.48	0.19	0.24
Number of different class means	0.17	0.24	0.34	0.37
Irrelevant variables	0.19	0.26	0.21	-0.11
Kurtosis	-0.12	-0.07	-0.27	0.09
Skewness	1.16	1.06	0.89	0.70
Probability to be deflected	-0.10	-0.16	-0.24	-0.61
Deflected direction	-2.17	-2.23	-1.00	-2.73

proposed classification rules for the DiSCo method so there is no general rule which one to use.

It can be concluded that LDA has best overall mean error. Classification trees perform well with deflection away from the class mean but having a large general deficit. The DiSCo method, having a good average result, is preferable with skewed data or differing class priors and a high percentage of deflected objects.

Mean values for the optimal thresholds are $S_{DR} = 0.67$ and 0.54 including and not including the neighbour regions while the averaged optimal S_{RR} are 1.88 and 1.75.

6 Summary

The introduced concept of **Characteristic Regions** allows the visualization of the class characteristics and so satisfies the aim of an easy comprehension and interpretation of the data. It also yields intuitive classification rules.
On simulated test data it outperformed classification trees and was almost as good as the linear discriminant analysis.

References

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. and Stone, C. (1984): *Classification and regression trees*. Chapman & Hall, New York.

FREEDMAN, D. and DIACONIS, P. (1981): On the histogram as a density estimator: L_2 theory. *Zeitschrift für Wahrscheinlichkeitsthoerie und verw. Gebiete, 57, 453–476.*

JOHNSON, N. L. (1949): Bivariate distributions based on simple translation systems. *Biometrika, 36, 149–176.*

PLACKETT, R., BURMAN, J. (1946): The design of optimum multifactorial experiments. *Biometrika, 33, 305–325.*

Density Estimation and Visualization for Data Containing Clusters of Unknown Structure

Alfred Ultsch

Databionics Research Group,
University of Marburg,
35032 Marburg, Germany

Abstract. A method for measuring the density of data sets that contain an unknown number of clusters of unknown sizes is proposed. This method, called Pareto Density Estimation (PDE), uses hyper spheres to estimate data density. The radius of the hyper spheres is derived from information optimal sets. PDE leads to a tool for the visualization of probability density distributions of variables (PDEplot). For Gaussian mixture data this is an optimal empirical density estimation. A new kind of visualization of the density structure of high dimensional data set, the P-Matrix is defined. The P-Matrix for a 79- dimensional data set from DNA array analysis is shown. The P-Matrix reveals local concentrations of data points representing similar gene expressions. The P-Matrix is also a very effective tool in the detection of clusters and outliers in data sets.

1 Introduction

To identify clusters in a data set it is sometimes not enough to consider distances between the data points. Consider, for example, the TwoDiamonds data set depicted in Figure 1. The data consists of two clusters of two dimensional points. Inside each "diamond" the values for each data point were drawn independently from uniform distributions. At the central region, marked with an arrow circle in Figure 1, the distances between the data

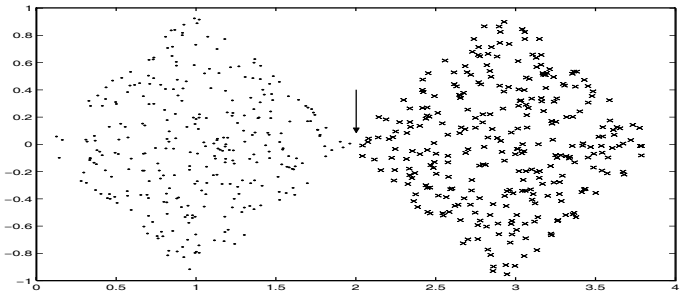


Fig. 1. The TwoDiamonds data set

points are very small. For distance based cluster algorithms it is hard to detect correct boundaries for the clusters. Distance oriented clustering methods such as single linkage, complete linkage, Ward etc. produce classification errors. The picture changes, however, when the data's density is regarded. The density at the touching point of the two diamonds is only half as big as the densities in the center regions of the clusters. This information may be used for a clustering of the data. Density based clustering algorithms have drawn much attention in the last years within the context of data mining, see for example (Xu et al (1998), Hinneburg (1998)). These algorithms call for methods to estimate the density of the data. In this paper we propose a method for density estimation that is optimal in an information theoretic sense. Furthermore we propose a one dimensional plot for empirical probability density and a method to visualize high dimensional density distributions.

2 Information optimal sets, Pareto Radius, PDE

Let S be a subset of a set of points and p denote the probability that a point belongs to S. The information of the set S can be calculated using Shannon's formula for information. Scaled to the range [0,1], the information of a set $I(S)$ is calculated as I(S) = - e p ln (p). An information optimal set is minimal in size but contains as much information as possible. To find such an optimal set size, define the unrealized potential $URP(S)$ of a set S as the Euclidian distance from the ideal point to $(p, I(S))$ of a given set. The ideal point (0,1) corresponds to a minimal set size producing 100% of information. Minimizing the unrealized potential URP results in an optimal set with p_u =20.13%. This set size produces 88% information. Subsets of the relative size p_u are called information optimal. The optimality of this set at about (20%, 80%) can serve as an explanation for the so called Pareto 80/20 law, which is empirically found in many domains. Let $d(x_i, x_j)$ be a dissimilarity measure defined on the set $E = x_1, ..., x_d$ of collected data. $N(x, r) = |\{x_i \in E \mid d(x, x_i) \leq r\}|$ is the number of points inside a sphere of radius r around x. The Pareto Radius r_p is a radius such that the median of the spheres around all data points is information optimal, i.e.: median$(N(x_i, r_p)) = p_u \cdot d$. This means the spheres contain in the average information optimal sets.
If a data set contains cluster, the Pareto Radius should be information optimal for each cluster. Let $v(k)$ denote the ratio of intra cluster distances to inter distances for k clusters in the data set. Then the optimal Pareto Radius is $r_p(k) = v(k) \cdot r_p$. In Ultsch (2003) an estimation procedure for $v(k)$ is described. The results of a large simulation study to find $v(k)$ is shown in Figure 2. For a given number k of clusters the circles give the mean of $v(k)$. The bordering lines indicate the interval in which $v(k)$ could be found with 95% probability. If the number of clusters k is known, $v(k)$ can be estimated as the mean in Figure 2. If k is unknown, $v = 1/3$ is covered by the 95% confidence interval for 3 up to 13 clusters. For only one or two clusters, $v(2) = 0.7$

is a good estimate. If the minimum number of clusters in a data set can be estimated, the lower bound of the 95% confidence interval is a good choice for v. From Figure 2 it can also be seen that all $v < 1/3$ cover a broad range of possible cluster numbers. Thus a rough estimate of the cluster number k is sufficient for the calculation of a suitable Pareto Radius. The calculation of a Pareto Radius for large data sets can be optimized by concentrating on the percentiles of distances. The Pareto Percentile pc_{par} is then the percentile of all distances closest to the Pareto Radius. As cluster corrected Pareto Radius the distance percentile closest to $v(k) \cdot pc_{par}$ is used. An empirical density estimate at a point x for data sets containing clusters is the number of points inside the sphere around x with radius $r_p(k)$. This density estimation is called Pareto Density Estimation (PDE).

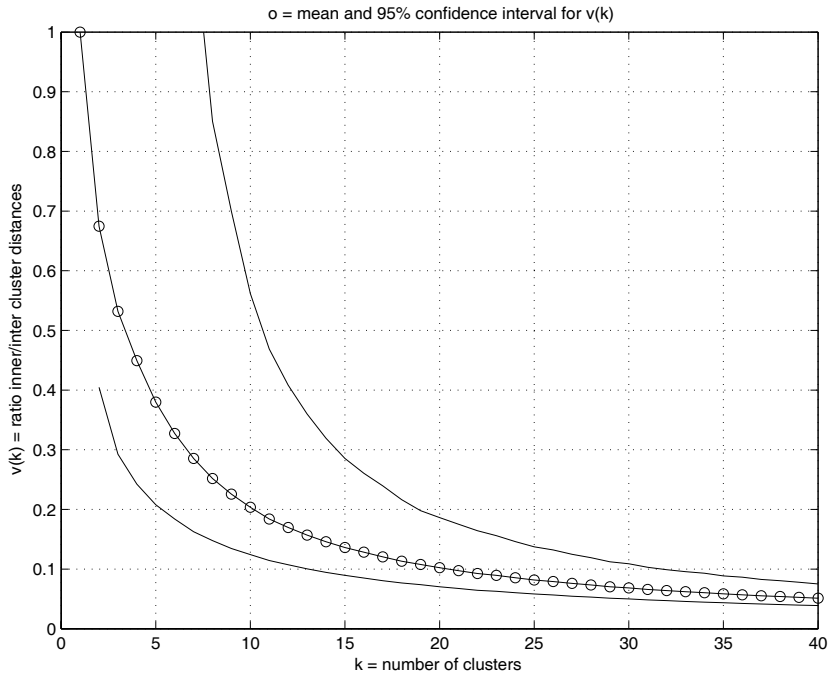


Fig. 2. Estimation of the ratio of intra/inter cluster distances

3 PDE in one dimension: PDEplot

For one dimensional data a probability density estimation can be calculated from PDE.

$$\mathrm{PPDE}(x) = \frac{\mathrm{N}(x, r_p)}{\text{area}} \qquad \text{where area is} \qquad \int_{-\infty}^{\infty} \mathrm{N}(x, r_p)dx \tag{1}$$

The denominator 'area' is approximated using the trapezoidal method on $(x_i, \mathrm{N}(x_i, rp))$. The formula (1) assures that the integral on PPDEplot(x) is equal to 1 to get a valid probability density function. Plotting PPDE(x) against x is the PDEplot.
PDEplots can be used for a closer look on distributions. In DNA microarray experiments using high density oligonucleotide arrays, such as the Affymetrics gene chip, it is important to visualize the distributions of gene expression rates (Parmigiani et al. (2003)). Gentleman and Carry implemented software, the so called "Expression Density Diagnostics", to compare empirical distributions to a given set of model distribution (Gentleman and Carry (2003)). Their probability density visualization allows, however, hardly a distinction between the presumably different distributions (see Gentleman and Carry (2003), p. 70, Figure 2.7). Figure 3 shows a PPDE plot of an Affymetrix data set of 124222 gene expressions for 7 liver and 7 brain cells of mice. The figure shows, that a decision for the origin of the cells (liver vs brain) can be based only on the different distributions. In histograms this is hardly seen. To find gene expression that differenciate brain vs. liver, the two different

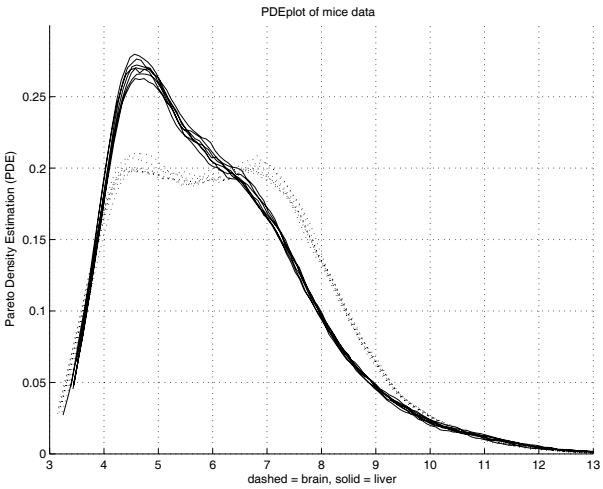


Fig. 3. PDEplot of DNA microarray data of liver and brain

distributions have to be adapted carefully. PDEplots allow to judge the quality of such adaptations. In this way PDE contribute to a successful search for differential expressed genes in DNA microarray experiments.

4 Measuring and visualization of density of high dimensional data

We used PDE on a data set of 1000 points with 120 variables. The variables describe molecules by the number of different atom types (Ghose and Crippen (1986)). It was known that there were at least five clusters in the data set. The Pareto Radius obtained for this data set was $r_p(k) = 3.6$. Figure 4 shows the Pareto Radius compared to hypersphere density estimations using other radii. It can be seen that for a large radius the density estimation oversmoothes the true density distribution. No structural information can be gained from such a density estimation. For a small radius e.g. $r = 2$ many of the spheres are empty. The density estimation with the Pareto Radius shows the most structural features of the data set. A large number of points, presumably inside a cluster have a comparably large density (around 250 points in the sphere). Then the density estimation drops to below 100 points for thinner regions of the data set. For all density estimations with hyperspheres this structural feature can be seen best when the Pareto Radius is used. Emergent SOM (ESOM) construct a nonlinear, topology preserv-

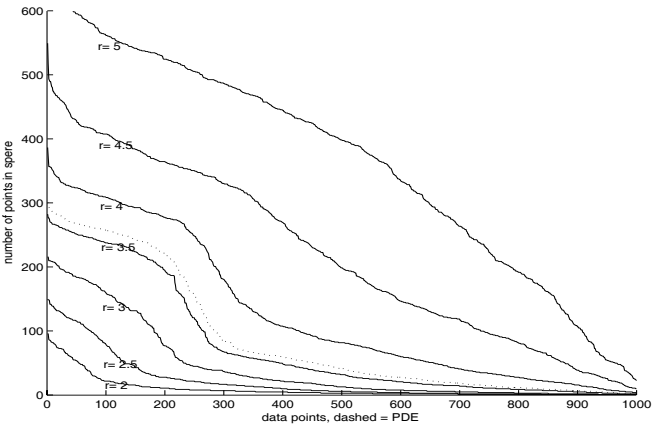


Fig. 4. Hypershere density estimation with different radii

ing mapping from high dimensional spaces to two dimensional grids (the map space)(Kohonen (1989)). A U-Matrix can be constructed on the map

space (Ultsch (2003c)). A U-Matrix displays the local dissimilarity between data points or interpolating points. The same map space can also be used for the display of density relationships. The PDE can be measured for each weight vector of a neuron on the map space. This gives a local density estimation in the input space. The display of these PDE measures as height values on top of the map space is a visualization of the density relationships in the high dimensional space. The properties of the SOM algorithm guarantee that local and global distance relationships between data points are appropriately represented on the map space. Such a display has been called a P-Matrix (Ultsch (2003c)). Figure 5 shows a P-matrix of Eisen's Yeast data

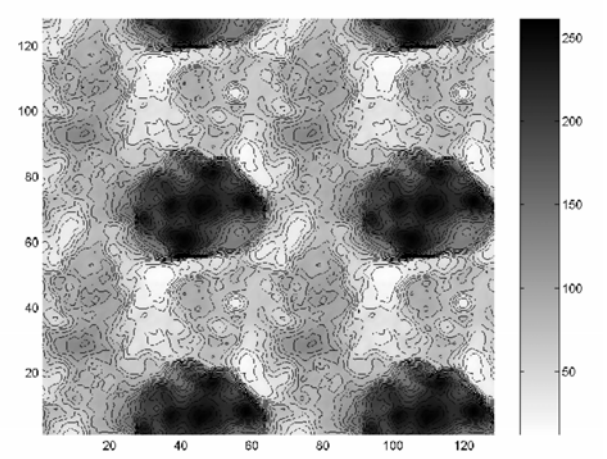
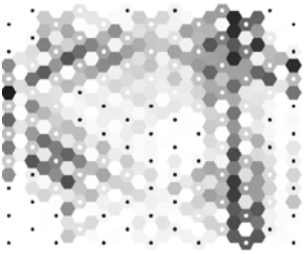


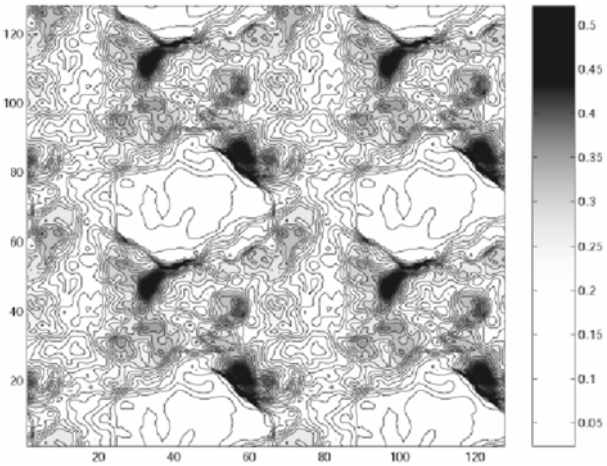
Fig. 5. P-Matrix of DNA microarray data

(Eisen et al. (1998)). The data set contains 2465 data points of 2465 gene expressions of yeast of dimension 79. The data is available from the web site "http://www-genome.stanford.edu". On the P-Matrix it can be seen that a substantial subset of the data points are mapped to locations where there is a big concentration of points. Compare the dark regions in Figure 5. There, the neighborhood numbers are around 400. Other regions, distant from the first have also a local density maximum of more than 250. This points to possible cluster structures. Some regions on the ESOM are also very under-populated. This is an indication for "outliers", i.e. singular special situations in the data set.

P-Matrices can also be used to enhance the visibility of cluster borders in a U-Matrix and to detect clusters in data sets. Figure 6 b shows a so called U*-Matrix which is the combination of a U-Matrix and a P-matrix (Ultsch (2003b)) in comparison to the display published by Kaski et al. (1998) on the same data set. Centering a SOM mapping on the point with highest PDE



(a) Kaski (1998)



(b) U*-Matrix

Fig. 6. Displays of Yeast data

results in a canonical view for SOM with borderless map space such as toroids (Ultsch (2003c)).

5 Summary

One of the goals of data mining is to discover clusters in empirical data. Distances are a prerequisite for the detection of clusters, but sometimes not enough for an automatic clustering. Data density is an alternative viewpoint on the data. Density considerations lead often to better cluster definition. The combination of both methods is hardly attempted. In this work a method for an efficient measurement of data density is presented. Pareto Density Estimation (PDE) is a method for the estimation of density functions using hyper spheres. The radius of the hyper spheres is derived from information optimal sets. The construction of the PDE from an empirical data set takes in particular into account that there might be an unknown number of clusters of

also unknown size in the set. Starting at an educated guess, the information on clusters discovered during the process of data mining can be employed in the method. A tool for the visualization of probability density distributions of variables, the PDEplot is defined. The usefulness of this tool is demonstrated on DNA array data. The visualization guides the search for better models for empirical distributions for this type of data. The usage of PDE to visualize the density relationships of high dimensional data sets leads to so called P-Matrices which are defined on the mapping space of emergent self-organizing maps (ESOM). A P-Matrix for a 79-dimensional DNA array data set is shown. The ESOM mapping preserves the data's topology. The P-Matrix reveals local concentrations of data points. This is a very useful tool in the detection of clusters and outliers in unknown data sets. Pareto Density Estimation, PDEplots for one dimensional data and the construction of P-matrices for high dimensional data have been implemented as MATLAB® routines. These routines may be obtained from the author (http://www.mathematik.uni-marburg.de/∼databionics/).

References

EISEN, M., SPELLMAN, P., BROWN, P.O. and BOTSTEIN, D. (1998): Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America, 95, 14863–14868.*

GENTLEMAN, R. and CAREY, V. (2003): Visualization and Annotation of Genomic Experiments. In: G. Parmigiani, E.S. Garrett, R.A. Irizarry, S.L. Zeger (Eds.): *The Analysis of Gene Expression Data.* Springer, New York, 46–72.

GHOSE, A.K. and CRIPPEN, G.M. (1986): Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity. *J. Comput. Chem., 7, 565–577.*

HINNEBURG, A. and KEIM, D.A. (1998): An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In: *Proc. 4rd Int. Conf. on Knowledge Discovery and Data Mining*, AAAI Press.

KASKI, S., NIKKILA, J. and KOHONEN, T. (1998): Methods for interpreting a self-organized map in data analysis. In: *Proc. 6th European Symposium on Artificial Neural Networks*, Brugges, Belgium.

PARMIGIANI, G., GARRETT, E.S., IRIZARRY, R.A. and ZEGER, S.L. (2003): *The Analysis of Gene Expression Data.* Springer, New York.

ULTSCH, A. (2003): Pareto Density Estimation: A Density Estimation for Knowledge Discovery. In: D. Baier and K.-D. Wernecke (Eds.): *Innovations in Classification, Data Science, and Information Systems*, Springer, Berlin, 91–98.

ULTSCH, A. (2003b): U*Clustering: automatic clustering on Emergent Self Organizing Feature Maps. *Technical Report Nr. 36*, Department of Computer Science University of Marburg.

ULTSCH, A. (2003c): Maps for the Visualization of high-dimensional Data Spaces. In: *Proc. Workshop on Self Organizing Maps WSOM03*, 225–230.

XU, X., ESTER, M., KRIEGEL, H.-P. and SANDER, J. (1998): A Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases. In: *Proceedings of the ICDE Conference.*

Hierarchical Mixture Models for Nested Data Structures

Jeroen K. Vermunt[1] and Jay Magidson[2]

[1] Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, Netherlands
[2] Statistical Innovations Inc., 375 Concord Avenue, Belmont, MA 02478, USA

Abstract. A hierarchical extension of the finite mixture model is presented that can be used for the analysis of nested data structures. The model permits a simultaneous model-based clustering of lower- and higher-level units. Lower-level observations within higher-level units are assumed to be mutually independent given cluster membership of the higher-level units. The proposed model can be seen as a finite mixture model in which the prior class membership probabilities are assumed to be random, which makes it very similar to the grade-of-membership (GoM) model. The new model is illustrated with an example from organizational psychology.

1 Introduction

Social science researchers, as researchers in other fields, are often confronted with nested or hierarchical data structures. Examples are data from employees belonging to the same organizations, individuals living in the same regions, customers of the same stores, repeated measures taken from the same individuals, and individuals belonging to the same primary sampling units in two-stage cluster samples.

This paper introduces an extension of the standard finite mixture model (McLachlan and Peel (2000)) that can take the hierarchical structure of a data set into account. Introducing random-effects in the model of interest is a common way to deal with dependent observations arising from nested data structures. It is well known that the finite mixture model is itself a nonparametric random-effects model (Aitkin (1999)). The solution that is proposed here is to introduce nonparametric random effects within a finite mixture model. That is, on top of a finite mixture model, we build another finite mixture model, which yields a model with a separate finite mixture distribution at each level of nesting.

When using the hierarchical mixture model for clustering, one obtains not only a clustering of lower-level units, but also a clustering of higher-level units. The clusters of higher-level units differ with respect to the prior probabilities corresponding to the lower-level clusters. This is similar to what is done in multiple-group latent class analysis, with the difference that we assume that each group belongs to one of a small number of clusters (latent classes) instead of estimating of a separate latent class distribution for each

group. The latter approach would amount to using a fixed-effect instead of a random-effects model.

Because it is not practical to estimate the hierarchical mixture using a standard EM algorithm, we propose a variant of EM that we call the upward-downward algorithm. This method uses the conditional independence assumption of the underlying graphical model for an efficient implementation of the E step.

2 Model formulation

2.1 Standard finite mixture model

Let y_{ik} denote the response of individual i on indicator, attribute, or item k. The number of cases is denoted by N, and the number of items by K. The latent class variable is denoted by x_i, a particular latent class by t, and the number of latent classes by T. Notation $\mathbf{y}_i$ is used to refer to the full response vector for case i. A finite mixture model can be defined as (McLachlan and Peel (2000))

$$f(\mathbf{y}_i) = \sum_{t=1}^{T} \pi(x_i = t)\, f(\mathbf{y}_i | x_i = t).$$

where $\pi(x_i = t)$ is the prior class membership probability corresponding to class t and $f(\mathbf{y}_i|x_i = t)$ is the class conditional density of $\mathbf{y}_i$. With continuous y_{ik}, we may take $f(\mathbf{y}_i|x_i = t)$ to be multivariate normal. If the indicators y_{ik} are categorical variables, we usually make the additional assumption that responses are independent given class membership (Lazarsfeld and Henry (1968)); that is,

$$f(\mathbf{y}_i|x_i = t) = \prod_{k=1}^{K} \pi(y_{ik}|x_i = t). \tag{1}$$

This assumption is justified if – as in our empirical example – the K items can be assumed to measure a single underlying dimension.

2.2 Hierarchical finite mixture model

For the hierarchical extension of the mixture model, we have to extend our notation to take into account the extra level of nesting. Let y_{ijk} denote the response of lower-level unit i within higher-level unit j on indicator k. The number of higher-level units is denoted by J, the number of lower-level units within higher-level unit j by n_j, and the number of items by K. Notation $\mathbf{y}_{ij}$ is used to refer to the full vector of responses of case i in group j, and $\mathbf{y}_j$ to refer to the full vector of responses for group j.

The latent class variable at the lower level is denoted by x_{ij}, a particular latent class by t, and the number of latent classes by T. The latent class

variable at the higher level is denoted by u_j, a particular latent class by m, and the number of latent classes by M.

The hierarchical mixture model consist of two parts. The first part connects the observations belonging to the same group. It has the following form:

$$f(\mathbf{y}_j) = \sum_{m=1}^{M} \pi(u_j = m)\, f(\mathbf{y}_j | u_j = m)$$
$$f(\mathbf{y}_j | u_j = m) = \prod_{i=1}^{n_j} f(\mathbf{y}_{ij} | u_j = m).$$

As can be seen, groups are assumed to belong to one of M latent classes with prior probabilities equal to $\pi(u_j = m)$ and observations within a group are assumed be mutually independent given class membership of the group. Note that this conditional independence assumption is similar to the assumption of the latent class model for categorical variables (see equation 1).

The second part of the model is similar to the structure of a standard finite mixture model, except for the fact that now we are dealing with $f(\mathbf{y}_{ij}|u_j = m)$ instead of $f(\mathbf{y}_i)$; that is, we have to define a density conditional on the class membership of the higher-level unit. This yields

$$f(\mathbf{y}_{ij} | u_j = m) = \sum_{t=1}^{T} \pi(x_{ij} = t | u_j = m)\ f(\mathbf{y}_{ij} | x_{ij} = t). \tag{2}$$

In the case of categorical y_{ijk}, we will again assume that

$$f(\mathbf{y}_{ij} | x_{ij} = t) = \prod_{k=1}^{K} \pi(y_{ijk} | x_{ij} = t).$$

If we compare the standard mixture model with the hierarchical mixture model, we see two important differences: 1] we not only obtain information on class membership of individuals, but also on class membership of groups and 2] groups are assumed to differ with respect to the prior distribution of their members across lower-level latent classes.

It should be noted that the hierarchical mixture model is a graphical model with a tree structure. The upper node is the discrete latent variable at the higher level. The intermediate nodes consist of the n_j discrete latent variables for the lower-level units belonging to higher-level unit j. These x_{ij} are mutually independent given u_j. The lower nodes contain the observed responses y_{ijk}, which in the latent class model are assumed to be mutually independent given x_{ij}.

3 Maximum likelihood estimation by an adapted EM algorithm

If we put the various model parts together, we obtain the following log-likelihood function for the hierarchical mixture model:

$$\begin{aligned}\log L &= \sum_{j=1}^{J} \log f(\mathbf{y}_j) \\ &= \sum_{j=1}^{J} \log \sum_{m=1}^{M} \pi(u_j = m) \prod_{i=1}^{n_j} \left[\sum_{t=1}^{T} \pi(x_{ij} = t | u_j = m)\ f(\mathbf{y}_{ij} | x_{ij} = t) \right].\end{aligned}$$

A natural way to solve the ML estimation problem is by means of the EM algorithm (Dempster et al. (1977)). The E step of the EM algorithm involves computing the expectation of the complete data log-likelihood, which in the hierarchical mixture model is of the form

$$\begin{aligned}E(\log L_c) = &\sum_{j=1}^{J} \sum_{m=1}^{M} P(u_j = m | \mathbf{y}_j) \log \pi(u_j = m) \\ &+ \sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{i=1}^{n_j} \sum_{x=1}^{T} P(u_j = m, x_{ij} = t | \mathbf{y}_j) \log \pi(x_{ij} = m | u_j = m) \\ &+ \sum_{j=1}^{J} \sum_{m=1}^{M} \sum_{i=1}^{n_j} \sum_{t=1}^{T} P(x_{ij} = t | \mathbf{y}_j) \log f(\mathbf{y}_{ij} | x_{ij} = t).\end{aligned}$$

This shows that, in fact, the E step involves obtaining the posterior probabilities $P(u_j = m, x_{ij} = t | \mathbf{y}_j)$ given the current estimates for the unknown model parameters. In the M step of the EM algorithm, the unknown model parameters are updated so that the expected complete data log-likelihood is maximized (or improved). This can be accomplished using standard complete data algorithms for ML estimation.

The implementation of the E step is more difficult than the M step. A standard implementation would involve computing the joint conditional expectation of the $n_j + 1$ latent variables for higher-level unit j, that is, the joint posterior distribution $P(u_j, x_{1j}, x_{2j}, ..., x_{n_j j} | \mathbf{y}_j)$ with $M \cdot T^{n_i}$ entries. Note that this amounts to computing the expectation of all the "missing data" for a higher-level unit. These joint posteriors would subsequently be collapsed to obtain the marginal posterior probabilities for each lower-level unit i within higher-level unit j. A drawback of this procedure is that computer storage and time increases exponentially with the number of lower-level units, which means that it can only be used with small n_j.

Fortunately, it turns out that it is possible to compute the n_j marginal posterior probability distributions $P(u_j = m, x_{ij} = t | \mathbf{y}_j)$ without going

through the full posterior distribution by making use of the conditional independence assumptions implied by the hierarchical mixture model. In that sense our procedure is similar to the forward-backward algorithm that can be used for the estimation of hidden Markov models with large numbers of time points (Baum et al. (1970)). In the upward-downward algorithm, first, latent variables are integrated out going from the lower to the higher levels. Subsequently, the relevant marginal posterior probabilities are computed going from the higher to the lower levels. This yields a procedure in which computer storage and time increases linearly with the number of lower-level observations instead of exponentially, as would have been the case with a standard EM algorithm.

The upward-downward algorithm makes use of the fact that

$$\begin{aligned} P(u_j = m, x_{ij} = t|\mathbf{y}_j) &= P(u_j = m|\mathbf{y}_j)P(x_{ij} = t|\mathbf{y}_j, u_j = m) \\ &= P(u_j = m|\mathbf{y}_j)P(x_{ij} = t|\mathbf{y}_{ij}, u_j = m); \end{aligned}$$

that is, given class membership of the group (u_j), class membership of the individuals (x_{ij}) is independent of the information of the other group members. The terms $P(u_j = m|\mathbf{y}_{ij})$ and $P(x_{ij} = t|\mathbf{y}_{ij}, u_j = m)$ are obtained as follows:

$$P(x_{ij} = t|\mathbf{y}_{ij}, u_j = m) = \frac{\pi(x_{ij} = t|u_j = m)f(\mathbf{y}_{ij}|x_{ij} = t)}{f(\mathbf{y}_{ij}|u_j = m)}$$

$$P(u_j = m|\mathbf{y}_j) = \frac{\pi(u_j = m)\prod_{i=1}^{n_j} P(\mathbf{y}_{ij}|u_j = m)}{f(\mathbf{y}_j)},$$

where $f(\mathbf{y}_{ij}|u_j = m) = \sum_{t=1}^{T} \pi(x_{ij} = t|u_j = m)f(\mathbf{y}_{ij}|x_{ij} = t)$ and $f(\mathbf{y}_j) = \sum_{m=1}^{M} \pi(u_j = m)\prod_{i=1}^{n_j} P(\mathbf{y}_{ij}|u_j = m)$.

In the upward part, we compute $f(x_{ij} = t, \mathbf{y}_{ij}|u_j = m)$ for each individual, collapse these over x_{ij} to obtain $f(\mathbf{y}_{ij}|u_j = m)$, and use these to obtain $P(u_j = m|\mathbf{y}_j)$ for each group. The downward part involves computing $P(u_j = m, x_{ij} = t|\mathbf{y}_{ij})$ for each individual using $P(u_j = m|\mathbf{y}_i)$ and $P(x_{ij} = t|\mathbf{y}_{ij}, u_j = m)$.

A practical problem in the implementation of the above upward-downward method is that underflows may occur in the computation of $P(u_j = m|\mathbf{y}_j)$. Such underflows can, however, easily be prevented by working on a log scale. The algorithm described here will be implemented in version 4.0 of the Latent GOLD program for finite mixture modeling (Vermunt and Magidson (2000)).

4 An empirical example

We will illustrate the hierarchical mixture model using data taken from a Dutch study on the effect of team characteristics on individual work conditions (Van Mierlo (2003)). A questionnaire was completed by 886 employees

from 88 teams of two organizations, a nursing home and a domiciliary care organization. Of interest for the illustration of the hierarchical mixture model is that employees are nested within (self-managing) teams, where the total number of observations per team ranged from 1 to 22.

Various aspects of work conditions were measured, one of which was the perceived task variety. The item wording of the five dichotomous items measuring perceived task variety is as follows (translated from Dutch):

1. Do you always do the same things in your work?
2. Does your work require creativity?
3. Is your work diverse?
4. Does your work make enough usage of your skills and capacities?
5. Is there enough variation in your work?

We had 36 cases with missing values on one or more of the indicators, but these cases can be retained in the analysis.

The model we use for these dichotomous response variables is an unrestricted latent class models. Besides a latent class model for the employees, we have to take into account the nested data structure. This is done by allowing teams to belong to clusters of teams that differ with respect to the prior distribution of the task-variety classes of employees. An alternative would have been to adopt a fixed-effects approach in which each team has its own prior latent class distribution. However, given the large number of higher-level units (88), this would yield a model with many parameters.

We fitted models with different numbers of classes of teams and different numbers of classes of employees within classes of teams. Table 1 reports the log-likelihood value, the number of parameters, and the BIC value for the estimated models. In the computation of BIC, we used the total number of employees (886) as the sample size. As can be seen, the very parsimonious model with two classes of teams and two classes of employees (within classes of teams) is the preferred model according to the BIC criterion.

Table 1. Testing results for the estimated models with the task-variety data

Teams	Employees	Log-likelihood	# Parameters	BIC value
1-class	1-class	-2797	5	5628
1-class	2-class	-2458	11	4991
1-class	3-class	-2444	17	5004
2-class	2-class	-2435	13	4958
2-class	3-class	-2419	20	4974
3-class	2-class	-2434	15	4970
3-class	3-class	-2417	23	4991

The estimated probability of giving a response that is in agreement with a high task variety ("no" for item 1 and "yes" for the other 4 indicators) equals

.51, .70, .97, .83, and .93 for the employees in the first latent class and .14, .17, .20, .42, and .17 for the second latent class. Thus, the first latent class can be called the high task-variety class and the second the low task-variety class.

Besides these two classes of employees we encountered two clusters of teams that differ in their team members' prior probability of belonging to the high task-variety class. In the first cluster of teams – containing 66% of the teams – this prior probability equals .79, whereas it is only .39 in the second cluster of teams. This shows that there are large differences between teams with respect to the perceived task variety of their employees. It also shows that the observations belonging to the same group are quite strongly correlated.

Whereas in this application the hierarchical structure arises from the nesting of individuals within groups, the proposed methodology is also useful in longitudinal studies: the higher-level units would then be individuals and the lower-level units measurement occasions or time points.

5 Variants and extensions

This paper presented the simplest form of the hierarchical mixture model. Several extensions and variants can be formulated. One important extension is the use of covariates affecting u_j, x_{ij}, or y_{ijk}. For example, assume that we have a set of P covariates affecting x_{ij} and that z_{ijp} denotes a particular covariate. In that case, we may use the following logit form for $\pi(x_{ij} = t|u_j = m, \mathbf{z}_{ij})$:

$$\pi(x_{ij} = t|u_j = m, \mathbf{z}_{ij}) = \frac{\exp(\gamma^m_{t0} + \sum_{p=1}^{P} \gamma_{tp}\, z_{ijp})}{\sum_{r=1}^{T} \exp(\gamma^m_{r0} + \sum_{p=1}^{P} \gamma_{rp}\, z_{ijp})}.$$

In equation (2), we implicitly assumed that u_j has no direct effect on $\mathbf{y}_{ij}$. In some applications one may wish to use an alternative structure for this equation. For example,

$$f(\mathbf{y}_{ij}|u_j = m) = \sum_{t=1}^{T} \pi(x_{ij} = t)\; f(\mathbf{y}_{ij}|x_{ij} = t, u_j = m),$$

which can be used for obtaining a three-level extension of the mixture regression model (see Vermunt (2004)). That is, a nonparametric random-coefficients model in which regression coefficients not only differ across clusters of lower-level units, but also across clusters of higher-level units.

The hierarchical mixture model is similar to the grade-of-membership (GoM) model (Manton et al. (1994)). As pointed out by Haberman (1995) and Esherova (2003), a GoM model can be defined as a latent class model with multiple exchangeable latent variables, which is exactly the same as

is done in the hierarchical mixture model. Whereas we model the variation in prior class membership probabilities by nonparametric random effects, a hierarchical mixture model with parametric random effects would be even more similar to the GoM model. Vermunt (2003) proposed such a variant in which the logit of $\pi(x_{ij} = t|u_j)$ is assumed to be normally distributed, which is the common specification for the random effects in logistic regression models. More specifically,

$$\pi(x_{ij} = t|u_j) = \frac{\exp(\gamma_t + \tau_t \cdot u_j)}{\sum_{r=1}^{T} \exp(\gamma_r + \tau_r \cdot u_j)}$$

with $u_j \sim N(0, 1)$.

Whereas the hierarchical mixture model presented in this paper contains only two levels of nesting, it is straightforward to extend the model and the upward-downward algorithm to three or more levels.

References

AITKIN, M. (1999): A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics, 55, 218–234.*

BAUM, L.E., PETRIE, T., SOULES, G. and WEISS, N. (1970): A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics, 41, 164–171.*

DEMPSTER, A.P., LAIRD. N.M. and RUBIN, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B, 39, 1–38.*

EROSHEVA, E.A. (2003): Partial membership models with application to disability survey data. In: H. Bozdogan (Ed.): *Statistical data mining and knowledge discovery.* Chapman and Hall/CRC, Boca Raton.

HABERMAN, S.J. (1995): Book review of "Statistical Applications Using Fuzzy Sets" by K.G. Manton, M.A. Woodbury and H.D. Dennis. *Journal of the American Statistical Association, 90, 1131-1133.*

LAZARSFELD, P.F. and HENRY, N.W. (1968): *Latent Structure Analysis.* Houghton Mifflin, Boston.

MANTON, K.G., WOODBURY, M.A. and TOLLEY H.D. (1994): *Statistical Applications Using Fuzzy sets.* Wiley, New York.

MCLACHLAN, G.J. and PEEL, D. (2000): *Finite Mixture models.* Wiley, New York.

VAN MIERLO, H. (2003). *Self-managing Teams and Psychological Well-being.* Phd. dissertation. Eindhoven University of Technology, The Netherlands.

VERMUNT, J.K. (2003): Multilevel latent class models. *Sociological Methodology, 33, 213–239.*

VERMUNT, J.K. (2004): An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica, 58, 220–233.*

VERMUNT, J.K. and MAGIDSON, J. (2000): *Latent GOLD 2.0 User's Guide.* Statistical Innovations, Belmont, MA.

Iterative Proportional Scaling Based on a Robust Start Estimator

Claudia Becker

Wirtschaftswissenschaftliche Fakultät,
Martin-Luther-Universität Halle-Wittenberg, 06099 Halle, Germany

Abstract. Model selection procedures in graphical modeling are essentially based on the estimation of covariance matrices under conditional independence restrictions. Such model selection procedures can react heavily on the presence of outlying observations. One reason for this might be that the covariance estimation is influenced by outliers. Hence, a robust procedure to estimate a covariance matrix under conditional independence restrictions is needed. As a first step to robustify the model building process in graphical modeling we propose to use a modified iterative proportional scaling algorithm, starting with a robust covariance estimator.

1 Introduction

Graphical modeling deals with detecting (conditional) independencies between variables. Model selection in this context implies the repeated estimation of covariance matrices under certain conditional independence restrictions. Kuhnt and Becker (2003) show how sensitive such model selection procedures react to the presence of outlying observations in the context of mixed graphical models. As estimating the covariance matrix is an essential part of these procedures it can be assumed that one reason for this sensitivity might lie in the sensitivity of covariance estimation against the influence of outliers. Hence, a robust procedure to estimate a covariance matrix under conditional independence restrictions is needed. In the case of a graphical covariance selection model based on a normal distribution assumption, the covariance estimation is usually performed by means of the so-called iterative proportional scaling algorithm (see Lauritzen (1996)), which uses the classical empirical covariance as a starting estimate and hence will inherently be nonrobust in nature. As a first step to robustify the model building process in graphical modeling we therefore propose to start the iteration with a robust covariance estimator like the minimum covariance determinant (MCD) estimator (Rousseeuw (1985), Rousseeuw and van Driessen (1999)) instead.

In the following section we briefly introduce the general setting of covariance selection models. Section 3 deals with the original version of the iterative proportional scaling algorithm while the new robustified approach is introduced and discussed in Section 4. In Section 5 we investigate the integration of the robustified approach into model selection strategies. We finish with some concluding remarks and put open questions arising from the presented results.

2 Covariance selection models

Graphical models are used to determine and at the same time visualize dependency structures between random variables (Cox and Wermuth (1996), Edwards (2000), Lauritzen (1996), Whittaker (1990)). A statistical model for a random vector $\mathbf{X} = (X_1, ..., X_p)^T$ is presented as the combination of a distributional assumption and a mathematical graph, where the vertices of the graph represent the variables $X_1, \ldots, X_p$. The dependency structure between the variables is related to the existence or non-existence of edges between the vertices. We restrict ourselves to the case of undirected graphs here, where edges are undirected lines, and the relationship expressed by an edge is mutual. Figure 1 shows an example of an undirected graph G with vertices $\{1, \ldots, 5\}$ representing random variates $X_1, \ldots, X_5$.

For a p-dimensional random vector $\mathbf{X} = (X_1, ..., X_p)^T$ and an undirected graph G with vertices $\{1, \ldots, p\}$ a graphical independence model consists of all distributions of $\mathbf{X}$, for which X_i and X_j are conditionally independent given all other variables, whenever G does not contain an edge between vertices i and j. As a special case, a covariance selection model additionally imposes a normality assumption, namely $\mathbf{X} \sim N(\mu, \Sigma)$. For the graph of Figure 1, we conclude that e.g. X_1 is conditionally independent of X_4 given (X_2, X_3, X_5).

In covariance selection models, the conditional independencies given by the graph are reflected in the inverse $K = \Sigma^{-1}$ of the covariance matrix Σ in the way that missing edges yield zeroes in K (Lauritzen (1996, p. 129)). For the graph of Figure 1, K would be of the following type:

$$K = \begin{bmatrix} k_{11} & k_{12} & k_{13} & 0 & 0 \\ k_{21} & k_{22} & k_{23} & 0 & 0 \\ k_{31} & k_{32} & k_{33} & k_{34} & k_{35} \\ 0 & 0 & k_{43} & k_{44} & k_{45} \\ 0 & 0 & k_{53} & k_{54} & k_{55} \end{bmatrix},$$

where K^{-1} has to be a valid covariance matrix, i.e. K^{-1} positive definite and symmetric.

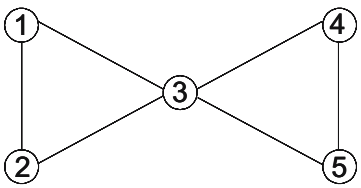


Fig. 1. Example of a graph.

3 Iterative proportional scaling (IPS)

Assume that for a data set $\mathbf{x}_1, \ldots, \mathbf{x}_n$ under investigation the graph structure is given, hence the conditional independencies are known. One can then estimate Σ with the maximum likelihood (ML) approach, restricted upon the zeroes in K according to the missing edges in the graph G. The solution of the ML equations cannot be given in closed form, but there exists an algorithm to compute the solution, the so-called iterative proportional scaling (IPS) algorithm (Lauritzen (1996), Speed and Kiiveri (1986)):

start estimate:	K_0 with zeroes according to G, K_0^{-1} valid covariance matrix (e.g. $K_0 = \mathcal{I}_p$, the $(p \times p)$-identity matrix)
iteration:	$K_{r+1} = (T_{C_1} \ldots T_{C_k})K_r$,
where	$T_C K_r = K_r + [n(SSD_{CC})^{-1} - (K_{r,CC}^{-1})^{-1}]^G$,
	$SSD = \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T$.

Here, C denotes a clique of G, a subset of maximum size of the vertices of G such that all vertices in C are joined by an edge. For a $(p \times p)$-matrix M the operation $M_{CC} = [m_{ij}]_{i,j \in C}$ means to extract the objects of M according to a clique C, whereas $[M_{CC}]^G$ stands for the opposite operation, namely to write the elements of M_{CC} back into a $(p \times p)$-matrix according to the positions of the clique elements. Each iteration step is performed over the set of all cliques $C_1, \ldots, C_k$ of G. Lauritzen (1996) shows that the series $\{K_r\}$ converges to the ML estimate $\widehat{K}$ of K. Inverting $\widehat{K}$ yields $\widehat{\Sigma}$.

As an example consider the data on the examination marks of $n = 88$ students in the $p = 5$ subjects mechanics (1), vectors (2), algebra (3), analysis (4), and statistics (5), as given in Mardia et al. (1979) and analyzed by Edwards (2000). The graph of Figure 1 was considered appropriate to reflect the dependency structure of these variables. Given this graph, the IPS algorithm results in

$$\widehat{K} = \begin{bmatrix} 0.005 & -0.002 & -0.003 & 0 & 0 \\ -0.002 & 0.011 & -0.006 & 0 & 0 \\ -0.003 & -0.006 & 0.029 & -0.008 & -0.005 \\ 0 & 0 & -0.008 & 0.010 & -0.002 \\ 0 & 0 & -0.005 & -0.002 & 0.007 \end{bmatrix},$$

corresponding to

$$\widehat{\Sigma} = \begin{bmatrix} 298.9 & 124.3 & 99.3 & 98.6 & 107.2 \\ 124.3 & 168.9 & 83.2 & 82.7 & 89.9 \\ 99.3 & 83.2 & 110.3 & 109.6 & 119.1 \\ 98.6 & 82.7 & 109.6 & 215.4 & 152.0 \\ 107.2 & 89.9 & 119.1 & 152.0 & 291.0 \end{bmatrix}.$$

If we now replace the first observation $\mathbf{x}_1 = (77, 82, 67, 67, 81)^T$ of the data set by $(\mathbf{770}, 82, 67, 67, 81)^T$, by this imposing a massive outlier into the data,

IPS comes up with

$$\widehat{K} = \begin{bmatrix} \mathbf{0.0001} & \mathbf{-0.0003} & \mathbf{-0.0001} & 0 & 0 \\ \mathbf{-0.0003} & 0.010 & -0.007 & 0 & 0 \\ \mathbf{-0.0001} & -0.007 & 0.028 & -0.008 & -0.005 \\ 0 & 0 & -0.008 & 0.010 & -0.002 \\ 0 & 0 & -0.005 & -0.002 & 0.007 \end{bmatrix},$$

corresponding to

$$\widehat{\Sigma} = \begin{bmatrix} \mathbf{6225.3} & \mathbf{368.9} & \mathbf{226.9} & \mathbf{225.4} & \mathbf{245.0} \\ \mathbf{368.9} & 168.9 & 83.2 & 82.7 & 89.9 \\ \mathbf{226.9} & 83.2 & 110.3 & 109.6 & 119.1 \\ \mathbf{225.4} & 82.7 & 109.6 & 215.4 & 152.0 \\ \mathbf{245.0} & 89.9 & 119.1 & 152.0 & 291.0 \end{bmatrix}.$$

Similar to other situations, the ML estimates of K and Σ under the restrictions given by G turn out to be not at all robust against disturbances of the data.

4 IPS robustified

To overcome the problem illustrated in the previous section, we propose to insert a robust estimator into the IPS algorithm:

start estimate:	K_0 with zeroes according to G, K_0^{-1} valid covariance matrix (e.g. $K_0 = \mathcal{I}_p$)
iteration:	$K_{r+1} = (T_{C_1} \dots T_{C_k})K_r$,
where	$T_C K_r = K_r + [m(MCD_{CC})^{-1} - (K_{r,CC}^{-1})^{-1}]^G$,
	$MCD = \sum_{i=1}^n w_i(\mathbf{x}_i - \mathbf{t}_{MCD})(\mathbf{x}_i - \mathbf{t}_{MCD})^T$,
	$m = \sum_{i=1}^n w_i$.

In this algorithm, $\mathbf{t}_{MCD}$ denotes the MCD location estimate according to Rousseeuw (1985). Observations whose Mahalanobis-type distance with respect to $\mathbf{t}_{MCD}$ and the corresponding raw MCD covariance estimate is too large ($> (\chi^2_{p;0.975})^{1/2}$) get weight $w_i = 0$, all other obserations get weight $w_i = 1$. Hence, up to a constant, MCD is the one-step reweighted MCD covariance estimate (Rousseeuw (1985), Rousseeuw and van Driessen (1999)). We call this modified IPS algorithm robustified iterative proportional scaling (RIPS).

In the students' marks example RIPS yields

$$\widehat{K}_{RIPS} = \begin{bmatrix} 0.007 & -0.002 & -0.006 & 0 & 0 \\ -0.002 & 0.011 & -0.007 & 0 & 0 \\ -0.006 & -0.007 & 0.041 & -0.008 & -0.007 \\ 0 & 0 & -0.008 & 0.013 & -0.003 \\ 0 & 0 & -0.007 & -0.003 & 0.008 \end{bmatrix},$$

for the undisturbed data which is similar to the IPS solution, which is also true for $\widehat{\Sigma}_{RIPS}$:

$$\widehat{\Sigma}_{RIPS} = \begin{bmatrix} 224.5 & 93.8 & 77.6 & 73.7 & 97.0 \\ 93.8 & 148.2 & 63.7 & 60.5 & 79.6 \\ 77.6 & 63.7 & 76.4 & 72.6 & 95.6 \\ 73.7 & 60.5 & 72.6 & 153.9 & 127.8 \\ 97.0 & 79.6 & 95.6 & 127.8 & 262.3 \end{bmatrix}.$$

Contrary to IPS, its robustified counterpart is not as much influenced by the outlier, ending up with

$$\widehat{K}_{RIPS} = \begin{bmatrix} 0.007 & -0.002 & -0.006 & 0 & 0 \\ -0.002 & 0.011 & -0.007 & 0 & 0 \\ -0.006 & -0.007 & 0.040 & -0.008 & -0.006 \\ 0 & 0 & -0.008 & 0.013 & -0.003 \\ 0 & 0 & -0.006 & -0.003 & 0.008 \end{bmatrix},$$

which yields

$$\widehat{\Sigma}_{RIPS} = \begin{bmatrix} 209.8 & 80.1 & 71.2 & 66.9 & 85.0 \\ 80.13 & 137.7 & 58.3 & 54.8 & 69.6 \\ 71.2 & 58.3 & 74.3 & 69.8 & 88.7 \\ 66.9 & 54.7 & 69.8 & 151.5 & 119.8 \\ 85.0 & 69.6 & 88.7 & 119.9 & 245.0 \end{bmatrix}.$$

From the results of this example, RIPS appears to be a sensible approach to robustify the covariance estimation in graphical covariance selection models.

5 Model selection with RIPS

Since usually the dependency structure and the graph corresponding to a data set is not known beforehand, model selection procedures are applied. Their goal is to find an appropriate dependency structure, which is as sparse as possible while at the same time consistent with the data. We will now investigate how the RIPS based estimators can be integrated into such procedures.

Model selection strategies in graphical modeling include several approaches like forward or backward selection, full model search based on information criteria, or the so-called Edwards-Havranek procedure (see Edwards (2000) for a detailed review). We investigate the backward selection procedure based on the deviance difference. This method starts with the saturated model corresponding to the complete graph, where any two vertices are joined by an edge. From this full model, claiming no conditional independencies, edges are successively removed by performing likelihood ratio tests of the actual model against all possible models containing one edge less. The edge whose removal leads to the largest p-value in the tests ("least

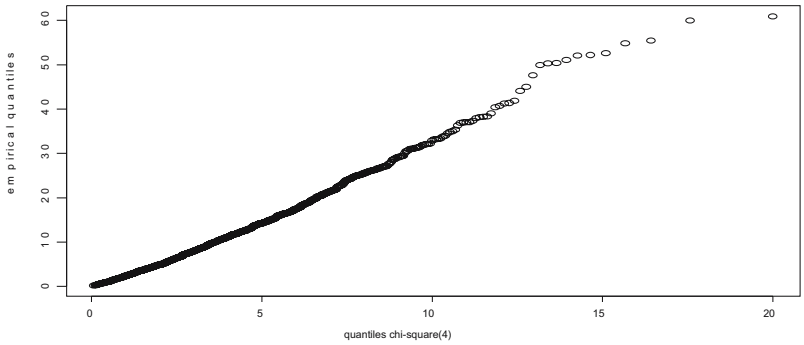


Fig. 2. Distribution of robustified D^2: qq-plot of χ^2 distribution.

significant edge") is then taken out, and the model is updated. This is continued until there are no more removable edges, meaning that all p-values lie below a given test level α. The test statistic is the deviance difference $D^2 = n(\ln\det\widehat{K}_{M_1} - \ln\det\widehat{K}_{M_0})$ between two successive models M_0 and M_1, $M_0 \subseteq M_1$, where M_0 and M_1 differ by one edge. Here, $\widehat{K}_M$ denotes the ML estimate of K under model M. The deviance difference can also more generally be used to test between two arbitrary models M_0 and M_1 which may differ by more than one edge. The test statistic is asymptotically χ^2 distributed with degrees of freedom equal to the difference in the number of edges between the two models.

Since the distributional result is based on ML estimation, if we use the RIPS based estimators instead of the ML estimators within the procedure, we have to check whether still the same χ^2 distribution can be used. Up to now, there do not exist any theoretical results. We performed some simulations to get a first impression. Figure 2 illustrates the results in an exemplary manner. The data of size $n = 1000$ were generated according to a covariance selection model reflected by G from Figure 1. The statistic D^2 was calculated for the test against the saturated model, where the RIPS based estimators for K were used. The two models differ by four edges, hence the original version of D^2 would follow a χ^2_4 distribution. Figure 2 shows the corresponding qq-plot of the distribution of our D^2. It seems that the distribution type is appropriate. Similar results came out with further simulations for other models. Detailed analyses of the histograms of the distributions of the new D^2 together with χ^2 densities revealed that rescaling the test statistic is necessary to get a sufficient approximation. We use the square root of a χ^2 quantile as the rescaling factor, which is motivated by the construction of the MCD estimator in the RIPS algorithm. Figure 3 shows the result for the data of Figure 2. Hence, we propose to use the RIPS based estimators $\widehat{K}_{RIPS}$ with test statistic

$$D^2_{RIPS} = \frac{n}{\sqrt{\chi^2_{r;0.95}}}(\ln\det\widehat{K}_{RIPS,M_1} - \ln\det\widehat{K}_{RIPS,M_0}),$$

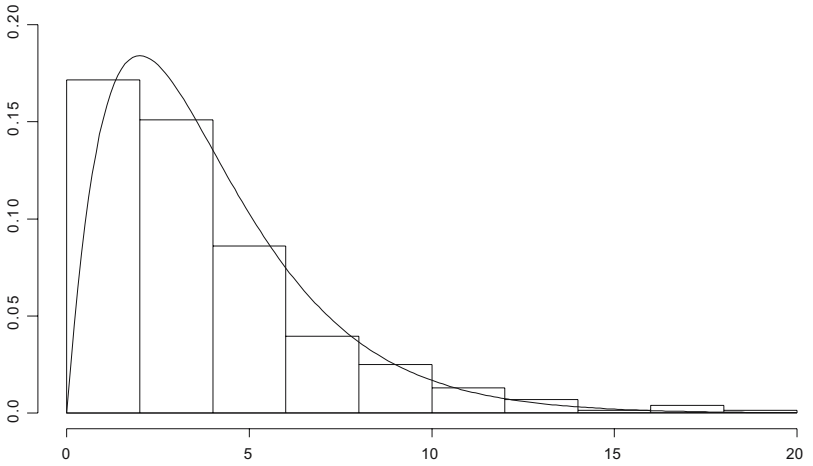


Fig. 3. Distribution of rescaled robustified D^2: histogram and density of χ^2.

where r denotes the number of edges different between models M_0 and M_1. Critical values can then be taken from χ^2_r. In backward selection we have $r = 1$.

To give an impression of the proposed procedure, we performed a small simulation study. We took a covariance selection model according to G from Figure 1 as basic model (model 0). Data sets of size $n = 1000$ were generated according to this model. We further considered three simple disturbances by replacing the first component of the first (model 1), the first 10 (model 2), and the first 25 (model 3) observations by a large value, in this way generating 0.1%, 1%, and 2.5% outliers, respectively. For each data set, the usual and the proposed RIPS based backward selection were performed with α chosen to be 5%. We generated 500 simulation runs for each model. The results are presented in Table 1. In the left part we see, how often the model

Table 1. Simulation results for model selection

	true model found in ... out of 500 runs		too many edges removed in ... out of 500 runs	
	ML backward	RIPS backward	ML backward	RIPS backward
model 0	364	145	0	0
model 1	381	151	0	0
model 2	250	159	191	0
model 3	38	153	435	0

selection procedures selected exactly the correct model. Obviously, the ML based procedure is much better under the null model and slight disturbances. However, already with 2.5% of extremely deviating observations the RIPS

based model selection beats ML. Moreover, as can be seen in the right part of the table, in outlier situations ML based backward selection tends to remove too many edges, hence claiming conditional independencies which do not hold. Here, RIPS is clearly advantageous.

6 Open questions

In this paper, we proposed a new approach for robustifying model selection in graphical covariance selection models. Up to now, there do not exist any theoretical results about this approach, but first experiences in a small simulation study confirm that it is worth investigating further. Still, there is a bunch of interesting research topics in this field like

- Does the RIPS algorithm converge, and to which solution?
- Can the asymptotic χ^2 distribution of D^2_{RIPS} be shown?
- Would some different test statistic be better?
- Is the MCD a good estimator in RIPS? What about other robust covariance estimators?
- How do we measure the robustness of model selection in graphical modeling? Is "counting edges" enough?

References

COX, D.R., and WERMUTH, N. (1996): *Multivariate Dependencies*. Chapman & Hall, London.

EDWARDS, D. (2000): *Introduction to Graphical Modelling*. 2nd ed. Springer, New York.

KUHNT, S. and BECKER, C. (2003): Sensitivity of Graphical Modeling Against Contamination. In: M. Schader, W. Gaul and M. Vichi (Eds.): *Between Data Science and Applied Data Analysis*. Springer, Berlin, 279–287.

LAURITZEN, S.L. (1996): *Graphical Models*. Clarendon Press, Oxford.

MARDIA, K.V., KENT, J.T. and BIBBY, J.M. (1979): *Multivariate Analysis*. Academic Press, London.

ROUSSEEUW, P.J. (1985): Multivariate Estimation with High Breakdown Point. In: W. Grossmann, G. Pflug, I. Vincze and W. Wertz (Eds.): *Mathematical Statistics and Applications*. Reidel, Dordrecht, 283–297.

ROUSSEEUW, P.J. and VAN DRIESSEN, K. (1999): A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics, 41*, 212–223.

SPEED, T.P. and KIIVERI, H. (1986): Gaussian Markov Distributions over Finite Graphs. *Annals of Statistics, 14*, 138–150.

WHITTAKER, J. (1990): *Graphical Models in Applied Mathematical Multivariate Statistics*. Wiley, Chichester.

Exploring Multivariate Data Structures with Local Principal Curves

Jochen Einbeck, Gerhard Tutz, and Ludger Evers

Institut für Statistik, Ludwig-Maximilians-Universität München,
Akademiestr.1, D-80799 München, Germany

Abstract. A new approach to find the underlying structure of a multidimensional data cloud is proposed, which is based on a localized version of principal components analysis. More specifically, we calculate a series of local centers of mass and move through the data in directions given by the first local principal axis. One obtains a smooth "local principal curve" passing through the "middle" of a multivariate data cloud. The concept adopts to branched curves by considering the second local principal axis. Since the algorithm is based on a simple eigendecomposition, computation is fast and easy.

1 Introduction

Principal components analysis (PCA) is a well established tool in dimension reduction. For a set of data $\mathbf{X} = (X_1, \ldots, X_n)^T$ with X_i in $\mathbb{R}^d$ the principal components provide a sequence of best linear approximations to that data. Specifically, let Σ be the empirical covariance matrix of $\mathbf{X}$, then the principal components decomposition is given by

$$\Sigma = \Gamma \Lambda \Gamma^T \tag{1}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is a diagonal matrix containing the ordered eigenvalues of Σ, with $\lambda_1 \geq \ldots \geq \lambda_d$, and Γ is an orthogonal matrix. The columns of $\Gamma = (\gamma_1, \ldots, \gamma_d)$ are the eigenvectors of Σ. The first eigenvector γ_1 maximizes the variance of $\mathbf{X}\gamma$ over all $\gamma \in \mathbb{R}^d$ with $||\gamma|| = 1$, the second eigenvector γ_2 maximizes the variance of $\mathbf{X}\gamma$ over all $\gamma \in \mathbb{R}^d$ with $||\gamma|| = 1$ which are orthogonal to γ_1, and so on. For illustration, we consider the location of scallops near the NE coast of the United States (Fig. 1; the data are included in the S+ SpatialStats Package). The first and second principal component axes, $g_j(t) = \mu + t\gamma_j \qquad (j = 1, 2,\ t \in \mathbb{R})$, with $\mu = \frac{1}{n}\sum_{i=1}^n X_i$, are also depicted. The principal axes unveil nicely the main directions in which the scallops spread out: the first from SW to NE, and the second from NW to SE. In the data cloud we clearly see two fields of scallops: one along the first principal axis and the other one along the second principal axis. The crossing of the axes is not positioned at the junction of the fields, since the default centering in PCA is at the over-all-center of mass of the data. Intuitively, one might determine the position of the crossing of the two fields by the point on the first principal axis where the spread on the second principal axis is maximal.

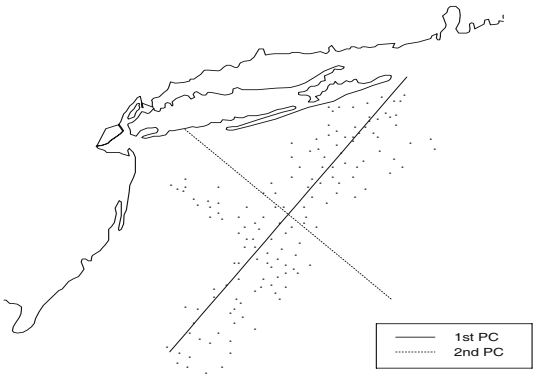


Fig. 1. First and second principal component through scallops near the NE coast of the USA.

In the following, we will go one step further and abandon the assumption of linearity, i.e. not only linear structures shall be described, but any form of multivariate curvaceous, possibly branched, connected or disconnected data structures. The goal is to find smooth nonparametric *local principal curves* passing through a data cloud. Therefore, it can be seen as a competitor to the principal curve algorithms from Hastie and Stuetzle (1989), Tibshirani (1992), Kégl et al. (2000), and Delicado (2001). Only the latter one is also based on the concept of localization. However, Delicado does not use local principal components, but rather local *principal directions*, which however cannot be calculated by a simple eigendecomposition. Principal directions are defined as vectors orthogonal to the hyperplane that locally minimize the variance of the data points projected on it. For a comparison of the principal curve algorithms we refer to Einbeck et al. (2003).

2 Local principal curves

Assume a data cloud $\mathbf{X} = (X_1, \ldots, X_n)^T$, where $X_i = (X_{i1}, \ldots, X_{id})^T$. We propose the following algorithm to find the local principal curve passing through $\mathbf{X}$:

Algorithm 1 (Local principal curves)

1. *Choose a set $S_0 \neq \emptyset$ of starting points. This may be done randomly, by hand, or by choosing the maximum/maxima of a kernel density estimate.*
2. *Draw without replacement a point $x_0 \in S_0$. Set $x = x_0$.*
3. *Calculate the local center of mass*

$$\mu^x = \frac{\sum_{i=1}^n K_H(X_i - x)X_i}{\sum_{i=1}^n K_H(X_i - x)}$$

at x, where $K_H(\cdot)$ is a $d-$dimensional kernel function and H a multivariate bandwidth matrix. Denote by μ_j^x the j-th element of μ^x.

4. *Estimate the local covariance matrix* $\Sigma^x = (\sigma^x_{jk})$ *at* x *via*

$$\sigma^x_{jk} = \sum_{i=1}^{n} w_i (X_{ij} - \mu^x_j)(X_{ik} - \mu^x_k)$$

with weights $w_i = K_H(X_i - x) / \sum_{i=1}^n K_H(X_i - x)$*, and* H *as in step 3. Let* γ^x *be the first column of the loadings matrix* Γ^x *computed locally at* x *in analogy to equation (1).*

5. *Update* x *by setting*

$$x := \mu^x + t_0 \gamma^x,$$

where t_0 *determines the step length.*

6. *Repeat steps 3 to 5 until the border of the data cloud is reached. This is the case when the sequence of* μ^x *remains approximately constant. Then set again* $x = x_0$*, set* $\gamma^x := -\gamma^x$ *and continue with step 5.*

7. *Repeat steps 2 to 6 as long as the set* S_0 *is not empty.*

The local principal curve (LPC) is given by the sequence of the μ^x. Note that, in step 5, one has to make sure that the orientation of the local eigenvector $\gamma^x_{(i)}$ after a number i of loops is the same as the local eigenvector $\gamma^x_{(i-1)}$ one loop before, and has to change its signum if $\gamma^x_{(i-1)} \circ \gamma^x_{(i)} < 0$, where $\circ$ denotes the scalar product.

In the sequel, we will extend the algorithm and look at local principal components of higher order. Let the term "k-th local eigenvalue" denote the k-th largest eigenvalue of Σ^x. The $k-$th local eigenvalues λ^x_k $(k \geq 2)$ are useful indicators for branching points.

Definition 1 (Branches of order θ and depth ϕ).

- The order θ of a branch of a LPC is the order of the local principal component which launched it. In other words, $\theta = k$ means that this branch of the LPC was induced by the k-th local eigenvalue. LPC's according to Algorithm 1 lead for all $x_0 \in S_0$ to a branch with $\theta = 1$.
- The depth ϕ of a branch is the number of junctions (plus 1) between the starting point and the branch. Thus, a branch of depth $\phi = \ell$ ($\ell \geq 2$) is launched by a high k-th ($k \geq 2$) local eigenvalue on a branch with $\phi = \ell - 1$. Algorithm 1 always yields curves of depth $\phi = 1$.
- Denote the maximum values of θ and ϕ used to construct a LPC by θ_{max} and ϕ_{max}, resp.

Obviously, $\theta_{max} = 1$ implies $\phi_{max} = 1$; $\theta_{max} \geq 2$ implies $\phi_{max} \geq 2$; and vice versa. The case $\theta_{max} \geq 3$ might be interesting for highdimensional and highly branched data structures. However, for the most applications it should be sufficient to have only one possible bifurcation at each point. Thus, we extend Algorithm 1 only to the case $\phi_{max} \geq 2, \theta_{max} = 2$:

Algorithm 2 (LPC with $\phi_{max} \geq 2, \theta_{max} = 2$)

Let $0 \leq \rho_0 \leq 1$ be a suitable constant, e.g. $\rho_0 = 0.5$.

1. *Construct a local principal curve α according to Algorithm 1. Compute the relation*
$$\rho^x = \frac{\lambda_2^x}{\lambda_1^x}$$
for all points x which were involved in the construction of α.
2. *Iterate for all $\phi = 2, \ldots, \phi_{max}$:*
   (a) *Let $\zeta_1, \ldots, \zeta_m$ denote all points x belonging to branches of depth $\phi - 1$ with $\rho^x > \rho_0$. If this condition is fulfilled for a series of neighboring points, take only one of them.*
   (b) *Iterate for $j = 1, \ldots, m$:*
      i. *Compute the second local eigenvector $\gamma_2^{\zeta_j}$.*
      ii. *Set $x := \mu^{\zeta_j} + 2t_0\gamma_2^{\zeta_j}$ and continue with Algorithm 1 at step 3. Afterwards, set $x := \mu^{\zeta_j} - 2t_0\gamma_2^{\zeta_j}$ and continue with Algorithm 1 at step 3.*

The factor 2 employed in 2.b.ii) for the construction of starting points of higher depth shall prevent that branches of second order fall immediately back to the branch of first order. In order to avoid superfluous or artificial branches one can apply a very simple form of pruning: If starting points of depth $\phi \geq 2$ fall in regions with negligible density, simply dismiss them.

3 Simulated data examples

The performance of the method shall be illustrated by means of some simulated examples. Our simulated data clouds resemble letters, keeping in mind that the recognition of hand-written characters is a possible application of principal curves (Kegl and Krzyzak (2002)). We consider noisy data structures in the shape of a "C", "E", and "K". In all examples, the set of starting points S_0 contains only one element x_0 which was chosen randomly. For the "C" only one branch of depth $\phi = 1$ is needed and thus Algorithm 1 is applied. The letter "E" requires to compute branches of depth up to $\phi = 2$. The letter "K" is even a little more complicated, and depending on the position of x_0 one needs branches of depth $\phi = 2$ or $\phi = 3$. Table 1 shows the setting of the simulation, and the parameter values used in the algorithm. We apply a bandwith matrix $H = h^2 \cdot I$, where I is the 2-dimensional identity matrix.

Table 1. Parameters for simulation and estimation of characters.

	Simulation		Estimation			
	σ	n	ϕ_{max}	θ_{max}	$h = t_0$	ρ_0
"C"	0.01	60	1	1	0.1	–
	0.1	60	1	1	0.15	–
"E"	0.01	100	2	2	0.1	0.4
	0.1	100	2	2	0.1	0.4
"K"	0.01	90	3	2	0.08	0.4
	0.07	90	2	2	0.15	0.4

The results are depicted in Fig. 2. The large amount of tuning parameters might give the impression that finding an appropriate curve might be quite cumbersome. In practice, however, there is only one crucial smoothing parameter: the bandwidth h. The parameter t_0 has certainly to be chosen as well, but it turned out to be a sensible choice setting it equal to the bandwidth. The parameters θ_{max} and ϕ_{max} depend directly on the data structure. The parameter ρ_0 does not play any role when $\theta_{max} = 1$, and will usually be situated in the small interval between 0.3 and 0.6. We illustrate the detection of branching points by means of the "E" with small noise. Fig. 3 shows the flow of the second local eigenvalue starting from the right bottom end of the "E" and rising to the right top end of it. One sees that the peaks are distinct and well localized, and thus useful for the detection of a bifurcation.

4 Real data examples

We return to the scallops example from the introduction. From the structure of these data it is immediately clear that one needs curves of second order

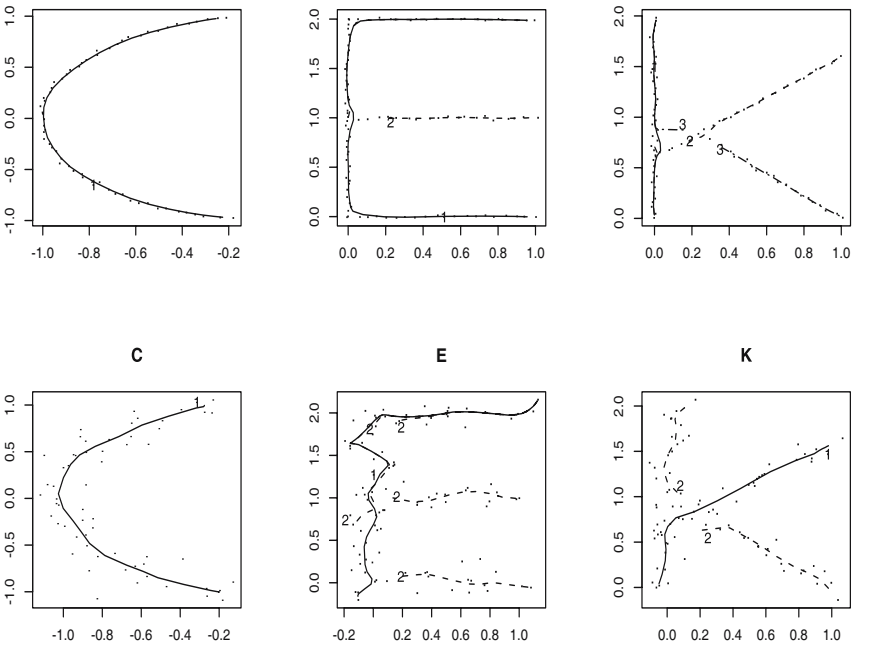


Fig. 2. LPC through letters with small (top) and large noise (bottom). Data points are depicted as "·". Branches of depth $\phi = 1$ are symbolized by a solid line, branches of depth $\phi = 2$ by a dashed line, and branches of depth $\phi = 3$ by a dashed-dotted line. The numbers indicate the starting points for branches of the corresponding depth.

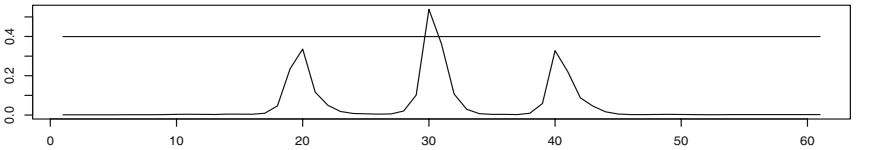


Fig. 3. Flow diagram of $\rho^x = \lambda_2^x/\lambda_1^x$ from the right bottom to the right top of the "E" with small noise. The horizontal line symbolizes the threshold $\rho_0 = 0.4$.

and depth, i.e. $\theta_{max} = 2$ and $\phi_{max} = 2$. Fig. 4 shows that the results of Algorithm 2 can differ for different starting points. Thus, it is natural to ask what the constructed curves represent. Scallops are known to like shallow ocean water. This suggests that the resulting local principal curves follow the ridges of underwater mountains. This hypothesis is confirmed by looking at contour plots from that area (Fig. 4 right). Obviously the left one of the two pictures represents nicely the underwater ridges: One small one from NW to SE (corresponding to the branch with $\phi = 2$), and one larger one from SW to NE (corresponding to the branch with $\phi = 1$). Certainly, the gap between

the two branches is not a real feature but is due to the factor 2 employed in step 2.b.ii) in Algorithm 2.

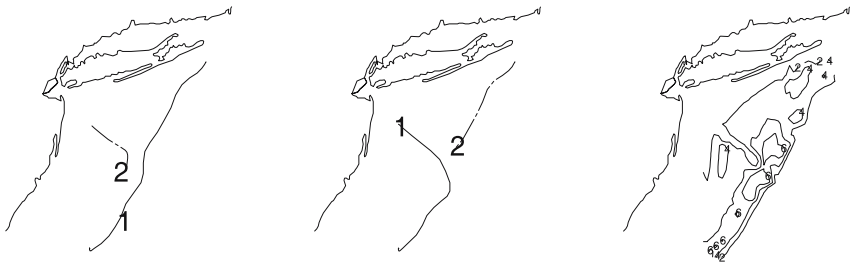


Fig. 4. Left, Middle: LPC of scallops data with bandwidth $h = 0.15$. Branches of depth $\phi = 1$ are launched by starting points "1" and branches of depth $\phi = 2$ start at points "2". Right: Contour plot of underwater plateaus. The numbers indicate the depth: High numbers mean shallow water. In all three pictures the NE coast line of the USA is plotted for orientation.

The scallops data are highly noisy, but not very far from linearity. We will provide one more real data example with data having small noise, but having a very complex nonparametric structure. The data are coordinates of European coastal resorts (taken from Diercke (1984)). Suppose one wants to reconstruct the European coast line given these sites. The European coast does not have mentionable ramifications, thus we use $\phi_{max} = \theta_{max} = 1$, but choose 10 starting points randomly. A typical result is shown in Fig. 5. Taking into account that Algorithm 1 does not have the notion about the shape of Europe that humans have, the coast is reconstructed nicely, although it failed to describe areas with very few data, as Albania, and highly chaotic regions as Schleswig-Holstein and Southern Denmark.

5 Conclusion

We demonstrated that local principal components can be effectively used to explore the structure of multivariate complex data structures. The method is especially useful for noisy spatial data as frequently met in geostatistics. The next step should be to reduce the dimensionality of the predictor space in a multivariate regression or classification problem by employing the local principal curve as low-dimensional, but highly informative predictor.

References

DELICADO (2001): Another Look at Principal Curves and Surfaces. *Journal of Multivariate Analysis, 77, 84–116.*

DIERCKE (1984): *Weltatlas.* Westermann, Braunschweig.

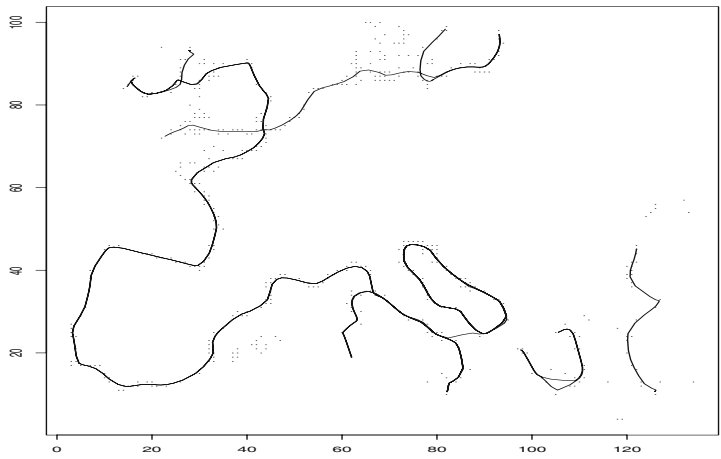


Fig. 5. LPC (solid line) through European coastal resorts (·). The positions are given on a digitalized 101×135 grid, and the applied bandwidth is $h = 2$, meaning that about 2 digits in each direction are considered for construction of the curve.

EINBECK, J., TUTZ, G. and EVERS, L. (2003): Local Principal Curves. *SFB386 Discussion Paper No. 320*, LMU München.

HASTIE, T. and STUETZLE, L. (1989): Principal Curves. *JASA, 84, 502–516.*

KÉGL, B. and KRZYZAK, A., (2002): Piecewise Linear Skeletonization using Principal Curves. *IEEE Trans. Patt. Anal. Mach. Intell., 24, 59–74.*

KÉGL, B., KRZYZAK, A., LINDER, T. and ZEGER, K. (2000): Learning and Design of Principal Curves, *IEEE Trans. Patt. Anal. Mach. Intell., 22, 281–297.*

TIBSHIRANI, R. (1992): Principal Curves Revisited. *Statistics and Computing, 2 , 183–190.*

A Three-way Multidimensional Scaling Approach to the Analysis of Judgments About Persons

Sabine Krolak–Schwerdt

Department of Psychology, Saarland University, 66041 Saarbrücken, Germany

Abstract. Judgments about persons may depend on (1) how coherently person attributes are linked within the stimulus person and (2) how strongly the given person information activates a social stereotype. These factors may determine the number of judgment dimensions, their salience and their relatedness. A three-way multidimensional scaling model is presented that measures these parameters and their change across stimulus persons or judgment conditions. The proposed approach involves a formal modelling of information integration in the judgment process. An application to experimental data shows the validity of the model.

1 Introduction

This paper is concerned with cognitive structures which underly judgments about persons and the presentation of a formal account to model judgment processes. Social judgments play a central role in the professional as well as private everyday life. Frequently, these judgments contribute to far reaching decisions about persons. Examples are eye witness testimonies at the courtyard, medical expert judgments or decisions about job applicant candidates. The question arises what a formal account of judgments should offer to adequately model the nature of judgments. In the following, this issue will be investigated and a corresponding data model will be presented. Subsequently, the usefulness of the proposed approach will be shown in an application to experimental judgment data.

2 The structure of judgments about persons

Social cognition research has shown that the structure of judgments may be described by a dimensional representation. That is, judgments consist of graduating stimuli along a number of continua (Anderson and Sedikides (1991)). Examples of judgment dimensions are evaluation or agreeableness of persons (Schneider (1973); McCrae and Costa (1985)). In a number of judgment conditions people make judgments based on all of the relevant information, weighted and combined into a dimension by an algebraic integration principle (Anderson (1974); Fishbein and Ajzen (1975)). For example, if a person is assessed as agreeable, each piece of information is checked according to its relevance and consistency to agreeableness. Subsequently, the information pieces are combined or integrated, either by adding them all up or by taking their average.

In this case, the mental representation used for making judgments depends on the diversity and coherence of the encountered person attributes (Asch and Zukier (1984)). Thus, cognitive structures underlying judgments reflect the co–occurrence of attributes within the stimulus. If the attribute information exhibits a consistent pattern of co–occurrences, a coherent and simple pattern of associations between attributes may be extracted resulting in a set of correlated judgment dimensions. Conversely, rather diverse person attributes lacking a conceivable pattern of co–occurrences cause a representation composed of independent dimensions.

Under many conditions, however, people use heuristic strategies that depart from a deliberate integration of each piece of information (Gigerenzer and Todd (1999)). One such strategy is to search through a small subset of the information cues and to base one's judgment on a few number of cues. The effect on the judgmental system is that only a small number of dimensions receive sufficient attention and will be integrated within the cognitive structure whereas others will be ignored. Thus, different attribute dimensions receive different weights of importance or saliences. Another strategy is to use a category or stereotype in the process of judging the target person. Stereotypes are cognitive structures that contain people's belief about person attributes and they involve illusory correlations of category membership and specific person attributes (Leyens, Yzerbyt and Schadron (1994)). As an example, females categorized as 'business women' are believed to be self–assertive and not agreeable. Thus, in referring to stereotypes, judgment dimensions which are truly independent domains may become correlated.

To adequately model the outlined nature of judgments, a formal account must incorporate the following components: (1) In deriving dimensions from judgment data, a formal account should reflect the combination principle of information integration. (2) The model has to specify parameters for the salience and relatedness of dimensions in each judgment condition. (3) The model has to provide statements about changes in these parameters across different conditions. The model to be presented in the following, termed 'SUMM–ID', was designed to adopt these requirements.

3 'SUMM–ID' model

Our approach belongs to the class of three–way multidimensional scaling techniques which were developed to model individual differences in the representation of proximities. In the following, the scalar product form of the approach will be outlined. The input data to the model are assumed to consist of a three–way data matrix $X = (x_{ijj'})$, $i = 1, \ldots, I$, $j, j' = 1, \ldots, J$, where I is the number of individuals or conditions and J the number of attributes. X can be thought of as comprising a set of $I(\geq 2)J \times J$ scalar products matrices. X_i, a slice of the three–way matrix, consists of scalar products between attributes j, j' for an individual or a condition i. The scalar products may derive from cognitive associations between the attributes.

The basic equation of the approach can be expressed as

$$X_i = BH_iB' + E_i \ , \tag{1}$$

where B is a $J \times P$ matrix specifying an attribute space or judgment configuration which is common to all individuals or conditions where P is the number of dimensions. H_i is a $P \times P$ symmetric matrix designating the nature of individual $i's$ representation of the judgment dimensions. Diagonal elements h_{ipp} of H_i correspond to weights applied to the judgment dimensions by individual i, while off–diagonal elements $h_{ipp'}$ are related to perceived relationships among the judgment dimensions p and p'. As Equation (1) shows, the matrix H_i, termed *'individual characteristic matrix'* (Tucker (1972)), transforms the common judgment space into the individual representation, and E_i collects the errors of approximation $e_{ijj'}$.

Thus, the model assumes that there is a common space represented by matrix B which underlies judgments in general. On the basis of the common space, the model allows for two kinds of distortions in individual representations. The first is that individuals may attach different weights to different judgment dimensions. The second is that individual representations may be rotated versions of the common space in which independent dimensions become correlated.

The final decomposition of the data matrix, as outlined in Equation (1), is accomplished in two steps. The first step is to implement the combination principle of information integration as well as to introduce parameters for individual weights. The second step is to optimize the representation as to the dimensionality of the judgment space and to introduce differential rotations of the common judgment dimensions.

In order to implement the combination principle in the first step, it is necessary to introduce a comparatively large number F of judgment dimensions (these will be condensed into an optimal number P in the second step). The judgment dimensions will be termed b_f in the following, $f = 1, \ldots, F$, and a corresponding set of dimensions a_f will be introduced representing weights for the individuals or conditions. The central feature is to base b_f on the introduction of sign vectors z_f for the attributes j, $z_{jf} \in \{-1, 1\}$, and, in an analogous way, to base a_f on sign vectors s_f for individuals i, $s_{if} \in \{-1, 1\}$, where

$$\sum_i \sum_j \sum_{j'} s_{if} z_{jf} z_{j'f} x_{ijj'} = \gamma_f := max \ . \tag{2}$$

In the equation, the scalar γ_f is a normalizing factor and the vector z_f indicates if an attribute j (j' respectively) consistently pertains to the judgment dimension b_f (in this case $z_{jf} = 1$) or if the attribute conveys information which is inconsistent to the judgment domain (in this case $z_{jf} = -1$). The vector z_f thus reflects the consistency or inconsistency of the attributes to the judgment domain f. In an analogous way, s_f indicates how individuals should be combined to derive the common judgment dimension.

The vectors z_f and s_f are the basis for the determination of dimension a_f and b_f in the following way:

$$\begin{aligned} a_{if} &= \tfrac{1}{\sqrt[3]{\gamma_f{}^2}} \quad u_{if} \ , \ \text{where } u_{if} = \textstyle\sum_j \sum_{j'} z_{jf} z_{j'f} x_{ijj'} \ , \\ b_{jf} &= \tfrac{1}{\sqrt[3]{\gamma_f{}^2}} \quad v_{jf} \ , \ \text{where } v_{jf} = \textstyle\sum_i \sum_{j'} s_{if} z_{j'f} x_{ijj'} \ , \\ & \qquad\qquad \text{and} \quad \gamma_f = \textstyle\sum_i s_{if} u_{if} = \sum_j z_{jf} v_{jf} \ . \end{aligned} \tag{3}$$

After the determination of a_f and b_f in this way, the procedure continues in computing the residual data $x^*_{ijj'} = x_{ijj'} - a_{if} b_{jf} b_{j'f}$ and repeating the extraction of dimensions according to (2) and (3) on the residual data until a sufficient amount of the variation in the data is accounted for by the representation. Dimensions b_f may be considered as those dimensions which are used for the combination of attributes in order to generate a judgment. Values a_{if} of the individuals i on the dimension a_f indicate how salient judgment dimension b_f is under the point–of–view of individual i.

As Equation (2) shows, the values of the sign vectors z_f and s_f are determined in such a way that the sum of the data values collected in a dimension will be maximized. To compute the sign vectors, an algorithm is used that alternates between the subject and the attribute mode. That is, given the sign vector of one mode, multiplying the data matrix (the residual data matrix, respectively) with this sign vector yields the signs of the other vector. This is due to the cyclical relation between the modes

$$\begin{aligned} \textstyle\sum_i \sum_j \sum_{j'} s_{if} z_{jf} z_{j'f} x_{ijj'} &= \textstyle\sum_i s_{if} u_{if} \ = \sum_i \mid u_{if} \mid \\ &= \textstyle\sum_j z_{jf} v_{jf} = \sum_j \mid v_{jf} \mid = \gamma_f \ , \end{aligned}$$

which also motivates the definition of the normalizing factor γ_f in Equation (3). The relation between the modes is used in the SUMM–ID algorithm by iteratively fixating one vector and estimating the other vector until the values of both sign vectors have stabilized. The procedure was first introduced by Orlik (1980).

After extraction of F dimensions the three–way data matrix is given by

$$x_{ijj'} = \sum_{f=1}^{F} a_{if} b_{jf} b_{j'f} + e_{ijj'} \ . \tag{4}$$

Arranging the values of an individual i in the derived dimensions in a diagonal $F \times F$ matrix A_i yields a matrix formulation

$$X_i = B_F A_i B'_F + E_i \tag{5}$$

where B_F is a $J \times F$ matrix with the dimensions $b_f, f = 1, \ldots, F$ as columns which represents the common judgment space and A_i reflects the dimensional weights under the point–of–view of individual i. Applying the individual weights to the common judgment space B_F yields the individual judgment space, $Y_i = B_F A_i^{\frac{1}{2}}$, which corresponds to differentially weighing the importance of the common judgment dimensions.

However, the representation obtained in step 1 has the drawback that the dimensions in the common judgment space are oblique or even linear dependent and thus do not provide a parsimonious description of the data. To derive a more compact and non–redundant representation the second step of the model involves a rotation of the common judgment space

$$\tilde{B} := B_F T \ , \tag{6}$$

where the orthonormal rotation matrix T evolves from the eigenvector–eigenvalue decomposition $B'_F B_F = T \Delta T'$. The rotated judgment space is inserted into the model equation from step 1 in the following way:

$$\begin{array}{rcccl} X_i & = & B_F & A_i & B'_F & +E_i \\ & = & B_F T & (T' A_i T) & T' B'_F & +E_i \\ & = & \tilde{B} & \tilde{H}_i & \tilde{B}' & +E_i \end{array}$$

Finally, omitting columns of $\tilde{B}$ with an eigenvalue of zero or near zero and corresponding rows and columns of $\tilde{H}_i$ yields the final representation that was introduced in Equation (1) in which the set of F oblique dimensions is condensed into P orthogonal and substantial dimensions in B and H_i.

The following relations to other three–way models should be noted. The explication of SUMM–ID step 1 in Equation (5) is formally equivalent to the INDSCAL model (Carroll and Chang (1970)). Thus, both methods fit the same model, but SUMM–ID step 1 finds the estimates by a particular algorithm which optimizes the representation of hypotheses regarding the data involving binary contrasts whereas INDSCAL optimizes the algorithmic fitting procedure. This has some consequences for empirical validity. Whereas INDSCAL has a considerable tendency towards degenerate solutions, this problem is avoided in SUMM–ID by introducing constraints which correspond better to the data. Furthermore, as compared to the INDSCAL algorithm the SUMM–ID estimates are very easy to obtain. The Tucker (1972) model and SUMM–ID step 2 have the introduction of the individual characteristic matrix H_i in common. In SUMM–ID step 2, H_i evolves from the product of a rotation matrix and dimensional weights with a_{if} as diagonal values and $t'_f t_{f'}$ as off–diagonal values. In contrast to the core matrix from the Tucker model, this offers a more straightforward interpretation in terms of saliences and perceived relationships between judgment dimensions, while the Tucker core matrix is confounded with other information such as variances of the judgment dimensions (cf. Kroonenberg (1983)).

Statements about changes in salience and correlations of judgment dimensions may be obtained by comparing the individual characteristic matrices H_i across individuals or conditions. This will be demonstrated in the following application of the SUMM–ID model to experimental judgment data.

4 Application

In an experiment on social judgments subjects received descriptions about persons as stimulus materials. The experiment had a factorial design where the first factor was stereotype activation. That is, in one condition a stereotype was activated by the description, while the other condition pertained to a non–stereotypical (individual) stimulus person. The second factor introduced coherence of person attributes with an unconnected list of statements about a person lacking consistency of attributes in one condition and a coherent description introducing patterns of co–occurences of person attributes in the other condition. After having read one of the descriptions, subjects' (number of subjects = 40) task was to judge the personality of the described person. Subjects received 16 rating scales for the rating task. The scales corresponded to the following judgment domains which have been shown as fundamental in social cognition (Schneider (1973); McCrae and Costa (1985)): (a) agreeableness (e.g., 'materialistic – idealistic'), (b) self-assertiveness (e.g., 'selfless – egoistic'), (c) evaluation (e.g., 'good – bad') and (d) dynamism (e.g., 'active – passive').

The data were preprocessed in the following way. For each of the four experimental conditions, the scales were standardized and data were aggregated across the ten subjects in each condition. Thus, $I = 4$ and $J = 16$ in this example. Subsequently, the distances between the scales were computed. Distances were transformed into scalar products by Torgerson's formula (cf. Carroll and Chang (1970)) and then subjected to the proposed scaling model.

We expected the following results. If the model is valid, then a common judgment space should occur which consists of four dimensions, each corresponding to one of the fundamental domains outlined above. Distortions of this space due to experimental conditions should be reflected in the individual characteristic matrix by diminished saliences h_{ipp} in the diagonal and increased correlations between judgment dimensions $h_{ipp'}$ in the off–diagonal due to the activation of a stereotype and the introduction of a coherent text description. The most extreme distortions should occur in the characteristic matrix for the stereotypical text condition.

The SUMM–ID solution accounted for 90% of the variance of the data after having extracted 16 dimensions in step 1. Thus, our account showed an excellent recovery of the data. In the following, the SUMM–ID step 2 representation is reported. In the common judgment space, after rotation four dimensions were obtained as substantial. Thus, $F = 16$ and $P = 4$ in this data analysis. To test the hypothesis that these correspond to the reported judgment domains, a procrustes analysis was conducted where the obtained dimensions were rotated to optimal agreement with the fundamental domains outlined above. Congruence coefficients (cr) were in the range between 0.85 and 0.91. We found that the first dimension corresponds to agreeableness (cr = 0.91) , the second reflects self-assertiveness to a certain degree (cr = 0.85), the third resembles evaluation (cr = 0.87) and the fourth is dynamism (cr

Table 1. Individual characteristic matrices of the SUMM–ID representation of the experimental judgment data.

Experimental		**Judgment dimensions:**			
condition:		Agreeable.	Self–Assert.	Evaluation	Dynamism
individual, list	Agreeableness	1.17			
	Self–Assertion	0.36	1.43		
	Evaluation	-0.33	-0.23	1.01	
	Dynamism	0.26	0.00	0.21	1.38
individual, text	Agreeableness	1.04			
	Self–Assertion	0.30	0.83		
	Evaluation	0.12	-0.34	1.15	
	Dynamism	-0.59	-0.31	-0.69	1.24
stereotypical, list	Agreeableness	1.13			
	Self–Assertion	0.44	0.84		
	Evaluation	-0.34	-0.41	1.01	
	Dynamism	-0.91	-0.18	-0.67	1.01
stereotypical, text	Agreeableness	0.68			
	Self–Assertion	1.00	0.46		
	Evaluation	-0.14	-0.75	1.32	
	Dynamism	-0.73	-0.13	-0.80	0.97

= 0.89). The individual characteristic matrices are shown in Table 1. The diagonal values h_{ipp} indicate that in the individual listwise description all four judgment dimensions are salient, whereas the activation of a stereotype and the introduction of a text each reduce the salience of the second dimension. The most extreme condition is the stereotypical text description, where the salience of all dimensions is reduced except of dimension 3 and this is evaluation.The off–diagonal values $h_{ipp'}$ show the following. As compared to the individual listwise description, the presentation of a stereotype and of a coherent text each yields substantial correlations of the dynamism dimension with the others. Again, the most extreme condition is the stereotypical text description, where a number of substantial correlations between all dimensions evolves.

5 Concluding remarks

In conclusion, then, the proposed scaling approach reflects the structure of the common judgment space and the expected distortions in every detail. That is, the parameters of the model were sensitive to manipulations of coherence and stereotype activation. The model shows (1) how the combination principle of information integration may be implemented and (2) how saliences and perceived relationships between judgment domains may be formally represented. Advantages of our approach compared to others are that the introduction of constraints in terms of binary contrasts helps to avoid degeneracy kwown

from the INDSCAL technique and that the introduction of conceptual principles concerning hypotheses on the data also yield an interpretation of the individual characteristic matrix which is improved over the one from Tucker's model.

At the outset, the approach was introduced as a data model for the analysis of judgments about persons. However, the basic principles of the model, that is information integration and judgment distortions by salience and illusory correlations, are very general principles of human judgment. As another example, research in the attitude domain (Fishbein and Ajzen (1975)) has shown that these principles underly the formation and use of attitudes.

Thus, the scope of the proposed model is not restricted to the social cognition domain, but may offer a more general formal modelling of judgment processes. Other domains of relevance may be expert judgments, or product judgments in marketing, or attitudes.

References

ANDERSON, N.H. (1974): Cognitive algebra: Integration theory applied to social attribution. In: L. Berkowitz (Ed.): *Advances in Experimental Social Psychology, Vol. 7*. Academic Press, New York, 1–100.

ANDERSON, C.A. and SEDIKIDES, C. (1991): Thinking about people: Contributions of a typological alternative to associationistic and dimensional models of person perception. *Journal of Personality and Social Psychology, 60, 203–217*.

ASCH, S.E. and ZUKIER, H. (1984): Thinking about persons. *Journal of Personality and Social Psychology, 46, 1230–1240*.

CARROLL, J.D. and CHANG, J.J. (1970): Analysis of individual differences in multidimensional scaling via an N–way generalization of Eckart–Young decomposition. *Psychometrika, 35, 283–319*.

FISHBEIN, M. and AJZEN, I. (1975): *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison–Wesley, Reading, MA.

GIGERENZER, G. and TODD, P.M. and the ABC Research Group (1999): *Simple heuristics that make us smart*. Oxford University press, New York.

KROONENBERG, P.M. (1983): *Three–mode principal component analysis*. DSWO Press, Leiden.

LEYENS, J.P., YZERBYT, V. and SCHADRON, G. (1994): *Stereotypes and social cognition*. Sage, London.

McCRAE, R.R. and COSTA, P.T. (1985): Updating Norman's "adequate taxonomy": Intelligence and personality dimensions in natural language and in questionnaires. *Journal of Personality and Social Psychology, 49, 710–721*.

ORLIK, P. (1980): Das SUMMAX–Modell der dreimodalen Faktorenanalyse mit interpretierbarer Kernmatrix. *Archiv für Psychologie, 133, 189–218*.

SCHNEIDER, D.J. (1973): Implicit personality theory: A review. *Psychological Bulletin, 79, 294–309*.

TUCKER, L.R. (1972): Relations between multidimensional scaling and three–mode factor analysis. *Psychometrika, 37, 3–27*.

Discovering Temporal Knowledge in Multivariate Time Series

Fabian Mörchen and Alfred Ultsch

Data Bionics Research Group
University of Marburg, D-35032 Marburg, Germany

Abstract. An overview of the *Time Series Knowledge Mining* framework to discover knowledge in multivariate time series is given. A hierarchy of temporal patterns, which are not a priori given, is discovered. The patterns are based on the rule language *Unification-based Temporal Grammar*. A semiotic hierarchy of temporal concepts is build in a bottom up manner from multivariate time instants. We describe the mining problem for each rule discovery step. Several of the steps can be performed with well known data mining algorithms. We present novel algorithms that perform two steps not covered by existing methods. First results on a dataset describing muscle activity during sports are presented.

1 Introduction

Many approaches in time series data mining concentrate on the compression of univariate time series (patterns) down to a few temporal features. The aim is often to speed up the search for *known* patterns in a time series database (see Hetland (2004) for an overview). The introduced techniques for time series abstraction and the accompanying similarity measures can often be used in other contexts of data mining and knowledge discovery, e.g. for searching *unknown* patterns or rules. Most rule generation approaches search for rules with a known consequent, that is some unknown pattern predicting a predefined event (Povinelli (2000)). In addition, the form of the possible patterns is often restricted by rule language syntax (see Hetland and Saetrom (2002) for a discussion). Very few approaches search for rules with an unknown antecedent part *and* an unknown consequent part (Saetrom and Hetland (2003), Höppner (2001)). Finally, few publications explicitly consider multivariate time series (Höppner (2002)).

Knowledge Discovery is the mining of previously unknown rules that are useful, understandable, interpretable, and can be validated and automatically evaluated (Ultsch (1999)). It is unlikely that one method will maintain good results on all problem domains. Rather, many data mining techniques need to be combined for this difficult process. In Guimaraes and Ultsch (1999) some early results of understandable patterns extracted from multivariate times series were presented. Here, we want to describe our new hierarchical time series rule mining framework Time Series Knowledge Mining (TSKM).

The rest of this paper is structured as follows. The data from sports medicine is described in Section 2. The temporal concepts expressible by the

rule language are explained in Section 3 using examples from the application. Section 4 defines the steps of the framework and gives details on two novel algorithms. The merits of the application, possible extensions of our work, and related methods are discussed in Section 5. Section 6 summarizes the paper.

2 Data

The TSKM method is currently applied to a multivariate time series from sports medicine. Three time series describe the activity of the leg muscles during In-Line Speed Skating measured with surface EMG (Electromyography) sensors. The current leg position is described by three angle sensors (Electrogoniometer), attached at the ankle, the knee, and the hip. Finally, there is a time series produced by an inertia switch, indicating the first ground contact.

3 Unification-based Temporal Grammar

The Unification-based Temporal Grammar (UTG) is a rule language developed especially for the description of patterns in multivariate time series (Ultsch (1996)). Unification-based Grammars are an extension of context free grammars with side conditions. They are formulated with first order logic and use unification. The UTG offers a hierarchical description of temporal concepts. This opens up unique possibilities in relevance feedback during the knowledge discovery process and in the interpretation of the results. An expert can focus on particularly interesting rules and discard valid but known rules before the next level constructs are searched. After obtaining the final results, an expert can zoom into each rule to learn about how it is composed and what it's meaning and consequences might be.

At each hierarchical level the grammar consists of semiotic triples: a unique symbol (syntax), a grammatical rule (semantic), and a user defined label (pragmatic). The unique symbols can be generated automatically during the mining process. The labels should be given by a domain expert for better interpretation. Due to lack of space we will only briefly describe the conceptual levels of the hierarchy (see also Figure 1) along with an example from the application. The basic ideas of the UTG were developed in Ultsch (1996) and applied in Guimaraes and Ultsch (1999). For a detailed description see Ultsch (2004).

A *Primitive Pattern* is a temporal atom with unit duration. It describes a state of the time series at the smallest time scale. For the muscle activity we found 3 to 5 states corresponding to subsets of *very low*, *low*, *medium*, *high*, and *very high*. For the leg position six typical sport movement phases, namely *stabilization*, *forward gliding*, *pre-acceleration* (of center of gravity), *preparation* (of foot contact), *foot placement*, *push-off*, and *leg swing* were

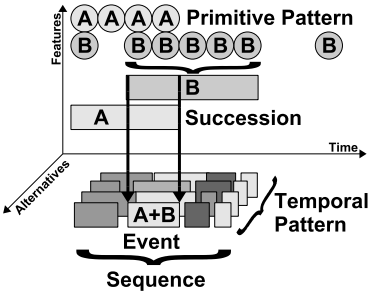


Fig. 1. UTG concepts

```
An Event is a 'Weight Transfer' if
  'Foot contact' is 'on'
coincides with
  'Movement' is 'glide'
coincides with
  'Medial Gastrocnemius' is 'high'
coincides with
  'Vastus Medialis' is 'high'
coincides with
  'Gluteus Maximus' is 'high'
.
```

Fig. 2. UTG Event rule

identified. The labeling (pragmatic) needs to be done in close cooperation with an expert to ensure meaningful results.

A *Succession* introduces the temporal concept of duration. It represents a time interval where nearly all time points have the same Primitive Pattern label. Short interruptions (*Transients*) of an otherwise persisting state should be removed. For the muscle activity interruptions shorter than 50ms were discarded because a change of the state at this time scale is physiologically not plausible. The movement phases are much longer in general, interruptions up to 100ms were removed.

An *Event* represents the temporal concept of coincidence. It represents a time interval where several Successions overlap. If the start points of all overlapping Succession are approximately equal and the same is true for the end points, the Event is called *synchronous*. The Events present in the skating data relate the current movement phase and the position of the foot to the activation of the muscles *at the same time*. One Event in the skating data corresponded to all three muscles being highly active during the forward gliding phase with the foot on the ground. This Event was labeled by the expert as the weight transfer from one leg to the other (see Figure 2). The five most frequent Events were labeled by the expert as follows: *active gliding* (G), *relaxation* (R), *anticipation* (A), *weight transfer* (W), *initial gliding* (I).

A *Sequence* introduces the temporal concept of order. A Sequence is composed of several Events occurring sequentially, but not necessarily with meeting end and start points. The three most frequent Sequences were (G,R,A), (G,R,A,W), and (G,R,A,W,I). They all have the same prefix (G,R,A) corresponding to the *contraction & relaxation* phase. The Events W and I complete the typical skating motion cycle, but are not always recognized due to measurement errors in the foot contact sensor.

A *Temporal Pattern* is the summary of several Sequences by allowing a set of Events at some positions of the pattern. Temporal Patterns represent the non-temporal concept of alternative. Since all Sequences were quite similar in this application, they were merged into a single Temporal Pattern describing the *typical motion cycle* during Inline-Speed Skating.

4 Time Series Knowledge Mining

The temporal knowledge discovery framework Time Series Knowledge Mining (TSKM) aims at finding interpretable symbolic rules describing interesting patterns in multivariate time series. We define the data models, mining task, and algorithms for each level of the framework. The levels correspond to the temporal concepts of the UTG and include some additional steps (see Figure 5. Some tasks can be solved with well known data mining algorithms, while other require new algorithms.

Aspects: The starting point of the TSKM is a multivariate time series, usually but not necessarily uniformly sampled. An expert should divide the features of the time series into possibly overlapping groups that are related w.r.t. the investigated problem domain. Each feature subset is called an Aspect and should be given a meaningful name. In the absence of such prior knowledge, one Aspect per time series can be used. Each Aspect is treated individually for the first steps of the framework.

Preprocessing and feature extraction techniques are applied to each Aspect or even each time series individually. This is a highly application dependent step.

Finding Primitive Patterns: The task of finding Primitive Patterns is the reduction of the time series to a series of states. The input data is a real or vector valued time series, the output is a time series of symbols for each atomic time interval. It is important, that each symbol is accompanied by a rule and a linguistic description to complete the semiotic triple.

Many discretization techniques can be used to find Primitive Patterns for univariate Aspects. Simple methods aggregate the values using histograms. Additionally down-sampling can be performed by aggregation over a time window, e.g. Lin et al. (2003). The symbols for the bins can easily be mapped to linguistic descriptions like *high* or *low*. A first order description method describes the current trend of a time series, e.g. Kadous (1999). Second order descriptions additionally incorporate the second derivative of the signal to distinguish convex from concave trends, e.g Höppner (2001).

For Aspects spanning several time series we propose to use clustering and rule generation on the spatial attributes. If the process alternates between several regimes or states, these regions should form clusters in the high dimensional space obtained disregarding the time attribute. In Guimaraes and Ultsch (1999) and for the identification of the skating movement phases Emergent Self-Organizing Maps (ESOM)(Ultsch (1999)) have been used to identify clusters. The rules for each cluster were generated using the Sig* Algorithm (Ultsch (1991)). The ESOM enables visual detection of outliers and arbitrarily shaped clusters and Sig* aims at understandable descriptions of the Primitive Patterns.

Finding Successions: The input data for finding Successions is a univariate symbolic time series of Primitive Patterns, the output consist of a univariate series of labeled intervals. The merging of consecutive Primitive Pat-

```
i := 2
while i < n
    // check symbols and duration
    if (s_{i-1} = s_{i+1}) and (d_i ≤ d_max)
    and (d_i ≤ r_max * (d_{i-1} + d_{i+1}))
        // merge 3 intervals
        d_i - 1 := Σ_{j=i-1}^{i+1} d_j
        ∀k ∈ {i, i+1} d_k := 0
        i := i + 2
    else
        i := i + 1
    end if
end while
// remove zero durations
S := S \ {(t_i, d_i, s_i) ∈ S | d_i = 0}
```

Fig. 3. SequentialTransientFilter

```
i := 2
while i < n
    s := i
    // search end
    while S(i)=S(i-1)
        i := i + 1
    end while
    // check duration
    if i - s ≥ min_d
        add Event on [s, i - 1]
    end if
    i := i + 1
end while
```

Fig. 4. FullEvents

terns into a Succession is straight forward. But with noisy data there are often interruptions of a state (*Transients*). Let a Succession interval be a triple of a start point t, a duration d, and a symbol s. Let the input Successions be $S = \{(t_i, d_i, s_i)\ i = 1..n\}$ with $t_i + d_i \leq t_{i+1}$ and $s_i \neq s_{i+1}$. Let d_{max} be the maximum absolute duration and r_{max} the maximum relative duration of a Transient. For the removal of Transients we propose the *SequentialTransientFilter* algorithm shown in Figure 3.

The time complexity of the algorithm is $O(n)$. A good choice for r_{max} is 0.5, i.e. the gap is allowed to be at most half as long as the surrounding segments together. The d_{max} parameter has to be chosen w.r.t to the application. Often, some knowledge on the minimum duration of a phenomena to be considered interesting is available.

Finding Events: Events represent the concept of coincidence, thus in this step all Aspects are considered simultaneously. The input data is a multivariate series of labeled intervals (Successions) and the output is a univariate series of labeled intervals (Events). Let S be a $k \times n$ matrix containing the symbols of the Successions from k Aspects at n time points. We use $S(i)$ for the i-th column of S and $S(i) = S(j)$ for element-wise equality. Let min_d be the minimum duration of an Event. The algorithm *FullEvents* shown in Figure 4 discovers all Events where Successions from all Aspects coincide.

The time complexity of the algorithm is $O(n)$. The d_{max} parameter can be chosen similar the maximum duration of Transients when finding Successions. The post-processing to identify synchronous Events is rather straight forward. For each Event the maximum difference between all start points of the participating Successions are checked against a threshold and the same is done for the end points. Additionally, the SequentialTransientFilter algorithm can be applied to the resulting Event series.

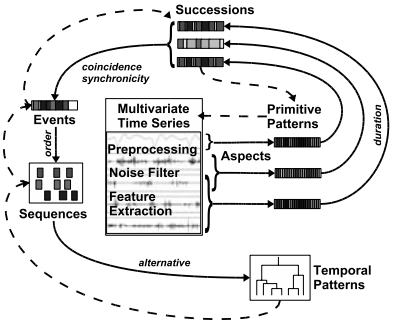


Fig. 5. The TSKM process

Finding Sequences: For the step of finding Sequences there is a large number of algorithms that could be utilized. The input data is a univariate series of labeled intervals and the output is a set of subsequences thereof. For moderately sized dataset we use a suffix trie (e.g. Vilo (1998)). Compared to a suffix tree, all egde labels in a trie have length one. Tries are larger, but can easier be queried for patterns with wild-cards. The trie stores all subsequences up to a maximum length and can be queried with frequency and length thresholds to find the most interesting patterns. For larger datasets more scalable and robust techniques from sequential pattern mining, e.g. Yang et al. (2002), can be used.

Finding Temporal Patterns: The Sequences often overlap. The last step of the framework is finding generalized Sequences, called Temporal Patterns. We propose to use clustering based on a string metric to find groups of similar Sequences. The Temporal Pattern can be generated by merging the patterns in a cluster using groups of symbols at positions where the patterns do not agree. We have successfully used hierarchical clustering based on the string edit distance with a dendrogram visualization.

5 Discussion

The Temporal Pattern found in the skating data provided new insights for the expert. The symbolic representation offers better interpretation capabilities on the interactions of different skeletal muscles than the raw EMG data. We identified the most important cyclical motion phases. The rule describing this phase can be expanded to provide more details. At the level of Temporal Patterns there is a Sequence of Events allowing some variations. Each Event is associated with a rule listing the coinciding muscle and movement states in form of the underlying Successions. Each movement Succession is linked to a Primitive Pattern with a rule describing the range of hip, knee, and ankle angles observed during this state. We plan to compare the patterns between several skaters and running speeds to investigate possible differences. Based

on background knowledge about the performance of the individual skaters this can lead to strategies for individualized training optimization.

One could criticize the manual interaction needed at some levels of our mining process, but we feel that a fully automated knowledge discovery is not desirable. We see the hierarchical decomposition into single temporal concepts as a great advantage of the TSKM. The separate stages offer unique possibilities for the expert to interpret, investigate and validate the discovered rules at different abstraction levels. The search space for the algorithms is smaller than when mining several concepts at the same time. Also, a large set of different algorithms can be plugged in the framework, e.g. segmentation to discover Successions or Hidden Markov Models to obtain Primitive Pattern to name just a few.

Usually only frequent Events and Sequences are kept while rare occurring patterns are discarded from further processing. Depending on the application, rare pattern might be important, however, and ranking should be done by a different interestingness measure.

While the data model for Events is currently univariate, we are experimenting with algorithms allowing overlapping Events involving less Aspects. However, this increases the number of Events found and makes mining Sequences more problematic.

There are only very few methods for rule discovery in multivariate time series. Last et al. (2001) use segmentation and feature extraction per segment. Association rules on adjacent intervals are mined using the Info-Fuzzy Network (IFN). The rule set is reduced using fuzzy theory. Höppner mines temporal rules in sequences of labeled intervals (Höppner (2001), Höppner (2002)), also obtained by segmentation and feature extraction. Patterns are expressed with Allen's interval logic (Allen (1983)) and mined with an Apriori algorithm. A comparison to the TSKM method on a conceptual and experimental level is planned.

6 Summary

We have presented our time series knowledge extraction framework TSKM. The hierarchal levels of the underlying rule language UTG cover the temporal concepts duration, coincidence, synchronicity and order at successive levels. Rules from each level are accompanied by linguistic descriptions, thus partial results can be interpreted and filtered by experts. We proposed algorithms for the mining stages including two new algorithms for mining duration and coincidence. First results of an application in sports medicine were mentioned.

Acknowledgements

We thank Dr. Olaf Hoos, Department of Sports Medicine, Philipps-University Marburg, for the data and interpretation of the mining results.

References

ALLEN, J. F. (1983): Maintaining knowledge about temporal intervals. *Comm. ACM, 26(11), 832–843.*

GUIMARAES, G. and ULTSCH, A. (1999): A method for temporal knowledge conversion. In: D. J. Hand, J. N. Kok and M. R. Berthold (Eds.): *Proc. of the 3rd Int. Symp. on Advances in Intelligent Data Analysis.* Amsterdam, Springer, 369–380.

HETLAND, M. L. (2004): A survey of recent methods for efficient retrieval of similar time sequences. In: M. Last, A. Kandel and H. Bunke (Eds.): *Data Mining in Time Series Databases.* World Scientific, 23–42.

HETLAND, M. L. and SAETROM, P. (2002): Temporal rule discovery using genetic programming and specialized hardware. In: A. Lotfi, J. Garibaldi, and R. John (Eds.): *Proc. of the 4th Int. Conf. on Recent Advances in Soft Computing (RASC)*, 182–188.

HÖPPNER, F. (2001): Discovery of temporal patterns – learning rules about the qualitative behaviour of time series. In: *Proc. of the 5th European PKDD.* Springer, Berlin, 192–203.

HÖPPNER, F. (2002): Learning dependencies in multivariate time series. *Proc. of the ECAI'02 Workshop on Knowledge Discovery in (Spatio-) Temporal Data, Lyon, France*, 25–31.

KADOUS, M. W. (1999): Learning comprehensible descriptions of multivariate time series. In: *Proc. 16th International Conf. on Machine Learning*, 454–463.

LAST, M., KLEIN, Y. and KANDEL, A. (2001): Knowledge discovery in time series databases. *IEEE Transactions on Systems, Man, and Cybernetics (31B).*

LIN, J., KEOGH, E., LONARDI, S. and CHIU, B. (2003): A symbolic representation of time series, with implications for streaming algorithms. In: *Proc. 8th ACM SIGMOD workshop DMKD 2003*, 2–11.

POVINELLI, R. J. (2000): Identifying temporal patterns for characterization and prediction of financial time series events. In: *Proc. International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining: TSDM 2000, Lyon, France*, 46–61.

SAETROM, P. and HETLAND, M. L. (2003): Unsupervised temporal rule mining with genetic programming and specialized hardware. In: *Proc. Int. Conf. on Machine Learning and Applications, ICMLA*, 145–151.

ULTSCH, A. (1991): Konnektionistische Modelle und ihre Integration mit wissensbasierten Systemen. *Forschungsbericht Informatik Nr. 396, Universität Dortmund, Habilitationsschrift.*

ULTSCH, A. (1996): Eine unifikationsbasierte Grammatik zur Beschreibung von komplexen Mustern in multivariaten Zeitreihen. *personal communication*

ULTSCH, A. (1999): Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series. In: Oja, E. and Kaski, S. (Eds.): *Kohonen Maps*, Elsevier, New York, 33–46.

ULTSCH, A. (2004): Unification-based temporal grammar. *Technical Report No. 37, Philipps-University Marburg, Germany.*

VILO, J. (1998): Discovering frequent patterns from strings. *Technical Report C-1998-9, Department of Computer Science, University of Helsinki.*

YANG, J., WANG, W., YU, P.S. and HAN, J. (2002): Mining long sequential patterns in a noisy environment. In: *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, 406–417.

A New Framework for Multidimensional Data Analysis

Shizuhiko Nishisato

The University of Toronto
252 Bloor Street West, Toronto, Ontario, Canada M5S 1V6
snishisato@oise.utoronto.ca

Abstract. Our common sense tell us that continuous data contain more information than categorized data. To prove it, however, is not that straightforward because most continuous variables are typically subjected to linear analysis, and categorized data to nonlinear analysis. This discrepancy prompts us to put both data types on a comparable basis, which leads to a number of problems, in particular, how to define information and how to capture both linear and nonlinear relations between variables both continuous and categorical. This paper proposes a general framework for both types of data so that we may look at the original statement on information.

1 Information in data

If X and Y are distributed as bivariate normal, its product-moment correlation (population correlation) is an upper limit for the correlation calculated from any categorized X and Y. In this case, using correlation as a measure of information, one may conclude that the continuous variables provide more information than their categorized counterparts. There is an important point to note, however, that once the bivariate normal distribution is assumed this assumption excludes the possibility of two variables ever being nonlinearly related: the Pearsonian (linear) correlation is the only parameter for the relationship.

There are several widely used measures of information such as Fisher's information, Kullback-Liebler information, Shannon's information and the sum of eigenvalues of the variance-covariance matrix. Out of them, the last one is often used for both continuous and categorical variables. However, the meaning of the eigenvalue for continuous data is often different from that for categorical data.

Considering that data are typically multidimensional, the most economical ways to describe data are principal component analysis (PCA) for continuous data and dual scaling (DS)(or, multiple-correspondence analysis, homogeneity analysis) for categorical data. Both PCA and DS are applications of singular value decomposition to the respective data types. Can we then

compare eigenvalues from PCA and DS in a meaningful manner? The answer is no. There is an important difference between PCA and DS: PCA considers a linear combination of variables; DS considers a linear combination of categories of variables. This difference is important in the sense that in PCA any variable (continuous or fixed-interval categorical) is represented as an axis in multidimensional space, while in DS categories of a variable are not restricted to lie on a straight line. Let us look at this difference first by using a numerical example.

2 Illustrative example

Let us borrow an example from Nishisato (2000, 2002)- please see the original papers for detailed analysis.

1. How would you rate your blood pressure?...(Low, Medium, High): coded 1, 2, 3
2. Do you get migraines?...(Rarely, Sometimes, Often): 1, 2, 3 (as above)
3. What is your age group?...(20-34; 35-49; 50-65): 1, 2, 3
4. How would you rate your daily level of anxiety?...(Low, Medium, High): 1, 2, 3
5. How would you rate your weight?...(Light, Medium, Heavy): 1, 2, 3
6. What about your height?...(Short, Medium, Tall): 1, 2, 3

Table 1 lists those Likert scores and also (1,0) response patterns. For data analysis, it is a widely used practice to regard Likert (fixed-interval) scores as continuous data, and treat them by methods for continuous variables, in the current case by PCA.

Following the current practice, PCA was applied to the left-hand side of the table, and DS to the corresponding response-pattern or indicator form on the right-hand side of the table.

The results of PCA show one cluster of blood pressure (BP), age (Age) and anxiety (Anx), meaning that as one gets older the blood pressure and the anxiety level tend to increase. In the orthogonal coordinate system, each variable is represented as an axis (i.e., the positions of three categories of BP lie on a straight line from the origin to the coordinates of BP) and this is the same whether fixed-interval categorical variables (the present case) or continuous variables are subjected to PCA. As such the correlation between two variables is defined as the cosine of the angles of the two axes. In contrast, DS does not restrict the positions of the categories of each variable and thus identifies clusters of any functionally related variables, linear or nonlinear. The DS results can be summarized as in Table 2.

Table 1. Likert Scores and Response Patterns

	PCA						DS					
	BP	Mig	Age	Anx	Wgt	Hgt	BP	Mig	Age	Anx	Wgt	Hgt
Subject	Q1	Q2	Q3	Q4	Q5	Q6	123	123	123	123	123	123
1	1	3	3	3	1	1	100	001	001	001	100	100
2	1	3	1	3	2	3	100	001	100	001	010	001
3	3	3	3	3	1	3	001	001	001	001	100	001
4	3	3	3	3	1	1	001	001	001	001	100	100
5	2	1	2	2	3	2	010	100	010	010	001	010
6	2	1	2	3	3	1	010	100	010	001	001	100
7	2	2	2	1	1	3	010	010	010	100	100	001
8	1	3	1	3	1	3	100	001	100	001	100	001
9	2	2	2	1	1	2	010	010	010	100	100	010
10	1	3	2	2	1	3	100	001	010	010	100	001
11	2	1	1	3	2	2	010	100	100	001	010	010
12	2	2	3	3	2	2	010	010	001	001	010	010
13	3	3	3	3	3	1	001	001	001	001	001	100
14	1	3	1	2	1	1	100	001	100	010	100	100
15	3	3	3	3	1	2	001	001	001	001	100	010

Table 2. Characteristics of Two DS Solutions

Component 1	**Component 2**
Association among [Low BP, High BP, Frequent Migraine, Old, High Anx, Short]	Association among [High BP, Old, Heavy, Short] [Low BP, Young, Tall]

A strong nonlinear association between BP and Mig (i.e., a frequent migraine occurs when BP is either low or high) was detected by DS, but it was completely ignored by PCA since the Pearsonian correlation between them was -0.06. Similarly, the above table shows a number of other nonlinear relations.

At this point, we should note the following characteristic of linear analysis. The correlation (-0.06) between BP and Mig indicates the extent to which it is similar to the perfect linear relationship. The value of -0.06 indicates that there is not much (linear) relation between them. PCA decomposes only a set of linear relations in data into orthogonal components, ignoring nonlinear relations. Notice that the value of -0.06 does not contain any information on what kind of nonlinear relationship is involved, but just almost the total absence of linear relation. Thus, when variables do not follow the multivariate normal distribution, PCA is not a method for data analysis, but is only a method for analyzing the linearly related portion of data. In contrast, DS decomposed whatever relations involved in data and as such it can always be said a method for data analysis.

Do these results indicate that DS captured more information than PCA? If the answer is yes, does it mean that categorized variables contain more information than continuous variables, contrary to our common sense?

3 Geometric model for categorical data

Let us introduce a common framework for both continuous and categorical data. Consider a variable with three categories (e.g., a multiple-choice question with three response options). The possible response patterns are (1,0,0), (0,1,0) and (0,0,1), which can be regarded as coordinates of three possible responses (Nishisato (2003a)). Each time, the question is answered by a subject, his or her response falls at one of these points. Once data are collected, the final locations of the three points are determined by the principle of the chi-square distance, as used by DS. Connecting the three points results in a triangle in two-dimensional space.

When a variable has more than two categories, therefore, we need more than one-dimensional space to represent it. Specifically speaking, a variable with p categories generally requires p-1-dimensional space. Thus, as has been widely done, approximating a multidimensional variable by a single component or two appears almost ill-founded. Rather the first approximation to a triangle in multidimensional space, for example, should be defined as its projection to the space with the first two principal axes.

Let us consider a continuous variable in the same way such that the number of its categories is equal to the number of distinct elements in the date set, say p*. Then, the variable can be represented as a (p*-1)-dimensional polyhedron. Admitting that this is not practical, we will find a common framework for both types of data.

4 Squared item-component correlation

Let us now pay attention to a statistic that can be used for multidimensional decomposition of data. One of them is the square of the item-component correlation, which indicates the degree of the contribution of a variable to a particular component. In terms of this statistic, our numerical examples of PCA and DS provide the decompositions of information in data as in Table 3.

In Table 3, "Eta2" indicates the correlation ratio, an objective function used by DS for optimization. Note that the sum of the square of item-component correlation over all components is equal to the number of categories minus one, that is, 2 in the above table of 3-category data. This sum is indicative of the degree of a polynomial functional relation, up to which DS can capture, that is, linear and quadratic relations in our example. Note

Table 3. Squared Item-Component Correlations of PCA and DS

(a)Squared Item-Component Correlation of PCA							
Dim	BP	Mig	Age	Anx	Wgt	Hgt	Sum(eigenvalue)
1	.59	.02	.55	.26	.24	.41	2.053
2	.03	.78	.23	.02	.58	.03	1.665
3	.23	.10	.10	.51	.02	.05	1.007
4	.05	.00	.00	.06	.03	.57	0.738
5	.05	.03	.11	.01	.10	.01	0.292
6	.04	.09	.06	.01	.05	.00	0.245
Sum	1	1	1	1	1	1	6

(b)Squared Item-Component Correlation of DS								
Dim	BP	Mig	Age	Anx	Wgt	Hgt	Eta2	Sum
1	.92	.93	.54	.41	.11	.36	.54	3.24
2	.38	.34	.22	.29	.52	.49	.37	2.25
3	.40	.48	.55	.46	.18	.01	.35	2.07
4	.02	.03	.40	.36	.83	.21	.31	1.84
5	.02	.01	.03	.31	.06	.35	.13	0.78
6	.10	.06	.03	.02	.02	.49	.12	0.72
7	.04	.08	.13	.06	.12	.03	.08	0.45
8	.05	.04	.06	.06	.07	.00	.05	0.28
9	.04	.01	.00	.03	.05	.05	.03	0.19
10	.01	.02	.03	.01	.03	.01	.02	0.10
11	.00	.01	.02	.01	.00	.00	.01	0.04
12	.00	.00	.00	.00	.00	.00	.00	0.00
Sum	2	2	2	2	2	2	2	12

therefore that DS of categorical data does not necessarily capture all nonlinear relations, but it is restricted to the number of categories of a variable.

5 Correlation between multidimensional variables

Suppose we have two categorical variables with m_i, m_j categories each. From two sets of response patterns, we can construct the contingency table of order $m_i x m_j$, which yields q (non-trivial) eigenvalues, where $q = min(m_i - 1, m_j - 1)$, say $\lambda_1, \lambda_2, ..., \lambda_q$ in the descending order.

Recently Nishisato (2004 in press) derived a measure of correlation, ν, between multidimensional variables, using forced classification (Nishisato (1984), Nishisato and Gaul (1991), Nishisato and Baba (1999)). He showed then that his measure is equal to the square root of the average eigenvalue of the contingency table, and that it is identical to Cramér's coefficient V. The interested readers are referred to the above paper.

6 Decomposition of information in data and total information

As was shown in the above examples of PCA and DS as applied to the Blood Pressure Data, the information in data can be decomposed into the contribution of the variable to each dimension (component) in terms of the squared item-component correlation coefficient (Nishisato (2003b)). When all the six variables are perfectly correlated to each other, those tables for PCA and DS would change respectively as in Table 4.

Table 4. Squared Item-Component Correlations of PCA and DS

(a)Squared Item-Component Correlation of PCA							
Dim	BP	Mig	Age	Anx	Wgt	Hgt	Sum(eigenvalue)
1	1.00	1.00	1.00	1.00	1.00	1.00	6.000
2	0.00	0.00	0.00	0.00	0.00	0.00	0.000
3	0.00	0.00	0.00	0.00	0.00	0.00	0.000
4	0.00	0.00	0.00	0.00	0.00	0.00	0.000
5	0.00	0.00	0.00	0.00	0.00	0.00	0.000
6	0.00	0.00	0.00	0.00	0.00	0.00	0.000
Sum	1	1	1	1	1	1	6

(b)Squared Item-Component Correlation of DS								
Dim	BP	Mig	Age	Anx	Wgt	Hgt	Eta2	Sum
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	6.00
2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	6.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sum	2	2	2	2	2	2	2	12

This means that when all the variables are perfectly correlated PCA yields only one axis to represent the variables and DS in this case of three category data needs only two dimensions where all six triangles are completely merged into one triangle.

Imagine what will happen when all the variables are mutually uncorrelated. In PCA, the above table changes to the identity matrix, meaning that

each variable is represented by its own axis. In DS, the triangle of the first variable occupies, for example, dimensions 1 and 2, where we see the entries of 1, the second variable dimensions 3 and 4, and so on, thus all six triangles span 12-dimensional space without any overlapping among them.

In the case of DS, the data set can be accommodated in two-dimensional space (perfect correlation) or 12-dimensional space (uncorrelated case), which seems to show a great discrepancy in information contained in data. If six items are perfectly correlated to one another, we need only one item to know everything about six items. What emerges from this is a proposal for a definition of information, not im terms of the sum of eigenvalues (variances), or the sum of six triangles in the current example, but in terms of the joint variance. This idea can be related to the terms used in set theory and information theory. In set theory, the traditional definition of the sum of the variances or eigenvalues corresponds to the sum of sets, while the new idea corresponds to the union of sets. In information theory, the traditional idea is the sum of entropies of variables, while the new idea is the joint entropy. Along these lines of thought, Nishisato (2003b) adjusted the sum of squares of item-component correlations (the sum of eigenvalues) by eliminating overlapping portion through two-way correlations, three-way correlations and so on up to the *n-th* order correlation, n being the number of variables, as we do in expressing the joint entropy in information theory, arriving at the following measure:

$$T(inf) = \sum_{j=1}^{n}\sum_{k=1}^{K} r^2{}_{jt(k)} - \sum_{i<j}^{n} r_{ij} + \sum_{i<j<k}^{n} r_{ijk} - \cdots + (-1)^{n-1} r_{123\cdots n}, \quad (1)$$

where

$$r_{123\cdots p} = \sum_{i=1}^{N} \frac{z_{1i}z_{2i}z_{3i}\cdots z_{pi}}{N}, \quad (2)$$

z_{ij} is the standardized score of subject i on item j, N is the number of subjects, and K is the total number of components. In practice, we would need to use a simplified expression as an approximation to the above expression since it is likely that high-order terms may not contribute much to the measure. Since this aspect of approximation depends on data, we are yet to see how it works in practice.

7 Conclusion

The current study proposes the following framework for multidimensional data analysis.
(1) Discretize (desensitize) continuous measurement so that nonlinear relations between variables may be captured by DS without difficulty.

(2) When a variable is described as a k-dimensional polyhedron, the first approximation to the variable should be defined as one from the fist k components, rather than the current practice of the first component.
(3) In view of (2), correlation between two categorical variables should be defined in k-dimensional space, rather than in one dimension.
(4) The total information in data should be defined as the union of polyhedrons or joint entropy of variables. (5) In view of (4), the information contained in k-dimensional space should be defined, not by the k-th eigenvalue, but by the union of variables in k-dimensional space. (6) In view of (4) and (5), the traditional hypothesis testing in multivariate analysis may have to be reconsidered since a function of eigenvalues will be modified by the new definition of information.

References

NISHISATO, S. (1984): Forced classification: A simple application of a quantification technique. *Psychometrika, 49, 25–36.*

NISHISATO, S. (2000): Data types and information: Beyond the current practice of data analysis. In Decker, R. and Gaul, W. (eds.), *Classification and information processing at the turn of the millennium.* Springer, Berlin, 40–51.

NISHISATO, S. (2002): Differences in data structures between continuous and categorical variables from dual scaling perspectives, and a suggestion for a unified mode of analysis. *Japanese Journal of Sensory Evaluation, 6, 89–94* (in Japanese).

NISHISATO, S. (2003a): Geometric perspectives of dual scaling for assessment of information in data. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano and J. Meulman (Eds.): *Recent Developments in Psychometrics.* Tokyo, Springer, 453–462.

NISHISATO, S. (2003b): Total information in multivariate data from dual scaling perspectives. *The Alberta Journal of Educational Research, XLIX, 244–251.*

NISHISATO, S. (2004 in press): Correlational structure of multiple choice data as viewed from dual scaling. A paper to appear in a book, edited by M.J.Greenacre and J. Blasius. (publisher yet unknown)

NISHISATO, S. and BABA, Y. (1999): On contingency, projection and forced classification of dual scaling. *Behaviormetrika, 26, 207–219.*

NISHISATO, S. and GAUL, W. (1990): An approach to marketing data analysis: The forced classification procedure of dual scaling. *Journal of Marketing Research, 27, 354–360.*

External Analysis of Two-mode Three-way Asymmetric Multidimensional Scaling

Akinori Okada[1] and Tadashi Imaizumi[2]

[1] Department of Industrial Relations, School of Social Relations, Rikkyo (St. Paul's) University, 3-34-1 Nishi Ikebukuro, Toshima-ku Tokyo, 171-8501 Japan
[2] School of Management and Information Sciences, Tama University, 4-4-1 Hijirigaoka, Tama city, Tokyo, 206-0022 Japan

Abstract. An external analysis of two-mode three-way (object×object×source) asymmetric multidimensional scaling is introduced, which is similar to the external analysis of INDSCAL. The present external analysis discloses the asymmetry of each object, and source differences in symmetric and in asymmetric proximity relationships among objects respectively for an externally given configuration of objects. The present external asymmetric multidimensional scaling is applied to the university enrollment flow among Japanese prefectures.

1 Introduction

Multidimensional scaling (MDS) usually analyzes proximities or preferences to obtain a spatial representation or a configuration of objects or of objects and sources. In the configuration, proximity relationships among objects are represented as interpoint distances, or preference relationships are represented as distances from ideal points to object points or as projections of object points on ideal vectors. The external analysis of MDS integrates proximities or preferences with an externally given configuration to obtain additional information. PREFMAP (Carroll (1972)) maps sources into an externally given configuration of objects as ideal vectors or ideal points by analyzing preferences for objects of sources. The external analysis of INDSCAL (Arabie et al. (1987)) derives a weight configuration of sources for an externally given group stimulus configuration of objects by analyzing a set of proximity matrices among objects where each matrix comes from a source.

The present external analysis of two-mode three-way asymmetric MDS is based on the predecessor (Okada and Imaizumi (1997)), and is similar to the external analysis of INDSCAL mentioned above. For an externally given configuration of objects, the present external analysis discloses (a) the asymmetry of each object, (b) source differences (differences among sources) in symmetric proximity relationships among objects, (c) the orthogonal rotation for the externally given configuration of objects, and (d) source differences in asymmetric proximity relationships among objects by analyzing two-mode three-way asymmetric proximities (object×object×source).

2 The method

The model and the method are external analysis versions of those combining the two earlier studies (Okada and Imaizumi (2000, 2002)), both were extended from those of the predecessor. While in Okada and Imaizumi (2000) the common object configuration was not rotated, in Okada and Imaizumi (2002) an orthogonal rotation was applied to the common object configuration. The combined model consists of the common object configuration, the symmetry weight, the asymmetry weight, and the orthogonal rotation matrix for the common object configuration. The common object configuration shows symmetric and asymmetric proximity relationships among objects which are common to all sources, where each object is represented as a point and a circle (sphere, hypersphere) centered at that point in a multidimensional Euclidean space. The radius of the circle of an object shows the asymmetry of the object; the larger the radius of an object is, the larger the similarity from the object to the other objects is, and the smaller the radius of an object is, the larger the similarity from the other objects to the object is. The larger the difference of the two radii of two objects is, the larger the asymmetry between the two objects is. When two objects have the radius of the same size, there is no asymmetry between the two objects. The symmetry weight shows source differences in symmetric proximity relationships among objects. The asymmetry weight shows source differences in asymmetric proximity relationships among objects.

Let n be the number of objects, N the number of sources, p the dimensionality of the externally given common object configuration, and s_{jki} the observed proximity from object j to object k for source i ($j, k = 1, \cdots, n; i = 1, \cdots, N$). It is assumed that for each source s_{jki} is monotonically related to m_{jki};

$$m_{jki} = d_{jki} - \frac{d_{jki} r_j}{\sqrt{\sum_{t'=1}^{p} \left[\frac{x^*_{jt'} - x^*_{kt'}}{u_i u^*_{t'}} \right]^2}} + \frac{d_{kji} r_k}{\sqrt{\sum_{t'=1}^{p} \left[\frac{x^*_{kt'} - x^*_{jt'}}{u_i u^*_{t'}} \right]^2}}, \tag{1}$$

where $x^*_{jt'}$ is the coordinate of object j on dimension t' derived by orthogonally rotating dimensions of the common object configuration, $r_j \geq 0$ is the radius representing object j, and d_{jki} is the distance between two points representing objects j and k in the configuration of objects for source i, which is defined as

$$d_{jki} = w_i \sqrt{\sum_{t=1}^{p} (x_{jt} - x_{kt})^2}, \tag{2}$$

and x_{jt} is the coordinate of object j on dimension t of the common object configuration. The symmetry weight $w_i \geq 0$ represents the salience of symmetric proximity relationships among objects for source i. There are two kinds of

the asymmetry weights $u_i \geq 0$ and $u^*_{t'} \geq 0$ (Okada and Imaizumi 2000). u_i represents the salience of asymmetric proximity relationships among objects for source i, and $u^*_{t'}$ represents the salience of asymmetric proximity relationships among objects along dimension t' of the orthogonally rotated common object configuration.

The procedure for deriving the common joint configuration, the symmetry weight, the asymmetry weight, and the orthogonal rotation matrix in the combined model is almost the same as those for Okada and Imaizumi (2000, 2002). A nonmetric iterative algorithm to derive the common object configuration (x_{jt}; $j = 1, \cdots, n$, $t = 1, \cdots, p$, and r_j; $j = 1, \cdots, n$), the symmetry weight (w_i; $i = 1, \cdots, N$), the asymmetry weight for the source (u_i; $i = 1, \cdots, N$), the asymmetry weight for the dimension ($u^*_{t'}$; $t' = 1, \cdots, p$), and the orthogonal rotation matrix from s_{jki} $(j, k[j \neq k] = 1, \cdots, n; i = 1, \cdots, N)$ was extended from the one for the predecessor. The badness-of-fit measure S Stress is defined as

$$S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\sum_{\substack{j=1\\ j\neq k}}^{n}\sum_{k=1}^{n}(m_{jki} - \hat{m}_{jki})^2 \Big/ \sum_{\substack{j=1\\ j\neq k}}^{n}\sum_{k=1}^{n}(m_{jki} - \bar{m}_i)^2\right]}, \qquad (3)$$

where $\hat{m}_{jki}$ is the monotone transformed s_{jki}, and $\bar{m}_i$ is the mean of m_{jki} for source i. The radius, the symmetry and the asymmetry weights, and the orthogonal rotation matrix which minimize the Stress are sought for a given dimensionality.

In the present external analysis model, the common object configuration is given externally, representing each object as a point without a radius. In this configuration, only symmetric proximity relationships among objects are shown, and asymmetric proximity relationships among objects are not shown. The externally given common object configuration is normalized in exactly the same manner as that of the predecessor, so that the origin is at the centroid of the points representing objects and the sum of squared coordinates of objects is equal to the number of objects (Okada and Imaizumi (1997), Equation (3.4)). Then, the radius, the symmetry weight, the asymmetry weight, and the orthogonal rotation matrix to be applied to the externally given common object configuration are derived exactly the same procedure mentioned above from the observed proximity s_{jki} $(j, k[j \neq k] = 1, \cdots, n; i = 1, \cdots, N)$. In the iterative process of the procedure, the externally given coordinates of the objects (x_{jt}; $j = 1, \cdots, n$, $t = 1, \cdots, p$) are unchanged.

3 An application

The present external analysis of asymmetric MDS was applied to the university enrollment flow data among 47 Japanese prefectures. The data, which

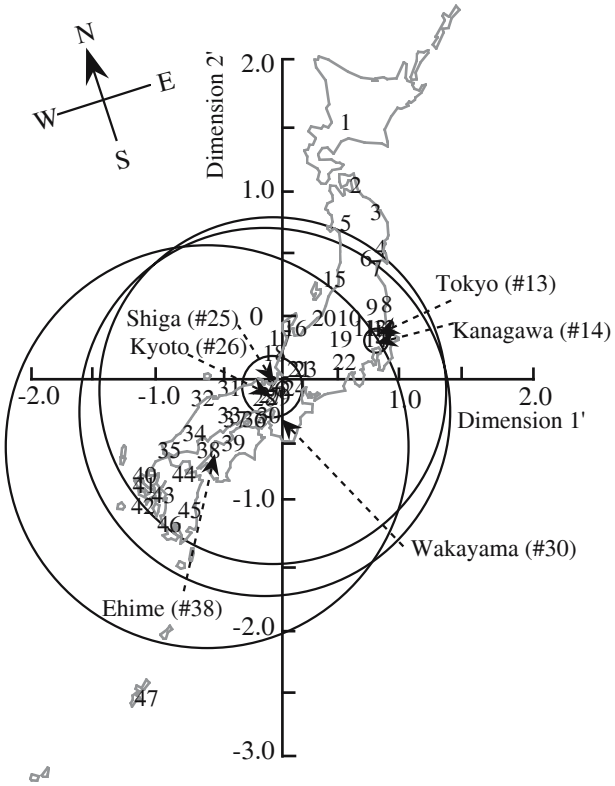


Fig. 1. Rotated common object configuration of the external analysis.

had been analyzed in Okada and Iwamoto (1996) by using the asymmetric cluster analysis extended from Hubert (1973), consist of two 47×47 tables. Each table corresponds to the data before and after the introduction of the JFSAT (Joint First Stage Achievement Test) in 1979. The (j, k) element of each table represents the number of enrollments into universities in prefecture k from high schools in prefecture j during four years from 1974 to 1977 for the first table and from 1981 to 1984 for the second table. The two tables are 47×47×2 two-mode three-way asymmetric proximities.

The longitude and the latitude of the capitol of each prefecture were used as the externally given two-dimensional common object configuration of 47 prefectures; i.e., a geographical map was used. Analyzing the two tables by using the two-dimensional configuration of prefectures as the externally given common object configuration, the radius of the prefecture, the symmetry weight for source or the table (before and after the JFSAT), the asymmetry weights for source and for dimension, and the orthogonal rotation matrix are derived.

Table 1. Radius

Prefecture	External	Internal	Prefecture	External	Internal
1 Hokkaido	0.649	0.202	25 Shiga	1.424	0.757
2 Aomori	0.649	0.268	26 Kyoto	0.247	0.064
3 Iwate	0.788	0.291	27 Osaka	0.648	0.299
4 Miyagi	0.282	0.055	28 Hyogo	0.997	0.462
5 Akita	1.020	0.394	29 Nara	0.782	0.284
6 Yamagata	0.984	0.382	30 Wakayama	1.493	0.807
7 Fukushima	0.534	0.155	31 Tottori	1.259	0.630
8 Ibaragi	0.415	0.075	32 Shimane	1.342	0.650
9 Tochigi	0.761	0.266	33 Okayama	1.180	0.524
10 Gunma	0.748	0.262	34 Hiroshima	1.153	0.513
11 Saitama	0.212	0.000	35 Yamaguchi	1.321	0.585
12 Chiba	0.287	0.040	36 Tokushima	1.231	0.601
13 Tokyo	0.000	0.016	37 Kagawa	1.382	0.679
14 Kanagawa	0.107	0.002	38 Ehime	1.609	0.757
15 Nigata	0.949	0.380	39 Kochi	1.204	0.558
16 Toyama	1.293	0.616	40 Fukuoka	0.857	0.371
17 Ishikawa	0.548	0.138	41 Saga	1.387	0.644
18 Fukui	1.212	0.589	42 Nagasaki	1.316	0.626
19 Yamanashi	0.529	0.109	43 Kumamoto	1.246	0.536
20 Nagano	0.852	0.325	44 Oita	1.218	0.540
21 Gifu	1.230	0.594	45 Miyazaki	1.270	0.583
22 Shizuoka	1.024	0.429	46 Kagoshima	1.244	0.586
23 Aichi	0.659	0.258	47 Okinawa	1.044	0.544
24 Mie	1.398	0.706			

The analysis was done in two-dimensional space, and the minimized Stress was 0.745. Figure 1 shows the rotated common object configuration. Dimensions $1'$ and $2'$ were derived by orthogonally rotating the dimensions of the map (the longitude and the latitude) 19.9 degrees clockwise. The longitude and the latitude coordinates are represented in the upper left corner of Figure 1 to show the orientation of the dimensions before the orthogonal rotation. Although in the present model each prefecture is represented as a point and a circle centered at that point, the circle was drawn only for prefectures having the smallest and largest three radii. The other circles were eliminated to avoid the complexity of the configuration and to more clearly show the location of prefectures. The horizontal dimension (Dimension $1'$) is almost parallel with the line connecting Tokyo (#13), Aichi (#23), and Kyoto (#26) which have universities being in high reputation, and constitute the most industrialized region as well. In addition, the horizontal dimension seems to differentiate two areas distinguished by Okada and Iwamoto (1996): one, whose center is Tokyo, consists of prefectures whose high school graduates have a stronger tendency of entering into universities in Tokyo, and the other, whose center is Kyoto, consists of prefectures whose high school graduates have a stronger

tendency of entering into universities in Kyoto. The vertical dimension (Dimension 2′) corresponds to the geographic distance from the centers of the two areas, and seems to represents differences among prefectures within each of the two areas.

The radius is shown at the column labeled external of Table 1. In the present application, the larger radius means the larger tendency of entering into universities in the other prefectures from high schools in the corresponding prefecture, and the smaller radius means the larger tendency of accepting high school graduates from the other prefectures into universities in the corresponding prefecture. Prefectures (Tokyo, Kanagawa (#14), and Kyoto) having universities into which many high school graduates from the other prefectures enter, have the smaller radii (Table 1). The radius of Tokyo is the smallest which is zero by the normalization.

Table 2 shows the symmetry and the asymmetry weights. The symmetry weight is smaller before the introduction of the JFSAT than after the JFSAT; suggesting the symmetry of the university enrollment flow among prefectures increased by the introduction of the JFSAT. This is also reflected in the asymmetry weight for the sources; the asymmetry weight before the JFSAT is larger than that after the JFSAT. The asymmetry weight for Dimension 1′ is larger than that for Dimension 2′. This suggests that the asymmetry between the two areas is larger than that within the area.

Table 2. Symmetry and asymmetry weights

Source	Symmetry weight for source
Before the JFSAT	1.000
After the JFSAT	1.068
Source	Asymmetry weight for source
Before the JFSAT	1.110
After the JFSAT	0.918
Dimension	Asymmetry weight for rotated dimension
Dimension 1′	1.013
Dimension 2′	0.987

4 Discussion

An external analysis of two-mode three-way asymmetric MDS was introduced and applied to the migration data from the high school to the university among Japanese prefectures. The results are consistent with those obtained in Okada and Iwamoto (1996).

To compare the present result with that obtained by using the internal (usual) analysis, the present data were analyzed by the internal two-mode

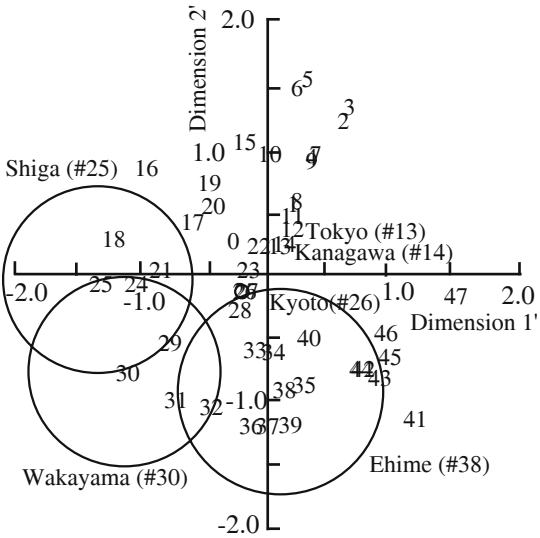


Fig. 2. Common object configuration of the internal analysis.

three-way asymmetric MDS using the same model as that of the present external analysis. The analyses with rational initial configuration and values were done by using the maximum dimensionality of five through nine and the minimum dimensionality of one (Okada and Imaizumi (1997, p. 216)). The analyses gave five different kinds of results for each dimensionality from five through one. The minimum Stress values of the five different kinds of results at each of five- through unidimensional spaces were 0.372, 0.379, 0.402, 0.455, and 0.610. The two-dimensional result was chosen as the solution. Comparing the minimized Stress values of the external (0.745) and the internal (0.455) analyses suggests that the goodness-of-fit deteriorated by using the geographical map of the country as the externally given common object configuration. Figure 2 shows the obtained common object configuration with radius for only the same prefectures having radii in Figure 1. The radius is shown at the column labeled internal of Table 1. In Figure 2 the map of the country is distorted; northern (#1-#7) and southern (#40-#47) prefectures are located closer to the center of the area (Tokyo and Kyoto) than in the geographical map.

The two dimensions seem to represent the same meaning as in the external analysis; the difference between two areas (Dimension 2′), and the difference within each of the two areas (Dimension 1′). Also, Dimension 2′ is almost parallel with the line connecting Tokyo, Aichi, and Kyoto. While the radii of the external analysis are larger than that of the internal analysis, two sets of radii shown in Table 1 are similar, and the correlation coefficient between the two sets is 0.98. Thus, although the asymmetric relationships of the university enrollment flow is more exaggerated in the external analysis

than in the internal analysis, almost the same characteristics of the asymmetric relationships among prefectures are represented in both the external and the internal analyses. The symmetry and asymmetry weights of the internal analysis have the same characteristics as those of the external analysis; both analyses suggest that the symmetry of the university enrollment flow increased after the introduction of the JFSAT, and the asymmetry between the two areas are larger than that within the area.

The ratios of the magnitude of symmetric and asymmetric components of the total sum of squared deviation of m_{jki} from $\bar{m}_i$ (Okada and Imaizumi (1997, p. 212)) were 0.653 and 0.347 for the external analysis, and 0.799 and 0.201 for the internal analysis. This suggests that the larger asymmetric component of the external analysis resulted in the deterioration of the goodness-of-fit. But the deterioration seems not to destroy the relationships of the university enrollment flow based on the two areas, and not to destroy characteristics of the asymmetric relationships. While the present external analysis has poorer fit to the data than the internal analysis has, the external analysis gives results which can be interpreted based on the geographical features of prefectures, which facilitates understanding of the results considerably.

References

ARABIE, P., CARROLL, J.D., and DESARBO, W.S. (1987): *Three-Way Scaling and Clustering*. Sage Publications, Newburry Park, U.S.A.

CARROLL, J.D. (1972): Individual Differences and Multidimensional Scaling. In: R.N. Shepard, A.K. Romney, and S.B. Nerlove (Eds.): *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences Vol. 1 Theory*. Seminar Press, New York, 105–155.

HUBERT, L. (1973): Min and Max Hierarchical Clustering Using Asymmetric Similarity Measures. *Psychometrika, 38, 63-72*.

OKADA, A. and IMAIZUMI, T. (1997): Asymmetric Multidimensional Scaling of Two-Mode Three-Way Proximities. *Journal of Classification, 14, 195–224*.

OKADA, A. and IMAIZUMI, T. (2000): Two-Mode Three-Way Asymmetric Multidimensional Scaling with Constraints on Asymmetry. In: R. Decker and W. Waul (Eds.): *Classification and Information Processing at the Turn of the Millennium*. Springer-Verlag, Berlin, 52–59.

OKADA, A. and IMAIZUMI, T. (2002): Two-mode three-way Multidimensional Scaling with Different Orientations of Dimensions for Symmetric and Asymmetric Relationships. In: S. Nishisato, Y. Baba, H. Bozdagan and K. Kanefuji (Eds.): *Measurement and Multivariate Analysis*. Springer-Verlag, Tokyo, 52–59.

OKADA, A. and IWAMOTO, T. (1996): University Enrollment Flow among the Japanese Prefectures: A Comparison Before and After the Joint First Stage Achievement Test by Asymmetric Cluster Analysis. *Behaviormetrika, 23, 169–185*.

The Relevance Vector Machine Under Covariate Measurement Error

David Rummel

Institut für Statistik,
Universität München, 80539 München, Germany

Abstract. This paper presents the application of two correction methods for covariate measurement error to nonparametric regression. We focus on a recent and due to its sparsity properties very promising smoothing approach coming from the area of machine learning, the Relevance Vector Machine (RVM), developed by Tipping (2000). Two correction methods for measurement error are then applied to the RVM: regression calibration (Carroll et al. (1995)) and the SIMEX method (Carroll et al. (1995)). We show why standard regression calibration fails and present a simulation study that indicates an improvement of the RVM regression in terms of bias when SIMEX correction is applied.

1 Introduction

Considering bivariate data $\{(x_i, t_i)\}_{i=1}^N \in \mathbb{R} \times \mathbb{R}$ including a single covariate x_i and a scalar response t_i (so-called target) for each observation i, we usually assume the targets being decomposable into a structural and a random part:

$$t_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, N,$$

where the random errors ϵ_i are independent samples from some noise process. From a statistical point of view one is now interested in estimating the functional relationship between the target variable and the covariates. In contrast to parametric modeling we will not impose a certain model structure on $f(\cdot)$, but allow for a flexible approximation faithful to the data, using the Relevance Vector Machine introduced by Tipping (2000).
Our special interest however lies in the behavior of the RVM under covariate measurement error. This aspect gains impact when we analyze real world data, where assumption of at least minor measurement error seems indispensable. Popular examples come from the area of medicine and epidemiology, where exposure to a certain radiation or nutrition habit is recorded.
Carroll et al. (1999) presented work on measurement error for (penalised) regression splines, but to our knowledge this problem has not yet been discussed in the context of RVM.
Section 2 contains the theory of the RVM, while in Section 3 we describe our approach to measurement error correction in the RVM context. Finally we present the results of a simulation study on the error correction applying the SIMEX method.

2 Nonparametric regression using the RVM

Kernel methods like the Support Vector Machines (Vapnik (1995)) became increasingly popular in the last years (Schölkopf and Smola (2002)). Against this background Tipping (2000, 2001) presented the Relevance Vector Machine which is related to the Support Vector Machines (SVM) but within a truly Bayesian framework.

2.1 The RVM model setup

Similar to the B-Spline, P-Spline and SVM approach, the RVM concept of approximation is fitting a sum of individually weighted, generally nonlinear basis functions to the data. The targets are assumed to be decomposable into

$$t_i = \sum_{j=0}^{N} w_j \cdot \phi_j(x_i) + \epsilon_i, \quad i = 1, \ldots, N. \tag{1}$$

We note that $\phi_0(x)$ is an intercept vector with associated weight w_0, while for $j \geq 1$ the basis function $\phi_j(x)$ is centered on x_j. The errors are assumed to be i.i.d. normally distributed

$$p(\boldsymbol{\epsilon}) = \prod_{i=1}^{N} \mathcal{N}(\epsilon_i | 0, \sigma^2). \tag{2}$$

Here and in any subsequent expression we omit the implicit condition on the covariate data $\{x_i\}$. The model (1) is overspecified since every covariate sample serves as a basis knot in this formulation. Tipping (2001) encodes the preference for a sparse model by placing a prior distribution on the $N + 1$ weights:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{j=0}^{N} \mathcal{N}(w_j | 0, \alpha_j^{-1}), \tag{3}$$

where $\mathbf{w}$ denotes the parameter vector including all $N+1$ model weights with each weight w_j being i.i.d. Gaussian with zero mean and individual variance α_j^{-1}.

To put the RVM into a fully Bayesian framework two more distributional assumptions are employed for the inverse variance parameters $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_N)^T$ and $\beta = \sigma^{-2}$. Tipping (2001) specifies Gamma (hyper-) priors for these scale parameters:

$$p(\boldsymbol{\alpha}) = \prod_{j=0}^{N} Gamma(\alpha_j | a, b) \text{ and } p(\beta) = Gamma(\beta | c, d), \tag{4}$$

setting the corresponding parameters $a = b = c = d = 0$; this is equivalent to specifying uniform distributions for $\boldsymbol{\alpha}$ and β on a logarithmic scale. Specification of an individual prior/hyperprior for every weight, follows an approach

from MacKay (1994), termed as *automatic relevance determination*, generally leading to sparse models.

2.2 Inference

Estimation of the unknown parameters $\mathbf{w}, \boldsymbol{\alpha}$ and β in a Bayesian framework is done via the posterior distribution of these parameters:

$$p(\mathbf{w}, \boldsymbol{\alpha}, \beta|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{w}, \boldsymbol{\alpha}, \beta)p(\mathbf{w}, \boldsymbol{\alpha}, \beta)}{p(\mathbf{t})}. \tag{5}$$

Analytic calculation of (5) is not feasible, since the normalizing integral $p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w}, \boldsymbol{\alpha}, \beta)p(\mathbf{w}, \boldsymbol{\alpha}, \beta)\, d\mathbf{w}d\boldsymbol{\alpha}d\beta$ cannot be computed.
Tipping (2001) suggests decomposition of the posterior distribution into

$$p(\mathbf{w}, \boldsymbol{\alpha}, \beta|\mathbf{t}) = p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \beta)p(\boldsymbol{\alpha}, \beta|\mathbf{t}),$$

with $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}, \beta)$ being Gaussian, see Tipping (2001) for details. However, the posterior of the hyperparameters $\boldsymbol{\alpha}, \beta$ cannot be stated and under the assumption of $p(\boldsymbol{\alpha})$ and $p(\beta)$ being uniform (over a logarithmic scale) Tipping (2001) just maximizes the marginal likelihood to find most probable hyperparameters:

$$p(\mathbf{t}|\boldsymbol{\alpha}, \beta) = (2\pi)^{-\frac{N}{2}} \mid C \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{t}^{\mathrm{T}}C^{-1}\mathbf{t}\right\}, \tag{6}$$

with covariance matrix $C := \beta^{-1}\mathrm{I} + \Phi A^{-1}\Phi^{\mathrm{T}}$ and design matrix $\Phi_{N\times N+1}$ consisting of the intercept vector and a set of columns, representing basis function values. Inherent parameters in the basis functions need to be set in advance or estimated via cross validation. In similar Bayesian models, this maximizing method is referred to as type-II maximum likelihood method.
For most probable values $\widehat{\boldsymbol{\alpha}}_{MP}$ and $\widehat{\beta}_{MP}$, based on (6), the weights are computed as the mean of the Gaussian posterior $p(\mathbf{w}|\mathbf{t}, \widehat{\boldsymbol{\alpha}}_{MP}, \widehat{\beta}_{MP})$.
Optimizing the unknown parameters is then an iterative cycling through two steps:

- estimate most probable values $(\widehat{\boldsymbol{\alpha}}_{MP}, \widehat{\beta}_{MP})$ via type-II ML based on the marginal likelihood (6), using actual estimates for the weights,
- update the weights as mean of the Gaussian posterior $p(\mathbf{w}|\mathbf{t}, \widehat{\boldsymbol{\alpha}}_{MP}, \widehat{\beta}_{MP})$.

For many weights the posterior becomes sharply peaked around zero and these weights will practically be set to zero during the optimization process. Thus the estimated model (1) shrinks to a sum of very few weighted basis functions. Tipping (2001) compares the performance of the RVM to the SVM and states comparable results for benchmark data sets in terms of accuracy. However, the major result was the superior sparsity of the RVM in using relevant basis functions for modeling the data.

3 Covariate measurement error and its correction

In many practical situations the variable X we observe is merely a surrogate for the variable ξ that cannot be observed directly. That happens when measurement of ξ is error prone, as it is the case when e.g. physical measurement is affected by the surrounding conditions. We stress that there is almost no kind of measurement that is free from potential measurement error. This error may seem negligible in some cases, however in a lot of cases it is not. Statistical analysis ignoring such inherent error is referred to as 'naive analysis' and Carroll et al. (1999) emphasize, that when measurement error is ignored 'conventional parametric and nonparametric techniques are no longer valid'. That is, the parameter estimates we get from the naive analysis are usually biased.
There are two standard approaches to error correction: Carroll et al. (1995) describe regression calibration and an approach based on simulation and extrapolation (SIMEX). Carroll et al. (1999) present one adoption of this SIMEX for nonparametric regression using (penalised) regression splines.

3.1 The classical error model

In order to take measurement error into account in our statistical analysis, we need a model relating the true covariate to the observable covariate.
Assume that there is a variable ξ one would like to include into the set of covariates but the device allows measurement merely under inclusion of a random error. A common model for that type of error process is:

$$X = \xi + \delta, \quad (\delta, \xi) \sim \text{indep.}, \quad \mathbb{E}(\delta) = 0, \tag{7}$$

which is frequently extended to $\delta \sim \mathcal{N}(0, \sigma_\delta^2)$ and $\xi \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2)$.
Measurement deviates randomly from the true value but is expected to be correct on average.

3.2 Error correction using regression calibration

Carroll et al. (1995) describe regression calibration, where the true but unobservable covariate ξ is replaced by $\mathbb{E}(\xi|X)$ and then a standard analysis is carried out. Computability of $\mathbb{E}(\xi|X)$ depends on the type of error model, but for the classical error model it is easily derived and given as:

$$\mathbb{E}(\xi|X) = \mu_\xi + \lambda \cdot (X - \mu_\xi), \quad \text{with} \quad \lambda := \frac{\sigma_X^2 - \sigma_\delta^2}{\sigma_X^2}. \tag{8}$$

The error variance σ_δ^2 needs to be known or estimated from e.g. validation/replication data and both $\mu_\xi (= \mu_X)$ and σ_X^2 can be estimated from the sample. This approach yields consistent parameter estimates in the linear regression.

We apply this method to the RVM by rewriting the regression of the target variable T on the observable covariate X in terms of the model weights:

$$\begin{aligned}\mathbb{E}(T|X) &= \mathbb{E}(\mathbb{E}(T|X,\xi)|X) \\ &= \mathbb{E}(\mathbb{E}(T|\xi)|X) \qquad (9) \\ &= \mathbb{E}\left(\left(\textstyle\sum_{j=0}^{N} w_j \phi_j(\xi)\right)|X\right) \qquad (10) \\ &= \textstyle\sum_{j=0}^{N} w_j \cdot \mathbb{E}\left(\phi_j(\xi)|X\right). \qquad (11)\end{aligned}$$

(In (9) it is assumed that the measurement error is independent of the target variable.) Now, if there is measurement error in the data, $\mathbb{E}\left(\phi_j(\xi)|X\right)$ should be used to estimate the weight parameters instead of naively using $\phi_j(X)$, which generally yields biased estimates. Note that only if $\phi(\cdot)$ is a linear basis $\mathbb{E}(\phi_j(\xi|X))$ in (11) simplifies to $\phi_j\left(\mathbb{E}(\xi|X)\right)$.
Ignoring nonlinearity of Gaussian basis functions and simply replacing the error prone X by $\mathbb{E}(\xi|X)$ from (8) into the RVM model gives the following modified basis function:

$$\begin{aligned}\phi_j\left(\mathbb{E}(\xi|X)\right) &= \exp(-\eta^2(\mathbb{E}(\xi|X) - \mathbb{E}(\xi|x_j))^2) \\ &= \exp(-(\eta\cdot\lambda)^2 \cdot (X - x_j)^2),\end{aligned}$$

where the parameter η determines the width of the Gaussian basis. Thus, inference of the weight parameters applying standard regression calibration is equivalent to a naive analysis using a modified basis kernel. We expect no genuine correction effect from merely using a wider (since $\lambda \leq 1$) Gaussian kernel and thus this regression calibration approach will not be considered in our simulation study.
Calibration of the basis functions i.e. replacing the unobservable $\phi_j(\xi)$ by $\mathbb{E}\left(\phi_j(\xi)|X\right)$ is another more promising approach and will be explored in forthcoming work.

3.3 Error correction using SIMEX

Carroll et al. (1999) describe the concept of SIMEX (SIMulation EXtrapolation) for nonparametric regression based on the 'classical' SIMEX (Carroll et al. (1995)). The effect of covariate measurement error on the estimated function is studied in a simulation study and afterwards an extrapolation on the error-free case is performed.
For the classical error model (7), random errors $\delta_i^* \sim N(0, \sigma_{\delta^*}^2)$ are generated and added to the observed $x_i, i = 1, \ldots, N$. Then a standard RVM analysis is performed using these 'new' data under additional error. Varying the error variances $\sigma_{\delta^*}^2 = c \cdot \sigma_\delta^2$ in multiples of the true measurement error variance allows us to study its effect on the prediction $\widehat{f}(\xi_k)$ at arbitrary points of interest $\xi_k, k = 1, \ldots, K$. Figure 1 illustrates the increasing attenuation of

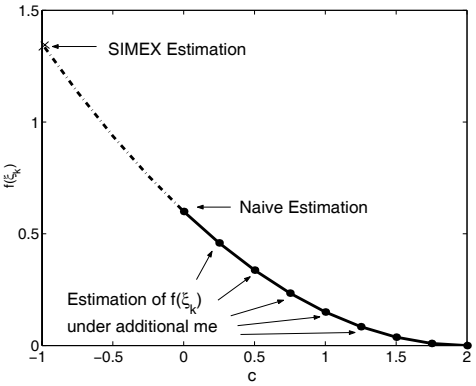


Fig. 1. By inflating the variance for artificially generated error by $c \times \sigma_\delta^2$, the effect of additional error on the estimation $\widehat{f}(\xi_k)$ can be studied. For $c = 0$ we use the original data in the analysis. The curve can be extrapolated to the case of zero measurement error by using quadratic regression.

$\widehat{f}(\xi_k)$ with increasing variance $\sigma_{\delta^*}^2$ of the additional error. Finally we extrapolate on the case of zero measurement error (for all K points of interest). We note that the true error variance σ_δ^2 again has to be estimated or known. Since in fact we generate multiple sets of errors for each chosen error variance SIMEX becomes a computational heavy method. This problem can be taken into account by starting with a subset of basis functions in the RVM to speed up the calculations.

3.4 Simulation results for the SIMEX

We extended an RVM program code by Michael Tipping to perform SIMEX correction. This original RVM code can be found at
`http://research.microsoft.com/mlp/RVM/relevance.htm`.
To check the performance of the SIMEX method, we ran 200 simulations with 201 samples generated from the sinc-function $f(\xi) = \sin(\xi)/\xi$ under Gaussian error $\epsilon \sim \mathcal{N}(0, 0.2^2)$ at data points $\xi \in \{-10, -9.9, \ldots, 10\}$. Measurement error was chosen to be Gaussian $\delta \sim \mathcal{N}(0, 1)$, where we assumed its variance $\sigma_\delta^2 = 1$ to be known for SIMEX.
Figure 2 displays the mean prediction function for the naive RVM and the SIMEX-corrected RVM. The true function and the mean prediction for the RVM using the error free samples from ξ are included as references. The mean naive approximation clearly over-smoothes the data, especially where the curvature of the true function is high. The SIMEX correction does obviously better, particularly at the center of the data but there is also a tendency to over-smooth at the fringe of the data. Figure 3 displays the mean bias and 95% error bars over 200 simulations for the SIMEX and the naive analysis. We identify a generally smaller bias of the SIMEX method but more expanded

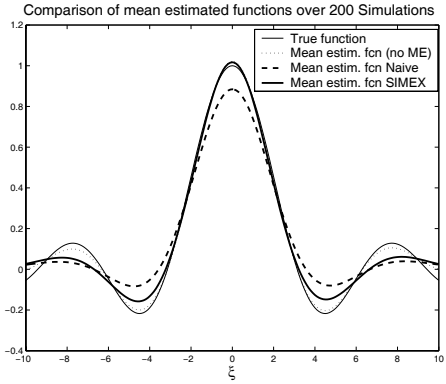


Fig. 2. Comparison of mean prediction using SIMEX and naive analysis under measurement error.

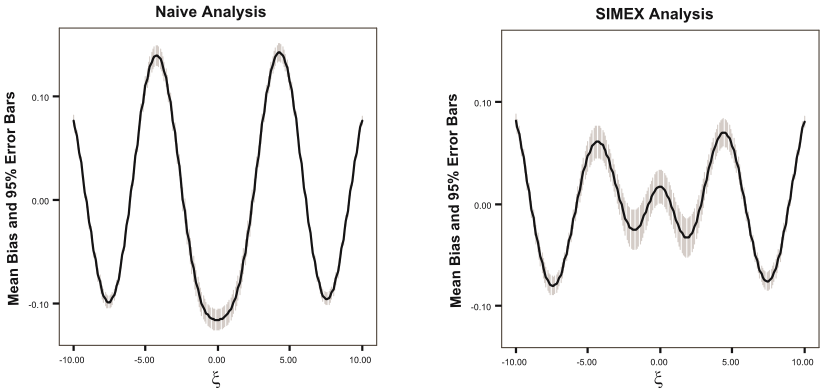


Fig. 3. Mean bias (solid line) and 95% error bars (= 95% confidence interval of the mean bias) of the estimated function using naive analysis and SIMEX, respectively.

error bars, especially at the center of the data. Thus by applying the SIMEX correction we gain in terms of unbiasedness but on the other hand seem to pay for that by higher variability of the estimates.

4 Discussion

In this work we presented a combination of two methods: nonparametric regression with the RVM and covariate measurement error correction using regression calibration and SIMEX, respectively. From the simulation study we see that the SIMEX correction works quite well in this basic case of modeling Gaussian responses and correcting for additive Gaussian measurement error in a single covariate. However, we must keep in mind the computational effort to perform SIMEX. We showed that, for the RVM with Gaussian basis,

inference of the weight parameters under standard regression calibration is equivalent to inference based on naively using the error-prone covariate within a wider basis kernel. A more sophisticated approach is calibration of the basis functions which requires knowledge of the measurement error distribution and involves integration over the nonlinear basis functions. This promising calibration approach will be explored in forthcoming work.

Acknowledgements

Financial support of the German Science Foundation DFG, Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen" is gratefully acknowledged.

References

CARROLL, R.J., RUPPERT, D. and STEFANSKI, L.A. (1995): *Measurement Error in Nonlinear Models*. Chapman & Hall/CRC, London.

CARROLL, R.J., MACA, J.D. and RUPPERT, D. (1999): Nonparametric regression in the presence of measurement error. *Biometrika, 86, 541–554.*

MACKAY, D.J.C. (1994). Bayesian non-linear modelling for the prediction competition. In: ASHRAE Transactions, V.100, Pt.2, ASHRAE, Atlanta Georgia, 1053–1062.

SCHÖLKOPF, B. and SMOLA, A.J. (2002): *Learning with Kernels*. MIT Press, Cambridge, MA.

TIPPING, M.E. (2000): The Relevance Vector Machine. In: S.A. Solla and T.K. Leen and K. R. Müller (Eds.): *Advances in Neural Information Processing Systems*. MIT Press, 652–658.

TIPPING, M.E. (2001): Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research, 1, 211–244.*

VAPNIK, V.(1995): *The Nature of Statistical Learning Theory*. Springer, New York.

Part III

Applications

A Contribution to the History of Seriation in Archaeology

Peter Ihm

Birkenweg 7, D-79183 Waldkirch, Germany *

Abstract. The honour to be the first who published the seriation of archaeological finds by formal methods is attributed by David Kendall (1964) to Sir W. M. Flinders Petrie (1899). According to Harold Driver (1965), an American anthropologist, the earliest numerical seriation studies are those of Kidder (1915), Kroeber (1916), and Spier (1917). It seems, however, that a general acceptance of formal seriation methods did not begin until the pioneering publications of Ford and Willey (1949) and G. W. Brainerd (1951) and W. S. Robinson (1951). Hole and Shaw published an algorithm for permutation search (1967), Elisséeff's (1965) and Goldmann's (1968) methods leading finally to correspondence analysis.

1 Introduction

Today advanced statistical methods are commonly used in archaeology. Mostly inferential and explorative methods are being used, and numerical classification and chronological seriation are primarily applied. While cluster analysis deals with the discovery of distinct groups, seriation aims to bring archaeological finds into chronological order. Only the latter will be subject of the present contribution. For the application of other aspects of data analysis in archaeology see Ihm (2001).

2 The early years

David Kendall (1964, 1971) attributes the merit to Sir W. M. Flinders Petrie (1899) to be the first who used formal seriation methods. The material Petrie analyzed was discovered between 1884 and 1899 in cemeteries near the river Nile. The graves contained dynastic pottery and many other objects. Decades later carbon dating assigned them to a period between 4000 B.C. and 2500 B.C. Since this evidence was not available to him at the time, he proposed a formal method for arranging the graves roughly in their chronological order using a criterion which he called the *concentration principle*. He selected a subsample of 900 graves and 804 types of pottery and did what we call today the 'diagonalisation' of a 900×804 contingency table *grave* $\times$ *type*, mostly subjectively and to some extend with elementary computations. However

* Illustrations have been omitted by a lack of space. The author will provide them on request.

Petrie's records were destroyed and it is now hard to understand what Petrie really did.

According to Harold Driver (1965), an American anthropologist, *"the earliest numerical seriation studies are those of Kidder (1915), Kroeber (1916), and Spier (1917). These men collected pottery fragments from a number of sites in the Southwestern United States, classified the sherds into artifact types, and arranged the raw frequencies or percentages of each type into series which they inferred were temporal sequences. The authenticity was later confirmed by stratification."*

It seems, however, that a general acceptance of formal seriation methods did not begin until the pioneering publications of G. W. Brainerd (1951) and W. S. Robinson (1951). Two years before, Ford and Willey published a manual technique which became the basis of formal methods developed by Elisséeff (1965, 1968, 1970) and Goldmann (1968) and their mathematicians. This led finally to correspondence analysis.

3 Mathematical models

The initial quantification of archaeological evidence deals with the primary translation of archaeological material into a descriptive, numerical language that can provide a starting point for its seriation. The objects of this analysis are single artifacts, *i.e.* items defined by a number of features, or collections of artifacts of different type and arranged in a table called data matrix. The data matrix presents quantified information where the units of concern are listed and described by their scores on a number of variables. In the following, either incidence matrices with presence/absence data or abundance matrices with frequencies (contingency tables) will be treated.

Denote the $m \times n$ data matrix by $\mathbf{P} = (p_{ik})$ with row and column sums

$$p_{i\cdot} := \sum_{k=1}^{n} p_{ik}, \qquad p_{\cdot k} := \sum_{i=1}^{m} p_{ik}, \qquad p_{\cdot\cdot} := \sum_{i=1}^{m}\sum_{k=1}^{n} p_{ik}.$$

Two common models of chronological variation are defined as follows:
Model I: Incidence matrix: Here $p_{ik} \in \{0, 1\}$. Per time unit one type appears $(0 \to 1)$ and another one disappears $(1 \to 0)$. The result is a band matrix, *e.g.*

$$\mathbf{P} = \begin{vmatrix} 1\,1\,1\,0\,0 \\ 0\,1\,1\,1\,0 \\ 0\,0\,1\,1\,1 \end{vmatrix}$$

In French publications the band matrix is also known as *scalogramme parfait* or *privilégié*. Similar models are given by the transpose of the above matrix or a mixture of both.
Model II: Abundance matrix: Here $p_{ik} \geq 0$. In dependence of time, the elements in each column increase to a maximum and then decrease, or the

elements increase, or the elements decrease. Here an example **P** from Kendall (1971) with:

$$\mathbf{P}^T = \begin{vmatrix} 4\,1\,1\,0\,0\,0 \\ 2\,2\,4\,4\,1\,1 \\ 0\,0\,2\,2\,3\,5 \end{vmatrix}$$

This property may be called *unimodality*.

4 The method of Brainerd and Robinson

In two didactically excellent articles, Brainerd (1951), an archaeologist, and Robinson (1951), a sociological statistician, proposed a method based on the concept that each type originates at a given time at a given place, is made in gradually increasing numbers as time goes on, then decreases in popularity until it becomes forgotten, never to recur. This is exactly Model II. Table 1 is based upon data from a Mani collection of Brainerd's. The deposits from three trenches I, II, III were taken stratigraphically, and the letters denote the order of the strata per trench, A indicating the top stratum, i.e. the latest one, B indicating the next below, and so on. Robinson expressed the row frequencies in percent, *i.e.* $p_{i\cdot} = 100$, and computed first the L_1-distance between each pair of rows, measuring their disagreement. The maximum value which the disagreement could have is 200. Then a suitable *agreement coefficient* between rows i and j will be

$$s_{ij} = 200 - \sum_{k=1}^{n} |p_{ik} - p_{jk}|. \tag{1}$$

If the collections are in correct order then the following relations between neighboring row elements of the matrix **S** will hold: For all i

$$s_{ij} \geq s_{i(j+1)}, \qquad \text{for all } j \geq i \quad \text{and} \quad s_{i(j-1)} \leq s_{ij}, \qquad \text{for all } j \leq i. \tag{2}$$

S is called a Robinson matrix. Not all matrices **S** of agreement coefficients can be made to satisfy (2) by rearranging the rows of **P**. Robinson used the number $c_{\mathbf{S}}$ of agreement coefficients not satisfying (2) as measure of deviation from unimodality, but he did not describe an algorithm for the ordering of rows and columns of **S** minimizing $c_{\mathbf{S}}$. However, if the deposits are chronologically arranged along the margins of the table, the totals for the rows and columns will show a typical pattern: Beginning at either end of the chronologically ordered series, the totals will rise progressively to a maximum, and then will decrease progressively to a minimum at the other end of the series. This fact may help to rearrange the matrix more or less 'by hand'. The rearranged abundance matrix is shown in Table 2.

Kendall (1971) proposed an alternative agreement measure instead of (1):

$$s_{ij} = \sum_{k=1}^{n} w_k \{min(p_{ik}, p_{jk})\}. \tag{3}$$

Table 1. Percentages of eight types of pottery in three stratified trenches.

Type Deposits	1	2	3	4	5	6	7	8
IIA	24.0	66.8	1.3	.0	.0	4.0	.0	3.9
IIB	1.4	.9	.0	.0	.0	.0	97.7	.0
IIC	.2	.0	.2	.0	.0	.0	99.3	.3
IA	11.3	.0	3.8	1.3	3.3	24.9	52.6	2.8
IB	.3	.0	.2	.2	.5	1.4	97.4	.0
IIIA	29.6	.0	14.1	.0	.0	7.0	.0	49.3
IIIB	54.3	3.5	14.0	1.8	5.3	7.0	12.3	1.8
IIIC	.0	.0	6.6	3.3	5.5	27.5	57.1	.0

Table 2. Percentages of eight types of pottery in three stratified trenches, after rearrangement of rows.

Type Deposits	1	2	3	4	5	6	7	8
IIA	24.0	66.8	1.3	.0	.0	4.0	.0	3.9
IIIA	29.6	.0	14.1	.0	.0	7.0	.0	49.3
IIIB	54.3	3.5	14.0	1.8	5.3	7.0	12.3	1.8
IA	11.3	.0	3.8	1.3	3.3	24.9	52.6	2.8
IIIC	.0	.0	6.6	3.3	5.5	27.5	57.1	.0
IB	.3	.0	.2	.2	.5	1.4	97.4	.0
IIB	1.4	.9	.0	.0	.0	.0	97.7	.0
IIC	.2	.0	.2	.0	.0	.0	99.3	.3

with strictly positive, but otherwise arbitrary numbers w_k. He showed that if there exists a permutation matrix $\mathbf{\Pi}$ such that $\mathbf{\Pi P}$ has unimodal columns, then $\mathbf{\Pi S \Pi}^T$ has also unimodal rows and columns *i.e.* is a Robinson matrix. However, he did not mention how a permutation matrix $\mathbf{\Pi}$ could be derived. Two mathematical methods were described by Gelfand (1971).

5 Permutation search

It was the merit of Frank Hole (archaeologist) and Mary Shaw (mathematician) (1967) to have overcome the difficulties of permutation search 'by hand' although their method did not become a standard in archaeological numerical methodology. In a worldwide distributed booklet they published an algorithm with only $m \times (m-1)/2 + m^2$ instead of $m!$ permutations to be tested. The algorithm is explained in full extension and applied to the Brainerd-Robinson data, leading to the same solution.

An application of permutation search was presented by Ileana Kivu-Sculy (1971) at the Mamaia Conference 1970 and applied to the seriation of certain inscriptions of the Hellenistic epoch in Romania.

Doran (1971), in his computer analysis of data from the La Tène cemetery at Münsingen-Rain, Switzerland, used Flinders Petrie's concentration prin-

ciple minimising by permutation search $\sum_{k=1}^{n} R_k$ where R_k is the number of rows from the first to the last entry equal 1 in the k-th column of the incidence matrix grave×type.

Another application is that of Simon Régnier (1977) who proposed also an alternative measure of deviation from unimodality.

6 Towards correspondence analysis

How a data matrix could be ordered 'by hand' and without computation of agreement coefficients was demonstrated by Ford.[1] He wrote:

Method of constructing a seriation graph. Frequencies of the types in each collection (=row) are drawn as bars along the top as graph paper strips. These are arranged to discover the type-frequency pattern and are fastened to a paper backing with paper clips. When the final arrangement has been determined, a finished drawing may be prepared. (Ford (1957), quoted from Ford (1962)).

Other technical devices have been constructed to move rows and columns mechanically. The task was simplified with the appearance of electronic computers. 'Scores' were computed for rows and columns and the data matrix arranged according to their value. This procedure lead finally to correspondence analysis.

Mathematically, correspondence analysis is the singular value decomposition of a standardised non-negative data matrix or contingency table. The name goes back to the French *Analyse factorielle des correspondances*. A short history can be found *e.g.* in Ihm and van Groenewoud (1984).

As in Ford's example, data matrices have been analyzed directly by different authors. Early examples are those of Vadim Elisséeff and Klaus Goldmann.

Elisséeff (1965, 1968, 1970) described an analysis of a sample of Chinese archaïc bronzes of type *Yeou large* assigned to a period between 1400 to 800 B.C. He distinguished 19 types, characterised by 16 stylistic and iconographic features, and tried to seriate them as band matrix which he called *scalogramme parfait*. He computed similarity measures between rows and columns of his data matrix type × feature and grouped and seriated the types until he arrived finally at the band matrix of table 3 [2].

I presented a principal component analysis of Elisséeff data in a seminar, 1961, and Brigitte Escofier-Cordier published a correspondence analysis in Benzécri's *Analyse des Données*, vol. 2, p. 321 seq. (1973).

Goldmann (1968) published a chronology of Bronze Age swords, ranging in date from 2000 to 1400 B.C., originating in South Eastern, Central, and Northern Europe. The seriation was based on a number of technical and

[1] According to Driver (1965, p. 320) Ford published this technique already in 1949 (Ford and Willey (1949))

[2] Elisséeff's example is reproduced in Ihm et al. (1978).

Table 3. Elisséeff's 'scalogramme parfait'; *x: feature present, blank: absent.*

Feature								
Type	3	5	15	13	4	6	16	14
A,F,B	x	x		x		x		
C,H		x		x		x	x	
R,J,O,K,S,M			x		x	x	x	
T,U				x	x	x		x
N,X,Z,P					x	x		x

decorative features. [3] In cooperation with the mathematician E. Kammerer he developed a seriation algorithm based on *reciprocal ranking* of rows and columns. The variables are the row and column indices i, k of an $m \times n$ incidence matrix $\mathbf{P} = (p_{ik})$ (artifact × feature) with entries 0 and 1 and row and column sums $p_{i\cdot}$ and $p_{\cdot k}$, respectively. First calculate

$$x_i = \sum_{k=1}^{n} k \cdot p_{ik}/p_{i\cdot} \qquad i = 1, \ldots, m$$

Then the rows of $\mathbf{P}$ are rearranged according to the rank of x_i. In the following step the columns of $\mathbf{P}$ are rearranged in a similar way in increasing order of

$$y_k = \sum_{i=1}^{m} i \cdot p_{ik}/p_{\cdot k} \qquad k = 1, \ldots, n$$

such that, e.g., column with the smallest y-value will be the first one *etc.* These two steps are iterated in turn until convergence. Ties require a special treatment. A disadvantage of this simple procedure was that it could enter into an infinite cycle and a special stopping rule was required. Nevertheless, it could be applied without a computer, and some users cut the table, as Ford did, into paper-strips and glued them on a backing paper. I was told that someone constructed even a simple device to replace row- or column-wise strings of wooden cubes with zero's and one's. Clearly, an experienced computer programmer would not have rearranged rows and columns before the final stop. It was sufficient to replace the type or location indices after each step by their ranks and to compute mean ranks instead of mean row and column numbers.

It took some time until it was understood that not *reciprocal ranking* but *reciprocal averaging* [4] was the method of choice, leading to the matrix equations

$$\mathbf{Py} = \rho\,\mathbf{Rx}, \tag{4}$$

$$\mathbf{P}^T\mathbf{x} = \rho\,\mathbf{Cy} \tag{5}$$

[3] Goldmann's data and results are reported in Ihm and van Groenewoud (1984, p. 30)

[4] After replacement of x_i and y_k by *e.g.* $\{x_i - min(x_i)\}/\{max(x_i) - min(x_i)\}$ and $\{y_k - min(y_k)\}/\{max(y_k) - min(y_k)\}$ to avoid convergence to the trivial solution.

with $\mathbf{R} = \text{diag}(p_{i.})$, $\mathbf{C} = \text{diag}(p_{.k})$, showing that reciprocal averaging is a solving algorithm of the singular value decomposition of the matrix

$$\mathbf{R}^{-1/2}\mathbf{P}\mathbf{C}^{-1/2} \tag{6}$$

with singular values ρ and singular vectors $\mathbf{R}^{1/2}\mathbf{x}$ and $\mathbf{C}^{1/2}\mathbf{y}$, leading to

$$\mathbf{P} = \mathbf{R}\sum_{\nu}\rho_{\nu}\mathbf{x}_{\nu}\mathbf{y}_{\nu}^{T}\mathbf{C} \tag{7}$$

with $\rho_{\nu} \geq \rho_{\nu+1}$ and $\rho_1 = 1$. The solutions of (4) and (5) maximize the canonical correlation coefficients. The singular vectors $\mathbf{x}_1$, $\mathbf{y}_1$ are trivial, and rows and columns of $\mathbf{P}$ are arranged in increasing value of the coordinates of $\mathbf{x}_2$ and $\mathbf{y}_2$. More details can be found *e.g.* in Benzécri (1973), Nishisato (1980), Greenacre (1984), Ihm and van Groenewoud (1984).

Today, instead of seriating rows and columns of a data matrix, $\mathbf{x}_3$ is plotted against $\mathbf{x}_2$, sometimes the plot of $\mathbf{y}_3$ against $\mathbf{y}_2$ is superposed. Some early users of the method were surprised that the plot did not show a more or less elliptical point cluster; the frequent parabolic point configuration became known as 'horseshoe'.

The elements p_{ik} of an ordered abundance matrix from Clarke (1970), plotted against i and k, show a pattern which can be approximated by a two-dimensional normal distribution.[5] A two-dimensional normal density $\phi(x, y)$ with unit variances and correlation coefficient $\rho > 0$ has Hermite polynomials $H_{\nu}(x)$, $H_{\nu}(y)$ as eigenfunctions, $\nu = 0, 1, \ldots$, where ν indicates the degree of the polynomial, $H_2(x)$ corresponding to $\mathbf{x}_3$ *etc.*. Ihm et al. (1978) fitted one-dimensional normal densities $\phi(x)$ to Clarke's data (see also Ihm (1976, 1981) and Ihm and van Groenewoud (1984)). Iwatsubo (1984) gives eigenvectors for some special cases of incidence matrices typical in correspondence analysis. Both cases explain the horseshoe.

Correspondence analysis became a useful tool for the chronological seriation of graves in historic cemeteries. As examples may serve (i) the analyses of two Merovingean cemeteries from Southwest Germany where the data analyses were carried out in cooperation with members of the GfKl (see Roth and Theune (1995), Sasse (2001)) and (ii) a collection of papers from various fields of archaeology and history, edited by Johannes Müller and Andreas Zimmermann (1997).

In a publication of 2001 Groenen and Poblome treat the problem of available extra information on artefact assemblages as *e.g.* the stratification $\text{A} \succ \text{B} \succ \text{C}$ in Robinson's example and describe how a *constrained correspondence analysis* can be obtained.

In my opinion, correspondence analysis became as popular because the maximization of a correlation coefficient is a principle common to many other

[5] Clarke's example is reproduced in Ihm et al. (1978) and Ihm and van Groenewoud (1984)

procedures and easily understood. It might not be generally known that correspondence analysis was first proposed by Hans Otto Hirschfeld (1935), better known as H. O. Hartley, who determined for a contingency table **P** 'scores' x_i, y_k such that both regressions are linear.

References

BENZECRI, J.-P. (1973): L'Analyse des Données. II. L'Analyse des Correspondances. Dunod, Paris.

BORILLO, M. et al. (Eds.)(1977): Raisonnement et Méthodes mathématiques en Archéologie. CNRS, Paris.

BRAINERD, G.W. (1951): The place of chronological ordering in archaeological analysis. *American Antiquity, 16, 301–313.*

CLARKE, D.L. (1970): Beaker pottery of Great Britain and Ireland. Vol. 1, Cambridge.

DIDAY, E. (Ed.) (1984): Data Analysis and Informatics, III. Elsevier Sciences Publishers (North Holland).

DORAN, J.E. (1971): Computer analysis of data from the La Tène cemetary of Münsingen-Rain. In: F.R. Hodson, D.G. Kendall and P. Tăutu (Eds.): *Mathematics in the Archaelogical and Historical Sciences*, 422–431.

DORAN, J.E. and HODSON, F.R. (Eds.) (1975): *Mathematics and Computers in Archaeology.* Edinburgh University Press, Edinburgh.

DRIVER, H. (1965): Survey of numerical classification in anthropology. In: D. Hymes (Ed.): *The Use of Computers in Anthropology*, 301–344.

ELISSEEFF, V. (1965): Possibilités du scalogramme dans l'étude des bronzes chinois. *Mathématiques et Sciences humaines n. 11.*

ELISSEEFF, V. (1968): De l'application des propriétés du scalogramme à l'étude des objets. In: *Calcul et formalisation dans les sciences de l'homme*, 107–120.

ELISSEEFF, V. (1970): Données de classement fournies par les scalogrammes privilégiés. In: J.-C. Gardin, M. Borillo (Eds.): *Archéologie et Calculateurs*, 177–186.

ESCOFIER-CORDIER, B. (1973): Recherche d'un scalogramme par l'analyse factorielle. In: J.P. Benzecri: *L'Analyse des Données. II. L'Analyse des Correspondances*, 321–325.

FORD, J.A. (1957): Método cuantitativo para determinar la cronología arqueológica. *Divulgaciones Etnológicas, 6*, 9–11.

FORD, J.A. (1962): A Quantitative Method for Deriving Cultural Chronology. Pan American Union, Washington, D. C.

FORD, J.A. and WILLEY, G.R. (1949): Surface survey of the Virù Valley, Peru. *Anthropological Papers of the American Museum of Natural History, 43, 1–89.*

GARDIN, J.-C. and BORILLO, M. (Eds.) (1970): Archéologie et Calculateurs. CNRS, Paris.

GELFAND, A.E. (1971): Rapid seriation methods with archaeological applications. In: F.R. Hodson, D.G. Kendall and P. Tăutu (Eds.): *Mathematics in the Archaelogical and Historical Sciences*, 186–201.

GOLDMANN, K. (1968): Zur Auswertung archäologischer Funde mit Hilfe von Computern. *Die Kunde, 19, 122–129.*

GREENACRE, M.J. (1984): Theory and Applications of Correspondence Analysis. *Academic Press, London.*

GROENEN, P.J.F. and POBLOME, J. (2001): Constrained correspondence analysis for seriation in archaeology applied to Sagalassos ceramic tablewares. In: M. Schwaiger and O. Opitz (Eds.): *Proc.* 25^{th} *Ann. Conf. of GfKl*, 90–97.

HIRSCHFELD, H.O. (1935): A connection between correlation and contingency. *Proc. Camb. Phil. Soc., 31, 520–525.*

HODSON, F.R., KENDALL, D.G. and TĂUTU, P. (Eds.) (1971): *Mathematics in the Archaelogical and Historical Sciences.* Edinburgh University Press, Edinburgh.

HOLE, F. and SHAW, M. (1967): Computer Analysis of Chronological Seriation. *Rice University Studies, 53(3), 1–166.*

HYMES. D. (Ed.) (1965): *The Use of Computers in Anthropology.* Mouton, The Hague.

IHM, P. (1976): Chronologische Seriation, ein Problem statistischer Schätzung. *IX. Congrès UISPP: Banques de données et méthodes formelles en achéologie et protohistoire.* Nice, 133–140.

IHM, P. (1981): The Gaussian model in chronological seriation. *X. Congreso UISPP: Manejo de datos y metodos matematicos de arqueologia.* Mexico City, 108–124.

IHM, P. (2001): Archäologie und historische Wissenschaften (AG ARCH). In: H.-H. Bock and P. Ihm (Eds.): *25 Jahre Gesellschaft für Klassifikation*, Shaker, Aachen, 66–71.

IHM, P. and VAN GROENEWOUD, H. (1984): Correspondence analysis and Gaussian ordination. *Compstat Lectures, 3, 1–60.*

IHM, P., LÜNING, J. and ZIMMERMANN, A. (1978): *Statistik in der Archäologie.* Rheinland Verlag, Köln.

IWATSUBO, S. (1984): The analytical solutions of eigenvalue problems in the case of applying optimal scoring methods to some types of data. In: E. Diday (Ed.): *Data Analysis and Informatics, III.* Elsevier Sciences Publishers (North Holland), 31–39.

KENDALL, D.G. (1964): A statistical approach to Flinders Petrie's sequence-dating. *Bulletin of the International Statistical Institute, 40, 657–681.*

KENDALL, D.G. (1971): Seriation from abundance matrices. In: F.R. Hodson, D.G. Kendall and P. Tăutu (Eds.): *Mathematics in the Archaeological and Historical Sciences*, 214–252.

KIDDER, A.V. (1915): Pottery of the Pejarito Plateau and of some adjacent regions in New Mexico. *American Anthropological Association, Memoir 2*, 407–462.

KIVU-SCULY, I. (1971): On the Hole-Shaw method of permutation search. In: F.R. Hodson, D.G. Kendall and P. TĂUTU (Eds.): *Mathematics in the Archaelogical and Historical Sciences*, 253–254.

KROEBER, A.L. (1916): Zuni potsherds. *Anthropological Papers of the American Museum of Natural History, 18, 1–38.*

MÜLLER, J. and ZIMMERMANN, A. (Eds.) (1997): *Archäologie und Korrespondenzanalyse. Beispiele, Fragen, Perspektiven.* Leidorf, Espelkamp.

NISHISATO, S. (1980): *Analysis of Categorial Data: Dual Scaling and its Applications.* Univ. of Toronto Press, Toronto.

PETRIE, W.M.F. (1899): Sequences in prehistoric remains. *Journal of the Anthropological Institute, 29, 295–301.*

REGNIER, S. (1977): Sériation des niveaux de plusieurs tranches de fouille dans une zone archéologique homogène. In: M. Borillo et al. (Eds.): *Raisonnement et Méthodes Mathématiques en Archéologie. CNRS, Paris.*

ROBINSON, W.S. (1951): A Method for chronologically ordering archaeological deposits. *American Antiquity, 16, 293–301.*

ROTH, H. and THEUNE, C. (1995): *Das frühmittelalterliche Gräberfeld bei Weingarten I.* K. Theiss Verlag, Stuttgart.

SASSE, B. (2001): *Ein frühmittelalterliches Reihengräberfeld bei Eichstetten am Kaiserstuhl.* K. Theiss Verlag, Stuttgart.

SCHWAIGER, M. and OPITZ, O. (Eds.) (2001): Exploratory Data Analysis in Empirical Research.Proc. 25^{th} Ann. Conf. of GfKl, Springer, Heidelberg.

SPIER, L. (1917): An outline for a chronology of Zuni ruins. *Anthropological Papers of the American Museum of Natural History, 18, 207–331.*

Model-based Cluster Analysis of Roman Bricks and Tiles from Worms and Rheinzabern

Hans-Joachim Mucha[1], Hans-Georg Bartel[2], and Jens Dolata[3]

[1] Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS), Mohrenstraße 39, 10117 Berlin, Germany
[2] Institut für Chemie, Humboldt-Universität zu Berlin, Brook-Taylor-Straße 2, 12489 Berlin, Germany
[3] Landesamt für Denkmalpflege Rheinland-Pfalz, Abt. Archäologie, Amt Mainz, Große Langgasse 29, 55116 Mainz, Germany

Abstract. Chemical analysis of ancient ceramics has been used frequently to support archaeological interpretation. Often the dimensionality in the data has been high. Therefore multivariate statistical techniques like cluster analysis have been applied. Successful applications of simple model-based Gaussian clustering of Roman bricks and tiles has been reported by Mucha et al. (2001). And now, more complex Gaussian models can be investigated because of an increase of sample size by new findings excavated in Boppard. Additionally these and previous successful simple models will be applied in a very local fashion considering two supposed brickyards only. Here, after giving a brief history of clustering Roman bricks and tiles, some cluster analysis models including different data transformations will be investigated in order to answer questions like: Is it possible to differentiate between brickyards of Rheinzabern and Worms on basis of chemical analysis? Do the bricks and tiles found in Boppard belong to the brickyards of Worms or Rheinzabern?

1 Introduction and task

Cluster analysis can support researchers in interesting application areas like archaeology. In the last decade, the advantage of cluster analysis (Mucha (1992)) as the most commonly used multivariate technique in archaeometry has been taken to investigate about 600 samples of bricks and tiles from the northern part of the former Roman Empire's province *Germania Superior*. There has been developed a complex model of history and relations of the brick and tile production by archaeologist and now it is proposed to consolidate these ideas. The aim of that clustering respecting 19 chemical elements measured with X-ray fluorescence analysis (XRF) was both to confirm supposed sites of brickyards and to find places of those ones that are not yet identified. The obtained results were published in a few tens previous papers, among them Bartel et al. (2003), Dolata (2000, 2001), Dolata et al. (2003), Mucha et al. (2001), and Werr (1998). The data itself was published by Dolata (2000) (see pages 53–67). As a result the following military brickyards could be established (in brackets the number of objects): Frankfurt-Nied (137), Groß-Krotzenburg (63), Rheinzabern (192), Straßburg-Königshofen (113),

Worms (19) and two with respect to their provenience not yet known ones (67 and 7 respectively). The corresponding bivariate nonparametric density estimation (Figure 1) shows two mountain chains. The four mountain heads on the left side could be identified with not yet known 1, Rheinzabern A and B, and Worms (from left to right). But there is not a sufficient possibility to distinguish the several tops, especially Worms proves almost to be only a slope of Rheinzabern B.

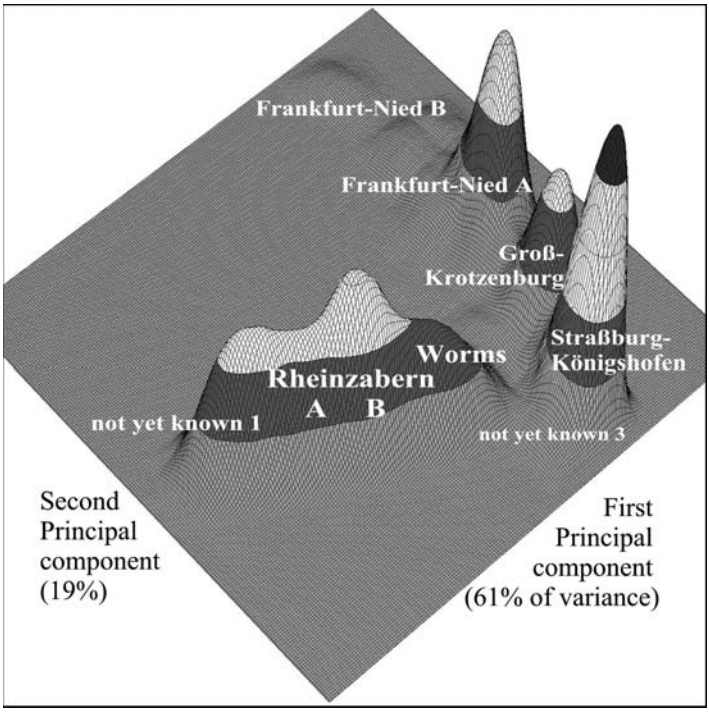


Fig. 1. Bivariate density estimation based on the first two principal components

Already there exist geochemical reference-data from XRF for the products belonging to two of these production-fields (Worms: $n = 19$, Rheinzabern: $n = 192$). Both brickyards are close together from geography, geology and chronology . The chemical components of the references are characteristic for the whole production of the two sites. Now an enlarged number of 47 samples analysed by XRF and archaeologically assigned to Worms was placed at our disposal (see Figure 2 for an example). The new data was published by Mucha et al. (2004). Therefore 66 objects from Worms are available altogether until now. Using this set of samples joined with that of the 192 ones assigned to Rheinzabern it could be asked if it is possible in a mathematical way to find a stable "Worms" - cluster that is well separated from the brickyard(s) in

Fig. 2. Brick (*later*) from Boppard (LDA Koblenz, Inv. 64/440 = Boppard ZS8)

Rheinzabern by its chemical components only. Thus the main question is: Do the bricks and tiles from Boppard belong without doubt to the Worms-references? Since the number of samples from Worms is more than three times greater than the number of variables the most complex model-based Gaussian clustering can be tried using different data pre-processing transformations.

2 Model-based Gaussian clustering

Here model-based Gaussian clustering techniques are applied in archaeometry in order to set up new hypotheses about the data under investigation. Concerning model-based clustering the paper of Banfield and Raftery (1993) gives a good insight into the topic. Let $\mathbf{X}$ be the $(I \times J)$-data matrix consisting of I observations and J variables. The most general model-based Gaussian clustering is if the covariance matrix $\mathbf{\Sigma}_k$ of each cluster k is allowed to vary completely. Then the log-likelihood is maximized whenever the partition P(I,K) of I observations into K clusters minimizes

$$Y_K = \sum_{k=1}^{K} n_k \log |\frac{\mathbf{W}_k}{n_k}|. \tag{1}$$

Herein $\mathbf{W}_k = \sum_{i \in C_k} (\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T$ is the sample cross-product matrix for the k-th cluster C_k, and $\overline{\mathbf{x}}_k$ is the usual maximum likelihood estimate of expectation values in cluster C_k. The cardinality of cluster C_k is denoted by n_k. When the covariance matrix of each cluster is constrained to be $\mathbf{\Sigma}_k = \lambda \mathbf{I}$, the well-known sum-of-squares criterion

$$V_K = \sum_{k=1}^{K} tr(\mathbf{W}_k), \tag{2}$$

has to be minimized. When the covariance matrix of each cluster is constrained to be $\mathbf{\Sigma}_k = \lambda_k \mathbf{I}$, the logarithmic sum-of-squares criterion

$$Z_K = \sum_{k=1}^{K} n_k \log tr(\mathbf{W}_k / n_k), \tag{3}$$

has to be minimized. The last two cluster analysis models are dependent on scales. In order to apply such models to data with quite different scales the variables have to be transformed. More generally, a transformation can be considered as weighting the variables. For instance, criterion (2) can be written as

$$V_K = \sum_{k=1}^{K} \sum_{i \in C_k} d_Q^2(\mathbf{x}_i, \overline{\mathbf{x}}_k), \tag{4}$$

where $d_Q^2(\mathbf{x}_i, \overline{\mathbf{x}}_k) = (\mathbf{x}_i - \overline{\mathbf{x}}_k)^T \mathbf{Q} (\mathbf{x}_i - \overline{\mathbf{x}}_k)$ is the squared weighted Euclidean distance with $\mathbf{Q}$ diagonal. Because of successful results (see the references concerning the history above, for instance Mucha et al. (2001)) we recommend either the use of the special weights

$$q_{jj} = 1/\overline{x}_j^2 \tag{5}$$

(see Underhill and Peisach (1985)) or the use of adaptive weights like diagonal elements proportional to the inverse pooled within-cluster variances

$$q_{jj} = 1/\overline{s}_j^2, \tag{6}$$

where

$$\overline{s}_j^2 = 1/K \sum_{k=1}^{K} \sum_{i=1}^{I} \delta_{ik} (x_{ij} - \overline{x}_{kj})^2. \tag{7}$$

is the pooled standard deviation of the variable j. The indicator function δ_{ik} is defined in the usual way: $\delta_{ik} = 1$, if observation i comes from cluster k, or $\delta_{ik} = 0$ otherwise.

The weights can be estimated in the adaptive K-means method in an iterative manner (Mucha (1992)). Figure 3 shows a low-dimensional projection of the data. Here, the adaptive K-means method presents two well separated clusters. It should be mentioned that the principal component analysis (PCA) based on the correlation matrix presents only one cloud of points (Mucha et al. (2004)). Additionally the two interesting objects "H880" and "G139" are marked (see the discusssion below).

In terms of transformations the special weights (5) can be obtained by preprocessing the original data matrix by dividing each column (variable) by its arithmetic mean. This quite simple transformation has the useful property that the relative variability in the original variables become the variability of the transformed ones. As a consequence of this transformation the original variables, measured in quite different scales, become comparable one with each other. The arithmetic mean of each new variable is equals 1. Moreover the variables preserve their different original variability and therefore have different influence (contribution to) on the sum of squares criterion (2) as well as the logarithmic sum of squares criterion (3). Obviously the results of data pre-processing are influenced by going from global data analysis of about 600 observations to the local one here. For instance, the chemical trace

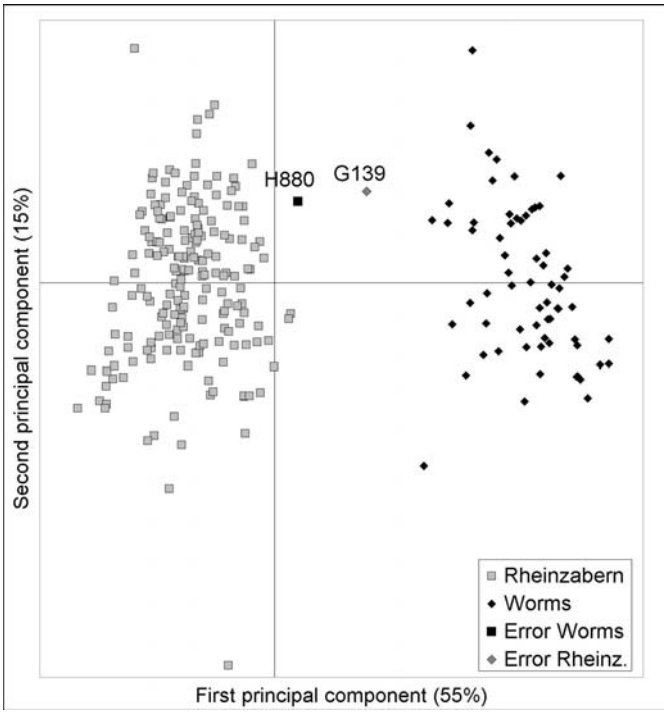


Fig. 3. PCA plot based on the covariance matrix after transforming by (6).

element Zr becomes much more important in the local cluster analysis (for details, see Mucha et al. (2004)). Figures 4, 5 and 6 show the same data as Figure 3. Additionally the two interesting objects "H880" and "H857" are marked (see the discussion below). The transformations above do not affect the most general (unconstrained) Gaussian model (1) and the "constant" within clusters covariance matrix model minimizing

$$U_K = |\sum_{k=1}^{K} \mathbf{W}_k|, \tag{8}$$

where it is supposed the covariance matrix is uniform across all clusters. Here nonlinear transformations of Box-Cox-type (Box and Cox (1964)) are recommended with the special cases of transforming logarithmically (Papageorgiou et al. (2001)) or taking the squareroots of the values.

3 Results and archaeological discussion

The logarithmic sum-of-squares criterion using the weights (5) performs best. That can be confirmed by simulations. From archeological point of view there

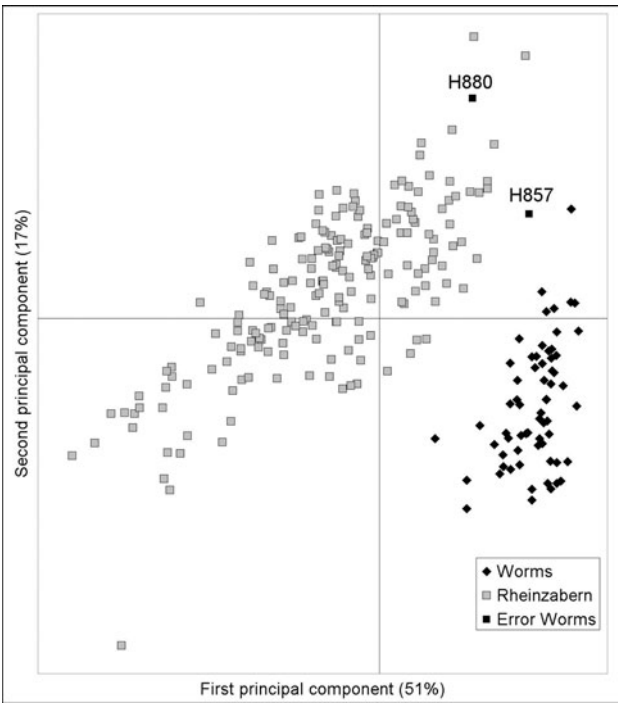


Fig. 4. PCA plot based on the covariance matrix after transforming by (5).

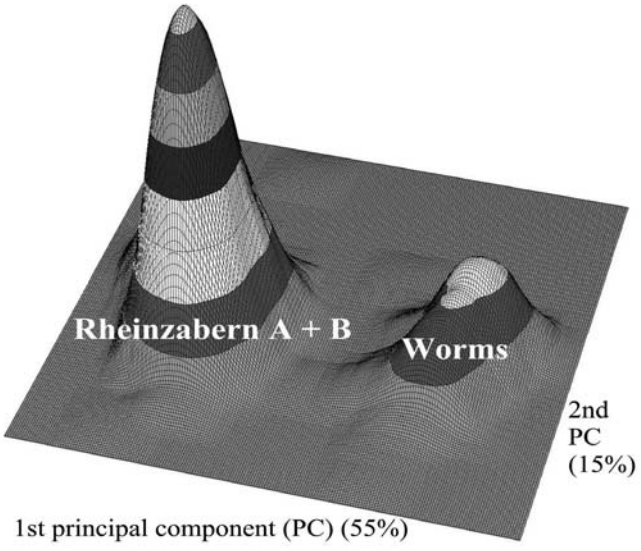


Fig. 5. Density plot based on the first two PC after transforming by (6).

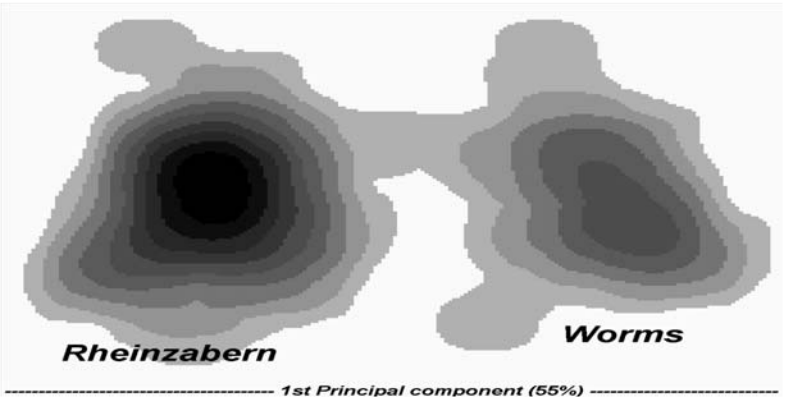


Fig. 6. Several cuts of density based on the first two PC after transforming by (6).

is one error only (see the next paragraph below). Figure 4 shows the clustering results. Similar results can be obtained by using the weights (6) in the simple models (2) and (3) (see Figure 3). The results of the most complex model (1) are quite instable and highly dependent on the initial partition. For instance, taking as initial partitions the ones from Figure 4 or Figure 3 the determinant criterion works best, but with at least six "errors". The model (8) performs similar as (1).

Almost all (except one) of the new analysed bricks and tiles from Boppard are geochemical homogeneous and belong without any doubt to the brickyard of Worms. These products present brick stamps of the so-called Flörsheim group of the *legio XXII Primigenia* (Dolata (2001)). Without exception all the different types of brick stamps (Boppard type 1 – 6) do belong to the references of Worms. The tile (*tubulus*) H 880 is an exception that was not furnished with a stamp. It could be assigned now to the area of Rheinzabern. That is confirmed by both clusterings (Figure 3 and Figure 4). Hence it follows that it is from a late repair. *Tubuli* are flue-tiles that are a part of the heating-system *hypocausis*. They were worn out by effects of heat and therefore they could have been replaced after years of utilization. The repair happened in a time where no supply from Worms was possible or the brickyards of Worms did no longer exist. In the epoch of the Emperor *Valentinianus I* (364–375 A.D.) the first century brickyard of Rheinzabern has been reopened. It is possible that the *milites menapii*, of which a single brick stamp has been found at Boppard, also have manufactured the *tubulus* examined. The following two interesting objects are classified different, see Figure 4 and Figure 3. H 857 is geochemically aside the other samples but belongs certainly to the Worms-products. Because of the well known inhomogeneity of coarse ware products (Werr (1998)) there is no problem to interpret that fact. G 139 has changed its membership from Rheinzabern-references to Worms. The sample is of a column for a *hypocausis* and presents no brick stamp. Because of broad

basis the new integration is the statistical better one. Here is no problem for archaeology to interpret these suggestions.

4 Conclusion

Simple Gaussian models perform best in the case of appropriate data transformation. Therefore the recommendation is adaptive weights transformation or "coefficient of variation" - transformation. Complex models are quite instable because of the high number of dimensions regarding the sample size. They perform well only in the case of low dimensionality of the data and if the structure in the data has been very clear (see Tubb et al. (1980) and Papageorgiou et al. (2001)). Logarithmic and square root data transformations perform similar in complex Gaussian models in this application.

References

BANFIELD, J.D. and RAFTERY, A.E. (1993): Model-Based Gaussian and non-Gaussian Clustering. *Biometrics, 49, 803–821.*

BARTEL, H.-G., MUCHA, H.-J., and DOLATA, J. (2003): Über eine Modifikation eines graphentheoretisch basierten partitionierenden Verfahrens der Clusteranalyse. *Match 48, 209–223.*

BOX, G.E.P. and COX, D.R. (1964): An analyis of transformations. *J. R. Statist. Soc. B 26, 211–252.*

DOLATA, J. (2000): *Römische Ziegelstempel aus Mainz und dem nördlichen Obergermanien.* Dissertation, Johann Wolfgang Goethe-Universität, Frankfurt.

DOLATA, J. (2001): Römische Ziegelstempel der sogenannten Flörsheimer Gruppe. *Flörsheimer Geschichtshefte 3, 4–7.*

DOLATA, J., MUCHA, H.-J., and BARTEL, H.-G. (2003): Archäologische und mathematisch-statistische Neuordnung der Orte römischer Baukeramikherstellung im nördlichen Obergermanien. *Xantener Berichte 13, 381–409.*

MUCHA, H.-J. (1992): *Clusteranalyse mit Mikrocomputern.* Akademie Verlag, Berlin.

MUCHA, H.-J., DOLATA, J., and BARTEL, H.-G. (2001): Validation of Results of Cluster Analysis of Roman Bricks and Tiles. In: W. Gaul and G. Ritter (Eds.): *Classification, Automation, and New Media.* Springer, Berlin, 471–478.

MUCHA, H.-J., BARTEL, H.-G., and DOLATA, J. (2004): Modellbasierte Clusteranalyse römischer Ziegel aus Worms und Rheinzabern. *Archäologische Informationen 26 (2), 471–480.*

PAPAGEORGIOU, I., BAXTER, M.J., and CAU, M.A. (2001): Model-based cluster analysis of artefact compositional data. *Archaeometry 43 (4), 571–588.*

TUBB, A., PARKER, A.J., and NICKLESS, G. (1980): The analysis of Romano-British pottery by atomic absorption spectrophotometry. *Archaeometry 22 (2), 153–171.*

UNDERHILL, L.G. and PEISACH, M. (1985): Correspondence analysis and its application in multielement trace analysis. *J. Trace and microprobe techniques 3 (1 & 2), 41–65.*

WERR, U. (1998): Grenzen der Aussagekraft chemischer Analytik für römische Baukeramik. *Archäometrie und Denkmalpflege, 96–98.*

Astronomical Object Classification and Parameter Estimation with the Gaia Galactic Survey Satellite

Coryn A.L. Bailer-Jones

Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

Abstract. Gaia is a cornerstone mission of the European Space Agency (ESA) which will undertake a detailed survey of over 10^9 stars in our Galaxy. This will generate an extensive, multivariate, heterogeneous data set which presents numerous problems in classification, regression and time series analysis. I give a brief overview of the characteristics and requirements of this project and the challenges it provides.

1 The Gaia Galactic survey mission

Gaia is a future satellite mission which will study our Galaxy in unprecedented detail (ESA (2000); Perryman et al. (2001)). Its objective is to study its composition, origin and ultimate evolution by determining the properties of over one thousand million stars in different populations across our entire Galaxy. One of the major contributions of Gaia is that it will measure distances to stars with much higher precision than is currently possible. Distance measurement is a very important (and difficult) task in astrophysics, as only with distances can we properly map structure in the Galaxy and determine fundamental stellar properties (e.g. absolute brightness). Gaia will also measure the space motions of stars in exquisite detail, which will be used in sophisticated dynamical models to map out the distribution of matter and is an important component in testing models of Galaxy formation.

2 Astrophysical data

Much of this so-called astrometric data from Gaia would be of little value if we did not know the intrinsic properties of the stars observed, quantities such as the temperature, mass, chemical compositions, radius etc. (collectively referred to as *Astrophysical Parameters*, or APs; see Bailer-Jones (2002b)). For this reason, Gaia is equipped with two photometric instruments which sample the stellar spectral energy distributions (or spectra) at discrete locations, producing *photometric* data. The first of these instruments measures the spectra at five locations, the second at about ten. (The optimization of these filter systems is ongoing; for more details on this and the sampling of stellar spectra, see my other contribution in these proceedings.) Together these data provide

a 15-dimensional data space from which we need to determine at least four APs for a very wide range of types of stars. For many of these stars we have (or can gather or simulate) reasonable quality pre-classified data which may be used as templates in a supervised classification/regression model (e.g. Hastie et al. (2001)), such as neural networks or minimum distance methods (Allende Prieto (2003); Bailer-Jones et al. (1998); Bailer-Jones (2002a); Folkes et al. (1996)).

The problem is, however, considerably more complex. First, each star is observed about 100 times over the course of the mission, and many of these stars are variable, i.e. their photometric measures vary over time on a range of time scales. This is both a problem and a benefit, because for some objects the way in which they vary is a significant source of information for determining their intrinsic properties (both the primary APs and additional characteristics). Second, not all of the objects which Gaia observes are stars. Gaia observes 'blind', that is, it observes every single object in the sky brighter than some level without any prior selection or information on what the objects are. Many of these objects will be single stars, but many other types of objects will be observed, including galaxies, quasars, asteroids and unresolved binary stars. Therefore, before we can even try to determine APs, we must perform a discrete classification to see whether the object is a type of star we are interested in. In some cases we can use morphological information, i.e. we get an image which is not simply a point source (typically for some – but not all – galaxies). Using this is of course an involved image classification problem in its own right (Naim et al. (1995)). In many cases we have no such morphological information, so we must perform the classification using the photometric data.

The classification problem is complicated further by the presence of a third instrument which will measure the entire stellar spectrum of each star over a narrow wavelength region. The spectrum covers some 500 elements. While we can certainly apply dimension reduction techniques to these data, it nonetheless provides considerably more independent information on the primary APs. Moreover there are several additional astrophysical parameters which we want to determine from these spectra. A particular challenge is combining these data with the two sets of photometric data.

3 Classification challenges

Gaia will produce a complex data set, the proper exploitation of which presents us with a number of significant challenges. The objectives can be summarised as follows:

- Discrete classification of objects: discriminate between single stars, multiple stars, galaxies, quasars, supernovae, asteroids etc.
- For single stars, determine their astrophysical parameters (APs), the exact number of which and the precision with which they should be estab-

lished depending on their type. There are four primary APs and several subsidiary ones. We will probably also want to perform a (discrete) classification of these stars into astrophysically relevant groups.
- Provide for the efficient identification of new types of objects for which we have no (or little) prior knowledge, i.e. employ unsupervised methods or outlier detection techniques.

The practical requirements of the classification system may be summarized as follows:

- Cope with missing data (e.g. due to partial instrument failure or 'down time') and deal with censored data (i.e. upper or lower limits on a measure due to the limited sensitivity or dynamic range of an instrument) in an unbiased fashion. In some cases an upper limit on a non-detection is an important indication of the type of star.
- Quantify uncertainties via probabilities of class membership; this must take account of the fact that some stars may be members of more than one class.
- Provide accurate estimates of AP uncertainties. The input data and the resulting APs will sometimes have correlated errors. Typically we have a good noise model for our instruments although the correlations between them are harder to characterize.
- The APs are not independent and they do not have an isolated effect on the data. For this reason, we cannot independently infer each AP in a multivariate regression.
- Cope with degeneracies. A degeneracy means that different objects can appear the same in the data space (within the expected measurement errors), especially at low signal-to-noise ratios. Some degeneracies are intrinsic and known to exist but have not been mapped out in detail. Degeneracies must be recognised and different classifications/sets of APs provided where appropriate (along with associated probabilities).
- Make efficient use of variability (time series) information. The classification systems should be insensitive to variability where it is not relevant (e.g. due to noise or errors) but recognise and exploit it where it is relevant (for certain types of stars).
- In some cases we have prior information on the APs of specific objects; making efficient use of this is a challenge.

Clearly, classification and parameter estimation with Gaia cannot be solved with a simple one-step approach. It will probably have to employ many different techniques operating in a hierarchical or iterative fashion. An outline framework for such an approach is shown in Fig. 1.

4 Outlook

Gaia will be launched in 2010 at a cost of some 450 million Euro. The data analysis – including the classification – will be undertaken by a dedicated

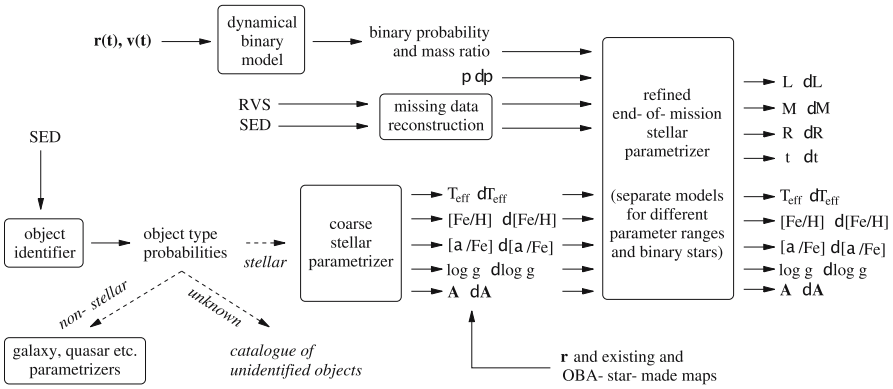


Fig. 1. A possible approach to the determination of physical stellar parameters with Gaia. SED = Spectral Energy Distribution (photometry and/or spectrum); nSED is the flux normalized version of this. RVS = Radial Velocity Spectrum, a high resolution spectrum in the region 850–875 nm containing important diagnostic information. r(t) and v(t) refer to the positional and kinematic information obtained as a function of time; the various parameters emerging on the right are the astrophysical parameters of interest, such as the effective temperature ($T_{\rm eff}$) and the chemical abundance ([Fe/H]). Not all elements of the system are shown. For example, parallax and proper motion information which only become available at the end of the mission are useful for identifying extragalactic objects, and variability is an important means of identifying a number of types of stars.

but geographically distributed consortium of astronomers, computer scientists and statisticians. For more information on the mission, see the Gaia web site at `http://www.esa.int/science/gaia`. The classification issues are being addressed by a dedicated working group, ICAP, which stands for *Identification, Classification and Astrophysical Parametrization*. Its web site, which gives more details on the problem, data and techniques currently being used, is `http://www.mpia.de/GAIA`

References

ALLENDE PRIETO, C. (2003): An automated system to classify stellar spectra – I. *Monthly Notices of the Royal Astronomical Society*, 339, 1111–1116.

BAILER-JONES, C.A.L., IRWIN, M. and VON HIPPEL, T. (1998): Automated classification of stellar spectra. II: Two-dimensional classification with neural networks and principal components analysis. *Monthly Notices of the Royal Astronomical Society*, 298, 361–377.

BAILER-JONES, C.A.L. (2002a): Automated stellar classification for large surveys: a review of methods and results. In: R. Gupta, H.P. Singh, C.A.L. Bailer-Jones (Eds.): *Automated Data Analysis in Astronomy*. Narosa Publishing House, New Delhi, 83–98.

BAILER-JONES, C.A.L. (2002b): Determination of stellar parameters with GAIA. *Astrophysics and Space Science*, 280, 21–29.

ESA (2000): GAIA: Composition, formation and evolution of the Galaxy, ESA-SCI(2000)4.

FOLKES, S.R., LAHAV, O. and MADDOX, S.J. (1996): An artificial neural network approach to the classification of galaxy spectra, *Monthly Notices of the Royal Astronomical Society*, 283, 651–665.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *The elements of statistical learning: data Mining, inference and prediction.* Springer, Heidelberg.

NAIM, A., LAHAV, O., SODRE JR., L. and STORRIE-LOMBARDI, M.C.(1995): Automated morphological classification of APM galaxies by supervised artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 275, 567–590.

PERRYMAN, M.A.C., DE BOER, K.S., GILMORE, G., HØG, E., LATTANZI, M.G., LINDGREN, L., LURI, X., MIGNARD, F., PACE, O., DE ZEEUW, P.T. (2001): Gaia: Composition, formation and evolution of the Galaxy. *Astronomy & Astrophysics*, 369, 339–363.

Design of Astronomical Filter Systems for Stellar Classification Using Evolutionary Algorithms

Coryn A.L. Bailer-Jones

Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany

Abstract. I present a novel method for designing filter systems for astrophysical surveys. The filter system is designed to optimally sample a stellar spectrum such that its astrophysical parameters (APs: temperature, chemical composition etc.) can be determined using supervised regression methods. The problem is addressed by casting it as an optimization problem: A figure-of-merit (FoM) is constructed which measures the ability of the filter system to 'separate', in a vectorial sense, stars with different APs; this FoM is then maximized with respect to the parameters of the filter system using an evolutionary algorithm. The resulting filter systems are found to be competitive in performance with conventionally designed systems.

1 Astrophysical context

Astrophysics relies on large statistical surveys of astronomical objects for advancing our understanding of the cosmos. In stellar astrophysics, for example, by measuring the spectral energy distribution (spectrum) of many different types of stars across our Galaxy we can gain insight into the formation and evolution of stars and of the Galaxy itself. Ideally we would obtain high quality spectra of literally billions of stars, from which we can determine stellar intrinsic properties, or *astrophysical parameters (APs)*, continuous quantities such as the temperature, chemical composition and surface gravity. However, for various technical reasons detailed spectroscopy on so many objects is not possible. Instead, we must limit ourselves to photometry, that is, coarsely sampling a spectrum at pre-defined locations with a filter system (for an example see Fig. 3). By analysing the spectrum of a specific star in detail, we could design a filter system which is adequate for determining the APs of that type of star to some desired accuracy. However, large surveys must observe *many* different types of stars with a *single* filter system. Hence this filter system must be some kind of optimal average system, the design of which is furthermore subject to numerous instrumental constraints.

A number of upcoming surveys are therefore faced with the difficult question of how to define their optimal filter system. Existing filter systems were designed for more specific purposes or for more restricted classes of objects, so are not appropriate for these new surveys. The 'conventional' approach to this problem is to manually modify existing systems based on the best

of our astrophysical knowledge. Yet given the numerous conflicting requirements placed on the filter system, this is unlikely to be very efficient or even successful. Moreover, one would still not know whether a better filter system could be constructed within the constraints.

In this paper I outline a more systematic approach to the filter design problem called Heuristic Filter Design (HFD). Given a set of stellar spectra (the *grid*) with known APs, we perform a directed search for a filter system which optimally samples these spectra such that we can best determine their various APs. The constraints of the problem are represented by the (fixed) instrument model, which defines the size of telescope, noise in detectors etc. Given a filter system, this instrument model allows us to calculate the amount of light (number of photons per unit wavelength) measured for each star in each filter, plus their expected errors. This is used to calculate the performance of the filter system (the FoM, section 2.2).

The spectral grid is designed to be representative of the stars of interest which will be observed in the survey. Here I use simulated spectra. We decide on a set of APs and use physical models to generate the corresponding spectra. This way we can populate the AP space with spectra as desired and can ensure we have appropriate neighbours (defined in section 2.2). In the application in section 3 I use a grid of 415 spectra showing variance in 4 APs.

This filter design problem is closely related to the issue of determining the J APs of a star from measurements in I filters. This latter problem is usually solved using supervised multivariate regression methods, that is, given a set of pre-classified filter data we apply a regression method (such as a neural network) to establish the data–AP mapping (Bailer-Jones (2002)). HFD can be seen as a partial inversion of this problem in which we essentially optimize the data space itself in order to simplify its topology with respect to the APs. This should increase the performance of an ideal regression model fitted to these data and/or permit a simpler model.

HFD is being used to aid the design of the filter system for a future astronomical survey (see the my other contribution in these proceedings). More details on HFD and its application can be found in Bailer-Jones (2004).

2 The optimization model

2.1 Parametrization

A filter is parametrized with three parameters: the central wavelength, c, the half-width at half maximum (HWHM), b, and the fractional integration (or exposure) time, t, i.e. the fraction of the total integration time available per star which is allocated to this filter. (The instrument model specifies the total time available per star.) The profile of a filter – the fraction of light transmitted at each wavelength, λ – is given by the generalized Gaussian

$$\Psi(\lambda) = \Psi_0 \exp\left[-(\ln 2)\left|\frac{\lambda - c}{b}\right|^{\gamma}\right] \quad . \tag{1}$$

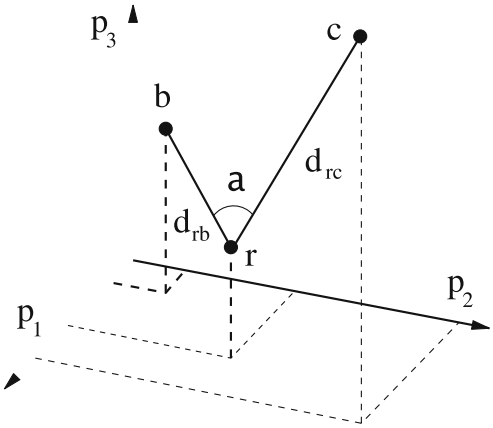


Fig. 1. A three-dimensional data (filter) space: p_i is the number of photons (i.e. brightness) collected in filter i. For star r in the grid we identify its two nearest neighbours (assuming the number of APs is two), b and c, each of which differs from r in just one of the APs. The scalar distance to these neighbours defines the AP-gradient and the angle between their vectors, $\sin\alpha$, the 'vector separation'. An optimal filter system (for r) has $\alpha = 90^\circ$ and the AP-gradients large.

This is Gaussian for $\gamma = 2$, and rectangular for $\gamma = \infty$. $\gamma = 8$ and $\Psi_0 = 0.9$ are used. For a system of I filters there are therefore $3I$ parameters which must be optimized. The optimization is performed within practical limits (set by the instrument): c and b are limited such that no part of any filter has a significant transmission (Ψ) outside of the wavelength range 2750–11250 Å. Additionally, the maximum HWHM is restricted to about 4000 Å. t must of course be $0.0 \leq t_i \leq 1.0$ and be normalized, $\sum_i t_i = 1.0$ (i labels a filter).

2.2 Figure-of-merit (fitness)

The I filters of any filter system define an I dimensional data space in which the measured objects (stars) reside (see Fig. 1), the units being photon counts observed in each filter. The location of any star is defined by its J APs. At any point in this space, each AP will therefore vary in a certain direction (the *principal direction*), and at a certain rate, the (scalar) *AP-gradient*. Using our pre-defined grid of stars, we can calculate, or at least approximate these. The ultimate purpose of the filter system is to enable us to determine these J APs. To do this, we clearly need $I \geq J$, but we must also ensure (1) that the AP-gradient is sufficiently large so that, given the signal-to-noise ratio (SNR) in the data, we can determine the AP to the desired precision, and (2) that the principal directions for each AP are mutually orthogonal, or as close to this as possible (otherwise the APs are partially degenerate). In other words, the goal of a filter system is to maximally 'separate' the different APs for the different stars in a vectorial sense.

These ideas are converted into a figure-or-merit, or fitness, as follows. For each star, r, in the grid, we find its J nearest neighbours, each of which differs from r in only one of the J APs (the grid can be constructed to ensure such neighbours exist). The relevant 'distance' between r and that neighbour differing in AP j (call it n_j), is the *AP-gradient* and is defined as

$$h_{r,n_j} = \frac{d_{r,n_j}}{|\Delta\phi_{r,n_j}|} \tag{2}$$

where d_{r,n_j} is the Euclidean distance[1] between r and n_j in and $\Delta\phi_{r,n_j}$ is their difference in AP j. Clearly, the larger h the better we have separated r and n_j. However, we must also minimise the degeneracy between the principal directions to these J neighbours, in other words, we want angle α in Fig. 1 to be as close to 90° as possible for all neighbour pairs. Combining these measures, we see that a useful figure-of-merit of separation is

$$x_{r,j,j'} = h_{r,n_j}\, h_{r,n_{j'}} \sin\alpha_{r,j,j'} \tag{3}$$

where j and j' label those neighbours which differ from r in APs j and j' respectively. Note that the above is simply the magnitude of the cross product between the two vectors. For J APs we have $J(J-1)/2$ pairs of neighbours and thus $J(J-1)/2$ terms like eqn 3. Summing these over all stars in the grid gives the final fitness which is to be maximized

$$F = \sum_{j,j'\neq j} \sum_{r} x_{r,j,j'} \quad . \tag{4}$$

(The actual fitness function is a slight modification which weights and transforms some of the terms to increase its sensitivity: see Bailer-Jones (2004)).

2.3 Evolutionary algorithm

An evolutionary algorithm (EA) is used for the optimization (e.g. Bäck and Schwefel (1993)). A population of 200 individuals is evolved over 200 generations. An outline of the algorithm is shown in Fig. 2. Natural selection is emulated using the 'roulette wheel' method, i.e. objects are selected with a probability directly proportional to their fitness (eqn. 4). Elitism is used, meaning that the E fittest individuals are always selected (and are still subject to probabilistic selection). In common with many other EA applications, this is found to improve performance. $E = 10$ is used in the results shown, although $E = 50$ actually ensures more consistent convergence (independence of initial conditions). The two search operators are recombination and mutation. Recombination involves swapping a randomly chosen filter between

[1] Distances between two points are divided by their combined error (obtained from the instrument model), so d_{r,n_j} is a dimensionless SNR.

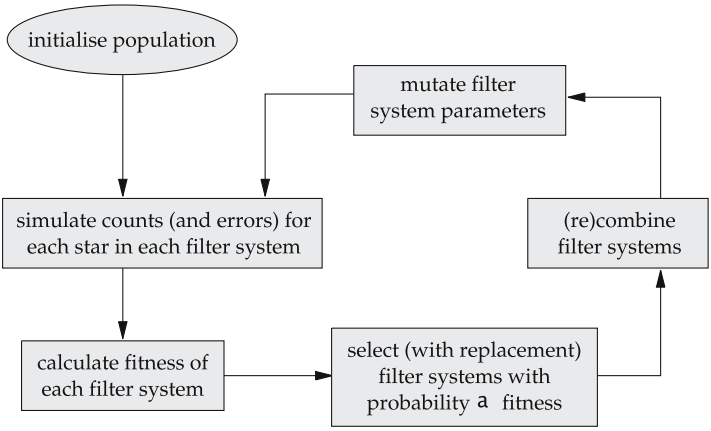


Fig. 2. Flow chart of the core aspects of the HFD optimization algorithm. A single loop represents a single iteration, i.e. the production of one new generation of filter systems.

two individuals. Mutation is implemented by adding a Gaussian random variable, $N(0, \sigma_c)$, to the central wavelength, c, and multiplying the HWHM, b, and fractional integration time, t, by $N(1, \sigma_h)$ and $N(1, \sigma_t)$ respectively. If a mutation would take a filter parameter out of bounds, then the mutation is rejected and that parameter passed on unchanged. The standard deviations σ_c, σ_h and σ_t were 500 Å, 0.5 and 0.25 respectively, and the mutation probability per parameter was 0.4. It was found that HFD was quite insensitive to the mutation probability (unless a very low probability is used, in which case there is rapid convergence to a poor local maximum) and to the standard deviations. The absence of recombination also made negligible impact.

3 Application, results and interpretation

HFD is applied to the design of a 10-filter system for determining four APs. Optimization was terminated after 200 iterations after which the fitness was found not to improve. The entire optimization was repeated 20 times from different initial (random) populations. The fittest filter system produced from this is shown in Fig. 3.

Inspection of the filter system shows that it consists of only seven filters, i.e. the optimization has 'turned off' three filters by setting their fractional integration times to zero. This is a recurrent feature. At low SNR it makes sense, because there is a penalty to be paid for retaining more filters (due to a constant noise source from the detectors). It is also interesting that the system has naturally self-regulated the widths (b) of the filters: in particular they are narrower than the maximum permitted by the limits of the optimization. This is encouraging, because on pure SNR grounds wider filters are better as they

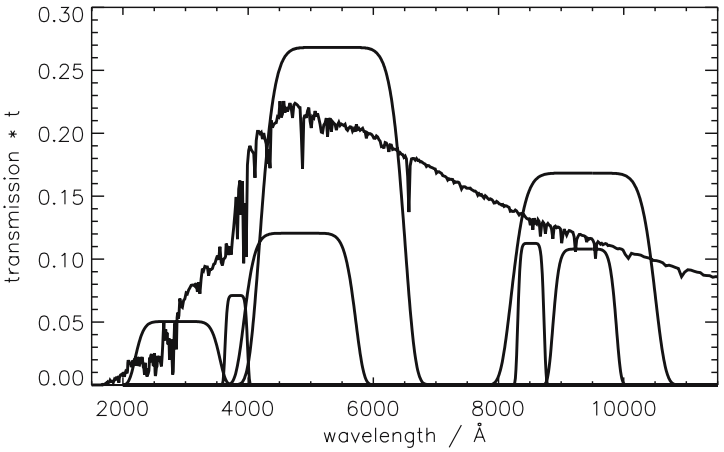


Fig. 3. An optimized filter system produced by HFD. Each of the seven filters is the plot of $\Psi(\lambda)$ from eqn. 1, but multiplied by the fractional integration time, t. Overplotted is an example of a stellar spectrum (number of photons vs. wavelength) arbitrarily scaled.

collect more photons (Poisson statistics). Yet beyond a certain width this is detrimental to the vector separation, and HFD has found this. The fact that the central wavelengths, c, cover the whole permitted wavelength range is expected from what we know about stellar spectra: a wide coverage is good for determining small changes in the slope of the spectrum. Some features are unexpected, for example the fact that the wide filter between and 8000 Å and 10 500 Å is almost equal to the sum of the two filters covering the same range. It could be that this is measuring small differences between the filters.

Astrophysically these filter systems are unconventional in two important respects. First, the filters are very broad compared to filters typically used for stellar parameter estimation. Narrow filters are able to isolate individual spectral features that we know are sensitive to specific APs. Certainly, in an ideal case, such narrow filters could better isolate specific signatures. But this implicitly assumes that we only have to deal with a narrow range of stellar types so that we could employ such specific filters. In contrast, HFD has been applied to a very broad grid of stars, as demanded by the planned surveys. Moreover, it has been applied to stellar parameters which can be demonstrated to have a broad band impact on the stellar spectrum (i.e. cause a variation which is coherent over a large wavelength range). In this case, broad band filters may be more efficient.

Second, the filters overlap in the wavelength domain. This is sometimes avoided, as it complicates the interpretation of plots of colour indices (a colour index is the ratio of the flux obtained in two filters). However, modern surveys employing many filters produce high dimensional data sets which cannot be so easily visualised, and probably contain much more information than low

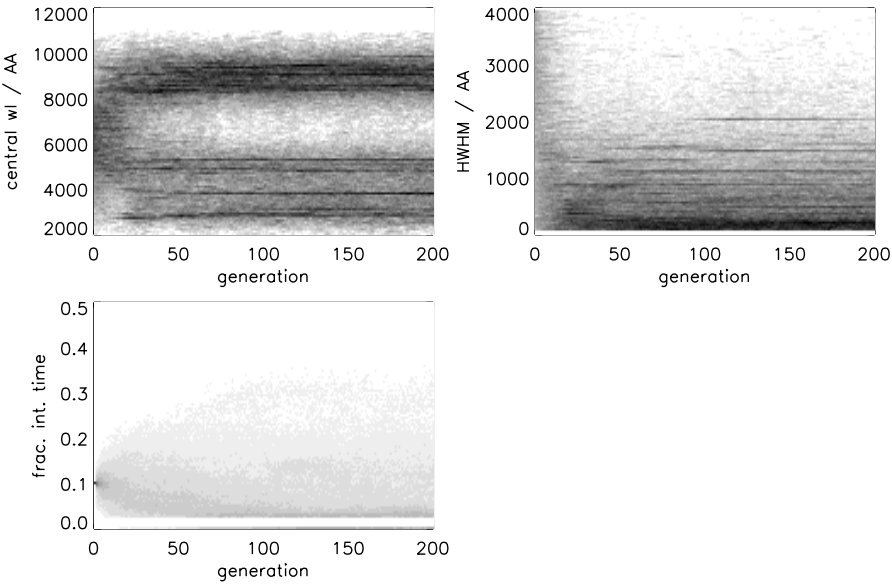


Fig. 4. Evolution of the three filter system parameters for a 10-filter system. At each generation there are ten points (one per filter) for each of the 200 filter systems for that filter parameter type, plotted as a grey scale.

dimensional slices through them. Thus the HFD results might be telling us that overlapping filters provide a more efficient sampling of stellar spectra than non-overlapping systems. This is not implausible, also given that each filter probably has a different relevance for determining each AP for each star, so the effective number of filters in each case is reduced.

When compared with alternative filter systems proposed for the same instrument, the HFD system performs much better in terms of overall fitness. This is mostly due to the broad filters and hence larger AP-gradients than conventional systems. One may suspect, therefore, that the HFD system suffers from low vector separation, because broad filters are generally less sensitive to *individual* APs. But the distribution of the vector separation terms shows this not to be the case (Bailer-Jones (2004)). Nonetheless, both the HFD and conventional systems continue to suffer from some limitations, and these will be addressed in future developments of the model.

It is interesting to follow the evolution of the filter system parameters during the optimization, as shown in Fig. 4. Looking first at the central wavelength (top left) we see that the filters occupy a fairly broad part of the parameter space for the first 20 or so iterations (generations). After about 20 iterations, some clear preferred regions appear which continue throughout the optimization and the region between 6000 and 8000 Å is disfavoured

throughout. Turning to the filter width (top right) we see that, although filters with a HWHM up to 4000 Å are permitted, after about 20 iterations the population is largely purged of filters wider than 2000 Å. A few dominant regions narrower than this stand out, but generally a range of widths are represented. The evolution of the fractional integration time (bottom panel) is quite different. They are initialized to equal values yet quickly diverge to cover the full range possible. Note the gap at low t. This is because a lower limit of 0.025 was imposed for practical reasons: filters allocated very little time will be ineffective due to low SNR. If a mutation takes t below 0.025 then t is set to zero (the thick line at the bottom). A positive mutation turns a filter on again. A maximum value of t of 0.4 is also imposed yet we see that HFD essentially self imposes a more stringent limit of about 0.3. Clearly it is inefficient if any one filter severely dominates the integration time budget.

4 Conclusions and future work

The Heuristic Filter Design model represents a systematic way for designing astronomical filters by casting this as a formal optimization problem. This makes it amenable to the extensive optimization literature. The current model is somewhat rudimentary, yet produces filter systems which are competitive with other systems designed for the same problem/instrument, at least according to the figure-of-merit developed here. The filters are somewhat unconventional – broad and overlapping – yet physically we can see why this may be preferred. Nonetheless, a number of improvements should be made to the model. First, the fitness function may be an oversimplification: it only accounts for linear variations in the data space and ignores any global degeneracies. It is also prone to 'overseparate' some stars or APs at the expense of others. Part of the problem here is that the fitness is a combination of fundamentally different terms with different scales, so the optimization is dependent on the weighting adopted (not discussed here; see Bailer-Jones (2004)). One way around this might be to use multiobjective optimization methods. In addition, more sophisticated genetic operators for search and selection could be employed, e.g. to make the search more directed, perhaps by explicitly incorporating astrophysical information.

References

BAILER-JONES, C.A.L. (2002): Automated stellar classification for large surveys: a review of methods and results. In: R. Gupta, H.P. Singh and C.A.L. Bailer-Jones (Eds.): *Automated Data Analysis in Astronomy*. Narosa Publishing House, New Delhi, 83–98.

BAILER-JONES, C.A.L. (2004): Evolutionary design of photometric systems and its application to Gaia. *Astronomy & Astrophysics*, 419, 385–403.

BÄCK, T. and SCHWEFEL, H.-P. (1993): An overview of evolutionary algorithms for parameter optimization. *Evolutionary Computation*, 1, 1–23.

Analyzing Microarray Data with the Generative Topographic Mapping Approach

Isabelle M. Grimmenstein[1], Karsten Quast[2], and Wolfgang Urfer[1]

[1] Fachbereich Statistik,
Universität Dortmund, 44221 Dortmund, Germany
[2] Boehringer Ingelheim Pharma GmbH & Co. KG,
88397 Biberach an der Riss, Germany

Abstract. The Generative Topographic Mapping (GTM) approach of Bishop et al. (1998) is proposed as an alternative to the Self-Organizing Map (SOM) approach of Kohonen (1998) for the analysis of gene expression data from microarrays. It is applied exemplarily to a microarray data set from renal tissue and the results are compared with those derived by SOM. Furthermore, enhancements for the application of the GTM methodology to microarray data are made.

1 Introduction

The technology of microarrays became very popular in the last years. It allows to monitor the expression patterns of thousands of genes from different cell types under diverse conditions simultaneously. This offers new perspectives for research in microbiology. It can help in understanding complex biochemical processes and gives a new perspective for the design of advanced therapies against diseases. However, the vast amount of data produced by microarray experiments makes efficient statistical tools necessary.

Depending on the research topic there are different approaches to analyze microarray data. If the aim is to discover inherent structures in the data (e.g. tumor classes or coregulated genes) without using prior knowledge, cluster analysis methods are appropriate. Such cluster analysis methods are often used in the exploratory phase of data analysis. Very popular for microarray data analysis is hierarchical clustering, which has however deficiencies in this context (see Tamayo et al. (1999)). Moreover, if also a visualization of the data is desired giving information about the relationships of the derived clusters to each other, other methods are more appropriate. The *S*elf-*O*rganizing *M*ap (SOM) approach by Kohonen (1982) proved to be well suited for such purposes. It was established in the field of microarray data analysis by two cornerstone papers of Tamayo et al. (1999) and Golub et al. (1999). It showed to cope well with high-dimensional and large data sets. However, to decipher the details of biochemical pathways like e.g enzyme pathways and the pathogenesis of diseases, other statistical methods are needed, which construct statistical models and also additional information to the gene expression values is needed. One approach to detect and model genetic pathways from

microarray data are e.g. Bayesian networks as applied in Grzegorczyk and Urfer (2004). Results from microarray data analysis derived by SOMs can help in generating hypotheses for such purposes.

Due to its heuristic nature though, the SOM approach exhibits also some deficiencies, which have been discussed for example in Grimmenstein et al. (2002) in the context of protein data analysis. It is not based on a statistical model and there exists no global optimization criterion. The convergence of the weight vectors is not guaranteed as well as a topographic ordering. The selection of the parameters employed has no theoretical basis. Different runs of the algorithm with different initializations and parameter settings can yield different results and there is no direct information about the reliability of the cluster assignments and topology preservation.

To overcome the main deficiencies of the SOM, the closely related *G*enerative *T*opographic *M*apping (GTM) approach by Bishop et al. (1998), being founded on a probabilistic framework, is proposed as an alternative. Its suitability for the analysis of microarray data is explored on an example set from renal tissue samples after describing the data structure and the theoretical framework. Due to restrictions of space only one data example is presented here. More data examples and more details can be found in Grimmenstein (2005). In conclusion, proposals for further enhancements of the GTM approach in the context of microarray data analysis are made.

2 Data structure

The acquisition of gene expression levels with microarrays is a complex process and includes many single steps from collection and preparation of samples, scanning the images on the arrays to determination of single expression values for each gene or expressed sequence tag (EST). All these steps have influence on the data quality. There are two common microarray technologies in use: cDNA arrays and oligonucleotide arrays by Affymetrix (cf. for more details Sebastiani et al. (2003)). The analysis methods described in the following are equally applicable to expression data derived by both technologies. The microarray data can be displayed in matrix form as in Table 1, where usually rows represent the genes and columns the different samples.

For higher-level analysis with GTM or SOM, the expression data have to be preprocessed. Common steps are normalization over different arrays to account for biases between arrays, the logarithmic transformation to even out highly skewed distributions and a normalization of the expression values for each gene to down weight genes with high variation across samples or to put the focus on the shape of the expression profiles rather than on absolute values. Additionally, a filtering step is usually performed, where genes with unreliable measurements, no expression or no significant variation across samples are filtered out from the data set to avoid that the relevant information for classification is obscured by too much noise. The selected preprocessing

Table 1. Matrix display of microarray data. Rows correspond to genes and columns to samples.

gene	**array 1**	$\cdots$	**array** b	$\cdots$	**array** B
1	y_{11}	$\cdots$	y_{1b}	$\cdots$	y_{1B}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
a	y_{a1}	$\cdots$	y_{ab}	$\cdots$	y_{aB}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
A	y_{A1}	$\cdots$	y_{Ab}	$\cdots$	y_{AB}

steps have as well influence on the classification and visualization results - especially the filtering step of genes, where the application of different methods can lead to different sets of selected genes. However, these preprocessing steps are not topic of this work and we have to assume that the evaluation and preprocessing steps yield data of sufficient quality for further analysis.

3 The GTM approach

A classification of microarray data can be performed either by genes or by samples depending on the research topic. Microarray data are classified by genes, if genes with similar expression profiles should be determined like e.g. coregulated genes in the cell cycle. A classification by samples is performed, if objects with similar expressions over diverse genes should be determined like e.g. in the case of tumor classification.

For classification and visualization purposes of gene expression data, we propose to use the Generative Topographic Mapping (GTM) approach by Bishop et al. (1998) as an alternative to the SOM approach by Kohonen (1982). The GTM approach is in spirit similar to the SOM, but has the advantage that it is founded on a probabilistic background and overcomes therefore essential deficiencies of the SOM. As a consequence of the probabilistic framework, a likelihood function forms a global optimization criterion not existing for SOMs, the convergence of the likelihood is guaranteed (at least to a local maximum), in contrast to the convergence of the weight vectors in SOMs, as well as the topographic ordering, which is also not guaranteed with SOMs. Additionally, direct information about the reliability of the cluster assignments of the data points is provided by posterior probabilities in contrast to SOMs.

With the GTM approach the higher dimensional data space $\subset \mathbb{R}^D$ is modelled by some latent variables $\mathbf{x}$ in a lower dimensional space $\subset \mathbb{R}^L, L < D$, which is analogous to the representation of the data points by the grid nodes of a map with SOMs. More precisely, GTM projects the latent space

into a non-Euclidean manifold within the data space of same dimension L as the latent space via a nonlinear mapping function $f(\cdot)$ and incorporates additionally a probability model to account for the variation of real data around the manifold. The mapping is thereby determined by a parameter matrix $\mathbf{W}$ and has to be optimized.

For visualization purposes the latent space is set to be two-dimensional, i.e. $L = 2$. For a data vector $\mathbf{y} \in \mathbb{R}^D$ it is assumed that it follows a spherical Gaussian distribution according to

$$\mathbf{y}|_{\mathbf{x},\mathbf{W},\sigma^2} \sim \mathcal{N}\left(f(\mathbf{x};\mathbf{W}), \sigma^2 \cdot \mathbf{I}\right) \tag{1}$$

being centered on $f(\mathbf{x};\mathbf{W})$ for a given latent point $\mathbf{x}$, parameter matrix $\mathbf{W}$ and variance parameter σ^2. For the distribution $p(\mathbf{x})$ over the two-dimensional latent space we assume in analogy to the SOM that it is just centered on K nodes of a regular grid, which is thus a sum of delta functions

$$p(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \delta(\mathbf{x} - \mathbf{x}_k). \tag{2}$$

Each such grid node $\mathbf{x}_k$, $k = 1, \ldots, K$, can be interpreted as standing for a cluster and being responsible for different parts of the data set. The projections of the grid nodes $f(\mathbf{x};\mathbf{W})$ form in data space the centers of different Gaussian distributions. As it is not known, according to the given classification problem, by which grid node $\mathbf{x}_k$ a data point $\mathbf{y}$ has been generated, the distribution $p(\mathbf{y}|\mathbf{W},\sigma^2)$ independent of latent points has to be considered. This is obtained by integrating over the latent variables, resulting in

$$p(\mathbf{y}|\mathbf{W},\sigma^2) = \int p(\mathbf{y}|\mathbf{x},\mathbf{W},\sigma^2) p(\mathbf{x})\, dx = \frac{1}{K} \sum_{k=1}^{K} p(\mathbf{y}|\mathbf{x}_k,\mathbf{W},\sigma^2), \tag{3}$$

a mixture of constrained Gaussian distributions. For N observations $\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{R}^D$ the joint density can be determined correspondingly by

$$p(\mathbf{y}_1, \ldots, \mathbf{y}_N|\mathbf{W},\sigma^2) = \prod_{n=1}^{N} \left\{ \frac{1}{K} \sum_{k=1}^{K} p(\mathbf{y}_n|\mathbf{x}_k,\mathbf{W},\sigma^2) \right\}, \tag{4}$$

if the simplifying assumption of independence between the observations is made. To determine an optimal mapping $f(\cdot)$ on the basis of the given data, the log likelihood of (4) is maximized with respect to the parameters $\mathbf{W}$ and σ^2. If the mapping function $f(\cdot)$ is chosen as a generalized linear regression model of the form

$$f(\mathbf{x};\mathbf{W}) = \mathbf{W}\boldsymbol{\phi}(\mathbf{x}) \tag{5}$$

with $\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))'$ as a vector of M fixed basis functions and $\mathbf{W}$ as a $D \times M$ parameter matrix, the maximum likelihood estimates for

$\mathbf{W}$ and σ^2 can be determined with the expectation maximization algorithm (Dempster et al. (1977)) in closed forms. The basis functions $\phi_j(\cdot)$, $j = 1, \ldots, M$, are thereby chosen in analogy to Bishop et al. (1998) as Gaussian functions being centered on a regular grid in latent space. However, also other functional forms are conceivable (see for more details Bishop et al. (1998)).

By applying Bayes' theorem, the posterior distribution of the latent points $\mathbf{x}_k$ is determined for each data point $\mathbf{y}_n$, $n = 1, \ldots, N$. For an unambiguous assignment of the data points to a latent point for classification and visualization purposes, the mode

$$\arg\max_k \; p(\mathbf{x}_k|\mathbf{y}_n, \mathbf{W}, \sigma^2), \tag{6}$$

is calculated. By comparing the mode with the posterior probabilities for other assignments, information about the reliability of the derived cluster assignment for a given data point is provided. On the map, several neighboring clusters can also represent together larger classes, if such a grouping is present in the data. The projected latent points $f(\mathbf{x}_k; \mathbf{W})$ in data space can be used for the determination of class profiles.

4 Application to a data set

For a comparative application of the SOM and GTM methodology we used a data set comprising gene expression values from altogether 28 samples from human kidney tissue acquired by Affymetrix GeneChips®. Thereby, 14 samples were from normal kidney and likewise 14 from renal tumors. The expression levels were determined for 10521 genes and expressed sequence tags. By the analysis of this data set with SOM and GTM it should be examined, if the inherent grouping can be retrieved on the gene expression level. The classification is therefore performed by samples. Further, subclassifications of the samples should be derived with a visualization of the relationships.

Before application of the two methods, the given expression data was preprocessed with standard methods as recommended in Section 2 by normalizing over arrays, logarithmic transformation (base 10), standardizing across samples for each gene and filtering of genes. 262 genes remained in the data set resulting in a 262-dimensional data space with 28 observations. For analysis with SOM we used the program package GeneCluster 2.0 (freely available on http://www.broad.mit.edu/cancer/software/gene
cluster2/gc2.html) and for analysis with GTM the GTM toolbox 1.02 (freely available on http://www.ncrg.aston.ac.uk/GTM/MATLAB_Impl.html).

The results with SOM showed that, apart from one sample from normal kidney, the other 27 samples were correctly classified together in one group containing the tumor samples and in another one containing the normal samples, also illustrated in Table 2. This result could be retrieved with a 1×2 SOM and also with larger map resolutions. However, the subclassifications

Table 2. Classification of 28 kidney samples with a 1×2 SOM (14 tumors, 14 normal). The second and third column contain respectively the distance of the samples to their corresponding weight vector forming cluster centers.

Cluster 1		**Cluster 2**	
Probe (#15)	Distance	Probe (#13)	Distance
tumor5	0.08723682	normal12	0.10456699
tumor2	0.088320196	normal9	0.11200833
tumor1	0.09527016	normal11	0.11891258
tumor9	0.12903029	normal13	0.12331128
tumor10	0.13737208	normal8	0.13159418
tumor12	0.14019221	normal3	0.13522816
tumor7	0.22082251	normal2	0.1356138
tumor14	0.2914098	normal10	0.16937995
tumor11	0.31995034	normal4	0.23670995
tumor8	0.32479638	normal5	0.23675716
tumor3	0.32805514	normal1	0.32501644
tumor13	0.35095918	normal14	0.3584385
tumor4	0.37606353	normal6	0.42434978
normal7	0.52952516		
tumor6	0.6459302		

of the samples from runs with different initializations and parameter settings showed some variation.

By applying the GTM method to the data set we could retrieve the main classification found by SOM, where one part of the map represents the tumor samples and the other part the normal samples. Also the wrong assignment of the same normal sample to the tumor samples could be retrieved (see Figure 1). However, the subclassifications of the samples deviated partly from the ones by SOM. As conclusion, the one normal sample which was wrongly assigned should be examined more thoroughly.

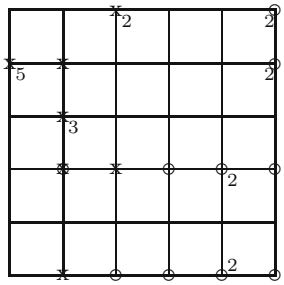


Fig. 1. Classification of 28 kidney samples by GTM on a 6×6 grid with sample numbers indicated, if more than one is assigned to a grid node. (○: normal, ×: tumor). It can be seen that the two categories are divided on different parts of the map.

5 Summary and outlook

We suggested in this work to use the GTM approach for the analysis of gene expression data for classification and visualization purposes instead of the widely used SOM approach in this context. The GTM approach is in spirit similar to the SOM, but overcomes because of its probabilistic framework some main drawbacks of the SOM, caused by its heuristic nature.

We compared the two methodologies on a data set with gene expression data from altogether 28 samples from renal tumors and normal kidney. The main inherent structure in the data set could be retrieved either with SOMs and with GTM. The subclassifications on the other hand showed some differences with the two methodologies and have to be further analyzed for their validity. The computational time needed with GTM was also considerably fast. Overall, the GTM approach with its proper probabilistic foundation and advantages seems to be a valid alternative to the SOM for the analysis of microarray data. For final conclusions about its suitability, it should be tested however on more and also larger gene expression data sets with known structure. As it is difficult to simulate microarray data with their complex inherent structure, real data should be used for this purpose.

Moreover, other distributions than Gaussians should be tried with the GTM methodology, which are potentially better suited for microarray data. Also a hierarchical approach for GTM could be used for gene expression data, if substructures should be discovered and related to each other, as a single classification step does not capture usually the whole information inherent in the data. Another aspect would be to take prior knowledge into consideration if available, which is quite often the case. The incorporation of these additional information could help in improving the classification and visualization results by GTM. In the case of classifying genes, additional information concerning the genes can be incorporated as extra components in the data vectors. These could be for example gene annotations from data bases in the internet regarding the location on the chromosome, as adjacent genes are often regulated conjointly, or the affiliation to functional groups, as proteins belonging to the same functional group often work in a similar way. When the classification of the gene expression data should be performed by samples, information concerning the samples has to be introduced in the data vectors like for example diagnostic information. The additional components introduced in the data vectors, could be either of continuous or of discrete nature. The incorporation of discrete variables would be straightforward with the probabilistic foundation of the GTM in contrast to the SOM, which is originally designed for continuous data. The incorporation of a binary variable would give for example data vectors with a conditional mixture distribution of the form

$$\mathbf{y}|_{\mathbf{x},\mathbf{W}_1,\sigma^2,\mathbf{W}_2} \sim \mathcal{N}\left(f(\mathbf{x};\mathbf{W}_1),\sigma^2\cdot\mathbf{I}\right)\times \mathsf{Ber}(\pi(\mathbf{x};\mathbf{W}_2)) \tag{7}$$

under the simplifying assumption of independence between the components, where $\mathbf{W}_1$ and $\mathbf{W}_2$ denote parameter matrices determining the mapping from latent to data space and $\pi(\mathbf{x}; \mathbf{W}_2)$ describes the probability parameter of the Bernoulli distribution. The applied probability model with the GTM approach could be further refined by introducing a weighting for the different components of the data vectors to give some components either more or less influence on the outcome. In the case of classifying genes this might be expedient, if samples with special properties should be highlighted or in the case of classifying samples genes with known importance like for example with special influence on a considered disease. A possible conditional distribution for the data vectors would be then

$$\mathbf{y}|_{\mathbf{x},\mathbf{W},\sigma^2} \sim \prod_i w_i \mathcal{N}\left(f_i(\mathbf{x}; \mathbf{W}), \sigma^2\right), \qquad \sum_i w_i = 1. \tag{8}$$

A combination of the two refinements of incorporating prior knowledge and using a weighting would be possible as well, like for example according to

$$\mathbf{y}|_{\mathbf{x},\mathbf{W}_1,\sigma^2,\mathbf{W}_2} \sim w\mathcal{N}\left(f(\mathbf{x}; \mathbf{W}_1), \sigma^2 \cdot \mathbf{I}\right) \times (1-w)\mathsf{Ber}(\pi(\mathbf{x}; \mathbf{W}_2)), \tag{9}$$

where the additional binary component is given a different weight as the gene expression values.

References

BISHOP, C.M., SVENSEN, M. and WILLIAMS, C.K.I. (1998): GTM: The Generative Topographic Mapping. *Neural Computation, 10, 215–234.*

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, B 39, 1–38.*

GOLUB, T.R., SLONIM, D.K., TAMAYO, P., HUARD, C., GAASENBEEK, M. et al. (1999): Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science, 286, 531–537.*

GRIMMENSTEIN, I.M., ANDRADE, M.A. and URFER, W. (2002): Identification of Conserved Regions in Protein Families by Self-Organizing Maps. *Technical Report 36/2002, SFB 475, Department of Statistics, University of Dortmund.*

GRIMMENSTEIN, I.M. (2005): *Development of Improved Topographic Mapping Methods in Bioinformatics.* PhD Thesis.

GRZEGORCZYK, M. and URFER, W. (2004): Determination of interacting genes in kidney tissues using Bayesian networks. *Forschungsbericht 2004/3, Department of Statistics, University of Dortmund.*

KOHONEN, T. (1982): Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics, 43, 59–69.*

SEBASTIANI, P., GUSSONI, E., KOHANE, I.S. and RAMONI, M.F. (2003): Statistical Challenges in Functional Genomics. *Statistical Science, 18, 33–70.*

TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S. et al. (1999): Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation. *Proceedings of the National Academy of Sciences, 96, 2907–2912.*

Test for a Change Point in Bernoulli Trials with Dependence

Joachim Krauth

Institute of Experimental Psychology
University of Düsseldorf, D-40225 Düsseldorf, Germany

Abstract. In Krauth (2003, 2004) we considered modified maximum likelihood estimates for the location of change points in Bernoulli sequences with first-order Markov dependence. Here, we address the more difficult problem of deriving in this situation a finite conditional conservative test for the existence of a change point. Our approach is based on the property of intercalary independence of Markov processes (Dufour and Torrès (2000)) and on the CUSUM statistic considered in Krauth (1999, 2000) in the case of independent binomial trials.

1 Introduction

In Krauth (2003) we derived modified ML estimates to identify change points and changed segments in Bernoulli trials with first-order Markov dependence. The results were extended in Krauth (2004) to the situation with multiple change points in an alternating-segments model and in a more general multiple change-points model. The results were applied to the analysis of DNA sequences.

It is a problem with our approach that it is certainly a valuable tool for detecting existing change points and changed segments but that it will yield also results in situations where no change points or changed segments, respectively, exist in reality. This problem can be addressed by using statistical tests indicating the presence of change points. Tests for one or two change points were studied in Krauth (1999, 2000) in the far simpler situation of independent binomial trials. However, the assumption of independence does not necessarily hold for DNA sequences. This was observed not only by Avery and Henderson (1999) but also in examples which were analyzed in Krauth (2003, 2004).

In the following, we consider only the case of the detection of one change point in a Bernoulli sequence with first-order Markov dependence. Based on the property of intercalary independence of Markov processes (Dufour and Torrès (2000)) we decompose the original Markov chain into subsequences of conditionally independent identically distributed Bernoulli variables. To these sequences a test described in Krauth (2000) is applied.

We are well aware that the test derived here is not the first and only test which has been proposed for detecting change points in DNA sequences. However, other published tests consider more restricted situations or are applied

without stating that they give valid results only if certain conditions are fulfilled. Corresponding tests for change points in independent Bernoulli trials were proposed not only by Krauth (1999, 2000) but also by Avery and Henderson (1999) and Halpern (1999, 2000). Other authors derived tests in the situation with dependent sequences, e.g. Johnson and Klotz (1974). However, these authors considered only the stationary case without change points.
An overview of statistical approaches used in DNA analysis has been given by Braun and Müller (1998). It seems that one of these methods nowadays is preferred above all in bioinformatics. This is the hidden Markov chain model proposed by Churchill (1989) for the segmentation of DNA sequences where the unknown parameters are estimated by means of the EM algorithm. However, as noted in Braun and Müller (1998) this approach assumes independent observations (though the unobserved states form a hidden Markov chain), requires large data sets and the EM algorithm may fail to find the global optimum. Other problems with this approach are indicated in Liu, Neuwald and Lawrence (1999), e.g. the inherent assumption that dublications and transpositions of segments of the gene are not permitted. Only under this assumption the recursive relationship comes to bear which is the key to the hidden Markov model.

2 Test problem

In Section 3 we consider the property of intercalary independence for first-order Markov chains. This property is described for a chain with an odd number (say $(2k+1)$) of trials. This is the reason why we formulate the following test problem for a sequence with $(2k+1)$ random variables. If an observed sequence consists of an even number of trials we omit the last trial. This loss of information will be of only minor importance with respect to the detection of a change point in an empirical sequence.
We consider a sequence of $(2k+1)$ random variables $X_1, ..., X_{2k+1} \in \{0,1\}$ where

$$P(X_i = 1) = 1 - P(X_i = 0) = \begin{cases} \pi_0 \text{ for } 1 \le i \le \tau \\ \pi_1 \text{ for } \tau + 1 \le i \le 2k+1 \end{cases}$$

with $k \in \{2, 3, ...\}$, $\tau \in \{2, ..., 2k-1\}, 0 < \pi_0, \pi_1 < 1$. Further, we define first-order transition probabilities
$\pi_{st,i} = P(X_i = t | X_{i-1} = s)$ for $2 \le i \le 2k+1, s, t \in \{0,1\}$.
If we assume stationarity of the transition probabilities before and after the change point (τ) we can consider the following reparametrization:
For $2 \le i \le \tau$:

$$\pi_{11}(0) := \pi_{11,i} =: \lambda_0, \qquad \pi_{10}(0) := \pi_{10,i} = 1 - \lambda_0,$$
$$\pi_{01}(0) := \pi_{01,i} = \frac{(1-\lambda_0)\pi_0}{1-\pi_0}, \quad \pi_{00}(0) := \pi_{00,i} = \frac{1 - 2\pi_0 + \lambda_0\pi_0}{1-\pi_0},$$

for $i = \tau + 1$:

$$\pi_{11}(\tau) := \pi_{11,\tau+1} =: \lambda_\tau, \qquad \pi_{10}(\tau) := \pi_{10,\tau+1} = 1 - \lambda_\tau,$$
$$\pi_{01}(\tau) := \pi_{01,\tau+1} = \frac{\pi_1 - \lambda_\tau \pi_0}{1 - \pi_0}, \ \pi_{00}(\tau) := \pi_{00,\tau+1} = \frac{1 - \pi_0 - \pi_1 + \lambda_1 \pi_0}{1 - \pi_0},$$

for $\tau + 2 \leq i \leq 2k + 1$:

$$\pi_{11}(1) := \pi_{11,i} =: \lambda_1, \qquad \pi_{10}(1) := \pi_{10,i} = 1 - \lambda_1,$$
$$\pi_{01}(1) := \pi_{01,i} = \frac{(1 - \lambda_1)\pi_1}{1 - \pi_1}, \ \pi_{00}(1) := \pi_{00,i} = \frac{1 - 2\pi_1 + \lambda_1 \pi_1}{1 - \pi_1}.$$

The following one-sided test problem is considered:
$\mathrm{H}_0 : \ \pi_1 = \pi_0, \ \lambda_0 = \lambda_\tau = \lambda_1$ vs. $\mathrm{H}_1 : \ \pi_1 < \pi_0$.
In Krauth (1999, 2000) we considered the test problem above in the special case of independence, i.e. for $\lambda_0 = \pi_0$, $\lambda_\tau = \lambda_1 = \pi_1$. We approximated the log likelihood ratio close to H_0 by means of a linearization and derived finally a cumulative sum statistic which is equivalent to similar statistics studied by Pettitt (1980), Worsley (1983) and Horváth (1989). Because we intend to use this test in the following we give here a short description of the procedure based on the presentation in Krauth (2000):
Let $W_1, ..., W_n \in \{0, 1\}$ with $n \in \{3, 4, ...\}$ independent random variables with

$$P(W_i = 1) = 1 - P(W_i = 0) = \begin{cases} \pi_0 \text{ for } 1 \leq i \leq \tau \\ \pi_1 \text{ for } \tau + 1 \leq i \leq n \end{cases}.$$

For the test problem $\mathrm{H}_0 : \ \pi_1 = \pi_0$ vs. $\mathrm{H}_1 : \ \pi_1 < \pi_0$ we consider the test statistic

$$T = \max_{1 \leq j \leq n-1} \{M_j - \frac{M_n}{n} j\}$$

where $M_j = \sum_{i=1}^{j} W_i, 1 \leq j \leq n$. We define

$$A_{j,u} = \{M_j - \frac{m_n}{n} j \geq u\} = \{M_j \geq u + \frac{m_n}{n} j\} \text{ for } 1 \leq j \leq n - 1$$

and derive under H_0 conditional on the observed value $M_n = m_n$ the conditional probability

$$P_{\mathrm{H}_0}(T \geq u | M_n = m_n) = P(\bigcup_{j=1}^{n-1} A_{j,u} | M_n = m_n).$$

In order to get an upper bound for this probability of a union we define

$$y_i := \max\{0, \lceil T + \frac{m_n}{n} i \rceil\} \text{ for } 1 \leq i \leq n - 1,$$

$$p_i := P_{\mathrm{H}_0}(A_{i,u} | M_n = m_n) = \binom{n}{m_n}^{-1} \sum_{s=y_i}^{\min\{m_n, i\}} \binom{i}{s} \binom{n-i}{m_n - s} \text{ for } 1 \leq i \leq n - 1,$$

$$p_{ij} := P_{\mathrm{H}_0}(A_{i,u} \cap A_{j,u} | M_n = m_n)$$
$$= \binom{n}{m_n}^{-1} \sum_{s=y_i}^{\min\{m_n, i\}} \sum_{t=\max\{s, y_j\}}^{\min\{m_n, j-i+s\}} \binom{i}{s} \binom{j-i}{t-s} \binom{n-j}{m_n - t} \text{ for } 1 \leq i < j \leq n - 1,$$

$p_{ij} := p_{ji}$ for $1 \le j < i \le n-1$.
A good upper bound for the probability of a union which was derived by Kounias (1968) yields

$$P_{H_0}(T \ge u | M_n = m_n) \le U := \min\{1, \sum_{i=1}^{n-1} p_i - \max_{1 \le j \le n-1} \sum_{\substack{i=1 \\ i \ne j}}^{n-1} p_{ij}\}.$$

We can reject H_0 for $U \le \alpha$. This results in a one-sided exact conditional conservative test for the existence of a change point in the situation with (independent) Bernoulli trials.

3 Intercalary independence of Markov processes

In Section 2 we considered a one-sided test problem, where we assumed under the null hypothesis (H_0) a homogeneous first-order Markov chain with $(2k+1)$ random variables $X_1, ..., X_{2k+1} \in \{0,1\}$, where $k \in \{2, 3, ...\}$, $P(X_i = 1) = 1 - P(X_i = 0) = \pi_0$ for $1 \le i \le 2k+1$, $0 < \pi_0 < 1$, $\pi_{st,i} = P(X_i = t | X_{i-1} = s)$ for $2 \le i \le 2k+1$ and $s, t \in \{0,1\}$, $\pi_{11} := \pi_{11}(0) = \pi_{11,i} = \lambda_0$, $\pi_{10} := \pi_{10}(0) = \pi_{10,i} = 1 - \lambda_0$, $\pi_{01} := \pi_{01}(0) = \pi_{01,i} = \dfrac{(1-\lambda_0)\pi_0}{1-\pi_0}$, $\pi_{00} := \pi_{00}(0) = \pi_{00,i} = \dfrac{1 - 2\pi_0 + \lambda_0 \pi_0}{1 - \pi_0}$ for $2 \le i \le 2k+1$.
For this chain the following properties hold for any choice of $(x_1, ..., x_{2k+1}) \in \{0,1\}^{2k+1}$:

$$\begin{aligned}
(i)\quad & P(\bigcap_{i=1}^{k} \{X_{2i} = x_{2i}\} | \bigcap_{j=1}^{k+1} \{X_{2j-1} = x_{2j-1}\}) \\
& = \prod_{i=1}^{k} P(X_{2i} = x_{2i} | \bigcap_{j=1}^{k+1} \{X_{2j-1} = x_{2j-1}\}) \\
(ii)\quad & P(X_{2i} = x_{2i} | \bigcap_{j=1}^{k+1} \{X_{2j-1} = x_{2j-1}\}) \\
& = P(X_{2i} = x_{2i} | X_{2i-1} = x_{2i-1}, X_{2i+1} = x_{2i+1}) \text{ for } 1 \le i \le k \\
(iii)\quad & P(X_{2i} = x_{2i} | X_{2i-1} = x_{2i-1}, X_{2i+1} = x_{2i+1}) \\
& = \frac{\pi_{x_{2i-1} x_{2i}} \pi_{x_{2i} x_{2i+1}}}{\pi_{x_{2i-1} 1} \pi_{1 x_{2i+1}} + \pi_{x_{2i-1} 0} \pi_{0 x_{2i+1}}} \text{ for } 1 \le i \le k \\
(iv)\quad & P(X_{2i} = 1 | X_{2i-1} = 1, X_{2i+1} = 1) \\
& = 1 - P(X_{2i} = 0 | X_{2i-1} = 1, X_{2i+1} = 1) = \frac{\pi_{11}^2}{\pi_{11}^2 + \pi_{10}\pi_{01}}, \\
& P(X_{2i} = 1 | X_{2i-1} = 0, X_{2i+1} = 0) \\
& = 1 - P(X_{2i} = 0 | X_{2i-1} = 0, X_{2i+1} = 0) = \frac{\pi_{01}\pi_{10}}{\pi_{01}\pi_{10} + \pi_{00}^2}, \\
& P(X_{2i} = 1 | X_{2i-1} = 1, X_{2i+1} = 0) = P(X_{2i} = 1 | X_{2i-1} = 0, X_{2i+1} = 1) \\
& = 1 - P(X_{2i} = 0 | X_{2i-1} = 1, X_{2i+1} = 0) \\
& = 1 - P(X_{2i} = 0 | X_{2i-1} = 0, X_{2i+1} = 1) = \frac{\pi_{11}}{\pi_{11} + \pi_{00}} \text{ for } 1 \le i \le k
\end{aligned}$$

The properties (i) and (ii) can be concluded from results in literature as special cases. The result (i) has been called "intercalary independence" by Dufour and Torrès (2000), and the result (ii) "truncation property". These authors prove these properties for general Markov processes of order p and generalize in this way more specific results by former authors. It is claimed that the proofs given are the first proofs for intercalary independence and the truncation property ever published. In Ogawara (1951) we find the remark that U.V. Linnik used intercalary independence (without proof) in 1949 in his proof of a central limit theorem (exact reference in Dufour and Torrès (2000)). Our result (i) corresponding to intercalary independence for first-order Markov chains is contained in Theorem 1 of Ogawara (1951) and Theorem 3.1 in Dufour and Torrès (2000). Our result (ii) corresponding to the truncation property for first-order Markov chains is contained in Theorem 3.2 of Dufour and Torrès (2000).
By expressing the right side of (ii) by the transition probabilities π_{11}, π_{10}, π_{01}, and π_{00} we derive (iii). In (iv) the eight possible values of (iii) are explicitly given.

4 Strategies for performing a test

If we want to analyse a sequence with $(2k+1)$ binary trials we are in the situation described above. In case of a sequence with $(2k+2)$ trials we omit the last one. Conditional on the $(k+1)$ values of X_{2j-1}, $1 \le j \le k+1$, the (k) variables X_{2i}, $1 \le i \le k$, are independent though not necessarily identically distributed according to the results (i) - (iv) above. This sequence of (k) independent variables can be split up into 3 subsequences of independent identically distributed random variables of lengths $k_{11}, k_{00}, k_{10,01}$ (k_{11}, k_{00}, $k_{10,01} \in \{0,1,2,...,k\}, k_{11}+k_{00}+k_{10,01}=k$):
$X_{2i}^{(11)}$ with $x_{2i-1}^{(11)} = x_{2i+1}^{(11)} = 1$ for $1 \le i \le k_{11}$,
$X_{2i}^{(00)}$ with $x_{2i-1}^{(00)} = x_{2i+1}^{(00)} = 0$ for $1 \le i \le k_{00}$,
$X_{2i}^{(10,01)}$ with $x_{2i-1}^{(10,01)} = 1$, $x_{2i+1}^{(10,01)} = 0$ or $x_{2i-1}^{(10,01)} = 0, x_{2i+1}^{(10,01)} = 1$
for $1 \le i \le k_{10,01}$.
In Section 2 we considered a one-sided test problem where even under the null hypothesis of no change point a first order Markov chain is assumed. The parameter describing the dependence of this chain can be considered as a nuisance parameter. We might estimate this parameter and derive an asymptotic test where in a real data situation we do not know whether the test is conservative or not. We preferred to derive an exact conditional conservative test applied to a sequence of conditionally independent variables. Unfortunately, the variables of this sequence are not identically distributed but form the three subsequences above with possibly different unknown parameters.
We can apply the one-sided finite conservative test with respect to the alternative hypothesis $H_1 : \pi_1 < \pi_0$ described in Section 2 to any of the three

subsequences. If we are interested in the one-sided hypothesis $H_1' : \pi_1 > \pi_0$ we consider the inverted sequence $X_{2(k-i+1)}$, $1 \leq i \leq k$, and proceed as above. However, if we are interested in the two-sided hypothesis $H_2' : \pi_1 \neq \pi_0$ we perform the test as well for a subsequence as for its inversion and reject H_0 if $U \leq \frac{\alpha}{2}$ for one of the two resulting bounds defined in Section 2.

In most empirical situations, e.g. in DNA analysis, we do not have a directed hypothesis and do not know which of the three subsequences with its inversion should be considered. A first idea for solving this problem might be to perform one-sided tests for all 6 subsequences and use one of the known procedures for multiple testing for 6 dependent tests, e.g. the Bonferroni procedure. However, due to the different lengths of the subsequences and the influence of the unknown parameters of the model the efficiencies of the 6 tests can differ considerably and due to the conservativeness of the tests upper bounds of 1 may occur frequently. Therefore, we propose to derive in a first stage modified maximum likelihood estimates of π_1 and π_0 as described in Krauth (2003) on the basis of the fixed sequence x_{2j-1}, $1 \leq j \leq k+1$. From this sequence we derive also the lengths of the 3 subsequences. Then we formulate a one-sided test problem based on the knowledge of these estimates and perform the corresponding test for the longest subsequence. Here, it is hoped that the estimates of π_0 and π_1 based on the sequence with $(k+1)$ trials are near to the corresponding true parameters in the original sequence with $(2k+1)$ trials and that the test has the highest efficiency for the longest subsequence.

An alternative approach considering the upper bounds for the 3 independent subsequences is based on the Tippett multiple test procedure (Tippett (1931)), where the smallest upper bound is compared with $1-(1-\alpha)^{1/3}$ instead of with α. In the case of a two-sided test problem the procedure is performed for the original 3 subsequences and also for the 3 inverted subsequences, both times with $(\frac{\alpha}{2})$ instead of α. In Sections 2-4 we assumed trials with first-order Markov dependence. This assumption can be weakened, at least in theory, by considering sequences with second- or even higher-order dependence. The main problem is that not only a far greater number of subsequences has to be considered but that in addition these subsequences are much shorter than in the case of first-order dependence. In view of the conservativeness of our test procedure we will have thus only a small chance to detect a change point.

5 Example

Just as in Krauth (2003, 2004) we consider the nucleotide sequence reported by Robb et al. (1998, Fig. 1). This is 1,200 *nt* in length, is constructed from overlapping clones and is based on the analysis of up to 181 mice embryos. Just as in Krauth (2004) we coded the letter A (corresponding to the purine adenine) by 1 and the other three letters (G = guanine, T = thymine, C =

cytosine) by 0 and generated in this way a binary sequence with 1,200 trials. After omitting the last trial we can choose $k = 599$. Fixing the $k + 1 = 600$ odd trials we get $k_{11} = 44$, $k_{00} = 332$, and $k_{10,01} = 223$ for the 599 even trials. The modified ML estimates of π_0 and π_1 based on the 600 odd trials are given by $\hat{\pi}_0 = .2354$, $\hat{\pi}_1 = .3762$. Because $\hat{\pi}_1 > \hat{\pi}_0$ holds, we have to consider an inverted (i) sequence. For the longest subsequence with $k_{00} = 332$ trials we get the upper bound $U^i_{00} = .0495$. This bound is smaller than $\alpha = .05$ and therefore we assume that a change point exists.

The other two bounds for the inverted sequence are given by $U^i_{11} = .5029$ and $U^i_{10,01} = .5413$. Tippett's procedure yields $1 - (1 - .05)^{1/3} = .0170$ and this is smaller than $U^i_{00} = .0495$, i.e. no significant result is obtained. The three bounds for the noninverted sequence are given by $U_{11} = .7677$, $U_{00} = 1$, and $U_{10,01} = 1$. Thus, neither the two-sided test for the (00)-sequence nor the comparison with the Bonferroni bound $.05/6 = .0083$ for all 6 (dependent) tests yield a significant result.

It should be emphasized that our procedure is based on the assumption that first we have at least the same ordering of $\hat{\pi}_0$ and $\hat{\pi}_1$ on the one side and of π_0 and π_1 on the other side, and second that the efficiency of the test increases with the length of the sequence of independent trials under consideration. It is not guaranteed that these assumptions are valid for empirical data. As a consequence it might be worthwhile to consider in addition all 6 possible upper bounds in case of a nonsignificant result.

In Section 1 we mentioned various former approaches for detecting change points in DNA sequences. However, if we cannot rule out that trials are dependent we cannot decide whether results for corresponding tests differ from our results because our test is rather conservative or whether those tests are not robust with respect to the presence of dependence.

References

AVERY, P.J. and HENDERSON, D.A. (1999): Fitting Markov Chain Models to Discrete State Series such as DNA Sequences. *Applied Statistics, 48, 53–61.*

BRAUN, J.V. and MÜLLER, H.G. (1998): Statistical Methods for DNA Sequence Segmentation. *Statistical Science, 13, 142–162.*

CHURCHILL, G.A. (1989): Stochastic Models for Heterogeneous DNA Sequences. *Bulletin of Mathematical Biology, 51, 79–94.*

DUFOUR, J.M. and TORRÈS, O. (2000): Markovian Processes, Two-Sided Autoregressions and Finite-Sample Inference for Stationary and Nonstationary Autoregressive Processes. *Journal of Econometrics, 98, 255–289.*

HALPERN, A.L. (1999): Minimally Selected p and Other Tests for a Single Abrupt Change Point in a Binary Sequence. *Biometrics, 55, 1044–1050.*

HALPERN, A.L. (2000): Multiple-Changepoint Testing for an Alternating Segments Model of a Binary Sequence. *Biometrics, 56, 903–908.*

HORVÁTH, J. (1989): The Limit Distributions of the Likelihood Ratio and Cumulative Sum Tests for a Change in a Binomial Probability. *Journal of Multivariate Analysis, 31, 148–159.*

JOHNSON, C.A. and KLOTZ, J.H. (1974): The Atom Probe and Markov Chain Statistics of Clustering. *Technometrics, 16, 483–493.*

KOUNIAS, E.G. (1968): Bounds for the Probability of a Union, with Applications. *Annals of Mathematical Statistics, 39, 2154–2158.*

KRAUTH, J. (1999): Discrete Scan Statistics for Detecting Change-Points in Binomial Sequences. In: W. Gaul and H. Locarek-Junge (Eds.): *Classification in the Information Age.* Springer, Heidelberg, 196–204.

KRAUTH, J. (2000): Detecting Change-Points in Aircraft Noise Effects. In: R. Decker and W. Gaul (Eds.): *Classification and Information Processing at the Turn of the Millenium.* Springer, Heidelberg, 386–395.

KRAUTH, J. (2003): Change-Points in Bernoulli Trials with Dependence. In: M. Schader, W. Gaul and M. Vichi (Eds.): *Between Data Science and Applied Data Analysis.* Springer, Heidelberg, 261–269.

KRAUTH, J. (2004): Multiple Change-Points and Alternating Segments in Binary Trials with Dependence. In: D. Baier and K.D. Wernecke (Eds.): *Innovations in Classification, Data Science, and Information Systems.* Springer, Heidelberg, 154–164.

LIU, J.S., NEUWALD, A.F. and LAWRENCE, C.E. (1999): Markovian Structures in Biological Sequence Alignments. *Journal of the American Statistical Association, 94, 1–15.*

OGAWARA, M. (1951): A Note on the Test of Serial Correlation Coefficients. *Annals of Mathematical Statistics, 22, 115–118.*

PETTITT, A.N. (1980): A Simple Cumulative Sum Type Statistic for the Change-Point Problem with Zero-One Observations. *Biometrika, 67, 79–84.*

ROBB, L., MIFSUD, L., HARTLEY, L., BIBEN, C., COPELAND, N.G., GILBERT, D.J., JENKINS, N.A. and HARVEY, R.P. (1998): Epicardin: A Novel Basic Helix-Loop-Helix Transcription Factor Gene Expressed in Epicardium, Branchial Arch Myoblasts, and Mesenchyme of Developing Lung, Gut, Kidney, and Gonads. *Developmental Dynamics, 213, 105–113.*

TIPPETT, L.H. (1931): *The Methods of Statistics.* Williams and Norgate, London.

WORSLEY, K.J. (1983): The Power of Likelihood Ratio and Cumulative Sum Tests for a Change in a Binomial Probability. *Biometrika, 70, 455–464.*

Data Mining in Protein Binding Cavities

Katrin Kupas and Alfred Ultsch

Data Bionics Research Group,
University of Marburg, D-35032 Marburg, Germany

Abstract. The molecular function of a protein is coupled to the binding of a substrate or an endogenous ligand to a well defined binding cavity. To detect functional relationships among proteins, their binding-site exposed physicochemical characteristics were described by assigning generic pseudocenters to the functional groups of the amino acids flanking a particular active site. These pseudocenters were assembled into small substructures and their spatial similarity with appropriate chemical properties was examined. If two substructures of two binding cavities are found to be similar, they form the basis for an expanded comparison of the complete cavities. Preliminary tests indicate the benefit of this method and motivate further studies.

1 Introduction

In a biological system multiple biochemical pathways are proceeded and regulated via the complementary recognition properties of proteins and their substrates. The ligand accommodates the binding cavity of the protein according to the lock-and-key principle. Two fold requirements are given: on the one hand, the ligand needs to fit sterically into the binding cavity of the protein. On the other hand, the spatial arrangement of ligand and receptor must correspond to a complementary physicochemical pattern.

The shape and function of a protein, e.g. of an enzyme together with its active site is not exclusively represented by a unique amino acid sequence. Accordingly, proteins with deviating amino acid sequence, even adopting a different folding pattern, can nevertheless exhibit related binding cavities to accommodate a ligand. Low sequence homology does not imply any conclusions on binding site differences or similarities. For this reason one has to regard the three-dimensional structure as a prerequisite for a reliable comparison of proteins. Such structures are available for many examples from X-ray crystallography. In literature, different methods based on the description of the spatial protein structures in terms of a reduced set of appropriate descriptors have been reported. In addition to the shape, it is required to code correctly the exposed physicochemical properties in a geometrical and also chemical sense.

In this paper, we describe a new algorithm to compare protein binding sites by the use of common local regions. These local regions form the basis for the further comparison of two binding cavities. Similar local regions

among sets of spatially arranged descriptors of two binding cavities provide a coordinate system which will be used in the next step to perform other substructure searches for related cavities. Once a convincing match is detected, it can be assumed that the two active centers are capable to bind similar ligands and thus exhibit related function.

The paper is organized as follows: In Chapter 2, other approaches to classify binding cavities are reviewed. Chapter 3 describes the underlying theory and concept of our algorithm used for cavity matching. The local region in descriptor space is defined. In Chapter 4, some preliminary results of a binding cavity matching are presented. Conclusions are given in Chapter 5.

2 Other approaches

Previously reported approaches to classify binding cavities can be assigned to three categories, according to the information they use for the classification:

1. Sequence alignments.
2. Comparison of folding patterns and secondary structure elements.
3. Comparison of 3D substructural epitopes.

Sequence alignments

If two proteins show high sequence identity one can assume structural and most likely also functional similarity among them. The mostly applied procedures were presented by Needleman and Wunsch (1970) and by Waterman (1984). Nevertheless, they are computationally and memory-wise quite demanding, so that often heuristic methods are used such as FASTA(Pearson and Lipman (1988), Pearson (1990)) and BLAST (Altschul et al. (1990)). These procedures do not find an optimal solution in all cases, but generally reveal good approximative results.

Comparison of folding patterns and secondary structure elements

In general, sequence alignment methods are only capable to detect relationships among proteins if sequence identity exceeds beyond 35%. To classify more distant proteins, information about their three-dimensional structure has to be incorporated to the comparison. Many methods, that establish classification and assignment of proteins to structural families, exploit global fold similarities. Hierarchical procedures have been developed which classify proteins according to their folding, their evolutionary ancestors, or according to their functional role. These systems operate either automatically or are dependent on manual intervention. Many of the classification schemes treat proteins as being composed by domains and classify them in terms of the properties of their individual domains (Ponting and Russell (2002)). Important approaches for such classifications are: SCOP(Murzin et al. (1995), Lo Conte et al. (2002)), CATH (Orengo et al. (1997, 2000)), FSSP/DALI (Holm

and Sander (1996), Holm (1998)), MMDB (Gibrat et al. (1996)). The ENZYME (Bairoch (2000)) and BRENDA database (Schomburg et al. (2002)) annotate proteins with respect to the catalyzed reaction.

Comparison of 3D substructural epitopes

Beyond these relationships, proteins can possess a similar function even if they do not have any sequence and/or folding homology in common. Accordingly, methods that compare proteins only with respect to their folding pattern cannot detect such similarities. Procedures which seek for similar substructures in proteins are better adapted to discover similarities in such cases.

The first group of algorithms comprises methods that scan protein structural databases in terms of pre-calculated or automatically generated templates. A typical example of such a template is the catalytic triad in serine proteases. A substantial advantage to restrict to relatively small templates is due to the fact that even large data collections can be scanned efficiently. Some of the best known procedures based on templates are ASSAM introduced by Artymiuk et al. (1993,2003), TESS/PROCAT by Wallace et al. (1996, 1997), PINTS by Stark and Russell (2003), DRESPAT by Wangikar et al. (2003) as well as the methods of Hamelryck (2003) and Kleywegt (1999).

The second group includes approaches to compare substructural epitopes of proteins which operate independent of any template definition. For the similarity search the whole proteins or substructures are used. The group of Ruth Nussinov and Haim Wolfson developed many approaches to compare entire receptor structures or substructures. The individual methods essentially differ whether the protein structure is represented by their C_α-atoms or grid points on their solvent-accessible surface, or by so-called "sparse critical points", a compressed description of the solvent-accessible surface. In each case, the different procedures use geometric hashing (Bachar et al. (1993)) for common substructure detection. They perform completely independent of sequence or fold homology. The approach of Rosen (1998) permits an automatic comparison of binding cavities. Kinoshita et al. (2003) use a graph-based algorithm to compare the surfaces of two proteins. Other methods, such as GENE FIT of Lehtonen et al. (1999) and the approach of Poirrette et al. (1997) use genetic algorithms to optimally superimpose proteins in identified substructure ranges.

3 Theory and algorithm

The algorithm builds on the approach of Schmitt et al. (2002). The physicochemical properties of the cavity-flanking residues are condensed into a restricted set of generic pseudocenters corresponding to five properties essential for molecular recognition: hydrogen-bond donor (DO), hydrogen-bond acceptor (AC), mixed donor/acceptor (DA), hydrophobic aliphatic (AL) and aro-

matic (PI). The pseudocenters express the features of the 20 different amino acids in terms of five well-placed physicochemical properties.

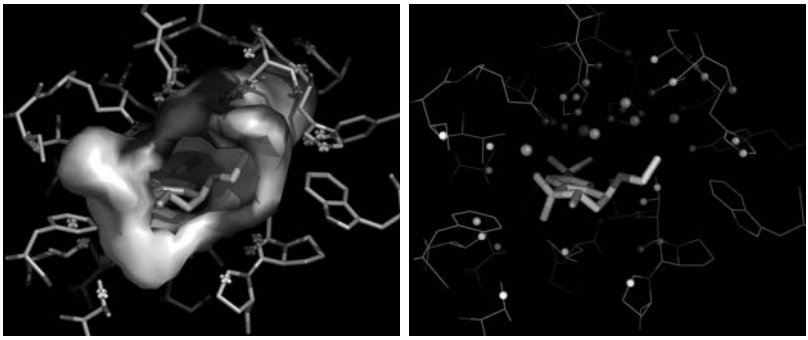


Fig. 1. Surface (left) and pseudocenters (right) of a binding cavity with bound ligand

The idea for this algorithm resides on the concept that common substructures of two binding cavities have an arrangement of these pseudocenters in common. Therefore local regions are regarded. They are composed by a center under consideration and the three nearest neighboring centers forming a pyramid. The pyramid was chosen as similarity measure because it corresponds to the smallest spatial unit spanned by these four centers.

Systematically every pseudocenter in a cavity is selected as the current center and forms a local region with its three nearest neighbors. Following the procedure, the binding cavities are partitioned by all local regions to be possibly inscribed. Accordingly, the mutual comparison of binding cavities is reduced to a multiple comparison of the different local regions.

Therefore, the spatial and physicochemical characteristics of the local regions are considered separately. The spatial features of the local regions were chosen under the aspects of using a minimum number of descriptors to identify local regions and producing minimal measuring errors. A set of six spatial descriptors was tested. Three of them, the height of the pyramid, the area of the triangle spanned by the three neighbor centers and the distance between the root point of the height and the barycenter of this triangle, have been chosen.

Every pair of local regions of two cavities with the same physicochemical and spatial properties forms the basis for the comparison of the two cavities. These two local regions are matched and the score of the appropriate overlay of the cavities is calculated.

1. Two pyramids with appropriate chemical and spatial characteristics of different binding cavities give rise to a coordinate transformation, which optimally superimposes both pyramids. That means if pyramid A consists

of the points ($A1, A2, A3, A4$) and pyramid B of the points ($B1, B2, B3, B4$), then a rotation/translation has to be found, so that the sum of the squares of the distances from A_i to B_i adopts a minimal value (Prokrustes analysis).
2. Subsequently this coordinate transformation is applied to all cavities. Then, every pair of pseudocenters of the two cavities, which mutually match chemically and fall close to each other beyond a threshold of 1 Å is counted. This number of successful matches is the score for this pair of pyramids.
3. The superpositioning is applied to all pairs of pyramids with the same physicochemical and sterical properties of these two cavities. The maximum of the resulting scores is determined.
4. Relating this maximum score to the "maximally achievable score", i.e. the number of pseudocenters in the smaller of the two cavities, gives an estimate of the maximally achievable score.

This algorithm has a set of advantages contrary to a consideration only of the individual pseudocenters. With a Prokrustes analysis concerning the individual pseudocenters the coordinate transformation must be accomplished for all pairs of chemically identical pseudocenters and the best match has to be calculated. By consideration of local regions four suitable pseudocenters are given, which have to be matched. Thus the number of computations for the superpositioning of the two cavities is reduced. Only those overlaps with a match of all four pseudocenters have to be computed. A local region composed of four centers gives a good initialization for the Prokrustes analysis. Fewer degrees of freedom exist for the coordinate transformation. A further advantage is that not all binding cavities of a data base have to be considered. Only those cavities containing a suitable local region come into consideration for the surface overlay of the cavities. The remaining cavities without a suitable local region are not consulted.

4 First results

The approach based on local regions for the comparison of protein active sites has been tested with four pairs of binding cavities with well known common substructures (Siemon (2001)). Other similarities between the proteins than these pairs were not expected. The proteins from where the binding cavities had been extracted are the following:
an Adenylate Kinase (1ake.2), an allosteric Chorismate Mutase (1csm.3), the Chorismate Mutase of E. Coli (1ecm.5), a Bovine-Actin-Profilin Complex (1hlu.1), a heat shock cognate Protein (1kay.1), Trypsin (1tpo.1), the Uridylate Kinase (1ukz.1) and Proteinase K (2prk.2) (Protein Data Base code (PDB)).
The pairs of proteins with well-known common substructures in their binding

cavities are 1ake.2/1ukz.1 (Kinases), 1csm.3/1ecm.5 (Isomerases), 1hlu.1/1kay.1 (Hydrolases) and 1tpo.1/2prk.1 (Serine Proteinases). The resulting scores after mutual match are shown in Figure 1.

Table 1. Resulting scores of a mutual comparison of four pairs of binding cavities with well-known common substructures

	1ake.2	1csm.3	1ecm.5	1hlu.1	1kay.1	1tpo.1	1ukz.1	2prk.1
1ake.2	—	21.1	20.7	11.5	13.5	19.4	**69.1**	26.2
1csm.3	21.1	—	**41.4**	14.0	21.1	12.3	17.5	9.5
1ecm.5	20.7	**41.4**	—	13.8	17.2	0.0	20.7	17.2
1hlu.1	11.5	14.0	13.8	—	**29.7**	14.9	12.7	11.9
1kay.1	13.5	21.0	17.2	**29.7**	—	17.9	13.6	21.4
1tpo.1	19.4	12.3	0.0	14.9	17.9	—	14.9	**28.6**
1ukz.1	**69.1**	17.5	20.7	12.7	13.6	14.9	—	19.1
2prk.1	26.2	9.5	17.2	11.9	21.4	**28.6**	19.1	—

The numbers are given as percentage with respect to the maximally achievable score (see section 3.3).

The table shows that those cavities which are known to possess common substructures also achieve the best scores, whereas the best fit found for the other cavities reveals in most of the cases significantly smaller values. The results have been examined by an expert. The coordinate transformations and the matching pseudocenters of the known pairs of binding cavities were identical with the estimated analogy.

5 Conclusions

We presented a new algorithm to find common substructures and compare protein binding cavities. The cavities are partitioned into small local regions with spatial and physicochemical properties. They are formed by pseudocenters assigned to five different physicochemical qualities. The local region exists of a center under consideration and its three nearest neighbors. The physicochemical characteristics of the local regions are the combination of the physicochemical attributes assigned to each pseudocenter. The spatial characteristics were described by the height of the pyramid, the area of the triangle spanned by the three neighbor centers and the distance between the root point of the height and the barycenter of this triangle.

The advantage of this algorithm is, that only those cavities are observed, that share a common local region, expressed in terms of the pyramid. Such substructures, which are represented by pseudocenters widely distributed over the cavity and so are not biologically relevant, are a priori excluded from the consideration. An advantage of the use of an ESOM for classifying protein

binding cavities is the fast comparison of one individual cavity with the entire database. All candidates of proteins of the whole database sharing in common similar local regions are identified in one step.

The comparison of four pairs of binding cavities with well-known common substructures led to promising results. It can be assumed that the approach of dividing protein binding cavities into local regions and comparing them is capable to detect similar substructures in the cavities.

References

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. and LIPMAN, D.J. (1990): Basic local alignment search tool. *J. Mol. Biol., 215, 403–410.*

ARTYMIUK, P.J., GRINDLEY, H.M., RICE, D.W. and WILLETT, P. (1993): Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J. Mol. Biol., 229, 707–721.*

ARTYMIUK, P.J., SPRIGGS, R.V. and WILLETT, P. (2003): Searching for patterns of amino acids in 3D protein structures. *J. Chem. Inf. Comput. Sci., 43, 412–421.*

BACHAR, O., FISCHER, D., NUSSINOV, R. and WOLFSON, H. (1993): A computer vision based technique for 3-D sequence-independent structural comparison of proteins. *Protein Eng., 6, 279–288.*

BAIROCH, A. (2000): The ENZYME database in 2000. *Nucleic Acids Res., 28, 304–305.*

GIBRAT, J.F., MADEJ, T. and BRYANT, S.H. (1996): Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol., 6, 377–385.*

HAMELRYCK, T. (2003): Efficient identification of side-chain patterns using a multidimensional index tree. *Proteins, 51, 96–108.*

HOLM, L. and SANDER, C. (1996): The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res., 24, 206–209.*

HOLM, S. (1998): Touring protein fold space with Dali/FSSP.

KINOSHITA, K. and NAKAMURA, H. (2003): Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci., 12, 1589–1595.*

KLEYWEYGT, G.J. (1999): Recognition of spatial motifs in protein structures. *J. Mol. Biol., 285, 1887–97.*

LEHTONEN, J. V., DENESSIOUK, K., MAY, A. C. and JOHNSON, M.S. (1999): Finding local structural similarities among families of unrelated protein structures: a generic nonlinear alignment algorithm. *Proteins, 34, 341–355.*

LO CONTE, L., BRENNER, S.E., HUBBARD, T.J., CHOTHIA, C. and MURZIN, A. G. (2002): SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res., 30, 264–267.*

MURZIN, A.G., BRENNER, S.E., HUBHARD, T. and CHOTHIA, C. (1995): SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol., 247, 536–540.*

NEEDLEMAN, S.B. and WUNSCH, C.D. (1970): A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol., 48, 443–453.*

ORENGO, C.A., MICHIE, A.D., JONES, S., JONES, D.T., SWINDELLS, M.B. and THORNTON, J.M. (1997): CATH–a hierarchic classification of protein domain structures. *Structure, 5, 1093–1108.*

ORENGO, C.A., PEARL, F.M., LEE, D., BRAY, J.E., SILLITOE, I., TODD, A.E., HARRISON, A.P. and THORNTON, J.M. (2000): Assigning genomic sequences to CATH. *Nucleic Acids Res,28, 277–82.*

PEARSON, W.R. and LIPMAN, D.J. (1988): Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A, 85, 2444–2448.*

PEARSON, W.R. (1990): Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol, 183, 63–98.*

POIRETTE, A.R., ARTYMIUK, P.J., RICE, D.W. and WILLETT, P. (1997): Comparison of protein surfaces using a genetic algorithm. *J. Comput. Aided Mol. Des., 11, 557–569.*

PONTING, C.P. and RUSSELL, R.R. (2002): The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct., 31, 45–71.*

ROSEN, M., LIN, S.L., WOLFSON, H. and NUSSINOV, R. (1998): Molecular shape comparisons in searches for active sites and functional similarity. *Protein Eng., 11, 263–277.*

RUSSELL, R.B. (1998): Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J. Mol. Biol., 279, 1211-1227.*

SCHMITT, S., KUHN, D. and KLEBE, G. (2002): A new method to detect related function among proteins independent of sequence and fold homology. *J. Mol. Biol., 323, 387–406.*

SCHOMBURG, I., CHANG, A. and SCHOMBURG, D. (2002): BRENDA, enzyme data and metabolic information. *Nucleic Acids. Res., 30, 47–49.*

SIEMON, R. (2001): Einige Werkzeuge zum Einsatz von selbstorganisierenden Neuronalen Netzen zur Strukturanalyse von Wirkstoff-Rezeptoren. *Diplomarbeit, 13.2.2001, FB Mathematik u. Informatik.*

STARK, A., SUNYAEV, S. and RUSSELL, R.B. (2003): A model for statistical significance of local similarities in structure. *J. Mol. Biol., 326, 1307–1316.*

STARK, A. and RUSSELL, R.B. (2003): Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures. *Nucleic Acids Res., 31, 3341–3344.*

WALLACE, A.C., LASKOWSKI, R.A. and THORNTON, J.M. (1996): Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci., 5, 1001–1013.*

WALLACE, A.C., BORKAKOTI, N. and THORNTON, J.M. (1997): TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci., 6, 2308–2323.*

WANGIKAR, P.P., TENDULKAR, A.V., RAMYA, S., MALI, D.N. and SARAWAGI, S. (2003): Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J. Mol. Biol., 326, 955–978.*

WATERMAN, M.S. (1984): General methods for sequence comparison. *Bull. Math. Biol., 46, 473–500.*

Classification of *In Vivo* Magnetic Resonance Spectra

Björn H. Menze[1], Michael Wormit[2], Peter Bachert[2], Matthias Lichy[2,3], Heinz-Peter Schlemmer[2,3], and Fred A. Hamprecht[1]

[1] Multidimensionale Bildverarbeitung,
Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR),
Universität Heidelberg, 69120 Heidelberg, Germany
[2] Deutsches Krebsforschungszentrum (dkfz), 69120 Heidelberg, Germany
[3] Radiologische Diagnostik, Universitätsklinik, 72076 Tübingen, Germany

Abstract. We present the results of a systematic and quantitative comparison of methods from pattern recognition for the analysis of clinical magnetic resonance spectra. The medical question being addressed is the detection of brain tumor. In this application we find regularized linear methods to be superior to more flexible methods such as support vector machines, neural networks or random forests. The best preprocessing method for our spectral data is a smoothing and subsampling approach.

1 Introduction

The use of magnetic resonance (MR) is a well established and widespread standard in medical imaging. Less known is the use of magnetic resonance spectroscopy (MRS) for the *in vivo* analysis of the cell metabolism.
Metabolites evoke a specific spectral pattern, which is characteristic for a number of tissue types. Changes in this spectral signature allow for a diagnosis of certain pathophysiologies.

For the extraction of diagnostic information a further processing of the data is indispensable. Due to their high potential of automation, we focus on methods of *pattern recognition* and *machine learning*.

Our aim is the detection of recurrent tumors after radiotherapy by means of MRS. In this application standard imaging methods usually fail. Remaining brain lesions cannot be diagnosed reliably based on the intensity images provided by ordinary imaging methods. Intracranial biopsies, being considered as *gold standard*, do not guarantee a fully reliable result either and are associated with a considerable lethal risk of up to one percent. As a consequence, biopsies are not applicable in routine follow-up examinations and any other complementary information such as the one inherent to MRS signals is desirable (Howe and Opstad (2003)). While nearly all clinical MR scanners are able to acquire MR spectra, the know-how for interpretating these data is still rare among radiologists. A reliable automated method has the potential of making MRS accessible to a wider group of clinical practitioners.

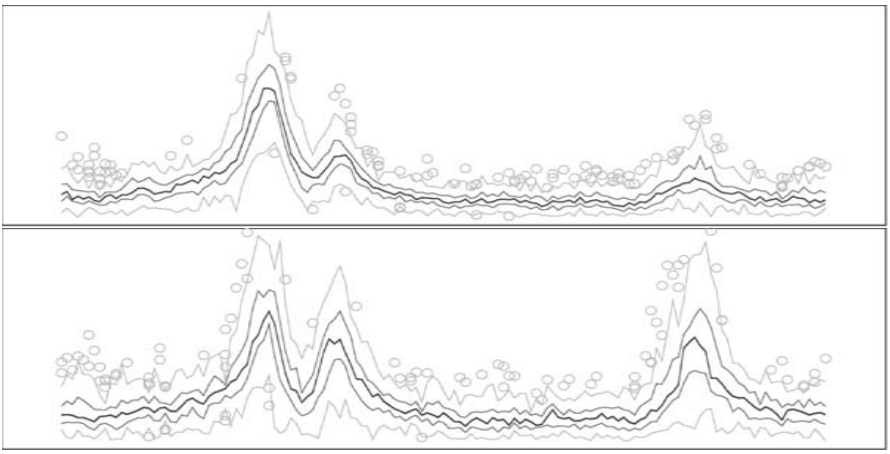


Fig. 1. Spectral pattern. Top: tumor group; bottom: non-tumor group. Central lines: median/quartiles; outer lines & dots: outlier

2 Data

2.1 General features

In vivo magnetic resonance spectra share a number of features with other spectral data. A high correlation of the P spectral channels is typical for this kind of data. As a consequence, the intrinsic dimensionality is low. In our case, only three to five resonance lines are observable (Fig. 1). As the acquisition time is limited in clinical practice, the signal-to-noise-ratio is poor. Some spectra are additionally corrupted by technical artifacts or uneven baselines. Generally, as in most medical studies of this kind, the number of observations N is much smaller than the number of explanatory variables P. In our data set, we have $N = 58$ and $P \leq 350$, depending on the preprocessing.

2.2 Details

The data set used in our survey comes from a retrospective study on the use of MRS in the evaluation of suspicious brain lesions after stereotactic radiotherapy (Schlemmer et al. (2001)). The spectra were acquired on a 1.5 Tesla MR Scanner at the German cancer research center (dkfz), Heidelberg, with long echo time (TE=135ms) by single-voxel-MRS sequences.

The study comprises a total of 58 spectra, recorded from 56 patients in a time span of several years. (Two patients participated twice in the study, at different time points.) The spectra fall into two classes: 30 of them stem from recurrent tumors, the remaining 28 from non-tumorous brain lesions (Fig. 1). The final assignment of a spectrum to either of these classes was confirmed by clinical follow-up examinations.

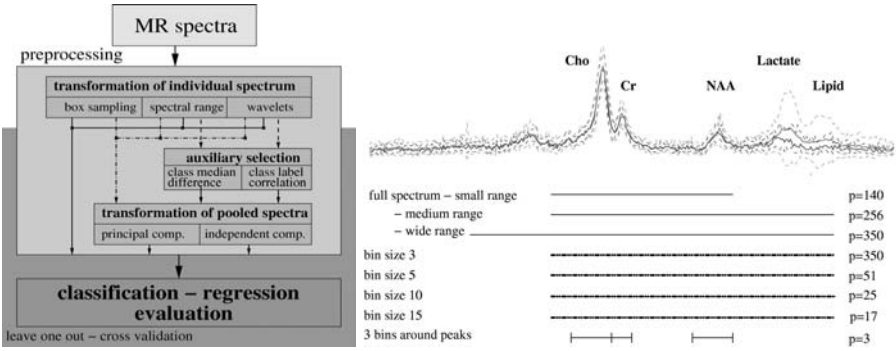


Fig. 2. Left: Flowchart of the different algorithms in the preprocessing step. All transformations involving the whole data set were part of the cross-validation process. Right: Plot of the tumor group, with an indication of the spectral regions chosen for the corresponding preprocessing; number of parameters P. Wavelet transformations were performed on the medium spectral range data set. Solid lines (upper/lower): mean/median; dashed lines (inner/outer): quartiles/variance.

Apart from filtering out the water resonance line and normalizing to the cumulative sum from the spectral region of the three most prominent peaks (choline, creatine and NAA resonances), no further preprocessing was performed on the absolute MR spectra.

3 Methods

We divide the algorithms applied to the data in two groups: preprocessing algorithms and classification algorithms. The algorithms in the first group disregard any class label information, while the latter use the group membership as an integral part.

Conceptually, the process of pattern recognition is often described as a sequence of feature extraction, feature selection and subsequent classification. For the purpose of feature extraction, the data are often transformed to a new space, the basis of which can be chosen independently of the data (as in wavelets) or dependent on the pooled observations (such as independent component analysis). Feature selection can either be explicit, as in a univariate preselection step, or implicit in the final classification. In feature extraction, a number of optional preprocessing methods were evaluated in a combinatorial way (cf. Fig. 2). In the end, we had about 50 differently preprocessed representations of the initial data set to evaluate the classifiers and regression methods on.

The question how to properly evaluate a benchmark of different classifiers on a small data set is yet unanswered, and hence the assessment of our results is anything but straightforward. Even for our limited problem it is

not clear, how to deduce general results from a wide range of combinations of preprocessing and classification without getting trapped in overfitting or overmodelling.

3.1 Evaluated algorithms

For the preprocessing, we applied a number of transformations individually to each spectrum. Some of these preprocessing algorithms were as simple as discarding certain spectral ranges (cf. fig. 2). The resulting data sets varied only in the number of spectral channels and the number of peaks visible in the graph of the spectral pattern. We also performed a smoothing and subsampling operation, called binning: it entails an accumulation of all the values from a certain number of adjacent spectral channels within a bin of predefined width, e.g. 3, 5, 10 or 15 channels. Another standard procedure in spectral preprocessing is to accumulate the spectral parameters into a single value within certain predefined spectral regions, e.g. around single resonance lines. For this we defined three bins around the three most prominent peaks. In addition, spectra were expressed in a dyadic wavelet basis (notably Daubechies-4 wavelets). Finally, continuous wavelets and wavelet packages were evaluated as possible preprocessing steps (for an overview see Fig. 2).

Also, there were transformations as adapted from the full data set. Principal component analysis (PCA) was used in conjunction with a follow-up regression step (principal component regression), while independent component analysis (ICA) was optionally performed on all data sets obtained from other preprocessing steps.

If necessary, an auxiliary selection was applied beforehand, in order to reduce the number of variables P approximately to the number of samples N. A ranking was performed according to the class label difference or the class label correlation of the single variables. In particular, it was optionally applied to the wavelet transformed data and the medium and wide range spectral vector data.

In the classification step, we evaluated fourteen different classification or regression methods. For the latter, a threshold was adjusted to obtain a binary result from the predicted values. Standard classifiers under study were *linear discriminant analysis* (Rao's LDA) and *k-nearest-neighbours* (knn). The first was also applied as stepwise LDA, with a F-value criterion. Regression methods using pooled data information were *principal component regression* (PCR) and *partial least squares* (PLS). Besides ordinary *multivariate linear regression* and *logistic regression* we used regularized multivariate linear regression methods, namely: *ridge regression* (here: being equivalent to *penalized discriminant analysis*), the *lasso*, *least angle regression* (LARS) and *forward selection*. From the classifiers, we evaluated *support vector machines* (using radial basis functions), feed-forward *neural networks* (nnet) and *random forests*. The hyperparameter of most of these methods was varied from

$\lambda = 1..12$ (see Table 1). Support vector machines, neural networks and random forests were evaluated with parameters varied in a grid search according to (Meyer et al. (2003)).

All computations were performed using the R computing language. For the algorithms mentioned above, we used the implementations available from the CRAN R repository (cran.r-project.org).

3.2 Benchmark settings

The optimization of real-world problems is rarely amenable to a one-stage-solution. A good representation of the problem in a low-level description usually has to be found in a first step, in order to obtain the desired high-level information in the following.

Proper benchmarking of different classifiers, even on one single data set of adequate size, is still an open question and subject to debate (Hothorn et al. (2003)). It is even more difficult to obtain results from an evaluation of processes that are composed of two essential parts.

Considering the small size of the data set, a naive selection of the best classifiers will easily result in overfitting or -modelling in spite of the cross-validation. So, having the size of the data set in mind, quantitative values should not be taken too literally and only allow for conclusions of a qualitative nature.

Within the given parameter range of each method, we assessed the classification error using the leave-one-out cross-validation. For the regression methods, we also evaluated the area under the receiver operator characteristic (ROC AuC), and the area under the precision-recall curve (PR AuC) as a performance measure.

The performance of each algorithm was measured using the best value obtained within the parameter range under study. This is probably overly optimistic, but can be understood in the light of the intrinsically low dimensional binary classification problem (see also Fig. 3). If possible (e.g. for the regression methods) top performing classifiers were checked for dimensionality and spectral interpretability.

To get a rough comparison of different preprocessing paradigms, we have pooled results as follows: at first we determined a subset of well performing classifiers (PCR, PLS and ridge regression showed to have the best overall performance with respect to our three measures). Then, we determined their optimal hyperparameters for each preprocessing scheme (binning with different sizes, wavelets, etc., cf. Fig.2). The number of correct and incorrect predictions in the leave-one-out cross-validation were evaluated for each method and interpreted as realizations from identical Bernoulli distributions. For each regression method, a Binomial distribution was fitted to these outcomes. Samples were drawn from the three distributions and concatenated into one list. Finally, the distribution of values in this list was visualized by the box-and-whisker plots in Fig. 3.

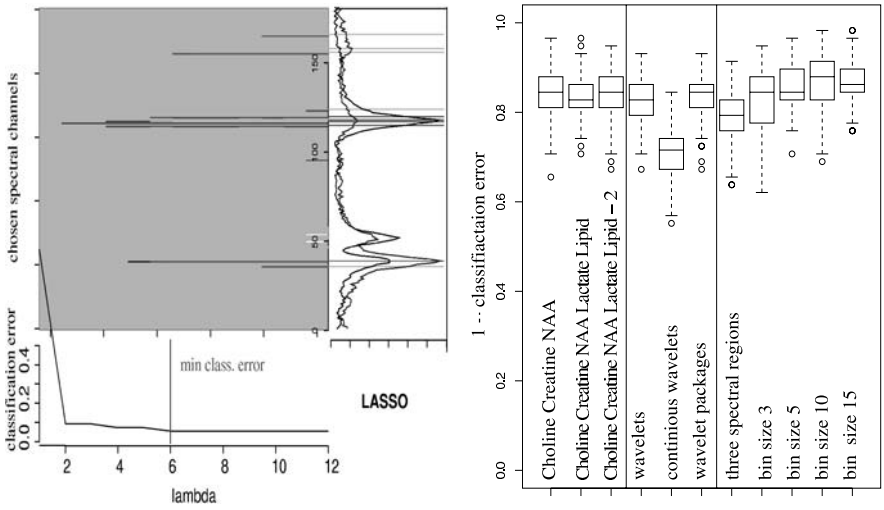


Fig. 3. Left: Spectral channels as chosen by lasso regression with corresponding classification error along hyperparameter lambda within tuning range. - Varying lambda values as determined by a - hypothetic - inner cross-validation loop would hardly affect the overall classification result, since the classification performance is nearly constant over lambda. Right: Classification performance pooled over the three top performing classifier for the given preprocessing. The only preprocessing to outperform plain spectral input (as in 'small' spectral range) is binning with an optimized bin width.

4 Results

Standard classifiers such as linear discriminant analysis or k-nearest- neighbours show rather moderate results and are outperformed by most of the other algorithms. Unconstrained regression methods only work well on the relatively low dimensional data sets. All kinds of shrinkage/regularized regressions perform considerably better, but a differentiation is difficult. On our data set principal component regression seems to perform best. Regardless of our wide grid search in the optimization of neural networks, we were seemingly not able to initialize this method correctly. Performance was bad throughout all data sets. Random forests and RBF-kernel support vector machines performed reasonably, especially after a prior dimensionality reduction. To summarize standard linear methods like PCA, PLS and ridge regression perform notably well throughout all three measures (classification accuracy (cf. Fig.4), ROC AuC, PR Auc). As measured by the ROC and PR, these three performed best on nearly any preprocessing under study.

All three spectral ranges yield a similar classification accuracy: Neither the use of lipid/lactate (as included in the medium range), nor the extension of the spectral region to the water peak (as in the wide range data set) changed the overall classification result.

	small	medium	wide	wavelet	cont w	w pack	p bins	bin 3	bin 5	bin 10	bin 15	Scale
ridge	0.83	0.84	0.84	0.84	<0.7	0.84	0.78	0.79	0.88	0.86	0.86	0.9
pls	0.81	0.83	0.79	0.83	<0.7	0.83	0.81	0.83	0.84	0.84	0.84	
pcr	0.88	0.84	0.88	0.83	<0.7	0.84	0.81	0.86	0.86	0.91	0.88	
forward	0.83	0.83	0.83	0.74	<0.7	0.83	0.81	0.81	0.88	0.88	0.9	
lasso	0.84	0.84	0.84	<0.7	0.83	0.83	0.81	0.81	0.84	0.88	0.88	
lars	0.84	0.84	0.84	<0.7	0.79	0.81	0.81	0.81	0.84	0.88	0.88	
knn	0.84	0.79	0.83	0.81	0.81	0.79	0.72	0.76	0.76	0.81	0.84	
lda	0.72	0.78	<0.7	0.83	0.79	0.79	0.78	<0.7	<0.7	0.76	0.79	
svm	0.81	0.78		0.76					0.88	0.88	0.88	
nn	<0.7	<0.7		<0.7					<0.7	<0.7	<0.7	
rndtree	0.71			0.78					0.86	0.87	0.87	
logit							0.76				0.72	0.7
reg							0.78			0.74	0.78	N.A.

Fig. 4. Partial overview of the results. Scale: classification accuracy. Classifier & regression methods: see text; preprocessing: small/medium/wide spectral ranges; wavelets, continuous wavelets, wavelet packages; bins around peaks, binning width 3/5/10/15. N.A.: no results available.

For an optimal bin width, smoothing and subsampling as performed by the binning approach, proved to be the best preprocessing.

The manual selection of bins around the visible peaks is somewhat worse than using the full spectral vector and seems to be – in our case – an inappropriate way of including *a priori* knowledge into the preprocessing.

The application of a wavelet transformation does not alter the classification performance compared to the raw spectrum. Without a preselection, the continuous wavelet transformation shows poor results. However, the use of wavelets that are smoother than the Daubechies 4 type we used, might result in better performance.

Generally, a preselection (from the auxiliary selection step) either on the wavelet transformed data or on the wide spectral ranges does not impair the classification performance. Nevertheless, it does not increase it over the performance of the respective smaller data sets (small spectral range, normal wavelet transform) either. A difference between the two univariate preselectors cannot be found.

The application of ICA does not improve the results, regardless of the number of mixing sources. Neither do the loadings of the independent components found by the algorithm help to interprete the data better than PCA, nor does the use of the ICA scores improve the following classification step compared to the performance obtained by PCA.

Table 1. Parameter range under study and optimal parameters on the differently preprocessed data (compare Fig. 4).

classifier/ regression	parameter range	preprocessing S	M	W	dw	cw	wp	bp	b3	b5	b10	b15	function name	from R package
ridge	$\lambda = 2^{-12..12}$	-4	-2	0	-3	-8	-3	-10	-1	-2	-1	2	gen.ridge	fda
PLS	$n = 1..12$	1	2	2	1	5	1	2	2	2	2	1	pls	pls.pcr
PCR	$n = 1..12$	9	3	8	6	8	6	2	3	5	7	2	pcr	pls.pcr
lasso	$n = 1..12$	6	6	6	8	5	8	3	7	10	5	5	lars	lars
lars	$n = 1..12$	6	6	6	7	7	7	3	7	9	5	5	lars	lars
forward	$n = 1..12$	6	6	6	8	8	8	3	5	9	6	6	lars	lars
knn	$k = 1..12$	6	8	8	6	3	6	3	2	1	10	7	knn	class
svm	$c = 2^{-5..12}$	0	0	-	2	-	-	-	-	2	4	4	svm	e1071
	$\gamma = 2^{-10..12}$	-8	-8	-	-10	-	-	-	-	-5	-10	-5		
nnet	$s = 1..5$	5	5	-	5	-	-	-	-	5	5	5	nnet	nnet
	$d = 0.1..1$	0.2	0.2	-	0.2	-	-	-	-	0.2	0.2	0.2		
randForest	$t = 25..200$	25	-	-	50	-	-	-	-	75	25	75	random-	random-
	$m = 1..7$	5	-	-	6	-	-	-	-	6	5	6	Forest	Forest
	$ns = 1..12$	10	-	-	4	-	-	-	-	6	2	10		
stepw.LDA	$n = 1..8$	2	5	2	-	-	-	2	4	3	3	2	(lda)	(MASS)
LDA	-												lda	MASS
regression	-												lm	base
logit	-												glm	glm

5 Conclusions

In preprocessing, the application of binning, a smoothing along the spectral vector in conjunction with a dimensionality reduction by subsampling, improves the overall result. Regularized regression methods perform well on this binary and balanced problem. We cannot find a need to use nonlinear, 'black-box' type models. This is of some importance, as in medical applications an interpretability of the diagnostic helper is of high value.

References

HOTHORN, T., LEISCH, F., ZEILEIS, A. and HORNIK, K. (2003): The design and analysis of benchmark experiments. Technical report, SFB Adaptive Informations Systems and Management in Economics and Management Science, TU Vienna.

HOWE, A.F. and OPSTAD, K. (2003): ^{1}H MR spectroscopy of brain tumours and masses. *NMR in Biomedicine,16(3), 123–131.*

MEYER, D., LEISCH, F. and HORNIK, K. (2003): The support vector machine under test. *Neurocomputing, 55, 169–186.*

SCHLEMMER, H.-P., BACHERT, P., HERFATH, K.K., ZUNA, I., DEBUS, J. and VAN KAICK, G. (2001): Proton MR spectroscopic evaluation of suspicious brain lesions after stereotactic radiotherapy. *American Journal of Neuroradiology, 22, 1316–1324.*

Modifying Microarray Analysis Methods for Categorical Data – SAM and PAM for SNPs

Holger Schwender

University of Dortmund*
Department of Statistics
44221 Dortmund, Germany

Abstract. Common and important tasks arising in microarray experiments are the identification of differentially expressed genes and the classification of biological samples. The SAM (Significance Analysis of Microarrays) procedure is a widely used method for dealing with the multiple testing problem concerned with the former task, whereas the PAM (Prediction Analysis of Microarrays) procedure is a method that can cope with the problems associated with the latter task.

In this presentation, we show how these two procedures developed for analyzing continuous gene expression data can be modified for the analysis of categorical SNP (Single Nucleotide Polymorphism) data.

1 Introduction

All humans share almost the same DNA. Only less than 0.1% of it differs between individuals. *Single nucleotide polymorphisms* (*SNPs*) are the most common type of such variations. They are single base pair positions at which different sequence alternatives exist in a population. To be considered a SNP, a variation has to occur in at least 1% of the population. Since almost any SNP exhibits only two variants, we assume that each SNP can take three realizations: "1" if both bases that explain the SNP (one base for each of the two haploid sets of chromosomes) are the more frequent variant in the population, "2" if one of the bases is the more frequent and the other is the less frequent variant, and "3" if both are the less frequent variant.

Advances in biotechnology have made it possible to measure the levels of hundreds or even thousands of SNPs simultaneously. Common and interesting statistical tasks arising in such experiments are the identification of SNPs whose distribution strongly differs under several conditions (e.g., case/control, different kinds of cancer), the classification of biological samples using the SNP data, and the determination of clusters of SNPs with coherent patterns.

Hence, the problems (e.g., that the number of variables is much larger than the number of observations) and the tasks in such experiments are similar to the problems and tasks in the analysis of gene expression data. It is therefore

* This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

a reasonable idea to apply procedures that are well established in microarray experiments to SNP data. Such methods however cannot be applied directly, since gene expression data are continuous, while SNP data are categorical.

In this paper, we exemplify how two well-known and widely used microarray analysis methods can be modified for the analysis of SNP data. This is, on the one hand, the *Significance Analysis of Microarrays* (*SAM*) proposed by Tusher et al. (2001), and on the other hand, the *Prediction Analysis of Microarrays* (*PAM*) suggested by Tibshirani et al. (2002a). The former procedure can be used for the identification of differentially expressed genes, while the latter is a classification method.

The paper is organized as follows. In Section 2, the False Discovery Rate (FDR), an error measure that is ideal for the testing situation in the analysis of both gene expression data and SNP data, is described. This error measure is estimated by the SAM procedure presented in Section 3. In Section 4, we show how SAM can be applied to SNP data. Section 5 contains a presentation of the PAM procedure which can be modified for SNP data as described in Section 6. Finally, Section 7 discusses our modifications.

2 Multiple testing and the false discovery rate

An important task in microarray experiments is the identification of differentially expressed genes, i.e. the identification of genes whose expression levels strongly differ under several conditions. Since the goal of such an analysis is the identification of a fairly large number of genes, typically a few hundred, for further analysis, one has to accept that some of these findings are false positives.

It has turned out that the *False Discovery Rate* (*FDR*) introduced by Benjamini and Hochberg (1995) is an ideal error measure for this testing situation. Denoting the number of identified genes by R and the number of false positives by V, the FDR is defined as

$$\mathrm{FDR} = \mathrm{E}\left(\frac{V}{R}\middle| R > 0\right)\mathrm{Prob}(R > 0),$$

where the FDR will be set to 0 if there is no significant finding, i.e. if $R = 0$.

The FDR can be estimated by

$$\widehat{\mathrm{FDR}}(\alpha) = \frac{\hat{\pi}_0 \alpha m}{\max\{\#\{p_i \leq \alpha\}, 1\}},$$

where α is the acceptable error rate, p_i is the (uncorrected) p-value of the ith test, $i = 1, \ldots, m$, and $\hat{\pi}_0$ is an estimate of the prior probability π_0 that a gene is not differentially expressed (Storey and Tibshirani (2001), and Storey (2003)).

There are several ways how π_0 can be estimated. Storey and Tibshirani (2003), for example, suggest to compute $\hat{\pi}_0$ by calculating $\hat{\pi}_0(\lambda) = \#\{p_i >$

$\lambda\}/\big((1-\lambda)m\big)$ for $\lambda = 0, 0.01, \ldots, 0.95$, and setting $\hat{\pi}_0$ to $\min\big\{h(1), 1\big\}$, where h is a natural cubic spline with three degrees of freedom fitted through the data points $\big(\lambda, \hat{\pi}_0(\lambda)\big)$ weighed by $1 - \lambda$.

3 Significance analysis of microarrays

The SAM (Significance Analysis of Microarrays) procedure suggested by Tusher et al. (2001) is a widely-used method for the identification of differentially expressed genes and the estimation of the FDR.

Given the expression levels x_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$, of m genes and n biological samples, and a response y_j for each of the samples, an expression score d_i for each gene i is computed, where d_i is an appropriate statistic for testing if there is an association between the expression levels of gene i and the responses.

In the original setting, for example, Tusher et al. (2001) consider two class unpaired data. In this case, the usual t-statistic could be an appropriate test statistic. There however is one problem concerned with the t-statistic that is especially encountered in microarray experiments: Genes with low expression levels. Since Tusher et al. (2001) want to avoid that such genes dominate the results of their SAM analysis, they add a small strictly positive constant, the so called *fudge factor*, to the denominator of the t-statistic, and use this modified t-statistic as expression score for each gene. For the computation and the effects of the fudge factor, see Schwender et al. (2003).

Given the set of d_i values, $i = 1, \ldots, m$, the SAM procedure described in the following can be used to identify "significantly" large expression scores, and hence to find differentially expressed genes:

1. Compute the observed order statistics $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(m)}$.
2. Draw B random permutations of the group labels. For each permutation b, $b = 1, \ldots, B$, compute the statistics d_i^b, and the corresponding order statistics $d_{(1)}^b \leq \ldots \leq d_{(m)}^b$. Estimate the expected order statistics by $\bar{d}_{(i)} = \sum_b d_{(i)}^b / B$, $i = 1, \ldots, m$.
3. Plot the observed order statistics d_i against the expected order statistics $\bar{d}_{(i)}$ to obtain the SAM plot (see Figure 1(a)).
4. Compute $i_0 = \underset{i=1,\ldots,m}{\arg\min}\Big\{\big|\bar{d}_{(i)}\big|\Big\}$.
5. For a fixed threshold $\Delta > 0$,
   (a) compute $i_1 = \min\limits_{i=i_0,\ldots,m}\big\{i : d_{(i)} - \bar{d}_{(i)} \geq \Delta\big\}$, and set $\text{cut}_{\text{up}}(\Delta) = d_{(i_1)}$, or if no such i_1 exists, set $i_1 = m + 1$, and $\text{cut}_{\text{up}}(\Delta) = \infty$,
   (b) find $i_2 = \max\limits_{i=1,\ldots,i_0}\big\{i : d_{(i)} - \bar{d}_{(i)} \leq -\Delta\big\}$, and set $\text{cut}_{\text{low}}(\Delta) = d_{(i_2)}$, or if no such i_2 exists, set $i_2 = 0$, and $\text{cut}_{\text{low}}(\Delta) = -\infty$,

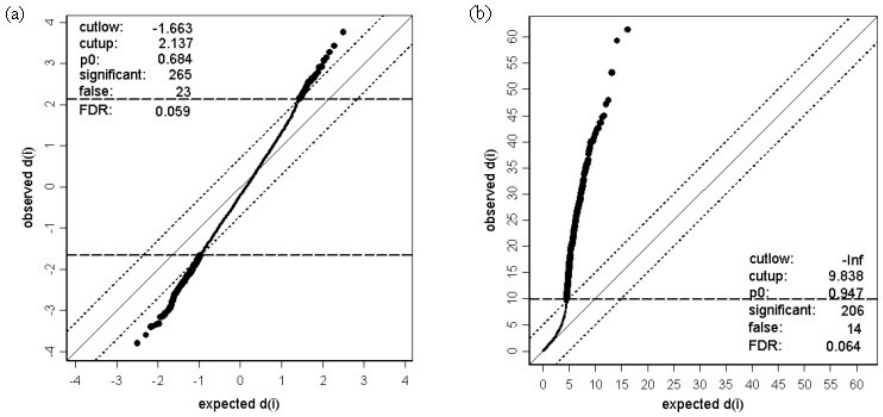


Fig. 1. SAM Plots. Plots of the ordered observed expression scores $d_{(i)}$ against the ordered expected expression scores $\bar{d}_{(i)}$ for (a) an analysis of two class unpaired gene expression data ($\Delta = 0.7$), and (b) an analysis of two class (unpaired) SNP data ($\Delta = 5$). The horizontal dashed lines represent the lower and upper cut-points, $\text{cut}_{\text{low}}(\Delta)$ and $\text{cut}_{\text{up}}(\Delta)$, respectively.

(c) call all genes with $d_i \geq \text{cut}_{\text{up}}(\Delta)$ positive significant, and all genes with $d_i \leq \text{cut}_{\text{low}}(\Delta)$ negative significant,

(d) estimate the FDR by

$$\widehat{\text{FDR}}(\Delta) = \hat{\pi}_0 \frac{(1/B)\sum_b \#\left\{d_i^b \notin \big(\text{cut}_{\text{low}}(\Delta), \text{cut}_{\text{up}}(\Delta)\big)\right\}}{\max\{i_2 + m - i_1 + 1, 1\}},$$

where $\hat{\pi}_0$ is the natural cubic spline based estimate described in Section 2.

6. Repeat step 5 for several values of the threshold Δ. Choose the value of Δ that provides the best balance between the number of identified genes and the estimated FDR.

4 SAM applied to single nucleotide polymorphisms

SAM has been developed for the analysis of (continuous) gene expression data. It is however easy to modify SAM for the analysis of (categorical) SNP data. Actually, the SAM algorithm presented in the previous section needn't to be modified. One only has to find an appropriate test statistic d. Such a score is given by the Pearson's χ^2-statistic

$$\chi_i^2 = \sum_{k=1}^{K}\sum_{t=1}^{3} \frac{\left(n_{kt}^{(i)} - \tilde{n}_{kt}^{(i)}\right)^2}{\tilde{n}_{kt}^{(i)}}, \tag{1}$$

Table 1. Contingency table for testing if the distribution of the 3 levels of SNP i, $i = 1, \ldots, m$, strongly differs between K groups. Note that the total number n and the group sizes $n_{k\cdot}$, $k = 1, \ldots, K$, are the same for all SNPs.

	1	2	3	$\sum$
Group 1	$n_{11}^{(i)}$	$n_{12}^{(i)}$	$n_{13}^{(i)}$	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Group K	$n_{K1}^{(i)}$	$n_{K2}^{(i)}$	$n_{K3}^{(i)}$	$n_{K\cdot}$
$\sum$	$n_{\cdot 1}^{(i)}$	$n_{\cdot 2}^{(i)}$	$n_{\cdot 3}^{(i)}$	n

where $\tilde{n}_{kt}^{(i)} = n_{k\cdot} n_{\cdot t}^{(i)}/n$ (see Table 1), that can be used to test if the distribution of the levels of SNP i, $i = 1, \ldots, m$, strongly differs under several conditions or between several groups, respectively. We therefore compute for each gene i the value of the χ^2-statistic and then set $d_i = \chi_i^2$.

Even though the SAM procedure actually assumes that the rejection region of the test statistic is two-sided, the algorithm presented in the previous section can also handle one-sided rejection regions as the rejection region that corresponds to Pearson's χ^2-statistic (1). In such a case the lower cut-point $\text{cut}_{\text{low}}(\Delta)$ is set to $-\infty$, and there thus is no negative significant gene. Figure 1(b) displays a SAM plot for the analysis of SNP data.

5 Prediction analysis of microarrays

The PAM (Prediction Analysis of Microarrays) procedure proposed by Tibshirani et al. (2002) is a discrimination method based on *nearest shrunken centroids* that can cope with high-dimensional classification problems.

Recall from Section 3 that x_{ij} is the expression level of gene i and sample j, $i = 1, \ldots, m$, $j = 1, \ldots, n$, and that a response y_j has been observed for each sample j. Since we here assume that the samples are independent and come from K different classes (with $K \ll n$), possible realizations of y_j are $1, \ldots, K$. Denoting furthermore the number of samples in class k by n_k, the PAM procedure works as follows:

1. For each gene i, $i = 1, \ldots, m$, compute the centroid $\bar{x}_{ik}$ of each class k, i.e. the average expression level of gene i in class k, $k = 1, \ldots, K$, and the overall centroid $\bar{x}_i = \sum_k n_k \bar{x}_{ik}/n$.

2. Compute

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k(s_i + s_0)},$$

where s_0 is the fudge factor (see Section 3), $m_k = \sqrt{1/n_k + 1/n}$, and

$$s_i^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{j:y_j=k}\big(x_{ij}-\bar{x}_{ik}\big)^2$$

is the pooled within-class variance of gene i.

3. For a set of values $\Theta > 0$, compute

$$d'_{ik} = \operatorname{sign}\big(d_{ik}\big)\big(|d_{ik}|-\Theta\big)\mathrm{I}\big(|d_{ik}|>\Theta\big), \tag{2}$$

and the shrunken centroids

$$\bar{x}'_{ik} = \bar{x}_i + m_k(s_i+s_0)d'_{ik}. \tag{3}$$

4. Choose the value of Θ that minimizes the misclassification error estimated by cross-validation (e.g., 10-fold cross-validation).

5. Given a new sample with expression levels $x^* = \big(x_1^*, x_2^*, \ldots, x_m^*\big)$,
   (a) compute the discrimination score

$$\delta_k(x^*) = \sum_{i=1}^{m}\frac{\big(x_i^*-\bar{x}'_{ik}\big)^2}{(s_i+s_0)^2} - 2\log\hat{\pi}_k$$

   for each class k, $k=1,\ldots,K$, where $\hat{\pi}_k$ is an estimate of the prior probability π_k of class k,

   (b) estimate the class probabilities $p_k(x^*)$, $k=1,\ldots,K$, by

$$\hat{p}_k(x^*) = \frac{\exp\big\{-0.5\delta_k(x^*)\big\}}{\sum_{h=1}^{K}\exp\big\{-0.5\delta_h(x^*)\big\}},$$

   (c) predict the class of the new observation by

$$\hat{k} = \underset{k=1,\ldots,K}{\arg\max}\ \hat{p}_k(x^*).$$

6 Prediction analysis of SNPs

Contrary to SAM, it is difficult to modify PAM for SNP data. Since we do not use shrunken centroids, we actually do not modify PAM. We only use some of its ideas.

For each SNP i, $i=1,\ldots,m$, the number $n_{\cdot t}^{(i)}$ of samples with SNP value t, $t=1,2,3$, is computed (cf. Table 1). Denoting the number of samples in class k, $k=1,\ldots,K$, having a SNP value t by $n_{kt}^{(i)}$, the value of

$$\chi_{ik}^2 = \sum_{t=1}^{3}\frac{\Big(n_{kt}^{(i)}-\tilde{n}_{kt}^{(i)}\Big)^2}{\tilde{n}_{kt}^{(i)}}$$

is computed to compare the distribution of the levels of SNP i in group k with the overall distribution of the levels of SNP i. Since $n_{\cdot t}^{(i)}$ of the n values of SNP i are t, we would expect that under the assumption that the group and the overall distribution are equal, $\tilde{n}_{kt}^{(i)} = n_{\cdot t}^{(i)}/n \cdot n_{k\cdot}$ of the $n_{k\cdot}$ samples in group k have a SNP value of t. Hence, $\tilde{n}_{kt}^{(i)}$ is the expected number that is also used in Section 4.

Following (2), we compute

$$\chi_{ik}^{2\prime} = \left(\chi_{ik}^2 - \Theta\right) \cdot \mathrm{I}\left(\chi_{ik}^2 > \Theta\right) \tag{4}$$

for a set of $\Theta > 0$, and choose the value of the shrinkage parameter Θ by (10-fold) cross-validation.

A new observation $x^* = (x_1^*, x_2^*, \ldots, x_m^*)$, where $x_i^* \in \{1, 2, 3\}$, $i = 1, \ldots, m$, is classified by computing the posterior probability

$$p\left(k|x_{\Theta}^*\right) = \frac{\pi_k p\left(x_{\Theta}^*|k\right)}{\sum_{h=1}^{K} \pi_h p\left(x_{\Theta}^*|h\right)},$$

where x_{Θ}^* denotes the subvector of x^* that contains the values of all SNPs with at least one non-zero $\chi_{ik}^{2\prime}$ value, and by predicting the class $\hat{k}$ of the new observation by

$$\hat{k} = \underset{k=1,\ldots,K}{\arg\max}\, p\left(k|x_{\Theta}^*\right).$$

7 Discussion

In this presentation, we have shown how procedures for analyzing continuous gene expression data can be modified for another type of genetic data, namely categorical SNP data.

The first method considered in this presentation is the SAM procedure that can be used for the identification of differentially expressed genes. SAM is relatively easy to modify for other kinds of data, since one only has to define a score for each gene/SNP that is suitable for testing if there is an association between the values of the genes/SNPs and a response variable. In the case of categorical SNP data, such an appropriate test statistic is given by the Pearson's χ^2-statistic.

As a second method, we have considered the PAM procedure. PAM is a discrimination method that uses nearest shrunken centroids. Since our approach for SNPs does not use such shrunken centroids, it is not really a modification of PAM. It however uses some of the ideas of PAM. In both procedures, a test statistic for each group and gene/SNP is computed, and genes/SNPs that are representative for the different groups are selected by successively reducing the values of the test statistics and by choosing the

reduction that minimizes the misclassification rate. This type of feature selection differs from more common approaches that screen genes by the significance of their corresponding test statistics.

A drawback of our versions of SAM and PAM modified for SNPs is that they require lots of samples. While in a microarray experiment the number of samples is typically much smaller than 50, we here should have far more than 50 samples.

The SAM version for SNPs will be contained in one of the next versions of our `R` (Ihaka and Gentleman 1996) package `siggenes` which can be downloaded from `http://www.bioconductor.org`.

References

BENJAMINI, Y. and HOCHBERG, Y. (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Ser. B 57, 289–300.*

IHAKA, R. and GENTLEMAN, R. (1996): R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics, 5, 299–314.*

SCHWENDER, H., KRAUSE, A. and ICKSTADT, K. (2003): Comparison of the Empirical Bayes and the Significance Analysis of Microarrays. *Technical Report, SFB 475, University of Dortmund, Germany.* `http://www.sfb475.uni-dortmund.de/berichte/tr44-03.pdf`.

STOREY, J. and TIBSHIRANI, R. (2001): Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays. *Technical Report 2001-28, Stanford University.* `http://faculty.washington.edu/~jstorey/papers/dep.pdf`.

STOREY, J. (2003): The positive False Discovery Rate: A Bayesian Interpretation and the q-value. *Annals of Statistics, 31, 2013–2035.*

STOREY, J. and TIBSHIRANI, R. (2003): Statistical Significance for Genome-wide Studies. *Proceedings of the National Academy of Sciences, 100, 9440–9445.*

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. and CHU, G. (2002): Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences, 99, 6567–6572.*

TUSHER, V.G., TIBSHIRANI, R. and CHU, G. (2001): Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proceedings of the National Academy of Science, 98, 5116–5121.*

Improving the Identification of Differentially Expressed Genes in cDNA Microarray Experiments*

Alfred Ultsch

Databionics Research Group
University of Marburg
35032 Marburg, Germany

Abstract. The identification of differentially expressed genes in DNA microarray experiments has led to promising results in DNA array analysis. The identification as well as many other methods in cDNA array analysis rely on correct calculations of differential colour intensity. It is shown that the calculation of logarithms of the ratio of the two color intensities (LogRatio) has several disadvantages. The effects of numerical instabilities and rounding errors are demonstrated on published data. As an alternative to LogRatio calculation, relative differences (RelDiff) are proposed. The stability against numerical and rounding errors of RelDiffs are demonstrated to be much better than for LogRatios. RelDiff values are linearly proportional to LogRatios for the range where genes are not differentially expressed. Relative differences map differential expression to a finite range. For most subsequent analysis this is a big advantage, in particular for the search of expression patterns. It has been reported that the variance of intensity measurements is a nonlinear function on intensity. This effect can be explained by an additive measurement error with constant variance. Applying the logarithm to such intensity measurements introduces the presumed nonlinear dependence. Thus in many cases no complicated variance stabilization transformation using nonlinear functions on the LogRatio expression values is necessary.

1 Introduction

In complementary DNA (cDNA) microarray experiments the data for each gene (spot) are two fluorescence intensity measurements (Parmigiani et al. (2003)). The measurements are produced by a mixture of two portions of mRNA labeled with two different fluorescent color dyes. One portion of the mRNA is labeled by the dye Cy5 producing a red fluorescence color (R), the other is marked by the dye Cy3 producing a green fluorescence color (G). The predominance of one of the colors indicates the relative abundance of the corresponding DNA sequence. This indicates the over expression of a particular gene. Equal intensities in red and green fluorescence at a spot indicate no particular over- or under expression of the corresponding gene.

* A longer version of this paper including more references can be obtained at www.mathematik.uni-marburg.de/∼ databionics

In most publications on microarray data the (binary) logarithm of the ratio R/G (LogRatio) is used, see Parmigiani et al. for an overview (Parmigiani et al. (2003)). There seem to be different types of arguments using a logarithmic transformation on intensity values: first, absolute differences are less meaningful, than relative ratios; second, convenience of visualisation; third, compensation for skewness of the distributions; four, lognormality of the distributions; five, stabilization of variance. The first argument is more an argument for the approach followed in this paper: to use relative differences. The convenience for plotting is a valid argument, if the colours are regarded separately. Skewness is compensated using log, but it is clear, that the distributions of intensities are not log normal distributed. So a log transformation changes the distribution but does not normalize it. The last argument, variance stabilisation is treated in detail in Chapter 4. In the following we will demonstrate, that using LogRatio has severe disadvantages and propose an alternative: relative differences.

2 Data sets, LogRatio, RelDiff

One of the cDNA data sets used consists of microarray experiments of Saccharomyces Cerevisiae. The focus of the experiment is the shift from anaerobic (fermentation) to aerobic (respiration) metabolism. At 7 time points during this diauxic shift the expression of 6153 genes is measured. We call this data the "DiauxicShift" data. The data is publicly available form the website *http://cmgm.stanford.edu/pbrown/explore/index.html*. Another data set was published by Eisen. The data consists of a set of 2465 gene expressions of yeast in 79 different experiments. The data is available from the website *http://www-genome.stanford.edu*. We refer to this data set as the "Yeast" data. The distribution of R and G is typically severely skewed. I.e. there are many small values and few very big values. The values for R and G, for example in the DiauxicShift data, range from 50 to 50.000.

The LogRatio is defined as LogRatio(R,G) = ld(R/G), where ld is the logarithm for basis two. If R is equal to G, then LogRatio equals 0. If LogRatio is greater then 0 (less then 0), then R is greater then G (R less then G). The distribution of LogRatios is centered around zero, but has a substantial number of values which are greater than one in absolute value. In the Yeast data set 2800 values i.e. 1.2% of LogRatios are greater than 2, 113 values i.e. 0.06% are greater than 4. In this paper we propose another value for the indication of differentially expressed genes, the relative difference (RelDiff). The relative difference is the ratio of the difference (R-G) to the mean intensity of the spot. RelDiff is defined as follows.

If R is equal to G, then RelDiff equals 0. If RelDiff is greater then 0 (less then 0), then R is greater then G (R less then G). The relative difference may also be measured in percent. This leads to the definition of RelDiff%(R,G).

$$\text{RelDiff(R,G)} = \frac{R-G}{\frac{1}{2}(R+G)} = 2 * \frac{R-G}{R+G};\ \text{RelDiff\%(R,G)} = \frac{R-G}{(R+G)} * 200\ [\%]$$

3 Comparison of LogRatio and RelDiff

Interpretation: The direct interpretation of LogRatio values is difficult except for powers of two. It is easy to see that a LogRatio of 1 means that the expression level of a particular gene is two fold. It is, however, not straight forward to see that a LogRatio value of 1.58 corresponds to a threefold, a LogRatio value of 3.322 to a ten fold over expression rate. For an interpretation of such LogRatio values one must be familiar with dual logarithms. The numerical values of RelDiff and in particular RelDiff% have a straight forward interpretation. Even an "odd" value of, for example 22.12%, for RelDiff has a direct interpretation. Such a value means that there is 22.12 percent more red color in the average luminosity of the particular spot.

Numerical stability: In many two color experiments almost all of the thousands of measured genes have an identical level of R and G. In order to investigate the numerical properties of the formulas above we assume that R is equal to G plus some small measurement error ε. I.e. we assume R = G + ε for some small error ε. For LogRatio we obtain: $\text{LogRatio}(G+\varepsilon, G) = ld\left(\frac{G+\varepsilon}{G}\right) = ld\left(1+\frac{\varepsilon}{G}\right)$.

This term results in very large negative values the closer ε gets to -G. For small values of G this might be the case. This means that there might be numerical instable LogRatio calculations. Furthermore for G close to zero the LogRatio may become very big. For RelDiff on the other hand we obtain: $\text{RelDiff}(G+\varepsilon, G) = 2 * \frac{G+\varepsilon-G}{G+\varepsilon+G} = \frac{2\varepsilon}{2G+\varepsilon} = \frac{\varepsilon}{G+0.5\varepsilon}$.

In this case ε must become as big as -2G in order to cause numerical problems. It can be concluded that RelDiff is twice as numerically stable as LogRatio. For G approaching zero, the RelDiff values approach 2. This means that the error for RelDiff is bound: $|\text{RelDiff}(G+\varepsilon, G)| \leq 2$.

For the DiauxicShift data we have observed the numerical properties empirically. R was set to G + EPS. The measurement error ε (EPS) was varied in the interval [0,10%] relative to the maximal value (max(G)) occurring in G. Figure 1 shows the resulting LogRatio and RelDiff values. The erroneous RelDiff values are consistently lower than the LogRatio errors. Figure 1 also demonstrates that the RelDiff errors are bound by 2 while the LogRatio error may become arbitrarily big.

In many two dye microarray experiments most measured intensities are very small. To analyze the numerical situation in this case we assume R = G + ε and G $\approx \varepsilon$. This gives: $\text{LogRatio}(\varepsilon+\varepsilon, \varepsilon) = ld\left(\frac{\varepsilon+\varepsilon}{\varepsilon}\right) = ld(2\varepsilon) - ld(\varepsilon)$. This is numerically instable! For ε close to zero, LogRatio values explode. For RelDiff on the other hand we get: $\text{RelDiff}(\varepsilon+\varepsilon, \varepsilon) = 2 * \frac{\varepsilon}{\varepsilon+\varepsilon+\varepsilon}$. Even if ε gets very close to zero, the denominator of the fraction in this term is always three times the numerator. This means: $\text{RelDiff}(\varepsilon+\varepsilon, \varepsilon) \cong \frac{2}{3} = 0.67$. The clear conclusion is that log ratios are numerically instable in particular for small intensities with almost equal values in the red and green intensities. Relative differences on the other hand are numerically stable with a maximum error value of 0.67, if both intensities are small and practically equal.

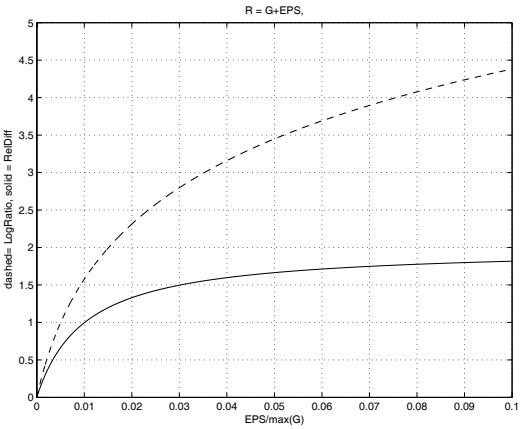


Fig. 1. LogRatios and RelDiff values for R =G+EPS in the DiauxicShift data

Rounding errors: The Yeast data is published as LogRatios with a precision of two digits after the decimal point. We examine the errors with respect to rounding by the exact calculation of LogRatio and RelDiff values, then round the values to a precision of one hundredth. From these rounded values the original value of G is reconstructed using the correct value of R. This gives a value G'. The difference between the value G' and the true value of G leads to the relative error (err) measured in percent of the true value. I.e. $\text{err} = \frac{G'-G}{G} * 100$ [%].

Figure 2 shows the relative error for the reconstructed values of G in the diauxic shift data. The left side of Figure 2 shows the error for LogRatio, the right side for rounded RelDiff. It can be seen that the error level may become very extreme for LogRatios. The conclusion is that rounding LogRatios to is considerable more critical than a rounding of RelDiff. To round LogRatios to two decimal digits, as done in Eisen et al's data is critical and may influence subsequent calculations severely. The same rounding has much less effects for RelDiff data.

Negative and zero values: In some two color microarray experiments the raw values measured for the intensities are close to the background values. Sometimes there is even more background intensity encountered than intensity inside a spot. The corrected difference between spot measurements and the background luminescence become negative in these cases. For such cases, the logarithm is undefined. This may lead to unwanted numerical errors or imaginary results in LogRatio. For RelDiff negative values are no problem. If one of the intensities is zero, this causes a numerical error for LogRatios. If R is zero, the logarithm gets arbitrarily big. If G is zero, the denominator of R/G causes an error. For RelDiff zero values in R or G are uncritical. The result is a meaningful value of RelDiff. If both intensities are zero, RelDiff is undefined. In this case, however, no intensity at all is measured in both col-

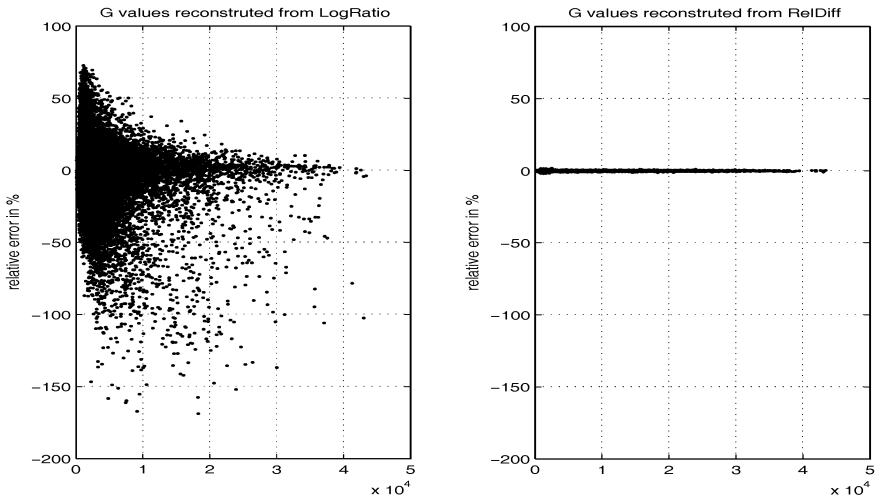
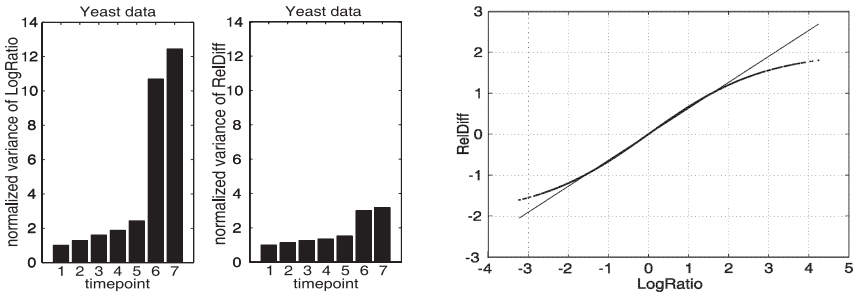


Fig. 2. LogRatio and RelDiff error values for R =G+EPS in the DiauxicShift data

ors. This case can be treated properly either by ignoring the measurement or setting the resulting RelDiff value to zero. The difficulties in the calculation of zero or negative logarithms might be the reason for the 3760 undefined values in the published Yeast data set.

Bounded in the limit: We will now investigate the properties of LogRatio and RelDiff for very large differences in color intensities. Let R >>G such that R±G≈R resp. ld(R)±ld(G)≈ld(R). We obtain: LogRatio(R,G) = $ld\left(\frac{R}{G}\right) = ld(R) - ld(G) \rightarrow ld(R)$. This means in particular, that there is no theoretical limit to the LogRatio values. Under the same assumptions for RelDiff holds: RelDiff(R,G) = $2 * \frac{R-G}{R+G} \rightarrow 2 * \frac{R}{R} = 2$.

In the same manner for G>>R Log Ratio goes to -ld(G) and RelDiff approaches -2. This means that RelDiff has a limited value range. Many two color DNA microarray experiments search for some similar expression of genes. For this a similarity measure is defined. Typical measures are Euclidian distances, correlation measures, Mahalanobis distances and others. Since the expression data of several experiments have to be compared, the variance of the data have to be taken into account. The limited scope of the RelDiff values is advantageous in this case since it limits the influence of very outlying values. In the DiauxicShift data, for example, the range in LogRatios varies by a factor up to 3.3 for the different times the experiment was performed during the diauxic shift. On the other hand the maximal ratio of the ranges of RelDiff is only 2.3, i.e. 45% less. In Figure 3 the variances of LogRatio and RelDiff values of the microrarrays measured at different time points during the diauxic shift can be compared. The variances are normalized such that the smallest variance is 1. Many authors use Euclidian distances to find similar expression patterns, for example, Hain and Ultsch (2002). If the different vari-



(a) Comparison of the variances (b) Proportionality

Fig. 3. Properties of LogRatio and RelDiff for the DiauxicShift data

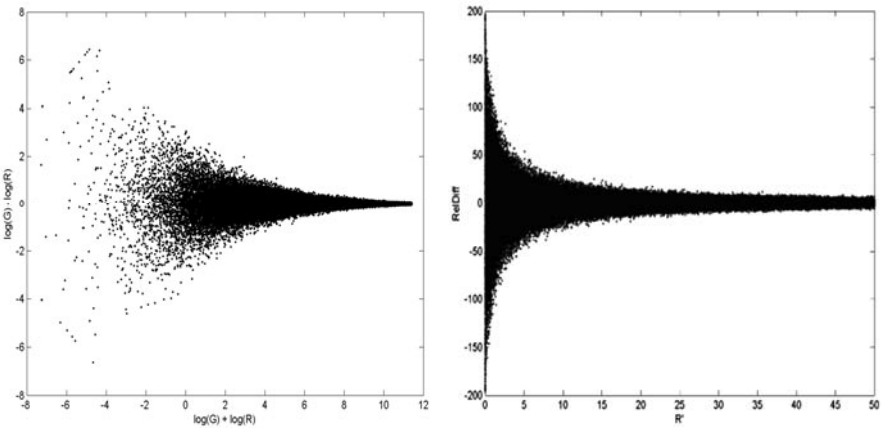
ances are not compensated, the Euclidian distance is mainly determined by the experiments with largest variance. To normalize by the empirical variance is, however, a big problem, since the LogRatio values are not normally distributed. The "fat tails" of the distributions (see Figure 2) invalidate simple variance calculations. The same holds for many other distance calculations, e.g. for the correlation distances. For RelDiff values this normalization problem is alleviated due to the natural limitation of the range. From Figure 3 it can be concluded, that for RelDiff a compensation for variance is less critical. In this experimental setting the variances differ by a factor of 3 in RelDiff. Log Ratio variances differ by factor of more than 12 in the same experiment!
Proportionality: The relationship between LogRatio values and RelDiff values is considered. The LogRatio values are almost directly proportional in the range of [-2;2]. A derivation of this can be obtained using approximations for logarithms. In practice this can be seen in Figure 3 for the DiauxicShift data. Within a range of 1/4 to 4 fold of gene expression the values of LogRatio are practically proportional to RelDiff values. Outside of this range the RelDiff values become over proportionally smaller in absolute values. This means that all subsequent calculations that rely on the relative differences of LogRatio values are still possible with RelDiff values. Differences in the analysis can only be expected, if the absolute values of extremely over- or under expression are important. The limitation to the range [-2 2] is mostly advantageous. In many publications gene expressions are depicted as color coded bars. For RelDiff data it is possible to assign a definite range. Extreme values are fixed to 2 for RelDiff. From the theoretical considerations above, it is also possible assign the range [-0.67 0.67] to a color indicating "hardly a differential expressed gene". A de facto convention for this color is yellow.

4 Stabilization of variance

Many researchers report, that the variance of gene intensities varies systematically with intensity. It is observed, that the variance reduces with increasing intensity. It is also noted that this phenomenon is a nonlinear function of log intensities. Complex mathematical transformations such as a combination of logarithms, squaring and squarerooting or synonymously the application of arcsinh have been proposed. We believe, however, that this is an effect artificially introduced by using LogRatio. Lee and O'Connell have estimated the variance at each expression level using local estimation methods (Parmigiani pp 163-184). They found that the variance Var(I) decays exponential with logarithmic intensity log(I). This means Var(I) = exp(-c log(I)), for some positive constant c. With a small algebraic calculation this gives:

$$\mathrm{Var(I)} = e^{-c\log(I)} = \left(e^{\log(I)}\right)^{-c} = I^{-c}$$

Assume an N(0,s) distributed measurement error for the intensities. Let R' and G' be the true measurements. The measured intensities R and G are then R = R' + e1 resp. G = G' + e2, with e1 and e2 drawn from N(0,s). We have simulated such measurements for 500 true measurements at each intensity level within the range 0.01 to 50. Figure 4 shows the so called MA plot. This plots the difference of the logarithmic intensities vs. their sums.



(a) MA plot of simulated data (b) RelDiff vs. true values

Fig. 4. Both (a) and (b) with an additive error of constant variance.

As it can be seen, the variance of this data depends nonlinearly on the intensity values. The similarity of this figure with published MA plots from

measured data is striking. The nonlinear decreasing dispersion of the measurements may therefore be explained by the application of the logarithm. Plotting the RelDiff values vs. the true intensities shows the I^{-c} obtained by the calculation above (see Figure 4). This is expected from the definition of RelDiff. For small intensities RelDiff is dominated by the denumerator, in particular by the inverse of the measurement error drawn from N(0,s).

5 Summary

This paper investigates the log ratio calculations for DNA array experiments using two color dyes. The calculation of a logarithm of the ratio of the two color intensities has several disadvantages. The numerical stability is questionable for measurement errors and for small intensities. It has also been shown that rounding errors are important for LogRatio calculations. Both effects have been demonstrated to play a critical role for published data. As an alternative to LogRatios the calculation of relative differences (RelDiff) is proposed. These values are directly proportional to LogRatio values when the two color intensities are about identical. In contrast to LogRatio, however, RelDiff values are bound. This is a big advantage if similarities on genes are calculated. For logarithms on intensities a nonlinear correlation of variance to intensities has been observed. Complicated variance stabilizing transformations have been proposed to compensate this. As shown here, this effect can also be explained by an additive error of constant variance for the intensity measurements. The nonlinearity is then produced mainly by the logarithmic transformation. The numerical stability for RelRiff is much better than for LogRatio. For small and almost equal intensities in both colors the error of RelDiff is finite and can be calculated precisely. The error for LogRatio may, however, become arbitrarily big for such measurements. Rounding has a much smaller influence on the precision of the obtained values for RelDiff than for LogRatios. Finally the values of RelDiff have a direct interpretation while LogRatio values require the knowledge of binary logarithms. In summary this paper shows that RelDiff is much better for DNA microarray analysis than LogRatio calculation.

References

HAIN, T. and ULTSCH, A. (2003): MDEAT - A new databionic evaluation and analysis tool to identify the virulence regulon of Listeria monocytogenes as a model system, *Proc. European Conference on Prokaryotic Genomes*, Göttingen, 2003.

PARMIGIANI, G. et al. (2003): *The Analysis of Gene Expression Data*. Springer, New York.

PhyNav: A Novel Approach to Reconstruct Large Phylogenies

Le Sy Vinh[1], Heiko A. Schmidt[1], and Arndt von Haeseler[1,2]

[1] NIC, Forschungszentrum Jülich, D-52425 Jülich, Germany
[2] Bioinformatik, HHU Düsseldorf, D-40225 Düsseldorf, Germany

Abstract. A novel method, PhyNav, is introduced to reconstruct the evolutionary relationship among contemporary species based on their genetic data. The key idea is the definition of the so-called minimal k-distance subset which contains most of the relevant phylogenetic information from the whole dataset. For this reduced subset the subtree is created faster and serves as a scaffold to construct the full tree. Because many minimal subsets exist the procedure is repeated several times and the best tree with respect to some optimality criterion is considered as the inferred phylogenetic tree. PhyNav gives encouraging results compared to other programs on both simulated and real datasets.

A program to reconstruct phylogenetic trees based on DNA or amino acid based is available (`http://www.bi.uni-duesseldorf.de/software/phynav/`).

1 Introduction

One objective in phylogenetic analysis is the reconstruction of the evolutionary relationship among contemporary species based on their genetic information. The relationship is described by an unrooted bifurcating tree on which the leaves represent contemporary species and the internal nodes can be thought of as speciation events. The total number of unrooted bifurcating trees with $n \geq 3$ leaves is $\prod_{i=3}^{n}(2i-5)$ (cf. Felsenstein (1978)). This number increases rapidly with n. For $n = 55$ sequences the number of trees exceeds the estimate of 10^{81} atoms in the known universe.

Commonly used tree reconstruction methods can be classified into three groups: (1) Minimum evolution methods (e.g., Rzhetsky and Nei (1993)), (2) maximum parsimony methods (e.g., Fitch (1971)), and (3) maximum likelihood methods (e.g., Felsenstein (1981)). Among these the maximum likelihood (ML) methods are statistically well founded and tend to give better results. An overview is given in Swofford et al. (1996) and Felsenstein (2004).

Here, we propose a new heuristic tree search strategy, which reduces the computational burden. Details how to construct minimal k-distance subsets are given in section 2. In section 3 we will describe the PhyNav algorithm and how it elucidates the landscape of possible optimal trees. The algorithm is then applied to simulated as well as biological data (Section 4).

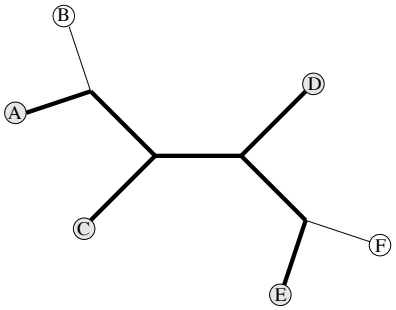


Fig. 1. An unrooted bifurcating tree of 6 species $\{A, B, C, D, E, F\}$. The bold tree is the scaffold with minimal 2-distance subset $\{A, C, D, E\}$.

2 Minimal *k*-distance subsets

First, we introduce the concept of *k-distance representatives*. A sequence s is said to be a k-distance representative for a sequence s' in a tree T if and only if their topological distance $d(s, s')$ in T, that is the number of branches on the path from s to s', is smaller or equal to $k \geq 0$. The smaller the value of k is the better a sequence s represents sequence s', and vice versa.

The k-distance representative sequence concept is now used to introduce minimal k-distance subsets. A subset S_k of sequences is called a *minimal k-distance subset* of an n-sequence set S if and only if the following two conditions hold:

1. For each sequence $s \in S$, there exists a sequence $s' \in S_k$ such that the sequence s' is a k-distance representative for the sequence s.
2. If we remove any sequence s' from S_k, S_k will violate the first condition. That means, the subset cannot be reduced any further.

The idea behind minimal k-distance subsets is that the phylogenetic information in the sequence subset S_k represents phylogenetic information from the whole set. $k = 3$ is a good choice according to our experience because it prevents the deletion of too many sequences as well as the removal of sequences that provide information to bridge long paths between distantly related subtrees.

A sequence $\bar{s} \notin S_k$ is then called a *remaining sequence*. The set $\overline{S_k} := S \backslash S_k$ of all such sequences, which remain to be added to S_k to obtain the full set S, is called *remaining set*.

Since $|S_k| \leq |S|$, the subtree T_k from subset S_k can usually be constructed in less time than the full tree. This subtree is used as a scaffold to build a full tree containing all sequences by adding all sequences $\bar{s} \in \overline{S_k}$.

For example, sequences A and B in the tree in Figure 1 are 2-distance representative of each other, as are E and F. The sequence subsets $\{A, C, D, E\}$,

$\{A, C, D, F\}$, $\{B, C, D, E\}$ and $\{B, C, D, F\}$ are minimal 2-distance subsets of the full set $\{A, B, C, D, E, F\}$.

3 The PhyNav algorithm

The Navigator algorithm is a three-step procedure: (1) the *Initial step*, (2) the *Navigator step*, and (3) the *Disembarking step*. We could use the algorithm with any objective function, e.g. maximum parsimony, maximum likelihood, to create a list of possible optimal trees. According to the objective function the best tree found is taken as the inferred phylogeny. In the PhyNav program we use the maximum likelihood principle because of its better accuracy.

Initial step: We employ some fast tree reconstruction method to create an initial tree. To that end, PhyNav uses the BIONJ (Gascuel (1997)) an improved Neighbor-Joining algorithm (Saitou and Nei (1987)) with the pairwise evolutionary distances and a fast nearest neighbor interchange (NNI) operation as described by Guindon and Gascuel (2003) to create the initial tree. This tree is then called the current best tree and denoted as T_{best}. T_{best} is used to construct the k-distance subsets.

Navigator step: Finds a minimal k-distance subset S_k and constructs the corresponding subtree T_k. Note, that there exist many minimal k-distance subsets and each can be determined in time of $O(n^2)$ (details are left out due to limited space). From the minimal k-distance subset S_k the corresponding subtree T_k could be created by several tree reconstruction methods. In PhyNav, T_k is created by optimizing the subtree T_{sub} of T_{best} induced by the leaves in S_k using NNI operations.

Disembarking step: Constructs the whole tree T based on the scaffold tree T_k using the k-distance information. To this end, PhyNav inserts the remaining sequences into the scaffold as follows: (1) assign T by T_k; (2) insert each remaining sequence $\bar{s} \in \overline{S_k}$ into an external branch e of T such that the corresponding leaf s_e adjacent to e is a k-distance representative for $\bar{s}$. If there are more than one external branches possible one branch is selected randomly; (3) apply NNI operations to T to compensate for incorrect placements. The new resulting whole tree is called intermediate tree. If the intermediate tree T has a better score than the current best tree T_{best}, replace T_{best} by T.

It cannot be guaranteed that T_k determined in the *Navigator step* is the optimal tree for S_k due to the use of heuristics. Even if T_k is the best tree it does not guarantee that tree T will be the optimal full tree. Hence, the *Navigator* and *Disembarking steps* are repeated several times. Then the program stops and the best tree T_{best} is considered as the final phylogenetic n-tree.

4 The efficiency of PhyNav

To measure the accuracy and the time-efficiency of PhyNav we reconstructed phylogenetic trees from simulated as well as biological datasets. The results

are compared to the results of other programs, in particular, Weighbor (Bruno et al. (2000); version 1.2) and PHYML (Guindon and Gascuel (2003); version 2.1).

Computing times were measured on a Linux PC Cluster with 2.0 GHz CPU and 512MB RAM.

4.1 Simulated datasets

Analysis

To evaluate the accuracy we performed simulations. To simulate realistic datasets we performed the simulations on a tree topology reconstructed from a real dataset. To that end an elongation factor (EF-1α) dataset with 43 sequences was used. The dataset as well as the tree was obtained from TreeBase (`http://www.treebase.org`, accession number S606, matrix accession number M932). The branch lengths of the tree topology were inferred using the TREE-PUZZLE package (Strimmer and von Haeseler (1996), Schmidt et al. (2002); version 5.1).

Based on that tree topology datasets were simulated using Seq-Gen (Rambaut and Grassly (1997); version 1.2.6) assuming the Kimura 2-parameter model with an transition:transversion ratio of 2.0 (Kimura (1980)). 1,000 datasets each were simulated with sequence lengths of 700 and 1000 bp.

The trees for simulated data sets were reconstructed using PhyNav, Weighbor (Bruno et al. (2000); version 1.2) and PHYML (Guindon and Gascuel (2003); version 2.1).

All programs were run with default options. The evolutionary model and its parameters were set to the simulation parameters. The PhyNav options were set to 5 repetitions and $k = 3$.

The results of the tree reconstructions were compared using two different methods. First the percentage of correctly reconstructed tree topologies was derived for each program and sequence length. To measure the variability of the results for each program, the Robinson-Foulds distance (Robinson and Foulds (1981)) was computed from each tree to the 'true tree' and the average was taken for each program and sequence length. The Robinson-Foulds distance is the number of splits (bipartitions) in the two trees, which occur in only one of the trees but not in the other. If the trees are identical their distance is zero.

Results for the simulated datasets

Tables 1(a) and 1(b) display the results for PhyNav, PHYML, and Weighbor. Both tables show that Weighbor (Table 1(c)) is out-performed by both PHYML and PhyNav. PHYML and PhyNav perform similarly well, both in the percentage of correctly reconstructed trees as well as in their average Robinson-Foulds distance to the 'true tree'. However, PhyNav shows slightly better values for all analyses.

Table 1. Results for the simulated datasets: (a) percentage of correctly reconstructed trees, (b) average Robinson-Foulds distance between the 'true tree' and the reconstructed trees, and (c) average runtime of tree reconstruction (1000 simulations per parameter setting).

(a) Percentage of correct trees.

	Weighbor	PHYML	PhyNav
700 bp	2.4	12.3	13.1
1000 bp	9.6	33.7	33.9

(b) Robinson-Foulds distance.

	Weighbor	PHYML	PhyNav
700 bp	7.57	4.09	3.96
1000 bp	4.62	2.11	2.07

(c) Average runtime.

	Weighbor	PHYML	PhyNav
700 bp	$3s$	$7s$	$52s$
1000 bp	$4s$	$9s$	$66s$

4.2 Biological datasets

Analysis

The PhyNav algorithm was applied to large biological datasets to test its efficiency on real datasets. Three datasets have been obtained from the PANDIT database version 7.6 (`http://www.ebi.ac.uk/goldman-srv/pandit/`; Whelan et al. (2003)). The first dataset consists of 76 Glyceraldehyde 3-phosphate dehydrogenase sequences with an alignment length of 633 bp (PF00044), the second of 105 sequences from the ATP synthase alpha/beta family (1821 bp, PF00006), and the last of 193 sequences with Calporin homology with an alignment of 465 bp (PF00307).

Since the true tree is usually not known for real datasets, the Robinson-Foulds distance cannot be used to measure the efficiency of algorithms. Therefore the likelihood value of the reconstructed trees is used to compare the methods.

Since Weighbor does not use likelihoods we only compare PHYML and PhyNav from the methods above. Note, that Weighbor already was outperformed in the simulation study (cf. 4.1). Additionally we wanted to use METAPIGA (Lemmon and Milinkovitch (2002)), another method for large datasets based on a genetic algorithm. Unfortunately the program crashed on all three datasets. Thus, only PHYML and PhyNav were used for comparison.

Table 2. Results from the biological datasets of 76 Glyceraldehyde 3-phosphate dehydrogenase sequences, of ATP synthase alpha/beta (105 seqs.), and of 193 Calporin homologs: (a) Log-likelihood values of the best reconstructed trees and (b) Runtimes of tree reconstruction consumed by the different methods. The PhyNav column presents the runtime of a single repetition.

(a) Log-likelihood values.

sequences	length	PHYML	PhyNav
76	633 bp	-32133	-32094
105	1821 bp	-88975	-88632
193	465 bp	-64919	-64794

(b) Runtimes.

sequences	length	PHYML	PhyNav		
			runtime	repetitions	(single repetition)
76	633 bp	40s	2529s	70	36s
105	1821 bp	117s	14413s	100	144s
193	465 bp	101s	22306s	200	116s

Results for the biological datasets

As explained above we use the likelihood values of the reconstructed trees to compare the efficiency of the two programs. According to the maximum likelihood framework (cf. for example Felsenstein (1981)) the tree with the higher likelihood value represents the more likely tree.

The log-likelihood values are given in Table 2(a). These results show that PhyNav always find a tree with a higher likelihood. The increase of the log likelihood ranged from 39 up to 343 units.

However, as Table 2(b) shows, the price to pay for better likelihood trees is an increase in computing time. Each single repetition in the algorithm has a time consumption comparable to the one run of PHYML.

5 Discussion and conclusion

We propose a new search strategy to optimize the objective function for large phylogenies. Starting from an initial tree the PhyNav method uses heuristics to reduce the number of sequences, to reconstruct scaffold trees, and to add again the remaining sequences. During these steps the constructed trees are optimized using fast NNI operations.

The suggested method produced better results on all dataset compared to Weighbor and PHYML. The tradeoff for better accuracy is of course the runtime. While Weighbor outperformed PHYML and PhyNav with respected to the runtime on the simulated datasets, PHYML is 7.5-fold faster than PhyNav. However, spending more time might be well acceptable, because the quality of the results increases.

On the biological datasets PhyNav showed much longer runtimes compared to PHYML. Nevertheless, the substantial increase of the likelihoods might well justify that this effort is worthwhile, since it is still far from the time consumptions demanded by classical ML methods like DNAML (Felsenstein (1993)).

The mechanism to add the remaining sequences of $\overline{S_k}$ to T_k cannot be expected to give the most accurate results. However, our way is simple and performs efficiently, especially since the NNI operations seem to well remove unfortunate placements during the construction of the full trees T. Additionally, it might be worth trying other algorithms like Important Quartet Puzzle (Vinh and von Haeseler (2004)) to add the remaing sequences.

PhyNav can be applied to large dataset. We analyzed an alignment of 1146 Ankyrin amino acid sequences (PF00023) downloaded from the PANDIT database version 12.0 (Whelan et al. (2003)). The PhyNav options were set to 1000 repetitions and $k = 3$ and the WAG model (Whelan and Goldman (2001)) was applied. PhyNav found a best tree with -74665 log likelihood and needed about 15 minutes per repitition. The whole computation took about 10 days.

Acknowledgments

We would like to acknowledge the use of supercomputing resources of the ZAM/NIC at Research Center Jülich.

References

BRUNO, W.J., SOCCI., N.D. and HALPERN, A.L. (2000): Weighted Neighbor Joining: A Likelihood Based-Approach to Distance-Based Phylogeny Reconstruction. *J. Mol. Evol., 17, 189–197.*

FELSENSTEIN, J. (1978): The number of evolutionary trees. *Syst. Zool., 27, 27–33.*

FELSENSTEIN, J. (1981): Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol., 17, 368–376.*

FELSENSTEIN, J. (1993): *PHYLIP (Phylogeny Inference Package) version 3.5c.* Department of Genetics, University of Washington, Seattle, distributed by the author.

FELSENSTEIN, J. (2004): *Inferring Phylogenies.* Sinauer Associates, Sunderland, Massachusetts.

FITCH, W.M. (1971): Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool., 20, 406–416.*

GASCUEL, O. (1997): BIONJ: An Improved Version of the NJ Algorithm Based on a Simple Model of Sequence Data. *Mol. Biol. Evol., 14, 685–695.*

GUINDON, S. and GASCUEL, O. (2003): A Simple, Fast and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Sys. Biol., 52, 696–704.*

KIMURA, M. (1980): A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences. *J. Mol. Evol., 16, 111–120.*

LEMMON, A.R. and MILINKOVITCH, M.C. (2002): The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. *Proc. Natl. Acad. Sci. USA, 99, 10516–10521.*

RAMBAUT, A. and GRASSLY, N.C. (1997): Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci., 13, 235–238.*

ROBINSON, D.R. and FOULDS, L.R. (1981): Comparison of phylogenetic trees. *Mathematical Biosciences, 53, 131–147.*

RZHETSKY, A. and NEI, M. (1993): Theoretical Foundation of the Minimum-Evolution Method of Phylogenetic Inference. *Mol. Biol. Evol., 10, 1073–1095.*

SAITOU, N. and NEI, M. (1987): The Neighbor–joining Method: A New Method for Reconstructing Phylogenetic Trees. *Mol. Biol. Evol., 4, 406–425.*

SCHMIDT, H.A., STRIMMER, K., VINGRON, M. and VON HAESELER, A. (2002): TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics, 18, 502–504.*

STRIMMER, K. and VON HAESELER, A. (1996): Quartet Puzzling: A Quartet Maximum–Likelihood Method for Reconstructing Tree Topologies. *Mol. Biol. Evol., 13, 964–969.*

SWOFFORD, D.L., OLSEN, G.J., WADDELL, P.J. and HILLIS, D.M. (1996): Phylogeny Reconstruction. In: D.M. Hillis, C. Moritz, and B.K. Mable (Eds.): *Molecular Systematics*, 2nd ed. Sinauer Associates, Sunderland, Massachusetts, 407–514.

VINH, L.S., and VON HAESELER, A.: IQPNNI: Moving fast through tree space and stopping in time. *Mol. Evol. Biol., 21, 1565–1571.*

WHELAN, S., DE BAKKER, P.I.W. and GOLDMAN, N. (2003): Pandit: a database of protein and associated nucleotide domains with inferred trees. *Bioinformatics, 19, 1556–1563.*

WHELAN, S. and GOLDMAN, N. (2001): A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum Likelihood Approach. *Mol. Biol. Evol., 18, 691–699.*

NewsRec, a Personal Recommendation System for News Websites

Christian Bomhardt and Wolfgang Gaul

Institut für Entscheidungstheorie und Unternehmensforschung,
Universität Karlsruhe (TH), 76128 Karlsruhe, Germany

Abstract. The behavior of how individuals select and read news depends on the underlying media for reproduction. Today, the use of news websites is increasing. Online readers usually have to click on abstracts or headlines in order to see full articles. This kind of selection of information is less pleasant than in traditional newspapers where glancing over the whole layout of double pages is possible. Personalization is a possible solution for this article selection problem. So far, most types of personalization are controlled by website owners. In our work, we discuss design aspects and empirical results of our personal recommendation system for news websites, which uses text classification techniques.

1 Introduction

The internet can be seen as an enabling technology for the development of innovative product ideas. File sharing services or online news papers are well-known examples. One handicap of web sites in general is the necessity to navigate through them by following hyperlinks - it is, e.g., not possible to quickly run over the pages of an online newspaper - in contrast to printed media. On the hyperlinked internet, information selection is therefore more "expensive" in terms of time required as every click results in a transfer delay. A common approach for this information selection problem is personalization. Mobasher et al. (1999) describe web personalization as "any action that makes the web experience of a user personalized to the user's taste". In Gaul et al. (2002) personalization is used as important feature for the characterization of recommender system output. Personalization methods can be categorized, e.g., as personalization by information filtering (e.g., on Lycos.com, registered users can choose their interests from given categories; on further visits, only information from the selected categories is displayed) and personalization by information supplementing (e.g., Amazon.com enhances the detail view pages of books with recommendations of additional/alternative books) where additional context specific information is provided. These personalization methods can use well structured input data from databases but are solely offered by website operators for their own websites. As not every website offers personalization, a limitation exists. If someone wishes personalized assistance independent of a special website, the help of browsing agents is an alternative. These agents can solely rely on webpages, as databases may

not exist or are not accessible by the public. As long as website management does not apply robot detection technologies (cp. Bomhardt et al. (2004)) in order to prevent robots from accessing their websites, browsing agents can be used.

Browsing agents are well-known from the literature (Middleton (2001)). One example is WebWatcher (Joachims et al. (1996)) which is a tour guide for the world wide web and assists users in browsing the www. It learns from the experiences of multiple users and given keywords using term frequency/inverse document frequency (TF-IDF) similarity measures. Letizia (Lieberman (1995)) is an agent that monitors user behavior. During idle times, Letizia autonomously and concurrently explores links available at the user's current position and tries to anticipate items of interest which are displayed upon request. Letizia uses a set of heuristics and TF similarities. While WebWatcher provides tours to many people and learns to become a specialist with respect to a particular web site, Personal WebWatcher (PWW) (Mladenić (2001)) accompanies a single user and considers her/his individual interests. It doesn't ask the user for any keywords or opinions about pages but instead solely records the addresses of pages requested by the user and highlights interesting hyperlinks. During a learning phase, the requested pages are analyzed and a model is built. For speed reasons (Mladenić (1999)), PWW uses the anchor text of a hyperlink and the text near by the hyperlink as input for prediction. The agent NewsWeeder (Lang (1995)) is specialized with respect to Usenet newsgroups. It is implemented as web-based newsgroup client which stores user ratings of articles, learns preferences, and builds personalized news collections. It combines content-based filtering and collaborative filtering while using TF-IDF similarities. WebMate (Chen and Sycara (1997)) is a personal agent for browsing and searching. WebMate automatically sorts documents into categories and tracks interesting ones per category, thus building a domain-specific knowledge base. A document is recommended if it is "close enough" to an interesting reference document of the detected document category. WebMate can spider a user-defined list or URLs to compile a personal newspaper or it can feed search engines with several top key words of the current profile and rate pages found. WebMate also uses TF-IDF similarities.

As we are not aware of any browsing agent that is specialized with respect to news websites, uses support vector machines as prediction method, is designed as a single user system, and is silently augmenting browsing experience, our implementation NewsRec (News Recommender) will be described in the following. In the next section requirements, system design, and implementation details of NewsRec are discussed. The used classification methods and their evaluation measures are described in section 3. Empirical results are presented in section 4. Our findings are summarized in section 5.

2 Requirements, system design, and implementation details

A personal recommendation system for news websites should be compatible with HTML-based news websites, should be usable with any browser, should contain a user-friendly interface that annotates hyperlinks with its recommendations instead of requiring explicit requests for recommendations, should contain domain-specific state-of-the-art prediction models, should be designed for single user application, and should not lead to noticeable delay during web browsing. NewsRec fulfills all mentioned aspects.

NewsRec is implemented as HTTP proxy server. All HTTP requests and responses pass through the proxy server which manages communication between web browser and internet. The interaction between NewsRec and the user (article labeling, requesting model updates) is realized via additional embedded HTML buttons. The user configurates desired hosts for which the recommendation engine should be used. If a webpage from such a website is requested, it is processed by NewsRec's recommendation engine. Otherwise, the request is forwarded to the internet. NewsRec's recommendation engine loads a requested webpage, searches for linked documents, requests and rates the linked documents, marks the links within the original webpage as interesting (+) or uninteresting (-), adds interaction buttons, and sends the page back to the web browser.

As news websites can contain many links sequentially requesting and rating them would be too slow. We addressed this speed problem by using a thread pool which issues many requests in parallel. This approach overcomes the time consuming summation of transfer delays and timeouts that occur if pages are requested sequentially. Another important implementation detail is the usage of a recommendation cache. It stores the rating of every examined webpage. As several webpages within one website can link to the same page, this reduces the number of pages that have to be investigated. Webpages are represented using the common bag-of-words approach. Here, memory saving data structures have to be taken into consideration in order to avoid time-consuming memory swapping. A dictionary maps between words and word IDs. These mappings are used frequently. As dictionaries can easily contain more than 50000 words, fast and efficient data structures are required. We selected a Ternary Tree (Bentley/Sedgewick (1997)) as dictionary and slightly modified it in order to meet our requirements. The thread pool together with the recommendation cache, memory saving data structures, and the fast dictonary structure lead to significant performance gains that enabled the consideration of the contents of linked webpages. It should be noted that the recommendation cache has to be cleared, if the prediction model is updated.

3 Website classification and evaluation measures

A webpage is written in HTML and consists of common text, surrounded by HTML tags. The layout of a usual news webpage contains fixed elements like logos, advertisements, components for navigation, and the text of the news article. Our basic idea is to transform a webpage into common text, on which well known text classification algorithms can be applied. We extract the relevant text - that is the part of the webpage that contains the article text - remove HTML tags like <B> and substitute or remove HTML entities like *ü &bpsp;*. Now, we have raw text for which appropriate transformations are required. The following notation is used:

n	number of documents
d_i	document i in text representation, $i = 1...n$
m	number of distinct words contained in all documents d_i $(i = 1...n)$
w_j	unique word j, $j = 1...m$
$TF(w_j, d_i)$	number of occurrences of word w_j in document d_i (term frequency)
$BIN(w_j, d_i) =$	1, if word w_j is contained in document d_i, 0 otherwise (binary)
IDF_j	$= log(\frac{n}{\sum_{i=1}^{n} BIN(w_j, d_i)})$ (inverse document frequency)
R_j	$= log(n) + \sum_{i=1}^{n} \frac{TF(w_j,d_i)}{\sum_{g=1}^{n} TF(w_j,d_g)} log(\frac{TF(w_j,d_i)}{\sum_{g=1}^{n} TF(w_j,d_g)})$ (redundancy)

A document d_i can be represented as document vector $\overrightarrow{d}_i = (d_{ij})$, where each component d_{ij} of $\overrightarrow{d}_i$ either contains $TF(w_j, d_i)$ (TF-notation), or $log(1 + TF(w_j, d_i))$ (LOG-notation), or $BIN(w_j, d_i)$ (BIN-notation). This first step is called frequency transformation. Term weighting is the next step, where each d_{ij} of the document vector is multiplied by a weight factor. This factor can be 1 (NOWEIGHTS-notation), IDF_j (IDF-notation) or R_j (RED-notation). The last step comprises the normalization of the document vector $\overrightarrow{d}_i$. It can be skipped (NONE-notation), or $\frac{1}{\sum_{j=1..m} d_{ij}}$ (L1-notation), or $\frac{1}{\sqrt{\sum_{j=1..m} d_{ij}^2}}$ (L2-notation) can be used. A selection of one frequency transformation, one term weighting, and one normalization scheme describes a preprocessing setting. We have not implemented frequency transformation via BIN as it has lead to poor results in the experiments performed by Cooley (1999). Weighting via RED is expensive in terms of resource usage and, according to Paaß et al. (2004), the advantage of redundancy weighting via IDF seems to be greater for larger documents, thus, we have not included the RED weighting scheme. We selected support vector machines (SVM) (Boser et al. (1992)) for prediction, as different researchers have found out that SVMs are well suited for text classification and outperform other methods

like naive bayes classifiers, C4.5, etc. (Joachims (2002), Dumais et al. (1998), Cooley (1999), Sebastiani (2002)).

Recall, precision, and the measure $F1$ ($F1 = \frac{2*recall*precision}{recall+precision}$) were selected as common evaluation measures from information retrieval. Recall is defined as the number of correctly predicted interesting documents divided by the total number of interesting documents. Precision is defined as the number of correctly predicted interesting documents divided by the total number of predicted interesting documents. Good recall values can be easily achieved in expense of poor precision values (think of predicting all documents as interesting) and vice versa. This is why the $F1$-metric is often used in practice. It is a balanced measure and is dominated by the smaller of the recall and precision values.

4 Empirical results

NewsRec was tested on the Heise news ticker (HEISE), which is maintained by the German computer magazine c't. As a first step, one of the authors used the news website during a period of 7 weeks and read and labeled 1265 articles. To avoid self fulfilling prophecies, the recommendation engine was deactivated during this time. 27% of the articles were indicated as interesting by personal inspection. The next step was the simulation and evaluation of a real-world scenario. We assume that a common user labels a certain number of articles and is then interested in the valuation of the next upcoming articles (e.g., based on achieved recall and precision values). Thus, we trained successive prediction models. The first model was trained on the first 50 documents and evaluated on the next 50 documents. The number of training documents was increased by the just evaluated 50 documents for every new model as long as there remained 50 unevaluated documents for the next application step. The achieved recall and precision values were micro-averaged in order to receive an overall prediction measure. This evaluation procedure was repeated for each preprocessing setting.

For our experiments, we fell back on the SVM implementation SVMlight by Joachims (1999) and used the linear kernel. We are aware of the fact that slightly better models based on the radial basis function (rbf) kernel (cp. Paaß et al. (2004) and Joachims (2002)) may exist, but rbf models require extensive fine tuning of model parameters. For an automatic system like NewsRec, the linear kernel turned out to be a good choice as it is less sensitive to parameter selection than the rbf kernel. Another advantage of the linear kernel is its speed.

Table 1 summarizes our findings. TF-IDF-L2 was the optimal preprocessing setting in terms of $F1$ and recall. Users which prefer high precision could select TF-IDF-NONE.

Going into more detail, table 2 contains the detailed recall and precision values for the best three preprocessing settings in terms of $F1$. Here, one can

see that model selection is not trivial, e.g., TF-NOWEIGHTS-L2 and TF-IDF-L2 achieve the same recall values, if trained on the first 50 documents. If trained on the first 100 documents, TF-NOWEIGHTS-L2 outperforms TF-IDF-L2 in terms of recall. Taking a look at the next application step with 150 training documents, TF-NOWEIGHTS-L2, LOG-IDF-L2, and TF-IDF-L2 achieve the same recall, but TF-NOWEIGHTS-L2 is outperformed in terms of precision by the two other preprocessing settings.

Another important aspect is the fact that recall and precision values are not constantly increasing but are oscillating up and down, instead. This is a result of the fact that articles on new subjects may come up. Notice, e.g., that the models built on 550 training documents altogether perform very poor. TF-NOWEIGHTS-L2 did not valuate any document as interesting (although nine where contained within the evaluation set) and therefore achieved 0% recall and 100% precision, because no uninteresting document was labeled interesting. LOG-IDF-L2 and TF-IDF-L2 valuated some documents as interesting which indeed were uninteresting which lead to 0% recall and 0% precision. For the 550 training documents, the best prediction quality was achieved with the LOG-NOWEIGHTS-NONE preprocessing setting (not contained in the table): 11,1% recall, 16,6% precision and 0.13 for $F1$. On the other hand, there exist models that achieve 100% recall (LOG-IDF-L2, 650 training documents) or 100% precision (TF-NOWEIGHTS-L2, 1000 training documents). Here, one can see that model selection on the basis of table 2 is a difficult task. Thus, we recommend TF-IDF-L2 according to the overall performance mentioned in table 1.

5 Conclusions and outlook

NewsRec is easy to use and enriches conventional browsing by augmented browsing with real add-on value. The results of similar tools - if reported - cannot be compared with the ones of NewsRec, as the approaches vary and so do the datasets. Nevertheless, we report results concerning WebMate and NewsWeeder to mediate a feeling for what can be expected. WebMate achieved 31% overall precision and NewsWeeder achieved 44% (dataset a) and 59% (dataset b) precision for the highest rated positive 10% of articles. Recall was not reported for these agents. If we use these precision values as benchmarks, then NewsRec - which achieved 49.2% overall recall and 55% overall precision - seems to compete favorable. Our results fall within a reasonable range and were confirmed on another dataset. One problem in the area just mentioned is the requirement of explicit user feedback. Other researchers therefore use implicit feedback (Mladenić (1999)). We think that implicit feedback alone is a weak indicator. Instead, we will address this problem in a forthcoming paper by using a hybrid (implicit and explicit) feedback approach.

Table 1. Micro-averaged prediction quality for different preprocessing settings. Values in parenthesis denote the best results for a single set of evaluation documents

Preprocessing setting	Overall Recall	Overall Precision	Overall $F1$
TF-NOWEIGHTS-NONE	45.01% (80%)	50.01% (85.71%)	0.4738 (0.70)
TF-NOWEIGHTS-L1	12.86% (100%)	40.00% (54%)	0.1946 (0.70)
TF-NOWEIGHTS-L2	48.55% (81%)	49.18% (100%)	0.4886(0.74)
LOG-NOWEIGHTS-NONE	45.98% (80%)	44.96% (82%)	0.4546 (0.67)
LOG-NOWEIGHTS-L1	14.47% (100%)	43.27% (100%)	0.2169 (0.71)
LOG-NOWEIGHTS-L2	47.27% (82%)	48.68% (100%)	0.4796 (0.78)
LOG-IDF-NONE	37.94% (100%)	55.14% (100%)	0.4495 (0.71)
LOG-IDF-L1	21.22% (100%)	50.77% (100%)	0.2993 (0.70)
LOG-IDF-L2	44.69% (100%)	52.85% (100%)	0.4843 (0.71)
TF-IDF-NONE	37.62% (80%)	60.62% (86%)	0.4643 (0.80)
TF-IDF-L1	12.86% (100%)	40.00% (54%)	0.1946 (0.70)
TF-IDF-L2	49.20% (82%)	55.04% (89%)	0.5196 (0.73)

Table 2. Detailed recall and precision values for the best three preprocessing settings

Number of training documents	TF-NOWEIGHTS-L2			LOG-IDF-L2			TF-IDF-L2		
	Recall	Precision	$F1$	Recall	Precision	$F1$	Recall	Precision	$F1$
50	81.48%	66.67%	0.73	55.56%	83.33%	0.67	81.48%	64.71%	0.72
100	61.54%	24.24%	0.35	46.15%	35.29%	0.40	53.85%	28.00%	0.37
150	46.15%	33.33%	0.39	46.15%	46.15%	0.46	46.15%	46.15%	0.46
200	56.25%	45.00%	0.50	37.50%	60.00%	0.46	56.25%	60.00%	0.58
250	64.29%	39.13%	0.49	42.86%	37.50%	0.40	57.14%	40.00%	0.47
300	62.50%	55.56%	0.59	37.50%	60.00%	0.46	56.25%	60.00%	0.58
350	18.18%	20.00%	0.19	9.09%	16.67%	0.12	9.09%	20.00%	0.12
400	50.00%	69.23%	0.58	55.56%	55.56%	0.56	66.67%	66.67%	0.67
450	57.14%	47.06%	0.52	57.14%	53.33%	0.55	57.14%	61.54%	0.59
500	75.00%	21.43%	0.33	25.00%	10.00%	0.14	50.00%	22.22%	0.31
550	0.00%	100.00%	0.00	0.00%	0.00%	-	0.00%	0.00%	-
600	42.86%	23.08%	0.30	42.86%	27.27%	0.33	42.86%	37.50%	0.40
650	80.00%	66.67%	0.73	100.00%	55.56%	0.71	80.00%	66.67%	0.73
700	53.85%	77.78%	0.64	46.15%	60.00%	0.52	46.15%	75.00%	0.57
750	55.56%	45.45%	0.50	66.67%	50.00%	0.57	44.44%	44.44%	0.44
800	15.38%	33.33%	0.21	23.08%	42.86%	0.30	7.69%	25.00%	0.12
850	8.33%	50.00%	0.14	41.67%	83.33%	0.56	41.67%	83.33%	0.56
900	25.00%	40.00%	0.31	25.00%	40.00%	0.31	25.00%	40.00%	0.31
950	35.71%	71.43%	0.48	35.71%	71.43%	0.48	35.71%	83.33%	0.50
1000	53.33%	100.00%	0.70	46.67%	77.78%	0.58	53.33%	88.89%	0.67
1050	23.81%	62.50%	0.34	38.10%	100.00%	0.55	38.10%	80.00%	0.52
1100	52.63%	83.33%	0.65	57.89%	73.33%	0.65	52.63%	76.92%	0.63
1150	55.56%	55.56%	0.56	55.56%	45.45%	0.50	44.44%	36.36%	0.40
1200	72.73%	66.67%	0.70	72.73%	53.33%	0.62	81.82%	64.29%	0.72

References

BENTLEY, J. and SEDGEWICK, R. (1997): Fast Algorithms for Sorting and Searching Strings.
http://www.cs.princeton.edu/~rs/strings/paper.pdf

BOMHARDT, C., GAUL, W. and SCHMIDT-THIEME, L. (2004): Web Robot Detection - Preprocessing Web Logfiles for Robot Detection, to appear.

BOSER, B., GUYON, I. and VAPNIK, V. (1992): A Training Algorithm for Optimal Margin Classifiers.
http://citeseer.nj.nec.com/boser92training.html

CHEN, L. and SYCARA, K. (1997): WebMate: A Personal Agent for Browsing and Searching. *http://citeseer.nj.nec.com/chen98webmate.html*

COOLEY, R. (1999): Classification of News Stories Using Support Vector Machines.
http://citeseer.nj.nec.com/cooley99classification.html

DUMAIS, S., PLAT, J., HECKERMAN, D. and SAHAMI, M. (1998): Inductive Learning Algorithms and Representation for Text Categorization.
http://robotics.stanford.edu/users/sahami/papers-dir/cikm98.pdf

GAUL, W., GEYER-SCHULZ, A., HAHSLER, M. and SCHMIDT-THIEME, L. (2002): eMarketting mittels Recommendersystemen. *MARKETING ZFP, 24. Jg. Spezialausgabe "E-Marketing" 2002, 47–55.*

HEISE: Heise News Ticker, `http://www.heise.de/ct`

JOACHIMS, T. (1999): Making Large-Scale SVM Learning Practical. Advances in Kernel Methods. In: B. Schölkopf, C. Burges and A. Smola (Ed.): *Support Vector Learning*, MIT-Press.

JOACHIMS, T. (2002): Learning to Classify Text Using Support Vector Machines. *Kluwer Academic Publishers*

JOACHIMS, T., FREITAG, D. and MITCHELL, T. (1996): WebWatcher: A Tour Guide for the World Wide Web.
http://citeseer.nj.nec.com/joachims96webwatcher.html

LANG, K. (1995): NewsWeeder: Learning to Filter Netnews.
http://citeseer.nj.nec.com/lang95newsweeder.html

LIEBERMAN, H. (1995): Letizia: An Agent That Assists Web Browsing.
http://lieber.www.media.mit.edu/people/lieber/Lieberary/Letizia/Letizia-AAAI/Letizia.ps

MIDDLETON, S. (2001): Interface Agents: A Review of the Field. *Technical Report Number: ECSTR-IAM01-001.*
http://www.ecs.soton.ac.uk/~sem99r/agent-survey.pdf

MLADENIĆ, D. (1999): *Machine Learning Used by Personal WebWatcher.*
`http://www-2.cs.cmu.edu/afs/cs/project/theo-4/text-learning/www/pww/papers/PWW/pwwACAI99.ps`

MLADENIĆ, D. (2001): Using Text Learning to Help Web Browsing.
http://www-ai.ijs.si/DunjaMladenic/papers/PWW/hci01Final.ps.gz

MOBASHER, B., COOLEY, R. and SRIVASTAVA, J. (1999): Automatic Personalization Based on Web Usage Mining.
http://citeseer.nj.nec.com/mobasher99automatic.html

PAAß, G., KINDERMANN, J. and LEOPOLD, E. (2004): Text Classification of News Articles with Support Vector Machines. In: S. Sirmakessis (Ed.): *Text Mining and its Applications*. Springer, Berlin, 53–64.

SEBASTIANI, F. (2002): Machine Learning in Automated Text Categorization. *ACM Computing Surveys, 34(1), 1–47.*

Clustering of Large Document Sets with Restricted Random Walks on Usage Histories

Markus Franke and Anke Thede

Institut für Informationswirtschaft und -management
Universität Karlsruhe (TH), 76128 Karlsruhe, Germany

Abstract. Due to their time complexity, conventional clustering methods often cannot cope with large data sets like bibliographic data in a scientific library. We will present a method for clustering library documents according to usage histories that is based on the exploration of object sets using restricted random walks.

We will show that, given the particularities of the data, the time complexity of the algorithm is linear. For our application, the algorithm has proven to work well with more than one million objects, from the point of view of efficiency as well as with respect to cluster quality.

1 Motivation

15 million documents – how to find similar ones? This is the question when designing an automated indexing system in a scientific library like the Universitätsbibliothek at Karlsruhe. Another question should be answered first: How exactly can similarities between documents be measured?

At Karlsruhe, a system for giving recommendations on books is operative (Geyer-Schulz et al. (2003)) that is based on applying Ehrenberg's (1988) repeat-buying theory to the purchase histories of documents. These purchase histories describe which combinations of books have been inspected by users of the OPAC web interface in one session. The documents accessed during a short time can be expected to have a strong coherence, a fact that is also used in market basket analysis.

Our motivation was to use the information inherent in these data sets in order to construct an automated complement for the tedious, error-prone and expensive task of manual indexing. However, normal clustering algorithms (cf. e.g. Bock (1974)) imply a superlinear complexity, leading to an explosion of the computation time when applied to data of such dimensions as we deal with here. The failure of clustering with single linkage algorithms has been described by Viegener (1997). Based on the ideas of Schöll and Paschinger (2002), we have developed a clustering method based on random walks on purchase history similarity graphs that allows to efficiently cluster document sets of this size (Franke (2003)).

In this paper, we will first describe the purchase histories serving as input data, then develop the ideas of the algorithm before addressing time complexity. Some results from the test runs, followed by an outlook on further research will conclude the text.

2 Clustering with purchase histories

We cluster the documents on the basis of similarity graphs. As similarity measure, we take usage histories: A purchase occasion is a user's request for a document detail page in the library's web interface. If the same user, in the same session, requests another document detail page, this is a cross-occurrence between the two documents. The results depend on the search behavior of the users, not on hypertext links between documents. As most users search by title, even new documents are quickly integrated. The input data for the clustering was derived from the log file of the library's WWW server and organized into "raw baskets" prior to the algorithm's execution. The raw basket of a document contains the documents it has been viewed with as well as the respective frequencies of these cross-occurrences.

Let V be the set of all documents whose raw basket contains at least one entry. From the contents of all raw baskets, we derive a symmetric and non-negative similarity measure $s(i, j)$ defined as the number of cross-occurrences between documents i and j. The self-similarity of an object can be defined arbitrarily, it is never used. $s(i, j)$ induces a finite weighted graph $G = (V, E, \omega)$ formed by the documents as vertices V and by edges E between documents with positive similarity, weighted with the similarity between their end points:

$$G = (V, E, \omega)\text{: } E = \{(i, j) \in V \times V | s(i, j) > 0, i \neq j\},\ \omega_{ij} = s(i, j) \qquad (1)$$

The basic idea of clustering with restricted random walks (RRW) is to execute a series of random walks on the graph such that, with growing walk length, the similarity between consecutively chosen nodes increases. Consequently, the later a pair of nodes is chosen in a walk, the more their content is related and the higher the probability that they belong to one cluster.

In the following algorithm, all random choices are based on a uniform probability distribution. A random walk on the set $V = \{1, \ldots n\}$ is a series $R = (i_0, i_1, \ldots)$, where $i_k \in V$ for $k \in \mathbb{N}_0$. We start by randomly choosing a starting node i_0 from the set V of all documents. For the first step, a successor i_1 is chosen at random from the document set

$$T_0 = \{j \in V | s(i_0, j) > 0, i_0 \neq j\} = \{j \in V | (i_0, j) \in E\} \qquad (2)$$

that have been viewed at least once together with i_0. In other words, we select one of the neighbors of i_0. For the m-th step, $m \geq 1$, the similarity $s_m := s(i_{m-1}, i_m)$ between the documents participating in this step is called the step length. For all further steps, we add the following restriction: The next object, i_2, is picked from those neighbors of i_1 having a higher similarity to i_1 than i_0 does. Formally, the $m + 1$-st object is picked from the set

$$T_m = \{j \in V | s(i_m, j) > s_m\} \qquad (3)$$

of nodes that can be reached via an edge with a greater weight than the preceding one. The s_m are then updated according to

$$s_{m+1} = s(i_m, i_{m+1}) \qquad (4)$$

Thus, s_m grows strictly monotonic until T_m is empty, i.e. a node is reached that has no neighbor j with $s(i_m, j) > s_m$. The resulting series is called a restricted random walk.

Since restricted random walks have an expected length of $O(\log n)$, we need to execute more than one walk in order to completely cover the document set. Random graph theory (Erdös and Renyi (1957)) suggests that, given n vertices, a number of $\frac{1}{2}n \log n + 10n$ randomly selected edges (corresponding to $O(\frac{1}{2}n + \frac{10n}{\log n})$ walks) is sufficient to obtain a connected graph and thus the necessary information for clustering with a probability of 99.995%. As a more simple solution, Schöll and Paschinger (2002) propose to start a walk from each of the documents. In this work, we followed their approach. As to the exact number of walks that need to be started to detect clusters with a given probability, further research will be undertaken.

In Schöll and Paschinger's formulation given above, the transition probabilities depend on the two last nodes, consequently, the RRW is not a Markov chain. In order to restore the Markov property, we formulate a Markov chain of order 2 on the set E of edges of the similarity graph: The set of states is now defined as $S = E \cup \{\Omega\}$ where E is the set of all possible steps from one object to another as defined in (1) and Ω is the "empty" or start/end state that we use to concatenate the single walks into an irreducible Markov chain. In this formulation, a restricted random walk R has the form $R = (\Omega, (i_0, i_1), (i_1, i_2), \ldots, (i_{f-1}, i_f), \Omega)$. The set of possible direct successors for a step (i, j) is now defined as

$$T_{ij} = \{(j, k) \in E | s(j, k) > s(i, j)\} \tag{5}$$

From this, we can derive the transition probabilities as follows:

Suppose that the Markov chain is in the empty state, Ω. Remember that, for the first step, we choose an object i_0 as a starting point with equal probability $\frac{1}{|V|}$ from the set V. We then select a second object from the set $T_0 = \{j \in V | s(i_0, j) > 0\} \backslash \{i_0\}$ (Equation (2)) of objects having a positive similarity $s(i_0, j)$ to i_0 with probability

$$\frac{1}{|\{j \in V | s(i_0, j) > 0\} \backslash \{i_0\}|} = \frac{1}{\deg(i_0)} \tag{6}$$

where $\deg(i_0)$ is the degree of node i_0, thus the number of edges incident to it. In the alternative formulation, we pick a step from $T_\Omega = E$. Thus, the probability of a step (i, j) being chosen after the initial state is

$$P((i, j)|\Omega) = \begin{cases} \frac{1}{|V| \deg(i)} & \text{if } (i, j) \in E \\ 0 & \text{else} \end{cases} \tag{7}$$

Analogously, the probability of a $(k, l) \in E \cup \{\Omega\}$ being chosen after (i, j) is

$$P((k, l)|(i, j)) = \begin{cases} \frac{1}{|T_{ij}|} & \text{if } (k, l) \in T_{ij} \\ 1 & \text{if } T_{ij} = \emptyset \text{ and } (k, l) = \Omega \\ 0 & \text{else} \end{cases} \tag{8}$$

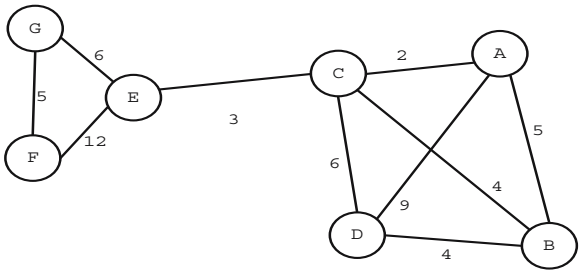


Fig. 1. An example similarity graph

The first condition models the uniformly distributed choice of the successor object from the set T_{ij}. The second part assures the return to the empty state Ω after the end of a walk. Other transitions are not possible.

Consider the example similarity graph in Figure 1 on which a walk is to be executed. From the set $T_\Omega = E = \{AB, AC, \ldots, EF, EG\}$ of possible steps, we pick one at random, AC. According to the formula (7), the probability of this event is $\frac{1}{|V|\deg(A)} = \frac{1}{7*3} = \frac{1}{21}$. We set $T_{AC} = \{CB, CD, CE\}$. We then pick CB, according to (8); the probability of this event is $\frac{1}{|T_{AC}|} = \frac{1}{3}$. We set $T_{CB} = \{BA\}$. Since only one step is left, we take it and set $T_{BA} = \{AD\}$. Once more, we take the only step left, AD. T_{AD} is empty because there is no neighbor node that is more similar to D than A. Consequently, the Markov chain enters Ω and the walk $ACBAD$ ends. Starting from other nodes, we get the walks, say BDC, CEF, $DBAD$, $ECBAD$, $FGEF$, and GFE.

The later a pair of objects appears in a walk, the higher is the significance of this event. On the other hand, walk lengths can differ considerably from each other like the ones in the example. In order to remove the influence of different walk lengths on the interpretation of the walk, we define the ratio $\frac{step\,number}{walk\,length}$ as the level of a step. Its purpose is to denote the relative position of a step in a walk. Thereby, it facilitates the comparison of the importance of steps coming from walks of differing lengths. For example, the step $(i_0, i_1) = AC$ in the first example walk has a level of $\frac{1}{4}$, since the walk comprises four steps. The higher the level of a step, the more meaning we attach to the occurrence of the nodes in the step.

For the cluster construction, three different approaches have been used in this article all of whom return a hierarchical clustering whose hierarchy cutoff levels are denoted by the variable l:

1. The original method by Schöll and Paschinger (2002) defined a cluster for the cutoff l as follows: For each step k, define a graph $G_k = (V, E_k)$ where V is the set of objects and E_k contains an edge for each pair of objects (i, j) iff i and j have been chosen in the k-th step of any walk, independent of the order in which they were visited. We then construct the union $H_l = \cup_{k=l}^{\infty} G_k$. Consequently, if two objects have been chosen in any step greater or equal l, they are neighbors in H_l. Clusters are defined

as components (connected subgraphs) of H_l. This means that two objects belong to the same cluster iff there exists a path in H_l between them.
In our example, this leads to edge sets $E_1 = \{AC, BD, CE, FG\}$, $E_2 = \{AB, BC, CD, EF, EG\}$, $E_3 = \{AB, AD, EF\}$, $E_4 = \{AD\}$ with alphabetically ordered edges. The graph H_3 has the edges $\{AB, AD, EF\}$ and thus the clusters $\{A, B, D\}, \{E, F\}$ and singletons $\{C\}, \{G\}$.
This approach leads to large clusters. For our purpose, this is not adequate: The algorithm produces clusters with tens of thousands of documents per cluster even on the highest level, which disqualifies it for the use in our specific scenario.

2. An alternative is the concept of the walk context we developed. The original method displays a sort of chaining effect, it connects – in our case – documents that have nothing in common via so-called bridge elements: Imagine a book about statistics and sociology. Its raw basket will contain some books that contain solely sociological matters and some others that only deal with statistics. A walk starting in the statistics region could end – via this bridge element – in the sociology section. This effect is much weaker than with single linkage clustering as described by Viegener (1997), but it is still visible. In order to limit the scope of this chaining, it is necessary to prune the search for components in the graph H_l. In the process of cluster identification, we put more emphasis on the context given by the course of the restricted random walks.
In order to find a cluster for a given object i at a given level l, we consider all walks where i participated in a step at a level greater than or equal l. The cluster consists of all objects that also appeared in those walks and had a step level greater or equal l.
In the example, a cluster for node D at level 1 is constructed from the walks $ABC\mathbf{AD}$, $B\mathbf{DC}$, $DB\mathbf{AD}$ and $ECB\mathbf{AD}$ (edges at level 1 in bold) with D at level 1. A and C also occur in these walks at level 1, so the cluster is the set $\{A, C, D\}$. Of course, the clustering resulting from walk contexts is no longer disjunctive since clusters partially overlap, but for our application this can even be desirable: Consider, once again, the book about statistics and sociology. In the cluster generated by this book, both books on statistics and sociology should be included. On the other hand, we do not want sociology books in a statistics cluster. This is the intuitive motivation for the use of the walk context: Consecutive steps in one walk have a lower risk of producing a chaining effect than whole components of the graph H_l and for our scenario, non-disjunctive clusters are desirable.
3. The bounded iterations variant is a compromise between the first two. We introduce another parameter, the number of iterations t, that determines whether this variant's results are closer to the walk context ($t = 0$) or the component clusters ($t = \infty$). For a given document i and a level l, we start by taking the walk context of i at level l and then, for t iterations, iteratively add all documents that are in the walk context – again at level l – of any of the documents already in the cluster. In our example,

we take the walk context cluster for D at level 1. We then add the walk contexts of A and C at level 1, that are $\{A, D\}$ and $\{C, D\}$, so the cluster is $\{A, C, D\}$.

3 Time complexity

A very attractive aspect of the RRW method lies in its low time complexity compared to the standard algorithms. Schöll and Paschinger (2002) give an average length of one walk of $O(\log n)$ and thus a complexity of $O(n \log n)$ for an object set of size n where log_2 seems to be a suitable approximation. With our data, the expected walk length is nearly 24.

However, in the case of library purchase histories, the data display one particularity that reduces the complexity to $O(n)$. The similarity matrix implied by the usage histories is extremely sparse. Among the 1.2 million documents having a raw basket there is none that exceeds the number of 2700 neighbors. Typically, an object has between 1 and 50 neighbors. Deducing from the experiences of the last years, the maximum number of neighbors seems to be a constant that shall be denoted by c. With this, the size of a raw basket is also bounded by a constant. Consequently, a RRW on the data has an average length of $O(\log c) = O(1)$. Although this conjecture still needs to be proven by further observation of the evolution of our usage histories, it is supported for the moment by our test runs of the algorithm, yielding an average length of about four which is much less than the theoretically expected length.

On the whole, the complexity of the RRW algorithm on these specific data is as follows (remember that c is a constant):

- Parsing baskets and building the partial similarity matrix: $O(nc) = O(n)$.
- Executing n consecutive random walks: $O(n \log c) = O(n)$.
- Finding clusters: depending on the method chosen (components or walk context), this value differs. For the walk context, the complexity of finding a cluster for a given document is $O(\log n)$ provided an efficient indexing system is used on the data.

Together, this yields a complexity of $O(n)$ for the preliminary works with the construction of the initial data base and $O(\log n)$ for the cluster generation.

4 Results

For a measurement of the quality of the clusters returned by the variants we tested the results against an external criterion, the library's manual classification that classifies documents into an hierarchy of categories. Documents can be assigned to several categories, leading to a quasi-hierarchical classification.

As a quality function, we decided to take the precision measure: For a cluster containing a certain document, we counted the number of documents in the cluster that share at least one category in the manual classification and

```
Duran, Benjamin S. and Odell, Patrick L. Cluster Analysis - A Survey
Anderberg, Michael R. Cluster analysis for applications
Everitt, Brian. Cluster analysis
Mather, Paul M. Cluster analysis
Benzécri, J.P. L'analyse des données
```

Fig. 2. A sample cluster for Duran and Odell's book on Cluster Analysis

divide it by the total number of documents in the cluster. The precision measure is preferable over the simple matching coefficient if, like in our case, the a priori classification cannot clearly state which documents are not related. The quality of the manual classification system at Karlsruhe differs strongly between the topics, depending on the person who entered the classification.

Since our main goal was a complementary indexing system, we focused on the precision of the results rather than on a high recall, since for indexing, quality is more important than quantity. Furthermore, the manual classification only covers about 30% of the documents, so that recall could not be tested in a sensible way.

As discussed, clustering with restricted random walks returns a (quasi-) hierarchical clustering. Since we are interested in a set of clusters rather than a hierarchy for indexing it is necessary to first find an optimal cutoff level that minimizes the deviation from the target, the manual classification. This was done with a small training sample (10% of the documents) for all three variants discussed in section 2. We used the RRW algorithm to construct clusters at levels 0, 0.25, 0.5, 0.75, and 1, then refined the search in promising areas up to a grid of 0.02. Finally, we conducted a hypothesis test on the whole document set, using the best method at the optimal level.

Before we discuss the numerical results, consider figure 2, showing the cluster that contains the book of Duran and Odell (1974). The cluster was constructed using the walk context variant at a step level of 0.88. As we can see, the precision is very high, all documents that were returned have a high similarity to each other and to the book by Duran and Odell, but the cluster size is relatively small given the current amount of literature on clustering.

The walk context gave the best results with a precision of 61.1 per cent at level .88 for the training sample, producing clusters of an average size of 3.14. The method of Schöll and Paschinger only reached 0.284 at level 1, while the bounded iterations yield a precision of 0.474 at level 1 for two iterations.

A χ^2-test (40 classes, $\alpha = 0.05$) revealed that the precision in the sample was normally distributed. Thus, the results of the walk context method were verified using Wald's (1966) sequential probability ratio test with H_0: *The cluster precision is normally distributed with mean* $\Theta = .61$ and $\alpha = 0.05$, $\beta = 0.05$, $\Theta_0 = 0.56$, $\Theta_1 = 0.61$. The variance was estimated from the sample. The advantage of Wald's method is a much smaller sample size for the same confidence, compared to standard methods. The test confirmed a precision of .61 after having tested the clusters of 5489 documents out of over 450,000 for having a cluster at level 0.88. For further details refer to Franke (2003).

5 Outlook

There are some possible enhancements and directions for future research: Instead of picking a successor from a uniform distribution, the probability of an object being chosen could directly depend on its similarity to the current one. This would cause faster walk convergence towards very similar documents but reduce the discriminatory power of the steps since the walks would be shorter.

The level as criterion for the construction of the hypergraph could be replaced by the step number, counted from the end of each walk. By using this variant, steps with high levels of long walks are weighted stronger, but those with middle weight are mixed with the start steps from short walks.

There is still some work to be done to gain insights into the stability of RRW clusters. We conjecture that it is strongly dependent on the number of walks, but the subject of convergence of the solutions and the number of walks that are necessary to detect clusters with a certain probability have not been investigated further yet due to the considerable computational complexity.

As our usage histories evolve, we will observe the raw basket sizes in order to verify if they can be bounded by a constant as proposed in section 3.

Another interesting aspect of our method is the possibility of detecting bridge elements with restricted random walks that still needs some research.

Acknowledgement. We gratefully acknowledge the funding of the project "Scientific Libraries in Information Markets" by the Deutsche Forschungsgemeinschaft within the scope of the research initiative "V^3D^2". It was on the data generated during this project that we could execute our random walks.

References

BOCK, H.H. (1974): *Automatische Klassifikation.* Vandenhoeck & Ruprecht, Göttingen.

DURAN, B.S. and ODELL, P.L. (1974): *Cluster Analysis - A Survey.* Springer, Berlin, Heidelberg, New York.

EHRENBERG, A.S.C. (1988): *Repeat-Buying: Facts, Theory and Applications.* Charles Griffin & Company Ltd, London.

ERDÖS, P. and RENYI, A. (1957): On Random Graphs I. *Publ. Math., 6, 290–297.*

FRANKE, M. (2003): *Clustering of Very Large Document Sets Using Random Walks.* Master's Thesis, Universität Karlsruhe (TH).

GEYER-SCHULZ, A.,NEUMANN, A. and THEDE, A. (2003): Others also Use: A Robust Recommender System for Scientific Libraries. In: T. Koch and I.T. Solvberg (Eds.): *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003.* Springer, Berlin, 113–125.

SCHÖLL, J. and PASCHINGER, P. (2002): Cluster Analysis with Restricted Random Walks. In: K. Jajuga, A. Sokołowski and H.H. Bock (Eds.): *Classification, Clustering, and Data Analysis.* Springer, Berlin, 113–120.

VIEGENER, J. (1997): *Inkrementelle, domänenunabhängige Thesauruserstellung in dokumentbasierten Informationssystemen durch Kombination von Konstruktionsverfahren.* infix, Sankt Augustin.

WALD, A. (1966): *Sequential Analysis.* John Wiley & Sons, New York.

Fuzzy Two-mode Clustering vs. Collaborative Filtering

Volker Schlecht and Wolfgang Gaul

Institut für Entscheidungstheorie und Unternehmensforschung,
Universität Karlsruhe (TH), 76128 Karlsruhe, Germany

Abstract. When users rate interesting objects one often gets two-mode data with missing values as result. In the area of recommender systems (automated) collaborative filtering has been used to analyze such kind of two-mode data. Like collaborative filtering (fuzzy) two-mode clustering can be applied to handle so far unknown ratings of users concerning objects of interest. The aim of this paper is to suggest a new algorithm for (fuzzy) two-mode clustering and compare it to collaborative filtering.

1 Introduction

Two-mode data depict relations between so-called first and second mode elements. These relations can describe the quantity of product j (second mode element) which customer i (first mode element) buys, the percentage of buyers that purchase a brand at time t+1 (second mode element) given that they bought a certain brand at time t (first mode element), or a rating concerning item j which person i has provided. Researchers from different fields, e.g., marketing (Schader and Gaul (1991), Gaul and Schader (1996), Baier et al. (1997)), psychology (Eckes and Orlik (1993), Ceulemans and Van Mechelen (2002), Krolak-Schwerdt (2003)), and biology (Li and Zha (2002), Kluger et al. (2003), Pollard and Van der Laan (2002), Jörnsten and Yu (2003)) have focused their attention on the analysis of two-mode data.

In applications one is often faced with the problem that some relations between elements of different modes have not been reported or are not available. A well-known example describes ratings concerning movies, which persons have provided. As it is unlikely that a person can rate every movie from a given list and as different persons will rate different movies it is of interest to recommend to persons such movies which they do not know (i.e., have not rated) yet but would probably like to see.

Collaborative Filtering (e.g., Breese et al. (1998), Herlocker et al. (1999)) manages to handle missing values by first comparing the person, for whom one wants to provide recommendations, to other people with respect to their ratings. According to Shardanand and Maes (1995) the basic assumption of collaborative filtering is that the more similar persons are to the person, for whom one wants to provide recommendations, the better the estimates of the missing values should be.

While classical methods of cluster analysis (Hartigan (1975)) rely on relationships ((dis)similarities, distances) between elements of just one mode, two-mode clustering can take information within and between modes into account.

If two-mode clustering is modified in such a way, that it can handle missing values, ratings of persons for certain items on the basis of cluster-memberships of both persons and items can be determined.

The aim of this paper is to compare the performance of a new algorithm for fuzzy two-mode clustering to the performance of collaborative filtering.

2 Two-mode data analysis

2.1 Memory-based Collaborative Filtering (CF)

Both model-based and memory-based algorithms for collaborative filtering exist for the analysis of two-mode data. The memory-based algorithms for collaborative filtering always utilize the whole user database to generate suggestions, whereas the model-based algorithms for collaborative filtering perform a two-step procedure. In the first step these algorithms operate over the entire user database to learn a model from which recommendations/suggestions are derived in the second step.

Today, most memory-based collaborative filtering techniques apply the Bravais-Pearson correlation coefficient, since it can easily be computed and outperforms known alternatives, e.g., vector similarity and Spearman rank correlation approaches (see Breese et al. (1998), Herlocker et al. (1999), Sarwar et al. (2000)). The Bravais-Pearson correlation coefficient describes the extent to which two different users are correlated with each other with respect to items which both of them have rated.

Let index i denote some user and index j denote an item. Furthermore, let s_{ij} be a rating, which user i has given concerning item j. Since most people rate only part of the items, the matrix $S{=}(s_{ij})$ contains missing values. We introduce J_i as the set of items, which user i has rated. Then

$$\bar{s}_i = \frac{1}{|J_i|} \sum_{j \in J_i} s_{ij}$$

is the average of all ratings that user i has provided and the adjusted Bravais Pearson correlation coefficient can be calculated:

$$w(a,i) = \frac{\sum_{j \in J_a \cap J_i} (s_{aj} - \bar{s}_a)(s_{ij} - \bar{s}_i)}{\sqrt{\sum_{j \in J_a} (s_{aj} - \bar{s}_a)^2 \sum_{j \in J_i} (s_{ij} - \bar{s}_i)^2}}$$

Here, the index a denotes the active user for whom one wants to provide recommendations.

With I_j as set of users that have provided a rating for item j the Bravais-Pearson correlation coefficient can be used to determine a so far unknown rating of user a concerning item j:

$$\hat{s}_{aj} = \bar{s}_a + \frac{\sum_{i \in I_j} w(a,i)(s_{ij} - \bar{s}_i)}{\sum_{i \in I_j} |w(a,i)|}$$

2.2 (Fuzzy) Two-Mode Clustering (FTMC)

Two-mode clustering is another well-known approach for the analysis of two-mode data. There are different approaches to two-mode clustering (e.g., DeSarbo (1982), DeSarbo et al. (1988), Gaul and Schader (1996)). Here, first mode elements and second mode elements are clustered simultaneously. Every first mode cluster is in some characteristic way associated with the second mode clusters and vice versa. Hence, every first mode cluster can be interpreted by looking at the second mode clusters it is connected with and vice versa. The following notation is used:

$i \in \{1, ..., I\}$ $[j \in \{1, ..., J\}]$	index of the first [second] mode elements,
$k \in \{1, ..., K\}$ $[l \in \{1, ..., L\}]$	index of the first [second] mode clusters,
$S = (s_{ij})$ $[\hat{S} = (\hat{s}_{ij})]$	observed [estimated] two-mode data matrix,
$P = (p_{ik})$ $[Q = (q_{jl})]$	matrix which describes the cluster-membership of the first [second] mode elements with p_{ik} $[q_{jl}] = \begin{cases} 1, \text{ if i [j] belongs to first [second] mode cluster k [l],} \\ 0, \text{ otherwise,} \end{cases}$
$W = (w_{kl})$	matrix of weights.

Two-mode clustering algorithms try to find the best-fitting estimator $\hat{S}$ for the given two-mode data matrix S (Gaul and Schader (1996), Baier et al. (1997)). A simple way for determining this best-fitting estimator $\hat{S}$ is $\hat{S} = PWQ'$, where the matrices P, W, and Q have to be alternatingly determined by minimizing the objective function

$$Z = \sum_{i=1}^{I} \sum_{j \in J_i} (s_{ij} - \hat{s}_{ij})^2,$$

where

$$\hat{s}_{ij} = \sum_{k=1}^{K} \sum_{l=1}^{L} p_{ik} w_{kl} q_{jl}.$$

Both overlapping and non-overlapping versions of this kind of algorithm for two-mode clustering exist.

In fuzzy two-mode cluster analysis the zero-one cluster-membership indicators are replaced by

$p_{ik} \in [0,1],\ i \in \{1,...,I\},\ k \in \{1,...,K\}$
$\forall\ i \in \{1,...,I\} : \sum_{k=1}^{K} p_{ik} = 1 (\geq 1)$

where p_{ik} denotes the degree of cluster-membership and the inequality "$\geq$" takes overlapping explicitly into account. Here, we just give the formulae for (p_{ik}), since the restrictions for (q_{jl}) are of equivalent form.

Gaul and Schader (1996) already mentioned that one could produce a fuzzy two-mode classification if one imposes appropriate restrictions on the matrices P and Q and applies a penalty-approach to compute P and Q in the context of an iterative procedure. In this paper another alternative, the "Delta"-Method for fuzzy two-mode clustering, is introduced.

3 The Delta-Method for fuzzy two-mode clustering

As starting-solution the Delta-Method uses a non-fuzzy non-overlapping two-mode classification. Because we have to deal with missing values now, the known formulae for deriving a starting-classification have to be adjusted to the new situation.

Starting from results obtained by non-overlapping non-fuzzy two-mode clustering the Delta-Method is able to produce quite similar classifications. However, the larger the numbers K and L of first and second mode clusters are chosen, the more the computation-time increases.

In the case of fuzzy two-mode cluster analysis one alternatingly determines P, W, and Q by minimizing the objective function

$$Z = \sum_{i=1}^{I} \sum_{j \in J_i} (s_{ij} - \sum_{k=1}^{K} \sum_{l=1}^{L} p_{ik} w_{kl} q_{jl})^2$$

where Newton's algorithm is used for the determination of $W{=}(w_{kl})$. Since equations are symmetric in (p_{ik}) and (q_{jl}) the idea for optimizing P and Q is the same. The optimization of $P{=}(p_{ik})$ is given in Table 1.

The general idea is to take the results of the non-overlapping non-fuzzy two-mode clustering algorithm provided by Baier et al. (1997) as starting solution and search for fuzzy cluster-membership values which improve the objective function Z.

According to Table 1, for every i the following steps have to be performed: First, we set Z_i^{best} and Z_i^{old} equal to a large number M and calculate

$$Z_i^{new} = \sum_{j \in J_i} (s_{ij} - \sum_{k=1}^{K} \sum_{l=1}^{L} p_{ik} w_{kl} q_{jl})^2.$$

During the iterations, as long as Z_i^{new} is smaller than Z_i^{old}, we set $Z_i^{old} = Z_i^{new}$. We determine the first mode cluster with the highest membership value and denote the index of this cluster by $k_{MAX} = arg\{max_{k \in \{1,...,K\}}\{p_{ik}\}\}$.

Table 1. Delta-Method (Optimization of P)

Starting-solution: results of non-overlapping non-fuzzy two-mode clustering $((p_{ik}),(w_{kl}),(q_{jl}))$

$i = 1$;
While$(i \leq I)\{$
$Z_i^{best} = Z_i^{old} = M$; M large
$Z_i^{new} = \sum_{j \in J_i}(s_{ij} - \sum_{k=1}^{K}\sum_{l=1}^{L} p_{ik}w_{kl}q_{jl})^2$;
While$(Z_i^{new} < Z_i^{old})$ {
$Z_i^{old} = Z_i^{new}$;
$k_{MAX} = arg\{max_{k \in \{1,\ldots,K\}}\{p_{ik}\}\}$;
$p_{ik_{MAX}} = p_{ik_{MAX}} - \Delta$; $(0 < \Delta \leq 0.1)$
$\bar{k} = 1$;
While$(\bar{k} \leq K)\{$
If$(\bar{k} \neq k_{MAX})$ {
$p_{i\bar{k}} = p_{i\bar{k}} + \Delta$; (if possible)
$Z_i^{temp} = \sum_{j \in J_i}(s_{ij} - \sum_{k=1}^{K}\sum_{l=1}^{L} p_{ik}w_{kl}q_{jl})^2$
If$(Z_i^{temp} < Z_i^{best})$ $\{Z_i^{best} = Z_i^{temp}; k^+ = \bar{k};\}$
$p_{i\bar{k}} = p_{i\bar{k}} - \Delta$;
}
$\bar{k} = \bar{k} + 1$;
}
If$(Z_i^{best} < Z_i^{old})$ { $Z_i^{new} = Z_i^{best}; p_{ik^+} = p_{ik^+} + \Delta$; }
Else { $p_{ik_{MAX}} = p_{ik_{MAX}} + \Delta$; }
}
i=i+1;
}

We set $p_{ik_{MAX}} = p_{ik_{MAX}} - \Delta$ and check for every $\bar{k}$ (unequal to k_{MAX}) whether

$$Z_i^{temp} = \sum_{j \in J_i}(s_{ij} - \sum_{k=1}^{K}\sum_{l=1}^{L} p_{ik}w_{kl}q_{jl})^2$$

can become smaller than Z_i^{best} when we set $p_{i\bar{k}} = p_{i\bar{k}} + \Delta$. If yes, we set $Z_i^{best} = Z_i^{temp}$ and $k^+ = \bar{k}$. Before we proceed to the next first mode cluster we subtract Δ from $p_{i\bar{k}}$, again.

We follow this procedure for all $\bar{k}$ and, finally, set $Z_i^{new} = Z_i^{best}$ and $p_{ik^+} = p_{ik^+} + \Delta$, if Z_i^{best} is smaller than Z_i^{old}, otherwise we add Δ again to $p_{ik_{MAX}}$.

4 Examples and comparisons

Collaborative filtering and fuzzy two-mode clustering were applied to the MovieLens data (available from http://www.ecn.purdue.edu/KDDCUP) collected from roughly 50000 users, who rated about 1500 films. This two-mode

data matrix has many missing values as the average number of movies, which one user has rated, is 46, while the average number of users, who rated the same movie, is 762. After having excluded all users that rated less than 10 movies and all movies which were rated by less than 10 users the following three data subsets were selected:

Data set 1 consisted of 60 users who rated 140 movies. In data set 2 a (1000 users $\times$ 634 movies)-matrix was analyzed.

Data set 3 was chosen in order to compare the ability of fuzzy two-mode clustering and collaborative filtering to cope with time-dependent appearance of data. From the (1000 users $\times$ 634 movies)-matrix we selected 980 users $\times$ 634 movies as training subset. For the rest of 20 users we divided the ratings (the movies) into two parts, part $\psi_{t<t_Z}$ consisted of ratings provided before time t_Z, the remaining part $\psi_{t>t_Z}$ contained the ratings that were not known up to time t_Z, i.e. data from $\psi_{t>t_Z}$ were treated as missing values.

We used the average absolute deviation

$$T = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{|J_i|}\sum_{j\in J_i}|s_{ij}-\hat{s}_{ij}| \qquad \text{(the smaller the better)}$$

and the variance accounted for

$$VAF = 1 - \frac{\sum_{i=1}^{I}\sum_{j\in J_i}(s_{ij}-\hat{s}_{ij})^2}{\sum_{i=1}^{I}\sum_{j\in J_i}(s_{ij}-\bar{s}_{..})^2} \quad \text{(the larger the better)}$$

for comparisons.

Tables 2, 3, and 4 show the results. With respect to data set 1 (see Table 2), fuzzy two-mode clustering (FTMC) outperformed collaborative filtering (CF) for the shown numbers K, L of first and second mode clusters in terms of VAF and T. With respect to data set 2 different numbers of first and second mode clusters were used. For $K = L = 8$ Table 3 shows that CF was still better than FTMC. K=10 and L=15 were needed for FTMC to show better performance than CF. (The results for $K = L$=10, $K = L$=12, $K = L$=15 are just shown to provide a feeling how FTMC improves.) Given the results of Tables 2 and 3 one could argue that if FTMC is based on "'large enough"' numbers of first and second mode clusters it will beat CF.

Finally, Table 4 shows how CF and FTMC behave on data set 3. Since it is much harder to estimate unknown values than to fit known ones, both methods perform worse if $\psi_{t>t_Z}$ is not known beforehand. Still fuzzy two-mode clustering is able to outperform collaborative filtering, if we choose the numbers K, L of first and second mode clusters large enough.

Table 2. Performance of CF and the Delta-Algorithm for Fuzzy Two-Mode Clustering (FTMC) (Data set 1: 60 users, 140 movies)

	CF	FTMC				
		K=L=5	K=L=7	K=L=8	K=L=10	K=L=12
VAF	0.543	0.658	0.748	0.775	0.837	0.868
T	0.128	0.127	0.109	0.102	0.087	0.079

Table 3. Performance of CF and the Delta-Algorithm for Fuzzy Two-Mode Clustering (FTMC) (Data set 2: 1000 users, 634 movies)

	CF	FTMC				
		K=10, L=15	K=L=8	K=L=10	K=L=12	K=L=15
VAF	0.569	0.597	0.531	0.572	0.607	0.638
T	0.129	0.128	0.138	0.132	0.126	0.121

Table 4. Performance of CF and the Delta-Algorithm for Fuzzy Two-Mode Clustering (FTMC) (Data set 3: Here, we used a training-set of 980 users and estimated the values of $\psi_{t>t_Z}$ for a test-set of 20 users.)

	CF	FTMC K=L=12	FTMC K=20, L=15	FTMC K=22, L=16
VAF	0.239	0.241	0.311	0.322
T	0.264	0.263	0.226	0.217

5 Conclusions

It is possible to derive versions of two-mode clustering, which can deal very well with missing values. One of these versions is the Delta-Algorithm for fuzzy two-mode clustering, which we introduced in this paper. Our results show that fuzzy two-mode clustering is able to outperform collaborative filtering with respect to the MovieLens data. Analyses of additional data sets and further developments of two-mode clustering variants are under consideration.

References

BAIER, D., GAUL, W., and SCHADER, M. (1997): Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring. In: R. Klar and O. Opitz (Eds.): *Classification and Knowledge Organziation*, Springer, Berlin, 557–566.

BREESE, J.S., HECKERMAN, D., and KADIE, C. (1998): Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, 43–52.

CEULEMANS, E. and VAN MECHELEN, I.: Tucker2 Hierarchical Classes Analysis. *Psychometrika* (in press).

DESARBO, W. S. (1982): GENNCLUS: New Models for General Nonhierarchical Clustering Analysis. *Psychometrika, 47, 446–449.*

DESARBO, W. S., DESOETE, G., CARROLL, J. D., and RAMASWAMY, V. (1988): A New Stochastic Ultrametric Tree Methodology for Assessing Competitive Market Structure. *Applied Stochastic Models and Data Analysis, 4, 185–204.*

ECKES, T. and ORLIK, P. (1993): An Error Variance Approach to Two-Mode Clustering. *Journal of Classification, 10, 51–74.*

GAUL, W. and SCHADER, W. (1996): A New Algorithm for Two-Mode Clustering. In: H.H. Bock and W. Polasek (Eds.): *Data Analysis and Information Systems*, Springer, Berlin, 15–23.

HARTIGAN, J. (1975): *Clustering Algorithms*. Wiley, New York.

HERLOCKER, J.L., KONSTAN, J.A., BORCHERS, A. and RIEDL, J. (1999): An Algorithmic Framework for Performing Collaborative Filtering. *Proceedings of the 22th Annual International AMC SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, 230–237.

JÖRNSTEN, R. and YU, B. (2003): Simultaneous Gene Clustering and Subset Selection for Sample Classification via MDL. *Bioinformatics 2003, 19, 1100–1109.*

KLUGER, Y., BASRI, R., CHANG, J.T. and GERSTEIN, M. (2003): Spectral Biclustering of Microarry Data: Coclustering Genes and Conditions. *Genome Research, 13, 703–716.*

KROLAK-SCHWERDT, S. (2003): Two-Mode Clustering Methods: Compare and Contrast. In: M. Schader, W. Gaul and M. Vichi (Eds.): *Between Data Science and Applied Data Analysis*, Springer, Berlin, 270–278.

LI, J. and ZHA, H. (2002): Simultaneous Classification and Feature Clustering Using Discriminant Vector Quantization With Applications to Microarray Data Analysis. *IEEE Computer Society Bioinformatics Conference 2002*, 246–255.

POLLARD, K.S. and VAN DER LAAN, M.J. (2002): Statistical Inference for Simultaneous Clustering of Gene Expression Data. *Mathematical Biosciences, 176, 99–121.*

SARWAR, W., KARYPIS, G., KONSTAN, J. and RIEDL, J. (2000): Analysis of Recommendation Algorithms for E-Commerce. In: *Proceedings of the ACME-Commerce 2000 Conference*, 158–167.

SCHADER, M. and GAUL, W. (1991): Pyramidal Clustering with Missing Values. In: E. Diday and Y. Lechevallier (Eds.): *Symbolic-Numeric Data Analysis and Learning, Nova Science Publishers*, 523–534.

SHARDANAND, U. and MAES, P. (1995): Social Information Filtering: Algorithms for Automating "Word of Mouth". In: *Proceedings of the CHI'95 Conference on Human Factors in Computing Systems*, 210–217.

Web Mining and Online Visibility

Nadine Schmidt-Mänz and Wolfgang Gaul

Institut für Entscheidungstheorie und Unternehmensforschung,
Universität Karlsruhe (TH), 76128 Karlsruhe, Germany

Abstract. In order to attract web visitors via the internet online activities have to be "visible" in the net. Thus, visibility measurement of web sites and strategies how to optimize Online Visibility are important. Here, web mining helps to define benchmarks with respect to competition and allows to calculate visibility indices as predictors for site traffic.
We use information like keyword density, incoming links, and ranking positions in search engines to measure Online Visibility. We also mention physical and psychological drivers of Online Visibility and describe the appropriateness of different concepts for measurement issues.

1 Introduction – "Why measurement of online visibility?"

Search engines appear to be very important to reach new visitors of web sites, because nearly 80% of all internet users find new web sites with the aid of search engines (Fischerländer (2003)). Results by Johnson (2002) also show that more active online shoppers tend to search across more sites and that the amount of online search is actually quite limited when internet surfers already have a special portfolio of web sites.
Therefore, it is very important to observe what could be called Online Visibility of web sites (see Drèze and Zufryden (2003) who have defined Online Visibility as the extent of presence of a brand or a product in the consumer's environment, e.g. by means of links from other web sites, online directories and search engines).
We suggest a measure called GOVis (Gage of Online Visibility) to keep track of the Online Visibility of web sites and to measure the success (unsuccessfulness) of conducted web site optimization.
In section 2 we shortly describe the web as a graph and focus on facts about human online searching and surfing behavior. We explain our measure of Online Visibility and main drivers to influence this phenomenon in section 3 while conclusions and some managerial implications are given in section 4.

2 (Human) Online search in a changing webgraph

The structure of the web is often compared with a haystack in which one tries to search for and find the needle. If the web is modeled as a directed

graph, the addition of new vertices and edges and the omission of old ones cause changings of the graphical structure.
Researchers try to build subgraphs of the complete underlying web graph and develop models to interpret evolving views on this dynamically altering net.
But what do persons really see when they are searching the web? They only get sub-subwebgraphs corresponding to their search efforts and requests which describe just static parts of the underlying situation.
In order to understand (human) online searching and surfing behavior and to derive managerial implications for web site owners adequate measures with respect to the underlying phenomena are needed.

2.1 The web as a graph

A web site consists of pages connected in a certain way. Their link structure can be described by the associated site graph. The web consists of sites and hyperlinks between certain pages within the same site (site subwebgraph) and across different sites. The pages can be seen as vertices in a directed graph and the hyperlinks as directed edges. If one tracks the web, one gets from vertice to vertice by following the directed edges. In the end one has information based on the structure of the tracked subgraph represented by its adjacency matrix. Given this information it is possible to calculate measures to characterize this subwebgraph.
Here, some facts about this microscopic view on the web have to be mentioned (Barabasi and Albert (1999), Broder et al. (2000)):
The average distance (also referred to as diameter) as number of links to get from any page to any other is about 19, if a path exists. The distribution of

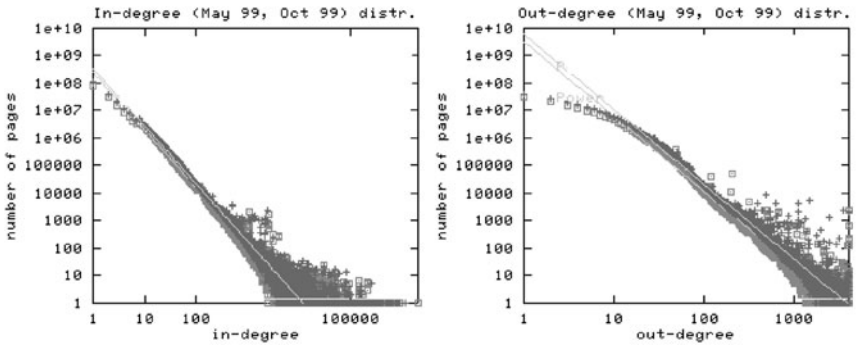


Fig. 1. Distribution of in- and outgoing links of pages

incoming and outgoing links of pages in the web is shown in figure 1 (Broder et al. (2000)) with the number of incoming and outgoing links on the x-axis

and the number of pages with corresponding in- and out-links on the y-axis, both depicted on logarithmic scales.We have used this shape to model the function $f(Z_L)$ in section 3.3.

2.2 (Human) Online searching and surfing behavior

Three main studies can be mentioned that report on online searching via web search engines by analyzing query logs: The Fireball, the Excite, and the AltaVista study (Hölscher (1998), Jansen et al. (2000), Silverstein et al. (1999)).
Conclusions of all three studies are nearly the same. The AltaVista study is based on the largest data set: one billion queries submitted to the main search engine over a 42-days period.
Facts about human online searching behavior corresponding to the AltaVista study can be summarized as follows: Nearly 77.6% of all query sessions consisted of only one request. 85.2% of the searchers examined only one result screen per query (7.5% two and 3.0% three screens). The average number of terms in a query adds up to 2.35 ($\sigma = 1.74$) and that of operators in a query to 0.41 ($\sigma = 1.11$). According to the total number of queries, 63.7% occurred only once. The most popular query was "sex" with an appearance of 1,551,477 times. This equals 2.7% of the total number of non-empty queries in the study.
People become also more efficient in using the web by navigating directly to

Table 1. Global Internet Usage (WebSideStory (2003))

Referral Type	**2002**	**2003**	trend
Direct Navigation	50.21%	65.48%	↗
Web Links	42.60%	21.04%	↘
Search Engines	07.18%	13.46%	↗

a web site they already know, see table 1. And they often use search engines to find new ones. Thus, search engines are effective instruments to reach new visitors or potential customers in the web business.

3 Measurement of Online Visibility

Online Visibility of a web site (or part of it) describes the extent to which it is recognizable or findable via normal searching strategies of web users.
Based on knowledge, e.g., about the link structure of the web as a graph, the functioning of ranking algorithms of search engines such as PageRank (Brin and Page (1998), Kleinberg (1999)), and human searching and surfing behavior, several impacts on Online Visibility can be defined.

3.1 Main drivers of Online Visibility

Online Visibility has to be composed of different visibility parts as, e.g., visibility via links from other web sites, visibility via listings in online directories, and visibility via search engines, to mention just the most important ones. Some information of this kind is already used by search engines within their strategies to place the most important web pages on top of corresponding listings. All in all two main kinds of drivers of Online Visibility can be identified:

1. Psychological Drivers of Online Visibility: This means that human online searching and surfing behavior and ways how humans interact with the internet or with search engines (e.g., only the first three result pages of search engines are normally inspected by browsing individuals) have to be taken into consideration.

2. Physical Drivers of Online Visibility: Physical drivers are such as links to a web site, banner ads, listings in search engines or directories etc.

Both psychological and physical drivers cause differences with respect to Online Visibility. To determine the real impact on Online Visibility one would have to subtract all overlappings from different visibility parts. For this reason, it is difficult to determine a precise measure. However, one can approximate a measure that takes the main phenomena mentioned into account.

3.2 Web data used for our sample

We used the Google search engine for data collection, because Google holds 73.4% of the share of the search engine market (percent of search requests), followed by Yahoo! with 5.5% (Webhits (2004)).

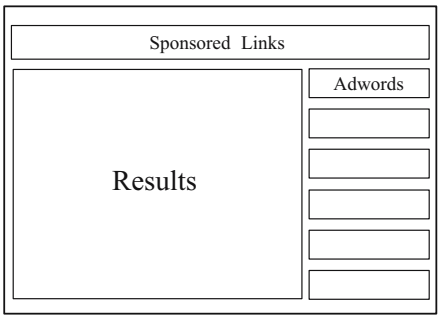


Fig. 2. Conventional result window of Google

If one considers a conventional result window of the Google search engine as it is shown in figure 2, one sees on the right hand side the so-called “Adword

Area", in the middle the result list of Urls corresponding to the search request, and up to two so-called "Sponsored Links" on top of the result screen. For our measurement of Online Visibility we keep track of the appearance of the Urls of interest in the result list and in the adword area for appropriate requests characterized by their keywords.
Additionally, we considered the AltaVista search engine for determining the number of incoming links as in AltaVista it is possible to exclude links from the home domain (link:www.xyz.com -host:www.xyz.com).
We excluded the measurement of Online Visibility in directories, on portals, in chat rooms or banner ads, etc. The reason is that it is not possible to measure OV, e.g., in directories in an impartial way (alphabetical order) and to take changes of banner ads into account without a huge amount of data from other webmasters (see, e.g. Drèze and Zufryden (2003) who incorporated expensive and time consuming information retrieval methods to calculate their measure, which is static and only a snap shot based on a selective situation in the web).

3.3 The measure GOVis

Obviously, there are many ways to try to formulate an Online Visibility measure but based on the reasons mentioned before our approach

$$GOVis(L) = \sum_{k=1}^{\sum_{n=1}^{N}\binom{N}{n}\cdot n!} \left[\alpha \cdot \sum_{p=1}^{2}\sum_{r=1}^{R} \frac{1}{e^{p-1}} \cdot X_{kpr} + \beta \cdot \sum_{p=1}^{2}\sum_{a=1}^{A} Y_{kpa} \right] + \gamma \cdot f(Z_L)$$

is a fast and cheap method and independent of third party data. Additionally, consecutive investigations can be performed. Here

* $\mathcal{K}$ is a set of interesting keywords for a query, with $\mid \mathcal{K} \mid = N$ (normally $N \leq 3$), $\sum_{n=1}^{N}\binom{N}{n} \cdot n!$ is the quantity of all ordered subsets of $\wp(\mathcal{K}) \setminus \{\emptyset\}$ and k is the kth subset of keywords with which a query in *Google* could be performed,
* p is the depth of the result pages of the search engine used (normally depth $p \leq 2$),
* R is the quantity of results per result page and r is the rth ranking position on the result pages (we used *Google* with $R = 10$),
* A is the maximum quantity of adwords per result page (*Google* standard is $A = 8$) and a is the ath adword ranking position,
* L is the corresponding URL, Z_L is the number of corresponding fan-ins,
* h_{kpr} is the hyperlink at the rth ranking position on the page with depth p by generating a query with the kth subset of keywords,
* w_{kpa} is the adword link at the ath adword ranking position on the page with depth p by generating a query with the kth subset of keywords,
* $X_{kpr} = \begin{cases} 1, & h_{kpr} \text{ links to } L \\ 0, & \text{otherwise} \end{cases}$ $\quad Y_{k1a} = \begin{cases} 1, & w_{kpa} \text{ links to } L \\ 0, & \text{otherwise} \end{cases}$
* $\alpha + \beta + \gamma = 1$ (these parameters help to adjust overlappings),
* and $f(Z_L)$ is a step function based on figure 1.

3.4 Results

We examined different branches: e.g., online book stores, erotic service web sites, automobile, and nonprofit web sites. We also observed "trendy" web sites such as the German home page of the movie "Lord of the Rings". The number of incoming links changed dependent on the branches observed (the number of incoming links of one book store decreased by 1,500 whereas the number of incoming links of erotic and nonprofit web sites were static in the last month (March 2004)). Incoming links of "trendy" web sites alter in accordance with the interest given to these web sites by press or television (e.g., the number of incoming links of the "Lord of the Rings" web site reincreased again after the academy awards of 11 "Oscars" in March 2004 up to 4,099 and fell down to 216 in April 2004).
The function $f(Z_L)$ that we used is based on the findings presented in figure 1, relative to $\log Z_L$, and absorbs changes in the number of incoming links to some extend (e.g., if a web site has already "many" incoming links, it doesn't matter if it gains or looses "some").
Figure 3 shows GOVis results of book stores ($\mathcal{K} = \{$dvds, roman, bestseller$\}$)

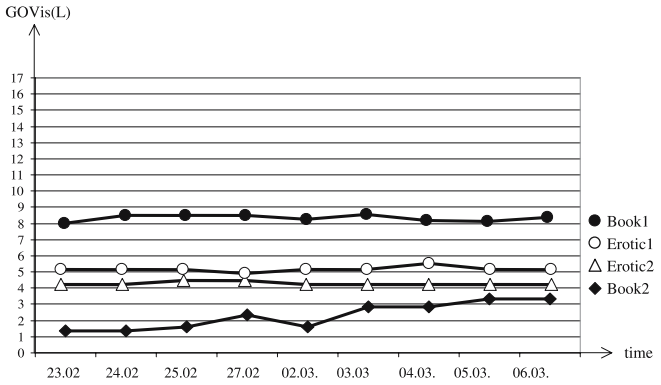


Fig. 3. $GOVis(L)$ for different web sites

and erotic service web sites ($\mathcal{K} = \{$erotik, sex, porno$\}$); $\alpha = 0.39$, $\beta = 0.01$, and $\gamma = 0.6$ was selected according to a scenario based on the numbers of table 1 (Take $\alpha \approx 13.46/35$, $\beta \approx 0.5/35$, and $\gamma \approx 21.04/35$).
One sees that the outcomes of $GOVis(L)$ for erotic web sites are very close to each other. Based on our measure managers of these sites can evaluate how their activities influence Online Visibility compared to salient competitors. For book stores we could select two example which demonstrate how different Online Visibility can be. Book1 is on an upper level of Online Visibility, because this web site has good listings in Google for nearly every keyword combination based on the chosen set $\mathcal{K}$ and has also many incoming links. Book2 can now try various activities and observe how it "best" (on-/offline

marketing campaigns) can approach this competitor (and, indeed, between February 23 and March 6, 2004, the GOVis measure has increased by two units). With the help of GOVis one can measure the "own" Online Visibility, but can also make online competitors visible to search for best practice examples of web sites and to derive hints for successful actions. Table 2 shows the

Table 2. Appearance of URLs for different branches from April to June 2004

Branch	#Urls in Total	#Urls in Results	#Urls in Adwords	% of Capacity of Adword Area
Book	502	377	117	48,9%
Erotic	1920	1058	853	91,55%
Automobile	565	337	215	99,6%
Nonprofit	1032	908	118	16%

appearance of Urls for different branches in online business. The appearance of URLs and the used part of the Adword Area helps to adjust the choice of α, β, and γ for GOVis. In our sample, e.g., the automobile and erotic branch is very active in the Adword Area.

4 Conclusion and managerial implications

In total, $GOVis(L)$ is qualified for revealing bench marks with respect to possible competitors and observing visibility changes over time in the web. It is also suited to get general impressions concerning how different branches use adwords or rely on incoming links.

However, visibility has to be measured in constant short time periods to get a deeper understanding of the rate with which the WWW or, more precisely, subwebgraphs are changing and how certain web activities influence the GOVis measure.

Based on the money spent on the optimization of web sites with respect to Online Visibility, one can observe the success (unsuccessfulness) of special arrangements with the help of GOVis. Although GOVis is only one suggestion to measure Online Visibility with the help of content visibility, adwords visibility, search engine visibility, and visibility based on incoming links (which, however, appear to be the most important instruments to account for Online Visibility), some managerial implications are obvious:

To improve Online Visibility it is important to follow different strategies. One question is, how the link structure of the web site is built up, and another one, how many links are pointing to different pages. An obvious strategy to generally improve the ranking in search engines is to be listed in online

directories. And, because there is an impact of the order of keywords in a query, the way of ordering the content of web pages has to be considered for long-time optimization of a corresponding web site. At first, however, web site owners have to find out which keywords are relevant for online searchers and web site content. For example, it is possible to observe the log files of corresponding web sites to detect search engine referrals including important keywords of searching persons. Another possibility is to sift through online keyword databases to compare already used keywords with descriptions or text content of the web site of interest to meet customer needs with respect to content, special topics or product descriptions. And if one detects potential competitors with GOVis, it is possible to analyze the sites of these competitors to find out whether their web appearance works better than the own one.

References

BARABASI, A. and ALBERT, R. (1999): Emergence of Scaling in Random Networks. *Science, 286, 509–512.*

BRIN, S. and PAGE, L. (1998): The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the 7th International World Wide Web Conference, WWW7, 107–117.*

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A. and WIENER, J. (2000): Graph Structure in the Web: Experiments and Models. *Computer Networks, 33, 309–320.*

DRÈZE, X. and ZUFRYDEN, F. (2003): The Measurement of Online Visibility and its Impact on Internet Traffic. *Journal of Interactive Marketing, 18(1), 20–37.*

FISCHERLÄNDER, S. (2003): Websites Google-gerecht - Ganz nach oben.*iX 08/2003, 84–87.*

HÖLSCHER, C. (1998): How Internet Experts Search for Information on the Web. In: H. Maurer and R.G. Olson (Eds.): *Proceedings of WebNet98 - World Conference of the WWW, Internet and Intranet.* AACE, Charlottesville, VA.

JANSEN, B., SPINK, A. and SARACEVIC, T. (2000): Real Life, Real Users, and Real Needs: A Study Analysis of User Queries on the Web. *Information Processing and Management, 36, 207–227.*

JOHNSON, E.J. (2002): On the Depth and Dynamics of Online Search Behavior. *Department of Marketing, Columbia Business School, Columbia University.*

KLEINBERG, J. (1999): Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM, 46(5), 604–632.*

SILVERSTEIN, C., HENZINGER, M., MARAIS, H. and MORICZ, M. (1999): Analysis of a Very Large AltaVista Query Log. *SIGIR Forum, 33(1), 599–621.*

Webhits (2004): Web-Barometer, Nutzung von Suchmaschinen, www.webhits.de, last visited April 14, 2004.

WebSideStory (2003): Search Engine Referrals Nearly Double Worldwide, According to WebSideStory, March 2003, www.websidestory.com, last visited March 18, 2004.

Analysis of Recommender System Usage by Multidimensional Scaling

Patrick Thoma and Wolfgang Gaul

Institut für Entscheidungstheorie und Unternehmensforschung,
Universität Karlsruhe (TH), 76128 Karlsruhe, Germany

Abstract. Recommender systems offer valuable information not only for web site visitors (who are supported during site navigation and/or buying process) but also for online shop owners (who can learn from the behavior of their web site visitors). We use data from large German online stores gathered between March and November 2003 to visualize search queries by customers together with products viewed most frequently or purchased most frequently. Comparisons of these visualizations lead to a better understanding of searching, viewing, and buying behavior of online shoppers and give important hints how to further improve the generation of recommendations.

1 Introduction

A recommender system can be defined as software that collects and aggregates information about site visitors (e.g., buying histories, products of interest, hints concerning desired/desirable search dimensions or FAQ) and their actual navigational and buying behavior and returns recommendations (e.g., based on customer demographics and/or past behavior of the actual visitor and/or user patterns of top sellers with fields of interest similar to those of the actual contact (Gaul and Schmidt-Thieme (2002)). A framework to classify such systems according to their input and output facilities can be found in Gaul et al. (2002). A substantial problem in the evaluation of recommender systems is that in real-life environments their influence to customers buying behavior can hardly be measured isolated from other effects, e.g., promotions, price reductions, etc. The research questions therefore was to analyze the acceptance and the functioning of recommender systems in a more qualitative way by making the site visitors interaction with it transparent.

In the following we will use multidimensional scaling (MDS) to visualize how online shoppers place search queries and react to recommender system output where MDS is the label for a class of methods that represent similarity or dissimilarity values with respect to pairs of objects as distances between corresponding points in a low-dimensional metric space (Borg and Groenen (1997)). The graphical display of the object representations as provided by MDS enables to literally "look" at the data and to explore structures visually. We used this technique to analyze recommender system usage on the basis of two data sets, collected from two different German online retail stores.

Both use a ruled-based recommender system with the same kernel functionality to support customers during the buying process. As a result relations between searching, viewing and buying behavior of the corresponding site visitors visualized in the underlying product space revealed valuable insights for product managers and recommender system engineers.

2 Methodology

A popular and classical technique to construct object representations in a low-dimensional space is MDS (Kruskal (1964)). The goodness of fit of the solution obtained within the different iterative steps of whatever method can be assessed by the so called *Kruskal stress*. We used an implementation of Kruskal's non-metric MDS which is available as the *isoMDS* function in the "MASS" library (Venables and Ripley (1997)) of the "R" software package. To overcome a weakness of classical MDS, i.e. the interpretation of the object positionings in low-dimensional spaces, we applied *property fitting*.

Let $O = \{o_1, \ldots, o_N\}$ be the set of objects and $b_n = (b_{n1}, \ldots, b_{nM})$ the representation of object o_n in the underlying M-dimensional target space, $n = 1, \ldots, N$. If additional information about the objects is given, e.g., in form of attribute vectors $a_p = (a_{1p}, \ldots, a_{Np})'$ for property p, $p = 1, \ldots, P$, one can construct property vectors $c_p = (c_{p1}, \ldots, c_{pM})$ in the M-dimensional space so that the projections

$$\hat{a}_{np} = \sum_{m=1}^{M} c_{pm} b_{nm} \tag{1}$$

of b_n onto c_p approximate the actual attribute values of the objects as good as possible with respect to the least squares criterion

$$\sum_{n=1}^{N} (\hat{a}_{np} - a_{np})^2. \tag{2}$$

Vector notation leads to $c_p = (B'B)^{-1}B'a_p$ with $B = (b_{nm})$. Quality of fit can be measured by correlation coefficients between $\hat{a}_{np}$ and a_{np}.

Similarly, we performed the subsequent transformation of the search queries $s_q = (s_{q1}, \ldots, s_{q\tilde{P}_q})'$. Here, $\tilde{p} = 1, \ldots, \tilde{P}_q, \tilde{P}_q \leq P$, indicates properties specified in the underlying query. In cases where a range of values was stated for a property p, e.g., for the price, we set $s_{q\tilde{p}}$ equal to the mean obtained from the lower and upper boundaries of the specified range. With given property vectors $c_{\tilde{p}}$ we looked for the representation $z_q = (z_{q1}, \ldots, z_{qM})'$ of s_q so that the projections

$$\hat{s}_{q\tilde{p}} = \sum_{m=1}^{M} z_{qm} c_{\tilde{p}m} \tag{3}$$

of z_q onto $c_{\tilde{p}}$ approximate the actual attribute values of the search profiles as good as possible with respect to the least squares criterion

$$\sum_{\tilde{p}=1}^{\tilde{P}_q} (\hat{s}_{q\tilde{p}} - s_{q\tilde{p}})^2. \tag{4}$$

Vector notation leads to $z_q = (C_q' C_q)^{-1} C_q' s_q$ with $C_q = (c_{\tilde{p}m})$.

3 Empirical results

3.1 The data sets

We collected system usage information from two large German online retail stores. The first data set contains searching, viewing and buying information of website visitors looking for notebooks, the second data set contains the respective information for washing machines. Both stores support their customers with the help of recommender systems for dedicated product domains. In both cases, users can specify a search query by defining the importance of and the desired value for every attribute from a list of given attributes. The recommender systems - as response - compute sorted lists of products which best fulfill the customers requirements using an internal ruled-based algorithm. The quality of the proposals is implicitly evaluated by counting the number of clicks on the images of the suggested products which lead to pages with additional, more detailed information about the products of interest and additionally, by counting how often corresponding products were put into electronic market baskets. These two events are in the following called "views" and "purchases".

The data set about notebooks was gathered between March and July 2003. Products were described by a list of 14 attributes of which we selected price, clockrate, ram, harddisk, display, drives, weight, interfaces, battery, and software for our analysis. Using these 10 properties a list of 307 different products could be recommended. The data set contained 7125 search queries of which 434 were empty which means that no values were specified for any attribute. Feedback information consisted of 15305 views and 509 purchases.

The data set for washing machines was collected from May until November 2003. In this case, the selected attributes were price, type (front or top loader), charge, effectiveness, maximum spin, programs, and extras. This data set contains 54 different products, 20024 search queries (thereof 14131 empty ones), 49696 views, and 758 purchases.

A comparison of these numbers shows the following: Online shoppers had a closer look at 2-3 products (notebooks: 2.15, washing machines: 2.50) per search query. The different number of empty search queries and the fact, that for notebooks, the "click-to-buy-ratio" (purchases/views) is more than twice the number for washing machines (notebooks: 3.3%, washing machines: 1.5%)

might be an indicator that site visitors looking for notebooks have already a clearer conception of what they are searching and thus are more willing to buy online than users looking for washing machines. In our analysis we tried to verify these assumptions of this kind.

3.2 Representation of products and search profiles

A first step of our analysis was to display the products together with the specified search queries in a two-dimensional space. Dissimilarities between objects were defined by first scaling all attribute vectors to $N(0,1)$ and then calculating euclidean distances using the *dist* function (The R development core team (2003)). As mentioned above, the object representations were obtained with the help of the *isoMDS* function. The *stress* measures of the so found solutions were sufficient (0.175) for notebooks and good (0.146) for washing machines. The correlation coefficients of the property vectors where rather high (> 0.7) in most cases, exceptions are interfaces (0.47) and battery (0.56) for notebooks and programs (0.56) for washing machines. To structure the many search queries, we grouped identical queries together and used the notation "search profiles" for identical ones. We selected search profiles with at least 10 queries and transformed them into the two-dimensional target space according to the methodology explained in section 2. The results are shown in figures 2 and 1. In figure 2 one can see, that notebooks and search profiles are relatively equally spread around the origin. From a management's point of view, this could be interpreted as hint that the product catalog for notebooks covers the range of customers' requirements quite well. This is not always the case as Gaul et al. (2004) revealed when analyzing a recommender system for digital cameras. Remarkable is always the group of search profiles forming a straight configuration in direction of the negative prolongation of the price vector (The price vector is marked with a dashed line.). These profiles stand for queries where a low price was the only specified attribute. From the comparatively high number of such profiles one can – not surprising – conclude that price is one of the most important search criteria and that many of the users are quite price sensitive since these profiles represent demand for products below the average price. For washing machines, figure 1 shows two more distinguishable lines of search profiles. These configurations result from queries that are identical in all but one attribute. The differentiating attribute is the number of extras like a "water-stop-function" or a "timer". Contrary to the notebooks' data set, most of the search profiles are situated in the area of above average prices. This could be a hint that online shoppers interested in washing machines are less price sensitive and more in favor of additional equipment of the products.

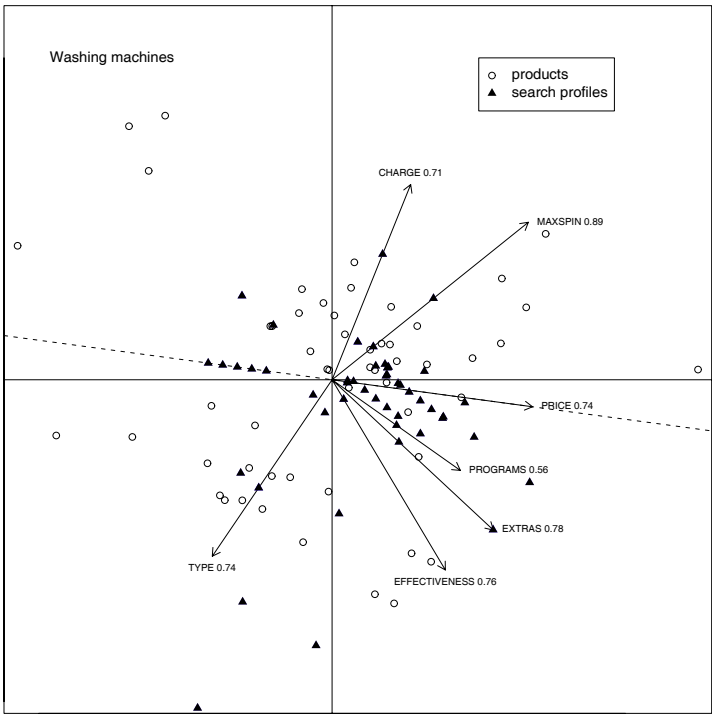


Fig. 1. Property vectors, search profiles, and products (Washing machines)

3.3 Analysis of system usage

In the second step, we analysed coherences between specified search profiles and subsequently viewed and/or purchased products. For this purpose, we determined for each profile the ten most frequently viewed products and connected their object representations in the target space with the representation of the search profile. This would lead to many spider like graphs. In figure 3 the situation is shown for two selected search profiles from the data set for washing machines. In both cases, most of the frequently viewed products are relatively close to the specified search profile. However, there are also products which have been viewed frequently but are positioned quite far away from the originally specified search profile. For the search profile in the 3rd quadrant these are the two products in the 4th quadrant. Looking at the underlying data shows that these two products are of different type (front- instead of top-loader). The opposite case also exists: There are products which seem to fulfill the requirements quite well according to their positioning in the spider graph but have not been viewed frequently. As users can only view products that have been recommended, this could be an indicator that these products were not on the recommendation list for the given search profile. Checking

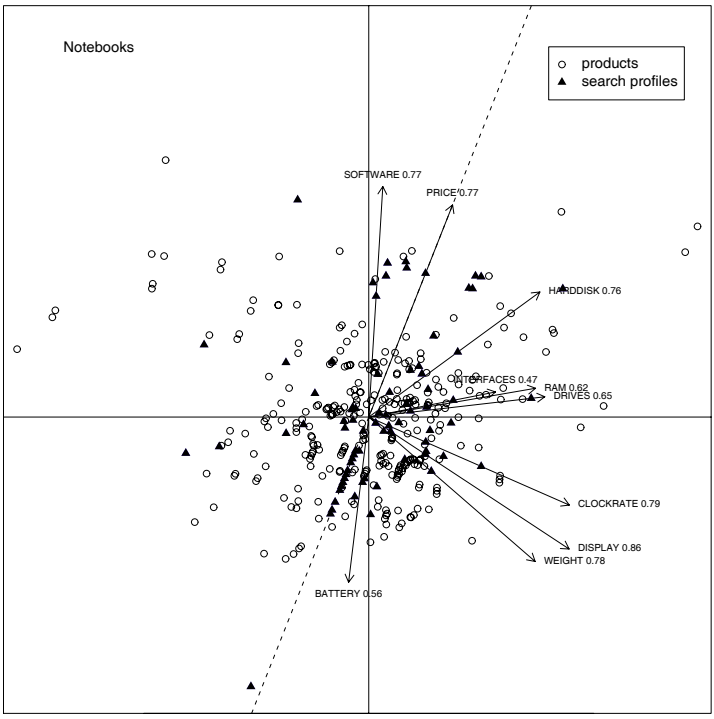


Fig. 2. Property vectors, search profiles, and products (Notebooks)

this with the actual recommendations - an information not available in our data sets - might help to improve recommendation generation.

The same analysis for notebooks showed a different picture (figure 4). In this case, the most frequently viewed products cannot really be characterized as always close to the search profile. From the underlying data, we can see that a desired price of 900 EURO was specified while the prices of the products viewed most frequently ranged from 979 EURO up to 1649 EURO. This means that online shoppers looking for notebooks are restrictive with respect to price when formulating their search queries, but are nevertheless interested in better equipped items - even if the price is far beyond the their preferred value.

This leads to the question how site visitors decide when it comes to buying. Figure 5 shows that for the selected search profile there is only one product in the category "purchased most frequently". Remarkably enough, this product is not among the ones of the category "viewed most frequently". Its representation has a medium distance from the representation of the search profile (in numbers: its price was 1249 EURO while 900 EURO were specified in the search query). This supports the assumption from section 3.1 that online shoppers looking for notebooks have already quite a good knowledge

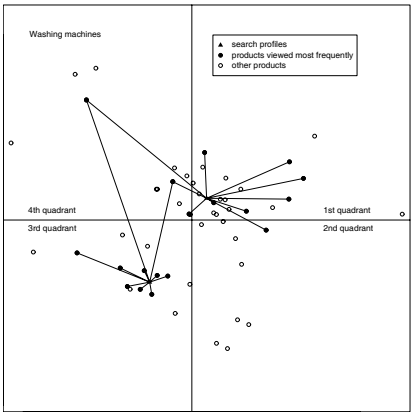


Fig. 3. Selected search profiles and products viewed most frequently (Washing machines)

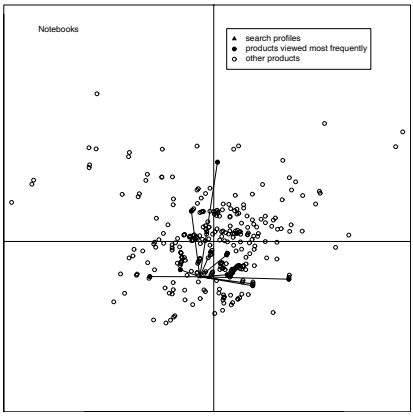


Fig. 4. Selected search profiles and products viewed most frequently (Notebooks)

about the product domain and do not rely too much on recommendations. For washing machines the situation is, again, different. Site visitors of washing machines bought to a much higher extent a product that was among the ones viewed most frequently (see figure 6). In the given example, no product was purchased after the search query in the 3rd quadrant had been specified. The products purchased on basis of the search query in the 1st quadrant are all relatively close and had been viewed frequently. This could be seen as a confirmation, that online shoppers looking for washing machines are more uncertain about their actual needs and therefore more willing to rely on suggestions of recommender systems.

4 Summary

In this paper two product domains were used to demonstrate, that Multidimensional Scaling can be a suitable tool to analyze recommender system usage. For the given data sets, products were represented in a two-dimensional space with sufficient/good stress measures and high correlation coefficients for additional property vectors. The presented methodology allowed us to subsequently transform the search profiles into the same target space and to display them together with the products in joint representations. The visualization of searching, viewing and buying behaviour of online shoppers showed remarkable differences between the two product domains. While notebook site visitors seem to have extensive knowledge concerning the product domain and to be rather independent from system recommendations, online shoppers looking for washing machines are more uncertain and therefore more willing to rely on recommender system's suggestions when they make their

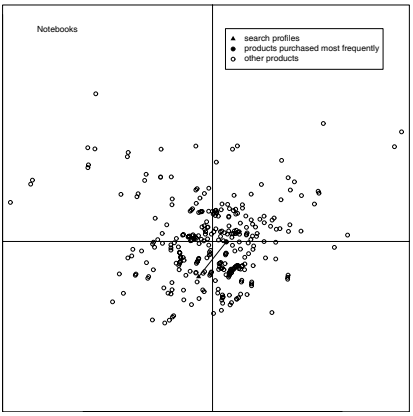


Fig. 5. Selected profiles and products purchased most frequently (Notebooks)

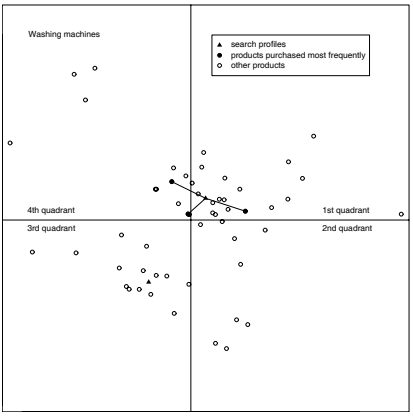


Fig. 6. Selected search profiles and products purchased most frequently (Washing machines)

buying decision. Analysis of this kind can create value for product managers and shop owners as well as for developers of recommender systems' software.

References

BORG, I. and GROENEN, P. (1997): *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York.

GAUL, W., GEYER-SCHULZ, A., SCHMIDT-THIEME, L. and HAHSLER, M. (2002): eMarketing mittels Recommendersystemen. *Marketing ZFP, 24, 47–55*.

GAUL, W. and SCHMIDT-THIEME, L. (2002): Recommender Systems Based on User Navigational Behavior in the Internet. *Behaviormetrika. 29, 1–22*.

GAUL, W., THOMA, P., SCHMIDT-THIEME, L., and VAN DEN BERGH, S. (2004): Visualizing Recommender System Results via Multidimensional Scaling, to appear in: *Proceedings of OR 2003 International Conference*. Springer, New York.

KRUSKAL, J. B. (1964): Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika, 29, 1–27*.

THE R DEVELOPMENT CORE TEAM (2003): *The R Environment for Statistical Computing and Graphics*.
http://cran.r-project.org/doc/manuals/fullrefman.pdf

VENABLES, W.N. and RIPLEY, B.D. (1997): *Modern Applied Statistics with S-PLUS*, 3rd ed. Springer, New York.

On a Combination of Convex Risk Minimization Methods

Andreas Christmann

University of Dortmund*
Department of Statistics
44221 Dortmund, Germany

Abstract. A combination of methods from modern statistical machine learning theory based on convex risk minimization is proposed. An interesting pair for such a combination is kernel logistic regression to estimate conditional probabilities and $\varepsilon-$support vector regression to estimate conditional expectations. A strategy based on this combination can be helpful to detect and to model high-dimensional dependency structures in complex data sets, e.g. for constructing insurance tariffs.

1 Introduction

In some applications regression problems can occur where the data set has the following characteristic features. (1) Most of the response values are zero. (2) The empirical distribution of the positive responses is extremely skewed to the right. (3) There is a complex and high dimensional dependency structure between explanatory variables. (4) The data sets to be analyzed are huge. (5) Some extreme high response values are rare events, but contribute enormously to the total sum of the response values. Sometimes there are only imprecise values available for some explanatory variables or some response values are only estimates. In this case robustness properties of the estimation techniques can be important, cf. Rousseeuw and Christmann (2003). Regression models to develop insurance tariffs or for scoring credit risks are examples where data sets can have such features.

In Section 2 a simple strategy is described that exploits knowledge of these features in order to detect and model hidden structure in such data sets. In Section 3 some facts of kernel logistic regression and $\varepsilon-$logistic regression are given. Section 4 briefly gives some results for applying the strategy to a large data set from an insurance project and Section 5 gives a discussion.

2 Strategy

Let Y denote the non-negative response variable and $x \in \mathbb{R}^p$ the vector of explanatory variables. In the first step we construct an additional stratification variable C by defining a small number of classes for the values of Y

* This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

with a high amount of interpretability. For example, define a discrete random variable C by

$$C = \begin{cases} 0, & \text{if } Y = 0 \quad \text{(no claim)} \\ 1, & \text{if } Y \in (0, c_1] \quad \text{(small claim)} \\ 2, & \text{if } Y \in (c_1, c_2] \quad \text{(medium claim)} \\ 3, & \text{if } Y \in (c_2, c_3] \quad \text{(high claim)} \\ 4, & \text{if } Y > c_3 \quad \text{(extreme claim).} \end{cases}$$

Of course, it depends on the application how many classes should be used and how reasonable boundary values can be defined. We will not address this problem here. Note that given the information that no event occurred, it holds that

$$\mathrm{E}(Y|C = 0, X = x) \equiv 0\,. \tag{1}$$

Using (1) we can write the conditional expectation of Y given $X = x$ by

$$\mathrm{E}(Y|X = x) = \mathrm{P}(C > 0|X = x) \times \sum_{c=1}^{k} \mathrm{P}(C = c|C > 0, X = x) \cdot \mathrm{E}(Y|C = c, X = x)\,, \tag{2}$$

and we denote this formula as strategy A. Note, that in (2) the summation starts with $c = 1$. Hence, it is only necessary to fit regression models to small subsets of the whole data set. However, one has to estimate the conditional probability $\mathrm{P}(C > 0|X = x)$ and the multi-class probabilities $\mathrm{P}(C = c|C > 0, X = x)$ for $c \in \{1, \ldots, k\}$, e.g. by a multinomial logistic regression model or by kernel logistic regression. If one splits the total data set into three subsets for training, validating, and testing, one only has to compute *predictions* for the conditional probabilities and the corresponding conditional expectations for *all* data points. Bias reduction techniques applied to the validation data set may be helpful to reduce a possible bias of the estimates.

From our point of view the indirect estimation of $\mathrm{E}(Y|X = x)$ via strategy A has practical and theoretical advantages over direct estimation of this quantity. The terms in (2) are interesting, because they contain additional information which can be important for example in the context of insurance tariffs: the probability $\mathrm{P}(C > 0|X = x)$ that a customer with characteristics x has at least one claim, the conditional probabilities $\mathrm{P}(C = c|C > 0, X = x)$ that a claim of size class c occurs, and the conditional expected claim size $\mathrm{E}(Y|C = c, X = x)$ given the event that the claim was within the size class c. The strategy circumvents the problem that most observed responses y_i are 0, but $\mathrm{P}(Y = 0|X = x) = 0$ for many classical approaches based on a gamma or log-normal distribution. A reduction of computation time is possible because we only have to fit regression models to a small subset of the data set. The estimation of conditional class probabilities for the whole data set is often much faster than fitting a regression model for the whole data set. It is possible, that different explanatory variables show a significant impact on

the response variable Y or on the conditional class probabilities for different classes defined by C. This can also result in a reduction of interaction terms. It is possible to use different variable selection methods for the $k+1$ classes. This can be especially important for the class of extreme events: because there may be only some hundreds or a few thousands of these rare events in the data set, it is in general impossible to use all explanatory variables for these data points. Finally, the strategies have the advantage that different techniques can be used for estimating the conditional class probabilities $\mathrm{P}(C = c|X = x)$ and for estimating the expectations $\mathrm{E}(Y|C = c, X = x)$ for different values of C. Examples for reasonable pairs are:

- Multinomial logistic regression + Gamma regression
- Robust logistic regression + semi-parametric regression
- Multinomial logistic regression + ε-Support Vector Regression (or $\nu-$SVR)
- Kernel logistic regression (KLR) + ε-Support Vector Regression
- Classification trees + regression trees
- A combination of the pairs given above, possibly also with methods from extreme value theory.

Even for data sets with several million of observations it is generally not possible to fit simultaneously all high-dimensional interaction terms with classical statistical methods such as logistic regression or gamma regression because the number of interaction terms increases too fast.

The combination of kernel logistic regression and $\varepsilon-$support vector regression described in the next section, both with the popular exponential radial basis function (RBF) kernel, has the advantage that important interaction terms are fitted automatically without the need to specify them manually.

We like to mention that some statistical software packages (e.g. R) may run into trouble in fitting multinomial logistic regression models for large and high dimensional data sets. Two reasons are that the dimension of the parameter vector can be quite high and that a data set with many discrete variables recoded into a large number of dummy variables will perhaps not fit into the memory of the computer. To avoid a multinomial logistic regression model one can consider all pairs and then use pairwise coupling, *cf.* Hastie and Tibshirani (1998).

Of course, the law of total probability offers alternatives to (2). The motivation for the following alternative, say strategy B, is that we first split the data into the groups 'no claim' versus 'claim' and then split the data with 'claim' into 'extreme claim' and the remaining classes:

$$\begin{aligned}\mathrm{E}(Y|X = x) = \mathrm{P}(C > 0|X = x) \cdot & \\ \{\mathrm{P}(C = k|C > 0, X = x) \cdot \mathrm{E}(Y|C = k, X = x) + & \\ [1 - \mathrm{P}(C = k|C > 0, X = x)] \cdot & \\ \sum_{c=1}^{k-1} \mathrm{P}(C = c|0 < C < k, X = x) \cdot \mathrm{E}(Y|C = c, X = x)\} . & \end{aligned} \tag{3}$$

This formula shares with (2) the property that it is only necessary to fit regression models to subsets of the whole data set. Of course, one can also interchange the steps in the above formula, which results in strategy C:

$$\begin{aligned} \mathrm{E}(Y|X=x) = {} & \mathrm{P}(C=k|X=x)\cdot \mathrm{E}(Y|C=k,X=x) \\ & + [1-\mathrm{P}(C=k|X=x)]\cdot\{\mathrm{P}(C>0|C\neq k,X=x)\cdot \\ & \sum_{c=1}^{k-1}\mathrm{P}(C=c|0<C<k,X=x)\cdot \mathrm{E}(Y|C=c,X=x)\}\,. \end{aligned} \tag{4}$$

Note that two big binary classification problems have to be solved in (4), whereas there is only one such problem in (3).

3 Kernel logistic regression and $\varepsilon-$support vector regression

In this section we briefly describe two modern methods based on convex risk minimization in the sense of Vapnik (1998), see also Schölkopf und Smola (2002).

In statistical machine learning the major goal is the estimation of a functional relationship $y_i \approx f(x_i)+b$ between an outcome y_i belonging to some set $\mathcal{Y}$ and a vector of explanatory variables $x_i = (x_{i,1},\ldots,x_{i,k})' \in \mathcal{X} \subseteq \mathbb{R}^p$. The function f and the intercept parameter b are unknown. The estimate of (f,b) is used to get predictions of an unobserved outcome y_{new} based on an observed value x_{new}. The classical assumption in machine learning is that the training data (x_i,y_i) are independent and identically generated from an underlying unknown distribution P for a pair of random variables (X_i,Y_i), $1\leq i\leq n$. In applications the training data set is often quite large, high dimensional and complex. The quality of the predictor $f(x_i)+b$ is measured by some loss function $L(y_i,f(x_i)+b)$. The goal is to find a predictor $f_{\mathrm{P}}(x_i)+b_{\mathrm{P}}$ which minimizes the expected loss, *i.e.*

$$\mathrm{E}_{\mathrm{P}}\, L(Y,f_{\mathrm{P}}(X)+b_{\mathrm{P}}) = \min_{f\in\mathcal{F},\, b\in\mathbb{R}} \mathrm{E}_{\mathrm{P}}\, L(Y,f(X)+b), \tag{5}$$

where $\mathrm{E}_{\mathrm{P}}\, L(Y,f(X)+b) = \int L(y,f(x)+b)d\mathrm{P}(x,y)$ denotes the expectation of L with respect to P. We have $y_i\in\mathcal{Y}:=\{-1,+1\}$ in the case of binary classification problems, and $y_i\in\mathcal{Y}\subseteq\mathbb{R}$ in regression problems.

As P is unknown, it is in general not possible to solve the problem (5). Vapnik (1998) proposed to estimate the pair (f,b) as the solution of an empirical regularized risk. His approach relies on three important ideas: (1) restrict the class of all functions f to a broad but still flexible subclass of functions belonging to a certain *Hilbert space*, (2) use a *convex* loss function L to avoid computational intractable problems which are NP-hard, and (3)

use a *regularizing term* to avoid overfitting. Let $L : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$ be an appropriate convex loss function. Estimate the pair (f, b) by solving the following empirical regularized risk minimization:

$$(\hat{f}_{n,\lambda}, \hat{b}_{n,\lambda}) = \arg \min_{f \in \mathcal{H},\, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i) + b) + \lambda \|f\|_{\mathcal{H}}^2, \tag{6}$$

where $\lambda > 0$ is a small regularization parameter, $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) of a kernel k, and b is an unknown real-valued offset. The problem (6) can be interpreted as a stochastic approximation of the minimization of the theoretical regularized risk, i.e.

$$(f_{\mathrm{P},\lambda}, b_{\mathrm{P},\lambda}) = \arg \min_{f \in \mathcal{H},\, b \in \mathbb{R}} \mathrm{E}_{\mathrm{P}}\, L(Y, f(X) + b) + \lambda \|f\|_{\mathcal{H}}^2 \,. \tag{7}$$

In practice, it is often numerically better to solve the dual problem of (6). In this problem the RKHS does not occur explicitly, instead the corresponding kernel is involved. The choice of the kernel k enables the above methods to efficiently estimate not only linear, but also non-linear functions. Of special importance is the exponential radial basis function (RBF) kernel

$$k(x, x') = \exp(-\gamma \|x - x'\|^2) \,, \quad \gamma > 0, \tag{8}$$

which is a universal kernel on every compact subset of $\mathbb{R}^d$.

If C can have more than two values, we consider several binary regression models for $C \in \{i, j\}$, and combine the estimation results in these models by the pairwise coupling approach proposed by Hastie and Tibshirani (1998).

For the case of binary classification, popular loss functions depend on y and (f, b) via $v = y(f(x) + b)$. Special cases are:

Support Vector Machine (SVM): $L(y, f(x)+b) = \max\{1 - y(f(x)+b), 0\}$
Least Squares SVM: $L(y, f(x) + b) = [1 - y(f(x) + b)]^2$
Kernel Logistic Regression: $L(y, f(x) + b) = \log(1 + \exp[-y(f(x) + b)])$

Kernel logistic regression has the advantage that it estimates

$$\log\Big(\frac{\mathrm{P}(Y = +1|X = x)}{\mathrm{P}(Y = -1|X = x)}\Big) \,,$$

i.e. $\mathrm{P}(Y = +1|X = x) = 1/(1+e^{-[f(x)+b]})$, such that scoring is possible. Note that the support vector machine 'only' estimates whether $\mathrm{P}(Y = +1|X = x)$ is above or below $\frac{1}{2}$.

For the case of regression, Vapnik (1998) proposed the $\varepsilon-$support vector regression ($\varepsilon-$SVR) which is based on the $\varepsilon-$insensitive loss function

$$L_\varepsilon(y, f(x) + b) = \max\{0, |y - [f(x) + b]| - \varepsilon\} \,,$$

for some $\varepsilon > 0$. Note that only residuals $y - [f(x) + b]$ lying outside of an $\varepsilon-$tube are penalized. Strongly related to $\varepsilon-$support vector regression is $\nu-$support vector regression, *cf.* Schölkopf und Smola (2002).

Christmann and Steinwart (2004) showed that many statistical machine learning methods based on Vapnik's convex risk minimization principle have - besides other good properties - also good robustness properties. Special cases are kernel logistic regression and the support vector machine.

4 Application

We use strategy A based on a combination of kernel logistic regression and $\varepsilon-$logistic regression for a large data set from the Verband öffentlicher Versicherer in Düsseldorf, Germany. The data set contains information about 4 million customers from 15 single German motor vehicle insurance companies. There are two response variables: the sum of all claim sizes of a policy holder within a year and the probability that the policy holder had at least one claim. The set of possible explanatory variables is quite large and most of the variables are measured on a nominal or ordinal scale. The data set also contains geographical information, see Christmann (2004) for more details. Approximately 95% of all policy holders had no claim, such that strategy A reduces the computation cost for $\varepsilon-$support vector regression substantially. It is interesting to note that less than 0.1% of all policy holders had a pure premium value higher than 50000 EUR, but they contribute almost 50% to the grand sum. Hence, the empirical distribution of the first response variable had a large atom in zero and is extremely skewed to the right for the positive values. The data set also fulfills other features listed in Section 1.

The combination of both nonparametric methods was helpful in automatically modelling and detecting hidden structure in this high-dimensional data set. For example, there is a non-monotone relationship between age of the main user, the probability of having at least one claim and the expected sum of claim sizes. Figure 1 shows the estimates of $\mathrm{E}(Y|X = x)$ and for the conditional probabilities stratified by gender and age of the main user. The conditional probability of a claim in the interval $(0, 2000]$ EUR given the event that a claim occurred, increases for people of at least 18 years, *cf.* the subplot for $\mathrm{P}(C = 1|C > 0, X = x)$ in Figure 1. This is in contrast to the corresponding subplots for medium, high or extreme events, see the subplots for $\mathrm{P}(C = c|C > 0, X = x)$, $c \in \{2, 3, 4\}$, in Figure 1. Especially the last two subplots show, that young people have a substantial higher probability in producing a claim than more elderly people. There is also a dependency structure between age of the main user, gender and both response variables. Using claim size classes defined by the variable C turned out to be helpful to investigate whether the impact of certain explanatory variables depends on the claim size. This actually happened for some variables.

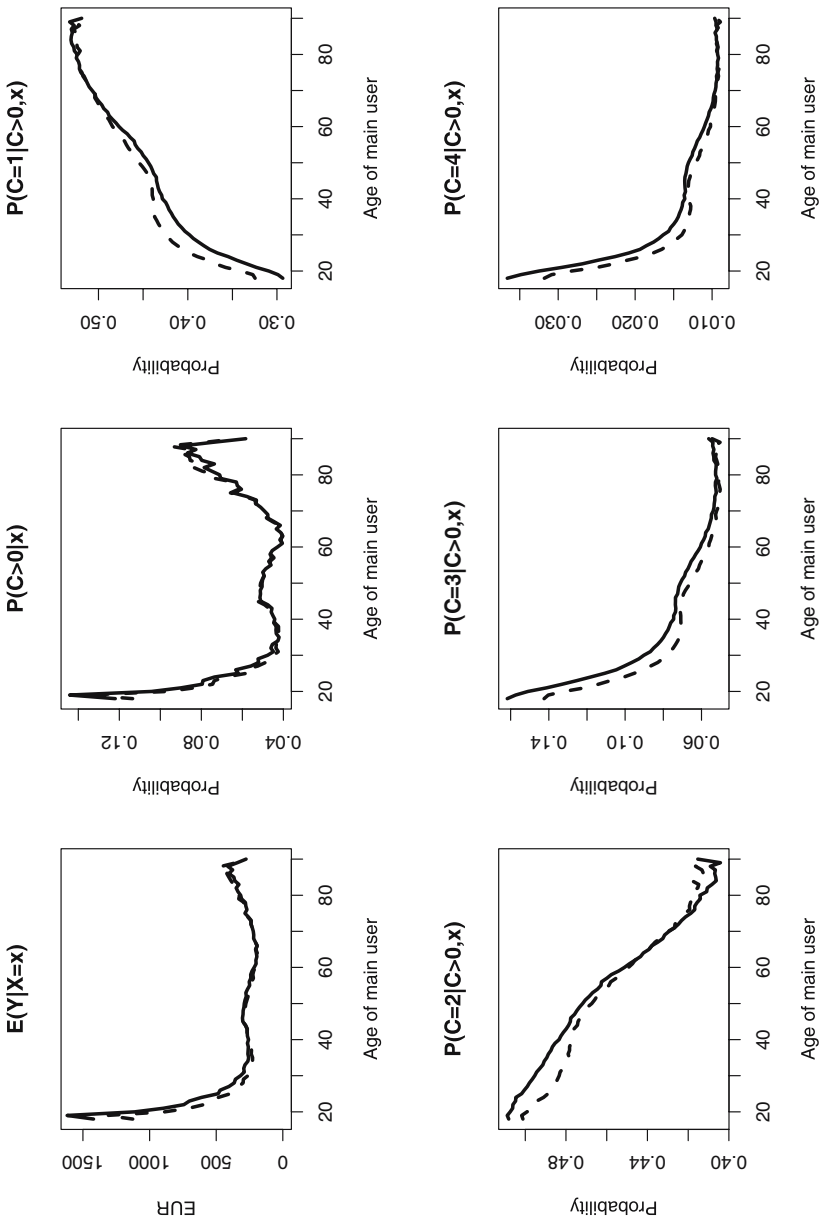


Fig. 1. Results of applying strategy A. Pure premium and conditional probabilities stratified by gender and age of the main user. Female: dashed. Male: solid.

5 Discussion

The strategy described in this paper has the goal to exploit knowledge about certain features often present in data sets from the insurance area. However, the strategy is not limited to this kind of application and it may also be helpful for credit risk scoring or in some data mining projects. A combination of modern statistical machine learning methods based on convex risk minimization can help to detect and to model complex high-dimensional dependency structures by using the kernel trick. Such dependency structures are often hard to model with classical statistical methods such as generalized linear models. An advantage of the strategy is that it is quite flexible because different estimation techniques, variable selection methods or parameter tuning can be used for the different event classes. Extreme value theory based on generalized Pareto distributions for the main knots of a tree can be helpful to model the extreme event class. For other methods from extreme value theory see e.g. Beirlant et al. (2002). The determination of the number of event classes and their endpoints will have an impact on the predictions, but a reasonable determination will depend on the concrete problem and no general rule seems to be available.

Acknowledgments

I am grateful to Mr. A. Wolfstein and Dr. W. Terbeck from the Verband öffentlicher Versicherer in Düsseldorf, Germany, for many helpful discussions.

References

BEIRLANT, J., DE WET, T., GOEGEBEUR, Y. (2002): Nonparametric Estimation of Extreme Conditional Quantiles. Katholieke Universiteit Leuven, Universitair Centrum voor Statistiek. Technical Report 2002–07.

CHRISTMANN, A. (2004): On a strategy to develop robust and simple tariffs from motor vehicle insurance data. University of Dortmund, SFB–475, TR–16/2004.

CHRISTMANN, A. and STEINWART, I. (2004): On robust properties of convex risk minimization methods for pattern recognition. *Journal of Machine Learning Research, 5, 1007–1034.*

HASTIE, T. and TIBSHIRANI, R. (1998): Classification by pairwise coupling. *Annals of Statistics, 26, 451–471.*

ROUSSEEUW, P.J. and CHRISTMANN, A. (2003). Robustness against separation and outliers in logistic regression. *Computational Statistics & Data Analysis, 43, 315–332.*

SCHÖLKOPF, B. and SMOLA, A. (2002): *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge.

VAPNIK, V. (1998). *Statistical Learning Theory.* Wiley, New York.

Credit Scoring Using Global and Local Statistical Models

Alexandra Schwarz and Gerhard Arminger

Department of Economics and Social Sciences,
University of Wuppertal, D-42097 Wuppertal, Germany

Abstract. This paper compares global and local statistical models that are used for the analysis of a complex data set of credit risks. The global model for discriminating clients with good or bad credit status depending on various customer attributes is based on logistic regression. In the local model, unsupervised learning algorithms are used to identify clusters of customers with homogeneous behavior. Afterwards, a model for credit scoring can be applied separately in the identified clusters. Both methods are evaluated with respect to practical constraints and asymmetric cost functions. It can be shown that local models are of higher discriminatory power which leads to more transparent and convincing decision rules for credit assessment.

1 Introduction

Most companies trading in durable consumer goods offer their customers the opportunity of an installment credit. Because of the increasing number of private arrears and insolvencies the modelling of consumer behavior becomes particular important in this field of consumer credits. The target is to minimize the costs resulting from the risk of the customers' financial unreliability which can be achieved by a credit assessment for predicting a customer's solvency.

In section 2 we briefly describe the data set analyzed in this paper. Section 3 deals with the basic ideas of credit scoring and refers to logistic discriminant analysis as a standard method for global classification and prediction. This type of model is based on all attributes that turned out to significantly influence the payment behavior in a certain sample. Thus the estimated parameters are equally applied to every rated customer. Therefore, it is a *global* scoring model. Furthermore we mention purposes and restrictions of credit analysis in practical applications and point out the differing classification rules. To incorporate the heterogeneity and dynamics of customer behavior we propose *local* models described in section 4. Unsupervised learning algorithms, such as k-means clustering and self-organizing maps (SOM), are applied to identify customer groups of homogeneous behavior on basis of their given attributes. Afterwards a model for credit scoring is applied to the identified clusters separately. Global and local models are developed on the same training sample. Finally, the models are evaluated with respect to their performance in the same test sample.

2 Description of the data set

The data set originates from a company selling durable consumer goods by direct sale. 25% of the customers use the opportunity of an installment credit. Among the contracts with twelve monthly installments more than 40% show deficient payment. For this company a credit assessment is established to predict whether a loan should be granted or not. The complete sample contains 5995 data sets of customers who decided to repay in twelve regular installments. As given in Table 1, it is randomly divided into a training and a test sample comprising two-thirds and one-third of all cases respectively. The training sample is used for pattern recognition and estimating model parameters whereas the test sample is applied to verify the discriminatory and predictive power of the global and local models. The credit status of a customer is measured by the observed process of repayment. If default or deficiency of repayment occur, the dependent binary coded variable TYPE is 1 and customer i is said to be not creditworthy. If the repayment is regular during the whole credit period the customer is said to be creditworthy and the variable TYPE is 0. The frequencies of the outcomes in the training and test sample are given in Table 1. The $p = 12$ classification variables provide

Table 1. Outcomes of payment behavior in training and test sample

	Training Sample	Test Sample	Total
TYPE = 1	1666 (41.7%)	810 (40.6%)	2476 (41.3%)
TYPE = 0	2333 (58.3%)	1186 (59.4%)	3519 (58.7%)
Total	3999 (66.7%)	1996 (33.3%)	5995 (100%)

information from the company's own database (age of the customer, term of delivery, amount of each installment and living in East or West Germany) as well as information made available by commercial providers of consumer and credit assessment data (e.g. social status, prevalent family structure and type of housing, former judicial and encashment procedures). All classification variables are subdivided into qualitative categories that are coded as dummy variables.

3 Global scoring model

This section introduces logistic discriminant analysis as a global model for measuring credit performance. Afterwards the adjustment of the classification rule under practical constraints is described.

3.1 Global scoring using logistic discriminant analysis

For new customers the credit status is unknown. For this reason a prospective customer i needs to be assigned to one of the two groups based on his p

individual attributes x_{ip}. So a decision rule $k(x_{ip})$ is needed which assigns customer i to the group he probably originates from. Discriminant analysis based on logistic regression (Arminger et al. (1997)) is a standard method for solving this classification problem. Here an estimation procedure with automatic variable selection is used which also considers interaction effects (Bonne (2000)). The linear predictor resulting from this logistic regression may directly be interpreted as the score of customer i:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip} \tag{1}$$

To predict a customer's solvency a threshold value c must be determined by which customer i is assessed to be not creditworthy. The resulting decision rule states negative credit status if $\hat{y}_i \geq c$. The following section deals with the determination of c in practical applications.

3.2 Classification rule under constraints

In credit assessment two classification errors occur: a customer with negative credit standing is assessed as creditworthy (with frequency h_{10}) or a customer with positive payment behavior is assessed as not creditworthy (h_{01}, cf. Table 2). The conventional classification rule minimizes the *total* number of misclassified objects which results in a balanced weighting of both classification errors. However, in business applications, the consideration of cost functions often requires unbalanced weighting of classification errors which should be incorporated in determining the threshold value c. For example we assume the following:

- To ensure the acceptance of the credit assessment inside the company, a maximum of 20% of the financially reliable customers should be wrongly refused.
- The measurement of the effect of credit analysis and its profit contribution is required. Therefore the resulting classification must be evaluated with regard to the varying costs caused by gaining and losing customers.

To incorporate the first constraint we define the β-error as the proportion of wrongly refused creditworthy applicants to all applicants refused: $\beta = h_{01}/h_{\cdot 1}$ and choose c on condition that $\beta \leq 0.2$ which results in the classification for the test sample given in Table 2 ($\beta = 0.1960$). For the second constraint, both possible misclassifications are associated with differing costs. A customer with negative credit standing that is assessed creditworthy causes loss and costs in the absence of acquittance or interest payments. Simultaneously expenses arise for dunning, encashment and processing. On the other hand, the wrong refusal of a credit application involves opportunity costs in the amount of lost assets and income from interest. It may also cause deterioration of customer relations. The classification rule for credit analysis should incorporate these varying costs to particular misclassification. In practice, at least the cost

Table 2. Classification for the training sample with $\beta \leq 0.2$

	$k(\mathbf{x}_i) = 1$	$k(\mathbf{x}_i) = 0$	$\sum$
$\text{TYPE}_i = 1$	1091 (h_{11})	575 (h_{10})	1666 ($h_{1.}$)
$\text{TYPE}_i = 0$	266 (h_{01})	2067 (h_{00})	2333 ($h_{0.}$)
$\sum$	1357 ($h_{.1}$)	2642 ($h_{.0}$)	3999 ($h_{..}$)

ratio must be known. Different asymmetric cost functions are conceivable. For example, the effect of the implementation of a credit assessment can be evaluated based on the expected increase in marginal return (mr) by analyzing the credit status of customer i (Bonne (2000)):

$$E(mr|\mathbf{x}_i) = C(1) \cdot P(\text{TYPE} = 1|\mathbf{x}_i) - C(0) \cdot (1 - P(\text{TYPE} = 1|\mathbf{x}_i)) \quad (2)$$

where $C(1)$ are costs and loss avoided by refusing customers that are not creditworthy, and C(0) measures the opportunity costs at the rate of lost profit by refusing customers that are creditworthy. $P(\text{TYPE} = 1|\mathbf{x}_i)$ is the conditional probability that the credit becomes deficient given the classification variables $\mathbf{x}_i$. Trading in high-valued consumer goods we usually suppose $C(1) > C(0)$, and in the analyzed case we assume a cost ratio of $C(1) : C(0) = 1 : 0.5$. Figure 1 shows the progression of the increase in marginal return according to the classification rules on basis of the linear predictor. The vertical line indicates the threshold value $\hat{y}_i \geq c$ under the restriction $\beta \leq 0.2$ ($\hat{y}_i \geq 0.4944$). Given the classification in Table 2 the increase in marginal return differs for varying cost ratios. It increases from 825 units for equal weighted costs to 958 at cost ratio 1:0.5. Fixing $E(mr|\mathbf{x}_i)$ to 825 on the other hand, the β-error decreases from 0.1960 to 0.1679 at cost ratio 1:0.5 as a lower refusal rate occurs.

In implementing global statistical models for credit analysis, a major difficulty arises. Often these models are not able to incorporate the heterogeneous composition of the population. For example in a certain application the estimated global model based on logistic regression rejects all applicants younger than 25 years as this attribute is highly significant concerning the customers' financial unreliability. In contrast to this we should request the credit assessment to filter out those applicants among the youngest customers which actually are solvent. We therefore suggest local scoring models to incorporate the heterogeneity of customer behavior in the underlying population.

4 Local scoring by two-stage classification

The suggested local scoring model consists of two steps. First unsupervised clustering algorithms are used to identify customer groups of homogeneous behavior within the classification variables. For this purpose we compare a self-organizing map, whose objects are afterwards clustered by the Ward algorithm, with a k-means cluster analysis. The second step consists in estimating

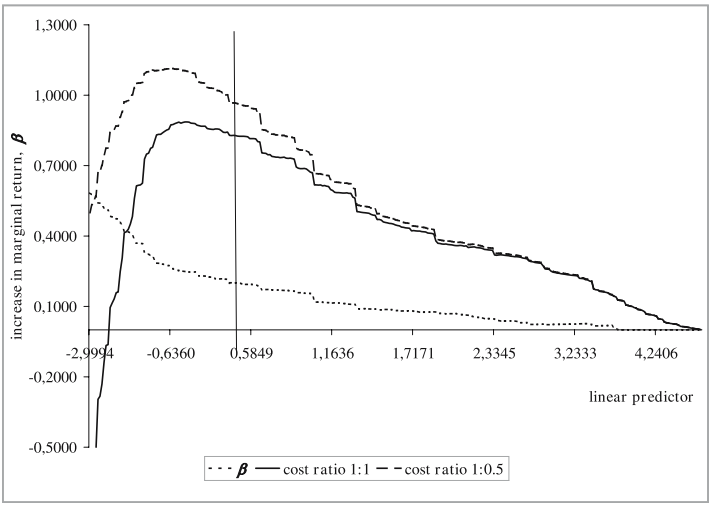


Fig. 1. Increase in marginal return (in thousand units) for different cost ratios

a model for predicting solvency based on logistic regression, as described in section 3, for every identified cluster separately.

4.1 Clustering using self-organizing maps

The self-organizing maps algorithm introduced by Kohonen (1982) is a powerful method for modelling customer behavior (Schmitt and Deboeck (1998)). Nevertheless, Kohonen maps are still rarely used in business applications. Self-organizing maps (SOM) show some specific characteristics among unsupervised learning algorithms. In particular they reduce a high dimensional feature space by transforming it onto a usually two-dimensional layer. Simultaneously SOM preserve the topology and structure of the input space which is basically characterized by the density of the input data and the ordering relation between input vectors. Below only a brief description concerning the application of this method according to Hastie et al. (2001) and Kohonen (1998) is given. For a detailed description concerning SOM and the Kohonen algorithm we refer to Kohonen (1982, 1984, and 1995).

A Kohonen self-organizing map (SOM) is an unsupervised artificial neural network that adapts itself responding to input signals based on the Kohonen algorithm. A SOM consists of K prototypes $\mathbf{m}_j,\ j = 1, \ldots, K$, which are initialized laying a uniform spread over a two-dimensional grid. Each of the K prototypes is parametrized with respect to a pair of integer coordinates $\mathbf{l}_j \in Q_1 \times Q_2$ with dimensions $Q_1 = \{1, 2, \ldots, q_1\}$ and $Q_2 = \{1, 2, \ldots, q_2\}$ which leads to the the map's size $K = q_1 \cdot q_2$. The sensible choice of q_1, q_2, and K respectively is left to the user. During the training process the observations $\mathbf{x}_i$ are presented to the map one at a time. Usually a certain number of

training cycles is carried out in which all the $\mathbf{x}_i$ are reapplied in succession. Presenting an observation $\mathbf{x}_i$ to the map we find the closest prototype $\mathbf{m}_j$ to $\mathbf{x}_i$ using a distance measure, e.g. the Euclidean norm. Then all neighbors $\mathbf{m}_k$ of this winning prototype $\mathbf{m}_j$ are moved closer to the data $\mathbf{x}_i$ via the update step

$$\mathbf{m}_k \leftarrow \mathbf{m}_k + \alpha h \left(\|\mathbf{l}_j - \mathbf{l}_k\|\right)(\mathbf{x}_i - \mathbf{m}_k) \tag{3}$$

with learning rate α and neighborhood function h. Typically α decreases during the training of the map from 1.0 to 0.0. Often a Gaussian neighborhood function is used (Kohonen (1998), Poddig and Sidorovitch (2001)):

$$h = \exp\left(-\|\mathbf{l}_j - \mathbf{l}_k\|^2 / 2\sigma^2\right) \tag{4}$$

with again σ continuously decreasing during the iterations. After completion of the training process, every observation vector $\mathbf{x}_i$ can be associated with the prototype it is mapped on according to the chosen distance measure, e.g. the Euclidean norm. This recall procedure can also be applied to accessory data sets of new customers which nevertheless should be part of the dynamic training process to achieve an adaptive and flexible map over time. We train a SOM with a large number of prototypes (1848) using the standardized input vectors of the training sample. Subsequently the prototypes of the resulting map are clustered by the hierarchical Ward cluster algorithm (Bacher (1996)).

4.2 K-means cluster analysis

K-means cluster analysis is used as a reference procedure for the local detection of underlying structure in customer behavior. SOM and k-means clustering are closely related to each other. If the neighborhood distance is chosen small enough so that each neighborhood contains only one prototype the SOM algorithm reduces to a dynamic version of k-means clustering (Hastie et al. (2001)). In addition the size of the Kohonen map, i.e. the number of prototypes, can be chosen as the determined number of k-means clusters (Poddig and Sidorovitch (2001)). As starting partitions in the k-means algorithm (Bacher (1996), Hastie et al. (2001)) we use the outcome of a Ward clustering. The initial cluster centroids are chosen at random and again all input vectors are applied standardized.

4.3 Evaluation of two-stage classification

Both procedures for identifying classes of homogeneous behavior detect four clusters. Within the particular local model every customer is first assigned to one of the detected clusters. Afterwards a local logistic regression model (cf. section 3) for classification and predicting the customer's solvency is estimated for each cluster separately. This involves the separate determination of the threshold value c with respect to an overall classification at optimal costs and an overall $\beta \leq 0.2$. The resulting classifications for the local models

Table 3. Classification for the training sample using local scoring models

k-means model

	$k(\mathbf{x}_i) = 1$	$k(\mathbf{x}_i) = 0$	$\sum$
$\text{TYPE}_i = 1$	959	707	1666
$\text{TYPE}_i = 0$	194	2139	2333
$\sum$	1153	2846	3999

SOM model

	$k(\mathbf{x}_i) = 1$	$k(\mathbf{x}_i) = 0$	$\sum$
$\text{TYPE}_i = 1$	1001	665	1666
$\text{TYPE}_i = 0$	149	2184	2333
$\sum$	1150	2849	3999

are given in Table 3. Showing an almost equal refusal rate (28.83% in the k-means model to 28.76% using SOM) the β-error for the local scoring model with SOM shows a much lower value ($\beta = 0.1296$) compared to the model using k-means clustering ($\beta = 0.1683$). By applying k-means clustering for structure detection the increase in marginal return for this classification is 765 units at cost ratio 1:1 and 862 at cost ratio 1:0.5. Using SOM clusters raises these results to 852 and 927 units.

5 Application to the test sample

In this section the developed local scoring models and the global model are evaluated concerning their performance in the test sample and their generalization ability. The particular threshold values are fixed according to the classification results in the training sample (cf. Table 1). The final classification results at cost ratio 1:0.5 are given in Table 4. The global scoring model shows a refusal rate of 31.5% of all applicants with an increase in marginal return of 430 units. But $\beta = 0.2019$ exceeds the value of maximal 20% wrongly refused applicants. Using the local k-means model for solvency prediction under the same constraints we achieve a lower increase in marginal return (400 units) accompanied by an acceptable β-error of 0.1798 at a refusal rate of 27.3%. Finally, the SOM model shows a refusal rate of 28.6% raising the increase in marginal return to 460 units. Simultaneously the β-error for this local model decreases to 0.1331.

Table 4. Performance of global, k-means and SOM model in the test sample

	global model		k-means model		SOM model	
	$k(\mathbf{x}_i) = 1$	$k(\mathbf{x}_i) = 0$	$k(\mathbf{x}_i) = 1$	$k(\mathbf{x}_i) = 0$	$k(\mathbf{x}_i) = 1$	$k(\mathbf{x}_i) = 0$
$\text{TYPE}_i = 1$ (810)	502	308	447	363	495	315
$\text{TYPE}_i = 0$ (1186)	127	1059	98	1088	76	1110
$\sum$ (1996)	629	1367	545	1451	571	1425

6 Conclusions

With respect to the regarded purposes and the resulting constraints the local statistical model using SOM clusters is superior to both the global and the local k-means model. By means of clustering a self-organizing map with a relatively large number of prototypes we achieve a segmentation into customer groups of homogeneous behavior. Referring to the underlying classification problem this leads to a fundamental reduction of misclassified objects, especially with regard to the number of wrongly refused creditworthy customers and the associated loss of profit.

Concerning the application of the self-organizing map methodology in credit management this is an ongoing work mainly depending on two important conditions. First the misclassification costs, or even their ratio to each other, must be known. Second for a sustainable measurement of dynamics and heterogeneity of customer behavior a regular revision and adjustment of the used model is required. Therefore training samples that provide customer information along several years, especially including the exact time of credit application, are needed.

References

ARMINGER, G. et al. (1997): Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Networks. *Computational Statistics, 12, 293–310.*

BACHER, J. (1996): *Clusteranalyse.* 2. Aufl. Oldenbourg, München.

BONNE, T. (2000): *Kostenorientierte Klassifikationsanalyse.* Eul, Lohmar.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *The Elements of Statistical Learning.* Springer, New York.

KOHONEN, T. (1982): Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43, 59–69.*

KOHONEN, T. (1984): *Self-Organization and Associative Memory.* Springer, Berlin.

KOHONEN, T. (1995): *Self-Organizing Maps.* Springer, Berlin.

KOHONEN, T. (1998): The SOM Methodology. In: G. Deboeck and T. Kohonen (Eds.): *Visual Explorations in Finance with Self-Organizing Maps.* Springer, London, 159–167.

PODDIG, T. and SIDOROVITCH, I. (2001): Künstliche Neuronale Netze: Überblick, Einsatzmöglichkeiten und Anwendungsprobleme. In: H. Hippner et al. (Eds.): *Handbuch Data Mining im Marketing.* Vieweg, Braunschweig, 363–402.

SCHMITT, B. and DEBOECK, G. (1998): Differential Patterns in Consumer Purchase Preferences using Self-Organizing Maps: A Case Study of China. In: G. Deboeck. and T. Kohonen (Eds.): *Visual Explorations in Finance with Self-Organizing Maps.* Springer, London, 141–157.

Informative Patterns for Credit Scoring: Support Vector Machines Preselect Data Subsets for Linear Discriminant Analysis

Ralf Stecking and Klaus B. Schebesch

Institut für Konjunktur- und Strukturforschung,
Universität Bremen, D-28359 Bremen, Germany

Abstract. Pertinent statistical methods for credit scoring can be very simple like e.g. linear discriminant analysis (LDA) or more sophisticated like e.g. support vector machines (SVM). There is mounting evidence of the consistent superiority of SVM over LDA or related methods on real world credit scoring problems. Methods like LDA are preferred by practitioners owing to the simplicity of the resulting decision function and owing to the ease of interpreting single input variables. Can one productively combine SVM and simpler methods? To this end, we use SVM as the preselection method. This subset preselection results in a final classification performance consistently above that of the simple methods used on the entire data.

1 Introduction

Credit scoring basically relies on a binary classification problem, which helps forecasting whether a credit applicant will default during the contract period. A credit applicant is described by many characteristics, resulting in a high dimensional data vector, which is labeled according to whether defaulting or non defaulting behavior was observed. Such labeled data vectors are used for determining a classification rule, usually by a method of statistical learning. The probable behavior of new credit applicants can then be read off by using the classification rule on their (as yet unlabeled) data.

Even for a largely unknown geometry of high dimensional (labeled) data, Support Vector Machines (SVM) are a very powerful statistical learning method for determining a successful classification rule (Schölkopf and Smola (2002)). In fact, for real life credit scoring data, Stecking and Schebesch (2003), Schebesch and Stecking (2003a) and Schebesch and Stecking (2003b) show that general SVM are superior to Linear Discriminant Analysis (LDA) and Logistic Regression in terms of out-of-sample prediction accuracy, both in the standard situation, but also under various more realistic circumstances, including unequal number of class representatives and asymmetric misclassification costs. Other recent work on credit scoring and SVM is Friedman (2002), Van Gestel et al. (2003) and Huang et al. (2004). From the practitioners point of view, a drawback of SVM is the need of a validation-parameterization cycle, more typical for testing new scientific methods but

less useful in routine applications with time constraints. Another observation of ours is that non-linear SVM applied to credit scoring data are not leading to an overwhelming improvement of the error rate when compared to linear SVM, probably owing to the relative sparsity of the data. However, the much simpler linear SVM is still very different from the traditionally used LDA, as is shown in this paper. Unlike LDA but similar to non-linear SVM, linear SVM uses margin-maximization of the class boundaries, which leads to a more robust separation rule in presence of noise and outliers. The very compact description of the class boundaries by some of the data points (i.e. certain support vectors of the linear SVM) can be used as a training subset preselection for LDA. The practitioner can then use LDA but also benefits from the more powerful results of the SVM. In this paper we show that the compact description of the class boundaries by certain support vectors of the linear SVM is a useful expert "data model" for credit scoring: pooled or proprietary data sets need not be revealed to a decentralized user (practitioner) who will employ traditional methods like LDA but who will benefit from the results of the SVM.

2 Linear SVM and LDA

A classification model for credit scoring uses $N > 0$ labeled training examples $\{x_i, y_i\}$, with vector $x_i \in \mathcal{R}^d$ describing credit applicant i by $d \gg 1$ characteristics, and with class label y_i indicating whether credit applicant i was "bad" (defaulting, $y_i = +1$) or "good" (non-defaulting, $y_i = -1$) during a contract period. A detailed description of the SVM classifier for such data is Schölkopf and Smola (2002) or Stecking and Schebesch (2003). Here we just state the optimization problem of the **linear SVM**. Figure 1 depicts margin maximization between the classes, for a stylized situation, where classes are linearly separable ($\langle .,. \rangle$ denotes scalar products). If all positive cases (i.e. x_i with $y_i = 1$) are positioned such that $\langle x_i, w \rangle + b \geq 1$ is verified (and $\langle x_i, w \rangle + b \leq -1$ for all negative cases), then linear separation by some hyperplane $\langle w, x \rangle + b = 0$ is possible. Owing to the fact that real life data can be both non-linear and noisy, maximizing the margin (or minimizing $||w||$) is now done allowing for some misclassification by means of slacks $\zeta \geq 0$ and parameter $C > 0$:

$$\min_{w,b,\zeta} C\sum_{i=1}^{N}\zeta_i + \frac{1}{2}\langle w, w\rangle \quad \text{s.t.} \quad y_i[\langle x_i, w\rangle + b] \geq 1 - \zeta_i, \;\; \zeta_i \geq 0.$$

The associated dual used for numerical computations reads

$$\max_{\alpha} \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle \quad \text{s.t.} \quad \sum_{i=1}^{N} y_i\alpha_i = 0, \quad C \geq \alpha_i \geq 0.$$

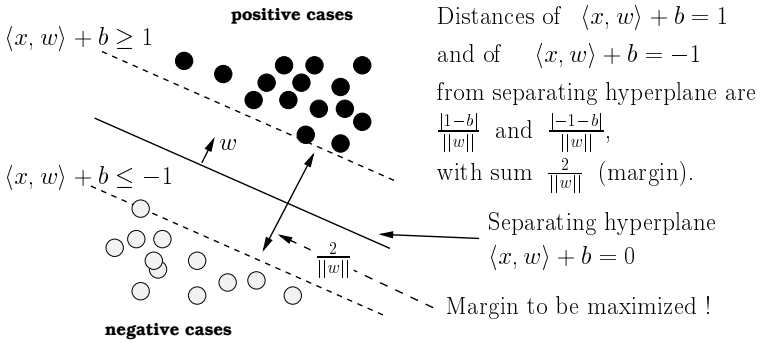


Fig. 1. Linear SVM separates classes and maximizes the margin.

The solution of the dual leads to an optimal decision rule for new cases x

$$y^*(x) \;=\; \mathbf{sign}\left\{ \sum_{i=1}^{N} \alpha_i^* y_i \langle x, x_i\rangle + b^* \right\}.$$

Take $N > d$ as is the case in credit scoring data. The $N^s \leq N$ **support vectors** are the set $\{x_i|\alpha_i^* > 0\}$ and the **"informative" support vectors** are the subset $\{x_i|0 < \alpha_i^* < C\}$, which delimit regions of reliable class predictability. The remaining support vectors are **critical**, i.e., they fall into the region, where predictability is not reliable. By varying C one can obtain at least $d + 1$ informative support vectors. They can be used as data input to a simple standard method like Linear Discriminant Analysis (LDA). However, LDA encounters singularity of the covariance matrix when using exactly $d+1$ data points in d dimensions. To circumvent problems of computation one can use the pseudo-inverse (Shashua (1999)). We choose to simply duplicate the labeled support vectors and randomly perturb the location of the replicated points in input space such that their maximum displacement is much smaller than the minimum distance between two empirical data points. LDA on this data approximates the SVM separation quite well. For linearly separable data, the separating hyperplane of the LDA on the support vectors and of the linear SVM are very similar and theoretically coincide when the class-wise data distributions are the same (Shashua (1999)). A more interesting situation is given if classes are not linearly separable, with effective misclassification of the linear SVM (i.e. $d + 1 < N^s < N$), as we also find in our real world credit scoring data. Figure 2 depicts the differences between using LDA and linear SVM (or a LDA on informative support vectors) on some two dimensional data, which are not linearly separable. Depending on the priors for the two classes, LDA would turn out a class separating line from the shaded area of the left column plots. Even for the first data example (upper row), the linear SVM selects the informative support vectors such that the slope of the linear discriminant on these data points differs from the LDA

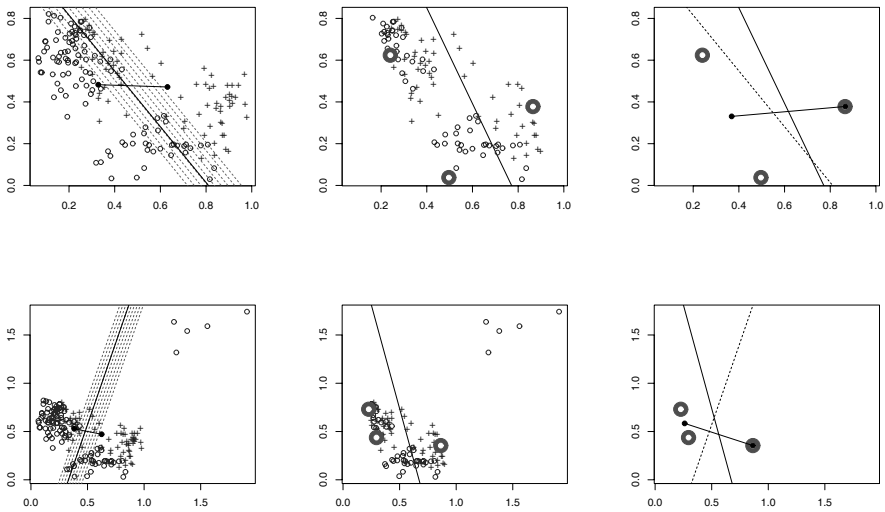


Fig. 2. Two examples of nonlinear binary classification problems (by row), with two dimensional inputs. Class labels of example data points are marked with circles with crosses respectively (left column plots). Separation by linear SVM (middle column plots): informative support vectors (big circles) and critical support vectors (small circles and crosses). Separation slopes obtained by SVM+LDA (long solid line) and that of LDA alone (long dashed line) are compared in the right column plots. The shorter lines in the left and right column plots connect the respective class means of informative support vectors.

on the original data (right column plot). The more extreme example (lower row) is formed by taking the first example with "outliers opposing the linear separation" added in the upper right corner (the new five distant points). Owing to the strong influence of these outliers on the class centers, the LDA is now tilting the slope of the separating line completely, while the linear SVM still separates the "bulk" of the linearly separable data more faithfully. The linear SVM is including the outliers into the set of critical support vectors (middle plot of lower row). Besides being high dimensional, our real life credit scoring data may exhibit distributional properties somewhere in between the presented examples. When classifying these data by SVM, quite many critical support vectors occur, indicating non linear class data (Stecking and Schebesch (2003), Schebesch and Stecking (2003a)). The rationale of using a SVM+LDA combination is the following: Imagine a data-pooling center which is collecting credit data from many financial firms, as presently practiced with credit card fraud data. The proprietary data should be hidden from the constituent members of the pool, but a data model (not revealing all data) can be forwarded. The support vectors of a linear or non-linear SVM

would in fact be such a model, which hence would enable members to use a more conventional method (e.g. LDA on support vectors from a linear SVM) to obtain high quality decisions implicitly based on much more representative data.

3 Subset preselection for LDA: Empirical results

We use a sample of 658 cases describing applicants for building and loan credits. A total of 323 out of 658 applicants could be classified as "defaulting", depending on their individual historical credit performance. The 335 remaining applicants are "non defaulting", respectively. The number of input variables is 16 (7 metric and 9 categorical). Due to coding of categorical variables as indicator variables each input pattern finally consists of 40 dimensions. In the following it is shown how to use Linear Support Vector Machines to detect informative patterns for Linear Discriminant Analysis. Then, three models are built, using credit scoring data: *(i)* Linear Support Vector Machine (SVM), *(ii)* Linear Discriminant Analysis (LDA), and *(iii)* LDA with subset preselection (LDA-SP). The classification results of these three models are compared and, finally, some advantages of our procedure are discussed.

3.1 About typical and critical subsets

SVM divide the patterns of the input space into (1) *informative*, (2) *typical* and (3) *critical* subsets. *Informative patterns* are essential support vectors and will be used as input for Linear Discriminant Analysis. *Typical patterns* can be separated with low (or zero) error by the SVM. *Critical patterns* cannot be separated well by the SVM. The classification error usually is high.

In case of *labeled data* classifying each pattern as informative, typical or a critical can be done simply with regard to the α_i of the SVM (see table 1).

Table 1. Overview of informative, typical and critical patterns.

Informative, typical and critical patterns		
Subset	*with regard to Lagrange multipliers* α_i	*with regard to SVM output*
Informative patterns	$0 < \alpha_i < C$	$\lvert\langle x_i, w\rangle + b\rvert = 1$
Typical patterns	$\alpha_i = 0$	$\lvert\langle x_i, w\rangle + b\rvert > 1$
Critical patterns	$\alpha_i = C$	$0 < \lvert\langle x_i, w\rangle + b\rvert < 1$
False patterns	$\in \{\text{Critical patterns}\}$	$y_i^*(\langle x_i, w\rangle + b) < 0$

In case of *unlabeled data* only the output of the SVM is given. Let $\langle x_i, w\rangle + b$ be the output of the SVM. Then a support vector is defined as being located exactly on the margin, i.e. $|\langle x_i, w\rangle + b| = 1$. All typical patterns are outside the margin ($|\langle x_i, w\rangle + b| > 1$), and the critical patterns are located within the margin $|\langle x_i, w\rangle + b| < 1$. In practical use informative patterns will be extracted by using the α_i of the SVM. Whether an (unlabeled) credit applicant belongs to the typical or critical subset then will be decided w.r.t. the SVM output.

3.2 LDA with subset preselection

A linear SVM with $C = 100$ was built to predict the state of credit ("good: non defaulting" or "bad: defaulting") using the credit scoring data set (658 cases) with 40 dimensional input vector to describe the credit applicants. By using SVM with linear kernel, upper bound C is the only parameter to be set in advance. By choosing C too small one gets a huge amount of critical patterns which is not desirable. With growing C the distribution of the patterns in between the subsets becomes stable, not changing anymore at above some fixed level. In our example this is the case for $C = 100$. Table 2 shows the classification results of the LDA with subset preselection. There is low error within the typical subset and high error within the critical subset (cf. table 1).

Table 2. Classification results for informative, typical and critical subsets. Informative patterns are selected w.r.t to Lagrange multipliers α_i, typical and critical patterns w.r.t. to SVM output.

LDA with subset preselection (LDA-SP)							
		Informative		*Typical*		*Critical*	
		good	bad	good	bad	good	bad
Observed	good	24	0	155	17	85	54
	bad	0	17	22	147	56	81
Region error	in %	0.00		11.44		39.86	
Total error	in %			22.64			

3.3 Comparing SVM, LDA and LDA-SP

LDA with subset preselection (LDA-SP) leads to a decision rule, which is a simple linear discriminant function. Through subset selection on the other hand, some of the advantages of the SVM are given to the LDA: a tight subset of the input space, that is not disturbed by mass informations or by

outliers, which entails all information needed to construct a useful decision rule. Furthermore, it is possible to label a credit applicant as typical or critical, indicating excellent or poor classification capability. But is SVM-LDA really competitive to SVM (and also to LDA)? One would assume, that the performance of the combined approach lies between the ones of SVM and LDA. Especially mass information of the full data set biases the classification function of the LDA away from the true boundary between good and bad credit applicants (Stecking and Schebesch (2003)). Table 3 shows intermediate performance (compared to SVM and LDA) of the LDA with subset preselection.

Table 3. Classification results for SVM, LDA and LDA-SP.

		SVM		*LDA*		*LDA-SP*	
		good	bad	good	bad	good	bad
Observed	good	259	76	247	88	264	71
	bad	71	252	69	254	78	245
Model error	in %	22.34		23.86		22.64	

3.4 Advantages of LDA with subset preselection

Subset preselection can be used in multiple ways: Using the informative patterns as input for LDA leads to a very simple and comprehensive back-end function, as simple as, but superior to the traditional LDA. Subset preselection divides credit applicants into the subsets of "critical" and "typical" patterns. The predicted classification of a typical pattern is much more reliable than that of a critical pattern. Consequently, the ratio of critical to typical credit applicants can hint at the reliability of the classification model given the data, and it can be used as a possible measure for the "predictability" of the whole data set. Furthermore, if you profile and compare typical good versus typical bad applicants it is easy to focus on the differences between both groups. Finally, as stated above, outlier and data error detection is also possible with subset preselection.

4 Conclusions

Searching for classification methods best suitable for credit scoring, we backtrack in this paper to using SVM with linear kernels. The informative support vectors of the linear SVM are stable data models of the original data. They can be used as compact but encompassing data input to more conventional

models like LDA by practitioners. For data which are not linearly separable (as are e.g. credit scoring data), such preselected data (i.e. informative support vectors) also lead to a linear separation which may differ considerably from that of directly applying LDA to the original data. Linear SVM more faithfully separates the "bulk" of the linearly separable data, being much less influenced by outliers which "oppose a linear separation". For real life credit scoring data it is shown that SVM can be successfully used to detect three kinds of subsets: informative, typical and critical. By estimating the coefficients of the LDA with the small number of informative patterns (support vectors) as the only inputs, a credit scoring performance is obtained, which usually lies near that of SVM or between that of SVM and LDA on original data. Furthermore, SVM can be easily adjusted to non standard situations most common to credit scoring, like the vastly different numbers of good and bad credit applicants and different misclassification costs (Schebesch and Stecking (2003a)). Such strong asymmetries increase the benefit of building a linear SVM by an expert modeler even further, but they still enable the faithful use of LDA as a backend by the practitioner.

References

FRIEDMAN, C. (2002): CreditModel Technical White Paper, *Standard & Poors Risk Solutions*, New York.

HUANG, Z., CHEN, H., HSU, C.-J., CHEN, W.-H. and WU, S. (2004): Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study. *Decision Support Systems (DSS), 37(4), 543–558.*

SCHEBESCH, K.B. and STECKING, R. (2003a): Support Vector Machines for Credit Scoring: Extension to Non Standard Cases. Submitted to Proceedings of the 27th Annual Conference of the GfKl 2003.

SCHEBESCH, K.B. and STECKING, R. (2003b): Support Vector Machines for Credit Applicants: Detecting Typical and Critical Regions. *Credit Scoring & Credit Control VIII*, Credit Research Center, University of Edinburgh, 3–5 September 2003, 13pp.

SCHÖLKOPF, B. and SMOLA, A. (2002): *Learning with Kernels.* The MIT Press, Cambridge.

SHASHUA, A. (1999): On the Relationship Between the Support Vector Machine for Classification and Sparsified Fisher's Linear Discriminant. *Neural Processing Letters 9*, 129–139.

STECKING, R. and SCHEBESCH, K.B. (2003): Support Vector Machines for Credit Scoring: Comparing to and Combining with some Traditional Classification Methods. In: M. Schader, W. Gaul and M. Vichi (Eds.): *Between Data Science and Applied Data Analysis.* Springer, Berlin, 604–612.

VAN GESTEL, T., BARSENS, B., GARCIA, J. and VAN DIJKE, P. (2003): A Support Vector Machine Approach to Credit Scoring. `http://www.defaultrisk.com/pp_score_25.htm`, 15pp.

Application of Support Vector Machines in a Life Assurance Environment

Sarel J. Steel[1] and Gertrud K. Hechter[2]

[1] Department of Statistics and Actuarial Science, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa
[2] SANLAM, P.O. Box 206, Sanlamhof, 7532, South Africa

Abstract. Since its introduction in Boser et al. (1992), the support vector machine has become a popular tool in a variety of classification and regression applications. In this paper we compare support vector machines and several more traditional statistical classification techniques when these techniques are applied to data from a life assurance environment. A measure proposed by Louw and Steel (2004) for ranking the input variables in a kernel method application is also applied to the data. We find that support vector machines are superior in terms of generalisation error to the traditional techniques, and that the information provided by the proposed measure of input variable importance can be utilised for reducing the number of input variables.

1 Introduction

Since the introduction of the support vector machine by Boser et al. (1992), this technique has become a popular tool for classification and regression. Although it was initially mainly applied in the machine learning community, the support vector machine is rapidly becoming a standard option for solving statistical problems (see for example Hastie et al. (2001) for a discussion of support vector machines from a statistical perspective). In this paper we report a study of support vector machines applied in a life assurance environment. Data miners and statisticians at Sanlam, a major South African financial services company, currently use standard statistical techniques to classify new policy applicants into one of two classes, viz. clients that will in future lapse one or more policies, and those that will not do so. The aim of our study was twofold: firstly, to determine whether support vector machines are capable of improving upon the standard techniques in terms of accurate classification, and secondly, to apply a measure of variable importance proposed by Louw and Steel (2004) to the data to rank the input variables in terms of their ability to separate the two groups when using a support vector machine.

The second aim deserves more comment. Although support vector machines typically classify very accurately, they do not provide a natural way of determining the relative importance of the different input variables. Ranking the input variables according to their importance serves several purposes: it

leads to better insight into the structure of the problem or data set, we obtain a parsimonious and sensible summary of the input variables, it is more cost-effective to work with fewer input variables, and using fewer input variables when classifying future cases may actually lead to more accurate results. See in this regard Guyon and Elisseeff (2003). In Section 4 of this paper we therefore follow Louw and Steel (2004) and show how the concept of alignment introduced by Cristianini et al. (2002) can be used to apply a new measure of input variable importance in a support vector machine context.

Regarding the remainder of this paper, in Section 2 we provide a very brief overview of support vector machine methodology. Section 3 contains a description of the problem context and the data. Section 4 introduces the new measure of variable importance, and illustrates its application to the data under consideration. The results of the comparative study are presented and discussed in Section 5.

2 Support vector machines

Support vector machines are sufficiently well known to make an extensive discussion of the underlying theory in this paper unnecessary. We therefore limit ourselves to a brief overview of the topic, introducing required notation along the way. Detailed discussions of the theory of support vector machines for classification and regression can be found in, amongst others, Schölkopf and Smola (2002).

Consider the following generic two-group classification problem. We observe a binary response variable $Y \in \{-1, +1\}$, together with classification or input variables $X_1, X_2, \cdots, X_p$. These variables are observed for $N = N_1 + N_2$ sample cases, with the first N_1 cases coming from population 1 and the remaining N_2 cases from population 2. The resulting training data set is therefore $\{(\boldsymbol{x_i}, y_i), i = 1, 2, \cdots, N\}$. Here, $\boldsymbol{x_i}$ is a p-component vector representing the values of $X_1, X_2, \cdots, X_p$ for case i in the sample. Our purpose is to use the training data to determine a rule that can be used to assign a new case with observed values of the predictor variables in a vector $\boldsymbol{x}$ to one of the two populations.

Application of a support vector machine to a given data set entails implicit transformation of the input data to a high dimensional feature space, followed by construction of a decision function for classification of future cases by fitting a hyperplane to the transformed data. Let Φ denote the transformation from input to feature space. Then the support vector machine classification function for a new case with input vector $\boldsymbol{x}$ is given by $sign\left\{b + \sum_{i=1}^{N} \alpha_i y_i \langle \Phi(\boldsymbol{x_i}), \Phi(\boldsymbol{x}) \rangle\right\}$. Here, b and $\alpha_1, \alpha_2, \cdots, \alpha_N$ are quantities determined by applying the support vector machine algorithm to the training data, while $\langle \Phi(\boldsymbol{x_i}), \Phi(\boldsymbol{x}) \rangle$ denotes the inner product between the (possibly infinite dimensional) feature vectors $\Phi(\boldsymbol{x_i})$ and $\Phi(\boldsymbol{x})$.

The so-called kernel trick is an essential element of support vector machine methodology in that it obviates explicit calculations in the feature space. This stratagem implies that the inner product $\langle \boldsymbol{\Phi}(\boldsymbol{x_i}), \boldsymbol{\Phi}(\boldsymbol{x}) \rangle$ can be replaced by $K(\boldsymbol{x_i}, \boldsymbol{x})$, where $K(.,.)$ is an appropriate kernel function. The SVM classification function therefore becomes $sign\left\{b + \sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x})\right\}$. Examples of popular kernel functions are the homogeneous polynomial kernel, $K(\boldsymbol{x_1}, \boldsymbol{x_2}) = \langle \boldsymbol{x_1}, \boldsymbol{x_2} \rangle^d$, where d is an integer, usually 2 or 3, the inhomogeneous polynomial kernel, $K(\boldsymbol{x_1}, \boldsymbol{x_2}) = (c + \langle \boldsymbol{x_1}, \boldsymbol{x_2} \rangle)^d$, with c a positive constant, and the Gaussian or radial basis function (RBF) kernel given by $K(\boldsymbol{x_1}, \boldsymbol{x_2}) = \exp(-\gamma \|\boldsymbol{x_1} - \boldsymbol{x_2}\|^2)$, where γ is a so-called kernel hyperparameter, and $\|\boldsymbol{x}\|^2 = \langle \boldsymbol{x}, \boldsymbol{x} \rangle = \sum_{i=1}^{p} x_i^2$. We restrict attention to the RBF kernel in the remainder of this paper. For more detailed discussions of kernel functions, see for example Herbrich (2001) or Schölkopf and Smola (2002).

Finally in this section, note that the quantities b and $\alpha_1, \alpha_2, \cdots, \alpha_N$ in the support vector machine decision function are typically found by solving a quadratic optimization problem. The objective function that is maximized during this optimization also contains a constant C, called a cost parameter, that guards against overfitting and consequent poor generalization ability of the support vector machine. Both the cost parameter C and the kernel hyperparameter γ have to be specified by the user from prior knowledge of the problem area, or determined from the training data. This is an issue that will arise again later in our discussion.

3 Problem context and the data

In the life assurance industry (as in most industries) client retention is very important. One aspect of client retention in this industry is the issue of early policy termination, in other words, lapses and surrenders. In essence a policy lapse usually means that there is a financial loss for the company, as the policy is terminated before the end of the period in which costs are recouped. Determining or predicting whether a client will lapse a policy is therefore very important for profitability, and it is crucial for the company to be able to predict in advance whether a client is likely to lapse a policy.

Several standard techniques are currently used at Sanlam to evaluate the risk of a policy lapsing. The aim of our study was to investigate the possible use of support vector machines in this context.

Regarding the data that were analysed, a random sample of 4851 policy lapse cases was selected from the client database. Another random sample of 4745 non-lapse cases was added, yielding a total sample size of 9596. The response (dependent) variable was an indicator of whether the person had lapsed a policy or not. There were 61 input (independent) variables, consisting of numerical and categorical variables. The categorical input variables were handled by introducing appropriately defined indicator variables.

The techniques that were applied to the data are Fisher's linear discriminant analysis (LDA), logistic regression, classification trees and support vector machines (SVMs). Software from R were used to perform the data analysis. The analysis included the following steps. The available data were randomly divided into a training set (typically, 75% of the total) and a test set (the remaining 25%). Each of the techniques mentioned above was applied to the training data, thereby obtaining four classification functions that were used to predict the lapse category of the data cases in the test data set. A test error was calculated for each technique as the percentage of test set misclassifications. The random split of the data into training and test data subsets was repeated 100 times, and an average test error was calculated for each of the techniques under investigation.

4 A measure of variable importance

An important property of support vector machines is that the input vectors $\boldsymbol{x_i}$ appear in the algorithm only as arguments of the kernel function, i.e. we encounter these vectors only in the form $K(\boldsymbol{x_i}, \boldsymbol{x_j}), i, j = 1, 2, \cdots, N$. Evaluating $K(\boldsymbol{x_i}, \boldsymbol{x_j})$ for $i, j = 1, 2, \cdots, N$, we are able to construct the so-called Gram matrix with ij-th entry $K(\boldsymbol{x_i}, \boldsymbol{x_j})$. When a support vector machine is applied to a two-group classification problem, the Gram matrix contains all the information provided by the input vectors $\boldsymbol{x_i}$. Since $K(\boldsymbol{x_i}, \boldsymbol{x_j})$ can be interpreted as a measure of the similarity between $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$, Cristianini et al. (2002) argue that an ideal Gram matrix would be of the form $\boldsymbol{yy'}$, where $\boldsymbol{y}$ is the N-component response input vector with -1 in the first N_1 positions and +1 in the remaining N_2 positions. They define the concept of (empirical) alignment between a given Gram matrix $\boldsymbol{G} = [K(\boldsymbol{x_i}, \boldsymbol{x_j})]$ and the ideal Gram matrix $\boldsymbol{yy'}$ by

$$A(\boldsymbol{G}, \boldsymbol{yy'}) = \frac{\langle \boldsymbol{G}, \boldsymbol{yy'} \rangle_F}{\sqrt{\langle \boldsymbol{G}, \boldsymbol{G} \rangle_F \langle \boldsymbol{yy'}, \boldsymbol{yy'} \rangle_F}} \ ,$$

where $\langle \boldsymbol{R}, \boldsymbol{S} \rangle_F = trace(\boldsymbol{RS})$ is the Frobenius inner product between the symmetric matrices $\boldsymbol{R}$ and $\boldsymbol{S}$. These authors investigate the properties of the alignment, the most important for our purpose being that a large value of the alignment is desirable, since this will typically lead to the support vector machine generalizing well, i.e. classifying new cases accurately.

Louw and Steel (2004) use the concept of alignment to define a quantity that reflects the importance of an input variable when fitting a support vector machine to a given data set. Consider in this regard the RBF kernel, and let $K_r(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp[-\gamma(x_{ir} - x_{jr})^2]$ with corresponding Gram matrix $\boldsymbol{G}_r, r = 1, 2, \cdots, p$. These are therefore the Gram matrices obtained by evaluating the kernel function on a single coordinate of the input vectors at a time. Louw and Steel (2004) suggest measuring the importance of variable X_r in terms of the alignment of $\boldsymbol{G}_r$ with the ideal Gram matrix $\boldsymbol{yy'}$, i.e. by calculating $A(\boldsymbol{G}_r, \boldsymbol{yy'})$. A large value of $A(\boldsymbol{G}_r, \boldsymbol{yy'})$ would imply that X_r

is an important input variable in the sense that it contributes significantly to separating the two populations under consideration. Whilst it may be difficult to quantify exactly what is meant by a large value of $A(\boldsymbol{G}_r, \boldsymbol{yy'})$ in this context, and further research is required on this aspect, it is clear that the values of $A(\boldsymbol{G}_r, \boldsymbol{yy'})$, for $r = 1, 2, \cdots, p$, can easily be used to rank the input variables in order of importance.

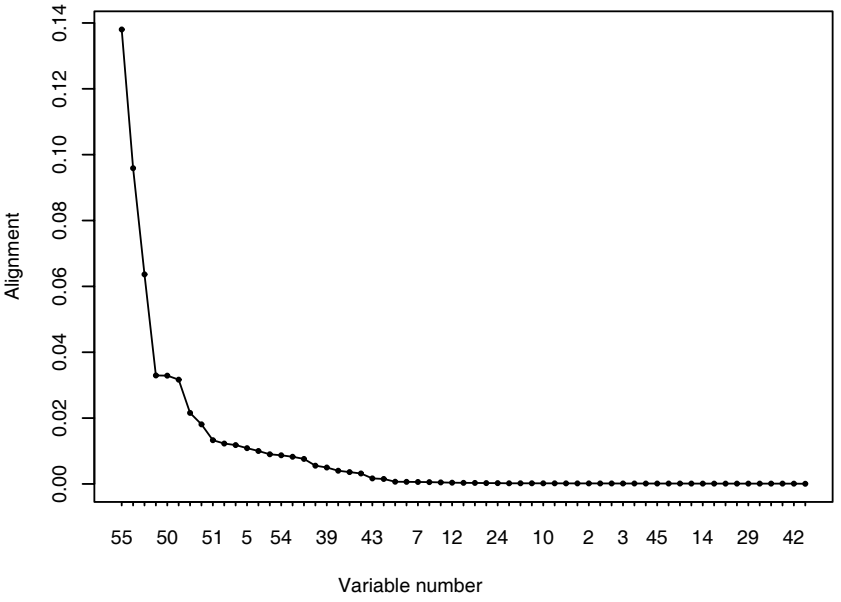


Fig. 1. Scree plot of ranked alignment values

There are several points deserving further attention that have to be made regarding the proposal to use $A(\boldsymbol{G}_r, \boldsymbol{yy'})$ as a measure of individual variable importance. Firstly, $A(\boldsymbol{G}_r, \boldsymbol{yy'})$ depends on the values of the kernel function hyperparameters. For the RBF kernel there is only a single hyperparameter, viz. γ. A decision therefore has to be made regarding the value of γ to use when calculating $A(\boldsymbol{G}_r, \boldsymbol{yy'})$. Louw and Steel (2004) report empirical evidence in favour of using a fixed value of γ, for example $\gamma = 1$. Although this seems to work well in the examples that they consider, other strategies to deal with this aspect clearly also need to be investigated. In our application of alignment to rank the variables in the data set we used $\gamma = 1$. A second

point regarding the proposed measure of variable importance concerns the possibility of using other more well known measures than $A(\boldsymbol{G}_r, \boldsymbol{yy}')$ for this purpose, for example the correlations between the input variables and the response. In this regard it should be borne in mind that by using a kernel function one is able to exploit highly non-linear relationships between the input variables and the binary response. It seems that a measure such as $A(\boldsymbol{G}_r, \boldsymbol{yy}')$ is able to capture such non-linear relationships, something which will be difficult if instead we calculate quantities such as the correlations between the independent variables and the response. A final question that deserves consideration is whether $A(\boldsymbol{G}_r, \boldsymbol{yy}')$ can be used for effective dimensionality reduction. This would of course have the advantage that only a subset of the original input variables need to be used in further analyses and it may even lead to better classification performance of the resulting rule. The crucial issue in this regard is how to decide on the number of input variables to retain. This question is similar to the problem of deciding on the number of principal components or factors to use when performing a principal component or factor analysis. One strategy could be to use a scree plot of the successive alignment values, a possibility that we explored for the Sanlam data set. Figure 1 shows the scree plot that was obtained by plotting the ranked $A(\boldsymbol{G}_r, \boldsymbol{yy}')$ values against the input variable numbers. Several competing subsets of input variables to be used in the classification of future cases can be identified from Figure 1, albeit somewhat subjectively. We therefore investigated the classification accuracy (using the procedure described in Section 3) of the competing techniques in our comparative study for the following subsets (identified here simply in terms of the variable numbers), and report on the results in the next section:

$A = \{55, 53, 47, 50, 30\}$
$B = \{55, 53, 47, 50, 30, 35, 40\}$
$C = \{55, 53, 47, 50, 30, 35, 40, 51, 52, 37, 5, 6, 34, 54, 15, 31\}$
$D = \{55, 53, 47, 50, 30, 35, 40, 51, 52, 37, 5, 6, 34, 54, 15, 31, 32, 39, 41, 44, 13\}$
$E =$ the full set of input variables.

5 Results

We report in two parts on the results: the effect on test error of the input variable subset used to train a classifier, and the relative sizes of the test errors corresponding to the four techniques that were investigated. Consider first the average test errors summarised in Table 1.

It is clear that for every technique the test error decreases as the input variable set increases in size, especially so when we move from variable set B to variable set C. Increasing the number of input variables beyond the 17 in variable set C causes the average test error to decrease still further, but now much more slowly than before. Which variable subset should be used? Overall it seems that variable set D provides almost the same accuracy in terms of

Table 1. Mean test errors for different sets of input variables

	LDA	Logistic regression	Classification trees	SVM
Variable set A	0.242	0.236	0.232	0.222
Variable set B	0.229	0.227	0.226	0.214
Variable set C	0.193	0.173	0.181	0.154
Variable set D	0.189	0.165	0.179	0.153
Variable set E	0.185	0.160	0.178	0.146

test error as variable set E. If one considers that the former contains only 22 variables compared to the 61 in the latter, the gain in parsimony and the potential cost saving if variable set D is used may well be worth the slightly higher generalisation error compared to variable set E. Of course, one could conduct a more detailed investigation of the connection between alignment and average test error by sequentially adding one input variable at a time and estimating the corresponding test error, but it seems that the impression gained from the scree plot in Figure 1, namely that little is to be gained by adding input variables beyond a certain point, is substantiated by the entries in Table 1. It is also interesting to note that this holds for each of the four techniques under consideration, although the alignment measure that was used to rank the input variables is based on the kernel function used in the support vector machine. Finally, basing a decision regarding the number of input variables on a scree plot contains a subjective element. Finding an objective criterion that can be used in this regard is still an open problem.

Moving on to the second aspect that was investigated in the study, we present in Figure 2 boxplots of the 100 test errors for each technique (based on variable set D). The superiority of the support vector machine is evident, with logistic regression, classification trees and linear discriminant analysis following in this order.

Nevertheless, although support vector machines are definitely an addition to the statistician's toolbox, it should be borne in mind that proper application of support vector machines requires significant input from the user as far as specifying values for the underlying parameters is concerned. In our study this entailed fairly extensive experimentation to find appropriate values for the kernel function hyperparameter, γ, and the cost complexity parameter, C. The reported results were obtained for a specific combination of (γ, C)-values. If different values of γ and/or C were to be used, it may well lead to appreciable deterioration in the performance of the support vector machine. In this sense the support vector machine is not an off-the-shelf procedure. Finally, the use of support vector machines is hampered by the fact that this technique is not yet part of the main statistical software packages. This state of affairs is however sure to change in the near future.

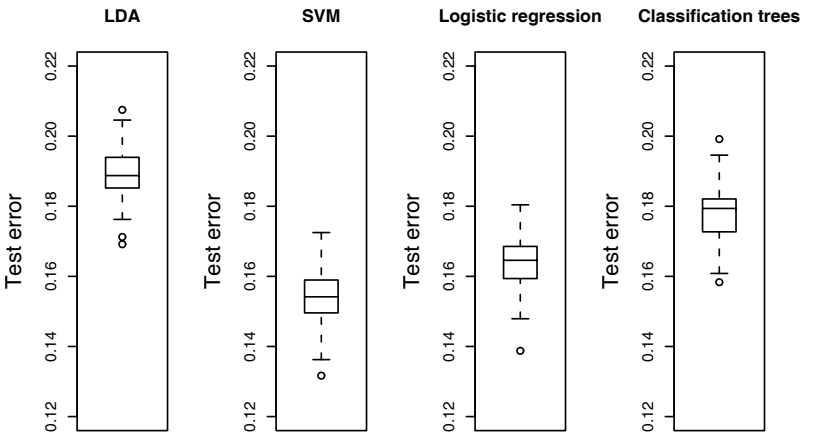


Fig. 2. Box plots of test errors of the four techniques

References

BOSER, B.E., GUYON, I.M. and VAPNIK, V.N. (1992): A Training Algorithm for Optimal Margin Classifiers. In: D. Haussler (Ed.): *5th Annual ACM Workshop on COLT*. ACM Press, Pittsburg PA.

CRISTIANINI, N., KANDOLA, J., ELISSEEFF, A. and SHAWE-TAYLOR, J. (2002): On Kernel-Target Alignment. In: T. Dietterich, S. Becker and D. Cohn (Eds.): *Neural Information Processing Systems, 14*. MIT Press, Cambridge.

GUYON, I.M. and ELISSEEFF, A. (2003): An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research, 3, 1157–1182*.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, Heidelberg.

HERBRICH, R. (2001): *Learning Kernel Classifiers*. MIT Press, London.

LOUW, N. and STEEL, S.J. (2004): Identifying Important Input Variables by Applying Alignment in Kernel Fisher Discriminant Analysis. Abstract submitted to the *19th International Workshop on Statistical Modelling, 4–8 July, Florence, Italy*.

R DEVELOPMENT CORE TEAM (2003): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

SCHÖLKOPF, B. and SMOLA, A.J. (2002): *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, London.

Continuous Market Risk Budgeting in Financial Institutions

Mario Straßberger

Chair of Finance, Banking, and Risk Management,
Friedrich-Schiller-University of Jena, 07743 Jena, Germany

Abstract. In this contribution we develop a profit & loss-dependent, continuous market risk budgeting approach for financial institutions. Based on standard modelling of financial market stochastics we provide a method of risk limit adjustment adopting the idea of synthetic portfolio insurance. Due to varying the strike price of an implicit synthetic put option we are able to keep within limits accepting a certain default probability.

1 Introduction

Modern risk management practices in financial institutions include more and more effective risk controlling procedures next to the pure measurement of risk. Instruments of active risk controlling are, for example, budgeting risks via setting concrete risk limits and hedging risks. Financial institutions may need risk management to reduce costs of external capital. They may need it to lower costs of financial distress and, by reducing earnings volatility, to avoid high taxes (Froot et al. (1993), Stulz (1996)). Specifically financial institutions face risk-based capital requirements so that hedging and budgeting risks may be preferred against raising additional capital. Motivations for risk controlling were first for the most part driven by the increasing magnitude of market risk and as a result, the Value-at-Risk (VaR) concept has become the standard tool to specify risk. But although tremendous efforts of academics and practitioners were undertaken to adequately measure Value-at-Risk, questions of how to control and specifically how to budget these measurable risk attracted surprisingly small interest.
But why should financial institutions even limit their risk-taking? The simple answer is: because their economic capital backing risky positions is restricted. Defining the institution's sustainability of risk as the outmost loss it could just bear by sure to maintain its solvability and long-term existence, we identify the implicit risk capital as reserve capital which is sufficient to cover unexpected losses with comparatively high probability. And why should the limitation of risk-taking depend on actual profit and loss? We argue that risk capital could be distinguished into different grades and that profit and loss is a part of the institution's short termed, first grade risk capital. Hence, together with cumulative losses the risk capital and the ability of risk-taking are decreasing (Merton and Perold (1993), Kupiec (1999)). Furthermore, by

incorporating shortfall constraints like Value-at-Risk limits into portfolio theory it was shown that risk becomes a function of the investor's risk aversion which again may depend on the amount of risk capital (Campbell et al. (2001)). Against that background, a primary goal of risk management may be to avoid so called lower-tail losses.
This paper enhances the ideas of Locarek-Junge et al. (2000) on an advanced modelling. Section 2 introduces the analyzing framework. Section 3 provides some remarks to the question of time and risk. The continuous market risk budgeting approach is presented in Section 4 and applied in a simple simulation study in Section 5.

2 Analysis framework

We consider a complete and arbitrage-free capital market where a single risky asset, e.g. a stock market index, is traded in continuous time. The price $S(t)$ of that asset at time t is modelled by a geometric Brownian motion

$$dS(t) = \mu S(t)dt + \sigma S(t)dz(t), \tag{1}$$

where μ denotes the drift of the asset value, σ the standard deviation of the asset value, and $dz(t)$ the standard Brownian motion. In the market there exists a risk free investment which's price $B(t)$ follows the dynamic

$$dB(t) = rB(t)dt, \tag{2}$$

where r denotes the continuous interest rate. A (weakly) expected-utility risk averse financial institution or its agent, respectively, only trades the considered risky asset in his portfolio. The market risk which the bank or financial institution faces can be identified as potential loss in portfolio value caused by price changes in the risky asset. As well known, Value-at-Risk, a measure of a portfolio's market risk, quantifies a loss bound that will not be exceeded by (positive) stochastic losses $L(T)$ with a specified probability p at a given time horizon T (Duffie and Pan (1997), Jorion (2000), Linsmeier and Pearson (2000), Locarek-Junge et al. (2002)). A general definition of Value-at-Risk is given by:

$$\mathrm{VaR}(p,T) := \inf\{l : l \geq 0, \Pr(L(T) \leq l) \geq p\}. \tag{3}$$

We assume the institution's risk management criterion is Value-at-Risk. Because in our analysis framework the conditional distribution of asset prices is log-normal and the conditional distribution of asset log-returns is normal, Value-at-Risk can be easily estimated using a method known as delta-normal approximation. This approach calculates Value-at-Risk at time t for a given progress in time, $t+T$, along the price process (1) to

$$\mathrm{VaR}(p,T) = S(t)\exp(rT) - S(t)\exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)T + z(p)\sigma\sqrt{T}\right), \tag{4}$$

where $z(p)$ is the p-quantile of the standard normal distribution (Dowd (2002), Duffie and Pan (1997)).
Although Value-at-Risk is not the latest of all possible risk measures, it is a fair approximation of risk, and it has become a widely used industry standard. In the case of financial institutions it can be motivated through capital requirements. Alternatively we could use other downside risk measures as, for example, the recent Conditional Value-at-Risk (Rockafellar and Uryasev (2002)). In brief, Conditional Value-at-Risk is an estimate of expected loss under the condition of loss is exceeding Value-at-Risk. Furthermore, Value-at-Risk was theoretically criticized (Szegö (2002)). It does not fulfill the coherence criteria set out by Artzner et al. (1999). But even though not in general, in our modelling framework Value-at-Risk is a coherent measure of risk and applies for our purposes.

3 Time dimension of risk limits

In the context of risk budgeting it is common practice in financial institutions to understand risk capital as to be available per business year. Even in academic literature the question arose of how an annually defined amount of risk capital would have to be calculated into a daily amount of risk capital. There was suggested to convert annual risk limits into daily risk limits using the known square root of time rule (Beeck et al. (1999)). Before we develop the dynamic risk budgeting model, in this section we discuss, whether risk capital or a risk limit should be defined as to be annual and whether it could be converted into sub-periods and vice versa.
For the first instance, the amount of risk capital is available independent from any time horizon. If the bank management provides an amount of risk capital for the business day that is smaller than the amount of risk capital available on the whole, than this reflects nothing but the management's attitude towards risk. The more risk averse the management is, the less it would feel up to put at stake per day. As shown by Campbell et al. (2001), the management's degree of risk aversion is captured by both the size and the confidence level of the chosen Value-at-Risk limit.
Next to economic reasons, this can be founded by expected-utility theory if we assume the bank management to have a concave utility function $u(C)$ of risk capital C, and to be risk averse with constant relative risk aversion $C\left(-u''(C)/u'(C)\right) = C\rho(C)$. Absolute risk aversion $\rho(C)$ then depends on the amount of risk capital available, which seems to be evident. The higher the amount of risk capital to dispose the lower the risk aversion, and the more risk is taken (Froot et al. (1993), Froot and Stein (1998)). Cumulative losses accompany decreasing risk capital. If relative risk aversion is constant then absolute risk aversion must increase together with cumulative losses which means lower acceptance of risk-taking and therefore lower risk limits.
What now is needed to incorporate into the theory is the dimension of time.

As well as one should additionally to the traditional criteria expected return and risk (as measured in terms of variance or down-side risk) account for time in modern portfolio theory, we additionally account for time in risk budgeting. In the figurative sense, the risk averse manager would prefer to put at stake a certain amount of risk capital within a certain period of time against the same amount of risk capital for a shorter period of time. Absolute risk aversion $\rho(C,T)$ then additionally depends on the time period T of risk bearing (Gollier and Zeckhauser (2002)). We argue that if a risk limit per business day or some other comparatively short time interval is smaller than the risk capital available, the rational is the preference of lower risk per unit of time.

4 Continuous risk budgeting

In this section we develop the continuous market risk budgeting approach. For the operational solution of the profit and loss dependent risk limiting problem we use the idea of portfolio insurance. Varying the strike price of the synthetic put option, we are able to move from perfect hedging to a hedge with accepted default probability.

Since the work of Rubinstein and Leland (1981), we know that options can exactly be replicated. They showed that a put option at every time can be duplicated by trading the underlying asset and the risk free investment, e.g. a (near) risk free government bond. That is, because in case of a geometric Brownian motion for the asset price process both the option and the underlying asset depend linearly on a single source of market risk. The duplication portfolio consists of a short position in the underlying asset and a long position in the risk free bond. Option delta is thereby calculated within the well known model of Black and Scholes (1973). So, in that "classical" option based portfolio insurance, instead of hedging portfolios with put options the hedging effect is achieved by dynamic reallocation of the capital between the risky asset and the risk free bond.

For simplicity, we assume options to be priced according to the Black-Scholes model. The market price $P(t) = P(S(t), X, r, T, \sigma)$ of a put option at time t equals

$$P(t) = X \exp(-rT)\Phi(d_1) - S(t)\Phi(d_2), \tag{5}$$

$$d_1 = \frac{\ln\left(\frac{X}{S(t)}\right) - \left(r - \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}},$$

$$d_2 = \frac{\ln\left(\frac{X}{S(t)}\right) - \left(r + \frac{\sigma^2}{2}\right)T}{\sigma\sqrt{T}},$$

where X denotes the strike price, and $\Phi(.)$ the cumulative standard normal distribution. We continuously calculate the delta factor of this put option as:

$$\Delta(t) = \frac{\partial P(S(t), X, r, T, \sigma)}{\partial S(t)} < 0. \tag{6}$$

The put delta expresses the sensitivity of the market price of the put option with respect to changes in the market price of the underlying asset. The reciprocal of the delta factor ceteris paribus indicates the number of put options needed to completely neutralize the price change per asset over the next infinitesimal time step. As is known, the put option can be duplicated by a portfolio consisting of a short position in the underlying asset amounting to

$$-\Delta(t)S(t) = S(t)\Phi(d_2),$$

and a long position in the risk free bond amounting to

$$B(t) = X \exp(-rT)\Phi(d_1).$$

Hence, for the synthetic put option it follows:

$$P(t) = B(t) + \Delta(t)S(t). \tag{7}$$

Because the synthetic put option itself partly consists of a short position in the asset, hedging with synthetic put options means reducing the risky position in assets in favor of a risk free position in bonds. The hedged portfolio then has a quota in the asset amounting to

$$a(t) = \frac{(1 + \Delta(t))S(t)}{(1 + \Delta(t))S(t) + B(t)}, \tag{8}$$

and a quota in the risk free bond amounting to $1-a(t)$. With our risk budgeting problem in mind, this portfolio insurance procedure is not implemented in real. We further just adopt the results obtained, leaving the risky position in the asset at the portfolio manager's charge, and adjusting the risk limit set by the management via:

$$\mathrm{VaR}(t) = a(t)(\mathrm{VaR}(t-1) - L(t)). \tag{9}$$

Market risk budgeting is made in such a way that the risk limit set decreases with cumulating losses and vice versa. We define the strike price of the synthetic put option as

$$X \in \left(0, \frac{\mathrm{VaR}(0)}{\exp(rT) - \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)T + z(p)\sigma\sqrt{T}\right)} - \mathrm{VaR}(0)\right], \tag{10}$$

whereby the upper bound of the interval marks the lowest accepted bound in portfolio value given the Value-at-Risk limit VaR(0) set at the beginning.

At time $t = 0$, the risky position maximally possible in the asset is much greater than the strike price of the synthetic put option. The option is well "out of the money" and its delta factor is near zero. At the same time, d_1 becomes very small, and hence the value of the cumulative standard normal distribution becomes nearly zero. From this it follows that $B(0)$ becomes zero, and $a(0)$ becomes one. Thus this means, the Value-at-Risk limit set is completely available at the beginning. Does the risky asset rise in market value the financial institution is making profits, and the risk limit expands at these profits because the risk capital increases at this amount. $B(t)$ stays unchanged at zero, and $a(t)$ stays unchanged at one. Whereas, does the risky asset fall in market value, and the financial institution is cumulating losses the synthetic put option moves more and more towards "at the money". Thereby both the delta factor and the quota in the asset of the hedged portfolio are declining. The Value-at-Risk limit is then reduced, ones due to the charging of losses and also due to the lower $a(t)$-factor.

Ahn et al. (1999) provide a model of optimally hedging a given market risk exposure under a Value-at-Risk constraint using options. In the same setting of a complete and arbitrage-free capital market they show that hedging costs are independent of the strike price of the put option used. The optimal strike price rather depends on the riskiness of the asset, the time horizon of the hedge, and the confidence level desired by the management.

Thus, by varying the strike price of the synthetic put option it must be possible to determine the level of portfolio protection. If the strike price equals the upper bound of the interval defined in (10) the asset position would be hedged at its lowest accepted value bound and the risk limit keeps that level. Reducing the strike price continually yields to accepting the fall below the lowest value bound of the asset position with increasing probability. We now can reduce the strike price as long as the default probability associated with the Value-at-Risk limit set is achieved. Then the risk limit will be violated with this probability.

5 Simulation analysis

We test the proposed continuous risk budgeting approach within our analysis framework. For the asset price process we assume $\mu = .0005$ and $\sigma = .015$. Furthermore, we set $S(0) = 100$ and $r = .03$. In 7,500 test runs we calculate price processes each with 256 time steps. In parallel we apply risk budgeting using an implicit synthetic put option with strike price at the lowest accepted bound in portfolio value. The Value-at-Risk limit is calculated at a five percent probability level. Comparing the dynamic behavior of the portfolio value the risk budgeting approach shows the expected properties. In Figure 1 we draw the resulting probability density of the portfolio profit & loss in the case of continuous market risk budgeting against the case of a constant risk limit over time.

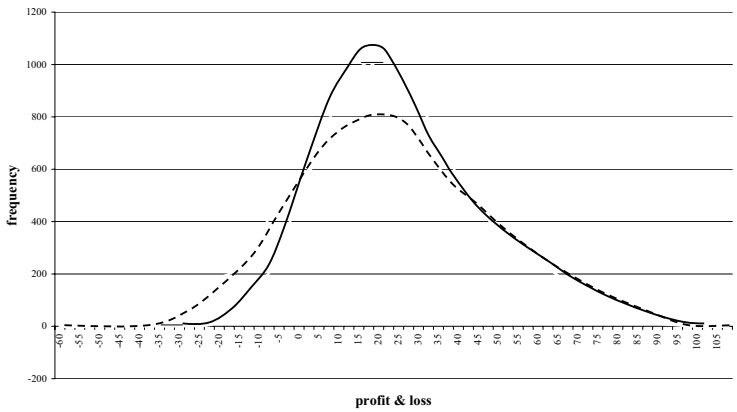


Fig. 1. Profit & loss density without (- - -) and with (—) continuous risk budgeting.

Applying the continuous risk budgeting strategy the distribution of portfolio profit & loss becomes more asymmetric. Skewness increases and kurtosis decreases. This is because of the portfolio insurance property. There is moved probability mass from the left tail of the distribution into its center.
In order to stay close to the desired Value-at-Risk probability we accept a default probability of the budgeting approach at the confidence level of the initial Value-at-Risk limit. We reduce the strike price of the implicit synthetic put option and find that a strike price of the half of the lowest asset value bound results in a default probability of about five percent. That equals the five percent probability level from risk calculation.
For financial institutions facing risk based capital requirements the situation has improved. Our approach does announce both reducing costs of capital and reducing probability of bankruptcy (For further analysis strengthen the aspect of cost reduction using knock-out- instead of plain-vanilla-put-options see Locarek-Junge et al. (2000)).

Acknowledgement

Much of these ideas arose at Dresden University of Technology's Chair of Finance and Financial Services. Thanks goes to Hermann Locarek-Junge for helpful suggestions and comments.

References

AHN, D.H., BOUDOUKH, J., RICHARDSON, M. and WHITELAW, R.F. (1999): Optimal Risk Management Using Options. *Journal of Finance, 54, 359–375.*

ARTZNER, P., DELBAEN, F., EBER, J.M. and HEATH, D. (1999): Coherent measures of risk. *Mathematical Finance, 9, 203–228.*

BEECK, H., JOHANNING, L. and RUDOLPH, B. (1999): Value-at-Risk-Limitstrukturen zur Steuerung und Begrenzung von Marktrisiken im Aktienbereich. *OR-Spektrum, 21, 259–286.*

BLACK, F. and SCHOLES, M. (1973): The pricing of options and corporate liabilities. *Journal of Political Economy, 81, 637–654.*

CAMPBELL, R., HUISMAN, R. and KOEDIJK, K. (2001): Optimal portfolio selection in a Value-at-Risk framework. *Journal of Banking & Finance, 25, 1789–1804.*

DOWD, K. (2002): *Measuring Market Risk.* John Wiley & Sons, Chichester.

DUFFIE, D. and PAN, J. (1997): An Overview of Value at Risk. *Journal of Derivatives, 4, 7–49.*

FROOT, K.A., SCHARFSTEIN, D.S. and STEIN, J.C. (1993): Risk management: coordinating corporate investment and financing policies. *Journal of Finance, 48, 1629–1658.*

FROOT, K.A. and STEIN, J.C. (1998): Risk management, capital budgeting, and capital structure policies for financial institutions: an integrated approach. *Journal of Financial Economics, 47, 55–82.*

GOLLIER, C. and ZECKHAUSER, R.J. (2002): Horizon Length and Portfolio Risk. *Journal of Risk and Uncertainty, 24, 195–212.*

JORION, P. (2000): *Value-at-Risk - The New Benchmark for Managing Financial Risk*, 2nd ed. McGraw-Hill, New York.

KUPIEC, P.H. (1999): Risk capital and VaR. *Journal of Derivatives, 6, 41–52.*

LINSMEIER, T.J. and PEARSON, N.D. (2000): Value at Risk. *Financial Analysts Journal, 56, 47–67.*

LOCAREK-JUNGE, H., PRINZLER, R. and STRAßBERGER, M. (2002): The estimation of market risk in portfolios of stocks and stock options. *Schmalenbach Business Review, 54 (Special issue 1), 171–189.*

LOCAREK-JUNGE, H., STRAßBERGER, M. and VOLLBEHR, H. (2000): Dynamische Limitsetzung. In: B. Rudolph and L. Johanning (Eds.): *Handbuch Risikomanagement.* Uhlenbruch, Bad Soden, 833–849.

MERTON, R.C. and PEROLD, A.F. (1993): Theory of risk capital in financial firms. *Journal of Applied Corporate Finance, 6, 16–32.*

ROCKAFELLAR, R.T. and URYASEV, S. (2002): Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance, 26, 1443–1471.*

RUBINSTEIN, M. and LELAND, H.E. (1981): Replicating options with positions in stock and cash. *Financial Analysts Journal, 37, 63–72.*

STULZ, R.M. (1996): Rethinking risk management. *Journal of Applied Corporate Finance, 9, 8–24.*

SZEGÖ, G. (2002): Measures of risk. *Journal of Banking & Finance, 26, 1253–1272.*

Smooth Correlation Estimation with Application to Portfolio Credit Risk

Rafael Weißbach[1] and Bernd Rosenow[2]

[1] Institut für Wirtschafts- und Sozialstatistik, University of Dortmund*
Department of Statistics, 44221 Dortmund, Germany
[2] Institut für Theoretische Physik, Universität zu Köln, 50923 Köln, Germany

Abstract. When estimating high-dimensional PD correlation matrices from short times series the estimation error hinders the detection of a signal. We smooth the empirical correlation matrix by reducing the dimension of the parameter space from quadratic to linear order with respect to the dimension of the underlying random vector. Using the method by Plerou et al. (2002) we present evidence for a one-factor model. Using the noise-reduced correlation matrix leads to increased security of the economic capital estimate as estimated using the credit risk portfolio model CreditRisk$^+$.

1 Introduction

Managing portfolio credit risk in a bank at first place requires sound and stable estimation of the loss distribution (for a given time horizon being typically one year). Special emphasis couches on high quantiles denoted by Credit Value-at-Risk (CreditVaR) and the resulting span between the expected loss and the CreditVaR, the Economic Capital, a scarce resource of a bank in general.

In large banks one key risk driver to be taken into account is concentration risk in industry sectors.

In CreditRisk$^+$ (Credit Suisse First Boston (CSFB) (1997)) concentration in sectors is modelled by a multiplicative random effect to the probability of default (PD) per counterpart or more general per risk entity. If sectors are set equal to the industry sectors - which seems to be common and straightforward practice - the sector variables X_k be can interpreted as "economy activity in sector k". In the technical document of CreditRisk$^+$ (Credit Suisse First Boston (CSFB) (1997)) the loss distribution is calculated for independent sector variables. The correlation of sector PD's was incorporated into the CreditRisk$^+$ framework by Bürgisser et al. (1999). Comparing the two situations of correlated and uncorrelated sectors clearly shows significant impact on the loss distribution.

While the authors of Rosenow et al. (2004) have focused on a conservative estimate of PD correlation, the emphasis of the present work is the detailed

* The work of Rafael Weißbach has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

comparison of different types of point estimators. We present a stable estimate for the PD correlation matrix without changing the portfolio credit risk assessment. To this end, we reduce the dimensionality of the parameter space by formulating and proving a model. The dimensionality of the estimation problem is reduced from $K(K-1)/2$ to K (in the example from 190 to 20). The observation of further data in time will change the correlation estimate smoothly as long as the one-factor model is applied. If evidence for a second latent factors is available it can be incorporated without a drastic change in the estimate.

The simplest model is the independence of the sector effects. In multivariate analysis tests for the covariance matrix to be the identity date back to the early 70's (see John (1971)). So far, testing was impossible when the sample size was smaller than the dimension of the distribution. The restriction was recently overcome by Ledoit and Wolf (2002). In credit risk in general and in CreditRisk$^+$ especially the effects of volatility and correlation are separated. We consider the correlation in this paper and we assume the volatility to be known. Hence the tests for the covariance matrix need to be applied to the correlation matrix. The observations have to be standardized to variance 1. Under this restriction we investigate the independence of sectorial economic activity with the result of a clear rejection of the independence for a segmentation of the German industry into 20 sectors.

The contributions of this paper are: (i) Using the ideas of Ledoit and Wolf (2002) to reject the independence of industry sectors with respect to the PD's; (ii) transforming the ideas of Plerou et al. (2002) from market risk to credit risk to generate a model and (iii) comparing the parameter estimates for the one-factor model of the "economy activity" to the empirical correlation matrix including impact as co-variable in CreditRisk$^+$.

The paper is organized as follows: Section 2 describes the generation of the sector variables and the sample data of insolvency rates in Germany. The test for independence of the latter sector variables and its results for the example are described in Section 3. The transfer of the ideas from mathematical physics to state the model is laid out in Section 3. The model and the estimation of parameters are presented in Section 5. Section 6 presents an algebraic approximation to the correlation matrix. In Section 7 the results of the data example are used to assess the impact on a realistic bank portfolio subject to credit risk using the CreditRisk$^+$ model.

2 The sector variable

A variable which measures the economic activity in various industry sectors is not observable at first place. The relevant concretization for the latent variable in CreditRisk$^+$ is a variable which quantifies the correlation between PD's of companies in the sectors.

The model is

$$P(A\ defaults) = p_A X_k$$

for a counterpart A belonging to sector k with individual (expected) PD p_A. The "relative sector economic activity" X_k has expectation 1.

We observe the insolvency rates per sector in past years $t = 1, \ldots, T$. The estimator of the sectorial PD in observation year t is:

$$\widehat{PD}_{kt} = \frac{\sharp\{A\ \in\ Sector\ k\ in\ year\ t\ defaulting\}}{\sharp\{\ A\ \in\ Sector\ k\ in\ year\ t\}}.$$

The relative PD movement is measured by

$$X_{kt} = \frac{\widehat{PD}_{kt} T}{\sum_{t=1}^{T} \widehat{PD}_{kt}}.$$

The common estimate of the covariance matrix is denoted by $S = \frac{1}{T}\sum_{t=1}^{n}(X_t - \bar{X})(X_t - \bar{X})'$ where $'$ denotes the transposed vector. The column vector X_t contains the K entries for the sectors and $\bar{X}$ is the mean over the X_t's. The empirical correlation matrix C is derived as usual with $C_{ij} = S_{ij}/\hat{\sigma}_{X_i}\hat{\sigma}_{X_j}$ and $\hat{\sigma}^2_{X_i} = S_{ii}$.

Example. The sample which will serve as illustration is the following: The Federal Statistical Office of Germany supplies default histories per industry sector. The industry sectors are defined in depth in the WZ93 key (Statistisches Bundesamt (1999)). We analyzed default quotes for the segmentation of the economy into 20 sectors, a magnitude common in banking trading-off concentration and granularity. The yearly data dated from 1994 till 2000.

The empirical correlation matrix is listed in Table 1 (left) in the Appendix.

3 Testing for independence

Ideas for testing the structure of the covariance matrix Σ of multivariate normally distributed random variables date back to the 70's (John (1971)) and were currently reviewed by Ledoit and Wolf (2002) with special dedication to small samples and large dimensions, a typical situation for credit risk. For testing the hypothesis $H_0 :\ \Sigma = I$ Ledoit and Wolf (2002) introduce the test statistic

$$W = \frac{1}{K} tr((S - I)^2) - \frac{K}{T}\left(\frac{1}{K} tr(S)\right)^2 + \frac{K}{T}$$

which is T-consistent and K-consistent with (T,K) limiting distribution $\frac{TK}{2} W \xrightarrow{D} \chi_{K(K+1)/2}$.

Because our interest focuses on the correlation matrix we argue as follows to deduce a level-α test. As we assume to know the variances we may normalize the data such that the covariance matrix of the generated time series

is the original correlation matrix. We apply the test and argue as follows. If we reject we know that either the diagonal elements are not 1 or/and the off-diagonal elements, i.e. the correlations are not 0. Because we know that the diagonal elements must be 1 we can conclude that the correlations are not negligible. In fact, we "wasted" power, the test is not admissible. Alternatively, we could have compared the variables on a one-by-one basis with tests for independence incurring the problem of multiple testing.

Note: In the CreditRisk$^+$ model the PD's are assumed to be Γ-distributed. However, there is no proof for the latter and the assumption is only of technical nature to enable an algebraic calculation of the probability generating function (see Credit Suisse First Boston (CSFB) (1997)). Although the assumption of normality of the "economic activity" factor $X = (X_1, \ldots, X_p)$ has the draw-back of enabling negative PD's we like to consider the case and have in a mind a truncated normality.

Example. For our example correlation matrix T is 7 and K is 20. The value of W is 5.81 and the critical value for level $\alpha = 0.05$ is 3.50. (The p-value is below 10^{-4}.) The calculations were performed using SAS/IML[1]. At a level of 5% one must reject the hypothesis of independent sectors.

4 Model generation

For the case of market risk Plerou et al. (2002) consider the case of high dimensional asset correlation estimation. They find that large eigenvalues of the empirical correlation matrix indicate a difference of the correlation matrix compared to the identity matrix implying independent dimensions. Note that a correlation matrix of independent variables has a K-fold Eigenvalue of 1.

They find an asymptotic boundary of

$$\lambda_{\pm} = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \tag{1}$$

with $Q = T/K$ for significant Eigenvalues where T denotes the number of - K-dimensional - observations. Eigenvalues outside the boundary indicate deviation from the hypotheses of independent sectors. The assumption of normally distributed random variable is crucial for the Wishart distribution of the empirical covariance. The same holds for the sample correlation matrix S under the assumption of variances of 1 or equivalently for known variances[2]. Following the argumentation of Plerou et al. (2002) we consider all eigenvalues under the threshold λ_+ to represent estimation noise.

[1] SAS and SAS/IML are registered trademarks of SAS Institute Inc. Carry, NC, USA.

[2] For the components of the largest eigenvalue's eigenvector they find that because of the akin positive quantity of the components the interpretation is a common influence by "the market" (see also Campbell et al. (1997)).

Example. For our data $\lambda_+ = 1+\frac{20}{7}+2\sqrt{\frac{20}{7}} = 7.24$. The ordered Eigenvalues are $10.38 > 4.60 > 2.01 > 1.26 > 0.96 > 0.78 > 1.71 \times 10^{-15} > 1.53 \times 10^{-15} > 1.21 \times 10^{-15} > 6.22 \times 10^{-16} > 5.57 \times 10^{-16} > 1.86 \times 10^{-16} > 3.6 \times 10^{-18} > -1.49 \times 10^{-16} > -3.73 \times 10^{-16} > -5.6 \times 10^{-16} > -8.24 \times 10^{-16} - 1.02 \times 10^{-15} > -1.26 \times 10^{-15} > -1.74 \times 10^{-15}$. The largest Eigenvalue of 10.38 is above the threshold (1) whereas the second largest 4.60 is below. That means their method advocates for one significant factor or equivalently for a one-factorial design.

5 A one-factor model

We start with the one-factor design:

$$X_{kt} = \alpha_k + \beta_k \tilde{X}_t + \gamma_k \varepsilon_{kt}, \tag{2}$$

for the dimensions $k = 1, \ldots, K$.

The α_k's and β_k's are given due to the restrictions $E(X_{kt}) = 1$ and $Var(X_{kt}) = \sigma_k^2$. The parameters γ_k must be estimated. However, to apply linear regression to estimate the K parameters we must restate the problem.

As we are interested in modelling correlations rather than covariances, we normalize the X_{it} such that they have the same, namely the average variance $\hat{\sigma}_X^2 = (1/K)\sum_{i=1}^K \hat{\sigma}_{X_i}^2$ and subtract the mean $Y_{it} = (X_{it} - 1)\frac{\hat{\sigma}_X}{\hat{\sigma}_{X_i}}$. The empirical correlation matrix is unchanged.

We model the correlations between relative PD movements by the one–factor model

$$Y_{it} = \delta_i \tilde{X}_t + \epsilon_{it} \tag{3}$$

The coefficients $\{\delta_i\}$ are then found by performing a linear regression without off-set.

How does one define the latent $\tilde{X}_t$? A simple average over the X_{kt}'s, $\tilde{X}_t \approx \frac{1}{K}\sum_{k=1}^K X_{kt}$, would not reflect the importance of the specific sector. We use the definition by Plerou et al. (2002). We diagonalize the empirical cross correlation matrix C and rank order its eigenvalues $\lambda_{(i)} < \lambda_{(i+1)}$. We use the components of the eigenvector $\mathbf{u}^{(K)}$ corresponding to the largest eigenvalue $\lambda_{(K)} = 11.46$ to define the factor time series

$$\tilde{X}_t = \sum_{k=1}^K u_k^{(K)} Y_{kt} \quad .$$

The point estimator can now be calculated under the assumption that the residuals $\{\epsilon_{i,t}\}$ are iid observations from uncorrelated random variables ϵ_i $i = 1, \ldots, K$, i.e. $Corr(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. Estimating the factor variance $\hat{\sigma}_{\tilde{X}}^2 = \frac{1}{T-1}\sum_{t=1}^T \tilde{X}_t^2$, one finds the point estimator for the cross correlation matrix as

$$C_{ij}^{1\mathrm{F}} = \Delta_{ij} + (1 - \Delta_{ij})\delta_i \delta_j \hat{\sigma}_Y^2 / \hat{\sigma}_X^2 \quad . \tag{4}$$

with Δ_{ij} as Kronecker-Δ. The correlation matrix is given in Table 1 (right) in the Appendix. Before comparing the impact of the one-factor model (2) in a portfolio we like to compare the derived correlation matrix (4) with the empirical correlation matrix in Table 1 (left). Although the maximal increase of a correlation caused by the one-factor model is 0.56 and the maximal decrease is 0.85 the mean change is only a decrease of 0.04. I.e. the overall level of correlation has not been altered. In order to assessing the tendency of individual changes we calculated the mean of the absolute changes; the result 0.18 demonstrates that the one-factor model does not change the correlation estimate dramatically.

6 Algebraic approximation

The derived estimate of the correlation matrix has almost the form $vector \times vector'$. The same form arises when directly applying the spectral decomposition of $C = \sum_{k=1}^{K} \lambda_k u^k u^{k'}$ with Eigenvalues λ_k and Eigenvectors u^k. If the largest Eigenvalue is dominant, the correlation matrix can be approximated by $C \approx \lambda_{(K)} u^{(K)} u^{(K)'} =: \tilde{C}$. The approximating matrix has two deficiencies. First, its trace is not K anymore but ${\lambda_{(K)}}^3$. In order to regaining the trace without interference of the positive semi-definiteness we multiply the matrix with the factor $K/\lambda_{(K)}$. Secondly, the diagonal elements of the achieved matrix are not 1 implying a change in the marginal variances because for the resulting variance-covariance estimate S^{alg} holds $S^{alg}_{kk} = \sigma_k^2 \tilde{C}_{kk}$ or in matrix notation $S^{alg} := K \times diag((\sigma_1, \ldots, \sigma_K)) u^{(K)} u^{(K)'} diag((\sigma_1, \ldots, \sigma_K))'$, where $diag(v)$ denotes for $v \in \mathbb{R}^K$ the matrix $\in \mathbb{R}^{K \times K}$ with v as diagonal and 0 else. In case of a matrix V, $diag(V)$ denotes the matrix with identical diagonal and 0 else. The implicit correlation matrix can be calculated by multiplying S^{alg} from left and right with the inverse of the square root of the diagonal version of S^{alg}. The procedure respects the necessary condition of a correlation matrix to be positive semi-definite (psd).

$$
\begin{aligned}
C^{alg} &:= diag(S^{alg})^{-\frac{1}{2}} S^{alg} diag(S^{alg})^{-\frac{1}{2}} \\
&= diag((|\sigma_1 u_1^{(K)}|, \ldots, |\sigma_K u_K^{(K)}|))^{-1} (\sigma_1 u_1^{(K)}, \ldots, \sigma_K u_K^{(K)})' \\
&\quad (\sigma_1 u_1^{(K)}, \ldots, \sigma_K u_K^{(K)}) diag((|\sigma_1 u_1^{(K)}|, \ldots, |\sigma_K u_K^{(K)}|))^{-1} \\
&= (sign(u_i^{(K)} u_j^{(K)}) 1)_{i,j=1,\ldots,K}. \qquad (5)
\end{aligned}
$$

Note, that the signs (denoted by $sign$) of the components of the Eigenvector $u^{(K)}$ to the largest Eigenvalue $\lambda_{(K)}$ vary. Alternatively, one could simply *set* the diagonal to 1, which may destroy the psd feasure, but interestingly

[3] Note that the trace of a product of matrices is invariant under cyclic interchanges and hence $K = trace(C) = trace(U \Lambda U^t) = trace(U^t U \Lambda) = trace(\Lambda) \neq trace(\lambda_{(K)} u^{(K)} u^{(K)'}) = \lambda_{(K)}$, because U is orthogonal.

equation (4) implies a similar method. Note that Rosenow et al. (2002) describe an akin approach using a principal component analysis.

As in Section 5 we like to compare the derived correlation matrix (5) with the empirical correlation matrix in Table 1 (left). Here, the maximal increase of a correlation caused by the one-factor model is 1.58 and the maximal decrease is 1.62 but the mean change is only a decrease of 0.02. I.e. the overall level of correlation has not been altered. In order to assessing the tendency of individual changes we calculated the mean of the absolute changes 0.60 demonstrating that algebraic manipulation does change the correlation estimate dramatically.

We did not apply this procedure in the following because of the unrealistic change in correlation structure. Additionally, the one-factor model has a more appealing interpretation: All economic activity in different branches X_k is linked via *one* latent global economic activity $\tilde{X}$, only the strength δ_k of the relation varies.

7 Impact on the practical performance

Using a portfolio we compare the results of the CreditRisk$^+$ calculation. The portfolio we study is realistic – although fictitious – for an international bank. It consists of around 5000 risk units distributed asymmetrically over 20 sectors with 20 to 500 counterparts per sector. The largest exposure is 500 mn Euro and the smallest exposure of 0.1 mn Euro. The counterpart specific default probability varies between 0.03% and 7%, the expected loss for the total portfolio is 121.1 mn Euro.

The competing correlation matrices are: (i) the original sample correlation, (ii) the correlation matrix assuming a one-factor model (4) and (iii) the algebraic approximation (5).

We calculated the loss distribution by using CreditRisk$^+$ and the method of Bürgisser et al. (1999) for integrating correlations. For the empirical correlation matrix C the CreditVaR is 957 mn Euro, whereas the CreditVaR for C^{1F} is 943 mn Euro. The difference is negligible demonstrating the one-factor model describes the data sufficiently in the context of portfolio credit risk.

Acknowledgement. The help of F. Altrock and the contribution of an anonymous referee are gratefully acknowledged.

References

BÜRGISSER, P., KURTH, A., WAGNER, A. and WOLF, M. (1999): Integrating Correlations. *Risk, 7, 57–60.*

CAMPBELL, J., LO, A.W. and MACKINLAY, A.C. (1997): *The Econometrics of Financial Markets.* Princeton University Press, Princeton.

CREDIT SUISSE FIRST BOSTON (CSFB) (1997): Credit Risk +: A Credit Risk Management Framework. *Technical document.*
JOHN, C. (1971): Some optimal multivariate tests. *Biometrika, 58, 123–127.*
LEDOIT, O. and WOLF, M. (2002): Some hypothsis tests for the covariance matrix when the dimansion is large compared to the sample size. *Annals of Statistics, 30, 1081–1102.*
PLEROU, V., GOPIKRISHNAN, P., ROSENOW, B., NUNES, L.A.N., GUHR, T. and STANLEY, H.E. (2002): A Random Matrix Approach to Cross-Correlations in Financial Data. *Physical Review Letters E, 65, 066126.*
ROSENOW, B., PLEROU, V., GOPIKRISHNAN, P., and STANLEY, H.E. (2002): Portfolio optimization and the random magnet problem. *Europhysics Letters, 59, 500–506.*
ROSENOW, B., WEISSBACH, R. and ALTROCK, F. (2004) Modelling Correlations in Portfolio Credit Risk, *arXiv:cond-mat/ 0401329 v1 19 Jan 2004.*
STATISTISCHES BUNDESAMT (1999): *Klassifikation der Wirtschaftszweige, Ausgaben 1993 (WZ93).* Statistisches Bundesamt, Wiesbaden.

A Appendix

Table 1. Left: Empirical correlation matrix. Right: One-factor correlation matrix.

1.00	-0.70	0.59	-0.71	0.49	0.11	-0.05	-0.26	-0.86	0.00	0.92	0.90	0.96	0.23	0.34	0.90	0.83	0.67	0.80	-0.72
-0.70	1.00	-0.69	0.41	-0.62	-0.61	-0.21	-0.37	0.43	-0.58	-0.67	-0.70	-0.71	-0.34	-0.14	-0.65	-0.52	-0.34	-0.83	0.62
0.59	-0.69	1.00	0.01	0.61	0.49	0.15	0.12	-0.41	0.37	0.46	0.73	0.72	0.43	0.08	0.59	0.47	0.20	0.54	-0.78
-0.71	0.41	0.01	1.00	0.08	0.27	0.28	0.45	0.80	0.26	-0.61	-0.47	-0.56	-0.03	-0.60	-0.69	-0.75	-0.84	-0.69	0.45
0.49	-0.62	0.61	0.08	1.00	0.47	0.48	0.60	-0.22	0.62	0.63	0.59	0.56	0.63	0.02	0.49	0.34	0.08	0.45	-0.12
0.11	-0.61	0.49	0.27	0.47	1.00	0.69	0.60	0.35	0.46	0.27	0.39	0.27	0.12	-0.44	0.08	-0.14	-0.40	0.36	-0.28
-0.05	-0.21	0.15	0.28	0.48	0.69	1.00	0.55	0.33	0.11	0.30	0.32	0.14	0.50	-0.06	0.16	0.01	-0.21	0.31	0.10
-0.26	-0.37	0.12	0.45	0.60	0.60	0.55	1.00	0.49	0.84	-0.05	-0.11	-0.20	0.40	-0.24	-0.20	-0.31	-0.41	0.05	0.38
-0.86	0.43	-0.41	0.80	-0.22	0.35	0.33	0.49	1.00	0.17	-0.68	-0.68	-0.78	-0.31	-0.69	-0.87	-0.93	-0.92	-0.67	0.63
0.00	-0.58	0.37	0.26	0.62	0.46	0.11	0.84	0.17	1.00	0.01	-0.01	-0.01	0.32	-0.18	-0.05	-0.14	-0.21	0.14	0.09
0.92	-0.67	0.46	-0.61	0.63	0.27	0.30	-0.05	-0.68	0.01	1.00	0.91	0.91	0.34	0.28	0.88	0.76	0.56	0.82	-0.50
0.90	-0.70	0.73	-0.47	0.59	0.39	0.32	-0.11	-0.68	-0.01	0.91	1.00	0.98	0.41	0.29	0.91	0.79	0.53	0.84	-0.76
0.96	-0.71	0.72	-0.56	0.56	0.27	0.14	-0.20	-0.78	-0.01	0.91	0.98	1.00	0.34	0.31	0.92	0.82	0.60	0.82	-0.79
0.23	-0.34	0.43	-0.03	0.63	0.12	0.50	0.40	-0.31	0.32	0.34	0.41	0.34	1.00	0.66	0.56	0.56	0.46	0.51	-0.10
0.34	-0.14	0.08	-0.60	0.02	-0.44	-0.06	-0.24	-0.69	-0.18	0.28	0.29	0.31	0.66	1.00	0.64	0.79	0.90	0.54	-0.24
0.90	-0.65	0.59	-0.69	0.49	0.08	0.16	-0.20	-0.87	-0.05	0.88	0.91	0.92	0.56	0.64	1.00	0.97	0.82	0.90	-0.69
0.83	-0.52	0.47	-0.75	0.34	-0.14	0.01	-0.31	-0.93	-0.14	0.76	0.79	0.82	0.56	0.79	0.97	1.00	0.93	0.83	-0.63
0.67	-0.34	0.20	-0.84	0.08	-0.40	-0.21	-0.41	-0.92	-0.21	0.56	0.53	0.60	0.46	0.90	0.82	0.93	1.00	0.70	-0.47
0.80	-0.83	0.54	-0.69	0.45	0.36	0.31	0.05	-0.67	0.14	0.82	0.84	0.82	0.51	0.54	0.90	0.83	0.70	1.00	-0.67
-0.72	0.62	-0.78	0.45	-0.12	-0.28	0.10	0.38	0.63	0.09	-0.50	-0.76	-0.79	-0.10	-0.24	-0.69	-0.63	-0.47	-0.67	1.00

1.00	-0.69	0.61	-0.65	0.48	0.15	0.11	-0.14	-0.80	0.04	0.82	0.86	0.89	0.47	0.53	0.92	0.88	0.73	0.86	-0.69
-0.69	1.00	-0.48	0.51	-0.38	-0.12	-0.09	0.11	0.64	-0.03	-0.65	-0.68	-0.70	-0.37	-0.42	-0.73	-0.70	-0.58	-0.68	0.55
0.61	-0.48	1.00	-0.45	0.34	0.10	0.08	-0.10	-0.56	0.03	0.57	0.60	0.62	0.33	0.37	0.64	0.61	0.51	0.60	-0.48
-0.65	0.51	-0.45	1.00	-0.36	-0.11	-0.08	0.11	0.59	-0.03	-0.61	-0.64	-0.65	-0.35	-0.39	-0.68	-0.65	-0.54	-0.64	0.51
0.48	-0.38	0.34	-0.36	1.00	0.08	0.06	-0.08	-0.44	0.02	0.45	0.48	0.49	0.26	0.29	0.51	0.49	0.41	0.48	-0.38
0.15	-0.12	0.10	-0.11	0.08	1.00	0.02	-0.02	-0.13	0.01	0.14	0.14	0.15	0.08	0.09	0.15	0.15	0.12	0.14	-0.12
0.11	-0.09	0.08	-0.08	0.06	0.02	1.00	-0.02	-0.10	0.00	0.11	0.11	0.11	0.06	0.07	0.12	0.11	0.09	0.11	-0.09
-0.14	0.11	-0.10	0.11	-0.08	-0.02	-0.02	1.00	0.13	-0.01	-0.13	-0.14	-0.14	-0.08	-0.09	-0.15	-0.14	-0.12	-0.14	0.11
-0.80	0.64	-0.56	0.59	-0.44	-0.13	-0.10	0.13	1.00	-0.04	-0.76	-0.79	-0.82	-0.44	-0.49	-0.85	-0.81	-0.68	-0.79	0.64
0.04	-0.03	0.03	-0.03	0.02	0.01	0.00	-0.01	-0.04	1.00	0.04	0.04	0.04	0.02	0.02	0.04	0.04	0.03	0.04	-0.03
0.82	-0.65	0.57	-0.61	0.45	0.14	0.11	-0.13	-0.76	0.04	1.00	0.81	0.83	0.45	0.50	0.87	0.83	0.69	0.81	-0.65
0.86	-0.68	0.60	-0.64	0.48	0.14	0.11	-0.14	-0.79	0.04	0.81	1.00	0.88	0.47	0.53	0.91	0.87	0.72	0.85	-0.69
0.89	-0.70	0.62	-0.65	0.49	0.15	0.11	-0.14	-0.82	0.04	0.83	0.88	1.00	0.48	0.54	0.94	0.89	0.74	0.87	-0.70
0.47	-0.37	0.33	-0.35	0.26	0.08	0.06	-0.08	-0.44	0.02	0.45	0.47	0.48	1.00	0.29	0.50	0.48	0.40	0.47	-0.38
0.53	-0.42	0.37	-0.39	0.29	0.09	0.07	-0.09	-0.49	0.02	0.50	0.53	0.54	0.29	1.00	0.56	0.54	0.45	0.53	-0.42
0.92	-0.73	0.64	-0.68	0.51	0.15	0.12	-0.15	-0.85	0.04	0.87	0.91	0.94	0.50	0.56	1.00	0.93	0.78	0.91	-0.74
0.88	-0.70	0.61	-0.65	0.49	0.15	0.11	-0.14	-0.81	0.04	0.83	0.87	0.89	0.48	0.54	0.93	1.00	0.74	0.87	-0.70
0.73	-0.58	0.51	-0.54	0.41	0.12	0.09	-0.12	-0.68	0.03	0.69	0.72	0.74	0.40	0.45	0.78	0.74	1.00	0.72	-0.58
0.86	-0.68	0.60	-0.64	0.48	0.14	0.11	-0.14	-0.79	0.04	0.81	0.85	0.87	0.47	0.53	0.91	0.87	0.72	1.00	-0.69
-0.69	0.55	-0.48	0.51	-0.38	-0.12	-0.09	0.11	0.64	-0.03	-0.65	-0.69	-0.70	-0.38	-0.42	-0.74	-0.70	-0.58	-0.69	1.00

How Many Lexical-semantic Relations are Necessary?

Dariusch Bagheri

Fachbereich II, Linguistische Datenverarbeitung,
Universität Trier, 54286 Trier, Germany

Abstract. In lexical semantics several meta-linguistic relations are used to model lexical structure. Their number and motivation vary from researcher to researcher. This article tries to show that one relation suffices to model the concept structure of the lexicon making use of intensional logic.

1 Introduction

In recent years great effort has been undertaken to build up representative cross-sections of the lexicon of a language. By far the best known result of these efforts is WordNet. As the name suggests the entries of the lexicon are linked by serveral lexical relations: hyponoymy/hyperonymy, synonymy, meronymy/holonymy, oppositions, and familiarity. Even though the relations are seen as primary they are not sufficient to distinguish different meanings. In addition so-called *glosses* are added which resemble very much customary definitions in defining dictionaries. The need for definitions is recognized as a flaw which should be overcome by the use of more relations or by a further division of the relations into subtypes.

But even the definitions of the most basic relations like hyponymy, hypernonymy and synonymy cannot be considered as uncontroversial. (e. g. Cruse (1986), Murphy (2003)). Another point of criticism is scalability of those lexical systems. Once the relations are established and the lexicon is compiled with many thousand entries it is in practice nearly impossible to add a new relation. Furthermore, because of the different types of relations, it is senseless to give overall mathematical characteristics of the lexical systems, as for example the distribution of the paths in the net. The question how many relations are needed to code the lexicon or whether the relations used are sufficient is never raised explicitly.

In order to elaborate the concept calculus in the next section first order predicate logic is used. The sentence operators conjunction (*and*, symbol $\wedge$), adjunction (*or*, symbol $\vee$), subjunction (*conditional, if-then*, symbol $\rightarrow$), bisubjunction (*biconditional, if-and-only-if*, symbol $\leftrightarrow$), and negation (symbol $\sim$) have the usual definitions. All variables and constants in the concept calculus are concepts. So a variable or constant c might stand for a concept $beautiful$ or an n-ary concept like $x\ is\ greater\ than\ y$.

2 Concept calculus

This concept calculus goes back to G. W. Leibniz. Kauppi (1967) condensed and improved the intensional calculus and put it into a modern form in terms of relations. Leibniz distinguished between logic purely based on concepts (intension), and a logic based on objects (extension). A definition constitutes the fundamental relation between concepts: the defined concept contains the defining one. Usually a concept contains several concepts.

A system of concepts is always made up of concepts of the same arity, i. e. the number of arguments that saturate a concept in the same sense a mathematical function is saturated by its arguments. This means that there are concept systems of concepts with arity 0, 1, 2 and so on. [1] These different systems can be connected so that higher arity systems are determined by systems of lower arity and that eventually all systems are determined by systems of arity one. The reason for this division of systems is that a one-place concept like *red* (something is red) cannot contain a two-place concept like for example *lighter* (red is lighter than blue). The following axioms and theorems hold for concepts irrespective of arity. This calculus is abbreviated as **BK** (*Begriffskalkül*/concept calculus). The laws in connection with concepts of an arity higher than one are taken into account by the relation calculus **RK** (*Relationenkalkül*). These laws must be considered if concepts of different arity are to be linked. Finally the application of a concept system has to be specified by another relation not being part of BK or RK. This relation specifies to what individuals or objects the concepts might be assigned. It is a binary relation with as its first argument a concept, and, as its second argument, an object or an ordered tuple of objects, depending on the concept's arity.

Following is a selection of the laws of the concepts calculus: [2]

$$a > b \quad (\textbf{containment}) \tag{1}$$

$$a \gtrdot b =_{def} a > b \wedge \sim b > a \quad \textit{and} \quad a \lessdot b =_{def} a < b \wedge \sim b < a \tag{2}$$

$$a < b =_{def} b > a \tag{3}$$

$$a = b =_{def} a > b \wedge b > a \quad (\textbf{identity}) \tag{4}$$

$$a \; \mathbf{H} \; b =_{def} \exists x(a > x \wedge b > x) \quad (\textbf{comparability}) \tag{5}$$

$$a \; \mathbf{I} \; b =_{def} \sim \exists x(a > x \wedge b > x) \tag{6}$$

$$a \curlywedge b =_{def} \exists x(x > a \wedge x > b) \quad (\textbf{compatibility}) \tag{7}$$

$$a \curlyvee b =_{def} \sim \exists x(x > a \wedge x > b) \tag{8}$$

$$c = a \odot b =_{def} \forall x(c > x \leftrightarrow a > x \wedge b > x) \quad (\textbf{product}) \tag{9}$$

---

[1] Concepts of arity 0 will not be considered here.
[2] For a deeper discussion cf. Kauppi (1967).

$$c = a \oplus b =_{def} \forall x(x > c \leftrightarrow x > a \wedge x > b) \quad (\textbf{sum}) \tag{10}$$

$$b = \bar{a} =_{def} \forall x(x > b \leftrightarrow x \curlyvee a) \quad (\textbf{negation}) \tag{11}$$

$$a > \bar{\bar{a}} \tag{12}$$

$$c = a \oslash b =_{def} \forall x(x > c \leftrightarrow x > a \wedge x \curlyvee b) \quad (\textbf{quotient}) \tag{13}$$

$$a \oslash b =_{def} a \oplus \bar{b} \tag{14}$$

The relation calculus RK is of great importance for the analysis of definitions and for defining concepts. The distinction is mostly ignored. Consider a definition of *triangle* which states that *a triangle has three sides.* Commonly this is judged to be a feature of *triangle,* but compared to the distinction drawn here between concepts of different arity, it must be stated that a triangle cannot contain *having three sides* because this is a three-place predicate of the form: *X has N Y.* A one-place predicate like triangle cannot contain a three-place predicate like *X has N Y.* It determines only one argument, that means the first argument is the concept *triangle*, the second an indivudual concept for *number* like *zero*, *one*, *two* etc., and the third an unspecified concept for what the first argument in question 'has': a *triangle has three sides.* What this means for a definition and its corresponding linguistic sign, i. e. a lexeme, will be considered later.

To distinguish these two kinds of relations it will be said that if a concept contains another one they are intensionally connected, if a concept of lower arity determines a place or places of a concept of higher arity then these concepts are R-logically connected. Eventually every place of a n-ary concept is determined by a concept of arity one. A n-ary relation r will be written as r_n. The determination of certain places means that those places are occupied by certain concepts and that these places are only applicable to individuals to which the determining concepts are applicable. The two-place concept *father* contains the two-place concept *parent* and is determined in its first place by the concept *male.* Formally this is notated like this

$$r_n \diagup \!\!\! \diagup {}^{a}_{k} \quad \text{with} \quad 0 < k \leq n$$

Determinations of several places simultaneously are notated like this

$$r_n \diagup \!\!\! \diagup {}^{a\ b\ c}_{k\ l\ m} \quad \text{with} \quad 0 < k, l, m \leq n$$

where k, l, m are the numbers of the places, and a, b, c are concepts determining these places. Another term which has to be introduced characterizes concepts which are components of n-ary concepts. It is called component relation (*Unterrelation*). For example, the concept *father* which has two places has three component relations: $father_1$ which is extensionally applicable to a person who is a father, $father_2$ which is applicable to the children (in the sense of descendants), and $father_{1\ 2}$ which is applicable to a pair of persons who stand in this relationship to each other.

3 Diagrammatic representation

Partially ordered set or lattice diagrams are good candidates for a graphical representation. But the arity of the concepts cannot be distinguished. The labels of the nodes are another disadvantage. This will be dealt with in the next section. For now the focus lies on the graphic represention of the higher arity concepts.

N-place concepts can only contain other concepts of the same arity. For example the two-place concept *father* contains the two-place concept *parent*, but cannot contain the one-place concept *male*, although *male* determines the first place of the concept *father*. The idea is now to change the nodes of higher arity concepts to circles containing nodes or other circles and to join the edges to the line of an outer or inner circle or node to express the containment relation appropriately. This looks now as shown in Figure 3. The edge starting at the oval labelled "father" down to the oval labelled "parent" represents a containment relation. The edge starting from leftmost inner node of father down to node "male" represents the containment relation of the first place of father. If the inner nodes of circles are not labelled with numbers conventionally the indexing goes from left to right, that is, the leftmost one represents the first place, the node next to this the second place etc.

This extension makes it possible to treat a rather complex example with respect to the arity of concepts: kinship relations. The definitions of the terms are not taken to be representative. Especially the affinal relations, i. e. marriage, are left out completely. First the definitions are given and then the diagram will be built up incrementally. Prime concepts are explicitly marked, as well as arity, which is given in brackets. Polysemy or homonymy are completely ignored in this example.

Male (1) undefined. **Female (1)** undefined. **Ancestor = descendant (2)** undefined. **Parent (2)** is an ancestor of his or her children. **Grandparent (3)** is a parent of a parent. **Father (2)** is a male parent. **Mother (2)** is a female parent. **Child (2)** is a descendant of parents. **Grandchild (3)** is child a of a child. **Son (2)** is a male child. **Daughter (2)** is a female child. **Siblings (3)** are children with the same parents. **Brother (3)** is a male child of the same parents. **Sister (3)** is a female child of the same parents. **Grandfahter (3)** is a father of a parent. **Grandmother (3)** is a mother of a parent. **Grandson (3)** is a son of a child. **Granddaughter (3)** is a daughter of a child. **Uncle (4)** is the brother of a parent. **Aunt (4)** is a sister of a parent. **Niece (4)** is a daughter of a brother or sister. **Nephew (4)** is a son of a brother or sister. **Cousin (5)** is a child of an uncle or aunt.

Extensionally one would expect, for example, the cousin relation to be a pair of individuals, not a concept of arity five. But imagine a family reunion with many members and the participants reasoning who stands in what relationship to others. To judge that two persons are cousins, they must figure out

who their parents are, and in which relation these parents stand to each other. This will become clearer in due course.

All concepts of arity one are prime concepts and therefore contain no other concepts. To start with the nuclear family the example of Figure 3 can be simply extended. The concepts of the nuclear family have arity two. The only prime concept is that of *ancestor/descendant*. Figure 1 shows the relations. The two nodes labelled "parent/child" and "ancestor/descendant"

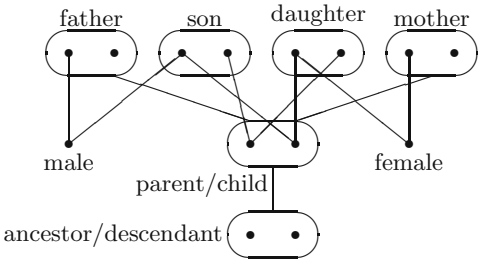


Fig. 1. Relation of the nuclear family.

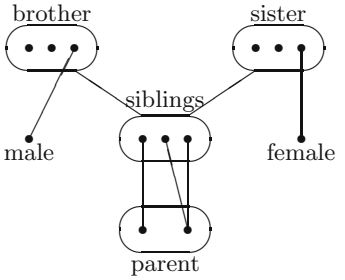


Fig. 2. Relation of brother and sister.

represent a concept, for which there are two linguistic expressions. These two expressions do not mean the same. As mentioned above, a higher arity concept has component concepts that possess part of the determined places of the original concept, or the same number of concepts but in different order. So "parent" refers to the concept $parent/child_{1\,2}$ and "child" to the component concept with inverted places $parent/child_{2\,1}$. As the places do not contain different non-identical concepts they can be represented in one node. The same holds for "ancestor/descendant". It is different, for example, in the case of the concepts *father* and *son*. Both first places are determined by the concept *male*. But as they contain different non-identical concepts they cannot be combined in one node. *Son* determines the second place of *parent/child* as male, whereas *father* determines the first place. As can be seen in the case of *son*, that contains *child*, as the above definition states, this is realized via explicit edges starting from the places of *son* and ending at the appropriate places in *parent/child*. There is no reason not to draw an extra node for *parent* and *child*, but to convey the identity-relation between these expressions and their concepts.

To complete the nuclear family the definitions for *brother*, *sister*, and *siblings* are displayed in Figure 2. The second and the third place of *siblings* determine the child relation to the same parent; the parent is specified in the first place. The third place, though, is determined by sex on the concept nodes for *brother* and *sister*.

The next step supplies the kinship concepts two generations above ego. Figure 5 shows these relations and leaves out, for clarity, the relations just

introduced. There are two component relations determining *grandparent/-child*: first place is the parent of the second place and the second place is the parent of the third place. The second place is a concept which is the sum of being a parent and a child. The other concepts are determined at the appropriate places like *father* and *son*. They could also be determined by the sex concepts *male* and *female*, but choosing the higher arity concepts reveals the structure much better.

The collateral four-place concepts are illustrated in Figure 6 and Figure 7. Here the first three-places of *uncle* are determined by the *brother* relation. The third place is the uncle, that is the brother of ego's parent. Ego's parent is determined in the second place by containment of the first place of *parent*, and ego in the fifth place of *uncle* is determined as the child of *parent*. The

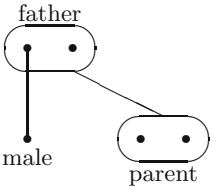


Fig. 3. The father relation.

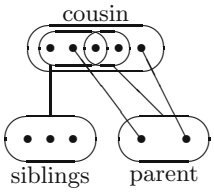


Fig. 4. The cousin relation.

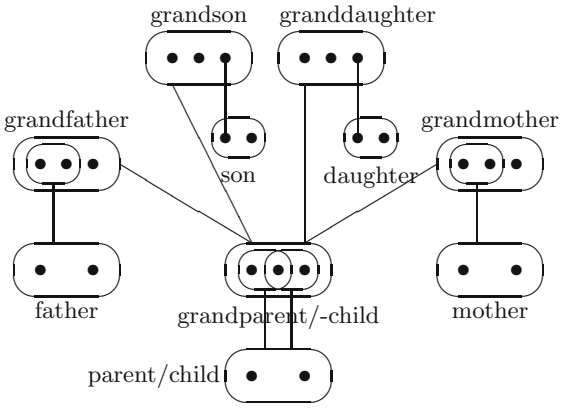


Fig. 5. Relation of kinship two generations above ego.

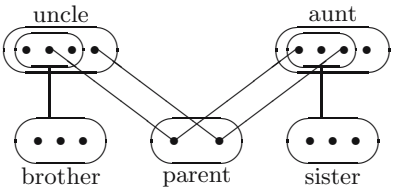


Fig. 6. Relation of uncle and aunt.

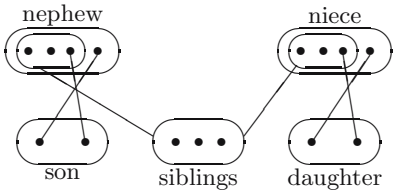


Fig. 7. Relation of nephew and niece.

nephew is determined in the first three places by *siblings*. The siblings' son is the nephew. This is determined by containment of the concept *son*. The third place of *nephew* thereby is determined by the second place of *son* (which is the *parent*), and the fifth place of *nephew* by the first place of *son*. Ego is here in the second place of *nephew*.

Finally the relations of the five-place concept will be considered. They are given in Figure 4. Ego is in the fourth place. His or her parents' sibling is the parent of the cousin, which is determined in the fifth-place.

All relations of the above definitions have now been put into diagrams. It is, of course, possible to assemble them all together in one diagram.

At the end of this section it is reasonable to give a brief account of how concepts are applied to extension, that is, the objects. A one-place concept can be applied to individuals only. It has to be assumed – and this is never part of the calculus – that a relation exists that enables the assignment of concepts to objects. This relation exists, usually, in the ability of a person or a system to use concepts correctly. This is not a question of logical or formal concern at all. N-ary concepts are applied to n-tuples of objects. To return to the example of family reunion: to find out whether two persons are cousins five people have to be found of whom the explicated relations hold.

4 Concept and linguistic sign

How the concept structure is built up by the definitions of the concepts is one side of the coin. The other side, which was another point of criticism, has not been addressed yet: How is the linguistic expression related to the concept? To stick to kinship relations: There are, for example, several expressions connected to the concept *father*: "father", "dad", "daddy", "pop", "old man". These terms, extensionally interpreted, do not denote different kinds of persons. But they are *used* in different circumstances for the same kind of persons. The term "old man" might be used among adolescent people, when they talk about their fathers. But they would not use this term when they speak to their fathers. The term "daddy" is only used by girls (Schusky (1972), 13) not by boys in referring to their father. There are many more aspects which determine the choice of expression. Especially kinship relations of different cultures are truly a treasure trove for very subtle differences of lexical coding. This cannot and should not be incorporated into the same concept system.

This insight rules out the possibility to label a concept like *father* with a set of expressions like "father", "dad", etc. Another concept system has to be assembled, coding the knowledge that is responsible for the choice of expression. The application of the concepts of this kind of concept system is not what is commonly understood by objects, viz. persons, entities, or everyday situations. The application now involves the concepts of the concept system of the 'real world' itself as objects, and the linguistic expressions as objects.

Also some 'real world' reference is necessary. When to use the term "daddy" would then be a concept having in one place a determination about social circumstances like *in family*, in the second place a concept about a concept of another concept system, the 'real world' concept system which is applied, the *father* concept, and finally in the third place the correct expression, "dad", to speak to a person to whom the concept of the second place applies in a situation like the one determined in the first place, *in family*.

5 Summary

This article introduced an intensional logic calculus and applied it to semantic analysis. It is suggested to substitute the popular meta-linguistic relations completely by the containment relation of the concept calculus, and to shift the mapping between lexicon and concept system to another concept system of language use.

The reduction to one type of relation enables research to develop mathematical characteristics which allow to draw conclusions about a language. Statistics about different types of relation are merely valuable in a technical sense. Their significance across the relation types is doubtful. The uniformity of the structure eases the scalability of implementation. The lexicon and the concept system can be extended by new lexemes and concepts without changing the overall structure of the system.

References

CRUSE, A. (1986): *Lexical Semantics*. Cambrigde Univ. Press, Cambridge.

FELLBAUM, C. (Ed.) (1998): *WordNet. An Electronic Lexical Database*. MIT Press, Cambrigde Massachusetts.

GANTER, B. and WILLE, R. (1996): *Formale Begriffsanalyse*. Springer, Berlin.

KAUPPI, R. (1967): Einführung in die Theorie der Begriffssysteme. *Acta Universitatis Tamperensis, Ser. A, 15*, Tampere.

MURPHY, M. L. (2003): *Semantic Relations and the Lexicon*. Cambridge Univ. Press, Cambridge.

PRISS, U. (1998): *Relational Concept Analysis. Semantic Structures in Dictionaries and Lexical Databases*. Shaker, Aachen.

SCHUSKY, E. L. (1972): *Manual for Kinship Analysis*. Holt, Rinehart and Winston, New York.

VOSSEN, P. (Ed.) (1998): *Special issue EuroWordNet*. Kluwer, Dordrecht. In: Computer and the Humanities. Vol. 32, 2/3.

WIERZBICKA, A. (1996): *Semantics. Primes and Universals*. Oxford Uinv. Press, Oxford.

WordNet 1.7.1. Princeton Universtity. `http://www.cogsci.princeton.edu/~wn/`

Automated Detection of Morphemes Using Distributional Measurements*

Christoph Benden

Department of Linguistics - Linguistic Data Processing,
University of Cologne, 50923 Köln (cbenden@spinfo.uni-koeln.de)

Abstract. To simply take the distribution of linguistic elements as a basis for analysis was the methodological prime of researchers of the so-called "American Structuralism". This paper deals with the detection of morphemes from a large corpus of German by simply applying a distributional procedure of counting the number of potential successors of a given sequence of letters of a word, a method reminiscent of proposals by Harris, Shannon and others. Morphemes can be heuristically read off by an increase in the potential successor count. Three different methods of identifying morpheme breaks are discussed and a proposal for improvement of the method by transforming graphemic to partial phonemic representation is put forward.

1 Overview and introduction

The paper deals with a method of detecting morphemes by segmenting words into parts – ideally morphemes – following a distribution-based algorithm originally developed during the 1940ies and 1950ies by Zellig S. Harris and other researchers.[2]
Segmenting words in linguistically valid units is a task that has been neglected in computational linguistics, especially for poorly documented languages (section 2). The historical background and a short overview of the distributional paradigm are given in section 3. The main part of the paper is dedicated to the basic method, demonstrated by a few examples, and to some direct refinements of the algorithm (sections 4, 5). In section 6 a proposal for refining the distributional analyses of the graphemic representations by (partly)

* A. Fenk pointed out to me that the method described does not strictly speaking use an "information theoretical measurement" as the original title suggested. I agree to this appraisal and accordingly replaced the term with "distributional measurements" which – ultimately for historical reasons – might be more appropriate. Thanks to Gustav Vella for painstaking corrections of my "Enklisch".

[2] To my knowledge, this algorithm has up to now not been aplied to a large corpus. Some work on distributional analysis has been done by Déjean (1998), but with a somewhat different focus. The properties of language exploited here are of course well known (e.g. Shannon (1950) and many more) and are somewhat remniscent of Markov processes.

converting them into a phonemic representation is suggested.
The corpus used for the following analyses consists of about 294.000 newspaper articles of different length, with a total of about 86 mil. tokens and about 1.8 mil. types. Here only those types with a token count $> 10 = 201.000$ were used.

2 Why bother with the segmentation of words at all?

In the context of information retrieval Manning & Schütze (1999, 132f.) concisely formulate why adequate segmentation is necessary in Natural Language Processing:

> [M]ost retrieval studies have been done on English – although recently there has been increasing multilingual work. English has very little morphology, and so the need for dealing intelligently with morphology is not acute. Many other languages have far richer systems of inflection and derivation, and then there is a pressing need for morphological analysis. A full-form lexicon for such languages, one that separately lists all inflected forms of all words, would simply be too large. [...] [I]n the languages with 'conjunctive' orthographies, morphological analysis is badly needed.

The problem grows if one takes into account that Manning and Schütze only refer to some kind of "item-and-arrangement" morphology without the everyday linguistic phenomena like morphophonology, (case) syncretism, ablaut etc.. As an admittedly rather extreme case compare Mohawk (Northern Iroquoian, Canada/USA). The word for "stove polish", for instance, (lit. 'one makes it shine by blackening that what makes heat (in) the house'), comsisting of at least 14 morphemes demonstrates the point (not every detail is pointed out here e.g. for the pronominal prefix *ion-*, for *-hon'tsihsta-* 'blacken', the quite complex 'epenthetic' *-tshera-*):

ion–	t(e)–	nonhs(a)–	'tarih–	(a)'t–	(a)hkw(a)–	tshera–
XxA(>NsU)–	SRFLX–	house–	heat–	CAUS–	INSTR–	E–

hon'tsihsta–	tshera–	hstar–	a'the–	't–	(á)hkw–	a'
blacken–	E–	???–	shine–	CAUS–	INSTR–	HAB

3 The historical background of research: Distributional analysis

Zellig S. Harris was the main figure of the so called "distributionalism", somewhat pejoratively also dubbed "taxonomic linguistics", a branch of "American structuralism" whose two other leading proponents were L. Bloomfield and E. Sapir. Harris characterizes the "distributional" program as follows:

> [E]ach language can be described in terms of a distributional structure, i.e. in terms of the occurrence of parts (ultimately sounds) relative to other parts, and [...] this description is complete without intrusion of other features such as history or meaning. [...] All elements in a language can be grouped into classes whose relative occurrence can be stated exactly. However, for the occurrence of a particular member of one class relative to a particular member of another class it would be necessary to speak in terms of probability, based on the frequency of that occurrence in a sample. (Harris (1954), 3f.)

Every statement about a linguistic element had to be made with respect to its distribution, mainly using substitutability of elements in fixed environments as a class-defining property. The main advantage is the fact that the degree of the categories thus (extensionally) established is the highest possible. This approach stands in sharp contrast to the traditional practice, which is still common, of mixing and inconsequently applying 'operational' criteria by e.g. determining parts-of-speech, where semantic, morphologcial and syntactic criteria are half-heartedly applied, although the principal necessity for a more reasonable classification, at least for scientific purposes, is always mentioned (cf. for German: Duden (1998), Eisenberg (1998), Engel (1988), Zifonun et al. (1997) etc.).
The results of distributional-analyses in the field of linguistic categorization are almost always more complex than traditional classifications, cf. Bergenholtz and Schaeder (1977), who suggest 51 part–of–speech–categories for German, a number which would be even higher if the distributional method they adopted would have been applied with more rigor:

> The criterion of distribution can be applied more or less consequently. A part-of-speech-system strictly developed on the basis of distributional criteria would probably contain far more than 100 part-of-speech-categories (Bergenholtz and Schaeder (1977), 14, my translation).

One should bear in mind that such a number of categories is not cognitively implausible, but scientifically difficult to handle, at least within conventional paradigms. This situation could be improved by complementing traditional linguistics with computational means. The present work is part of such an attempt.

4 Basic method

The proposals roughly follow the original suggestions by Harris (1951, 1954). "Grapheme" was not the unit originally focussed on by Harris, but it is used here because the corpus consists of written language.

1. Of the possible combinations of graphemes of a natural language L, only a tiny fraction is used (partly due to phonological restrictions, but mostly for no systematic reason: these combinations are simply not in use).

2. Let G be the set of graphemes L possesses. The occurrence of any grapheme $g \in G$ in an arbitrary position of a sequence of elements $s = g_1$, ..., g_n is not random: It depends on the valid sequences of graphemes forming the morphemes L possesses.
3. Assumption: With growing length of s, the number of different g that can follow the graphemes constituting s (called successor values, SV for short), and which thereby form (part of) a valid verbal sequence of L, tends to decrease.
4. If (III) formulates a valid tendency for the distribution of graphemes within a morpheme, (a) words cosisting of more than one morpheme should show an overall decreasing number of successors for increasing number of graphemes (b) with increasing counts or local maxima of the count of possible successors at the boundaries between two morphemes, that is as SV of the last grapheme of the first morpheme under consideration.
5. Assumption: For any sequence of graphemes s, the higher the index i of g_i becomes, the less significant the SV becomes with respect to a morpheme break. This specifies assumption (4a) because the tendency for longer words is for the later SV to become one, cf. (2).

As an illustration of (IVa, IVb) we will look at the SV for *Vorstellung* in (1). Complete SV-analyses for two selections of the corpus had to be performed beforehand. "Selection 1" refers to the values for a selection of all words of the corpus, "Selection 2" refers to a selection of words with more than 10 tokens (Only Selection 2 will be referred to in the following; the SVs which, at first sight, may appear improbably high, like 36, 30 etc. but correspond to the given data since graphemes like *á*, *é* etc., as well as the end of word are also taken into account).

(1)

v	o	r	s	t	e	l	l	u	n	g	
36	30	31	15	12	8	4	9	2	4	7	SV for Selection 1
30	21	28	9	7	3	1	5	1	1	3	SV for Selection 2

(IVa) and (V) might be illustrated for *Hausmüllverbrennungsanlage* in (2) a word that had to be chosen for illustrative purposes because the proposed morpheme breaks are often better than purported by (V), cf. *Abgeschlossenheitsbescheinigungen* in (3).

(2)

h	a	u	s	m	ü	l	l	v	e	r	b	r	e	n	n	u	n	g	s	a	n	l	a	g	e
31	28	17	26	6	1	1	4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

(3)

a	b	g	e	s	c	h	l	o	s	s	e	n	h	e	i	t	s	b	e	s	c	h	e	i	n	i	g	u	n	g	e	n
29	29	9	22	9	1	11	4	1	1	1	1	4	1	1	1	2	3	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1

First algorithm. The first proposal is not really the search for local maxima but it simply starts with the SV of the leftmost element and sets the

morpheme borders after the first grapheme whose SV is higher than that of its predecessor.
The method yields some unwanted results, but is more plausible than would be thought at first sight because it accounts for the "Fugenelement" *-s-* in *Abgeschlossenheit-s-bescheinigungen*, which none of the other algorithms do, by simply looking at the SV. The "Fugenelement"–pattern is easy to explain: While the stem *Abgeschlossenheit* only allows for derivation and inflection, that is for a small number of extensions, the stem for composition, just derived by adding *-s-*, allows for potentially every grapheme to follow it in compounds – although accidentally only three graphemes actually do turn out to follow *Abgeschlossenheits-*.

5 Refinements of the evaluation

A few words on the evaluation of the results seems appropriate. There does not appear to be a general method to evaluate the quality of the results, other than by previously establishing a complete list of morphemes with other means .
So points for refinements are "hand-picked" and only a few characteristic sequences and positions are discussed in the following (compound words with special features, short (= one- or two-segmental) prefixes at the beginning of a word, etc.).
The first experiment showed a weakness e.g. with regards to the suffix *-heit-* which is sometimes spelled with a final *-d* in the corpus (because of the Afrikaans orthography of *Apartheit*):

(4)

a	p	a	r	t	h	e	i	d
29	16	5	2	4	1	1	2	7

The word is segmented as *apart-hei-d* because of the SV of *-i-* which is higher than the SV of *-e-*, but the final *-d* has an even higher successor-value than the *-i-*. We do not come up with a suffix *-heit* but with a mutual suffix *-hei*.
The number of wrong segmentations in this guise was 66 types/6314 tokens.
Apparently, one should not simply search for the first grapheme with an SV higher than that of its predecessor but for the first local maximum of SV:

Second algorithm. The second proposal is a slightly enhanced version of the first experiment. Instead of taking the first grapheme with a SV higher than that of its predecessor, the next local maximum is looked up and this grapheme is taken to be the last grapheme of the morpheme under consideration. The "local maximum" is defined as the first SV that has no greater SV for the graphemes immediately to the left or right. This proposal seems to realize the original idea of Harris best.

For words like *Apartheid* we now get the better segmentation *apart-heid*. Although the erroneous segmentations are depleted (17 types/2331 tokens), we get stuck with words like *Eigenheit*, cf.:

(5)
a	p	a	r	t	h	e	i	d
29	16	5	2	4	1	1	2	7

(6)
e	i	g	e	n	h	e	i	t
30	19	4	2	20	2	1	2	2

It seems reasonable to assume that the farther right in the sequence of SV one gets, the higher the need of a careful evaluation of what constitutes a morpheme break and what doesn't. The next variant of the algorithm changes the search for local maxima to the search for ranges of local maxima like in *Eigenheit...*, *Reinheit...*, *Gottheit...* .

Third algorithm. For the third proposal, the algorithm searches from left to right for increasing values for the SV until a local maximum is reached and places the break after this maximum or at the end of the word, if it is reached: `... 5 3 1 1 2 4 | 2 ...` . Alternatively, it looks for sequences of SV that constitute ranges of local maxima. The morpheme break is placed after such a range or at the end of the word if it is reached: `... 5 3 1 1 2 3 3 | 2 ...` .

The result for the third proposal seems to be the optimum that can be achieved by a direct evaluation of the SV-measure for *-heit* (11 types/1570 tokens).
The three proposals for drawing morpheme breaks for large parts of the words yield the same result. The "Fugenelement" was an example (cf. segmentation of *Ver-band-s-ge-meinde* by the first algorithm vs. *Ver-bands-ge-meinde* by the second and third algorithm).
Another case of interest might be the following: *verschaffen* is segmented differently by the second algorithm (*ver-sch-aff-en*) and by the third algorithm (*ver-sch-affen*) because of the existence of an erroneous *verschaffen.* (with a final dot, due to inadequate preprocessing that only takes dots followed by white-space as sentence boundary).

(7a)
v	e	r	s	c	h	a	f	f	e	n
30	24	30	15	1	13	7	1	2	2	2

(7b)
v	e	r	s	c	h	a	f	f	e	n	.
30	24	30	15	1	13	7	1	2	2	2	1

The different algorithms show different behavior, partly wanted, partly unwanted. It is a matter of future work to refine the algorithms by combining the wanted effects of the evaluation procedures. One should bear in mind that these results have been obtained without the use of a parser that reapplies already established units.

6 Transferring graphemic to phonemic representation

An interesting problem is the predicted morpheme *sch*, clearly no candiate for a German morpheme. This problem is basically a mapping problem: Between

graphemic and phonemic representations in alphabetic writing systems, a mapping relation should be such that for any grapheme one can specify the respective phoneme and vice versa. As is well known, this mapping relation is somewhat defective (for an overview compare Eisenberg (1998)).

By looking at the SV for the word initial grapheme sequence <sch> we get segmentations that are implausible and should be resolved (the segmentations (not the SV!) are the same for the three algorithms), cf. (8).

(8a)	s	c	h	o	n		(8b)	v	e	r	s	c	h	o	b		(8c)	w	u	n	s	c	h	
	34	14	18	15	12			30	24	30	15	1	13	5	2			28	17	2	3	1	14	

Fourth algorithm. Following a table of grapheme-phoneme-correspondences for the language in question, replace the graphemic representations by the respective phonemic representation (or an arbitrary symbol that acts as a placeholder). Use as the SV of the new element the last value of the replaced grapheme, or, if it is a sequence of graphemes, the value of the last grapheme replaced.
We thereby get the segmentations in (8), using <S> as replacement for <sch>.

(9a)	S	o	n				(9b)	v	e	r	S	o	b				(9c)	w	u	n	S			
	18	15	12					30	24	30	13	5	2					28	17	2	14			

The grapheme-phoneme-correspondences are usually not that trivial. Simply replacing the <sch> by a single symbol, say <S>, in German is a relatively reliable process (disregarding diminutives like *Häschen*, *Küsschen*). Replacing the <S> for the correct instances of <s> in the digraphs <sp> or <st> needs a high level of sophistication, above all because of the high productivity of compounding in German. Nonetheless, it appears that the effort to exploit the relatively transparent correspondences between graphemes and phonemes in alphabetically written languages will improve the results.

7 Concluding remarks

The method presented for the segmentation of words in linguistically relevant units, that is morphemes, is essentially distributional and, although not explicitly stated, exclusively uses a computational version of the substitution-test (i.e. which grapheme *c*an occur in a given position of a sequence and which grapheme *g'* can replace *g*, yielding a correct (partial) grapheme sequence of the given language).
SV-evaluations provide a simple method to segment words into morphemes ("simple" not only with respect to the method itself but also with respect to the implementation). Only universal (and orthographic – but this belongs to another area) assumptions have been used, thus the method is easily transferable to data of other dialects or languages. We envisage to next apply the methods to the agglutinative languages Basque and Turkish.
SV-evaluations are especially advisable in contexts, where

1. Non-idiosyncratic (non-language-specific) methods should be applied to gain a possibly new view and understanding of already 'established' analyses. SV-evaluation seems to be a good starting point.
2. No large lexicons, morphological descriptions etc. are available, to reach an at least rough overview of the morphological setup of a language (one of the main aims of distributional analyses from the outset).

One hypothesis of Harris (1968) was that the imbalance of the distribution of vowels and consonants due to the syllabic setup of languages would lead to imbalances in the analysis. Although not further discussed, this does not seem to be the case (dipthongs might prove problematic but haven't been tested yet).
Of course, many improvements are conceivable:

1. Improvements of the 'feeding' components such as the deployed preprocessor, the corpus itself etc.
2. Further elaboration of the evaluation algorithms; up to now only direct countings and direct segmentation have been taken into account.
   - A parser which uses already well established morphemes to segment longer, suspiciously unsegmented, sequences; this way the analysis develops by repeated 'self precisioning'.
   - Rising values in the SV-sequences seem to be meaningful (as well as overproportional descent for absolute prefixes not discussed yet). These could be brought to use via more complex mathematical evaluation, for instance, the comparison of the proportional gradient of different rising sequences.
   - By reversing the graphemic sequence of a word one will be able to analyze the suffixes of longer words, a proposal originating from Harris and mentioned by Déjean (1998) as an improving factor.
3. Further elaboration of the reduction from the graphemic to a phonemic representation will surely render better results.

References

BERGENHOLTZ, H. and SCHAEDER, B. (1977): *Die Wortarten des Deutschen*. Klett, Stuttgart.

DÉJEAN, H. (1998): Morphemes as Necessary Concepts for Structures Discovery from Untagged Corpora. *Workshop on Paradigms and Grounding in Natural Language Learning*. Adelaide, 295–299.

EISENBERG, P. (1998): *Grundriß der deutschen Grammatik. Band 1: Das Wort*. Metzler, Stuttgart.

HARRIS, Z. (1951): *Methods in Structural Linguistics*. University of Chicago Press, Chicago.

HARRIS, Z. (1954): Distributional Structure. *Word, 10.2-3, 146–162*.

MANNING, C. D. and SCHÜTZE, H. (1999): *Foundations of Statistical Natural Language Processing*. MIT-Press, Cambridge, MA.

SHANNON, C. E. (1950): Prediction and Entropy of Printed English. *Bell System Technical Journal, 3, 50–64*.

Classification of Author and/or Genre? The Impact of Word Length

Emmerich Kelih[1], Gordana Antić[2], Peter Grzybek[1], and Ernst Stadlober[2]

[1] Department for Slavic Studies, University Graz, A-8010 Graz, Austria
[2] Department for Statistics, Technical University Graz, A-8010 Graz, Austria

Abstract. 190 Russian texts – letters and poems by three different authors – are analyzed as to their word length. The basic question concerns the quantitative classification of these texts as to authorship or as to text sort. By way of multivariate analyses it is shown that word length is a characteristic of genre, rather than of authorship.[3]

1 Word length and the quantitative description of text(s) and author(s)

This study focuses on word length. Word length is a central characteristic in the framework of quantitatively oriented linguistics. In fact, the study of word length can be traced back to a hundred year long tradition (as to a historical and methodological survey of these studies, cf. Grzybek (2004)). Knowing this historical background, it is evident that word length, as it is studied today, is no isolated characteristic.[4]

The basic question of the present study is to what degree word length may contribute to the discrimination of authors and genres. An answer to this question will not only shed light on specific factors influencing word length; it will also provide an argument if word length is an appropriate variable to describe an author's individual style, or the stylistic traits of specific genres.

The discussion of these questions has a history of its own: as opposed to the field of *quantitative typology of texts* (cf. Alekseev (1988), Pieper (1979)), approaches in the realm of *stylometry* (cf. Martynenko (1988)) assume that the individual style of texts and/or authors can be quantitatively described. Part of this research has concentrated on the question of authorship attribution, particularly applying quantitative methods to decide

[3] This study has been conducted in context of research project # 15485 (Word Length Frequencies in Slavic Texts), financially supported by the Austrian Research Fund (FWF); cf.: `http://www-gewi.uni-graz.at/quanta`.

[4] Within a synergetic approach, word length is closely interrelated with other linguistic levels and units, and it is well known that word length interacts, e.g., with the number of phonemes (in a given inventory), with lexicon size (cf. Köhler (1986)), with polysemy (cf. Altmann et al. (1982)), or word length and word frequency (Strauss et al. (2004), with a survey of the Zipfian tradition).

doubtful cases of authorship (cf. Marusenko (1990)). In a way, these approaches have paved the way for contemporary research in the field of computer linguistics, where related problems are being discussed under the heading of automatic authorship attribution and text categorization. The status of this contemporary research may be characterized by two tendencies. On the one hand, word length is not at all taken into consideration; in this case, researchers assume word length to be a "low-level phenomenon" (cf. Stamatatos et al. (2001), 195), which leads to no reliable results, neither for text categorization nor for authorship attribution. On the other hand, word length is taken into account as one possible variable among others (such as, e.g., sentence length, lexical type-token ratio, adverb counts, etc.) for multivariate discriminant analyses (vgl. Karlgren and Cutting (1994)). As to this line of research, there are a number of methodological problems which have not been sufficiently reflected:

1. More often than not, word length has been measured as the number of characters per word; it is a well-known fact, however, that for most languages, measuring word length as the number of characters (letter, graphemes) per word is no appropriate procedure leading to erroneous results due to the instability of the graphemic system (cf. Kelih and Grzybek (2004));
2. Most of the studies in this field do not analyze the impact of word length as a variable in its own right, but only as part of some undifferentiated pool of variables.

This situation gives rise to a new systematic study of word length as a possible discriminating variable for authorship attribution and/or text categorization, paying due attention to and avoiding the methodological flaws of the studies mentioned above.

2 A case study: text basis and analytical options

With regard to the problems discussed above, the present study proceeds as follows:

a. Word length is measured as the number of syllables per word; 'word' is thus understood as an orthographical-phonological unit, the systematic changes of which, depending on linguistic definitions, are well known as well (cf. Antić et al. (2004)).
b. Discriminant analyses are undertaken, taking into consideration only variables which are directly related to or derived from the frequency distribution of x-syllable words in a given text.

In the present study, the word length of 190 Russian texts is analyzed. These texts are systematically chosen in order to design a balanced study, based on an approximately equal number of two different text types, written

by three different authors. By way of multivariate methods, the role of word length as a characteristic of authorship or of text type shall be studied.[5]

In order to study the relevance of word length on the level of text sorts and authors, respectively, ca. 30 texts written by three well-known Russian authors each (A. S. Puškin, D. Charms, and A.A. Achmatova) in two different sorts of text (poems and letters), are considered. On the basis of this text sample, a number of different analytical options are at our disposal (cf. Figure 1). These options include the discrimination

- of authors within a given genre (i.e., studying only letters or poems, respectively);
- of different texts sorts written by different authors (e.g., Charms' private letters in contrast to Achmatova's poems);
- of text sorts without paying attention to authorship.

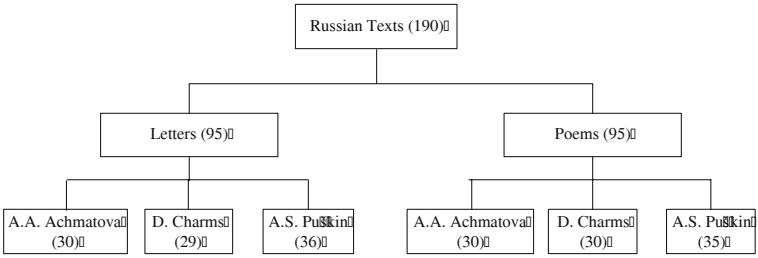


Fig. 1. Graphical Representation of the Text Data Base

3 Methods of text discrimination

As to the discrimination of author and/or text, we want to concentrate on the impact of word length, only. Therefore, from our pool of 30 possible discrimination variables, all those variables which are related to other characteristics of a text (such as, e.g., text length), will be excluded, as well as variables which, though primarily characterizing word length, have such factors as indirect components.[6]

[5] The text basis is part of the text data base developed in the research project mentioned above.

[6] Text length is, of course, an important characteristic of a text, and has well been used in other studies on authorship or genre discrimination (cf. Djuzelic (2002)). Although in our case, the average text length of the letters ($\bar{x} = 238.20, s = 170.37$) does not significantly differ from that of the poems ($\bar{x} = 204.37, s = 178.59$) – as can be shown by a Mann/Whitney U-Test ($z = -1.56, p = 0.12$) – we have focused on word length, only, in order to strictly control the impact of this variable.

3.1 Quantitative measures for text analysis

Each text contains N words (w_i for $i = 1, 2, \ldots, N$). Word length (x_i) is measured in the number of syllables per word ($x_i = j$ where $i = 1, 2, \ldots N; j = 1, 2, \ldots K$). Actually we are dealing with words of $1, 2, 3, \ldots$, or K syllables. Words are divided into K frequency classes; f_j refers to the number of elements that belong to the same class (absolute frequencies). Texts will be quantitatively described by a number of measures reflecting the moments of the word length frequency distribution.

Not all variables which possibly describe the distribution are equally important for our study; our aim was to find a minimal set of variables, relevant for discriminant analyses (thus having the strongest classification power). On the basis of our empirical tests, we obtained a set of six variables, which are appropriate for our purposes. The definitions of these six variables are listed in Table 1.

Table 1. Six statistical measures characterizing 190 Russian texts

Variable	Formula	Explanation
m_2	$= s_0^2 = 1/N \cdot \sum_{i=1}^{N} (x_i - \bar{x})^2$	empirical variance of the word length
m_4	$= 1/N \cdot \sum_{i=1}^{N} (x_i - \bar{x})^4$	fourth central moment
v	$= s_0/m_1$	coefficient of variation
d	$= m_2/(m_1 - 1)$	quotient of dispersion
o_i	$= m_2/m_1$	first criterion of Ord
p_4	$= f_4/N$	relative proportion of 4-syllable words

Every text, now, can be seen as a statistical object incorporating its information in the six variables listed in Table 1. Thus, the quantitative description of a given text j, belonging to group i, is given by an observation vector of dimension 6 (for $i = 1, 2\,;\, j = 1, \ldots, 95$):

$$\mathbf{x}_{ij} = (m_2(i,j), m_4(i,j), v(i,j), d(i,j), o_i(i,j), p_4(i,j))$$

For each group, the mean values of the six variables are combined in the mean vector of the same dimension (for $i = 1, 2$):

$$\bar{\mathbf{x}}_i = \left(\bar{m}_2(i), \bar{m}_4(i), \bar{v}(i), \bar{d}(i), \bar{o}_i(i), \bar{p}_4(i)\right)$$

Table 2 represents one example, including two Russian texts with all six statistical values discussed above. Actually, there are 95 texts from both genres in our text corpus. $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ denote the mean vector for the text groups, i.e., *letters* and *poems*, respectively, and they are calculated for all 95 texts of each group.

Table 2. Six statistical measures of two Russian texts for both text types

Text type	m_2	m_4	v	d	o_i	p_4
Letter #1	1.26	6.53	0.55	1.23	0.62	0.07
Letter #2	1.37	7.07	0.50	1.01	0.58	0.16
$n_1 = 95$; $\bar{\mathbf{x}}_1 =$	(1.47	7.86	0.53	1.17	0.64	0.11)
Poem #96	0.81	2.04	0.45	0.83	0.41	0.04
Poem #97	0.86	2.92	0.49	0.97	0.46	0.04
$n_2 = 95$ $\bar{\mathbf{x}}_2 =$	(0.92	2.57	0.47	0.88	0.45	0.06)

3.2 Discriminant analysis

In a first step, the texts are discriminated along the category of 'author', only. In this case, each of our three authors – A.A. Achmatova {A}; D. Charms {C}; A.S. Puškin {P}) – is treated as a separate class, and no genre distinction is taken into consideration. As can be seen from Table 3[1], this results in a percentage of only 38.4% correctly discriminated texts.

As can also be seen from Table 3[2], this poor result can be improved up to a percentage of 56%, if 'genre' is additionally taken into consideration. In the next step concentrating on one particular text group (i.e., either letters or poems), and testing each combination of two authors, one obtains definitely better results between 63% and 77% (cf. Table 3[3,4]). Finally concentrating on one individual author, only, and juxtaposing letters vs. poems, one gets even better results up to a percentage of 82% to 93% correctly classified texts (cf. Table 3[5]).

This overall result is a clear indication for word length being dependent on the type of text, rather than on authorship (i.e. being a good variable for text categorization, rather than authorship attribution).

3.3 Statistical distance as a measure for data discrimination

Given these findings, it is important to see which relevant variables are appropriate for discriminant analyses. The univariate distance is an important measure for separating data corpora into two different text groups. Let us assume that the texts are independent samples $(x_{1_1}, \ldots, x_{1_{95}})$,$(x_{2_1}, \ldots, x_{2_{95}})$ of two distributions, which have possibly different theoretical means μ_i and the same variance σ^2. The theoretical means will be estimated by the arithmetic mean $\bar{x}_i$ of the sample, and the variance by pooling the two empirical variances s_i^2 of the sample as follows:

$$s_{pool}^2 = \frac{1}{n_1 + n_2 - 2} \left((n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2 \right)$$

The univariate statistical distance D is given as:

$$D\left(\bar{x}_1, \bar{x}_2\right) = \frac{|\bar{x}_1 - \bar{x}_2|}{s_{pool}}$$

Table 3. Discriminant Analyses: Author vs. Genre

	Text Type	Author	Classification
1		{A}{C}{P}	38.40%
2	{Letters}{Poems}	{A}{C}{P}	46.30%
	{Letters}	{A}{C}{P}	55.80%
	{Poems}	{A}{C}{P}	54.70%
3	{Letters}	{A}{C}	62.70%
		{A}{P}	71.20%
		{C}{P}	67.70%
4	{Poems}	{A}{C}	76.70%
		{A}{P}	0.00%
		{C}{P}	73.80%
5	{Letters}{Poems}	{A}	81.70%
		{C}	93.00%
		{P}	93.20%

The distance D between two groups is thus defined as the distance between the group centers (means), standardized by the pooled variance. Table 4 contains mean values, standard deviations and univariate statistical distances for all six variables; also, results are given for all pairwise comparisons between these two text groups.

Table 4. Means, standard deviations and univariate statistical distances for pairwise comparisons (letters vs. poems)

Variable	Text type	$\bar{x}_1 \mid \bar{x}_2$	$s_1 \mid s_2$	$D(\bar{x}_1, \bar{x}_2)$
m_2	Letter	1.47	0.43	5.20
	Poem	0.92	0.17	
m_4	Letter	7.86	6.75	0.23
	Poem	2.57	1.09	
v	Letter	0.53	0.06	24.87
	Poem	0.47	0.03	
d	Letter	1.17	0.15	16.53
	Poem	0.88	0.11	
o_i	Letter	0.64	0.11	23.66
	Poem	0.45	0.06	
p_4	Letter	0.11	0.04	**36.17**
	Poem	0.06	0.03	

Table 4 shows the highest distance value D, based on the variable p_4 (i.e., the relative frequency of 4-syllable words). This means that variable p_4 has the strongest power for the separation of our text corpus into two groups: p_4 thus is the best discriminator for these two text groups.

The fourth central moment (m_4) has the lowest discrimination power, what implies a bad separation. The reason for this is the fact that although variable m_4 has the highest mean value, it has as large statistical deviation, which keeps the distance relatively small. Knowing that these two text groups remarkably differ as to the proportion of 4-syllable words, this result was to be expected. With variable p_4 alone, up to 76.3% of our texts can be correctly classified: combining p_4 with variable d, the percentage of correctly classified items improves to 89.5%. In Figure 2, variable p_4 is plotted against variable d for the two categories *letters* and *poems*.

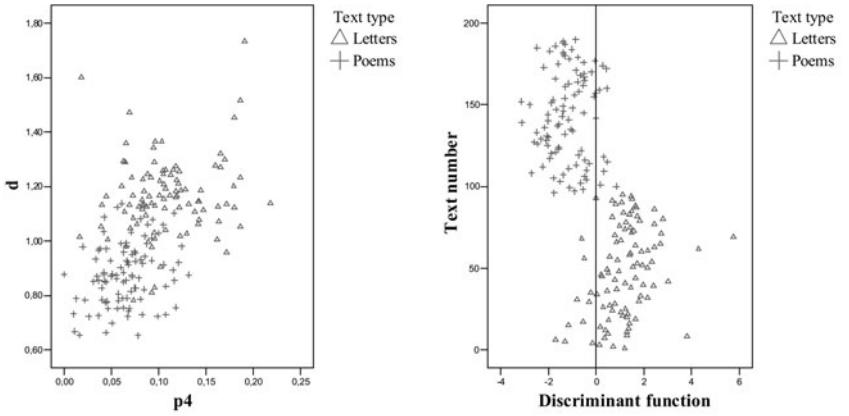


Fig. 2. Left scatter plot p_4 vs. d; right separation of letters and poems

Figure 2 illustrates the fact that it is possible to separate *letters* from *poems*. The linear discriminant function is calculated as a linear combination of relevant variables. In our case, the set of six variables is reduced to a set of two relevant variables, namely, p_4 and d. Figure 2 also shows the good separation power of the discriminant function. The cut point between the two groups is represented by the vertical line in 0, which marks the separation. Each point represents a text; the text numbers can be seen on the y-axis. Every text has different values of p_4 and d, so the value of the discriminant function is also different for each text: we can see two clearly separated groups. We can notice that only nine *poems* and eleven *letters* are misclassified. This corresponds to a high percentage of correct classifications, which sum up to 90.5%, or 88%, respectively.

4 Summary

Our study clearly shows that word length, if properly defined as the number of syllables per word, has a strong discriminating power for text categorization: with only two variables, a percentage of up to 90% correctly discriminated texts can be obtained. As opposed to this, word length does not seem to play an important role as to questions of authorship attribution.

References

ALEKSEEV, P.M. (1988): *Kvantitativnaja lingvistika teksta*. Leningrad.

ALTMANN, G., BEÖTHY, E. and BEST, K.H. (1982): Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung, 35, 537–543.*

ANTIĆ, G., KELIH, E. and GRZYBEK, P. (2004): Zero-syllable Words in Determining Word Length. In: P. Grzybek (Ed.): *Contributions to the Science of Language. Word Length Studies and Related Issues.* New York, Springer. [In print]

DJUZELIC, M. (2002): *Einflussfaktoren auf die Wortlänge und ihrer Häufigkeitsverteilung am Beispiel von Texten slowenischer Sprache.* Dipl. Arbeit, TU Graz.

GRZYBEK, P. (2004): History and Methodology of Word Length Studies: The State of the Art. In: P. Grzybek (Ed.): *Contributions to the Science of Language: Word Length Studies and Related Issues.* New York, Springer. [In print]

KARLGREN, J. and CUTTING, D. (1994): Recognizing text genres with simple metrics using discriminant analysis. In: M. Nagao (Ed.): *Proceedings of COLING, 94, 1071–1075.*
`http://www.sics.se/~jussi/Papers/1994_Coling_Kyoto_1/cmplglixcol.ps`

KELIH, E. and GRZYBEK, P. (2004): Wortlänge in Silben und Graphemen. [In prep.]

KÖHLER, R. (1986): *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik.* Studienverlag Brockmeyer, Bochum. [= Quantitative Linguistics; 31]

MARTYNENKO, G.Ja. (1988): *Osnovy stilemetrii.* Leningrad.

MARUSENKO, M.A. (1990): *Atribucija anonimnych i psevdoanonimnych literaturnych proizvedenij metodami raspoznavanija obrazov.* Leningrad.

PIEPER, U. (1979): *Über die Aussagekraft statistischer Methoden für die linguistische Stilanalyse.* Tübingen, Narr. (= Ars linguistica, 5)

STAMATATOS, E.; FAKOTAKIS, N. and KOKKINAKIS, G. (2001): Computer-Based Authorship Attribution Without Lexical Measures. *Computers and the Humanities, 35, 193–214.*

STRAUSS, U., GRZYBEK, P. and ALTMANN, G. (2004): Word length and word frequency. In: P. Grzybek (Ed.): *Contributions to the Science of Language. Word Length Studies and Related Issues.* New York, Springer. [In print]

Some Historical Remarks on Library Classification – a Short Introduction to the Science of Library Classification

Bernd Lorenz

Fachbereich Archiv- und Bibliothekswesen
Fachhochschule für öffentliche Verwaltung und Rechtspflege

Abstract. Classification as a human activity in general becomes a scientific activity in librarianship. There are famous examples of this history of classification among them the schemes of Conrad Gesner (1548) and the Princeton University Library (1901). In present time we find a number of new tasks and obligations in this field.

1 Introduction

Librarians and library science scholars were possibly the first information specialists who developed theoretical approaches, practical tools and techniques for organizing and retrieving bibliographic documents and also the bibliographic data about them, above all more essentially information about their subject contents. Indeed "It may not bee too far fetched to say that the history of theoretical classification began with the division of knowledge into the knowledge of good and the knowledge of evil". With these words Ernest Cushing Richardson (1930) cites the text of 1 Moses 2,9 as the starting point of thinking and practicing classification in his basic work "classification" (written nearly one hundred years ago).

Moreover, we must remind Antony Flew's remarkable volume "An Introduction to Western Philosophy. Ideas and Argument from Plato to Sartre". Flew (1971) chooses the title "Classification as a human activity" as the title of a chapter of his book and coins with this title nearly a philosophical program.

Indeed a library is "a unique type of human organisation" (KASHYAP (2003)) and its classification, as a human activity in general, becomes a scientific activity in librarianship during the epochs of history. The need of organizing the contents of texts, books, libraries and so on requires thinking and competent working above all, not only an imitation of some classification of sciences and humanities.

Structuring and subject cataloguing was a permanent challenge in the history of libraries and very much related to the physical location of a book in shelves, stock-rooms, etc. Some examples of this history of classifications within the libraries are listed chronologically below.

In fact, the books in the great Byzantinic respectively Arabic libraries were sorted by subject in shelves or rooms. Smaller libraries only grouped the books by the main categories (e. g. clerical and secular, liturgical and dogmatic opuses, etc.) or other criteria (Lorenz (2003b, pp. 57 and 64)).

2 Classified arrangement in monastery libraries of the Middle Ages

In the Middle Ages - at least in the beginning - monastic libraries became the important part in librarianship. "In der That, ein Kloster ohne Bücher ist wie eine Festung ohne Waffen. Daher waren Mönche zun¨chst darauf bedacht, eine möglichst groe Bibliothek anzulegen." (Wetzel 1877) [Cite: Really a monastery without books is like a fort without arms. Due to this the monks considered to set up a library as large as possible.] This reflection started with the mighty deed of a whole catalogue, consisting of a sequence of single catalogues of different libraries. Maybe the Benedictine monks of St. Emmeram in Regensburg especially abbot Albert († 1358) started the distinguished attempt unique in the German librarianship of the 14th century, to compile the complete registration of all books of all the monasteries for monks in Regensburg in one single volume.
The subject order of the registration of literature in the medieval catalogues and projects followed also former examples, one is the "Biblionomia" by Richard de Fournival († 1260) using academic aspects and therefore his library consisted of three sections: philosophy, medicine/jurisprudence (scientiae lucrativae) and theology (starting with the profane literature).
Richard's combination of "septem artes liberales" and aristotelic-scholastic classification of science with the three university faculties at the top is known since the 13th century as a common and conventional scheme although it refers often to the shelving in medieval libraries.

3 Classified arrangement in private libraries of the Middle Ages

Besides monastic libraries there were also some important private libraries (Ludwig (1997), Lorenz (1997a)). One focus is the catalogue of the professor of medicine Amplonius Ratingk (1263/4-1435) in the 15th century who sorted his library according to 12 subjects, whereas the Nuremberg physician Hartmann Schedel (1440-1514) subdivided his catalogue into 22 subjects in accordance to the Richard de Fournival system. In contrary to the medieval usage the artes liberales were here at the beginning. Ratingk and Schedel also grouped the theological literature at the end of the classification scheme.

This scheme with its 22 topics displayed its own history. Three hundred years later it appeared again - as a product either of fortune or of decision - and was used by Ernst Gottfried Baldinger (1738 - 1804), professor of

medicine in Marburg/Lahn. Due to his large collection the main topics were subdivided.

4 Classified arrangement in the late Middle Ages and at the beginning of modern times

From the beginning, librarians and readers use as their basis the consistent classified arrangement - with some few local variations. The pattern was Bible, Fathers of the Church, other theology, profane literature. Within the subjects the arrangement was different. A correlation between the library classification and the academic classification of Isidor of Sevilla, Hrabanus Maurus or Vinzenz of Beauvais cannot be observed nor was the Bibliionomia of Richard de Fournival adapted. It is a moot point whether this opus which is cited in only one manuscript was publically known or only by its author/creator. Shelving could also be seen as a tool of subject cataloguing and the fact that the order of the books is irrelevant to inventory as long as they are in the same place for audit.
Already in the middle ages the basics of the modern three part cataloguing were set up: Shelf, author, subject.
With timely changes the classification was used in accordance with the faculty departments. The main categories from of the Middle Ages like theology, law and medicine were kept untouched whereas arts was split into different disciplines some of which became new main categories and the rest were subcategories.

As a famous example we reproduce here the system (created 1548 - 1549) of Conrad Gesner (Zürich, bibliographer and physician):

1. Grammatica	8. Astronomica	15. Metaphysica
2. Dialectica	9. Astrologia	16. Ethica
3. Rhetorica	10. Historica	17. Oeconomia
4. Poetica	11. Geographia	18. Politica
5. Arithmetica	12. Divinatio et Magia	19. Juriprudentia
6. Geometrica	13. Artes literates	20. Medicina
7. Musica	14. Physica	21. Theologia

5 Cataloguing in the 18th century

In the Renaissance time many libraries just needed the classified arrangement for browsing. Often catalogues existed but rather for inventory function. Many catalogues registered location and subject at the same time. Without any dramatic changes but with slight modification they were used in this epoch for indexing as well.
Step by step the raw systematic catalogues changed to dignified ones best seen at the university library of Göttingen in the 18th century. Classification

was not yet a common task but introduced at several libraries in different versions as, for example, the fine structured subject classification in Göttingen or as group marks number in Milan - in contrast to the individual marks for every book.

6 Systematic cataloguing in the 18th century

Two main aspects were relevant for the development of modern systematic cataloguing as it is still used today: the growing number of books in libraries and the change in the scientific and educational system, as to say the beginning of modern university, in the age of Enlightenment.

The universal philosophy of science in this era (eg. Leibniz) demanded "man müsse schon beim Betrachten einer Büchersammlung die ganze Literargeschichte wie in einem Spiegel aufgefangen vor sich sehen"(Legipontius (1747); see also Naud (1627)) [cite: Looking at a book collection should already give you an impression of the complete history of science and literature]. The libraries in the Baroque buildings represent glance and thinking of the era and - in this way - present the books excellently: This was the main aspect for the librarians much more than cataloguing.
In some cases additional numbers of classes were adequate, elsewhere reorganisation was done by systematic aspects. In 1694 Christoph Hendreich, a library worker at the elector's library at Cölln an der Spree (=Berlin), replaced the 6 main categories with 46 new ones - a nearly revolutionary act during the history of sciences and the history of classification. Meanwhile at Göttingen's university library a voluminous systematic catalogue was set up using the book marks mentioned above.

7 Subject cataloguing in the 19th century

At the beginning of the 19th century nearly every library had to reorganize its book shelving and indexing system. This was necessary either because of the recurring tightness in the classified arrangement and the full written catalogues, or because of the ongoing accretion of books due to secularisation. One exception again was the university library of Göttingen where cataloguing was done in an exemplary way by Heyne and Reußin 1776–1790 and so Göttingen became a centre of discussion in the progress of German libraries. Some theoreticians in the 19th century accentuated that systematic cataloguing did not really matter as long as it only reproduced the arrangement of books. For these theoreticians, writing up a systematic catalogue had a minor significance in the daily work of librarian. The classified arrangement - originating in the common medieval location of a - not so great - number of books - spread out to most German libraries in the course of time. The adaptation of the Göttingen archetype was wide, but seldom perfect. In detail it was not practicable and did not succeed completely.

But already in the 18th century the modern arrangement by groups was alerted to the libraries in southern Germany due to the reorganisation of the Munich State Library and the scripts of Martin Schrettinger (a Benedictine monk before the secularization; then catalogue specialist at the Bavarian State Library).
There was also another trend that contrasted the Göttingen model of a centralistic library for common use, the more and more differentiating learned fields set up their own specific libraries in the middle of the 19th century.
The classification was refined in this time, but cumulated also in the big catalogue systems of the 19th century - a heredity of the time of Enlightenment. Examples may be Berlin, Darmstadt and Halle/Saale. As a rule, the subject catalogue of a scientific library in the 19th and 20th century was a systematic register of the location of the books. A great difference is obvious in this catalogue between social and natural sciences. In social science, a detailed and specialized, nearly canonical grouping is used. In contrast only a rough classification in small number existed in natural science, technology and other upcoming fields.
The dichotomy social versus natural science first appeared in the library classifications of the 19th century. Until the end of the 18th century this separation was not usual.

In this context the meaning of the alphabetic order is of special interest, but not really observed by historians of classification. Often the alphabetic order was integrated in the systematic registers as an isolated application. These solutions were often found in subjects or parts of subjects where a systematic was not theoretically approved in detail. Much more important is the fact that in accordance to the upcoming natural science, technology was nearly not mentioned in the library systematic which was dominated by the social sciences and the humanities. An early exception represents the classification of the Princeton University Library (1901):

0. General
1. Historical sciences
2. Language and literature
3. Modern languages and literatures
4. Arts
5. Theology
6. Philosophy and education
7. Sociology
8. Natural Sciences
9. Technology

8 Subject cataloguing in the 20th century

After the First World War the time of traditional subject cataloguing with references to the location seemed to be over because no library was capable of regrouping their books on shelves to meet all requirements without changing the signatures. Information management in a library was mainly considered a practical task and not alone a speculative issue of the theory of science.
In an international context the Dewey Decimal Classification (DDC) that was created in the end of the 19th century, came up. It dissected science into 10

parts in a very formal way. With these premises Hans Wilhelm Eppelsheimer (†1972) and his colleagues at the city library of Mainz created a location free subject catalogue. The hierarchical systematic of science was replaced by the coexistence of the subjects. This was done by clearance through standardisation: Repeating parts of literature were put in a chronology. Eppelsheimer's invention was introduced in the state library of Darmstadt (1932), whereas the "analytical subject catalogue" published 1931 by Hans Trebst from Dresden had found no practical application.

In present time there are five main streams:

1. Systematic shelving of large numbers of books - similarly to the situation one and a half centuries ago - requires revised methods of ordering for open access to book collections.
2. International respectively transnational use of classifications such as DDC, UDC (Universal Decimal Classification) and the Regensburg Classification (Lorenz (1997b), Lorenz (2003a)) (at the same time: shelf classification) demands a continuing discussion of developments and prospects for the major schemes (including the methodology of faceted classification).
3. Relationship between theory and application demands an exchange between semantic structures / terminology and the technical development.
4. Study of classification theories and systems could stimulate and enlighten discussion in a period of turbulent changes in the world of learning and in the organisation of knowledge.
5. Classification research as a basis of information policy demands new steps to the information technology in general.

Today the - longwinded discussed - projects for a "Unified Classification" (De Grolier (1991)) have been finished, but the real situation of parallel existing classification systems demands a trend to concordances.

Indeed the science of library classification (Losee (1993)) represents an important part of the library and information science and of the philosophy of science - with many needs, plans and visions.

References

DE GROLIER, E. (1991): Some Notes on the Question of a So-Called "Unified Classification". In: R. Fugmann (Ed.): *Tools for Knowledge Organization and the Human Interface, 2.* Ergon, Würzburg, 85–108.

FLEW, A. (1971): *An Introduction to Western Philosophy.* Thames and Hudson, Indianapolis.

KASHYAP, M. M. (2003): Likeness Between Ranganathan's Postulations Approach to Knowledge Classification and Entity Relationship Data Modelling Approach. *Knowledge Organization, 30/1, 1–19.*

LEGIPONTIUS, O. (1747): *Dissertationes philologico-bibliographicae.* translated by Georg Ley. - Gabriel Naud in his Advis pour dresser une bibliotheque (Paris 1627. Reprint Leipzig 1963).

LORENZ, B. (1997a): Humanistische Bildung und fachliches Wissen. Privatbibliothek deutscher Ärzte, Erster Teil. *Philobiblon, 41, 127–152.*

LORENZ, B. (1997b): The Regensburg Classification: A Short Survey. *Cataloging and Classification Quarterly, 25, 39–49.*

LORENZ, B. (1998): Humanistische Bildung und fachliches Wissen. Privatbibliothek deutscher Ärzte, Zweiter Teil. *Philobiblon, 42, 253–300.*

LORENZ, B. (2003a): *Handbuch zur Regensburger Verbundklassifikation.* Harrassowitz, Wiesbaden.

LORENZ, B. (2003b) *Systematische Aufstellung in Vergangenheit und Gegenwart.* Harrassowitz, Wiesbaden.

LOSEE, R. M. Jr. (1993): Seven Fundamental Questions for the Science of Library Classification, *Knowledge Organization, 20/2, 65–70.*

LUDWIG, W. (1997): Téodore de Bèze und Heinrich Rantzau über ihre Bücherliebe. *Philologus, 141, 141–144.*

RICHARDSON, E. C. (1930): *Classification, Theoretical and Practical. Together with an Appendix containing an Essay towards a Bibliographical History of Systems of Classification*, 3rd Edition. Wilson, New York.

WETZEL, F. X. (1877): *Die Wissenschaft und Kunst im Kloster St. Gallen im 9. und 10. Jahrhundert.*

Automatic Validation of Hierarchical Cluster Analysis with Application in Dialectometry

Hans-Joachim Mucha[1] and Edgar Haimerl[2]

[1] Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS), Mohrenstraße 39, 10117 Berlin, Germany
[2] Institut für Romanistik, Universität Salzburg, Akademiestraße 24, 5024 Salzburg, Austria

Abstract. Successful applications of hierarchical cluster analysis in the area of quantitative linguistics were reported in the pioneering works by Goebl (1982, 1984, 1994). Often the dimensionality of linguistic data is high. Therefore multivariate statistical techniques like cluster analysis can to some degree support the researcher. However there is much room left for heuristics. Cluster analysis methods can be generalized by taking weights of observations into account. Using special weights leads to well-known resampling techniques. Here we offer an automatic validation technique for hierarchical cluster analysis that can be considered as a so-called built-in validation of the number of clusters and of each cluster itself, respectively. Furthermore this built-in validation can be used to find the appropriate cluster analysis model. As an illustration of an application in linguistics, the validation of results of hierarchical clustering based on the adjusted *Rand*'s measure is presented.

1 Introduction

Cluster analysis has several synonyms like numerical taxonomy (Goebl (1982)), segmentation or unsupervised learning. It aims at finding interesting partitions or hierarchies without taking any background knowledge into account (Kaufman and Rousseeuw (1990), Mucha (1992), Banfield and Raftery (1993)). Hierarchical clustering is in some sense more general than partitional clustering because a hierarchy (this is usually the result of a hierarchical cluster analysis) is a sequence of nested partitions. Here a partition is treated as an elementary component of a hierarchy. In the following, partitions P(I,K) of the set of I objects (observations) into K non-empty clusters (subsets, groups) C_k are considered, $k = 1, 2, ..., K$. The clusters are assumed pairwise disjoint and a partition is an exhaustive subdivision. In this paper a general way of validation of hierarchies will be recommended.

Some model-based clustering techniques can be expressed in terms of pairwise data clustering (Fraley (1996), Mucha et al. (2002)). Starting from pair-wise distances one can carry out both hierarchical and partitional clustering (Späth (1985)). A generalised form using weighted observations can be given. Otherwise it is well-known that the principle of weighting of observations is a key idea in data mining for handling cores (representatives of dense regions) and outliers (Mucha et al. (2002)). In the case of outliers one has to

downweight them in order to reduce their influence. Special weights are used for resampling purposes in the proposed automatic validation technique that is applied to linguistic data.

2 Pair-wise data clustering

Let $\mathbf{X}$ be the $(I \times J)$-data matrix under investigation consisting of I observations (objects) and J variables. In view of the application in linguistics below using the well-known *Ward* method, let us consider the well-known sum-of-squares criterion

$$V_K = \sum_{k=1}^{K} tr(\mathbf{W}_k), \tag{1}$$

that has to be minimized concerning the partition P(I,K). Herein $\mathbf{W}_k = \sum_{i \in C_k} (\mathbf{x}_i - \overline{\mathbf{x}}_k)(\mathbf{x}_i - \overline{\mathbf{x}}_k)^T$ is the sample cross-product matrix for the k-th cluster C_k, and $\overline{\mathbf{x}}_k$ is the usual maximum likelihood estimate of expectation values in cluster C_k.

Criterion (1) can be written in the following equivalent form without explicit specification of cluster centres $\overline{\mathbf{x}}_k$

$$V_K = \sum_{k=1}^{K} 1/n_k \sum_{i \in C_k} \sum_{l \in C_k, l > i} d_{il}, \tag{2}$$

where n_k is the cardinality of cluster C_k, and

$$d_{il} = d(\mathbf{x}_i, \mathbf{x}_l) = (\mathbf{x}_i - \mathbf{x}_l)^T (\mathbf{x}_i - \mathbf{x}_l)$$

is the pair-wise squared Euclidean distance between two observations i and l. This criterion can be minimized for a single partition P(I,K) by exchanging observations between clusters (Späth (1985)). This is equivalent to *k-means* clustering. Otherwise the hierarchical *Ward* method (Ward (1963)) minimizes (2) in a stepwise manner by agglomerative hierarchical clustering. Mucha et al. (2002) presented other model-based cluster analysis in the pair-wise distances fashion.

Another benefit of clustering based on pairwise distances over clustering that is based directly on the $(I \times J)$-data matrix $\mathbf{X}$ is the more general meaning of distances. For instance, distances allow cluster analysis of mixed data (quantitative and qualitative data, see, for example, Gower (1971)). By doing so exploratory results can be obtained that are at least of practical use.

The expression in (2) can be generalized to

$$V_K = \sum_{k=1}^{K} \frac{1}{M_k} \sum_{i \in C_k} m_i \sum_{l \in C_k, l > i} m_l d_{il}, \tag{3}$$

by using positive weights of observations, where $M_k = \sum_{i \in C_k} m_i$ and m_i denote the weight of cluster C_k and the weight of observation i, respectively.

3 Resampling techniques based on weights of observations

In the following, the weights m_i will be used for resampling purposes. For instance, considering the right hand side of equation (3) one can see obviously the independence of weights m_i and M_k, respectively, from the pair-wise distances d_{il}. The latter one exists in any case and is independent from the weighting the the observations. That means once a distance matrix is figured out it will be unchanged in simulations. One has to change the weights only for simulation purposes. For example, the well-known bootstrap resampling technique can be formulated by choosing the following weights of observations:

$$m_i = \begin{cases} n \text{ if observation } i \text{ is drawn } n \text{ times} \\ 0 \text{ otherwise} \end{cases}$$

Here $I = \sum_i m_i$ holds in the bootstrap-resampling with replacement. Other resampling techniques can be described in a similar fashion by introducing weights. In the following let us focus on effective simulations based on pair-wise distances. Moreover the stability of every hierarchical cluster analysis method based on pair-wise distances can be investigated by assigning the following special weights of observations:

$$m_i = \begin{cases} c \text{ if observation } i \text{ is drawn randomly } (c > 0) \\ 0 \text{ otherwise} \end{cases}$$

This resampling technique is without replacement. Usually c equals 1. The observations with $m_i > 0$ are called active objects whereas the ones with $m_i = 0$ are called supplementary objects. The latter ones do not affect the cluster analysis in any way. However, as an option of our software, they can be allocated after clustering into the partitions and hierarchies according to their distance values. This can be done, for instance, by k nearest neighbour classification.

4 *Rand*'s measure for comparing partitions

Partitions are basic results of cluster analysis that cover also hierarchies. Therefore comparing partitions becomes a basic and general tool for validation of cluster analysis results. The key approach for comparing partitions is based on the comparison of object pairs concerning their class membership (Rand (1971)). For instance, to compare two partitions P(I,K) and Q(I,L), the *Rand* index $R^* = (a+d)/\binom{I}{2}$ (similarity index) can be applied. Here a and d count the pair-wise matches which are good in the sense of similarity (correspondence), see Table 1. Equivalently, *Rand*'s index R^* can be expressed by using a contingency table obtained by crossing directly the two partitions P and Q:

$$R^* = [\binom{I}{2} + 2\sum_{k=1}^{K}\sum_{l=1}^{L}\binom{n_{kl}}{2} - \sum_{k=1}^{K}\binom{n_{k+}}{2} - \sum_{l=1}^{L}\binom{n_{+l}}{2}]/\binom{I}{2}.$$

	Partition Q	
Partition P	Same cluster	Different clusters
Same cluster	a	b
Different clusters	c	d

Table 1. Contingency table of pairs of observations concerning two partitions

		Partition Q						
		1	*2*	...	*l*	...	*L*	Sum
	1	n_{11}	n_{12}	...	n_{1l}	...	n_{1L}	n_{1+}
	2	n_{21}	n_{22}	...	n_{2l}	...	n_{2L}	n_{2+}
	...	...	...	...	...	...	...	...
P	*k*	n_{k1}	n_{k2}	...	n_{kl}	...	n_{kL}	n_{k+}
	...							
	K	n_{K1}	n_{K2}	...	n_{Kl}	...	n_{KL}	n_{K+}
	Sum	n_{+1}	n_{+2}	...	n_{+l}	...	n_{+L}	$I = n_{++}$

Table 2. Contingency table by crossing two partitions P and Q

Table 2 shows such a contingency table with elements n_{kl}. At the right hand side and at the bottom there are the marginal sums n_{k+} and n_{+l}, respectively. The contingency table has the important advantage over Table 1 that the stability of every single cluster can be investigated additionally. Moreover the reliability of each observation can be assessed based on the framework of the investigation of the stability of clusters (see the application below).

The measure R^* is dependent on the number of clusters K. The higher K the higher R^* becomes in average. In order to avoid this disadvantage Hubert and Arabie (1985) recommended the adjusted *Rand* index R based on the assumption of the generalized hypergeometric model:

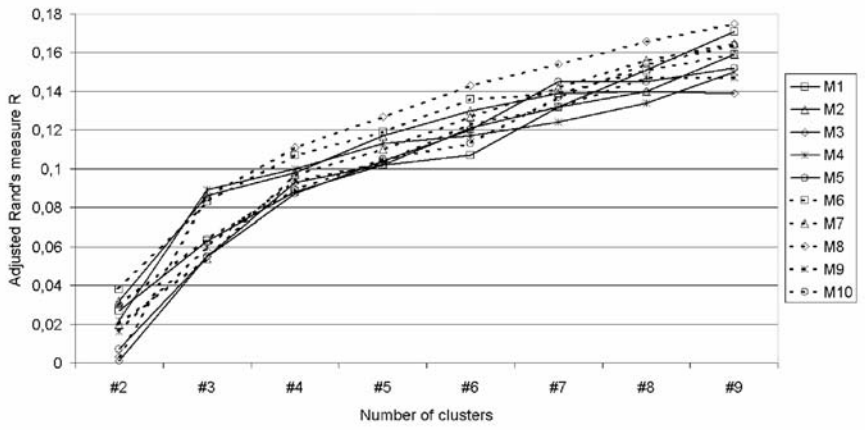


Fig. 1. Medians of adjusted Rand's index R for one standard normal (20 dimensions).

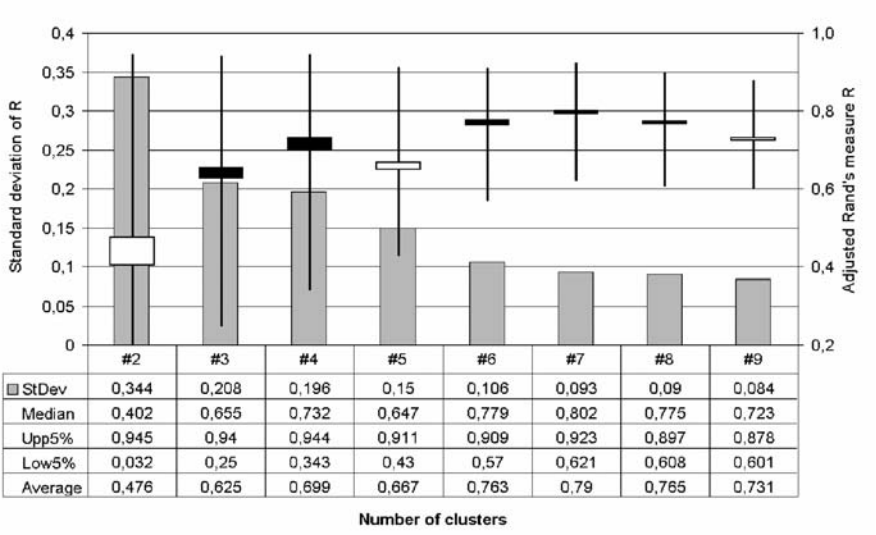
$$R = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L}\binom{n_{kl}}{2} - [\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{l=1}^{L}\binom{n_{+l}}{2}]/\binom{I}{2}}{1/2[\sum_{k=1}^{K}\binom{n_{k+}}{2} + \sum_{l=1}^{L}\binom{n_{+l}}{2}] - [\sum_{k=1}^{K}\binom{n_{k+}}{2}\sum_{l=1}^{L}\binom{n_{+l}}{2}]/\binom{I}{2}}. \quad (4)$$

This measure suits better for the decision about the number of clusters K than R^* because it takes the value 0 when the index R^* equals its expected value for each $k, k = 2, 3, ..., K$. Depending on the options of assignment of supplementary observations after clustering (see the previous section below), the measures R and R^* are figured out based either on all I observations or on a smaller number of observations (= sum over all m_i).

5 A simulation study

Hierarchical clustering gives a unique solution (hierarchy). In this paper the focus is on the investigation of such a unique solution and not on model selection. A unique solution is in opposition to some iterative method like *k-means* clustering that lead to locally optimal solutions depending on initial partitions.

Now let us investigate a data set "without cluster structure". The data is drawn from a multivariate normal distribution with unit covariance matrix. The number of dimensions equals 20, and the number of observations equals 250. Figure 1 shows a set of medians of the adjusted *Rand* index versus the number of clusters. Each median is obtained from 250 adjusted *Rand* values from bootstrap samples (250 replications) of a multivariate sample. Obviously, the index values are located near above zero with an increasing trend to higher values when the number of clusters increases.



	#2	#3	#4	#5	#6	#7	#8	#9
StDev	0,344	0,208	0,196	0,15	0,106	0,093	0,09	0,084
Median	0,402	0,655	0,732	0,647	0,779	0,802	0,775	0,723
Upp5%	0,945	0,94	0,944	0,911	0,909	0,923	0,897	0,878
Low5%	0,032	0,25	0,343	0,43	0,57	0,621	0,608	0,601
Average	0,476	0,625	0,699	0,667	0,763	0,79	0,765	0,731

Fig. 2. Statistics of the adjusted Rand's measure R versus the number of clusters.

6 Application in quantitative linguistics

This is an application in dialectometry, see Haimerl (2004) for more details of this project. The linguistic data consists of 217 regions (observations) with a high number of variables (3899 classified maps with a total number of 20177 taxats, for details see Bauer (2003)). Here the number of clusters, the stability of each cluster and the reliability of the cluster membership of each region are assessed. 250 simulations (and thus 250 cluster analyses) were carried out.

Figure 2 shows both the most important numerical results concerning the adjusted *Rand* index and a corresponding graphical representation of these univariate statistics. Here this measure is computed on the basis of active observations only. The reading of this figure is as follows. The axis at the left hand side and the bars in the graphic are assigned to the standard deviation of R, whereas the axis at the right hand side and the box-plots are assigned to other statistics of R (Median, Average, upper and lower 5 percent quantile). The median of R for $K = 7$ takes the maximum value. That means, the seven cluster solution is the most stable one. It can be confirmed in a high degree for almost all samples. For more than seven clusters the median (or alternatively the average) of the adjusted *Rand*'s values becomes much lower. Therefore the number of cluster $K = 7$ is most likely.

Figure 3 presents the final result of *Ward*'s clustering of 217 regions. The data are based on the linguistic atlas *ALD I* (Goebl (1998)). The polygon map shows the spatial structure of the clustering result into 7 clusters: The area of investigation is obviously divided into 7 sub areas marked with different grey scales and numbers from 1 to 7. Those locations with hatching patterns

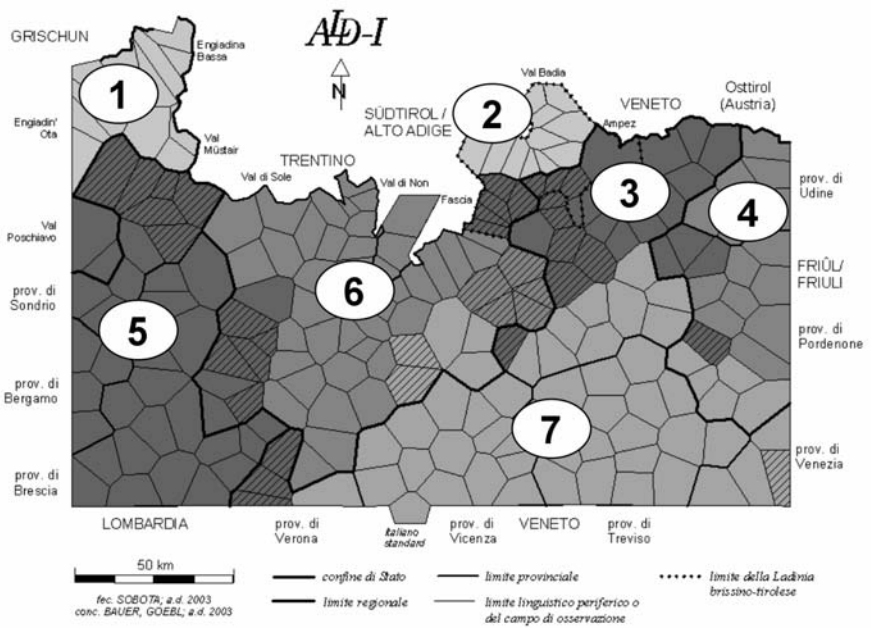


Fig. 3. Linguistic map of cluster membership with information about the reliability of cluster membership of regions. The clusters are marked by numbers.

have been identified as not stable whereas those without hatching are stable with respect to the interval of depth of coverage. The groups with number 1, 2 and 4 are highly stable groups that do not contain any unstable location. This result is what a linguist might expect: The Ladin areas in Grisons (1) and in the northern part of the Dolimitic Ladinia (2) and the area Friuli (4) are clearly separated from the bordering regions. The other four areas have unstable locations at their borders; polygons with hatching at the border between Veneto (3 and 7) and Trentino (6) on one side and between Trentino (6) and Lombardy (5) on the other side. This analysis could be the basis for a deeper linguistic investigation in unstable zones or into significant borders as between Friuli and Veneto that is out of scope of this article.

7 Conclusions

The principle of weighting of observations is a key idea for the built-in validation technique for hierarchical clustering. Using special weights leads to well-known resampling techniques. The proposed automatic validation techniques based on comparison of partitions is especially recommended for investigating the results of hierarchical clustering. Moreover this built-in validation is very easy to apply. Because hierarchical cluster analysis presents "nice" results (dendrograms) independent from the existence of real clusters it is higly

recommended to validate the number of clusters, the stability of each cluster, and the reliability of each observation.

References

BANFIELD, J.D. and RAFTERY, A.E. (1993): Model-Based Gaussian and non-Gaussian Clustering. *Biometrics, 49, 803–821.*

BAUER, R. (2003): Dolomitenladinische Ähnlichkeitsprofile aus dem Gadertal; ein Werkstattbericht zur dialektometrischen Analyse des ALD-I. *Ladinia XXVI-XXVII (2002-2003), 209–250.*

FRALEY, C. (1996): Algorithms for model-based Gaussian Hierarchical Clustering. *Technical Report, 311.* Department of Statistics, University of Washington, Seattle.

GOEBL, H. (1982): *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie.* Verlag der Öst. Akademie der Wissenschaften, Wien.

GOEBL, H. (1984): *Dialektometrische Studien anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, vol. 1 (vol. 2 and 3 contain maps and tables). Max Niemeyer, Tübingen.

GOEBL, H. (1994): Die Dialektale Gliederung Ladiniens aus der Sicht der Ladiner. Eine Pilotstudie zum Problem der geolinguistischen "Mental Maps". *Ladinia XVII, 59–95.*

GOEBL, H. (Ed.) (1998): *Atlante linguistico del ladino dolomitico e dei dialetti limitrofi I (ALD I) - Sprachatlas des Dolomitenladinischen und angrenzender Dialekte I.* Dr. Ludwig Reichert Verlag, Wiesbaden.

GOWER, J.C. (1971): A General Coefficient of Similarity and some of its Properties. *Biometrics, 27, 857–874.*

HAIMERL, E. (2004): Das Dialektometrieprojekt der Universität Salzburg. (in German and English). http://ald.sbg.ac.at/dm

HUBERT, L.J. and ARABIE, P. (1985): Comparing Partitions. *Journal of Classification, 2, 193–218.*

KAUFMAN, L. and ROUSSEEUW, P.J. (1990): *Finding Groups in Data.* Wiley, New York.

MUCHA, H. -J. (1992): *Clusteranalyse mit Mikrocomputern.* Akademie Verlag, Berlin.

MUCHA, H. -J., SIMON, U. and BRÜGGEMANN, R. (2002): Model-based Cluster Analysis Applied to Flow Cytometry Data of Phytoplankton. *Weierstraß-Institute for Applied Analysis and Stochastic, Technical Report No. 5.* http://www.wias-berlin.de/.

RAND, W.M. (1971): Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association, 66, 846–850.*

SPÄTH, H. (1985): *Cluster Dissection and Analysis.* Ellis Horwood, Chichester.

WARD, J.H. (1963): Hierarchical Grouping Methods to Optimise an Objective Function. *JASA, 58, 235–244.*

Discovering the Senses of an Ambiguous Word by Clustering its Local Contexts

Reinhard Rapp

Johannes Gutenberg-Universität Mainz,
Fachbereich Angewandte Sprach- und Kulturwissenschaft,
76711 Germersheim, Germany

Abstract. As has been shown recently, it is possible to automatically discover the senses of an ambiguous word by statistically analyzing its contextual behavior in a large text corpus. However, this kind of research is still at an early stage. The results need to be improved and there is considerable disagreement on methodological issues. For example, although most researchers use clustering approaches for word sense induction, it is not clear what statistical features the clustering should be based on. Whereas so far most researchers cluster global co-occurrence vectors that reflect the overall behavior of a word in a corpus, in this paper we argue that it is more appropriate to use local context vectors. We support our view by comparing both approaches and by discussing their strengths and weaknesses.

1 Introduction

Many problems in statistical natural language processing can be successfully approached by using methods that rely on feature vectors. Since entities in natural language tend to be ambiguous, the feature vectors that we derive from text corpora can be assumed to be mixtures of the vectors of some underlying unambiguous entities. The problem in understanding and simulating natural language is that we can only observe and study the complicated behavior of the ambiguous entities, whereas the presumably simpler behavior of the underlying unambiguous entities remains hidden.

To be more concrete, let us look at word meaning. In this case, the ambiguous entities we consider are words, the unambiguous entities are their senses, and as the relevant features the co-occurring words can be used. Looking at co-occurrences is appropriate as it has been shown that the meanings of a word are well reflected in its lexical neighborhood (Schütze (1997)), that is, the neighbors of a word can be considered to be its features.

Past work on word senses has concentrated on disambiguation, that is, on choosing among a predefined set of senses when given an ambiguous word in context. In contrast, the problem that we consider in this paper is word sense induction, which is the automatic discovery of the possible senses for a given word. Several recent papers, e.g. Pantel and Lin (2002), Neill (2002), Dorow and Widdows (2003), Rapp (2003), and Rapp (2004) give evidence that sense induction now also attracts attention.

Despite many differences, most approaches to sense induction that have been published so far have a common limitation: They rely on global co-occurrence vectors, i.e. on vectors that have been derived from an entire corpus. Since most words are semantically ambiguous, this means that these vectors reflect the sum of the contextual behavior of a word's underlying senses, i.e. they are mixtures of all senses occurring in the corpus.
When starting from global vectors the task of sense induction requires determining the co-occurrence vectors of the senses given the co-occurrence vectors of the ambiguous words. As reconstructing the sense vectors from the mixtures is difficult and often suffers from the sparse data problem, the question is if we really need to base our work on mixtures or if there is some way to directly observe the contextual behavior of the senses thereby avoiding the mixing beforehand. Our suggestion is to look at local instead of global co-occurrence vectors. As can be seen from human performance, in almost all cases the local context of an ambiguous word is sufficient to disambiguate its sense. From this observation we conclude that the local context of a word usually carries no ambiguities. The aim of this paper is to show how this approach, whose application tends to be affected adversely by the sparse-data problem, can be successfully exploited for word sense induction.

2 Approach

Our computations are performed on a term/context-matrix that is based on the concordance of a word. This is an important difference to approaches that use co-occurrence- or term/document-matrices as the vectors in these types of matrices reflect the overall behavior of a word in an entire corpus. That is, they are mixes of all senses of a word, whereas in a term/context-matrix each vector relates to a single sense.
An example of a term/context-matrix is shown in table 1. It relates to the ambiguous word *palm* with its *tree* and *hand* senses. If we assume that our corpus has six occurrences of *palm*, i.e. there are six local contexts, then we can derive six local co-occurrence vectors for *palm*. Considering only strong associations to *palm*, these vectors could, for example, look as shown in the table.
The dots in the matrix indicate if the respective word occurs in a context or not. We use binary vectors since we assume short contexts where words usually occur only once. By looking at the matrix it is easy to see that contexts c1, c3, and c6 seem to relate to the *hand* sense of *palm*, whereas contexts c2, c4, and c5 relate to its *tree* sense. Our intuitions can be resembled by using a method for computing vector similarities, for example the cosine coefficient or the (binary) Jaccard-measure. If we then applied an appropriate clustering algorithm to the context vectors, we should obtain the expected two clusters, and the words closest to the geometric centers of the clusters should be good descriptors of each sense.

Table 1. Term/context matrix for the word *palm*.

	c1	c2	c3	c4	c5	c6
arm	•		•			
beach		•			•	
coconut		•		•	•	
finger	•					
hand	•		•			•
shoulder	•					•
tree		•		•		

However, as matrices of the above type can be quite large and extremely sparse, clustering is a difficult task, and common algorithms often deliver sub-optimal results. Fortunately, the problems of matrix size and sparseness can be minimized by reducing the dimensionality of the matrix. An appropriate algebraic method that has the capability to reduce the dimensionality of a rectangular or square matrix in an optimal way is *singular value decomposition* (SVD). As shown by Landauer and Dumais (1997), Schütze (1997), Rapp (2003), and others, by reducing the dimensionality a generalization effect can be achieved that often improves the results.

As this method is rather sophisticated, we can not go into the details here. A good description can be found in Landauer and Dumais (1997). The essence is that by computing the singular values of a matrix and by truncating the smaller ones, SVD allows to significantly reduce the number of columns, thereby (in a least squares sense) optimally preserving the euclidean distances (and angles) between the rows (Schütze (1997), 191). Alternatively, it is also possible to reduce the number of rows thereby preserving the distances between the columns.

The approach that we suggest in this paper involves reducing the number of columns (contexts) and then applying a clustering algorithm to the rows (words) of the resulting matrix. This works well since it is a strength of SVD to reduce the effects of sampling errors and to close gaps in the data.

3 Algorithm

Our computations are based on a partially lemmatized version of the British National Corpus (BNC) which has the function words removed (Rapp (2002)). With partial lemmatization we mean that only those words in the corpus have been replaced by their root forms that according to a large lexicon of English can be unambiguously assigned to a stem. This makes the corpus more manageable, the computations faster, and reduces the sparse data problem without introducing many errors (other than those resulting from omissions in the lexicon). Our vocabulary consists of all 374240 different word forms occurring in this corpus after lemmatization.

Next we have to specify how we define the context of a word. Since the doc-

uments in the BNC are rather long (average sample size is 24274 words), it seems advisable to choose shorter contexts, for example sentences, paragraphs, or text windows of a fixed size. We decided to use text windows of ±20 words around the given word. Since function words were removed from our corpus, this corresponds to a larger window size of approximately ±40 words in the original corpus if we assume that roughly every second word is a function word.

Based on the list of 12 ambiguous words provided by Yarowsky (1995) which is shown in table 2 we created the concordances for these words, with the lines in the concordances each relating to one context window of ±20 words. From the concordances we computed 12 term/context-matrices whose binary entries indicate if a word occurs in a particular context or not (as exemplified in table 1 for the word *palm*). Assuming that a context word's discriminative power is highly correlated with its association strength to the ambiguous word, in each matrix we removed all words that are not among the top 30 first order associations.

The selection of these first order associations was conducted fully automatically using an algorithm that had been developed to simulate human associative behavior. In no case has there been any manual intervention in selecting these words. As described in a previous paper (Rapp (2002)) the computations in this algorithm are based on the log-likelihood ratio. However, as it was observed that in association experiments conducted with human subjects there was a strong preference towards words that are in the middle of the frequency range (Rapp (1996), 52), before ranking the words we multiplied the log-likelihood values with a triangular shaped function as shown in figure 1 that depends on word frequency. This leads to a considerably better agreement of the computed associations with human intuitions[1]. To give an impression of typical results obtained with this algorithm, the top 30 associations to the words *palm* and *poach* are shown in figures 2 and 3.

Table 2. Ambiguous words and their senses as provided by Yarowsky (1995).

WORDS	SENSES	WORD	SENSES
axes	grid - tools	palm	tree - hand
bass	fish - music	plant	living - factory
crane	bird - machine	poach	steal - boil
drug	medicine - narcotic	sake	benefit - drink
duty	tax - obligation	space	volume - outer
motion	legal - physical	tank	vehicle - container

[1] The agreement between the simulation program and a group of test persons is actually better than the average agreement among the humans: For 31 out of 100 stimulus words the predicted response is equal to the response most frequently given by the subjects. This compares to an average of only 28 such responses given by an average subject. Also, on average 13.5% of the human subjects give the response predicted in the simulation, whereas only 12.6% give the answer chosen by another subject.

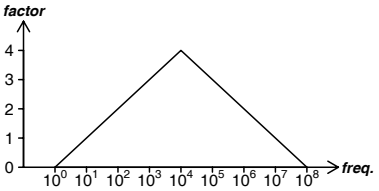


Fig. 1. Triangular function used for computing word associations.

Given that our term/context matrices are very sparse with each of their individual entries seeming somewhat arbitrary, it is necessary to detect the regularities in the patterns. For this purpose we applied the SVD to each of the matrices, thereby reducing their number of columns to the three main dimensions. This number of dimensions may seem low. However, since our matrices of local contexts (derived from the concordance of a word) are much smaller and usually more sparse than the global co-occurrence matrices used elsewhere, it is clear that we had to use fewer than the 100 to 300 dimensions used in other studies. Interestingly, as there are strong dependencies in our data, it turned out that in some cases it was not even possible to compute more than three singular values. Therefore, we decided to use three dimensions for all matrices (with the exception of *space*, where only two singular values could be computed).

The last step in our procedure involves applying a clustering algorithm to the 30 words in each matrix. For our condensed matrices of 3 rows and 30 columns this is a relatively simple task. We decided to use the hierarchical clustering algorithm readily provided in the MATLAB (MATrix LABoratory) programming language. After some testing with the various similarity functions and linkage types available in MATLAB, we finally opted for the cosine coefficient and single linkage which is the combination that apparently gave the best results.

4 Results

Let us exemplify our results by looking at a typical example. Figure 2 shows the dendrogram for *poach* as obtained after applying the algorithm described in the previous section to a dimensionality-reduced term/context matrix. The two main clusters in the dendrogram nicely distinguish between the two senses of *poach*, namely *boil* and *steal*. The left branch of the hierarchical tree consists of words related to cooking, the right one mainly contains words related to the unauthorized killing of wildlife in Africa which apparently is an important topic in the BNC. This example nicely demonstrates what distinguishes the clustering of local contexts from the clustering of global co-occurrence vectors. To see this, let us bring our attention to the various species of animals that are among the top 30 associations to *poach*. Some of them seem

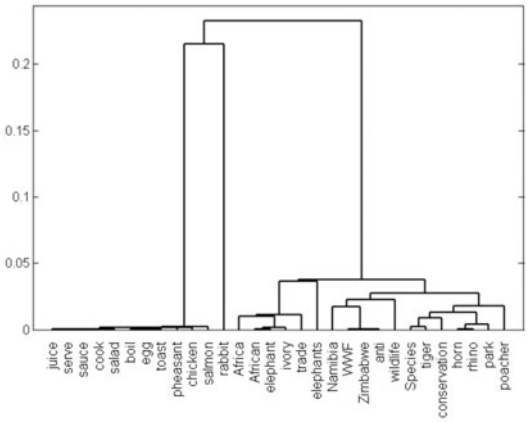


Fig. 2. Clustering results for *poach* with SVD.

more often affected by cooking (pheasant, chicken, salmon), others by poaching (elephant, tiger, rhino). According to the diagram, only the rabbit can be likewise the subject of both activities and in this way confirms its familiar role as a victim to all kinds of actions. However, there is still some hope of survival for the poor rabbit in our corpus-based reality, as luckily its affinity to cooking is lower than it is for the chicken, and to poaching it is lower than it is for the rhino.

The important thing is that by clustering local contexts our algorithm was able to separate the different kinds of animals according to their relationship to *poach*. If we instead clustered global vectors, it would most likely be impossible to obtain this separation. Note that what we exemplified here for animals applies to all linkage decisions made by the algorithm, i.e. all decisions must be seen from the perspective of the underlying ambiguous word. This implies that often the clustering may be counterintuitive from the global perspective that as humans we tend to have when looking at isolated words. In short, the clusters shown in figure 2 can only be understood if the ambiguous words they are derived from are known.

Having shown that the clustering of local contexts is very specific with respect to the given word, let us now show that the clustering of global co-occurrence vectors does not have this desirable property. To illustrate this, figure 3 shows the clustering results for *poach* based on global co-occurrence vectors[2]. The dendrogram looks intuitively plausible as it places related words

[2] These are the details of the computation: Our starting point was a global co-occurrence matrix of 30 rows, with each row corresponding to one of the top 30 first order associations to *poach*, and several thousand columns, each relating to one of the observed context words in the lemmatized BNC that have a corpus frequency of 20 or higher. The window size for counting co-occurrences was ±20 words. We used a simple transformation function that incremented each co-occurrence count by one and then dampened it by computing the logarithm.

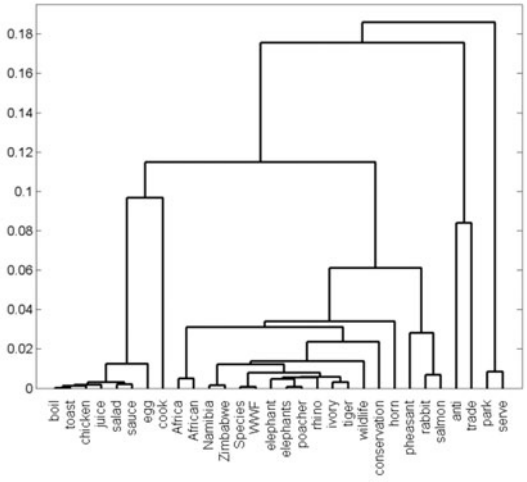


Fig. 3. Clustering results for *poach* using global co-occurrence vectors and SVD.

closely together, e.g. *boil* and *toast*, *salad* and *sauce*, *elephant* and *elephants*, or *Namibia* and *Zimbabwe*. However, the overall appearance of the diagram is different from figure 2, so which one is better?

Our answer is: It depends on the task. If a human subject is asked to cluster the list of 30 words according to their semantic relatedness (not telling that they are all associations to *poach*), then something similar to figure 3 might come out. If, on the other hand, the subject is explicitly asked to cluster the same list of words according to their relationship to *poach* (possibly even specifying the two main senses of *poach*), then this should result in something like figure 2.

In our view it is important to see that these are two distinct tasks. The first task is related to thesaurus construction, the second to sense induction. This distinction was not always made clear in previous papers, which led to some confusion. As there is some correlation between the outcome of the two tasks, mixups can easily go through unnoticed: A program that actually solves task 1 may be presented as a solution to task 2 with reasonable results.

5 Conclusions and prospects

From the observations described above we conclude that avoiding the mixture of senses, i.e. clustering local context vectors instead of global co-occurrence vectors, is a good way to deal with the problem of word sense induction. The clustering of global vectors is inappropriate, as the global co-occurrence

For better comparison with the previous results (based on local contexts), before conducting the hierarchical clustering an SVD-step was performed which reduced the number of columns to three.

vectors are usually dominated by word usage in less relevant contexts (i.e. contexts not containing the ambiguous word). However, there is a pitfall, as the matrices of local vectors are extremely sparse. Fortunately, our simulations suggest that computing the main dimensions of a matrix through SVD solves the problem of sparseness and leads to significant improvements.
Although the results seem useful even for practical purposes, we can not claim that our algorithm is capable of finding all the fine grained distinctions that are listed in manually created dictionaries such as the Longman Dictionary of Contemporary English (LDOCE), or in lexical databases such as WordNet. However, we see many possibilities for further improvements: For example, we can consider a larger corpus, various window types and transformation functions, a wider selection of context words, or we can take a word's part of speech into account.

Acknowledgements

I would like to thank Manfred Wettler and Raz Tamir for interesting discussions, Hinrich Schütze and Mike Berry for the SVD software, and the DFG for financially supporting this work.

References

DOROW, B. and WIDDOWS, D. (2003): Discovering corpus-specific word senses. In: *Proceedings of EACL 2003*, Budapest, conference companion (research notes and demos), 79–82.

LANDAUER, T. K. and DUMAIS, S. T. (1997): A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104(2), 211–240.*

NEILL, D. B. (2002): *Fully Automatic Word Sense Induction by Semantic Clustering.* Cambridge University, Master's Thesis, M.Phil. in Computer Speech.

PANTEL, P. and LIN, D. (2002): Discovering word senses from text. In: *Proceedings of ACM SIGKDD*, Edmonton, 613–619.

RAPP, R. (1996): *Die Berechnung von Assoziationen.* Olms, Hildesheim.

RAPP, R. (2002): The computation of word associations: comparing syntagmatic and paradigmatic approaches. In: *Proceedings of 19th COLING*, Taipei, ROC, Vol. 2, 821–827.

RAPP, R. (2003): Word sense discovery based on sense descriptor dissimilarity. In: *Proceedings of the Ninth Machine Translation Summit*, New Orleans, 315–322.

RAPP, R. (2004): Mining text for word senses using independent component analysis. In: M.W. Berry, U. Dayal, C. Kamath, D. Skillicorn (Eds.): *Proceedings of the SIAM International Conference on Data Mining*, Lake Buena Vista, Florida.

SCHÜTZE, H. (1997): *Ambiguity Resolution in Language Learning: Computational and Cognitive Models.* Stanford: CSLI Publications.

YAROWSKY, D. (1995): Unsupervised word sense disambiguation rivaling supervised methods. In: *Proceedings of the 33rd Meeting of the ACL*, Cambridge, MA, 189–196.

Document Management and the Development of Information Spaces

Ulfert Rist

IABG mbH, VG15
85521 Ottobrunn, Germany

Abstract. Through the use of formal document structures, for example paragraphs and tables, steps are shown on how to use these to extract information in the course of the automatic recognition of the contents of OpenOffice text documents and HTML documents as part of a document management project. It is possible to create formal graphs that structure the document-related information space based on a given information model by using a natural language processing chain and a wrapping procedure. A combined text and layout analysis is carried out with open source components that aims at representing information as a semantic network in a formal and visualizable manner. Scalable ways of retrieving information and processing knowledge are produced by uniting document-related information spaces to form thematic domains.

1 Starting point and task

A given information model serves to support the interoperability of electronic information and simulation systems. The information categories of the data model underlying the information model (referred to as a core data model, hereinafter CDM) represent the lowest semantic level of a military ontology, in which connection a uniform semantic description is given of the user data. Due to the general design of the CDM the seamless integration of themes of other domains is possible. The goal of text-related information extraction consists in the projection of text information to constituents of the information model. The generated information constructs are used to fill the CDM database support. In this connection, the integration of efficient individual technologies that require minimum tailoring comes to the fore.

2 Implementation

From the layout perspective, there is initially a concentration on flow texts and tables. The DOM (document object model) of an OpenOffice text document[1] is accessible via the OpenOffice API. A Python class can access DOM elements, such as the paragraph (flow text), headline, list etc. within an

[1] In the OpenOffice open source environment, a MS Word document can be read without any difficulty. URL: http://de.openoffice.org/.

OpenOffice application. At the same time, partially structured HTML documents are migrated. From the perspective of the software architecture regarding OpenOffice, the relevant parts of a document undergo cross-computer processing with a Python XML server[2] which sends an XML copy as an inquiry to applications on other computers and, after processing, receives and allocates the result.

3 Representation of the information space

Prime words (abbreviated to *primes*), e.g. for artifacts, that are linked to one another through relations are derived from the CDM in the information model. The semantics of primes is clearly defined in the CDM. The relations also have a semantic dimension that is usually realized as a verbal context pattern. As a directed graph, the prime-relation-prime structure grammatically organizes a subject-verb context-object structure. Standardization takes place on the uppermost level based on the following rule: Subject and object positions are realized through basic forms on a syntactic level. Verbal constructs are in the third person singular syntactically (e.g. *is-used-to-describe*, *is-made-up-through*). Examples of further standardization rules are deleting leading determiners in noun phrases, separating prepositions from prepositional phrases and adding the same to the verbal expression.

Graphs can be constructed to form complex graph structures that make up a semantic network on a topological level. In its entirety, the information model forms an information space that is structured by all the primes and contexts involved the constituents. From the perspective of information retrieval, the CDM database support underlying the information model provides a basis for filing and for searching for extracted text information. The information space of an individual document is represented by the extracted graphs. A thematic domain is represented as a cross-document information space, i.e. as a union of the information spaces involved.

4 Processing flow text

As natural language expressions occur fairly rarely as simple linguistic sentences (see below), several processing steps are usually required to transfer a sentence from a German real world text into a number of derived graphs. For this purpose, a natural language processing tool chain has been set up. English graphs are produced as a main result. The tool chain consists of several autonomous modules steered by a Python integration layer.

The sentence-boundary detector serves sentence boundary disambiguation of flow texts that are processed to form lists of sentences. The provision of fragments of flow text at sentence level is essential for sentence-related

[2] URL: http://www.xmlblaster.org.

parsing later. An appropriate rule-based and lexicon-based Python class was realized within only few man days. In the course of the further development of modules, the following approaches are used as possible ways of optimising modules: Frey (2002) describes a non-rule-based method of detecting sentence boundaries with high hit ratios using a neural network. A language-independent, statistically-based connection between abbreviation recognition and a subsequent least effort sentence boundary detection system is presented by Kiss and Strunk (2003).

Apart from colloquial abbreviations, technical expressions must be put into an expanded form. At present, a list of military abbreviations which has around 10,000 entries in XML format provides the required data basis for this. In addition, a decision must be made in one correction step for inflecting parts of speech as to the syntactical form in which a lexeme appears in a phrase, i.e. which case and number are realized in an individual case. For example, noun phrases are thus to be formed correctly, e.g. *Der Kdr PzBtl 592* to *Der Kommandeur des Panzerbataillons 592* (The commander of the tank battalion 592).

The part of speech tagger (POS tagger) is used to determine parts of speech based on the Stuttgart-Tübingen tagset (STTS) that has a scope of 54 tags (Schiller (1995)). In addition, basic forms are to be detected to standardize graphs. For this purpose, extensive, tagged material is presented to a teachable tagger in a preliminary processing step. This type of tagger can be implemented without much effort as an instance of a Python class in the NLTK module (natural language toolkit, Loper and Bird (2002)). Technical lexicons and a module for the provision of basic forms are currently being set up. A syntactical parser generates the analysis of a natural language expression. With the aid of a syntactic analysis, syntactical units can be detected that assume certain grammatical functions in a sentence, such as subjects and objects, that are usually realized as noun phrases (NP). The in-house development of a parser for complete sentence analysis has proved to be an impractical undertaking in the present project framework due to the considerable effort required. The results on variabilities in tag patterns gained from an informal text type study clearly show how difficult this task is. In the study on text types, ten flow texts of different genres in ASCII format were analysed which each had a scope of about 100,000 kilobytes. A high degree of variability of sentence patterns was shown at tag pattern level (see Table 2). An average sentence is thus composed of approximately 21 to 22 tagged parts (including punctuation). About 99 of 100 sentences are uniquely structured at a tag pattern level (UTP portion in text).[3]

Only 0.512 percent of all sentences were discovered on average using a sentence grammar, which required a few man days to develop. In order to achieve

[3] The noticeably long sentences in fairy tales (FT) are due to frequently embedded rhymes. The detection of the start and end of these would have required special rules, which was dispensed with in this case.

Table 1. Texts of the study on text types

Text type	Abbreviation
Fairy tales	FT
Bible	BB
Bild Zeitung (yellow press, nationwide)	BZ
Abendzeitung (tabloid paper, local)	AZ
Süddeutsche Zeitung (daily newspaper, nationwide)	SZ
Spiegel (weekly magazine)	SP
IABG study report	IB
PhD thesis (psychology)	PT
Online consumer texts by financial service providers	FI
German Federal Armed Forces (online)	GA

Table 2. Text type-related tag pattern

Text type	Sentences	Unique tag patterns (UTP)	Tags per sentence	UTP portion in text
FT	682	665	30.41	97.5%
BB	879	862	23.20	98.1%
AZ	779	774	20.64	99.4%
BZ	1241	1205	14.58	97.1%
SZ	790	788	20.09	99.7%
SP	867	866	18.83	99.9%
IB	545	532	26.03	97.6%
PT	604	598	24.95	99.0%
FI	852	842	17.24	99.8%
GA	779	776	18.83	99,6%
Sum	8018	7908	21.48	98.6%

a better text structure detection rate with the underlying variability in the shortest possible time, the method of partial parsing was used as an alternative. The method of partial parsing (also called *light parsing* or *chunking*), which requires less effort compared to complete parsing, can be combined with other methods, if required, such as topological field analysis (Müller and Ule (2001)). Chunking was designed to recognise noun phrases. In turn, this was restricted to non-recurrent base noun phrases, whereby sequences were permitted as a recurrent expansion. The wording of the grammatical rules was carried out using the Python module NLTK (see above) at tag pattern level.

In figures, it was possible to detect the syntax of about half of the text structures (see Table 3) using only a few man days. The next task is to derive the graphs from the detected syntactical structures that match the information model. As it is not possible to automatically decide how lex-

Table 3. Text type-related NP analysis

Text type	NPs per sentence	Tags per NP	NP portion in text
FT	6.59	1.71	37.01%
BB	5.68	1.77	43.35%
AZ	5.01	2.09	51.46%
BZ	3.48	2.01	48.68%
SZ	4.79	2.12	50.60%
SP	4.40	2.09	48.73%
IB	5.91	2.46	55.78%
PT	5.72	2.11	48.75%
FI	4.25	2.01	51.51%
GA	4.60	2.24	54.50%
Sum	5.04	2.06	49.04%

emes are to be allocated to the valence framework of sentences without any knowledge of the morpho-syntactical information of these lexemes, automatic processing ends here. The XML output of the flow text in NP chunks is thus analysed by the editor to identify the grammatical function of noun phrases (e.g. subject of the noun phrase). Therefore, the XML document can be further developed either as an excerpt from a raw text or, after SGML tailoring, with the open source tool Alembic Workbench[4] (Day et al. (1997)). A graph pattern can be set up as a tagging pattern through the Alembic-supported functionality of the definition of relations. A text tagged with Alembic relations is subsequently read out by a Python class. The relations extract in graph format is passed on to the translation module. Literal word-for-word translations are carried out using multi-lingual catalogues of technical terminology, e.g. Lexis (Federal Office of Language). Standardised German graphs are translated into English graphs semi-automatically through the provision of several possible translations. In addition, synonym wrappings can be derived for expressions with WordNet and an appropriate Python bridge, as well as hierarchical abstractions, which increases the possibility of matching the primes of the information model. Initial experience gained shows the use of extensive lexicons, special alignment methods (see Manning and Schütze (2000)) and systematic paraphrasing (Barzilay (2003)).

5 Processing partially structured documents

For tables, reading out each table field of the lines contained in the table is obvious, whereby semantic context information for table fields in the body of the table can be found in the fields at the head of the table. Apart from automatically generated tables with a systematic distribution of white space,

[4] URL: http://www.mitre.org/tech/alembic-workbench/.

there are also numerous irregularities in manually generated tables in flow text format. HTML tables may also contain nested layout structures where table lines are used as list elements, as the following example on the equipment of Virginia class cruisers[5] shows (edges of table marked). Wrapping is

Table 4. Detection systems of Virginia class cruisers

DETECTION SYSTEMS		
Radar System:		
	Air Search:	ITT SPS 48C or 48D/E, 3D
		Lockheed SPS 40B or Raytheon SPS 49(V) 5
	Surface:	SC Cardion SPS 55
	Navigation:	Raytheon SPS 64(V) 9
	Fire Control:	2 SPG 5 1D
		SPG 60 D
		SPQ 9A
Sonar System:		EDO/GE SOS 53A (Bow Mounted)

ideal as the technology for creating partially structured HTML documents. The LAPIS wrapper employed in the project (Miller (2002)) uses TCL scripts (tool command language) and a text constraint language developed specially for this purpose in a Java environment. The interesting thing about LAPIS is the graphic user interface which has WYSIWYG functionality and the ability to work on texts with TCL with the aid of text constraint patterns. Aided by a command line-oriented API, LAPIS can be used as a pattern matching machine even without a GUI (graphical user interface). The extracted information material is semantically tagged with XML. The complex data entries of an XML fragment such as the following have to be broken down into several consecutive steps in order to be used in the CDM database support.

```
<data_section theme="DETECTION SYSTEMS">
    <sub_section theme="Radar System">
        <sub_section_data>
            <item_key>Air Search<item_key>
            <item_value>ITT SPS 48C or 48D/E, 3D</item_value>
            <item_value>Lockheed SPS 40B or
                        Raytheon SPS 49(V) 5</item_value>
        </sub_section_data>
    </sub_section>
    [...]
</data_section>
```

[5] URL: http://www.nasog.net/datasheets/warships/cruisers/Virginia_Class.htm

The anonymous relation *stands-in-relation-to* can be transferred to talking contexts such as *is-a-kind-of* through the tabular generic relation between the field in the head of the table and the field in the body of the table. Details of the directory tree can also be used: The Virginia class association between cruiser and warship ensues from the names of the directory tree where the HTML file is located. Key-value pairs can be expanded to graph format. Presentation as an IDEF1X diagram offers itself for the visualisation of graphs in the information space.

Table 5. Graphs derived from an HTML table

Subject	Context	Object
Virginia class	is-classified-as	cruiser
cruiser	is-a-kind-of	warship
Virginia class	has-as-a-component	detection system
radar system	is-a-kind-of	detection system
air search	requires	radar system
ITT SPS 48C *or* 48D/E, 3D	is-classified-as	radar system
Lockheed SPS 40B *or* Raytheon SPS 49(V) 5	is-classified-as	radar system
air search	requires	ITT SPS 48C *or* 48D/E, 3D
air search	requires	Lockheed SPS 40B *or* Raytheon SPS 49(V) 5

6 Summary and outlook

The approach adopted for information extraction is to be viewed as a demonstration showing the directions that are conceivable for the development of estimates on costs and results related to a particular project, apart from furnishing proof of the feasibility of the approach. If one understands information extraction to be a mechanism that generates meaningful insertions for variable positions in patterns (Manning and Schütze (2000), 376), it is possible to favourably assess the conceptual approach adopted in the project. From the perspective of project management, the major obstacle of linguistic processing comes to the fore for operative use which, however, can be overcome with sufficient resources. One only needs to consider the development and care of electronic dictionaries, grammars on sentence analysis and methods on the dissolution of references in text discourse (e.g. anaphora). In this connection, the use of automatic procedures is recommended, for example, the automatic compilation of lexicons (Zernik (1991)). The use of open source

components must be consistently increased. The functionality of the LAPIS wrapper is up for expansion through the integration of additional parsers, e.g. a parser for NP fine analysis which can be embedded in the pattern tree as an additional text constraint pattern.

References

DAY, D. et al. (1997): Mixed-Initiative Development of Language Processing Systems. In: *Fifth Conference on Applied Natural Language Processing*. Association for Computational Linguistics, Washington D.C.
URL: http://www.mitre.org/tech/alembic-workbench/ANLP97-bigger.html.

FREY, M. (2002): The Role of Data Representation in Sentence Boundary Disambiguation with Neural Networks. *FKIE-Bericht Nr. 46*, Forschungsgesellschaft für Angewandte Naturwissenschaften e. V. (FGAN), Wachtberg.

KISS, T. and STRUNK, J. (2003): Viewing sentence boundary detection as collocation identification. In: S. Busemann, S. (Ed.): *Konvens 2002 Tagungsband*. DFKI, Saarbrücken, 75–82. URL: http://www.linguistics.ruhr-uni-bochum.de/∼kiss/publications/07v-kiss.pdf.

LOPER, E. and BIRD, S. (2002): NLTK: The Natural Language Toolkit. In: *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Association for Computational Linguistics, Philadelphia.
URL: http://arxiv.org/PS_cache/cs/pdf/0205/0205028.pdf.

MANNING, C. D. and SCHÜTZE, H. (2000): *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts and London, England.

MÜLLER, F. H. and ULE, T. (2001): Satzklammer annotieren und Tags korrigieren – Ein mehrstufiges "Top-Down-Bottom-Up"-System zur flachen, robusten Annotierung von Sätzen im Deutschen. In: H. Lobin (Ed.): *Proceedings der GLDV-Frühjahrstagung 2001*. Universität Gießen, 235–244.
URL: http://www.uni-giessen.de/germanistik/asd/gldv2001/proceedings/pdf/GLDV2001-mueller.pdf.

MILLER, R. C. (2002): *Lightweight Structure in Text*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.
URL: http://www-2.cs.cmu.edu/∼rcm/papers/thesis/thesis.pdf.

SCHILLER, A. et al. (1995): *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Universität Stuttgart and Universität Tübingen.
URL: http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts-guide.ps.gz.

ZERNIK, U. (Ed.) (1991): *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Erlbaum, Hillsdale, New Jersey.

Stochastic Ranking and the Volatility "Croissant": A Sensitivity Analysis of Economic Rankings

Helmut Berrer, Christian Helmenstein, and Wolfgang Polasek

(berrer, helmen, polasek)@ihs.ac.at
Institute for Advanced Studies (IHS),
1060 Wien, Austria

Abstract. Rankings of countries are calculated using I indicator variables. Clearly, any ranking based on an index depends on the weights used, and therefore we conduct a sensitivity analysis on the weights of the index to obtain a measure for the volatility of the performance rankings. The weights are simulated from uniform and beta distributions on the simplex. As a result we observe a volatility "croissant": Countries in the top and the bottom of the ranking are less volatile than in the middle of the ranking. The methodology is shown for the standardized performance ranking (SPR) and the rank performance ranking (RPR).

1 Introduction

For companies there exists a single performance or success variable e.g. profit or EBIT (earnings before interest & tax). National economies are too complex to be rated by a single variable. Furthermore, there exists a huge amount of potential variables that measure performance. One of the most common methods to judge performance is the index method, where several factors are condensed into one single number describing the overall economic performance.

Index construction has become very popular over the last decades: This is due to two developments. The liberalisation process in many areas of the modern economies increases the competitions between regions or other economic units and people want to know where they stand. Also, modern technologies allow the collection of a broad spectrum of data and to make them available on the internet.

To motivate our approach we will review the index approaches of the most popular international studies: In the year 2001 the "Department of Trade and Industry" in England was interested in measuring the regional performance by using a wide range of indicators. This so-called DTI (2001) index includes among other variables the establishment of an enterprise and survival rate, occupation, creation of value per person employed according to industries, average income, gross domestic product per head and expenditures for education and infrastructure. The Milken Institute established several economic indices to measure the performance of the US regions, e.g. the Capital Access

Index, the Best Performing City Index, Best Places Index, State Technology and Science Index and the Knowledge-Based Economy Index.

The Progressive Policy Institute (PPI) constructed "The State New Economy Index" using the following weighting methodology: Raw scores were calculated for each state for each of the 21 indicators. In the composite analysis, the indicators were weighted so that closely correlated ones did not influence the results. In addition, to measure the magnitude of differences between states and not just their ranks, in each indicator, scores were based on the standard deviation from the mean score of all of the states. The 21 indicators are summed up into five sub-indices, using different weights.

The World Economic Forum publishes the Global Competitiveness Index (GCI) with the fundamental objective of evaluating the economic competitiveness of a large sample of countries. The GCI uses both hard (publicly available) data and data from the World Economic Forum's Survey (in total 11 different indicators with different weights) to estimate three "component indices" that capture the three pillars of growth: "Technology Index", "Public Institutions Index", and "Macroeconomic Environment Index". The three components are then combined to calculate the overall GCI. The IMD (International Institute for Management Development) started in 1989 to publish the World Competitiveness Yearbook comparing 60 countries on more than 300 criteria.

These examples show that there are various approaches to measure different performance behaviour by using indices for many fields and especially for national economies. Therefore the main goal of this paper is not to find the ultimate "best economic index", but rather to establish a framework to test the stability of already existing indices. For demonstration reasons we construct two different indices/rankings using 21 variables to describe the performance of 28 European countries.

2 Index definition and ranking

Most index calculations are based on a 2-stage approach, after the set of variables is defined $\{V_1, ..., V_I\}$. To create a ranking (or index) of different indicators we first have to make the variables "commensurable" (by data pre-processing), i.e. comparable for aggregation. In the second stage this transformed variables are aggregated by a set of weights. To guarantee the robustness of the final results we suggest two different data pre-processings for the index generation.

Standardized performance index and ranking

First we standardized the absolute values for each variable. Based on this data, the index was calculated for each country (standardized performance index' - SPI'). The index is therefore a weighted sum of I variables z_i with

equal mean and variance for each of the C countries, providing that the different scales and data ranges of the original variables V_i do not influence the result. Define the z-score of a country for variable i as:

$$z_i(c) := \frac{V_i(c) - Mean(V_i)}{stdev(V_i)}, \qquad c = 1, \ldots, C, \quad i = 1, \ldots, I. \tag{1}$$

$$SPI'(c) := \sum_{i=1}^{I} w_i z_i(c), \qquad c = 1, \ldots, C. \tag{2}$$

Where $(w_1, ..., w_I)$ are positive weights summing up to 1.

The SPI' does not have the form of an index, i.e. the values can be positive or negative[1].

The standardized performance ranking (SPR) is obtained by ordering the index values of the first stage:

$$SPR(c) := 1 + \sum_{l=1}^{C} 1_{[SPI'(c)<SPI'(l)]}, \qquad c = 1, \ldots, C. \tag{3}$$

Rank performance index and ranking

For the second rank-based index construction we need a ranking for each variable V_i, so that we get an (increasing) ordering of countries for every variable (O_i).

$$O_i(c) := 1 + \sum_{l=1}^{C} 1_{[V_i(c)<V_i(l)]}, \quad i = 1, \ldots, I, \quad c = 1, \ldots, C. \tag{4}$$

The rank performance index' (RPI')[2] is in the second stage a weighted sum of the indicator ranks $O_i(c)$ of country c:

$$RPI'(c) := \sum_{i=1}^{I} w_i O_i(c), \qquad c = 1, \ldots, C. \tag{5}$$

[1] A possible index construction could be:

$$SPI(c) := (1 + SPI'(c)) * 100, \qquad \forall c = 1, \ldots, C.$$

An index value (with equal index weights) of e.g. 106 means that the country possesses on average z-scores which are 6 percent higher in terms of standard deviation (=1 for all z-scores) than the average of the whole sample of countries. But since we are primarily interested in the relative position of a country and the volatility of this rank, this is just a side note.

[2] As in the SPI case the RPI' does not have the conventional form.

Again the rank performance ranking (RPR) is derived by ordering the achieved index values:

$$RPR(c) := 1 + \sum_{l=1}^{C} 1_{[RPI'(c)>RPI'(l)]}, \qquad c = 1, \ldots, C. \tag{6}$$

In a first step all weights of the index are set to a fixed (equal) value, a common approach for indices. The sensitivity of the index is calculated through stochastic simulations.

3 Data

We used data statistics from Eurostat including the 15 EU-member states, the 10 accession countries and the non-EU countries Romania, Bulgaria and Norway.

Table 1. Eurostat indicator variables and codes

Code	Variable	Label
caa13584	Population increase	Population
cba13584	Proportion of population aged 65 and over	100 - Non-Retirement
cca23312	Total population having completed at least upper secondary education	Education
ccb13584	Employment rate total	Employment1
ccb32528	Long-term unemployment	-Employment2
daa11536	Gross value added at basic prices	GVA
dab12048	Gross fixed capital formation (investments)	Investment
dab13072	External balance of goods and services	Balance
dad11024	Current taxes on income, wealth, etc.	-Tax1
dad13072	Taxes on production and imports	-Tax2
dbc11536	Short-term interest rates Three-month interbank rates	-Short Interest
dbc12048	Long-term interest rates	-Long Interest
dbc12560	Share price indices. Rebased	Shares (Growth rate)
eb011	GDP per capita in PPS	GDP capita
eb012	Real GDP growth rate	GDP growth
eb021	Labour productivity per person employed	Productivity1
eb022	Labour productivity per hour worked	Productivity2
eb040	Inflation rate	-Stability
eb060	Public balance	Budget
eb070	General government debt	Debt
eca10000	Research and development expenditure by sector	R&D

The 21 indicator variables[3] in Table 1 contain information about the population and its education, productivity, financial situation and the national economy.

It follows from the index definition that all variables have to be aligned in a "concordant" way, that a higher value denotes a better performance for this indicator. Therefore we transformed the variable X= "Proportion of population aged 65 and over" into the variable Non-Retirement (by the complement 100-X). The indicators "Long-term unemployment", "Current taxes on income, wealth, etc.", "Taxes on production and imports", "Short-term interest rates 3-month interbank rates", "Long-term interest rates", "Inflation rate" and "General government debt" are pre-processed by a multiplication of -1. Finally the variable "Share price indices; Rebased" was transformed into "Share price indices; Growth" using the annual growth rate.

Some indicator statistics after this first data pre-processing are listed in Table 2, indicating the need for a standardization of the indicator values before calculating an index in general.

Table 2. Indicator statistics 2002

Label	mean	stdev	min	max
Non-Retirement	85.2	1.9	81.8	88.8
Employment1	60.5	13.6	50.7	75.9
GVA	3.2	2.1	0.6	8.3
Investment	21.5	4.7	4.4	28.8
Balance	0.1	7.7	-11.2	17.4
Short Interest	-7.0	7.3	-41.3	-3.3
GDP capita	82.9	39.5	24.4	194.3
GDP growth	2.8	2.2	-1.2	7.9
Stability	-4.9	6.2	-34.5	-1.2
Budget	-0.6	4.5	-7.3	15.0
Debt	-49.4	27.4	-109.5	-4.8

4 Sensitivity analysis by randomised weights

Starting from the initial (equal) weight setting we construct randomly derived weights under the following conditions:

- The weight vector has to be uniformly distributed in the space of possible weight vectors.

[3] In the case of missing values for a country the average of the last 3 years was used. In the case of missing variables the average of the EU-15 or the acceding countries was used.

- Each component of the random weight vector has the same expected value and variance.
- The weights sum up to 1.

The Dirichlet distribution with identical parameters $\alpha_j = 1$ ($\forall j = 1, \ldots, n$) is a convenient specification to obtain uniformly distributed stochastic weights:

$$\theta \sim Dirichlet(1, \ldots, 1). \tag{7}$$

The marginal distributions of each component are Beta-distributed (see Gelman et al. (1995)), i.e.

$$\theta_j \sim Beta(1, n-1), \qquad j = 1, \ldots, n. \tag{8}$$

$$E(\theta_j) = \frac{1}{n}, \qquad \forall j = 1, \ldots, n. \tag{9}$$

$$var(\theta_j) = \frac{n-1}{n^2(n+1)}, \qquad j = 1, \ldots, n. \tag{10}$$

To generate samples of the Dirichlet distribution we use the following approach. First we draw x_j from a gamma distribution with shape parameter $\alpha_j = 1$:

$$x_j \sim \Gamma(1, 1), \qquad j = 1, \ldots, n. \tag{11}$$

and then we calculate θ_j random weigths of the n gamma variables:

$$\theta_j = \frac{x_j}{\sum_{j=1}^{n} x_j} \qquad j = 1, \ldots, n. \tag{12}$$

The θ_j are the randomised weights on the simplex for the sensitivity analysis.

We produced 10.000 different random weight vectors and calculated the mean and the standard deviation of each component. The averages range between 0.047061 and 0.04859 and the standard deviations between 0.044166 and 0.046781. This obviously only marginally differ from the theoretical values specified in Eq. 9 $Average = 1/21 \approx 0.04761905$ and Eq. 10 $Stdev = \sqrt{20/(21^2 * 22)} \approx 0.04540298$, ensuring that each indicator is treated equally.

5 Ranking results

The advantage of the stochastic ranking procedure is that by simulating weight we can calculate next to the mean of the simulated rank also an uncertainty measure of the ranking, e.g. by the standard deviation. The standard deviation depends on the "volatility" of the z-scores (or the indicator

rankings) for each country, and on the number of countries in the current "ranking (index) neighbourhood" of the respective country.

In Figures 1 and 2 we can see the (μ, σ) plot of the average ranks versus the standard deviation of this ranks is looking like a croissant. In the top and the bottom the volatility of the stochastic ranking procedure is small while the volatility is highest in the middle. This implies that changes in similar future ranking of countries which lie in the middle of the index range are more likely than on both end.

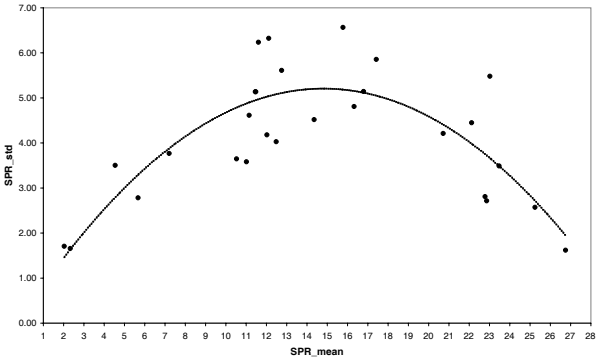


Fig. 1. Stochastic ranking for 2002: By data standardization (SPR).

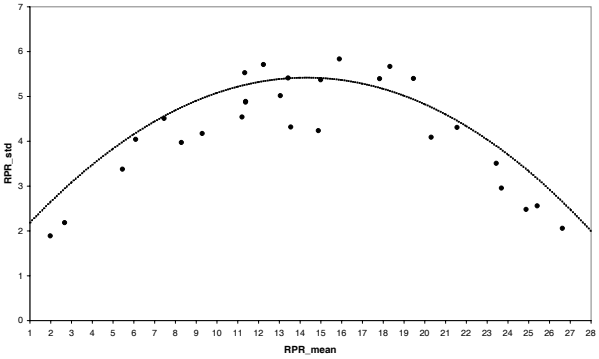


Fig. 2. Stochastic ranking for 2002: By indicator ordering (RPR).

To get a final ranking of the countries we average over the simulated stochastic weights. The rankings produced in the SPR and RPR case are very similar which is expressed by very high correlation coefficients between the two methods (see Table 3). The Pearson correlation between the average ranking of the RPR and the SPR, - but also the Spearman correlation between

the final ranking of the RPR and SPR, - always stays above 0.94 (for the period 2000 to 2002).

Table 3. Correlation between average and final rank for SPR and RPR

	Average rank	Final rank
2000	0.970	0.958
2001	0.966	0.950
2002	0.961	0.945

6 Conclusions

The ranking of countries according to indices is very popular but only a few studies also report the uncertainty associated with the ranking methods. In this paper we have explored two types of indices based on z-scores of the indicators, the linear SPR and the rank-based RPR. We have evaluated the standard deviations of these rankings based on the randomisation of the index weights (generated by a Dirichlet distribution). We find that these two stochastic ranking methods have similar average and final rankings, but the uncertainty in terms of standard deviation is larger for the RPR. Plotting these results in a (μ, σ)-diagramm shows the phenomenon of a volatility croissant. Countries that are ranked best or worst have small standard deviations while countries in the middle have large standard deviations. These results of the sensitivity analysis are quite robust with respect to the index method or the type of stochastic weight simulations.

References

DTI, DEPARTMENT OF TRADE AND INDUSTRY (2001): Regional Competitiveness Indicators. DTI, London. http://www.dti.gov.uk/

EUROSTAT http://europa.eu.int/comm/eurostat/

GELMAN, A., CARLIN, J.B., STERN, H.S. and RUBIN, D.B. (1995):*Bayesian Data Analysis*. Chapman & Hall, London.

IMD, INTERNATIONAL INSTITUTE FOR MANAGEMENT DEVELOPMENT: World Competitiveness Yearbook, Lausanne. http://www02.imd.ch/wcy/

MILKEN INSTITUTE, Santa Monica, CA. http://www.milkeninstitute.org/research/research.taf?cat=indexes

PPI, PROGRESSIVE POLICY INSTITUTE, Washington, D.C.: The Metropolitan New Economy Index. http://www.neweconomyindex.org/states/index.html

WEF, WORLD ECONOMIC FORUM (Ed.) (1999): The Global Competitiveness Report. World Economic Forum, Oxford University Press, New York. http://www.weforum.org/

Importance Assessment of Correlated Predictors in Business Cycles Classification

Daniel Enache and Claus Weihs

University of Dortmund*
Department of Statistics
44221 Dortmund, Germany

Abstract. When trying to interpret estimated parameters the researcher is interested in the (relative) importance of the individual predictors. However, if the predictors are highly correlated, the interpretation of coefficients, e.g. as economic "multipliers", is not applicable in standard regression or classification models. The goal of this paper is to develop a procedure to obtain such measures of importance for classification methods and to apply them to models for the classification of german business cycle phases.

1 Problem

1.1 Introduction

Multivariate classification of the four business cycle phases upswing, upper turning point, downswing, and lower turning point is often performed by linear discriminant analysis (LDA, cf. Meyer and Weinberg (1975)) and by time series analysis methods (e.g. Krolzig (1997)). Lately, other classification methods, like quadratic discriminant analysis, classification trees, artificial neural networks and support vector machines, have also been applied to this problem (e.g. Garczarek and Weihs (2002)) and new classification methods have been developed to solve this problem (e.g. Röhl et al. (2002)).

Heilemann and Münch (1996) reduced the stylized facts to a set of 13 important variables (see also Theis et al. (1999), Weihs and Garczarek (2002)):

* This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

Y yearly growth rate of the gross national product (GNP)
C yearly growth rate of private consumption
GD government deficit as a proportion of the GNP
L yearly growth rate of the number of wage and salary earners
X netto exports as a proportion of the GNP
M1 yearly growth rate of money supply
IE yearly growth rate of equipment investments
IC yearly growth rate of construction investments
LC yearly growth rate of labour unit costs
PY yearly growth rate of GNP price deflator
PC yearly growth rate of the consumer price index
RS nominal short term interest rate
RL real long term interest rate

In analyzing business cycles it is important not only to obtain good predictions, but also to measure the influence of the individual variables to get an impression of their importance. In linear models, the measures of influence are usually the regression weights. Under ceteris-paribus assumptions, these weights measure how much the dependent variable changes if the independent variable is varied by a certain amount. In economic contexts these regression coefficients are called "multipliers".

Similar to these regression models, the coefficients of linear classification equations, like LDA, can be interpreted as influences on the probability of being classified into the selected class. For the interpretation of the coefficients in linear regression or classification models it is crucial, that the independent variables are uncorrelated. Unfortunately, several "stylized facts" appear to be highly correlated, which prevents the interpretation of the coefficients as "multipliers", since the ceteris-paribus assumption is not realistic.

1.2 Measures of importance

This paper focuses on the importance of individual variables on the classification of business cycle phases. According to the focus of an analysis, two types of statistical importance should be distinguished: the importance with respect to model selection and the importance with respect to value change.

The first type is based on all the measures used for model selection, like F-values (e.g. F-to-enter, F-to-remove), R^2, etc. (e.g. Rencher (1995)). The second type is based on the measures specifying value changes of the dependent variable if the predictor variable changes its value. For regression and linear classification models these are usually the coefficients to be estimated.

2 Correlated predictors in regression models

2.1 Overview

In order to develop an approach for measuring the importance of correlated predictors in classification models, it can be useful to discuss some results in

regression models, because a lot of research concerning correlated predictors has been done in this area. The assumption of uncorrelated predictors is often not appropriate for macro economic data. In fact, some of the variables are highly correlated. This usually leads to correlated regression coefficients, which are not easily interpreted (Assenmacher (2002)) and the coefficients cannot be interpreted as "multipliers".

In a regression model containing correlated predictor variables, there exist several approaches to handle highly correlated variables:

The first type of methods transforms the predictor variables to eliminate the correlations. For example orthogonalization of predictors, which is often carried out by sequential regression (e.g. Kruskal (1987)). Such methods allow the interpretation of the variables. On the other hand the coefficients highly depend on the order of variables entered into the model.

The second type of methods corrects the coefficients using a scalar shrinkage parameter, like ridge regression (e.g. Hoerl and Kennard (1969)). The drawback of these methods, is that the scalar added to the main diagonal of the covariance matrix does not improve interpretability.

The third type of methods tries to collect correlated variables into latent variable, thus reducing dimensionality. These methods often use principal component regression models (e.g. Hawkins (1973)). The principal components representation does also not allow the interpretation of the single variable effect under ceteris paribus conditions. Models which combine ridge regression and principal component regression have also been presented (eg. Stone and Brooks (1990)), but have the same drawbacks as the individual approaches.

2.2 Orthogonalization

Because the focus of this paper is on the interpretation of the single variable's influences, an orthogonalization method which is based upon sequential regression has been used here to address the multicollinearity problem. The disadvantage of this approach is, that the coefficients highly depend on the order, in which the variables have been entered into the model.

Based upon Kruskal's (1987) idea, Fickel (2002) proposes an algorithm to overcome this disadvantage. It estimates sequential regression models for every sequence i out of the $p!$ possible variable sequences. From these estimations the coefficients $\hat{\beta}_{ij}$ and the increase in the coefficient of determination $(\Delta R^2)_{ij}$ are stored for each variable j and each sequence i. Then for each variable

$$\hat{\gamma}_j = \frac{1}{p!}\sum_{i=1}^{p!}\hat{\beta}_{ij}\ , \quad \hat{\delta}_j = \frac{1}{p!}\sum_{i=1}^{p!}(\Delta R^2)_{ij} \tag{1}$$

are estimated. The average coefficient $\hat{\gamma}_j$ is a measure of importance for value change and can be interpreted as average effect of variable j, when all other

variables are held constant (like "multipliers") and the average R^2-increment $\hat{\delta}_j$ as relative importance of variable j for model selection.

The additional level shift of the residuals $\boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ introduced in Fickel's paper in order to scale the residuals to the same level as the original variables is not used here, since the business cycle data consist of growth rates.

3 Correlated predictors in classification models

3.1 Orthogonalization

Most classification models provide a method to estimate the membership probability of each class k, $k = 1, \ldots, K$, $p_{Mod}(k|\boldsymbol{X})$, using specific density assumptions and specific estimation criteria.

The approach of this paper is to use a linear probability model to estimate the importance of the correlated variables for the estimated class membership probabilities:

$$\hat{p}_{Mod}(k|\boldsymbol{X}) = \alpha_k + \boldsymbol{X}\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}_k, \quad k = 1, \ldots, K. \tag{2}$$

where $\boldsymbol{X}$ is the matrix containing a sample of the random variable $\boldsymbol{x}$. This is done by estimation of the K discriminant functions by an appropriate classification model Mod and of the posterior probabilities $\hat{p}_{Mod}(k|\boldsymbol{x}_i)$. These are then used as dependent variables in K linear regression models, one for each class.

Fickel's (2002) method can now be applied to each of the individual regression equations (2). For each class $k = 1, \ldots, K$, all $p!$ possible variable sequences are used to estimate sequential regression models. Then the importance measures $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$ are computed from the estimation results.

This orthogonalization procedure can be used for a great variety of classification methods. The only requirement is that the classification is based upon a membership function $m_{Mod}(k|\boldsymbol{x})$ or, even better, upon the estimated posterior probability function $\hat{p}_{Mod}(k|\boldsymbol{x})$. The usage of the posterior probability instead of the membership function $m(k|\boldsymbol{x})$ enables the comparability of the results of different classification methods. In this paper the results of a linear discriminant analysis and a multinomial logit are compared.

3.2 Using a large number of variables

The computation time of the proposed method increases excessively with the number of variables. The reason is that all possible $p!$ permutations of variable sequences have to be evaluated. One possible way to deal with this problem and to obtain interesting results is to choose a random subset of the $p!$ variable sequences.

The chosen subset must be uniformly distributed among the permutations of variable sequences. If the sample of variable sequences is big enough, the means of the coefficients and R^2-increments will be estimated well enough.

3.3 Results for the business cycle model

Using a random selection of variable orderings, all 13 "stylized facts" can now be used for the analysis. Instead of evaluating all 13! = 6,227,020,800 possible permutations of variable sequences, which takes a very long time even on fast computers, only 50,000 randomly chosen variable sequences are used for the analysis. The estimated total correlation matrix for all 13 stylized facts is

Table 1. Empirical correlation matrix for the 13 "stylized facts".

	Y	C	GD	L	X	M1	IE	IC	LC	PY	PC	RS	RL
Y	1.000												
C	0.776	1.000											
GD	0.403	0.387	1.000										
L	0.737	0.657	0.365	1.000									
X	-0.123	-0.154	-0.195	-0.074	1.000								
M1	0.318	0.423	-0.098	0.198	0.169	1.000							
IE	0.742	0.647	0.257	0.669	-0.160	0.314	1.000						
IC	0.680	0.518	0.279	0.505	-0.042	0.176	0.388	1.000					
LC	-0.170	0.108	0.180	0.153	-0.328	-0.139	-0.087	-0.179	1.000				
PY	-0.176	0.012	0.034	0.048	-0.257	0.000	-0.093	-0.175	0.868	1.000			
PC	-0.352	-0.347	-0.193	-0.203	-0.294	-0.143	-0.367	-0.270	0.567	0.723	1.000		
RS	-0.241	-0.308	-0.181	0.051	0.071	-0.359	-0.269	-0.187	0.426	0.493	0.616	1.000	
RL	-0.094	-0.365	-0.322	-0.226	0.201	-0.197	-0.185	-0.099	-0.656	-0.656	-0.118	0.156	1.000

shown in Table 1. A few variable pairs, like PY and M1 as well as C and PY are almost not correlated, but for most variable pairs, correlations are in effect. The highest positive correlations have LC and PY with 0.868 and C and Y with 0.776. RL and LC (-0.656) and RL and PY (-0.656) have the highest negative correlations. The variables GD, X and M1 do not have very high correlations with other variables.

Please note that bivariate correlations give only a vague impression of the underlying multicollinearity, which should be reflected in the corrections made by the orthogonalization procedure. Table 2 shows the estimated coefficients for the upswing class. The first column $\hat{\beta}_j$ contains the standard regression coefficients for the comparison to the importance measure γ_j obtained by the orthogonalization procedure. The most important variables for model selection are RS, C, PC, PY, LC, and IE. PY and RL have been corrected strongly in LDA and in the logit model. Table 3 shows the estimated coefficients for the upper turning point class. The most important variables are C, L, and Y. The greatest correction has been made for Y in LDA and for X in the logit model. RL and LC have also been corrected substantially in both models. Table 4 shows the estimated coefficients for the downswing class. The most important variables for model selection are RS, LC, PY, IE, and RL. The most substantial corrections can be observed for PY and RL in both models. Table 5 shows the estimated coefficients for the lower turning point class. The variables most important for model selection are L and Y. The strongest corrections are observed for RL and for PY in both models.

Table 2. Estimated coefficients for the upswing class.

	LDA			logit		
	$\hat{\beta}_j$	$\hat{\gamma}_j$	$\hat{\delta}_j$	$\hat{\beta}_j$	$\hat{\gamma}_j$	$\hat{\delta}_j$
const.	0.881	0.497	0.000	0.830	0.464	0.000
Y	0.042	0.030	0.031	0.034	0.024	0.023
C	-0.117	-0.081	**0.115**	-0.104	-0.077	**0.093**
GD	-0.025	-0.012	0.007	-0.034	-0.024	0.014
L	0.101	0.058	0.035	0.092	0.046	0.024
X	0.053	0.036	0.034	0.051	0.035	0.029
M1	-0.030	-0.017	0.033	-0.032	-0.017	0.030
IE	0.012	0.016	**0.073**	0.013	0.016	**0.065**
IC	0.000	0.001	0.006	0.001	-0.000	0.004
LC	-0.008	-0.035	**0.076**	-0.022	-0.045	**0.083**
PY	0.162	*-0.006*	**0.075**	0.157	*-0.008*	**0.071**
PC	-0.074	-0.085	**0.102**	-0.054	-0.071	**0.076**
RS	-0.161	-0.105	**0.191**	-0.154	-0.100	**0.156**
RL	0.104	*0.014*	0.025	0.100	*0.023*	0.025

Table 3. Estimated coefficients for the upper turning point class.

	LDA			logit		
	$\hat{\beta}_j$	$\hat{\gamma}_j$	$\hat{\delta}_j$	$\hat{\beta}_j$	$\hat{\gamma}_j$	$\hat{\delta}_j$
const.	-0.356	0.077	0.000	-0.292	0.093	0.000
Y	-0.017	*0.014*	**0.075**	-0.005	0.023	**0.078**
C	0.056	0.050	**0.149**	0.048	0.046	**0.111**
GD	-0.009	-0.012	0.016	-0.004	-0.006	0.007
L	0.055	0.066	**0.132**	0.043	0.064	**0.109**
X	-0.005	0.003	0.005	-0.015	*-0.068*	0.005
M1	0.011	0.014	0.063	0.011	0.013	0.042
IE	0.000	0.004	0.046	0.003	0.007	0.061
IC	-0.001	0.002	0.024	-0.002	0.002	0.024
LC	-0.029	*-0.014*	0.017	-0.031	*-0.014*	0.013
PY	-0.014	-0.003	0.010	-0.013	0.010	0.009
PC	0.022	0.012	0.012	0.025	0.018	0.013
RS	0.045	0.037	0.069	0.054	0.041	0.062
RL	0.003	*0.026*	0.023	-0.016	*0.013*	0.011

4 Discussion and outlook

An orthogonalization procedure has been proposed for classification models with correlated predictor variables. The procedure has been applied to west german business cycle data to model the four cycle phases upswing, upper turning point, downswing, and lower turning point. For 13 pre-selected stylized facts the classification models linear discriminant analysis and multinomial logit have been compared.

Table 4. Estimated coefficients for the downswing class.

	LDA			logit		
	$\hat{\beta}_j$	$\hat{\gamma}_j$	$\hat{\delta}_j$	$\hat{\beta}_j$	$\hat{\gamma}_j$	$\hat{\delta}_j$
const.	1.154	0.252	0.000	1.136	0.259	0.000
Y	0.006	-0.013	0.019	0.021	-0.007	0.018
C	0.033	0.022	0.019	0.021	0.015	0.013
GD	0.005	0.015	0.010	0.002	0.015	0.011
L	-0.046	0.002	0.017	-0.017	0.024	0.020
X	-0.051	-0.033	0.035	-0.026	-0.015	0.012
M1	0.008	-0.010	0.036	0.004	-0.013	0.036
IE	-0.015	-0.017	**0.092**	-0.018	-0.019	**0.106**
IC	-0.003	0.000	0.004	-0.004	-0.001	0.004
LC	0.041	0.053	**0.111**	0.064	0.067	**0.130**
PY	-0.336	*-0.094*	**0.093**	-0.348	*-0.108*	**0.090**
PC	0.058	0.043	0.050	0.072	0.047	0.044
RS	0.179	0.110	**0.244**	0.145	0.094	**0.167**
RL	-0.269	*-0.107*	**0.075**	-0.251	*-0.103*	**0.065**

Table 5. Estimated coefficients for the lower turning point class.

	LDA			logit		
	$\hat{\beta}_j$	$\hat{\gamma}_j$	$\hat{\delta}_j$	$\hat{\beta}_j$	$\hat{\gamma}_j$	$\hat{\delta}_j$
const.	-0.679	0.169	0.000	-0.673	0.186	0.000
Y	-0.032	-0.031	**0.070**	-0.050	-0.041	**0.075**
C	0.028	0.009	0.031	0.035	0.016	0.031
GD	0.029	0.010	0.016	0.037	0.015	0.018
L	-0.109	-0.126	**0.264**	-0.119	-0.134	**0.238**
X	0.003	-0.007	0.008	-0.010	-0.013	0.008
M1	0.010	0.013	0.033	0.016	0.016	0.039
IE	0.002	-0.003	0.041	0.002	-0.004	0.038
IC	0.003	-0.004	0.030	0.007	-0.002	0.022
LC	-0.005	-0.004	0.014	-0.012	-0.008	0.012
PY	0.188	*0.103*	0.066	0.203	*0.105*	0.051
PC	-0.006	0.030	0.041	-0.043	0.005	0.020
RS	-0.063	-0.042	0.063	-0.045	-0.034	0.038
RL	0.162	*0.068*	0.043	0.168	*0.069*	0.035

For model selection, RS, PY, LC, and IE seem to be important for both the upswing and downswing phases, whereas Y and L seem to have more importance for the turning point phases. For the upswing class additionally the consumption related variables C and PC seem to be important for model selection (C even for the upper turning point class) and RL seems to be characteristic for the downswing class.

The orthogonalization procedure corrects the coefficients for the predictors in such a way, that these corrected coefficients can be interpreted similar to "multipliers". The procedure also allows to compare different classification

models, as LDA and multinomial logit for this paper. Apparently, the results for both methods are similar for the business cycle data.

Similar comparisons will be done using other classification methods like Support Vector Machines. Also, the approximation of using a linear probability model should be overcome. These two threads will be followed during further research in this area.

References

ASSENMACHER, W. (2002): *Einführung in die Ökonometrie*, 6. Aufl. Oldenbourg Verlag, München.

FICKEL, N. (2002): Regression Analysis of Extremely Multicollinear Data. In: W. Gaul and G. Ritter (Eds.): *Classification, Automation, and New Media*. Springer, Berlin, 67–74.

GARCZAREK, U.M. and WEIHS, C. (2002): Incorporating background knowledge for better prediction of cycle phases. To be published in Knowledge and Information Systems.

HAWKINS, D.M. (1973): On the Investigation of Alternative Regressions by Principal Component Analysis. *Journal of the Royal Statistical Society, Series C: Applied Statistics, 22(1), 275–286*.

HEILEMANN, U. and MÜNCH, H.J. (1996): West german business cycles 1963–1994: A multivariate discriminant analysis. In: *CIRET-Conference in Singapore, CIRET-Studien 50*.

HOERL, A.E. and KENNARD, R.W. (1969): Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics, 12(1), 55–67*.

KROLZIG, H.-M. (1997): *Markov-Switching Vector Autoregressions. Modelling, Statistical Inference and Application to Business Cycle Analysis*. Springer, Berlin.

KRUSKAL, W. (1987): Relative importance by averaging over orderings. *The American Statistician, 41, 6–10*.

MEYER, J.R. and WEINBERG, D.H. (1975): On the classification of economic fluctuations. *Explorations in Economic Research, 2, 167–202*.

RENCHER, A. C. (1995): *Methods of Multivariate Analysis*. Wiley, New York.

RÖHL, M. C. WEIHS, C., and THEIS, W. (2002): Direct minimization of error rates in multivariate classification. *Computational Statistics, 17, 29–46*.

STONE, M. and BROOKS, R.J. (1990): Continuum Regression: Cross-validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Component Analysis. *Journal of the Royal Statistical Society, Series B, 52(2), 237–269*.

THEIS, W., VOGTLÄNDER, K. and WEIHS, C. (1999): Descriptive Studies on Stylized Facts of the German Business Cycle. Sonderforschungsbereich 475, Technical Report 45/1999, Universität Dortmund.

WEIHS, C. and GARCZAREK, U. (2002): Stability of multivariate representation of business cycles over time. Sonderforschungsbereich 475, Technical Report 20/2002, Universität Dortmund.

Economic Freedom in the 25-Member European Union: Insights Using Classification Tools

Clifford W. Sell *

Fachhochschule Würzburg-Schweinfurt
Münzstr. 12
97070 Würzburg, Germany

Abstract. In 2004, ten additional countries join the European Union. As a result, the nature of the community and its member countries are predicted to change, including the economic freedom of individuals and organizations. This study uses classification tools to look at the Economic Freedom of the World index (EFI). Patterns of economic freedom are quite different between the current and the acceding EU members. On average, economic freedom in Europe has a good chance of increasing as a result the expansion.

1 Introduction

On May 1, 2004, ten additional countries join the 15-member European Union. As a result, the nature of the community and its countries are predicted to change. Change is particularly expected for both the current and the acceding members with respect to institutional and legal arrangements. A common research proxy for a country's set of institutional and legal arrangements is the Economic Freedom of the World index (EFI) by the Fraser Institute, the Heritage Foundation, and 50 think tanks worldwide (Gwartney and Lawson (2002)).

This index measures more than the economic freedom of individuals and organizations. It is of general interest because of the importance of economic freedom to economic growth (Easton and Walker (1997)). Countries with higher economic freedom also tend to enjoy higher economic growth. Thus, if Europe is interested in accelerating its rate of economic growth, the EU has to make certain that the enlargement of the EU expands rather than restricts economic freedom. A good basis for such a development exists: individuals and organizations in the European Union have higher average economic freedom than their counterparts in the new members.

According to the European Commission (2003, p.3), "never before have (the candidates) been so thoroughly prepared, with a sweeping transformation of the economies and societies." The EFI scores underline this assessment; most of the new members rate well with respect to many of the EFI

* e-mail: cliffsell@t-online.de

variables, in particular in the area of government size and involvement. In some respects, the current EU members can also learn from the joining countries. The enlargement can thus be seen as a chance for both the current and the acceding EU members to increase economic freedom.

2 Data description and distance measures

The current Economic Freedom of the World annual report shows the rankings for the year 2000 for 123 economies. The three top-ranked countries are Hong Kong, Singapore, and the United States. Section 2.1 lists the five 'areas' of the freedom index, which in sum contain 21 variables. In order to get a feeling for the data, the averages for the current and the new EU countries, as well as the variable scores for five particular countries, are reported for all 21 variables. Section 2.2 takes a look at the Euclidian distances between the countries.

2.1 Description of the economic freedom index data

The economic freedom index covers five areas: (1) size of government, (2) legal system and property rights, (3) sound money, (4) freedom to trade with foreigners, and (5) regulation. Each of these areas, in turn, consists of three to five variables. The variables themselves can take on values between ten, associated with the most economic freedom, and zero, associated with no economic freedom.

Within each area, the variables are equally weighted. To arrive at the overall index value, again all groups are equally weighted. Each variable thus receives a different weight in the overall index: Variable 2-A, for example, has a weight of 1/5 in its area and the area has a weight of 1/5 in the index, giving it an overall weight of 1/25 in the index. Variable 5-A, however, has a weight of 1/3 in its area and the are has a weight of 1/5 in the index, giving variable 5-A an overall weight of 1/15 in the index.

When a value for a variable is missing, that variable is omitted in the calculation of the area value and the remaining variables are equally weighted. When an area value is missing, that area is omitted in the calculation of the EFI and the remaining areas are equally weighted. Because of missing values, the averages across countries and across areas do not yield the same numbers. For an alternative weighting schemes, see Guggiola (2002).

To provide an overview of the variables and values of the EFI, Table 1 shows the variable scores for three current EU members (Germany, Ireland and Spain) and two new EU members (Czech Republic and Lithuania). The 15 current members receive an average score of 7.5 (average EFI world rank of 19) and the ten new members a score of 6.4 (average world rank of 61).

The current EU members score highly with respect to sound money, the legal system, and the freedom to trade. The new EU members also score

Table 1. Economic Freedom Index (EFI) example country scores

Area	Description of Variable	EU15	GER	IRE	SPA	**EU+10**	CZE	LIT
1-A	Gov. consumption	**3.8**	4.6	5.3	5.2	**4.9**	4.1	4.4
1-B	Transfers/subsidies	**4.5**	4.5	4.2	5.2	**5.1**	2.4	6.9
1-C	Gov. enterprises/investment	**6.5**	6.0	10.0	4.0	**4.6**	8.0	6.0
1-D	Top marginal tax rate	**2.0**	2.0	5.0	4.0	**5.9**	7.0	7.0
1	Size of government	**4.6**	4.3	6.1	4.6	**5.1**	5.3	6.1
2-A	Judiciary independence	**8.1**	9.4	8.7	7.5	**5.7**	6.0	M
2-B	Impartial courts	**8.2**	9.2	9.2	8.0	**5.5**	4.5	M
2-C	Intellectual property protect.	**7.9**	8.8	7.0	7.2	**4.7**	5.6	3.2
2-D	Military in politics	**9.6**	10.0	10.0	8.3	**9.2**	10.0	8.3
2-E	Law and order	**9.0**	8.3	10.0	6.7	**7.7**	8.3	8.3
2	Legal system/property rights	**8.5**	9.1	9.0	7.5	**6.9**	6.9	6.6
3-A	Money growth	**8.9**	9.1	9.9	8.5	**7.5**	9.3	8.4
3-B	Std. dev. of annual inflation	**9.6**	9.7	9.2	9.5	**7.7**	8.5	3.8
3-C	Annual inflation	**9.4**	9.6	8.9	9.3	**8.8**	9.2	9.7
3-D	Foreign currency bank accts.	**10.0**	10.0	10.0	10.0	**5.0**	10.0	5.0
3	Sound money	**9.5**	9.6	9.5	9.3	**7.2**	9.2	6.7
4-A	Tariffs	**9.0**	9.0	9.0	9.0	**7.8**	6.7	8.3
4-B	Regulatory trade barriers	**9.1**	8.9	8.9	8.7	**7.2**	7.9	5.3
4-C	Size of trade sector	**5.1**	5.6	8.2	5.6	**5.6**	7.4	5.7
4-D	Black mkt. exchange rates	**10.0**	10.0	10.0	10.0	**9.9**	10.0	10.0
4-E	Capital market controls	**8.5**	9.5	8.6	8.0	**5.0**	7.0	7.8
4	Freedom trade w. foreigners	**8.3**	8.6	8.9	8.3	**7.2**	7.8	7.4
5-A	Credit market regulation	**8.3**	7.5	8.1	8.1	**6.8**	5.7	6.2
5-B	Labor market regulation	**4.4**	2.9	5.3	5.3	**4.6**	5.2	4.2
5-C	Business regulation	**7.5**	7.8	7.8	6.9	**6.4**	6.1	6.2
5	Regulation	**6.7**	6.1	7.1	6.8	**5.8**	5.7	5.6
EFI	Economic freedom index	**7.5**	7.5	8.1	7.3	**6.4**	7.0	6.5

highest with respect to the variables in these three areas, but score somewhat lower than the current 15 members. For the size of government, however, the 10 new EU members score higher on average than the current 15 members. For example, the Czech Republic and Lithuania have governments that are relatively smaller and less intrusive than those of Germany and Spain.

2.2 Distance measures

Measuring the distance between countries' EFI scores is one way to quantify the similarities and differences between countries. There are a number of such measures, the most well-known of which is the Pearson correlation coefficient. However, the correlation coefficient only measures linear association. The relationships between the particular 21 variables of the economic freedom index are not likely to be linear.

For a variety of reasons, any measure of dissimilarity between the entities i and j (in this case countries) should satisfy the three properties (for a more

detailed discussion, see Sell (2001)):

$$D_{ij} \geq 0 \qquad D_{ii} = 0 \qquad D_{ij} = D_{ji} \tag{1}$$

A dissimilarity measure is a metric if it also satisfies the following property, called triangle inequality, between entities h, i, and j. Some researchers consider being a metric to be a precondition of using a dissimilarity measure in cluster analysis.

$$D_{ij} \leq D_{ik} + D_{jk} \tag{2}$$

One geometrically appealing measure that fulfils the conditions of a metric, while accounting for non-linearities, is Euclidian distance, which takes the square root of the sums the squared distances between countries i and j over all 21 variables k (Everitt (1993), Gordon(1999)).

$$D_{ij} = \Big(\sum_{k=1}^{K}(x_{ik} - x_{jk})^2\Big)^{\frac{1}{2}} \tag{3}$$

For this study, the Euclidian distances are calculated based on the relative weights of the 21 variables in the freedom index. If fewer variables are available for any of the countries, only those variables existing for both countries of the country pair are included. Figure 1 provides a stylized overview of the

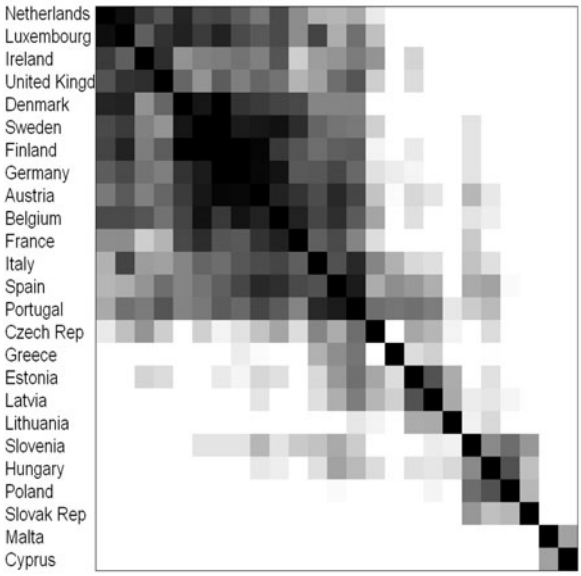


Fig. 1. Proximity Pattern

values of the Euclidian distances, with countries displayed on the horizontal and vertical axes. Country pairs with high similarity (low Euclidian distance) are dark, while pairs with a high Euclidian distance are light. The dark box covering the upper left quarter of the display shows that the countries from the Netherlands to Portugal are more similar than the entire set of countries. This group contains 14 of the 15 current EU members (the exception is Greece). Three smaller groups with high similarity are along the diagonal, for example one group containing the three Baltic countries.

3 Cluster analysis methods and cluster patterns

As the description of the proximities between countries illustrates, it is cumbersome to look at, let alone interpret, all 300 Euclidian distances in detail. It is, however, possible to impose a structure over this distance pattern using cluster analysis, which is discussed in section 3.1. Section 3.2 then describes the empirical cluster patterns.

3.1 Cluster analysis methods

The term cluster analysis covers a variety of methods for dividing data into homogeneous groups. They typically involve the following four steps:

Step 1: Choice of the measures that characterize the entities to be clustered. In the case of grouping countries cross-sectionally by their economic freedom, the alternatives are to either use the five areas or the 21 variables of the EFI. A larger number of variables results in greater stability of the results because missing or wrong data points exert a smaller influence. For this study, the 21 variables (or the number of variables available for any country pair) published for 2000 carry the same weights as in the index, as described in section 2.1.

Step 2: Calculation of the proximity or distance between the entities to be clustered. As discussed in section 2.2, this study uses the Euclidian distance.

Step 3: Recovery of clusters from the proximity matrix. There are a large number of theoretically appealing or empirically useful clustering algorithms from which to choose. For the purposes of this study, the results of the Monte Carlo studies presented in Milligan (1996) are used to select the so-called Ward algorithm. This clustering method initially allocates each country to a singleton cluster and then merges clusters in such an order that at each step the increase of the within-cluster sum of squared distances to the cluster means is minimized. The fusions of this clustering process can be displayed in a dendrogram, or tree diagram, which shows the order of fusions and the deviations from the cluster means.

Step 4: Evaluation and interpretation of the classification results. First the appropriate number of clusters from the statistically significant cluster

partitions has to be determined. In this case, the number of clusters was chosen that yields the best improvement in the deviates from the cluster means over the next smaller number of clusters. The detailed results are discussed below.

3.2 Empirical cluster patterns

The process of clustering the Euclidian distances of the 21 EFI variable scores for the 25 EU countries with the Ward algorithm is shown in figure 2. On the vertical axis are the 25 countries. On the horizontal axis is the loss of information when grouping, expressed as the average distance to the cluster mean. The dark shading shows the cluster members when the dataset is partitioned into six groups. This dendrogram shows that the countries fall

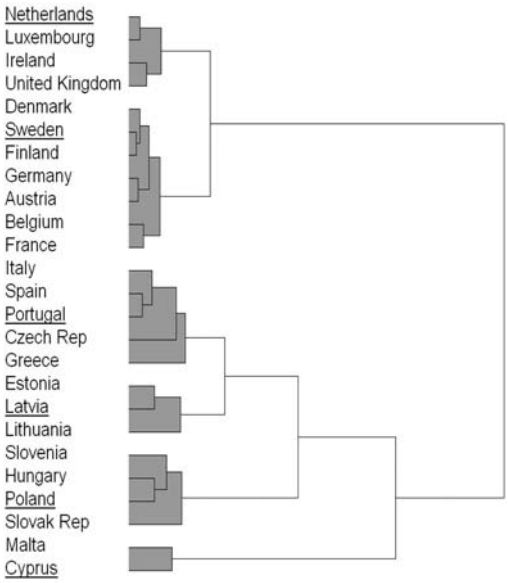


Fig. 2. Dendrogram

into (at least) two distinct groups: The Northwestern European countries (from the Netherlands to France) and the Southeastern European countries (from Italy to Cyprus). While the Northwestern group can be partitioned into two groups that are quite similar, the Southeastern group decomposes stagewise into four groups, some of which are quite heterogeneous.

Table 2 shows the average year 2000 scores of the six groups with respect to each of the five areas of the EFI. Group 1 (8.2): Netherlands, Luxembourg, Ireland, United Kingdom. This group has the highest average value of the

EFI. It also has the highest average scores for all of the five areas of the index, except with respect to the size of government. Group 2 (7.6): Denmark, Finland, Sweden, Belgium, Germany, Austria, France. The countries in this group receive the second highest overall score and are quite similar (but a little less free) than the countries in the first group. When looking at this area in detail, both groups' low average score for government size results from high government consumption, large transfers and subsidies, and a high top marginal tax rate. All countries in these two groups are current EU members.

Table 2. Average cluster scores of the five Economic Freedom Index areas for six country groups clustered using the Ward algorithm on Euclidian distances

Description of Variable	Grp 1	Grp 2	Grp 3	Grp 4	Grp 5	Grp 6	EU 15	EU +10
1 Size of government	5.4	3.7	5.3	6.0	4.0	5.9	4.6	5.1
2 Legal system	9.3	9.0	7.1	6.7	6.6	8.8	8.5	6.9
3 Sound money	9.6	9.5	9.3	7.5	6.6	7.0	9.5	7.2
4 Freedom to trade	8.7	8.4	7.9	7.8	7.1	5.4	8.3	7.2
5 Regulation	7.6	6.8	6.0	5.9	5.9	7.4	6.7	5.8
Economic Freedom Index	8.2	7.6	7.2	6.9	6.1	6.6	7.5	6.4

Group 3 (7.2): Italy, Spain, Portugal, Czech Republic, Greece. This group contains the current EU members with the (relatively) lowest economic freedom and the acceding EU member with the highest economic freedom. In 2000, these countries have sound money but have large governments. Group 4 (6.9): Estonia, Latvia, Lithuania. This group contains the Baltic countries, which are not very different from the countries in Group 3. These countries score higher than almost all of the current EU members for the size of government, with an average score of 6.0 (only Greece, the United Kingdom, and Ireland have smaller governments).

Group 5 (6.1): Slovenia, Hungary, Poland, Slovakia. This group has the lowest average EFI value. With the exception of the freedom to trade, these countries score lower than all other countries in every area of the EFI. Group 6 (6.6): Malta, Cyprus. These two small countries are the most different from the current and new EU members. In particular, the freedom to trade is more restricted in these countries. The results indicate that the countries in the last two groups need to change considerably to reach the current EU average.

Table 2 also shows the average scores for each area of the Economic Freedom Index for the current 15 and the acceding 10 EU countries. With respect to four areas, the current members score higher than the newcomers (in descending order of difference): Sound money, legal system, freedom to trade and regulation. With respect to government activity, however, individuals and organizations in the acceding countries enjoy greater economic freedom.

4 Conclusion and outlook

This study examines the similarities and differences of the current and the acceding European Union countries based on the Economic Freedom of the World Index. It calculates the Euclidian distances between 25 EU countries based on the 21 variables in the EFI and then assigns countries to groups using the Ward clustering algorithm. The ensuing groups reveal differing patterns of economic freedom. The groups also show a split into the Northwestern and the Southeastern countries of Europe.

Economic freedom is generally higher in the current EU countries, especially with respect to sound money, the legal system, freedom to trade and regulation. If the new members take steps to the change the institutions and legal frameworks in each of these four areas, it will increase economic freedom in Europe in many ways.

The average size of government of the current EU members has grown over the past 30 years. The opposite is true for the new EU members previously located behind the Iron Curtain, who have used the regime change to dramatically reduce the size and intrusiveness of their governments. If the current EU countries can learn from the new members with respect to government, especially government consumption and tax rates, economic freedom will also increase for individuals and organizations in the current EU countries.

References

COMMISSION OF THE EUROPEAN COMMUNITIES (2003): *Continuing Enlargement*. Brussels.

EASTON, S.T. and WALKER, M.A. (1997): Income, Growth, and Economic Freedom. *American Economic Review, 87, 328–382*.

EVERITT, B.S. (1993): *Cluster Analysis*. Arnold, London.

GORDON, A.D. (1999): *Classification*. Chapman and Hall, Boca Raton, FL.

GUGGIOLA, G. (2002): *The European Economic Freedom Index*. Working Paper, Centro Einaudi, Torino.

GWARTNEY, J.D. and LAWSON R.A. (2002): *Economic Freedom in the World: Annual Report 2002*. Fraser Institute, Vancouver, BC.

MILLIGAN, G.W. (1996): Clustering Validation: Results and Implications for Applied Analyses. In: P. Arabie, L. Hubert and G. De Soete (Eds.): *Clustering and Classification*. World Scientific Press, River Edge, NJ, 341–375.

SELL, C.W. (2001): *Finance Applications of Cluster Analysis*. Ph.D. Dissertation, Washington State University.

Intercultural Consumer Classifications in E-Commerce

Hans H. Bauer, Marcus M. Neumann and Frank Huber

Lehrstuhl für ABWL und Marketing II,
Universität Mannheim,
D-68131 Mannheim, Germany

Abstract. Global consumer typologies are an effective instrument for identifying regional consumer clusters and addressing different client needs in a focused fashion. The objective of this study is to examine whether the international online users are a homogeneous target group, or if it is possible to identify segments by means of selected criteria for constructing typologies. To answer the research question through an online survey in which interviewees participated from the cultural areas France, Germany and the US, theoretically secured constructs of the purchasing behaviour in the internet were obtained as well as different cluster analyses carried out. The results show that the internet users can be divided into three clusters - the risk aversive doubters, the open minded online shoppers and the reserved information seekers.

1 Introduction

To comply with the requirements of the strong international competition, the instruments of company policies have to be focused on the individual specificities of consumers. Consumer typologies are an effective instrument to comply with the requirement of identifying and addressing different consumer clusters despite of globally addressing markets. Only few typologies of internet users exist, such as the user typologies of the *Boston Consulting Group* or the one of *McKinsey* (Fritz (2001)), although the international competition in e-commerce intensifies due to the inherent ubiquity, the possibility of synchronically addressing many consumers and the lower market entry barriers of the distribution channel internet.

2 The concept of construction consumer typologies

The objective of a construction of typologies is to divide the totality of all individuals in groups that are as homogeneous as possible within a group, and as heterogeneous as possible between groups (Hair et al. (1995)). *Freter* (1983, p.43) understands under the construction of consumer typologies the division of buyers into subgroups as well as the treatment of one or more of these subgroups by means of segment specific marketing programs.

3 Characteristics for constructing typologies relevant for E-Commerce

3.1 Requirements regarding criteria used for constructing typologies

The present study organizes criteria for constructing typologies according to consumer behaviour by taking into account latent constructs of purchasing behaviour as features for segmentation. Because of the time stability and the inherent durability as criteria for segmentation, psycho graphical constructs are also being applied in the survey as criteria for the construction of typologies. To comply with the maxim of dividing a market into distinguishable and sound significant segments, and to guarantee methodological quality, the selection of relevant criteria for the construction of typologies in the frame of the envisioned research context will be discussed below.

3.2 Selected constructs for a classification

In the context of the construction of consumer typologies, the personality construct has an extraordinary position. Until today, the NEO-Five-Factor-Inventory (Costa and McCrae (1989)) appears among the most important measuring tools of personality research regarding the holistic systematization of the human personality structure. The results of the scientific research confirm that primarily the personality dimensions extraversion and neuroticism determine consumer behaviour in the internet (Kini and Chobineh (1998)). Several publications show that considerable intercultural differences exist in regard of the extent of these personality dimensions. Therefore, these personality dimensions represent a criterion for the construction of typologies in the context of this study.

Trust is a key element in successful customer relationship management (Hess (1995)). *Moorman, Deshpand* and *Zaltman* (1993, p. 82) define trust "as a willingness to rely on an exchange partner in whom one has confidence". A mayor reason for not shopping over the internet is the lack of trust by consumers into the medium and the shops selling their goods over the internet (Hoffman et al. (1999)). Due to its high relevance in consumer behaviour and its elementary importance for electronic commerce, trust is taken as additional typology criteria.

The construct of perceived risk contains all negative consequences of a purchase for a consumer which cannot be anticipated (Cunningham (1967)). *Cunningham* (1967) describes the perceived risk according to its components "Uncertainty" and "Consequences" or "dimension of loss". The key assumption of risk theory is the hypothesis that every individual has a limit of tolerance in terms of risk perception which is being determined by personal characteristics (Blackwell et al. (2001)). Action is therefore only triggered, when this individual's limit of tolerance is being crossed. In these cases, the

consumer feels a need to implement risk reducing actions. It has to be shown that perceived risk reduces the willingness to buy goods over the internet (Tan (1999)). The presented considerations give reason for selecting this construct as a typology criterion.

Attitudes have been used in market segmentation very successfully. The popularity of the attitude construct in market segmentation has a number of reasons. First of all, actions and intention can be forecasted by the positive or negative attitude towards an object (Lingenfelder and Loevenich (2003)). Second, results of attitude segmentation can be very helpful for designing marketing instruments. Third, attitudes are relatively constant over time. Due to these aspects, attitude toward online-shopping shall be added as a further typology criterion.

Especially in retailing research, shopping enjoyment plays an important role, since it determines the choice of the shopping location strongly and has a significant influence on willingness to buy of a consumer (Järvenpää and Tractinsky (1999)). The relevance of this construct for this study is argued because of its influence on the choice of shopping location and consumer loyalty. The results of scientific surveys show that shopping pleasure increases the chance of repeat buying (Eighmey and McCord (1998)). Due to these aspects, shopping pleasure is being added as typology criteria within this study.

Willingness to buy can be defined as a hypothetical construct that states how likely a person sees a purchase of a good concerning the shopping situation. It therefore considers the subjective judgement of the whole behavioural situation. The construct of willingness to buy allows a researcher to identify potential customers and reasons for positive or negative shopping intentions, since it not only represents the esteem of an individual but also holds the subjective judgement of the behavioural situation. *Kamakura and Wedel* (1995) argues that market segmentation based on this construct would be successful.

4 Empirical survey of the typology theory

4.1 Survey design and data collection

The data of this study was collected by an online questionnaire. Seven point Likert scales were used for testing the measurement models of the observed constructs (1= strongly agree; 7= strongly disagree). For the questionnaire, only existing inventories were used. To ensure a similar understanding of the wording across different cultures and to avoid semantic discrepancies the back translation method was used, as it has been suggested by *Berry* (1993). All measurement modells have been tested with a pretest of 87 probands. The final selection of these items was based on pretest-results of Cronbach's Alpha, item-to-total correlation and a Varimax-rotated exploratory factor analysis considering the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO).

A consequent recognition of fit indices (Coeff. Alpha>0.7; KMO>0.7) guaranteed reliability and validity of the construct measurement (Gerbing and Anderson (1988)). Results of a correlation analysis suggested that all constructs were uncorrelated (r<0.4).
The online survey was implemented from August 2nd until September 1st 2003. Of the 1011 probands, 45.8% were females and 54.8% males; 338 were from Germany, 337 were from the USA and 336 were French. The average age was 35.64 years within a range from 14 to 82 years. A descriptive analysis of the different country samples was made to measure a similar set of probands.

Table 1. Statistical value of the constructs.

Construct	**Based on**	**No. of Items**	**KMO**	**Coeff. Alpha α**	**Variance**
Extraversion	Costa and McCrae (1989)	8	0.786	0.7202	39.60%
Neuroticism	Costa and McCrae (1989)	8	0.836	0.7715	36.62%
Trust	Hess (1995)	9	0.906	0.8823	56.47%
Attitude (online-shopping)	Chen and Wells (1999)	6	0.864	0.9082	69.42%
Perceived risk	Järvenpää and Tractinsky (1999)	6	0.806	0.8201	53.23%
Shopping enjoyment	Ailawadi et al. (2001)	5	0.738	0.7970	55.67%
Willingness to buy	Baker et al. (1992)	4	0.812	0.8590	73.50%

4.2 A typology of online customers

An appropriate multivariate analysis method for this objective is the Centroid Clustering Analysis as it is provided in SPSS 11.5. Due to the large data sample, a four step cluster analysis was used to evaluate online customer segments. In the first step, the data input is being selected. The seven theory constructs, presented in chapter 2, are selected as input variables. In a second step, the cluster centres were built. The clusters were identified by running several k-means iteration processes, which used the average value for every variable as starting values, to be able to select the cases in a way internal homogeneity is being minimized (Hair et al. (1995)).
The process is repeated until the maximum number of iterations provided by the researcher is reached or until an additional selection of the objects provides no or little improvement to the former solution (Hair et al. (1995)).
The third step contains the determination of the cluster number. Since the data sample contains partial samples of the three different countries, a three cluster solution was calculated first and then benchmarked against a four and

five cluster solution. To judge the solution, the sum of the squared Euclidean distance was used. Since homogeneity was enhanced with an increasing number of clusters, a decision for the three cluster solution was made. After that the Centroids of the three clusters were assigned as starting centres for the iterative partionalizing k-means algorithm which provided the final solution. In the fourth step of the clustering process the observed cases are being assigned to the clusters. This is done by calculating groups within the cases, which maximize the internal homogeneity within the determined clusters. The final cluster centres of the study are provided in table 2. To test the first of the

Table 2. Mean values of the constructs for the different clusters (n = number of probands).

Construct	Cluster mean value (n)		
	1 (154)	**2** (400)	**3** (457)
Extraversion	3.68	3.04	3.83
Neuroticism	4.36	4.96	4.74
Trust	3.36	5.45	4.63
Attitude (online-shopping)	3.44	5.83	4.90
Percieved risk	4.87	2.61	3.45
Shopping enjoyment	2.93	5.77	4.47
Willingness to buy	2.83	6.15	5.31

solutions in terms if distinctions, a discriminant analysis was used. Due to the confirmatory character of this analysis, the clustering characteristics are defined as independent variables, while the cluster membership represents the nominally scaled dependent variable. The solution shows that 97.7% of the original groups' participants were classified correctly and confirms the very good fit of the three cluster solution, since the hit rate of the minimum criteria (97.7%) topped the maximum chance criteria (45.2%) clearly. To judge the classification capacity of a discriminant function correctly, the calculated hit rate has to be compared to the hit rate of a randomly structure of the elements (Hair et al. (1995)). A value of 0.15 for the multivariate Wilks-Lambda suggests a significant separation between the clusters by the discriminant function. The significance of the F-Values demonstrates that all related characteristic variables separate between the three groups at a significance level of 5%. Furthermore an analysis of variance has been used to evaluate whether the cluster centres are different. The results show to be highly significant on a 5% level for all constructs, so that the cluster solution can be interpreted.

- Cluster 1 "risk averse doubters"

Cluster 1 contains 154 (15.2%) of the Internet users. The evaluation of the averages for the personality dimension neuroticism and extraversion shows

Table 3. Results of the discriminant analysis (value in brackets = value of significance).

Discriminant function	**Variance**	**Canonical correlation**	**Wilks Lambda**	$\chi^2(p)$		**hit rate**	
1	96.2%	0.899	0.15	1809.57 (0.00)		97.7%	
2	3.8%	0.375					
		Univariate F (p)	**df**	**Coefficient standard.**		**Coefficient**	
				1. Fct.	**2. Fct.**	**1. Fct.**	**2. Fct.**
Neuroticism		62.05 (0.00)	2; 1008	-0.24	0.74	-0.11	0.63
Extraversion		24.47 (0.00)	2; 1008	0.13	0.28	0.11	0.01
Trust		348.49 (0.00)	2; 1008	0.26	-0.21	0.41	-0.04
Attitude (online-shopping)		427.79 (0.00)	2; 1008	0.34	-0.05	0.45	-0.04
Percieved risk		400.90 (0.00)	2; 1008	-0.38	-0.06	-0.43	-0.03
Shopping enjoyment		611.61 (0.00)	2; 1008	0.45	-0.48	0.53	-0.29
Willingness to buy		807.56 (0.00)	2; 1008	0.48	-0.48	0.61	0.59

that these people are extremely careful, reserved and usually sceptical against new experiences. Concerning the clustering variables, especially the highest perceived risk (4.87) and the lowest trust into an online-shop stand out. The averages of the latent buying relevant constructs, like "attitude towards online-shopping" and "willingness to buy" show the smallest values as well. The relative low shopping pleasure values support the point that Cluster 1 probands are critical about online shopping. In terms of national differences, this cluster is dominated by French probands (66.3%), while Germans (20.1%) and Americans are represented much less.

- Cluster 2 "open minded online shoppers"

Cluster 2 contains 400 Internet users and represents 39.6% of the total sample. The people in this cluster show little fear in all live situations ("neuroticism", 3.04) and are very open minded toward new things (extraversion, 4.96). The probands in this cluster show the lowest perceived risk when online shopping and at the same time the highest trust against an online shop. Therefore they show a very positive attitude (5.83) and the highest willingness to buy (6.15). The high rated shopping pleasure (5.77) the strong affinity of this cluster for online shopping. The cultural comparison of this cluster shows that Americans are represented most (44.5%) while only 22.0% are French. In terms of usage, this cluster has the strongest percentage of online-shoppers (97%), of which even 62% shop often or very often.

- Cluster 3 "reserved information seekers"

Cluster 3 contains 457 probands (45.2%). The analysis of the averages for neuroticism (3.83) and extraversion (4.74) shows that these people are in av-

erage careful and reserved. A positive opinion against online shopping (4.90), over average shopping pleasure (4.47), reasonable trust in an online-shop (4.63) and a relatively high willingness to buy (5.31) proof that this cluster is generally open to shop goods over the internet. Analysing the passive variable shows that the representatives of this cluster use the internet mainly to search for information (94%) and pre-purchase products evaluations 84%. Compatible with this is the high usage of search engines (95%). The cluster contains mainly German citizens (37.9%), while Americans are less represented (30.2). The solution is plausible and a comparison of averages of the different typology characteristics shows strong numeric differences. The strongest characteristic of the cluster solution is the cultural differences of the different clusters, however the country didn't stand as a typology criteria and only represented a passive variable for the interpretation of the clusters. The cluster "risk averse doubters" contains 66.2% French but only 13.6% Americans. On opposite to that, the cluster "open minded online shoppers" is being dominated by American probands, while only 22.0% of French internet-users are represented.

5 Conclusion

From a marketing perspective the identified typology refines former approaches of classifications by adding personality dimensions and cultural determinants in combination with the shopping behaviour of online consumers. The cluster solution provides significant differences between the three countries. Based on the three cluster solution, useful management implications can be set up that allow a better satisfaction of customers by providing a closer fit between a company's goods and services and their better understood heterogeneous customer needs. The main findings are the identification of different clusters of internet users, which can be of good use for shaping internet marketing, because the stability over time due to its cultural and personality characteristics.

This study shows that there is no homogeneous cyber community and that the adoption of marketing activities to the needs and expectations of customers must take cultural differences into consideration. However the three cluster solution clearly proved culturally shaped customer segments, every identified segment contains users of all three countries.

References

AILAWADI, K. L., NESLIN, S. A. and GEDENK, K. (2001): Pursuing the Value-Conscious Consumer: Store Brand Versus National Brand Promotions. *Journal of Marketing, 65, 71–89.*

BAKER J., LEVY M. and GREWAL D. (1992): An Experimental Approach to Making Retail Store Environmental Decisions. *Journal of Retailing, 68, 445–460.*

BERRY, J.W. (1993): An ecological approach to understanding cognition across cultures. In: J. Altarriba (Ed.): *A Cross-Cultural Approach to Cognitive Psychology*. Elsevier Science Publishers, Amsterdam, 361–375.

BLACKWELL, R., MINIARD, P. and ENGEL, J. (2001): *Consumer Behavior*, ITPS Thompson Learning, Fort Worth.

CHEN, Q. and WELLS, W. (1999): Attitude toward the Site. *Journal of Advertising, 39, 27–37*.

COSTA, P.JR. and MCCRAE, R. (1989): *NEO PI/FFI Manual Supplement*. Psychological Assessment Resources, Odessa.

CUNNINGHAM, S.M. (1967): he Major Dimensions of Perceived Risk. In: E.D.F. Cox (Ed.): *Risk Taking and Information Handling in Consumer Behavior*, Harvard University Press, Boston, 82–111.

EIGHMEY, J. and MCCORD, L. (1998): Adding value in the information age: Uses and gratifications of sites on the WWW. *Journal of Business Research, 41(3), 187–194*.

FRETER, H. (1983): *Marktsegmentierung.*,Kohlhammer, Stuttgart.

FRITZ, W. (2001): *Internet-Marketing und E-Commerce*. Gabler, Wiesbaden.

GERBING, D. and ANDERSON, J. (1988): An Updated Paradigm for Scale Development Incorporating Undimensionality and its Assessment. *Journal of Marketing Research, 25(2), 186–192*.

HAIR, J.F., ANDERSON, R.E., TATHAM, R.L. and BLACK, W. C. (1995): *Multivariate Data Analysis*. Prentice-Hall, Englewood Cliffs 1995.

HESS, J. (1995): Construction and Assessment of a Scale to Measure Consumer Trust. In: B.B. Stern and G.M. Zinkhan (Eds.): *AMA Educator's Proceedings*. Chicago, 20–26.

HOFFMAN, D.L., NOVAK,T.P. and PERALTA, M. (1999): Building consumer trust online. *Communications of the ACM, 42(4) 80–85*.

JÄRVENPÄÄ, S.L. and TRACTINSKY, N. (1999): Consumer Trust in an internet store: A Cross-Cultural Validation. *JCMC, 5(2)*.

KAMAKURA, W. and WEDEL, M. (1995): Life-style segmentation with tailored interviewing. *Journal of Marketing Research, 32(8), 308–317*.

KINI, A. and CHOOBINEH, J. (1998): Trust in electronic commerce. *Proceedings of the Thirty-first Hawaii international Conference on System Science*. Maui, 51–61.

LINGENFELDER, M. and LOEVENICH, P. (2003): Identifikation und Auswahl von Zielgruppen im E-Commerce. *Marketing ZFP, 25, 119–131*.

MOORMAN, C., DESHPAND, R. and ZALTMAN, G. (1993): Factors Affecting Trust in Market Research Relationships. *Journal of Marketing, 57, 81–101*.

TAN, S.J. (1999): Strategies for reducing consumers risk aversion in Internet shopping. *Journal of Consumer Marketing, 16, 163–180*.

Reservation Price Estimation by Adaptive Conjoint Analysis

Christoph Breidert[1], Michael Hahsler[1], and Lars Schmidt-Thieme[2]

[1] Department of Information Business,
Vienna University of Economics and Business Administration,
1090 Vienna, Austria
[2] Computer-based New Media group,
Institute for Computer Science,
University of Freiburg, 79110 Freiburg, Germany

Abstract. Though reservation prices are needed for many business decision processes, e.g., pricing new products, it often turns out to be difficult to measure them. Many researchers reuse conjoint analysis data with price as an attribute for this task (e.g., Kohli and Mahajan (1991)). In this setting the information if a consumer buys a product at all is not elicited which makes reservation price estimation impossible. We propose an additional interview scene at the end of the adaptive conjoint analysis (Johnson (1987)) to estimate reservation prices for all product configurations. This will be achieved by the usage of product stimuli as well as price scales that are adapted for each proband to reflect individual choice behavior. We present preliminary results from an ongoing large-sample conjoint interview of customers of a major mobile phone retailer in Germany.

1 Introduction

Pricing products is a difficult task for every business. Thorough knowledge of the demand in the market is necessary to predict the different effects that arise from the pricing strategy as well as from the set price for a product: Customer switching effects, cannibalization effects, and market expansion or contraction effects. Many of these effects can be analyzed for different strategies using the reservation prices of the participants in the market (Jedidi and Zhang (2002)). Varian (2003, p. 4) defines the reservation price as follows:

> The reservation price is the highest price that a given person will accept and still purchase the good. In other words, a person's reservation price is the price at which he or she is just indifferent between purchasing or not purchasing the good.

However, this definition is different from the definition of reservation price used by other authors. For example, Kohli and Mahajan (1991) define the reservation price for their study as the price for a product such that an individual switches away from her most preferred product. To our knowledge

Jedidi and Zhang (2002) are the only researchers who have applied the economic definition of reservation price in combination with a conjoint study on product pricing.

In this paper we present a novel approach to estimate the economic reservation price using the popular conjoint analysis. We do not incorporate price as an attribute in the conjoint analysis but we introduce price by an additional choice-based scene after the conjoint analysis. The paper is organized as follows: In section 2 we identify shortcomings of the estimation of reservation prices using only data from conjoint analysis. In section 3 we present our novel approach and its foundation in economic theory. In section 4 we outline an application of the method for a mobile phone retailer. We conclude with a short discussion of further research.

2 Conjoint analysis for reservation price estimation

Conjoint analysis and especially adaptive conjoint analysis (ACA) (Johnson (1987)) is a popular tool in marketing research to survey consumers' preferences for products that are seen as the combination of several attributes which have different levels. With conjoint analysis utility-scores for the attribute levels are estimated that reflect the respondents valuations of the inclusion, exclusion or degree of the levels.

The major approach in pricing studies by conjoint analysis is incorporating the price as an additional attribute (e.g., Green and Srinivasan (1990), Orme (2001)). The attribute price is then assigned a part-worth utility as the other attributes and some interpolation heuristics are applied. To estimate reservation prices several studies using conjoint data are found in the current literature (e.g., Kohli and Mahajan (1991), Jedidi and Zhang (2002)). In these studies authors try to estimate reservation prices from previously acquired conjoint data which include price as an attribute. However, these approaches have the following shortcomings:

1. Conjoint analysis only measures the preference structure for the analyzed product configurations. If the individual would really purchase at a given price is not elicited. Therefore, reservation price in an economic sense cannot be measured.
2. The *number-of-levels* and the *range* effect are well-known in conjoint analysis (Verlegh et al. (2002)). If the number of attribute levels or the range covered by the attribute levels is increased by the researcher, the perceived importance of that attribute also increases. These effects are especially problematic for pricing studies in which often a large number of different prices is surveyed.

In the following we address these issues by excluding the price from the conjoint analysis and estimate the reservation price with an additional interview scene which also allows for non-purchases.

3 Reservation price estimation based on economic theory

Following Varian (2003, p. 63) a utility function for two products X and Γ can be formulated as

$$U(x, \gamma) = u_X(x) + u_\Gamma(\gamma). \tag{1}$$

Hereby x is the amount of product X, for which the reservation price of one specific individual is to be estimated, and γ denotes the amount of the so-called composite product Γ. Varian (2003, p. 21) defines the composite good as everything that the consumer might want to buy other than good X. By definition the amount of money not spent on good X is spent on good Γ. Note, that the composite good is arbitrarily divisible and also includes the possibility to save money for later consumption.

The reservation price for a good X is defined as "the price at which the consumer is just indifferent between consuming good X or not consuming it" (Varian (2003, pp. 108–109)). Therefore, the reservation price p_X^* for one unit of product X is found, when the customer is indifferent between purchasing or not purchasing the product. Formally, indifference can be expressed by the following condition

$$U(1, \gamma) = U(0, \gamma') \qquad \text{where} \qquad \gamma' > \gamma. \tag{2}$$

On the left hand side of this equation the utility is given for an individual who consumes one unit of product X and consumes some amount of the composite good. On the right hand side of the equation the individual does not consume product X and therefore consumes a greater amount of the composite good denoted by γ'.

When consuming the goods X and Γ at the unit prices p_X and p_Γ each consumer is confronted with an individual budget constraint which can be defined as $m = p_X x + p_\Gamma \gamma$. Since the composite good is defined to be arbitrarily divisible, we can set the price for one unit of Γ to 1 (Varian (2003, p. 21)). For the consumption and non consumption of one unit of product X the following equations derived from the budget constraint hold

$$\gamma = m - p_X \tag{3}$$

$$\gamma' = m. \tag{4}$$

We only consider the case of buying one or zero units of X. For zero units no utility is derived from X ($u_X(0) = 0$). For the sake of formal simplicity let u_X denote the utility of consuming one unit of product X, that is $u_X := u_X(1)$. Using the utility function in equation 1 the condition for indifference in equation 2 can be rewritten as

$$u_X + u_\Gamma(\gamma) = u_\Gamma(\gamma'). \tag{5}$$

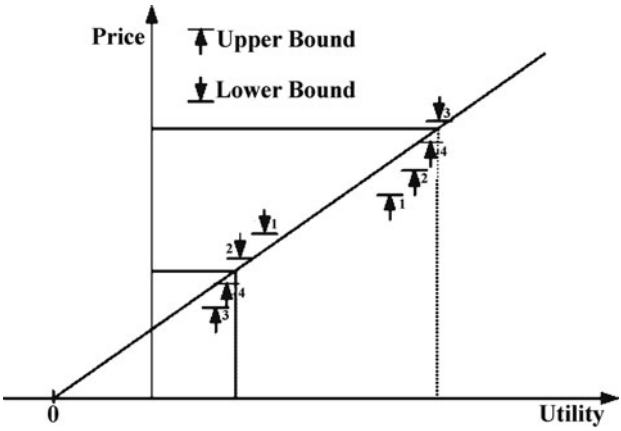


Fig. 1. Estimation of a high and a low reservation price point in the PE scene from observed upper- or a lower bounds of the price (denoted by the arrows).

When consuming the composite good, an individual will certainly always choose a combination that gives her the highest utility for her budget. Since the composite good is arbitrarily divisible and constructed from all possible goods (except good X), the consumer will face a large number of different combinations which equally have the same highest utility per price ratio k. Therefore, for $u_\Gamma(\gamma)$ a linear function with slope k and an intercept $u_\Gamma(0) = 0$ can be used (compare Jedidi and Zhang (2002)).

$$u_X + k \cdot \gamma = k \cdot \gamma' \tag{6}$$

Applying the budget constraint from equations 3 and 4 the following condition for the consumption of one unit of product X at the reservation price p_X^* can be formulated

$$u_X + k \cdot (m - p_X^*) = k \cdot m. \tag{7}$$

Applying some simple arithmetics to the equation m can be eliminated. Then, if the utility and the reservation price for one unit of product X is known, the slope k of the utility function of the composite product Γ can be calculated by

$$k = \frac{u_X}{p_X^*}. \tag{8}$$

Economically, the factor k represents the exchange rate between utility and money. With the factor k the reservation price for any product configuration for which the utility is known can be calculated. Note, that this calculation is based on ratio-scaled absolute utility but the conjoint analysis only produces interval-scaled utility-scores for products.

find-reservationprice-point$(u, p, \Delta u, \Delta p, \Delta p_{stop}, s_{max}, i)$:
$b^+ := (u^+, p^+) := \emptyset, b^- := (u^-, p^-) := \emptyset, j := 1$
while $b^+ = \emptyset$ or $b^- = \emptyset$ **do**
  if purchase(product$(u), p)$
    $b^- := (u, p)$
    $(u, p) := (u + \Delta u, p + j\Delta p)$
  else
    $b^+ := (u, p)$
    $(u, p) := (u - \Delta u, p - j\Delta p)$
  fi
  $j := j + i$
od
while $p^+ - p^- > \Delta p_{stop}$ and s_{max}-- > 0 **do**
  $(u, p) := (\frac{1}{2}(u^- + u^+), \frac{1}{2}(p^- + p^+))$
  if purchase(product$(u), p)$
    $b^- := (u, p)$
  else
    $b^+ := (u, p)$
  fi
od
return $(\frac{1}{2}(u^- + u^+), \frac{1}{2}(p^- + p^+))$

Fig. 2. Search algorithm for a reservation price point.

However, in the following we will show how to transform the interval-scaled utility-scores to ratio-scaled absolute utility while estimating the factor k. For this transformation we append the new *Price Estimation scene* (PE scene) at the end of the adaptive conjoint analysis. At this point all part-worth utilities are already estimated by the conjoint analysis and the utility-scores for all attribute combinations can be calculated. The PE scene is a choice-based scene where we offer the proband several times a different product at a dynamically set price and he or she has the option to accept the offer or leave it. With these questions we iteratively search for two reservation price points in the *utility* $\times$ *price* space. As shown in figure 1 with every question we find an upper or lower bound for price at a certain utility. Once we have found the reservation prices for two different products a straight line through the two points gives us an estimate for the factor k. At the same time we get an intercept with the utility axis which represents the conjoint analysis utility-score for the price 0 which by economic theory must correspond to an absolute utility of 0. Therefore, as shown in figure 1, we can assign an origin to the utility axis and utility is now ratio-scaled as necessary for reservation price estimation.

The algorithm for the estimation of the reservation price of one product combination is presented in figure 2. The function *product*(u) chooses the product configuration closest to a desired utility u from the list of all possible

combinations. Function $purchase(product(u), p)$ asks whether the user would buy the product chosen by $product(u)$ at a given price p.

The first while loop in figure 2 starts with an initial guess (u, p). The algorithm tries to box the probands utility/price exchange ratio by locating an upper and a lower bound (b^+, b^-), i.e., a price point at which the proband would purchase for a given utility and one at which the proband would decline to purchase. In the second loop of the algorithm this interval is gradually narrowed by a bisection search. The bisection search terminates when the found interval, in which the reservation price lies, is narrowed to a predefined accuracy Δp_{stop}. To limit the maximal number of purchasing decisions a participant has to make, a second termination condition restricts the algorithm to a predefined maximal number of search steps s_{max}.

If only two reservation price points, (u_1, p_1) and (u_2, p_2), are used, the utility/price exchange ratio can be easily found by $k = (u_2 - u_1)/(p_2 - p_1)$. When $n > 2$ reservation price points (u_i, p_i) are used the utility/price exchange ratio can be found by least squares fitting.

The possibility that the procedure influences the respondent's behavior needs attention. To avoid this influence the respondent should be explicitly asked to view each offer independently. Furthermore, the respondent can be presented product combinations from the n reservation price estimations in randomized or alternating order (e.g. alternating high utility with low utility combinations), such that the influence is minimized.

4 Application of the method

We implemented the PE scene in the modular framework of the Java Adaptive Conjoint tool (jAC version 1.1, Schmidt-Thieme (2004)) and incorporated it in a study designed for the NOKIA online-shop in the German market for mobile phones and accessories. In this shop customers are offered suitable telephone enhancements at discounted price on the purchase of a telephone. In terms of Pigou (1920) this strategy can be described as price discrimination of the third degree, because the shown telephone enhancements are only offered to a certain group of people at a lower price. The strategy can also be viewed as a mixed-bundling strategy as described by Adams and Yellen (1976). The telephone is offered together with enhancement at a discount, but the products can also be purchased individually without a discount. At the moment the marketing experts of the online-shop set the discounts for the telephone enhancements manually in view of the cost structure and sales information of the different products.

To enable the online-shop to optimize the pricing strategy we estimate the reservation prices of customers at the individual level. First, we use the adaptive conjoint analysis to estimate the part-worth utilities of all attribute levels excluding the price information. And then, we use the Price Estimation scene to estimate the reservation prices.

Table 1. Estimated reservation prices for a sample proband.

Utility	Reservation Price	Extra Charger	Car Accessory	Headset	Leather Case
17,64	171,46 EUR	ACP-12E	LCH-12	HDW-2	CNT-327
17,26	168,59 EUR	DCV-14	LCH-12	HDW-2	CNT-327
17,05	167,01 EUR	ACP-12E	-	HDW-2	CNT-327
16,67	164,14 EUR	DCV-14	-	HDW-2	CNT-327
16,60	163,63 EUR	ACP-12E	MBC-15S	HDW-2	CNT-327
16,22	160,77 EUR	DCV-14	MBC-15S	HDW-2	CNT-327
15,99	159,05 EUR	ACP-12E	LCH-12	HS-3	CNT-327
15,97	158,93 EUR	ACP-12E	LCH-12	HDW-2	-
15,61	156,19 EUR	DCV-14	LCH-12	HS-3	CNT-327
15,59	156,06 EUR	DCV-14	LCH-12	HDW-2	-
15,43	154,87 EUR	-	LCH-12	HDW-2	CNT-327
15,40	154,60 EUR	ACP-12E	-	HS-3	CNT-327

A large-sample online study will be carried out with the newsletter recipients of the NOKIA online-shop later this year. Here we only show how the procedure works by presenting results from a single sample participant. We searched for two reservation price points (with utilities around the 0.25 and 0.75 quantiles). Utility increments Δu were chosen to allow 20 steps in the search procedure. Δp_{stop} was set to 2,- EUR. Initial guesses for prices, price increments, and increase in step length were set by domain experts. Table 1 contains a subset of the results for the sample proband. The exchange rate between utility (measured by the conjoint analysis) and reservation price was estimated to $utility = 0.13 \cdot price - 5.22$ (rounded values). The stimuli of the conjoint analysis consisted of a fixed telephone and contract with different additionally bundled components.

From a single interview we can estimate reservation prices for all product combinations at the individual level. However, we can also aggregate the data to estimate reservation prices at market-level. To avoid the problem of preference heterogeneity we can segment the customers by self-selection, i.e., the preference for a certain phone type (business, fun, etc.), demographic variables, or characteristics of the self-explicated task of the adaptive conjoint analysis (Moore et al. (1998)). For these, more homogeneous groups distributions of reservation prices can be estimated. By applying an appropriate choice rule market reaction at different prices can be predicted.

5 Conclusion and further research

The approach presented in this paper addresses shortcomings of traditional pricing studies with conjoint analysis that arise from including price as an attribute in the study. We exclude price from the conjoint analysis and estimate it in an additional interview scene. With this procedure the *number-of-levels effect* and the *range* effect do not occur for price. Furthermore, by the use of

a no-purchase option we can also measure the reservation price as defined in economic theory.

The new approach needs to be tested in a real setting which will be done in a large-sample reservation price survey together with the NOKIA online-shop. Further research is also necessary to compare this approach to traditional pricing studies and other techniques of reservation price estimation as described by Sattler and Nitschke (2003).

Finally, it has to be noted that the presented approach is not bound to conjoint analysis. Any estimation method that delivers preference information for products and product combinations relatively scaled at the individual level can be combined with our new estimation scene.

References

ADAMS, W.J. and YELLEN, J.L. (1976): Commodity Bundling and the Burden of Monopoly. *Quarterly Journal of Economics, 90, 457-498.*

GREEN, P.E. and SRINIVASAN, V. (1990): Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice. *Journal of Marketing, 54, 3-19.*

JEDIDI, K. and ZHANG, Z.J. (2002): Augmenting Conjoint Analysis to Estimate Consumer Reservation Price. *Management Science, 48, 1350–1368.*

JOHNSON, R.M. (1987): Adaptive Conjoint Analysis. *Sawtooth Software Conference on Perceptual Mapping, Conjoint Analysis, and Computer Interviewing. Sawtooth Software Inc., 253–265.*

KOHLI, R. and MAHAJAN, V. (1991): A Reservation-Price Model for Optimal Pricing of Multiattribute Products in Conjoint Analysis. *Journal of Marketing Research, 28, 347–354.*

MOORE, W.L., and GRAY-LEE, J. and LOUVIERE J.J. (1998): A Cross-Validity Comparison of Conjoint Analysis and Choice Models at Different Levels of Aggregation. *Marketing Letters, 9(2), 195–207.*

ORME, B. (2001): Assessing the Monetary Value of Attribute Levels with Conjoint Analysis: Warnings and Suggestions. *Research Paper Series. Sawtooth Software Inc.*

PIGOU, A.C. (1920): *The Economics of Welfare.* McMillan Press, London.

SATTLER, H. and NITSCHKE, T. (2003): Ein empirischer Vergleich von Instrumenten zur Erhebung von Zahlungsbereitschaften, *Zeitschrift für betriebswirtschaftliche Forschung (ZfbF), 55, 364–381.*

SCHMIDT-THIEME, L. (2004): jAC – A Modular Framework for Online Adaptive Conjoint Analysis, *Technical report, University of Freiburg. http://www.informatik.uni-freiburg.de/cgnm/jac.*

VARIAN, H.R. (2003): *Intermediate Economics - A Modern Approach.* W. W. Norton and Company, New York, London.

VERLEGH, P.W., SCHIFFERSTEIN, H. N. and WITTINK, D.R. (2002): Range and Number-of-Levels Effects in Derived and Stated Measures of Attribute Importance. *Marketing Letters, 13, 41-52.*

Estimating Reservation Prices for Product Bundles Based on Paired Comparison Data

Bernd Stauß and Wolfgang Gaul

Institut für Entscheidungstheorie und Unternehmensforschung
Universität Karlsruhe (TH), 76128 Karlsruhe, Germany

Abstract. Reservation prices have evolved as important tool for designing and pricing new products or bundles of products where a reservation price for an item can be interpreted as maximum amount of money that a consumer is willing to pay for that item. In this paper, focusing on product bundles, two types of data collection - an already known and a new one - based on direct elicitation of reservation prices using paired comparison data are discussed. Variants of conjoint analysis that were proposed so far in this context are used, an explicit evaluation of two methods is described, and an example by means of empirical data from a seat system offered by a German car manufacturer is used as demonstration of the applicability of the methodology suggested.

1 Introduction

Bundling, i.e., the joint offer of two or more different products or services that are sold at a unique bundle price has experienced growing attention in marketing since it was first discussed in the early sixties in a primary legal context (Stigler (1963)). Most of the work that was done on this topic was rather theoretical analysis (Adams and Yellen (1976)) than normative modeling. However, there are two main exceptions, i.e., the mathematical approach of the bundle pricing problem by Hanson and Martin (1990) and an alternative formulation by Stauß and Gaul (2004) that facilitates modeling and allows for an intuitive solution heuristics. The incorporation of the reservation price concept is common to all formal modeling where reservation prices normally are given on a product rather than an attribute level[1]. Reservation prices may be seen as the maximum amount of money that someone is willing to pay for respective products or, equivalently, (parts of) bundles. Though reservation prices are often identified as an essential determinant of bundling strategies, only a few authors discuss the estimation problem for the unknown (individual) reservation prices (Aust (1995); Jedidi and Zhang (2002) or Jedidi et al. (2003)). This is somewhat astonishing as many authors explicitly use some special types of distributions (e.g. normal or unit) underlying reservation prices and analyze their influence on bundling outcomes.

[1] Chung and Rao (2003) discuss a general framework of preference models for bundles and provide an overview of attribute based models.

The reservation price concept is closely related to the well known theory of "willingness to pay" (WTP) which has its roots in economic literature and was originally used in the context of evaluation tasks for public or environmental resources. WTP has the amenable property that it implicitly transforms individuals' preferences into a monetary valued utility measure, which seems to be particularly appealing if market prices for the respective resources are not available. Therefore, WTP became a famous construct within projects of administration authorities, since it allows for a direct comparison between costs and utility. A common (and probably the most favored) practice used to find out WTP is contingent valuation first employed by Davis (1963) and is described,e.g., by Mitchell and Carson (1989, p. 3).

2 Gathering data for conjoint measurement

For quite some time, marketing literature ignored the need for tailored estimation procedures of reservations prices though the use of data analysis techniques for utility estimation became more prevalent. Finally, in the late eighties and early nineties both directions - the direct contingent valuation method and decompositional data analysis, namely conjoint measurement - were merged by the work of Cameron and James (1987) as well as Reardon and Pathak (1990).

2.1 Direct vs. indirect elicitation of reservation prices

Basically, two different techniques exist for eliciting reservation prices within the conjoint analysis framework[2] depending on the functional form of the underlying preference model.

Suppose that the preference of a consumer i for a bundle k (which is defined by a set of components $\mathcal{B}_k$) may be described by a mixed model consisting of a part worth formulation with respect to the components j that form the respective bundle (referred to as $u_{ik} = \sum_{j \in \mathcal{B}_k} \beta_{ij}$ where β_{ij} indicates consumer's i part worth for component j) and a vector model with respect to bundle prices. Thus, if α_i is consumer's i price sensitivity, $u_{ik} - \alpha_i p_k$ measures consumer's i utility for bundle k at a price p_k. Traditional conjoint analysis can be used to estimate the price sensitivity parameters α_i so that we get an estimate of the reservation price r_{ik} of bundle k by consumer i via $\frac{u_{ik}}{\alpha_i}$.

On the other hand, there is an alternative way of measuring reservation prices. Since $\frac{u_{ik}}{\alpha_i} - p_k$ is just a rescaling of consumer's i utility for a bundle k usually referred to as "consumer surplus" (Kalish and Nelson (1991)), one may ask a consumer directly at what price p_k^* he would be indifferent to

[2] See, e.g., Aust and Gaul (1995) or Baier and Gaul (2003) and the references cited there for conjoint analysis applications to product designs.

buying bundle k or doing nothing at all.[3] Suppose further that if the consumer doesn't purchase anything the utility he gets is zero. Thus, $r_{ik} = \frac{u_{ik}}{\alpha_i} = p_k^*$ holds, i.e., instead of processing rating or rank data we may use metric price data directly. Obviously the conjoint design becomes more tractable, since price has not to be modeled explicitly any more and the number of "attributes" is reduced by one. However, consumers are obliged to make valuations on a dollar-metric scale which is considered to be much more difficult than the task of just ranking or rating objects.

2.2 Relative direct elicitation of reservation prices

The direct elicitation method is known to produce estimates of high internal validity for the part worths (Kalish and Nelson (1991)). However, there is some reservation about the direct elicitation of reservation prices primarily due to the poor predictive validity of part worth estimates as indicated by the Kalish and Nelson (1991) study. The reason for this may be due to the difficulty of the underlying valuation tasks that may cause an inherent inconsistency of the data. Thus, a modification of the evaluation tasks by using paired comparisons appears to constitute a potentially promising way to improve validity. We may therefore ask consumers which amount of money would make them indifferent between two bundles k and m: $r_{ik} - r_{im} = \Delta_{ikm}$. This necessitates the use of a difference design $D = (d_{cj})_{c=1,\ldots,C;j=1,\ldots,J}$ which - of course - has to guarantee for unbiased estimates of the part worths. The difference design is evidently dummy coded as in traditional conjoint designs, where $d_{cj} = 1$ (-1) indicates if product j is part of the right (left) bundle in the c-th comparison. The case that product j is contained in both bundles or isn't part of any bundle is dummy coded as $d_{cj} = 0$. There has been some work done on the generation of difference designs for bundles. Hausruckinger and Herker (1992) proposed a method for designing an orthogonal difference design from Addleman's 2^{15} plan that yields at least 24 valuation tasks. Aust (1995) has defined a reference profile which is part of every paired comparison. He reported pretty worse validity of part worth estimates.

Consequently, this motivates our own study that compares the just described two different methods for direct paired comparison reservation price elicitation.

[3] There has been a lively discussion concerning an adequate definition of the choice alternative. We and many others use a so-called no-purchase-option while, e.g., Kohli and Mahajan (1991) suppose that the choice alternative consists of a status quo product chosen previously by the consumer. However, the valuation against a status quo product seems to be somewhat problematic, since an alteration of the status quo price would also result in a change of the reservation price for the new product.

3 Study design and application situation

From the arguments above, the generation of difference designs has the advantage that it facilitates evaluation tasks while, simultaneously, the paired comparison situation decreases the potential threat of exhaustion.

For an application data were collected with respect to a seat system assembled for a German car manufacturer. The seat system offers several (additional) features or options such as seat heating device, memory function etc. Using previous orders we identified seven seat options which seemed to be quite important for customers as well as the firm. In order to make the study design more elaborated departing from the complete 2^7- design a fractional factorial 2^{7-4}-design X was constructed and two kinds of difference designs were derived. The first one is called reference design D_{ref} where a set of bundles constructed by means of the design X is compared with a fixed reference profile, i.e. $(0, 1, 0, 1, 1, 1, 0)$. Constructing D_{ref} in this way is due to Aust (1995, pp. 183), however, in general no orthogonal design will be obtained. The second design is called partition design D_{part} that forms a bipartition from the set of products or components by defining the two bundles in each comparison of the difference design. It is evident that the second difference design is constructed in such a way that it provides an orthogonal design matrix which allows for an uncorrelated estimation of part worths. The two design matrices given below on basis of the dummy coded difference designs mentioned before are

$$D_{ref} = \begin{pmatrix} 0 & -1 & 0 & -1 & -1 & -1 & 0 \\ 0 & -1 & 1 & -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & -1 & 0 & 0 & 0 & -1 & 1 \\ 1 & -1 & 1 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & -1 & -1 & -1 & 1 \end{pmatrix}, \quad D_{part} = \begin{pmatrix} -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & -1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & 1 \\ -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{pmatrix},$$

however, due to confidential reasons we do not resolve the coding of the designs here. It should be noted that the results obtained from the reference design depend on the reference profile which was selected in accordance with suggestions given by the car manufacturer.

Data collection was partly done in the firm's customer center where the respective cars are surrendered to their new owners and on the fringes of a firm's celebration event. Respondents were asked which type of car they had bought, how much they had paid for that car and which amount of money they would be willing to spend on additional options. After that they had to compare eight pairs of bundles that were constructed according to D_{ref} (study (1)) or D_{part} (study (2)) on a dollar-metric scale. The answers of 30 persons with respect to study 1 and 22 responses from study 2 could be used. Finally, the respondents were presented three holdout bundles that were arbitrarily chosen and identical across all respondents. Since price wasn't modeled explicitly as bundle attribute, we did without pricing the holdouts

and just asked the respondents to rank them according to their (monetary) preference for the bundles.

4 Results

As already mentioned, in advance of the actual valuation tasks in both parts of our studies we asked the respondents how much money they would spend at most on additional trim, what type of car they had bought, and which price they paid for it. Based on cars' price lists we were able to recalculate the amounts of money that were actually spent on optional equipment. A paired t-test showed a significant bias between self-explained and implicitly stated amounts across the two studies ($t(51) = 2.591, p < 0.05$). Indeed, respondents' reported average amount spent on options was about 13.5% of cars' base price while their true expenditure added up to 17.0% in average. Thus, we assume that direct absolute elicitation of reservation prices would result in biased estimates, a fact that further motivates the use of the suggested relative elicitation methods.

Some results of our study are presented in tables 1 and 2 and compared to known results from Kalish and Nelson (1991) and Aust (1995). However, since the models used show different degrees of freedom and the construction as well as the number of holdout stimuli designed for the different kinds of reservation price estimation vary the reported numbers can only serve as hints for the ranges in which internal and predictive validity values can be expected. As can be seen from table 1, internal validity indicated by means of mean R^2 and R^2_{adj} is almost equally high for our own studies and comparable to the results of Kalish and Nelson (1991). However, the adjusted R^2 are quite low for our studies which is not surprising, since the underlying models were designed to minimize the burden of paired comparisons for the respondents which resulted in smallest numbers for the degrees of freedom. We also assessed models' predictive validity by using three measures

Table 1. Internal validity measures

		scale	design	mean R^2	(R^2_{adj})	n
Kalish and Nelson (1991)	(KNRk)	Rank	-	0.91*		68
	(KNRt)	Rating	-	0.88*		97
	(KN$a)	Dollar	absolute	0.96*		89
Aust (1995)[4]	(A$a)	Dollar	absolute	-	(0.95)	158
	(A$r)	Dollar	relative	-	(0.79)	158
Own study results	(study 1)	Dollar	reference	0.94	(0.56)	22
	(study 2)	Dollar	partition	0.95	(0.57)	30

* The reported measure is Spearman's Rank Correlation.

4) Study by Aust (1995), pp. 182 – 191.

which partly allows for a comparison with the results of Kalish and Nelson (1991) or Aust (1995). First, from the ranking of the holdouts and the implicitly defined rank order determined by predicted reservation prices for the respective holdouts we calculated individual Spearman's rank correlations and corresponding means. Second, the fractions of correctly predicted first choices were calculated and, finally, we determined the relative numbers of correctly predicted total rankings. Table 2 depicts the three measures for the above mentioned studies including our own results.

The reference-design (study 1) reveals a quite worse mean rank correlation that falls even below the value for the direct elicitation of reservation prices (KN$a-design) reported by Kalish and Nelson (1991). This result could probably be attenuated by an alternative choice of the reference profile. Actually, nothing is known from the literature concerning the choice of optimal reference stimuli. Since the strategy to repeat data collection with different reference profiles and select a reference-design that shows best results is not practicable further research is needed, here. However, the partition-design (study 2) performs substantially better (with mean rank correlation of 0.56) than the reference-design. The fraction of correctly predicted first choices is also different in the two situations. The partition-design leads to correct first rank prediction in almost two of three cases, whereas the reference-design results in a 41% fraction of correctly predicted first choices that is actually just above the expectation of randomly assigned rankings. The partition-design results in a fraction of one half of correctly predicted total rankings.

In summary, the partition-design performed substantially better than the reference-design so that considerable concern about the predictive strength of the latter method remains. This is indeed what we would have expected having former arguments in mind. Thus, in conjunction with the results from the cited former studies the partition-design actually appears to be a valuable alternative in conjoint designs for estimating bundles' reservation prices.

Table 2. Predictive validity

		validity measures			
		mean rank corr.	correct first rank	cor. tot. ranking	n
Kalish and Nelson (1991)[5]	(KNRk)	0.57	62%	-	68
	(KNRt)	0.49	62%	-	97
	(KN$a)	0.43	46%	-	89
Aust (1995)[6]	(ARg)	0.50	-	-	120
	(A$a)	0.65	-	-	120
Own study results[7]	(study 1)	0.26	41%	41%	22
	(study 2)	0.56	63%	50%	30

5) 4 holdout stimuli, 6) 9 holdout stimuli, 7) 3 holdout stimuli. We refer to a study by Aust (1995), pp. 191 – 196 based on rating (dollar-metric) data as ARg (A$a).

5 Discussion

Though the use of reservation prices is prevalent in new product design and bundling, respectively, only a few references are known concerning the development of adequate estimation techniques for them. A promising research direction combines contingent valuation techniques with metric conjoint analysis. However, the absolute valuation tasks on a dollar-metric scale are known to be difficult to manage by respondents and may - as our results show - lead to downward biased self-reported willingness to pay measures. Thus, a differential analysis by means of paired comparisons seems to provide a promising way to facilitate valuation tasks. We compared two conjoint designs for paired comparison data, the so-called reference technique and the proposed partition design. The results show a high level of internal validity of both models. Based on the selected models it was accepted that the adjusted R^2 is quite low. As expected, the more elaborated partition-design provides more valid part worth estimates than its counterpart, the reference design. We suppose that this is primarily due to the construction of the pairwise evaluation tasks because the proposed D_{part} difference design allows for an uncorrelated estimation of part worth, and the evaluation tasks became more interesting to consumers as many completely different pairs of stimuli had to be compared. However, it cannot be ruled out that the results favoring the partition design could also partly be assigned to an increased efficiency compared to the reference design.[8]

There are some caveats with respect to the study that should be mentioned. First, the sample size is quite small for both designs (30/22 respondents). This is mainly due to the fact that customers receiving their cars at the customer center normally arrive just a few minutes before they attend a guided tour, so just a small period of time remains where customers can be interviewed. Second, the conjoint analysis results are based on a small number of degrees of freedom. However, this is a frequently observed problem in most conjoint studies (Wittink and Cattin (1989)). As an advantage, the valuation tasks became more tractable and interviewing time could be kept short, thus, we do not expect any bias due to symptoms of fatique.

In conclusion, we want to stress the considerable differences in predictive validity of both studies which suggests great potentials in getting better estimates via partition-designs making relative reservation price estimation more elaborate.

References

ADAMS, W.J. and YELLEN, Y.L. (1976): Comodity Bundling and the Burden of Monopoly. *Quarterly Journal of Economics, 90, 475–498.*

[8] D_{part} outperforms D_{ref} by means of minimizing the generalized variance of the parameter estimates for the pre-specified model (D-optimality).

AUST, E. (1995): *Simultane Conjoint Analyse, Benefitsegmentierung, Produktlinien- und Preisgestaltung*. Thesis, University of Karlsruhe.

AUST, E. and GAUL, W. (1995): A Unified Approach to Benefit Segmentation and Product Line Design Based on Rank Order Conjoint Data. In: W. Gaul and D. Pfeifer (Eds): *From Data to Knowledge. Studies in Classification, Data Analysis, and Knowledge Organization.*, Springer, Berlin, 289–297.

BAIER, D. and GAUL, W. (2003): Market Simulation Using a Probabilistic Ideal Vector Model for Conjoint Data. In: A. Gustafson, A. Herrmann, and F. Huber (Eds.): *Conjoint Measurement-Methods and Applications*, 3rd ed. Springer, Berlin, 123–146.

CAMERON, T.A. and JAMES, M.D. (1987): Estimating Willingness to Pay From Survey Data: An Alternative Pre-Test-Market Evaluation Procedure. *Journal of Marketing Research, 24, 389–395*.

CHUNG, J. and RAO, V.R. (2003): A General Choice Model for Bundles with Multi-Category Products: Application to Market Segmentation and Optimal Pricing for Bundles. *Journal of Marketing Research, 40, 115–130*.

DAVIS, R.K. (1963): Recreation Planning as an Economic Problem. *Natural Resources Journal, 3, 239–249*.

HANSON, W.A. and MARTIN, R.K. (1990): Optimal Bundle Pricing. *Management Science, 36, 155–174*.

HAUSRUCKINGER, G. and HERKER, A. (1992): Die Konstruktion von Schätzdesigns für Conjoint Analysen auf der Basis von Paarvergleichen. *Marketing ZFP, 14, 99–110*.

JEDIDI, K., SHARAN, J. and PUNEET, M. (2003): Measuring Heterogeneous Reservation Prices for Product Bundles. *Marketing Science, 22, 107–130*.

JEDIDI, K. and ZHANG, Z. J. (2002): Augmenting Conjoint Analysis to Estimate Consumer Reservation Price. *Management Science, 48, 1350–1368*.

KALISH, K. and NELSON, P. (1991): A Comparison of Ranking, Rating and Reservation Price Measure in Conjoint Analysis. *Marketing Letters, 2, 327–335*.

KOHLI, R. and MAHAJAN, V. (1991): A Reservation-Price Model for Optimal Pricing of Multiattribute Products in Conjoint Analysis. *Journal of Marketing Research, 28, 347–354*.

MITCHELL, R. C. and CARSON, R.T. (1989): *Using Surveys to Value Public Goods: The Contingent Valuation Method*. Resources of the Future, Washington.

REARDON, G. and PATHAK, D. S. (1990): Segmenting The Antihistamine Market: An Invetigation of Consumer Preferences. *Journal of Health Care Marketing, 10, 23–33*.

STAUSS, B. and GAUL, W. (2004): Product Bundling as a Marketing Application. In: D. Ahr, R. Fahrion, M. Oswald and G.Reinelt, (Eds.): *Operations Research Proceedings 2003*. Springer-Verlag, Berlin, 221–228.

STIGLER, G. (1963): United States vs. Loew's Inc: A Note on Block-Booking. *Supreme Court Review, 152, 152–157*.

WITTINK, D.R. and CATTIN, P. (1989): Commercial Use of Conjoint Analysis: An Update, *Journal of Marketing, 53, 91–96*.

Classification of Perceived Musical Intervals

Jobst P. Fricke

Department of Systematic Musicology,
University of Cologne,
50923 Cologne, Germany

Abstract. Tests were devised in which subjects were asked to judge the size of musical intervals in a musical context of pairs of successive intervals and chords performed by either harpsichord or violins. The judgements focused on the pitch intonation of one of the notes. Since subjects cannot base their judgements on beats since they are inaudible, results thus differ for one and the same interval depending on the musical context. Discrimination tools were applied in order to ascertain the significance of these differences. Furthermore, the fact that there is a region with a certain extent on the frequency continuum for 'in tune' intonation and that there is a region with constant interval perception (the latter can be interpreted as a phenomenon of categorical perception)–both contradict current consonant theories based on beats and roughness.

1 Background

By applying new statistical methods we can, in some cases, obtain new insights, even reconsidering data from earlier experiments.The data on which this article is based were collected 40 years ago in an experimental study of the categorical perception of musical intervals. The intervals were embedded in a musical context in order to provide them with a musical definition. This context was determined unambiguously (for listeners familiarized with Western music and with rock and pop music), just as a note which is represented by one and the same key on the keyboards is used regardless of whether its function is that of an ascending leading note (within the harmonic function of a dominant) or that of a suspension descending towards harmonic resolution.

Previous experiments which intended to base theories of consonance on listeners' judgements of consonance presented the drawback that, quite often, they only used isolated intervals produced by sine waves or by synthetic sounds. This experimental situation, quite removed from reality, was only of limited relevance. This can be atrributed to two factors:

I Rigid, fixed sounds, produced by individual pure tones or assembled from synthesized rows of harmonics, produce beats and combination tones, which, when we hear mistuned intervals composed of integral-digit frequency ratios, provoke a disturbance which is perceived by the listeners as 'roughness' (certain theories of dissonance are based to a large extent on such effects (Plomp and Levelt (1965)), which were designated earlier as

sekundäre Klangerscheinungen (Scheminzky (1935), 553), i.e. 'secondary timbre attributes').

II When an interval is played in isolation, the listener is not provided with any indication of its concrete application within a musical context. For instance, the listener is not informed whether a note within the interval occurs as a transition (in ascending voice-leading) or as a suspension; as a chromatic sharpening or as a dominant seventh. In other words, the interval is not provided with its own corresponding designation - neither in terms of chromatic alteration (sharp or flat), nor in terms of diatonic use (transition, leading note or suspension).

If an experiment tests judgements of consonance by limiting itself to exploring the phenomenon of beats, it is ignoring large areas of musical practice. This is not only due to the restricted situation which is necessarily required by any experiment in order to maintain test conditions and test variables under control and within grasp. Rather, such an approach results from a series of preconceptions which reflect an historical bias. Experiments constructed on the basis of the phenomenon of beats tend to lend an inordinate amount of importance to the musical repertoire written for reed organ (Fricke (2002), 107), for flute duets (recorder) and for organ (Helmholtz (1896), 511; Beilage XVIII p. 664). Such instruments, when playing intervals which are slightly out-of-tune (due to the fact that they result from integral-digit frequency ratios), produce combination tones which also manifest the phenomenon of 'beats' as a side-effect. These beats consist in modulation tones resulting from non-linear effects, and they can be perceived only under quite singular and limited conditions (Fricke (1996)). They have only gained importance in laboratory experiments which have used static timbres (those of tuning forks, electronically synthesized sounds and sinus tones), yet they have no relevance within the context of musical practice (Fricke (1980), 165).

When instruments play together, as in an orchestra, there are no combinations tones. Particularly in string ensembles there are no beats at all. The 'roughness' is already present due to the nonstationary acoustical phenomena produced by bowing (which hardly permit the emergence of beats). Vibrato, as well, excludes - by its own nature - any emergence of beats whatsoever (Fricke (1988), 69). For these reasons, it would be advisable to construct test situations that are closer to reality - tests in which beats would play a negligible role, or none at all, as a criterion for decision-making in interval judgements.

Experiments featuring isolated intervals also tend to create a situation which is too far removed from the reality of everyday musical practice. Nowadays, such experiments cannot be justified by evoking the necessity of restricting the number of test variables, but, rather, they reflect an absolute phenomena, which, supposedly, one should be able to test and judge 'logically' on the basis of isolated intervals (Plomp and Levelt (1965), Miskiewicz and Rogalla (2003)).

If we want to gather information about interval categories and their optimal boundaries, then we should favor experiments featuring perceptual judgements rather than those which measure performance. Experiments measuring performance (Abraham (1923), Dahlback (1958), Lottermoser and Meyer (1969), Shackford (1962), Kopiez (2003)) do indeed contain a) the desired, intended shifts in intonation as a tendencial recognizable component - yet they feature not only such shifts, but also b) the statistical fluctuations of intonations whose practical realization is not always optimally successful. Furthermore, 'reproductive' experiments, as opposed to perceptual experiments, still feature c) the tolerance in hearing which allows for a certain accepted variance in the area of intended target frequency. This phenomenon is termed *Zurechthören*, or 'hearing aright'. We must also bear in mind that d) any reproductive musical instrument also has, by its own nature, peculiar vicissitudes which can lead to certain imperfections in intonation. These four overlapping factors cannot be separated from one another when we attempt to analyze an individual case. This is why we should favor experimental situations which require the subjects to judge intervals, and not to perform them.

In the experiment on which this article is based, intervallic distances were varied and played for the subjects so that they could be judged. The first goal was to find out which is the optimal intonation in a certain specific case of performance. We were also interested in how much divergence the subjects tolerated in relation to 'optimal' intonation, and where the line was drawn between 'still barely acceptable' divergence and unacceptable divergence.
This approach was thus intended to collect information concerning interval categories and their resulting classification.

2 Experimental setting

The tests we carried out forty years ago attempted to reflect real musical situations as closely as possible, by applying the following criteria:
The acoustical material consisted in pairs of two- and three-part chords played by either harpsichord or a group of violins. The violin examples were recorded in three different versions: each part played by one soloist without vibrato, then each part played by a soloist with vibrato and, finally, by an ensemble of ca. 10 violins. The musical examples consisted in pairs of successive intervals and chords. From these, we have chosen the following for statistical analysis (Fig. 1): a suspension of a minor sixth resolves into the fifth; a dominant chord featuring an augmented fifth resolves into the tonic (Fig. 1, Ex. 2a, 2b); the augmented fourth, being an essential component of the dominant seventh chord, resolves into the sixth in the tonic; also, the corresponding diminished fifth (in the dominant seventh chord) resolves into the third in the tonic (Fig. 1, Ex. 1a, 1b). Thus, in all of these examples, the dissonances resolve into consonances following the rules of classical harmony.

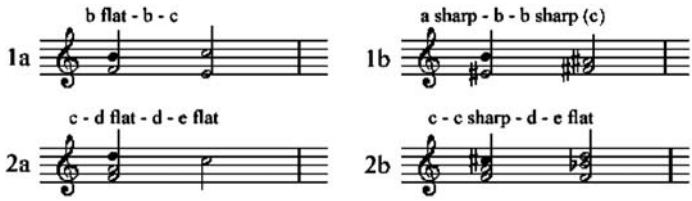


Fig. 1. Intervals and chords were heard in four different instrumental versions. The filled (black) note indicates the tone varied in pitch. Range of pitch variation: from b flat to c or from c to e flat resp..

Electronic loops of the recorded sounds were presented to experts, able to express the size of musical intervals: lecturer of music theory, professionals of instrumental teaching, leaders of choirs and orchestres, students in musicology, about 25 test persons per item. The play-back-frequency of the vibrato-or ensemble-played sounds, respectively, were measured by the oscilloscope matching method. Oscilloscope trigger frequencies and the control frequency for the period of repetition of the electronic loop were controlled by electronic counters with a maximal error of 0.1% . The maximal over all relative error was $\pm$ 0.002 corresponding to a 30th of a tempered semitone. When hearing these successions of chords in which one note was varied in its frequency, the subjects were asked to decide between:

a) optimal, perfectly in-tune intonation
b) tolerated, 'still barely acceptable' divergence from optimal intonation.

For theoretical reasons alone, it could be expected that the judgements of a) would not result in 'fixed points', but, rather, in 'stretches' or sections which, in terms of position and extension, are typical for the interval category which they represent.

With the help of the "b)" judgements, interval categories should then become visible, since the boundaries of 'acceptable divergence' represent the categorical area within which the interval is defined as such. The interval's extension within the categorical boundaries and its position on the relative frequency scale both serve as characteristic attributes for the function which it fulfills within a certain musical context. In particular, the categorical boundary which separates the 'dissonance' from its subsequent resolution into an adjacent 'consonance' will be characteristically displaced in the direction of the melodic step which the listener expects to hear next.

The tests were designed so that the note with variable intonation, which was the one to be judged, was the same one in both contexts. The question can thus be formulated as follows:

I. In the 'fifth-fourth' test (Ex. 1) contrasting the diminished fifth and the augmented fourth, is the leading note 'b', when optimally in tune, judged as being significantly different from the note 'b' in the role of seventh? In the 'fifth-sixth' test (Ex. 2) contrasting the augmented fifth and the

minor sixth, is the note 'c sharp' in the role of augmented fifth, when played optimally in tune, judged as being significantly different from the note of (enharmonically identical) 'd flat' in the role of minor sixth?

II. Do these findings imply significant alterations of the interval category as well? Is the interval category for diminished fifth ('e sharp' - 'b') located in the same position as the interval category for augmented fourth ('f' - 'b') - or are their respective extensions and relative frequency positions significantly divergent from one another? Do similar findings apply in the 'fifth-sixth' test?

III. With four different instrumental versions of the same successive pair of chords, the following question could be clarified: Is the 'c sharp' in the role of fifth, played on the harpsichord, judged as significantly different from the same note played in an ensemble of violins? Does the same apply to the 'd flat' in the role of sixth, or not?

3 Results

The first suggested statistical procedure which is appropriate to these alternatives is the decision tree. In the case of the 'fifth-fourth' test, Question No. I received a statistically significant answer, both in the violin version as well as in the harpsichord version. The corresponding tree consists only of one decision: "optimal intonation located at ≥ 610 cents". In the violin version we even attain a separation of 100% for this rule. They can also be analyzed and separated by means of linear discriminant analysis (LDA). Here, as well, the positions for optimal intonation of 'b' as seventh and 'b' as third can be separated from one another on a level of high significance. This applies to both varieties of instrumental timbre, although it turned out that the violin version attains better separation of interval categories than the harpsichord version. Analysis of variance confirms these results: p-value $< 2.2 \cdot 10^{-16}$ for the violins version, $< 3 \cdot 10^{-15}$ for the harpsichord version. The estimated Bayes error rate of LDA is about 0.2% and 3% respectively. Thus, Question No. I has been answered satisfactorily in the case of the 'fifth-fourth' test.
Certain typical tendencies were found relating to the width of the optimal intonation area and its displacement: the correlation is mildly negative, i.e. those subjects who tend to prefer sharp intonation (and, consequentially, a narrow interval between the dissonant tone and its resolution) have a relatively clear conception of the position to which their judgement if intonation as "optimal" applies. In the cases of these subjects, the area of 'optimal' intonation is very narrow (between 5 and 15 cents). This group's typical values for the resulting resolution from an optimal intonation position lie between 88 and 68 cents.

Significant results were also attained for the extension of interval category and the displacement of the categorical boundary. The decision tree is not an adequate tool for interpreting the results from violin examples, since they

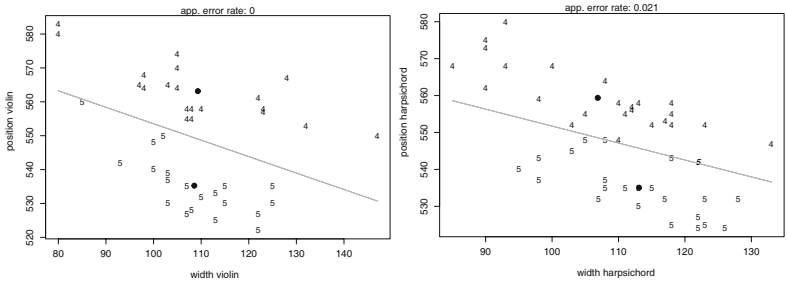


Fig. 2. LDA of judgements concerning pitch position and overall width of interval categories resulting from Examples 1 ('fifth-fourth' test) in the violin (left) and the harpsichord version (right).

are grouped diagonally (Fig. 2 (left)). Only the harpsichord examples yield a degree of differentiation which is sufficiently clear (prediction error: 6.4 corresponding to 10.6% for a decision in favor of the fourth, if position ≥ 550 cents), as one can see directly in the judgement results shown in Fig. 2 (right). However, using LDA we can separate, on a level of high significance, the scatterplots resulting both from the violin examples and from the harpsichord examples (Fig. 2). Thus, the questions in Point II have also been answered satisfactorily. Furthermore, the degree of inclination of the separation lines indicate the following correlation: the narrower the interval category, the more its boundary tends to be displaced toward the resolving note's pitch. This applies more strongly to the violin examples than to the harpsichord examples. Typical values for resulting resolution intervals lie only 30 to 15 cents beyond the categorical boundary.

Due to their strong internal correlations, the results from the 'fifth-sixth' test can only be separated adequately by using LDA. High significance is obtained for the result from the violin examples featured in Fig. 3 (left) when they are analyzed separately: apparent error rate 0.0%. In this case, analysis of variance also yields significant results (estimated Bayes error rate of classification about 3%).

However, the categories resulting from the harpsichord examples do not differentiate the intonation of the fifth from the sixth interval as clearly: apparent error rate 14%. The p-value from the harpsichord result t-test of LDA scores is $< 3.5 \cdot 10^{-8}$ as compared to the overall result from violin and harpsichord, which are of $2.8 \cdot 10^{-13}$. The decision tree is not applicable as an analysis tool in this case either, since the scatterplots are grouped diagonally along the two-dimensional diagram.

The results for Question No. III were tested on the optimal intonations of the 'fifth-sixth' test. The decision tree only yields positive results for the interval of minor sixth. LDA analysis points in the same direction. The explanation for this problem can be found by analyzing the subjects' individual tendency of judgement, visible in Fig. 3 (right), and which can be interpreted as follows: Those subjects who, in one case, prefer a large displacement of

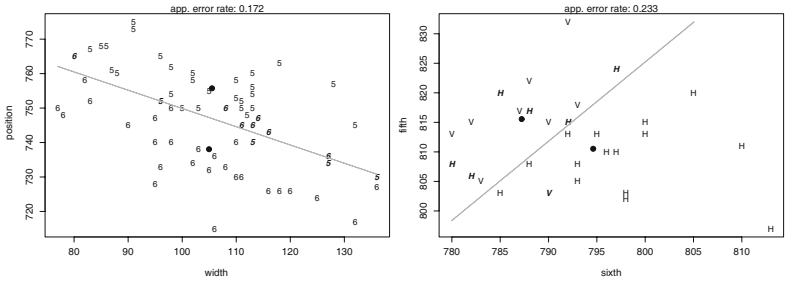


Fig. 3. Left: LDA of judgements concerning pitch position and overall width of interval categories resulting from Examples 2 ('fifth-sixth' test) in all versions (harpsichord and violins). Right: Comparison of judgements of violin versions (V) as opposed to harpsichord version (H) for Examples 2 ('fifth-sixth' test) computed by means of LDA.

intonation based on equal temperament, also do so in the other case. In other words, the larger the augmented fifth, the smaller the minor sixth tends to be. This context-dependent influence is stronger in violin examples than in harpsichord examples (as we have already often seen in the cases cited above). Here, the hypothesis that there is a significant difference based on instrumental timbre can only be confirmed in the case of the minor sixth; the effect of musical (harmonic) context is differentiated according to which instrumental timbre is chosen for the minor sixth. On the other hand, however, the values for the augmented fifth show no difference between violin and harpsichord examples - in all cases, intonation is strongly preferred in a displacement towards chromatic sharpening, leaning toward the upper note of resolution. This result would need to be interpreted as follows: the effect of the double leading note produced by chromatic sharpening of the fifth is so strong that the listener demands exaggerated chromatic sharpening of intonation in all of the examples. The position for exact, 'in-tune' intonation is, on the average, 15 cents higher than the equal-temperament value - in harpsichord examples as well. When compared to the harmonically 'pure' value, the demanded chromatic sharpening even attains an average of 21 cents.

4 Conclusion

Applying the decision tree and the method of LDA, it has become possible here to provide statistical evidence for hitherto suspected differences of intonation which are solely dependent on musical context. Differences of intonation (levelled off in equal temperament) occur not only in the case of enharmonic changes, but also when the same note plays different harmonic roles. Still, the most important result is that it is not systems of (Pythagorean or harmonically pure) temperament that determine the practice of intonation, but, rather, psychological factors as a result of musical context. 'Striving' tendencies of intonation realized in half-note steps which are 'too' small,

exaggerations which expand the distance in augmented intervals (the tritone, the fifth) and which contract the size of diminished fifths, etc., are the norm – and such tendencies were already outlined three quarters of a century ago (Abraham (1923)). The degree of non-stationary vibrations present in musically produced sounds (performed here by a violin ensemble, by individual violins with or without vibrato and on a harpsichord) determines to what extent these psychological tendencies prevail.

Acknowledgement
I am indebted to Karsten Lübke and Claus Weihs for their statistical work.

References

ABRAHAM, O. (1923): Tonometrische Untersuchungen an einem deutschen Volkslied. *: Psychologische Forschung, 4, 1–22.*

DAHLBACK, K. (1958): *New Methods in Vocal Folk Music Research.* Oslo University Press, Oslo.

FRICKE, J.P. (1980): Hindemiths theoretische Grundlegung der Kompositionstechnik in seiner "Unterweisung im Tonsatz". In: D. Altenburg (Ed.): *Ars Musica - Musica Scientia.* Gitarre & Laute Verlag, Köln, 159–170.

FRICKE, J.P. (1988): Klangbreite und Tonempfindung. Bedingugungen kategorialer Wahrnehmung aufgrund experimenteller Untersuchung der Intonation. In: K.-E. Behne, G. Kleinen and H. de la Motte-Haber (Eds): *Musikpsychologie. Jahrbuch der deutschen Gesellschaft für Musikpsychologie*, Vol. 5. Noetzel, Wolhelmshaven, 67–87.

FRICKE, J.P. (1996): Kombinationstöne. In: L. Finscher (Ed.): *Musik in Geschichte und Gegenwart*, 2nd ed., Vol. 5. Kassel et al., Verlage Bärenreiter, Metzler, 482–486.

FRICKE, J.P. (2002): Pitch Bending und das Harmonium als Reininstrument. In: M. Lustig (Ed.): *Harmonium und Handharmonika. 20. Musikinstrumentenbau-Symposium 1999.* Stiftung Kloster Michaelstein, Blankenburg, 105–116.

HELMHOLTZ, H. v. (1896): *Die Lehre von den Tonempfindungen*, 5th ed. Vieweg, Braunschweig.

KOPIEZ, R. (2003): Intonation of Harmonic Intervals: Adaptability of Expert Musicians to Equal Temperament and Just Intonation. *Music Perception, 20, 383–410.*

LOTTERMOSER, W. and MEYER, Fr.-J. (1960): Frequenzmessungen an gesungenen Akkorden. *Acustica, 10, 181–184.*

MISKIEWICZ, A. and ROGALLA, T. (2003): Roughness and Dissonance of Musical Dyads. In: R. Kopiez et al. (Eds.): *ESCOM 5 CDROM, Code 108.*

PLOMP, R. and LEVELT, W.J.M. (1965): Tonal Consonance and Critical Bandwidth. *Journal Acoustical Society of America, 38, 548–560.*

SCHEMINZKY, F. (1935): *Die Welt des Schalles.* Das Berglandbuch, Graz/Wien/Leipzig/Berlin.

SHACKFORD, Ch. (1962): Some Aspects of Perception. II. Interval Sizes and Tonal Dynamics in Performance, III. Addenda. *Journal of Music Theory 6, 66–90, 295–303.*

In Search of Variables Distinguishing Low and High Achievers in Music Sight Reading Task

Reinhard Kopiez[1], Claus Weihs[2], Uwe Ligges[2], and Ji In Lee[1]

[1] Hanover University of Music and Drama
Institute for Research in Music Education
30175 Hanover, Germany
[2] University of Dortmund*
Department of Statistics
44221 Dortmund, Germany

Abstract. The unrehearsed performance of music, called 'sight reading' (SR), is a basic skill for all musicians. Despite the merits of expertise theory, there is no comprehensive model which can classify subjects into high and low performance groups. This study is the first that classifies subjects and is based on an extensive experiment measuring the total SR performance of 52 piano students. Classification methods (cluster analysis, classification tree, linear discriminant analysis) were applied. Results of a linear discriminant analysis revealed a 2-class solution with 4 predictors (predictive error: 15%).

1 Background

Sight reading (SR) is a functional skill which is required by all musicians. It is not only of particular interest for musical occupations such as the piano accompanist, the conductor, or the repetiteur, but is also one of the five basic performance skills every musician should acquire. Obtained from path analysis, McPherson (1993) defines these skills as follows: to perform a repertoire of rehearsed music, to perform music from memory (where music was memorised using notation and then recreated aurally), to play by ear (where music was both learned and reproduced aurally), to improvise in both 'stylistically conceived' and 'freely conceived' idioms, to sight read music without prior rehearsal. This skill is characterized by high demands on the performer's capacity to process highly complex visual input (the score) under the constraints of real-time and without the opportunity for error correction. However, up until now, there has been no feasible theory of SR which considers all relevant factors such as practice-related variables (e.g. expertise), speed of information processing (e.g. mental speed), or psycho-motor speed (e.g. speed or repeated finger movements such as trills). The differences between individuals in sight reading achievement have not yet been fully explained. From

* The work of Claus Weihs and Uwe Ligges has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

previous studies we already know that there are a number of skills which are relevant for the explanation of differences in sight reading performance. A study by Kornicke (1992; 1995), based on 73 piano students, revealed the following influential variables: (1) aural imagery, (2) sight-reading experience (quantity, frequency, and range of sight-reading), (3) cognitive style of field dependence/field independence (important for males), (4) style of thinking measured by the Myers-Briggs Type Indicator, (5) external locus of control (important for males). Another influential study by Lehmann and Ericsson (1996), based on expertise theory, measured the performance of 16 expert pianists and revealed the following variables as best predictors in a multiple regression analysis: (1) accumulated amount of time spent on accompanying-related activities, (2) size of accompanying repertoire.

Against the background of an adjusted R^2-value of 0.65 obtained from a previous multiple regression analysis (see Lee (in press)), our study tries a different methodological approach and is guided by the following hypothesis: due to the limited number of subjects ($n = 52$), a method of data analysis which searches for clusters, representing a sufficient number of cases and separable with an acceptable predictive error, is more valid than analytical methods which search for sophisticated linear or non-linear relationships. To achieve our aim, we applied classification methods for the first time in sight reading research to uncover those variables or variable combinations which best contribute to the classification of SR performance classes. This study is only a first approach to the application of classification methods to sight reading performance. A more detailed publication is in preparation (see Kopiez et al. (in preparation)).

2 Method

Subjects

52 piano students (28 females, 24 males) from the Hanover University of Music and Drama served as subjects (mean age = 24.56, standard deviation = 4.9). These students had to have piano as a major subject or had to be experts in chamber music or accompanying.

Material

Sight reading task

For the sight reading task, the paradigm of a pre-recorded pacing melody was used (Lehmann et al. (1993)). Stimulus consisted of 2 warm-up pieces and 5 pieces with increasing complexity. This method created time constraints which forced the subjects to play in tempo. These were taken from existing piano sight reading literature (UNISA (no date)), and a composer rewrote these pieces for a solo melody and piano accompaniment. The pre-recorded

solo melody was played strictly in time by a violinist, and tempo indications were given by clicks before each piece, which were also pre-recorded.

Measurement of predictor variables

Selected predictor variables were derived from sight reading literature and divided into 3 groups: (a) general cognitive skills (such as short term and working memory), (b) elementary cognitive skills (such as simple reaction time and speed of information processing) and (c) practice-related skills (such as general piano expertise, inner hearing ability and accumulated hours of sight reading expertise). In total, there were 27 single predictors considered (Table 1; for a detailed description of the measurement of variables see Lee (in press)).

Table 1. List of 27 independent variables used for classification of sight reading performance.

Variable name	Variable label
ACHSRE10	Accumulated hours of SR expertise up to age 10
ACHSRE15	Accumulated hours of SR expertise up to age 15
ACHSRE18	Accumulated hours of SR expertise up to age 18
ACHSRETT	Accumulated hours of SR expertise total
ACHPSO10	Accumulated hours of solo practice up to age 10
ACHPSO15	Accumulated hours of solo practice up to age 15
ACHPSO18	Accumulated hours of solo practice up to age 18
ACHPSOTT	Accumulated hours of solo practice total
ACYPLE10	Accumulated hours of piano lessons up to age 10
ACYPLE15	Accumulated hours of piano lessons up to age 15
ACYPLE18	Accumulated hours of piano lessons up to age 18
ACYPLETT	Accumulated hours of piano lessons total
IH.DPRIM	Inner hearing score (d')
STMSMT	Short term music-specific memory (no. of notes)
NUMCONTS	Number connection Test (s)
RAVENMDS	Raven D matrices (no. of correct items)
TOTTRAVE	Total time for Raven's D matrix (s)
STM.PER	Short term memory (mean % of correct items)
WM.PERC	Working memory (mean % of correct items)
PICRT.ME	Reaction time picture (median in ms)
SNDRT.ME	Reaction time sound (median in ms)
ITI.LRHZ	Inter tap interval for both hands (median in Hz)
TR131HZ	Trill speed over 15 s, f.c.[1] 1-3, 1. trial (median in Hz)
TR132HZ	Trill speed over 15 s, f.c.[1] 1-3, 2. trial (median in Hz)
TR341HZ	Trill speed over 15 s, f.c.[1] 3-4, 1. trial (median in Hz)
TR342HZ	Trill speed over 15 s, f.c.[1] 3-4, 2. trial (median in Hz)
CCUM	Tapping lateralization coefficient (< 1.7 = non right-handed)

[1] f.c. = finger combination; all trills were played with the right hand.

Procedure

Subjects were required to accompany the pre-recorded violin part on a MIDI piano. Accompaniment was recorded onto a PC using the sequencer Software 'Cubase'. Retrospective interviews and measurement of predictor variables were carried out after the sight reading tasks (for a detailed description of the entire procedure and devices see Lee (in press)). The entire procedure lasted about 3 hours.

Scoring for the sight reading performances (target variable) was done using a researcher-developed computer program called 'MidiCompare' (Dixon (2002)). This program matches the pitches of a subject's recorded sight reading performance with the score. The output shows the number of matches within an adjustable critical time frame of ± 0.25 s. For this analysis, the total performance score of each subject for both hands, as a percentage, was used.

3 Results

The main aim of the classification analysis was to find variables which can classify cases with respect to the target variable 'total sight reading performance' with a minimum predictive error. Calculation of statistical analyses was done using the open source software 'R' (R Development Core Team (2004)). The analysis is work in progress and in this paper we will only show a classification into 2 classes. A more in-depth analysis is in preparation (Kopiez et al. (in preparation)).

The 2-class solution

Cluster analysis and 2-class LDA

Analysis commenced with a separation of subjects into 2 classes by means of a cluster analysis (method: k-means). All 27 predictor variables were included and the total sight reading performance was used as the target variable. Group boundaries were determined by the mean of the two cluster centre values, resulting in two groups (0–66%, 66–100% performance). This separation of performance data into ranges of the lower two-thirds and the upper one-third, with group sizes of 33 and 19 subjects, is reasonable. Cases were classified by stepwise linear discriminant analysis (LDA; method: stepwise with 4-fold cross-validation, direction: both, stop criterion: error improvement $< 5\%$).

Four separating variables (CCUM, NUMCONTS, TR342HZ, SNDRT.ME) were revealed as classifying variables. Classification was successful with a total predictive error of 0.15 (4-fold cross-validated). Figure 1 shows a differentiated picture of the apparent error for each combination of the selected separating variables. Class boundaries are indicated by the grey classification

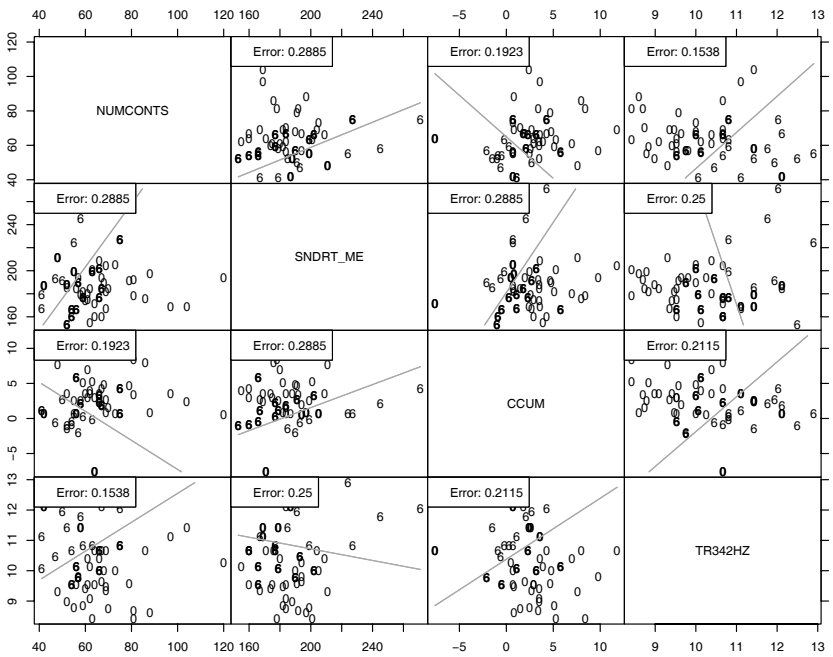


Fig. 1. Error matrix scatterplot for 2-classes LDA (0, 66, 100) with total sight reading performance as target variable. Total predictive error (4-fold cross-validated): 0.15. Grey line indicates class boundaries. Case-allocation to classes is indicated by symbol '0' for 0–66% and '6' for 66–100%. Bold symbols indicate false classification of cases to the respective class. The apparent error for each combination of separating variables is indicated in the upper left corner of each box.

line. Apparent error ranges from 0.154 (variable combination NUMCONTS-TR342HZ) to 0.288 (variable combinations NUMCONTS-SNDRT.ME and CCUM-SNDRT.ME). Despite the acceptable total apparent error of 0.15, we can see that particular variable combinations differ in error rate: on the one hand, the combination of an elementary cognitive skill, such as simple reaction time, in an auditory task (SNDRT.ME) with a psychomotor skill component (speed trill TR342HZ) reveals that a subject with a slower trill speed (< 11 Hz) and a shorter reaction time (< 200 ms) can be classified to the upper third performance class (66–100%) with an apparent error of 0.25. On the other hand, a combination of right-handedness (CCUM > 1.7) and a relatively slow mental speed (NUMCONTS > 60 s) also classifies subjects to the upper third performance class.

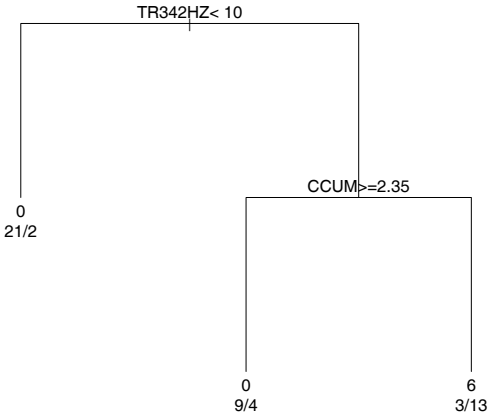


Fig. 2. Classification tree for the 2-class solution (0, 66, 100).

2-class classification tree

A clearer, but less differentiated picture of classifying variables is given by the 2-class classification tree (Figure 4). All 27 independent variables were included in the 2-class tree analysis leading to a tree with a relatively high predictive error of 0.44 (10-fold cross-validated). The left side of the first branch allocates those subjects to the lower performance group (0–66%) who show a slow trill over 15 seconds with the 3rd and 4th finger. Those subjects who could trill faster than 10 Hz were sorted into the right branch, and in the next step they were sorted into the high performance group (66–100%), in the case of a tendency to non-right-handedness (the CCUM value should be smaller than 1.7). 13 out of 16 subjects could be classified to the high performer class by these two criteria.

4 Discussion

Our motivation to look for an alternative method of performance prediction, by using classification procedures, was the existence of an unexplained variance of 35% in the multiple regression analysis. In this study we could demonstrate that classification of sight reading performance is a useful method of data analysis and results in an acceptable predictive error. The 2-class classification tree emphasizes the subject's psychomotor speed and handedness. The 2-class LDA also emphasizes speed-related factors such as simple reaction time, trill speed and cognitive speed as measured by the number connection test. The first surprise was that at the 2-class level, solutions did not show evidence of expertise-related factors as useful classifiers. As a second surprise, handedness (measured by the lateralization coefficient) was considered. Thus,

in a first rough approach we might conclude that 'speed matters'. However, future analyses will reveal how predictors are intertwined and where sight reading expertise unfolds its influence. Another question to answer in future analyses is whether below average performance in one predictor can be compensated for by above average performance in another predictor. This will be a completely new approach to a new insight into the structure of the fascinating skill of sight reading.

References

DIXON, S. (2002): *Midicompare [computersoftware]*. Austrian Institute for Artificial Intelligence, Vienna.

KOPIEZ, R., WEIHS, C., LIGGES, U. and LEE, J. I. (in preparation): Classification of low and high performers in a musical sight reading task.

KORNICKE, L. E. (1992): *An exploratory study of individual difference variables in piano sight-reading achievement* (Doctoral dissertation, Indiana University, 1992). DAI-A 53/12, p. 4125, Jun 1993.

KORNICKE, L. E. (1995): An exploratory study of individual difference variables in piano sight-reading achievement. *Quarterly Journal of Music Teaching and Learning, 6(1), 56–79.*

LEE, J. I. (in press): *Component skills involved in sight reading music*. Peter Lang, Frankfurt a. M.

LEHMANN, A. C. and ERICSSON, K. A. (1993): Sight-reading ability of expert pianists in the context of piano accompanying. *Psychomusicology, 12(2), 182–195.*

LEHMANN, A. C. and ERICSSON, K. A. (1996): Performance without preparation: Structure and acquisition of expert sight-reading and accompanying performance. *Psycho-musicology, 15(1-2), 1–29.*

MCPHERSON, G. E. (1993): *Factors and abilities influencing the development of visual, aural and creative performance skills in music and their educational implications* (Doctoral dissertation, University of Sydney - Australia, 1993). DAI-A 54/04, p. 1277, Oct 1993.

R DEVELOPMENT CORE TEAM (2004): *R: A language and environment for statistical computing*. R Foundation for Statistical computing, Vienna, `http://CRAN.R-project.org`.

UNISA (no date): *Playing at sight (piano)* (1-8). University of South Africa, Pretoria.

Automatic Feature Extraction from Large Time Series

Ingo Mierswa

Univ. Dortmund, Computer Science VIII, Germany
mierswa@ls8.cs.uni-dortmund.de

Abstract. The classification of high dimensional data like time series requires the efficient extraction of meaningful features. The systematization of statistical methods allows automatic approaches to combine these methods and construct a method tree which delivers suitable features. It can be shown that the combination of efficient methods also works efficiently, which is especially necessary for the feature extraction from large value series. The transformation from raw series data to feature vectors is illustrated by different classification tasks in the domain of audio data.

1 Introduction

Each instance for a numerical learning algorithm is described by the values of a given set of features. The learning scheme should find a hypothesis which allows the classification of unseen data (Mitchell (1996) and Witten and Frank (2000)). Transforming the given representation may ease learning such that a simple learning algorithm can solve the problem and provide better results (Morik (2000) and Pyle (1999)).

Music is a real-valued function of time. Therefore, audio data can be seen as an univariate value series. The amplitude a_i for each time point i (*sample point*) is given. A three-minute mono song consists of $44100Hz \cdot 180s \approx 8 \cdot 10^6$ values. The classification of these high dimensional series requires the extraction of features, so that a classification scheme can make use of the feature vectors instead of the large series data. By extracting the small feature vectors, both the improvement of results (Liu and Motoda (1998) and Ritthoff et al. (2002)) and a strong data compression is expected.

We have to face up with two problems: first, the great amount of data requires efficiently working methods to extract the features and second, it is not always clear which is the meaning of the extracted features. The automatic selection and combination of the best methods for feature extraction would be very useful.

The next section introduces a systematization of statistical methods, which allows automatic feature extraction from value series data. In section 3, an automatic approach for feature extraction based on genetic programming is presented and the runtime is analyzed. In section 4 the feature extraction from audio data is described and the results are discussed.

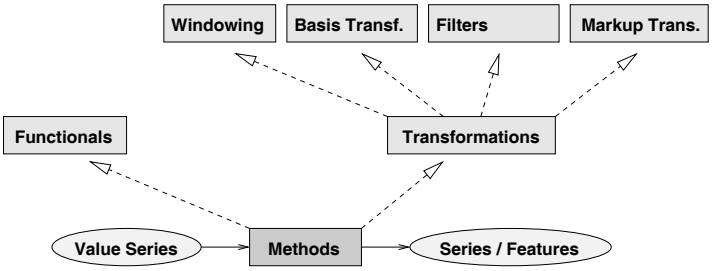


Fig. 1. All methods for value series analysis can be divided into groups according to their output. Transformations can be further divided in basis transformations, filters, mark-up transformations, and windowing.

2 Systematization of statistical methods

A systematization of statistical methods must be powerful enough to cover all known and future methods and it must be precise enough to allow automatic approaches to select and combine methods to find the optimal set of extracted features. A *method* is defined in an operator based way: it gets a value series as input and applies an arbitrary operation in order to deliver a result. This result turns out to be a good criterion to divide the methods into groups. We distinguish between:

Transformations: All methods which deliver a value series as output, i. e. a mapping $t : F \to F$ for a function space F of value series $(x_i)_{i\in\{1,\ldots,n\}}$.

Functions: All methods which deliver single values without any order, i. e. a mapping $f : F \to \mathbb{R}^m$ between a function space F and real numbers.

Transformations, which change the series itself without generating features, can be divided into several groups like *basis transformations* (e. g. Fourier transformation or state space reconstruction), *filters* (e. g. window functions or difference filter), and *mark-up transformations* (e. g. finding intervals in the series). Chains built from an arbitrary number of transformations and ending with a function deliver the desired features. Figure 1 shows the systematization.

2.1 Windowing extends the method space

In order to divide the existing methods for value series analysis (Bradley (1999) and Schlittgen and Streitberg (1997)) into the specified groups, a particular transformation requires a special treatment. With the aid of a *Windowing* operator a bunch of further transformations can be simulated and created:

Windowing: Given a value series $(x_i)_{i\in\{1,\ldots,n\}}$ with length n. A transformation is called *Windowing* if a window of size w is moved with step

size s over the series and in each window the value of a function f is calculated:

$$y_j = f((x_i)_{i \in \{(j \cdot s+1,\ldots,j \cdot s+w\}}).$$

The values y_j form a new series $(y_j)_{j \in \{0,\ldots,\lfloor (n-w)/s \rfloor\}}$.

If f is an average function, this definition of a windowing includes the well known moving average filters. But we take a step forward and allow all functions for windowing and additionally allow any number of transformations before we calculate the value of the function[1]. For large value series we must ensure that a windowing which uses efficient methods to calculate the values y_j also is an efficient method, i.e. has polynomial runtime. The overlap of a windowing is defined as $g = \frac{w}{s}$. Each windowing creates $\frac{n-w}{s} + 1 = \frac{n}{s} - g + 1$ windows. A windowing performing transformations and a function with runtime $O(n^2)$ on each window has an overall runtime of

$$\left(\frac{n}{s} - g + 1\right) \cdot w^2 = gnw - gw^2 + w^2.$$

To estimate the worst case we consider a windowing with step size $s = 1$. For a realistic overlap of $g = 2$ the runtime is $2nw - w^2$ which is smaller than n^2 for all window sizes $w < n$. The maximum amount of multiple used values is reached for a window size of $w = \frac{n}{2}$ and therefore an overlap of also $g = \frac{n}{2}$. For this worst case the runtime is

$$gnw - gw^2 + w^2 = n \cdot \left(\frac{n}{2}\right)^2 - \left(\frac{n}{2}\right)^3 + \left(\frac{n}{2}\right)^2 = \frac{n^3}{8} + \frac{n^2}{4}.$$

The runtime has a greater power in n but is still efficient. Similar calculations for the runtimes of other methods show that the usage of windowing with a realistic overlap like $g = 2$ always result in a smaller runtime than the application of the methods on the complete series.

2.2 Method trees for feature extraction

As we have mentioned, the extracted features are the result of a chain of transformations and a function at its end. A windowing is also a transformation. But this particular transformation performs other methods on windows to form a new series. One can see these methods as children of a windowing, which leads to the model of *method trees* for feature extraction.

Figure 2 shows an example for a method tree. The tree is traversed with a depth first search. The windowing is the root of a new tree, whose children are invoked once for each window. The dashed lines show the parent-child connection in the tree and the solid lines stand for the data flow. The last child in the chain is an average function which delivers the features "average and variance of the maximum frequency in the progression of time".

[1] Actually we can do a windowing without a function but with transformations only. This should only be done for windowings with $w = s$ and is called *piecewise filtering*.

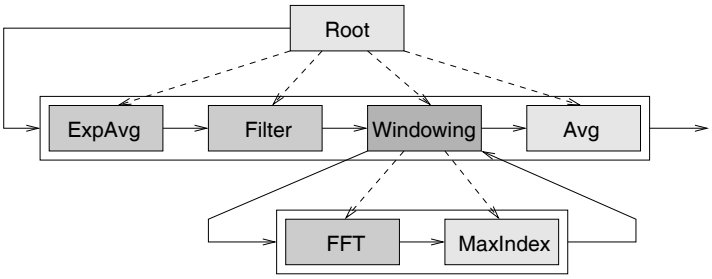


Fig. 2. A method tree that extracts the feature "average and variance of the maximum frequency in the progression of time". The windowing method is the root of a new tree.

2.3 Dynamic windowing in method trees

Each method tree provides one or several features. The tree structure emerges from the nesting of windowing methods. It is quite clear that it is impossible to nest two windowings with the same window size w. Since the children of the parent windowing work on windows of size w, a nested windowing with window size w creates a series with length 1 which is actually not a series anymore. Therefore the overlap of each windowing method should be fixed, which makes sure that the windowing works efficiently for realistic overlaps. The size of the windows must be *dynamic* and is defined as $w = \frac{n}{d}$ for a parameter $d \in \{2, \ldots, \frac{n}{2}\}$.

Dynamic windowing: Given a series $(x_i)_{i\in\{1,\ldots,n\}}$ with length n and a parameter $d \in \{2, \ldots, \frac{n}{2}\}$. A windowing with overlap g, window size $w = \frac{n}{d}$ and step size $s = \frac{n}{gd}$ is called *dynamic windowing*.

A method tree built of dynamic windowings and other transformations and functions has a maximum depth of $\log_d n - 1$. A dynamic windowing divides the series in windows with size $\frac{n}{d}$. Therefore, each dynamic windowing which is a child of another windowing works only on $\frac{n}{d}$ values and builds windows with size $\frac{n}{d^2}$. After $\log_d n - 1$ nested windowings each window has

$$\frac{n}{d^{\log_d n - 1}} = \frac{n}{\frac{d^{\log_d n}}{d}} = \frac{n \cdot d}{n} = d$$

values. Another windowing would reduce the number of values for the next child to 1.

Now we are able to analyze the runtime of method trees on a value series with length n. The worst case runtime of all transformations and functions is given as $L(k)$ on k values. The runtime of a dynamic windowing is

$$\left(\frac{n}{s} - g + 1\right) \cdot L\left(\frac{n}{d}\right) = (g(d-1)+1) \cdot L\left(\frac{n}{d}\right).$$

We add another dynamic windowing as a child and replace $L(\frac{n}{d})$ by $(g(d-1)+1) \cdot L(\frac{n}{d^2})$ which leads to $(g(d-1)+1)^2 \cdot L(\frac{n}{d^2})$. We iterate these steps which delivers

$$(g(d-1)+1)^i \cdot L\left(\frac{n}{d^i}\right)$$

as runtime of a method tree with depth i. We have shown that each method tree has a maximum depth of $\log_d n - 1$ which results in a runtime of

$$(g(d-1)+1)^{\log_d n-1} \cdot L\left(\frac{n}{d^{\log_d n-1}}\right) = (g(d-1)+1)^{\log_d n-1} \cdot L(d).$$

A method tree based on methods with a worst case runtime of $O(n^2)$ therefore has an overall runtime of

$$(g(d-1)+1)^{\log_d n-1} \cdot d^2 = \frac{d^2}{g(d-1)+1} \cdot n^{\frac{1}{\log_{g(d-1)+1} d}}.$$

It has been shown that the runtime is never exponential, for realistic dynamic windowings with overlap $g = 2$ and $d = 2$ the runtime is $\frac{4}{3} \cdot n^{1.585}$ which is always smaller than n^2. Hence, method trees built from efficient methods are efficient too.

3 Automatic feature extraction

We have fulfilled two premises for automatic approaches for feature extraction: the search space is structured and the elements of this space work efficiently and extract features from high dimensional data in polynomial time. Now we introduce a simple way for automatic feature extraction based on *genetic programming*. The search space in which the algorithm tries to find the optimum is the space of all method trees which can be created with the given transformations and functions. Each individual is a method tree and the tree which provides the best features for the classification task at hand is delivered as the result.

The first step is to create a population consisting of a number of individuals. Here, randomly created method trees are used as individuals. Then the same steps are performed repeatedly until a termination criterion is satisfied. *Mutations* are operations which derive a new individual from one other individual. The probability for small distances between parent and child should be greater than for great distances. We use *generating mutation*, which randomly creates a new method and adds it at an adequate place in the method tree, *removing mutation*, which removes a randomly chosen method from the method tree (windowing with overlap $g > 1$ must contain a function), and *changing mutation*, which changes a randomly chosen method and replaces it with a method from the same group (transformation or function). Another typical operation for evolutionary algorithms is *crossover*, where a new individual is derived by combining the informations about several parents. Crossover is realized by transfering subtrees of the same type.

In each generation the individuals are evaluated with a k-fold cross validation with respect to the learning task at hand. First, the method tree which should be evaluated is used to extract the features from the data. Then the performance of the learning task is estimated with an inner k-fold cross validation. The transformed data is divided into k parts, on $k-1$ parts a classifier is trained and on the last part it is applied. Individuals with a greater performance (fitness) will have a higher probability to survive.

Another fact is interesting for working with high-dimensional data: the building of method trees with genetic programming in order to extract an optimal feature set is like training the optimal feature extraction. We have *two* phases of training. The first phase is the training of a method tree for feature extraction which can be done on a subset of the data. This is especially useful for the great amount of data the high dimensionality brings. Then the best method tree is applied on the complete data and the second training phase starts: a hypothesis is learned from the feature vectors created by the method tree.

4 Experiments

We used the discussed approach to extract features from audio data for three different classification tasks:

1. genre classification *classic* and *popular music*: CLA/POP
2. genre classification *techno* and *popular music*: TEC/POP
3. classification of user preference: USER_1, USER_2, and USER_3

The first one is considered an easy task, the other two problems seem to be much harder. CLA/POP contains 100 instances, TEC/POP contains 80 instances, and the USER data sets 50 instances for each class. The methods are implemented within a generic framework for value series preprocessing like the one demanded in Morik and Liedtke (2000). The experiments were done with the learning environment YALE[2] (Fischer et al. (2002) and Mierswa et al. (2003)). Feature extraction for a 60 second sample of music lasts approximately 20 seconds using a 1600 MHz CPU.

The following features were extracted from the data sets: average loudness, average distance and variance between extreme values, average distance and variance between zero crossings, tempo and variance of autocorrelation, k highest peaks after a Fourier transformation, gradient of a linear regression function of the frequency spectrum, fraction of geometric and arithmetic average of the spectrum, fraction of maximum and arithmetic average of the spectrum, average and variance of the strongest frequency in the progressing of time, average and variance of the angles after a state space reconstruction, and average and variance of the distances after a state space reconstruction.

[2] `http://yale.cs.uni-dortmund.de`

Table 1. The classification error for the different classification tasks. For each task, an optimal subset of features was selected.

	CLA/POP	**TEC/POP**	**USER$_1$**	**USER$_2$**	**USER$_3$**
C4.5	1.67%	12.13%	8.12%	9.89%	5.02%
SVM	1.82%	13.22%	7.69%	9.44%	4.81%

They are described in detail in Mierswa (2003) and were collected during several runs of the genetic programming approach for the classification of audio data. The population size was 10, each mutation probability was 0.2, and crossover probability was 0.4. The maximum number of generations was 100. With a genetic algorithm (Ritthoff et al. (2002)), a subset of these features were selected for each classification task.

The application of a simple 1-R-Learner (Holte (1993)) delivers different features for the data sets. For the classification of CLA/POP, the variance of the distances after a state space reconstruction is the best feature. The created rule correctly classifies 184 of the 200 instances. In the domain TEC/POP the variance of the difference between the extreme values of the series was selected as best feature. The knowledge of this feature alone allows the correct classification of 121 of the 160 instances. Further results of the selection among the audio features are discussed in Mierswa (2003).

4.1 Results

The learning schemes used were the decision tree learner C4.5 (Quinlan (1993)) and a support vector machine (SVM) with a linear kernel function (Joachims (1999) and Rueping (2000)). The classification error was estimated with a 10-fold cross validation. The confidence for decision tree inducing was 0.25 with a minimum leaf size of 2. The support vector machine MySVM was used with default parameters. Table 1 shows the results for the classification tasks. The genre classification CLA/POP can be done with an error of 1.67%. The more difficult classification of user preferences can be handled with classification errors between 4% and 10%. The genre classification TEC/POP is the hardest discipline among these classification tasks. But the predictions were done with an error of 12.13%.

5 Conclusion

The methods of value series analysis can be divided into groups and systematized. Together with an extended concept of windowing operators these methods can build method trees for feature extraction. It has been shown that the windowing of efficient methods also is efficient, the same applies for method trees. The systematization and the efficiency of the methods allow automatic approaches to extract an optimal set of features from value series.

The discussed approach is based on genetic programming. The individuals are method trees which work on a subset of the data and are mutated and recombined. The result is a method tree which can be used on the complete data set for feature extraction.

Audio data can be seen as time series with an extraordinary length. We have seen a set of features which was automatically extracted from audio data. This leads to an error of nearly 1% for the genre classification classic/popular music and an user preference classification error below 10%.

References

BRADLEY, E. (1999): Intelligent Data Analysis: An Introduction. In: M. Berthold and D. Hand (Eds.): *Intelligent Data Analysis: An Introduction*. Springer, Berlin.

FISCHER, S. and KLINKENBERG, R. and MIERSWA, I. and RITTHOFF, O. (2002): YALE: Yet Another Learning Environment. *Technical report CI-136/02, University of Dortmund*

HOLTE, R. C. (1993): Very simple classification rules perform well on most commonly used datasets. *Journal of Machine Learning, 11, 63–90.*

JOACHIMS, T. (1999): Making large-Scale SVM Learning Practical. In: *Advances in Kernel Methods - Support Vector Learning*. MIT Press.

LIU, H. and MOTODA, H. (1998): *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. Kluwer.

MIERSWA, I. (2003): Beatles vs. Bach: Merkmalsextraktion im Phasenraum von Audiodaten. In: *LLWA 03 - Tagungsband der GI-Workshop-Woche Lernen - Lehren - Wissen - Adaptivität*.

MIERSWA, I. and KLINKENBERG, R. and FISCHER, S. and RITTHOFF, O. (2003): A Flexible Platform for Knowledge Discovery Experiments: YALE – Yet Another Learning Environment. In: *LLWA 03 - Tagungsband der GI-Workshop-Woche Lernen - Lehren - Wissen - Adaptivität*.

MITCHELL, M. T. (1996): *Machine Learning*. McGraw Hill, New York.

MORIK, K. (2000): The Representation Race – Preprocessing for Handling Time Phenomena. In: *Proc. of the 11th European Conference on Machine Learning*. Springer, Berlin, 4–19.

MORIK, K. and LIEDTKE, H. (2000): Learning about time. *MiningMart Deliverable No. 3, University of Dortmund*.

PYLE, D. (1999): *Data Preparation for Data Mining*. Morgan Kaufmann.

QUINLAN, R. (1993): *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Diego.

RITTHOFF, O. and KLINKENBERG, R. and FISCHER, S. and MIERSWA, I. (2002): A Hybrid Approach to Feature Selection and Generation Using an Evolutionary Algorithm. *Technical report CI-127/02, University of Dortmund*.

RÜPING, S. (2000): *mySVM - Manual*. University of Dortmund.

SCHLITTGEN, R. and STREITBERG, B. H. J. (1997): *Zeitreihenanalyse*. Oldenbourg Verlag München.

WITTEN, I. H. and FRANK, E. (2000): *Data Mining*. Morgan Kaufmann, San Diego.

Identification of Musical Instruments by Means of the Hough-Transformation

Christian Röver[1], Frank Klefenz[2], and Claus Weihs[1]

[1] University of Dortmund*
Department of Statistics
44221 Dortmund, Germany
roever@statistik.uni-dortmund.de
[2] Fraunhofer-Institut für Digitale Medientechnologie
Langewiesener Straße 22
98693 Ilmenau, Germany

Abstract. In order to distinguish between the sounds of different musical instruments, certain instrument-specific sound features have to be extracted from the time series representing a given recorded sound.
The Hough Transform is a pattern recognition procedure that is usually applied to detect specific curves or shapes in digital pictures (Shapiro (1978)). Due to some similarity between pattern recognition and statistical curve fitting problems, it may as well be applied to sound data (as a special case of time series data).
The transformation is parameterized to detect sinusoidal curve sections in a digitized sound, the motivation being that certain sounds might be identified by certain oscillation patterns. The returned (transformed) data is the timepoints and amplitudes of detected sinusoids, so the result of the transformation is another '*condensed*' time series.
This specific Hough Transform is then applied to sounds played by different musical instruments. The generated data is investigated for features that are specific for the musical instrument that played the sound. Several classification methods are tried out to distinguish between the instruments and it turns out that RDA (a hybrid method combining LDA and QDA) (Friedman (1989)) performs best. The resulting error rate is better than those achieved by humans (Bruderer (2003)).

1 The Hough-transform

The Hough-transform was originally developed to detect straight lines in (noisy) digital images, and was then later generalized to arbitrary lines or shapes. The procedure has similarities to regression methods, the common problem being to derive line parameters from points lying on that line. The Hough-transform is very robust to outliers, points that are not on the line have little influence on the estimation. It is even possible to fit several different lines independently at the same time (Shapiro (1978)).
Here the Hough-transform is applied to digitized sounds — as a special case

* The work of Christian Röver and Claus Weihs has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

of time series data — the question being, whether this yields a useful sound characterization. We will check this by trying to identify musical instruments by the sounds they play.
The motivation to apply the Hough-transform to sounds is that recently a computer chip has been developed that is able to perform the numerically expensive algorithm in real-time.

2 Application to sound data

2.1 Digital sounds

A sound is a periodic oscillation over time, as shown in Fig. 1. In this case the sound frequency (pitch) is 440 Hz, so the oscillation period is $\frac{1}{440} = 0.0023$ seconds, as indicated by the bar in the upper left. A digital sound

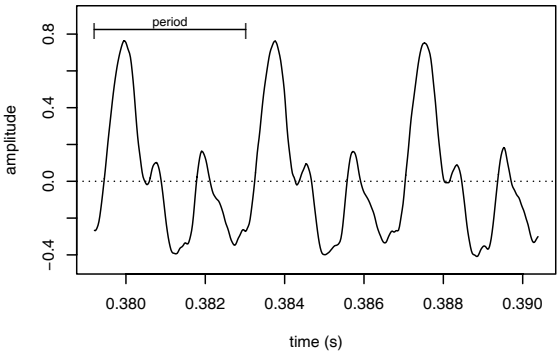


Fig. 1. Periodic oscillation of a sound.

recording is a discrete approximation of the original sound. The recording quality is determined by the resolution of this approximation: CD-tracks are recorded with a *sampling rate* of 44.1 kHz and a *resolution* of 16 bit, so the approximating step function has 44100 steps per second and each step height may take one out of $2^{16} \approx 65000$ values between 1 and -1.
So, statistically spoken, a digital sound is an equidistant time series.

2.2 Motivation: signal edges

The motivation to apply the Hough-transform is that a sound might have a specific oscillation pattern by which it can be identified. In order to catch the pattern features, we concentrate on the so-called *signal edges*, that is, the ascending oscillation sections rising from the time axis as indicated in Fig. 2. We will try to detect these signal edges by fitting appropriate curves

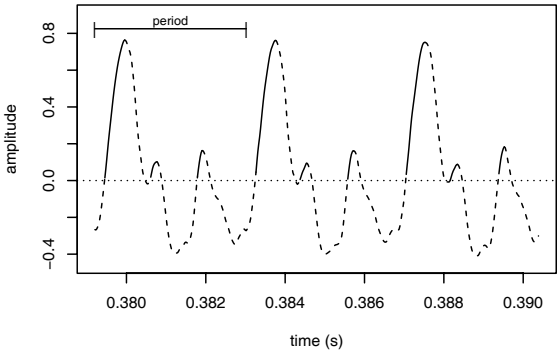


Fig. 2. Signal edges of a sound.

to the sound samples, and then see whether a sound can be classified by the generated sequence of signal edges.

2.3 Parametrization

The Hough-transform was then set to detect sinusoidal signal edges, that is, curves of the form

$$f(t) = A \cdot \sin(2\pi c \cdot t - \phi) \qquad (\phi \leq t \leq \phi + \tfrac{1}{4c})$$

are fit to the sound samples. Variable parameters are amplitude A (≥ 1) and phase difference ϕ (≥ 0); the center frequency c is fixed. The function is sketched in Fig. 3: A stretches the signal edge in the direction of the y-axis and so controls amplitude and slope, while ϕ places the edge along the time axis. Due to the transform procedure, both parameters take only discrete values: amplitudes are divided into 32 bins, and the phase difference resolution is defined by the sound sampling rate (44.1 kHz).
The transformation was then applied to the first 0.7 seconds of each sound, so for longer sound samples not the complete sound is captured in the transformed data.

2.4 Resulting data format

The result of transforming a digitized sound is another time series of amplitudes (A) and phase differences (ϕ); an example is given in Tab. 1: phase differences may be expressed in seconds or sample-indices, and the amplitude can be given in absolute values or bin-numbers. Note that low bin-numbers refer to high amplitudes (steep signal edges) and vice versa.
Fig. 4 shows the transformed data (amplitudes vs. time) for 4 different sounds, the left two played by a piano, and the right ones played on a trumpet. You

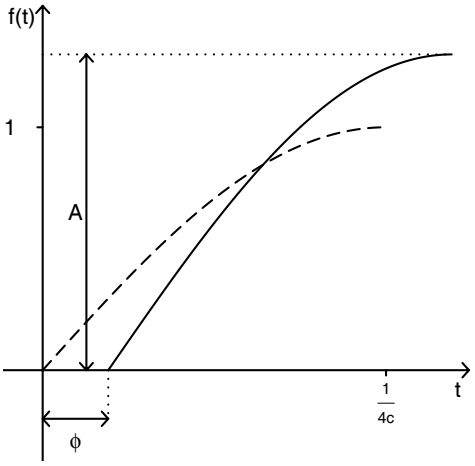


Fig. 3. The fitted signal edge.

	phase difference ϕ		amplitude A	
Nr.	sample	seconds	class-nr.	value
⋮	⋮	⋮	⋮	⋮
104	16731	0.3793881	28	1.163636
105	16838	0.3818141	31	1.049180
106	16894	0.3830841	22	1.488372
107	19896	0.3831291	25	1.306122
108	17004	0.3855781	30	1.084746
109	17065	0.3869611	27	1.207547
110	17173	0.3894101	31	1.049180
⋮	⋮	⋮	⋮	⋮

Table 1. Data format after transformation.

can clearly see similarities within the same instrument and differences across different instruments.
The next problem is now to derive characteristics from these time series that allow for classification of sounds.

3 Classification

3.1 Approaches

In general, two approaches were tried out to summarize the transformed data. The first question was whether the (overall) frequencies of amplitudes may yield a sufficient 'spectrum-like' sound characterization. The second approach

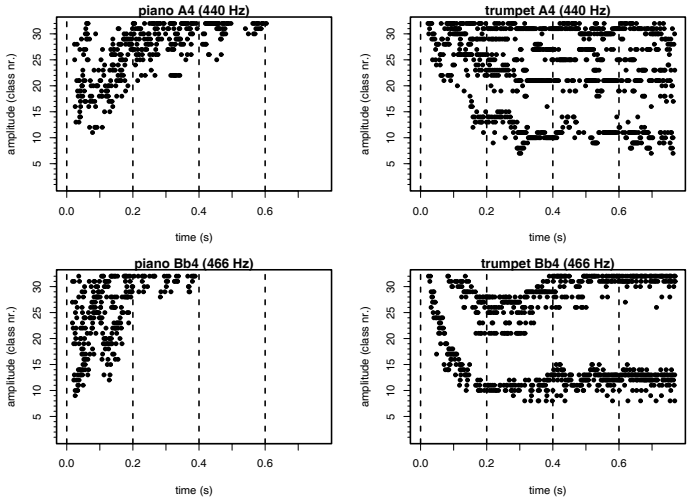


Fig. 4. Different instruments playing at same pitches; left: piano, right: trumpet.

was to derive other characteristics from the transformed time series (not only from amplitudes, but also from frequencies $f_i := \frac{1}{\phi_i - \phi_{i-1}}$).
The first approach uses 33 variables for classification (32 amplitude bins plus pitch), for the second approach 62 potential discriminators were derived from the transformed sound (for examples see results in section 6).

3.2 Data set

The investigated data set consisted of 1987 sounds played by different instruments and with pitches of each sound given. There were 62 sound sequences at subsequent pitches; different instruments covered different frequency bands, overall these spanned a range from A0 to C8 (27.5 to 4186 Hz). Sequences played by the same or very similar instruments were grouped together, like piano at different volumes or bassoon and contrabassoon. Finally, the set consisted of 25 instrument classes (Opolko and Wapnick (1987)).

3.3 Methods

The classification methods applied were:

- LDA: Linear Discriminant Analysis
- QDA: Quadratic Discriminant Analysis
- naive Bayes
- RDA: Regularized Discriminant Analysis
- Support Vector Machine
- Classification Tree
- k-NN: k-Nearest-Neighbour

Most methods should be well known except for RDA, which may require some explanation (for Classification Trees see Venables and Ripley (2002), for other methods see Hastie et al. (2001)).

Regularized Discriminant Analysis (RDA) is a hybrid method including LDA and QDA and was proposed by Friedman (1989). Assumptions and procedure are as in QDA, that is, group distributions are conditionally normal and the groups differ by their means and (co-)variances. But instead of using the usual groupwise covariance estimates, the covariance is manipulated using two parameters (λ and γ); first a convex combination is computed:

$$\hat{\Sigma}_k^{\mathrm{RDA}} = \lambda\hat{\Sigma}^{\mathrm{LDA}} + (1-\lambda)\hat{\Sigma}_k^{\mathrm{QDA}} \qquad (0 \le \lambda \le 1)$$

So the covariance estimate is a combination of the pooled ($\hat{\Sigma}^{\mathrm{LDA}}$) and the individual group covariances ($\hat{\Sigma}_k^{\mathrm{QDA}}$); for $\lambda = 1$ it is equal to LDA, and for $\lambda = 0$ it equals QDA. The second parameter γ then allows to shift the estimate towards an identity matrix, but this turned out not to improve error rates, so we restricted ourselves to using λ only and set γ to zero. Thus the covariance estimate simplifies to the above formula.

3.4 Variable selection

Variable selection is necessary for the second approach (characterizing variables), but not appropriate for the first (amplitude frequencies only). Also, classification trees select variables themselves.
For all other methods, variables were then selected applying the same principle (analogous to stepwise regression): Variables were selected step-by-step starting with pitch only and then in each step including the variable that improves the error rate (estimated by cross-validation) most.

3.5 Results

The best classification was achieved using 11 characterizing variables and applying RDA, which resulted in a misclassification rate of 26.1%. Using just the amplitude frequencies, the best error rate was only 66%, using k-Nearest-Neighbour.
The 11 discriminating features leading to the final error rate (26.1%) were:

- pitch
- waiting time for first edge and sound duration
- signal edge rate (per second)
- mean, variance and shape of amplitude distribution
- trend of amplitudes
- mean and variance of frequency distribution
- correlation of amplitude and frequency

%	ba	be	ce	cl	cr	eb	eg	ed	ef	fl	fr	gk	ma	ob	pi	sx	sy	tb	tp	tp	tu	vb	vp	vi	xy	Σ
bassoon	78	0	2	1	0	1	0	0	0	0	0	0	0	1	0	0	2	9	0	0	6	0	0	0	0	22
bells	0	95	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	5
cello	6	0	72	3	0	0	4	3	0	0	0	0	0	1	0	4	0	0	2	0	5	0	0	0	0	28
clarinet	2	0	3	52	0	0	0	8	0	2	1	0	0	7	0	10	0	3	7	0	1	1	0	3	0	48
crotales	0	0	0	0	97	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	3
elecbass	0	0	0	0	0	80	7	0	4	0	0	0	2	0	4	0	0	0	0	0	2	1	0	0	0	20
elecguitar	1	6	8	1	0	12	53	1	2	0	1	0	0	0	4	1	0	1	0	0	1	6	0	1	1	47
elecguitar-*distd.*	0	0	0	1	0	3	0	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5
elecguitar-*fh.*	0	0	0	0	0	12	3	0	73	0	0	0	0	0	3	0	0	0	0	0	0	1	8	0	0	27
flute	1	0	1	1	0	0	0	0	0	69	0	0	0	3	0	2	0	3	3	0	8	2	0	6	0	31
french horn	0	0	0	2	0	0	0	0	0	0	90	0	0	4	0	2	0	0	2	0	0	0	0	0	0	10
glockenspiel	0	0	0	0	11	0	0	0	0	0	0	83	0	0	1	0	0	0	2	0	0	0	0	0	2	17
marimba	0	0	0	0	0	8	0	0	0	0	0	0	61	0	1	0	0	0	0	0	0	0	3	0	26	39
oboe/enghorn	0	0	0	9	0	0	0	0	0	5	2	0	0	70	0	2	0	2	7	1	0	0	0	2	0	30
piano	6	1	1	0	0	7	3	0	1	0	0	2	10	0	55	0	0	0	0	0	0	4	2	0	8	45
saxophone	8	0	10	11	0	0	0	0	0	0	7	0	0	6	0	46	0	3	6	0	0	2	0	0	0	54
synthbass	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	98	0	0	0	0	0	0	0	0	2
trombone	4	0	0	7	0	0	0	1	0	3	0	0	0	3	0	0	0	73	7	0	0	0	0	1	0	27
trumpet	0	0	1	2	0	0	0	0	0	4	5	0	0	8	0	2	0	7	68	0	0	3	0	0	0	32
trumpet-*csto*	0	0	0	3	0	0	0	3	0	0	0	0	0	3	0	0	0	0	0	90	0	0	0	0	0	10
tuba	3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	95	0	0	0	0	5
vibraphone	0	2	1	1	0	5	8	0	2	9	1	0	3	0	0	0	0	0	1	1	7	57	0	0	1	43
violin-*pizzicato*	0	0	0	0	0	2	0	0	5	0	0	1	6	0	2	0	0	0	0	0	0	0	84	0	0	16
violin/viola	2	0	2	6	0	0	0	0	2	7	1	0	3	16	0	1	0	7	1	2	1	0	1	48	1	52
xylophone	0	0	0	0	0	0	0	0	1	0	0	5	23	0	2	0	0	0	0	0	0	2	0	0	66	34

total misclassification rate: 26.1%

Table 2. Confusion matrix for RDA using 11 variables (percentages).

The error rates are shown in detail in the confusion matrix (Table 2): each line corresponds to one instrument and shows how it was classified (in percentages); the main diagonal shows correct classifications, the off-diagonal elements show false classifications. The last column gives the total (instrument-wise) error rate.

For example, you can see that xylophone and marimba get confused with each other, and that there are certain instruments that are classified well (bells), while others are not clearly identified (saxophone).

Closer examination of the transformed data suggested that tuning of Hough-transformation settings might lead to further improvement of classification results. For further details see Röver (2003).

3.6 Comparing the results

The misclassification rate achieved by pure guessing would be $\frac{24}{25} = 96\%$. Error rates achieved by humans or other automatic classification approaches have previously been investigated in other experiments; in roughly comparable problem settings (with regards to number of instruments) rates for humans are quoted at 44%, and for automatic classification these range from 19–7.2% (Bruderer (2003)).

Note that in this study only the first 0.7 seconds of a sound were used, whereas usually complete sounds are evaluated for recognition. Other approaches often use features like envelope characteristics or fourier frequencies for classification.

4 Conclusions

Application of the Hough-transform to digitized sounds yields a useful sound characterization; the generated data allows to distinguish between sounds played by different instruments. Classification of 25 instruments leads to an error rate of 26.1%.
The misclassification rate so far is better than those achieved by humans, but still worse than for other automatic approaches. Further tuning of transform settings and application to complete sounds (longer than 0.7 seconds) might still improve the procedure.

References

BRUDERER, M.J. (2003): *Automatic recognition of musical instruments*. Masters Thesis, Ecole Polytechnique Fédérale de Lausanne.

FRIEDMAN, J.H. (1989): Regularized Discriminant Analysis. *Journal of the American Statistical Association, 84(405), 165–175*.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *The elements of statistical learning; data mining, inference, and prediction*. Springer-Verlag, New York.

OPOLKO, F. and WAPNICK, J. (1987): McGill University Master Samples (*CD-Set*). See `http://www.music.mcgill.ca/resources/mums/html/`

RÖVER, C. (2003): *Musikinstrumentenerkennung mit Hilfe der Hough-Transformation*. Diploma Thesis, Universität Dortmund.

SHAPIRO, S.D. (1978): Feature Space Transforms for Curve Detection. *Pattern Recognition, 10, 129–143*.

VENABLES, W.N. and RIPLEY, B.D. (2002): *Modern Applied Statistics with S*, 4th ed. Springer-Verlag, New York.

Support Vector Machines for Bass and Snare Drum Recognition

Dirk Van Steelant[1], Koen Tanghe[2], Sven Degroeve[3], Bernard De Baets[1], Marc Leman[2], and Jean-Pierre Martens[4]

[1] Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, 9000 Ghent, Belgium
[2] Department of Musicology (IPEM), Ghent University, Belgium
[3] Department of Bioinformatics, Ghent University, Belgium
[4] Department of Electronics and Information Systems (ELIS), Ghent University, Belgium

Abstract. In this paper we attempt to extract information concerning percussive instruments from a musical audio signal. High-dimensional vectors of descriptors are computed from the signal and classified by means of Support Vector Machines (SVM). We investigate the performance on 2 important classes of drum sounds in Western popular music: bass and snare drums, possibly overlapping. The results are encouraging: SVM achieve a high accuracy and F_1-measure, with linear kernels performing (nearly) as good as Gaussian kernels, but requiring 1000 times less computation time.

1 Introduction

With the explosive growth of the amount of digital music available on the Internet, Musical Information Retrieval has become a topic that has attracted the attention of researchers in a wide range of disciplines. The quality of content-based retrieval however depends heavily on how well the individual components for representation and matching of the data perform. Most existing commercial music information retrieval systems use text as the main supplier of meta-data of music, such as the name of the artist/performer and the title of the song. For text, rapid matching methods are available and are applied extensively in search engines on the World Wide Web. However, as soon as such meta-data is incomplete or unavailable, all of the existing commercial systems will fail to deliver.

The MAMI-project (Musical Audio MIning) aims at working out methodologies and software tools for content-based audio-mining by bundling the efforts of musicologists, engineers, mathematicians and computer scientists. MAMI is centered on the 'query-by-imitation' paradigm, where users can retrieve a musical piece by means of its sound characteristics, either by describing, playing or vocally imitating the piece.

In order to supply a ranked list of candidate songs to the user, the system has to match an intermediate representation of the (melodic or rhythmic)

input with a similar representation of all the songs in the database; this will typically be done by means of a (time-consuming) dynamic programming technique. To speed up the query, any additional information that can narrow down the search space is welcome; not only meta-data, but also a description of the content of the target song or the musical genre to which it belongs can be used for this purpose.

A user study (Lesaffre et al. (2003)) has shown that when users are asked to imitate a song they are familiar with, some of them will reproduce the rhythmic structure of the piece. This is one of the motivations for analyzing the percussive content of musical audio; if a transcription can be obtained, it can be matched with the description delivered by the user, used as a feature for genre classification or provide valuable information for the determination of beat, tempo and rhythmic structure.

For the recognition of drum sounds three levels of difficulty can be distinguished: (i) Isolated drum sounds; (ii) Overlapping drum sounds; (iii) Overlapping drum sounds layered with other instruments and voices.

Obtaining a full transcription of the percussive content of musical audio is a challenging task and, to our best knowledge, has never been attempted using SVM. We will therefore concentrate on two important classes of sounds (omnipresent in Western popular music): bass drums (typically low-pitched and strongly indicating the beat) and snare drums (with highly noisy components, delivering important clues about the metrical structure of the song). In this paper we will concentrate on musical audio situated at the first and second level, since Virtanen (2001) has shown recently that it is possible to extract drum tracks from musical audio by subtracting the harmonic parts from the signal.

The rest of this paper is organized as follows. In Section 2 we give an overview of previous work. In Section 3 we describe how data were generated using samples gathered from commercial CD's, standard MIDI songs and sequencer software. In Section 4 relevant descriptors for audio data are presented and in Section 5 Support Vector Machines are formally introduced. In Section 6 we report results for two experiments and in Section 5 we comment on these results and give directions for future research.

2 Previous work

A recent overview of classification techniques for musical instrument sounds in general can be found in Herrera-Boyer et al. (2003). Percussive instruments represent a special case as they can be considered to be pitch-independent, so their appearance throughout a musical piece is much more constant. Although this makes them good candidates for localization/classification, they only represent a small part of previous research and in most cases only recognition of isolated sounds is investigated.

McDonald and Tsang (1997) use Spectral Centre Trajectories to classify percussive sounds but tests are only conducted on a very small database. In Zils et al. (2002) a percussion transcription is obtained by an analysis by synthesis technique, whereby the sound searched for is gradually synthesized from the signal; a success rate of over 75% is reported. A large-scale study in Herrera et al. (2003) uses different subsets of temporal and spectral descriptors (up to 207) for the recognition of thirty different classes of isolated percussion instruments. K-NN, Kernel Density (KD) estimation, canonical discriminant analysis and decision trees (C4.5) were investigated as classification techniques. KD combined with correlation-based feature selection yielded a 85% hit rate.

3 Data gathering

We have gathered samples belonging to two classes of percussive instruments (bass drum and snare drum) from commercial sample CD's. Such CD's typically indicate the class to which a sound belongs by the name of the sample or its location in the directory structure, but this information is not always equally reliable. Listening to the sounds we realized that some of them were mixed with other (percussive) instruments and therefore we had them classified by two users; only the samples that were considered to be "pure" and correctly classified by both users were retained. This yielded 656 bass drums and 604 snare drums; in all classes samples of the acoustic as well as of the electronic type were selected.

To gather realistic data, MIDI (Musical Instrument Digital Interface) files were exploited. Standard MIDI assigns classes of instruments to predefined tracks which makes it possible for an electronic sound device supporting standard MIDI to play songs with its own internal sounds. From 32 songs in standard MIDI format we selected 16 measures of the drum track. These 32 files were loaded into a sequencer program; 8 variations for each track were generated by selecting at random pairs of bass and snare drum from the set of samples while the other drum sounds were drawn from a standard MIDI drum set. The audio generated by playing back the MIDI files using these sets of drum sounds was recorded, yielding 256 audio files in total. The isolated drum sounds were added to the data set. This yielded a positive/negative example ratio of 1472/2508 for the bass drums and 1315/2729 for the snare drums. All files were saved as mono wave files sampled at 44.1 KHz.

In order not to introduce any errors due to the incorrect localization of events, we did not perform any onset detection but instead used the timing and labelling information available in the MIDI files to determine at what position in time descriptors need to be extracted and whether an event is a positive or negative instance for our binary classifiers. The information in the MIDI files thus represents the "ground truth" for the corresponding recorded audio renderings.

4 Descriptors for audio

Digital audio corresponds to a very high data rate (88 Kbyte/s for mono CD quality). To arrive at a manageable data rate, one needs to select descriptors that capture the characteristics of the audio while suppressing details that are redundant for the problem at hand. This data reduction will typically be done by sliding a window with a fixed step over the raw audio signal (e.g. a 20 ms window every 10 ms) and by computing at every step descriptors over that window.

The events we are trying to classify do not have a fixed length; the bass drums in our database for example have a duration ranging from 71 ms to 1.892 s. Although SVM are able to handle variable temporal representations by applying specific kernels, e.g. Shimodaira et al. (2001), determining the end of an event in musical audio (offset detection) is difficult. We therefore decided to use a fixed context at the beginning of the events over which descriptors are to be computed. In Section 6 we determine the most appropriate context length for each class. In order not to confuse the binary classifiers, we excluded any negative examples that lie within the range of 50 ms of a positive example.

A first set of descriptors concerns the energy in the signal computed by means of a Root Mean Square (RMS) formula. When inspecting the accumulated spectra of hundreds of bass drums and snare drums, it can be seen that the spectral energy distributions of these different sounds are located in more or less distinct frequency bands (although not completely separated). Hence we divided the spectrum into three frequency bands and computed energy-related descriptors over these bands: RMS in the whole signal, RMS per frequency band, ratio of RMS to overall RMS (per band) and RMS per band relative to RMS of other bands (1 to 2, 1 to 3 and 2 to 3).

Temporal descriptors are computed on the sample signal. The following descriptors were withheld: Zero Crossing Rate (ZCR): number of times per second the signal changes sign; Crest Factor: ratio of maximum absolute value sample signal to RMS in the segment; Temporal Centroid: the center of gravity of the distribution of the absolute values of the samples in the window. Spectral descriptors are computed using the Fast Fourier Transform, which converts the time domain data into the frequency domain: spectral centroid, skewness and kurtosis; and the spectral rolloff.

Logan (2000) shows that Mel Frequency Cepstral Coefficients (MFCC), short-term spectral-based features widely used for speech recognition, are appropriate as a representation for music by examining the functionality of a music/speech discriminator. MFCC are especially interesting for complex music analysis because they combine low-dimensionality and the ability to discriminate between different spectral content. The amount of detail in the description depends on the number of coefficients extracted; for our experiments 12 coefficients were computed. The temporal deployment of these descriptors is further captured by computing their first and second order derivatives. As a window size of 20 ms and frame step of 10 ms for the extrac-

tion of this kind of descriptors is often advised, we used these settings and we computed the mean and standard deviation of the coefficients and their first and second order derivatives over the context.

5 Support Vector Machines

Formally, a data set T contains l instances $\mathbf{x_i}$ $(i = 1, \ldots, l)$ with each $\mathbf{x_i}$ labelled as $y_i = 1$ or $y_i = -1$ (known as *classes*), indicating a positive or negative instance, respectively. Each index x_{ij} $(j = 1, \ldots, n)$ in vector $\mathbf{x_i}$ is a descriptor as described above.

The Support Vector Machine (Vapnik (1995)) is a data-driven method for solving two-class classification tasks. The Linear SVM (LSVM) separates the two classes in T with a hyperplane in the input space such that:

(a) the "largest" possible fraction of instances of the same class are on the same side of the hyperplane, and
(b) the distance of either class from the hyperplane is maximal.

The prediction of an LSVM for an unseen instance $\mathbf{z}$ is given by the decision function

$$pred(\mathbf{z}) = \mathrm{sgn}(\mathbf{w} \cdot \mathbf{z} + b)\,. \tag{1}$$

The hyperplane is computed by means of a vector of Lagrange multipliers α maximizing

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i\, \alpha_j\, y_i\, y_j\, (\mathbf{x_i} \cdot \mathbf{x_j})\,,$$

subject to:

$$0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^{l} \alpha_i\, y_i = 0\,, \tag{2}$$

where C is a parameter set by the user to regulate the effect of outliers and noise, i.e. it defines the meaning of the word "largest" in (a). Some tolerance (denoted as ϵ) on the constraints in Equation 2 is acceptable.

A function K (called a kernel function) maps the descriptors in T, called the input space, into a feature space defined by K in which then a linear class separation is performed. For the LSVM this mapping is a linear mapping:

$$K(\mathbf{x_i}, \mathbf{x_j}) = \mathbf{x_i} \cdot \mathbf{x_j}\,. \tag{3}$$

The non-linear mapping used in this paper is the Gaussian-SVM (GSVM)

$$K(\mathbf{x_i}, \mathbf{x_j}) = e^{-|\mathbf{x_i} - \mathbf{x_j}|^2/\gamma^2}\,. \tag{4}$$

After calculating the α_i's in (2), the decision function (1) becomes:

$$pred(\mathbf{z}) = \text{sgn}(\sum_{i=1}^{l} \alpha_i \, y_i \, K(\mathbf{x_i}, \mathbf{z}) + b) \, . \tag{5}$$

An instance x_i for which α_i is not zero is called a Support Vector (SV). Note that the prediction calculated in (5) uses the support vectors only. As such, the support vectors are those instances that are closest to the decision boundary in the feature space.

All SVM in our experiments were trained using SVM$^{\text{light}}$ 5.0 (Joachims (1999)[1]) in classification mode with all parameters at their default values, except for C and the kernel-related parameter γ. The data were scaled so that every descriptor lies within the range $[-1, 1]$.

6 Experiments and results

In order to determine an appropriate context length for the two classes of drum sounds we computed the descriptors over various lengths (50, 70, 100, 140, 170 and 200 ms) and performed 3-Fold Cross Validation (3-FCV) using LSVM with $C = 2^i$ $(i = -8, \ldots, 0, \ldots, 10)$. As a performance measure we combined the obtained average precision and recall into

$$F_\beta = \frac{(\beta^2 + 1) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}$$

with β a user-controlled parameter expressing the preference for either high precision or high recall ($\beta = 1$ in the sequel). Table 1 shows the best F_1 for various context lengths. For both bass drum and snare drum the best performance was obtained using a context length of 100 ms.

Table 1. F_1 with 3-FCV on the whole data set for different context lengths

Context (ms)	50	70	100	140	170	200
F_1 BD	93.91	95.11	**95.15**	94.94	94.59	94.64
F_1 SD	97.39	97.69	**98.18**	97.61	97.30	96.58

Using the obtained context lengths, we investigated the difference in performance between a linear and (the more powerful) Gaussian kernel. It needs to be pointed out that there is no guarantee that the optimal context length for LSVM is also optimal for GSVM; ongoing research will have to clarify this point.

[1] http://svmlight.joachims.org/

The data were split into a 87.5% training set (for model selection and training) and a 12.5% test set (while respecting the balance between positive and negative examples). We had the optimal parameters for Gaussian kernels (C, γ) established by *looms* (Lee and Lin (2000)[2]) which estimates the leave-one-out error rate over a grid of candidate values using a loose stopping criterion in the optimization phase. For LSVM we obtained the optimal C using 3-FCV on the training set and F_1 as performance measure. The results in Table 2 also contain overall accuracy and the number of support vectors for the obtained models. These results show a very minor difference in performance for BD and no improvement at all for SD; despite bigger computational effort for model selection and the fact that the resulting model is more complex (the number of support vectors has almost doubled), exactly the same misclassifications are done with the Gaussian kernel as with the linear one.

Table 2. Classification of the 12.5% test set using LSVM (model selection by 3-FCV) and GSVM (model selection by estimating leave-one-out error).

	LSVM		GSVM	
	BD	SD	BD	SD
C	0.5	8	0.25	8
γ	-	-	0.256	0.064
accuracy	94.98	97.63	95.58	97.63
F_1	93.30	96.34	94.12	96.34
#SV	391	179	965	350

7 Conclusions and future work

The results show that our audio descriptors and SVM classifiers combine well into a technique for the recognition of drum sounds in an audio signal. We expect that the methodology can be extended for the detection of a wider range of percussive instruments.

The observation that linear kernels perform only slightly worse than the Gaussian ones is an important finding for applications in time-critical environments. An LSVM with approximately 1300 support vectors classifies 5000 examples (89-dimensional) in less than 10 ms while it takes a GSVM with the same number of SV close to 10 s (done on a mobile Pentium III 1.2 GHz with 256 DDR RAM); this difference could turn out to be crucial in a real-time system that, besides classification, also needs to perform onset detection and compute appropriate descriptors.

[2] http://www.csie.ntu.edu.tw/~cjlin/looms/

As there is a vast amount of candidate descriptors for the modelling of audio and various ways of encoding them, future research should try to extend the set of descriptors and at the same time, for the sake of simplicity, reduce it by means of variable selection methods (e.g. as in Degroeve et al. (2002)). Findings related to what kind of descriptors are relevant for the recognition of percussion would also provide interesting feedback to researchers in the field of musicology and perceptual psychology.

Acknowledgements

This research is funded by the Flemish Institute for the Promotion of Scientific and Technical Research in Industry.

References

DEGROEVE, S., DE BAETS, B., VAN DE PEER, Y. and ROUZE, P. (2002): Feature Subset Selection for Splice Site Prediction. *Bioinformatics, 18, 75–83.*

HERRERA, P., DEHAMEL, A. and GOUYON, F. (2003): Automatic Labeling of Unpitched Percussion Sounds. In: *Proc. 114th Convention of the Audio Engineering Society, Amsterdam, The Netherlands.*

HERRERA-BOYER, P., PEETERS, G. and DUBNOV, S. (2003): Automatic Classification of Musical Instrument Sounds. *Journal of New Music Research, 32, 3–21.*

JOACHIMS, T. (1999): Making Large-scale SVM Learning Practical, In: B. Schölkopf, C. Burges and A. Smola (Eds.): *Advances in Kernel Methods - Support Vector Learning.* MIT Press.

LEE, J.-H. and LIN, C.-J. (2000): Automatic Model Selection for Support Vector Machines. Technical report, Dept. of Computer Science and Information Engineering, National Taiwan University.

LESAFFRE, M., MOELANTS, D., LEMAN, M., DE BAETS, B., DE MEYER, H., MARTENS, G. and MARTENS, J.-P. (2003): User Behavior in the Spontaneous Reproduction of Musical Pieces by Vocal Query. In: *Proc. 5th Triennial ESCOM Conference, Hannover, Germany.*

LOGAN, B. (2000): Mel Frequency Cepstral Coefficients for Music Modelling. In: *Proc. Internat. Symposium on Music Information Retrieval, Plymouth, MA, USA.* 23–25.

MCDONALD, S. and TSANG, C.P. (1997): Percussive Sound Identification using Spectral Centre Trajectories. Technical Report Departmental Conference, Yanchep.

SHIMODAIRA, H., NOMA, K., NAKAI, K. and SAGAYAMA, S. (2001): Support Vector Machine with Dynamic Time-alignment Kernel for Speech Recognition. In: *Proc. Eurospeech, Aalborg, Denmark.*

VAPNIK, V.N. (1995): *The Nature of Statistical Learning Theory.* Springer-Verlag.

VIRTANEN, T. (2001): Audio Signal Modeling with Sinusoids plus Noise. MSc thesis, Tampere University of Technology.

ZILS, A., PACHET, F., DELERUE, O. and GOUYON, F. (2002): Automatic Extraction of Drum Tracks from Polyphonic Music Signals. In: *Proc. 2nd Internat. Conference on Web Delivering of Music, Darmstadt, Germany.*

Register Classification by Timbre

Claus Weihs[1], Christoph Reuter[2], and Uwe Ligges[1]

[1] University of Dortmund*, Department of Statistics 44221 Dortmund, Germany
[2] Musikwissenschaftliches Institut, Universität Wien, A-1090 Wien, Austria

Abstract. The aim of this analysis is the demonstration that the high and the low musical register (Soprano, Alto vs. Tenor, Bass) can be identified by timbre, i.e. after pitch information is eliminated from the spectrum. This is achieved by means of pitch free characteristics of spectral densities of voices and instruments, namely by means of masses and widths of peaks of the first 13 partials (cp. Weihs and Ligges (2003b)).

Different analyses based on the tones in the classical song "Tochter Zion" composed by G.F. Händel are presented. Results are very promising. E.g., if the characteristics are averaged over all tones, then female and male singers can be easily distinguished without any error (prediction error of 0%)! Moreover, stepwise linear discriminant analysis can be used to separate even the females together with 28 high instruments ("playing" the Alto version of the song) from the males together with 20 low instruments (playing the Bass version) with a prediction error of 4%. Also, individual tones are analysed, and the statistical results are discussed and interpreted from acoustics point of view.

1 Introduction

Sound characteristics of orchestra instruments derived from spectra are currently a very important research topic (see, e.g., Reuter (1996, 2002)). The sound characterization of voices has, however, many more facets than for instruments because of the sound variation in dependence of technical level and emotional expression (see, e.g., Kleber (2002)).

During a former analysis of singing performances (cp. Weihs and Ligges (2003b)) it appeared that register can be identified from the spectrum even after elimination of pitch information. In this paper this observation is assessed by means of a systematic analysis not only based on singing performances but also on corresponding tones of high and low pitched instruments. The aim of this analysis is the demonstration that the high and the low musical register (Soprano, Alto vs. Tenor, Bass) can be identified by timbre, i.e. by the spectrum after pitch information is eliminated. To this end, pitch independent characteristics of spectral densities of instruments and voices are generated. As in the voice prints introduced in Weihs and Ligges (2003b) we use masses and widths of peaks of the first 13 partials, i.e. of the fundamental and the first 12 overtones. These characteristics are computed for representatives of all tones involved in the classical song "Tochter Zion" composed

* The work of Claus Weihs and Uwe Ligges has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

by G.F. Händel. For the singing performances the first representative of each note was chosen, for the instruments the representatives were chosen from the "McGill University Master Samples" (see section 2). These data were analysed with Linear Discriminant Analysis (LDA) and decision trees (see section 3). The results are very promising (see section 4). Some acoustics' explanations of our findings are given in section 5.

2 Data

The analyses of this paper are based on time series data from an experiment with 17 singers performing the classical song "Tochter Zion" (Händel) to a standardized piano accompaniment played back by headphones (cp. Weihs et al. (2001)). The singers could choose between two accompaniment versions transposed by a third in order to take into account the different voice types (Soprano and Tenor vs. Alto and Bass). Voice and piano were recorded at different channels in CD quality, i.e. the amplitude of the corresponding vibrations was recorded with constant sampling rate 44100 hertz in 16-bit format. The audio data sets were transformed by means of a computer program into wave data sets. For time series analysis the waves were reduced to 11025 Hz (in order to restrict the number of data), and standardized to the interval $[-1, 1]$. Since the volume of recording was already controlled individually, a comparison of the absolute loudness of the different recordings was not sensible anyway. Therefore, by our standardization no additional information was lost.

Since our analyses are based on characteristics derived from tones corresponding to single notes, we used a suitable segmentation procedure (Ligges et al. (2002)) in order to get data of segmented tones corresponding to notes. The periodograms (cp. Brockwell and Davis (1991)) used for the analyses described in this paper were calculated from overlapping sections of 2048 observations, overlap starting in the middle of the preceding section. This way, we get roughly $11 (= 2 \cdot (11025/2048))$ periodograms per second of sound, whereas the duration of the whole song is roughly 60 seconds. These periodograms are classified to notes, and the notes are smoothed by means of double median smoothing. Based on the smoothed series of notes, begin and end of sung notes are decided upon. For further analysis the first representative of the notes with identical pitch in the song was chosen. This leads to 9 different representatives per voice in "Tochter Zion".

The notes involved in the analyzed song were also identified in the "McGill University Master Samples" either in the Alto or in the Bass version for the following instruments:

Alto version (McGill notation): *aflute-vib, bells, cello-bv, clari-bfl, clari-efl, elecguitar1, elecguitar4, enghorn, flute-flu, flute-vib, frehorn, frehorn-m, marimba, oboe, piano-ld, piano-pl, piano-sft, sax-alt, tromb-ten, trump-ba, trump-c, trump-csto, vibra-bow, vibra-hm, viola-bv, viola-mv, violin-bv,*

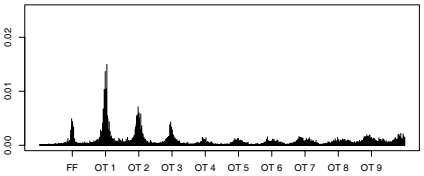


Fig. 1. Pitch independent periodogram (professional bass singer).

violin-mv.

Bass version: *bassoon, bflute-flu, bflute-vib, cello-bv, elecbass1, elecbass5, elecbass6, elecguitar1, elecguitar2, elecguitar4, frehorn, frehorn-m, marimba, piano-ld, piano-pl, piano-sft, tromb-ten, tromb-tenm, tuba, viola-mv.*
Thus, 28 high instruments and 20 low instruments were chosen together with 10 high female singers and 7 male.

From the periodogram corresponding to each tone corresponding to an identified note voice print characteristics are derived (cp. Weihs and Ligges (2003b)). For our purpose we only use the size and the shape corresponding to the first 13 partials, i.e. to the fundamental frequency and the first 12 overtones, in a pitch independent periodogram (cp. Figure 1). In order to measure the size of the peaks in the spectrum, the mass (weight) of the peaks of the partials are determined as the sum of the percentage shares of those parts of the corresponding peak in the spectrum which are higher than a pre-specified threshold. The shape of a peak cannot easily be described. Therefore, we only use one simple characteristic of the shape, namely the width of the peak of the partials. The width of a peak is measured by the half tone distance between the smallest and the biggest frequency of the peak with a spectral height above a pre-specified threshold. Overall, every tone is characterized by the above 26 characteristics which are used as a basis for classification. For details on the computation of the measures see Güttner (2001). Note that pitch information is eliminated in that the frequencies corresponding to fundamentals and overtones are ignored in the pitch independent periodogram. Mass is measured as a percentage (%), whereas width is measured in parts of halftones (pht). Figure 2 illustrates the voice print corresponding to the whole song "Tochter Zion" for a particular singer. For masses and widths boxplots are indicating variation over the involved tones. For the analyses of this paper we ignore halftone distance and formant intensity (cp. Weihs and Ligges (2003b)), and use the other characteristics of the voice print for individual tones, as well as averaged characteristics over all involved tones, leading to only one value for each characteristic per singer or instrument.

3 Classification methods

On these data we applied supervised classification methods (see, e.g., Michie et al. (1994)) trying to reproduce the pre-defined grouping by means of classi-

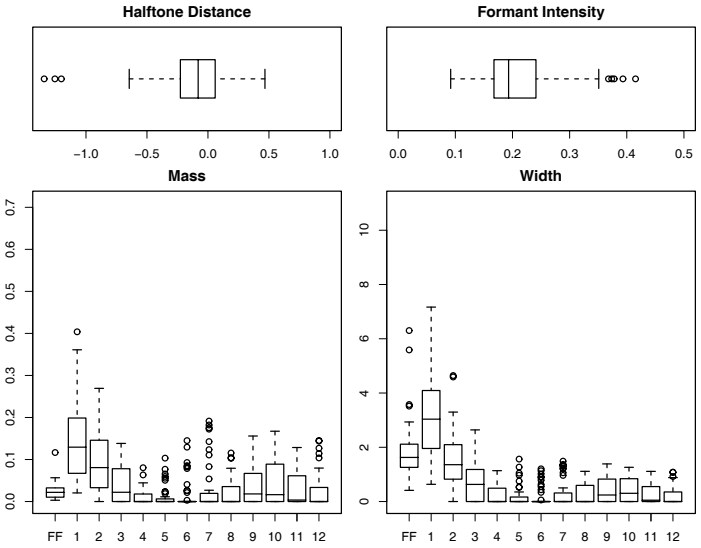


Fig. 2. Voice print of professional bass singer.

fication rules from the chosen voice print characteristics. We applied the easily interpretable classification tree (more specifically RPART by Therneau and Atkinson (1997)) and the well-known statistical linear discrimination analysis (LDA) to our data. These two classification methods are often identified to be adequate for quite different situations. For such methods the classification quality can, e.g., be measured by means of the misclassification rate, i.e. the ratio of the wrongly classified cases to the overall number of cases, which will be estimated by cross-validation.

4 Results

4.1 Individual tones, voices only

Let us start with the analysis of individual tones. If one restricts oneself to voices, then the best classification with only an error rate of 9.2% (estimated by 10-fold cross-validation) resulted from using only MassFF, MassOT01, WidthFF, WidthOT01 as predictors in LDA. The classification is detailed in Table 1. Obviously, the middle voice types Alto and Tenor generate the most errors. The results even show that the four characteristics MassFF, MassOT01, WidthFF, WidthOT01 are more appropriate for prediction of register than all 26 characteristics together (12.4% error). Thus, there are characteristics that deliver prediction irrelevant information for the classification rule. The prediction error of 9% of the individual notes appear to be acceptable. The most important characteristics for separation of high and low

Table 1. Classifying individual tones of voices with `LDA(MassFF, MassOT01, WidthFF, WidthOT01)`.

	high	low	error
Soprano	33	3	0.083
Alto	47	7	0.130
Tenor	3	24	0.111
Bass	1	35	0.028

voices are MassFF and WidthFF with 8.5% apparent error rate. However, the groups are not very well separated even for these characteristics. MassFF alone is not sufficient for prediction (21.6% error).

In the following we will mainly concentrate on reporting of the results of `LDA(MassFF, MassOT01, WidthFF, WidthOT01)`. Other results will only be mentioned in comparison. Note, however, that decision trees were never competitive.

4.2 Individual tones, voices and instruments

Considering the voices together with the instruments, the error rate of `LDA(MassFF, MassOT01, WidthFF, WidthOT01)` is roughly doubled, namely from 9.2% to 17.1% of the individual notes (estimated by 10-fold cross-validation). The only instruments which are predominantly misclassified are bass French horn and bass-marimba with 72% and 89% error, correspondingly. Again, the characteristics MassFF and WidthFF separate high and low particularly well (20.7% apparent error rate). However, the combination MassOT01 and WidthOT01 is even somewhat better (19.9%). Separation of groups is even worse than for voices alone. MassFF alone is, again, not sufficient for prediction (38.1% prediction error). Note, however, that LDA based on all 26 characteristics leads to the distinctly best error rate (14.2%). Here only bass-marimba is particularly bad predicted.

4.3 Averaged tones, voices only

After averaging the characteristics of the individual tones, i.e. using only one value for each characteristic per voice, prediction is possible without any error (0% error estimated by 17-fold cross-validation) using the classification rule based on `LDA(MassFF, MassOT01, WidthFF, WidthOT01)`. The apparent error rate is 0% for three pairs of characteristics, namely for "MassFF, WidthFF", "MassOT01, WidthOT01", and "WidthFF, WidthOT01". Again, MassFF alone is not sufficient for prediction (error rate = 11.8%).

4.4 Averaged tones, voices and instruments

If instruments are considered also, then the error rate is only increasing to 4.6% for `LDA(MassFF, MassOT01, WidthFF, WidthOT01)` (estimated by 65-

Table 2. Classification of voices and instruments based on averaged characteristics.

	LDA(MassFF, MassOT01, WidthFF, WidthOT01)			LDA(all charact.)		
	high	low	error	high	low	error
Soprano	4	0	0.000	4	0	0.000
Alto	6	0	0.000	6	0	0.000
A-instr.	28	0	0.000	28	0	0.000
Tenor	0	3	0.000	1	2	0.333
Bass	0	4	0.000	0	4	0.000
B-instr.	3	17	0.150	1	19	0.050

fold cross-validation, i.e. by leave-one-out cross validation). Only the low instruments cannot be predicted perfectly (see Table 2). When considering all characteristics the corresponding error rate of the LDA classification rule is somewhat decreasing to 3.1%. In the case of `LDA(MassFF, MassOT01, WidthFF, WidthOT01)` only three bass-instruments are wrongly predicted as high, namely French horn (stomped and not stomped) and Marimba. Using LDA with all characteristics only Marimba and one Tenor singer was wrongly classified. The scatterplot matrix shows that the variable pair "MassOT01, WidthOT01" leads to the smallest apparent error rate (see Figure 3). Again, using only MassFF for prediction is not sufficient (41.5% error!).

5 Acoustics

Our findings are well supported by acoustics. Some explanations are the following. The relatively small opening of the human mouth acts as a high pass filter, i.e. the lower the tone the less the mass of the fundamental relative to the 1st overtone. This was already found in the middle of the last century (s. Scheminzky (1943), 428). From this it, e.g., follows that sopranos have more mass in the fundamental than basses. Moreover, synthesizing the fundamental together with a 18 dB weaker 1st overtone plus a vibrato typical for singing voices (6 Hz, 1–2% lift) leads to the impression of a soprano voice (Voigt and Reuter (1998), 18–20). Thus, the fundamental together with the 1st overtone is enough to produce voice similar tones. Overall, the fundamental and the 1st overtone appear to be important candidates for the separation of high an low register for voices.

Sopranos nearly always use head voice with strong fundamentals, basses nearly always chest voice. Altos and Tenors change between the two types of register, which leads to errors in register prediction. Therefore, overlap of registers occur for altos and tenors, and these voice cannot be attached to only one type of register in the case of individual tones.

Most music instruments are too small for a strong production of their lowest fundamentals. Thus, the fundamental has the more mass the higher

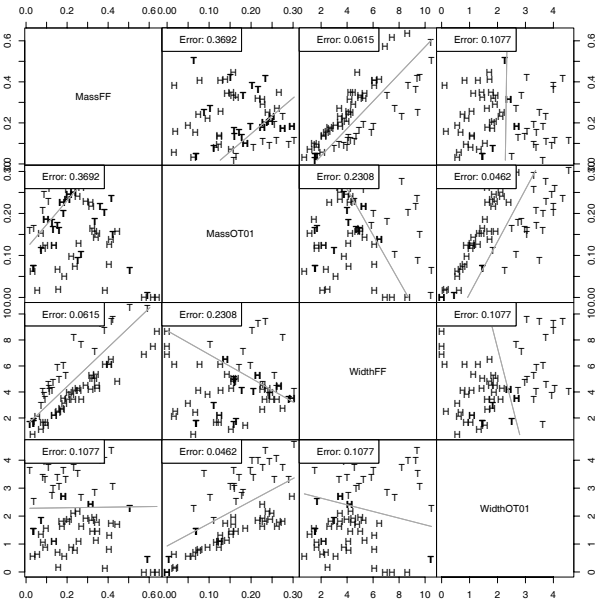


Fig. 3. Scatterplot matrix of MassFF, MassOT01, WidthFF, WidthOT01 with class separating lines and apparent error rates for voices and instruments based on averaged characteristics (H = high, T = low).

the tone is, and a strong fundamental relative to the 1st overtone indicates a high register for music instruments. The most problems occurred with French horn and Marimba. However, comparing French horn and Bassoon in their low register both instruments have similar spectral properties, e.g. a strong formant area 300–500 Hz. For both instruments the fundamental reaches the formant area with increasing pitch, however, slowly for the French horn, and abruptly for the Bassoon (Reuter (2002), 263, 327). Thus, the change between a strong fundamental and a strong 1st overtone is more exact for the Bassoon, leading to a lower error rate. For the Marimba in its low register partials are not harmonic so that the impression of the fundamental is built by a residual tone not included in the spectrum (Hall (1997), 176). This causes the problems with classification. Overall, following these arguments, except for French horn and Marimba the fundamental and the 1st overtone appear to be good indicators for register.

6 Conclusion

Altogether, the found characteristics lead to astonishingly well prediction of register. Individual tones are predicted correctly in more than 90% of the cases for the sung tones, and classification is only somewhat worse if

instruments are included in the analysis. Even better, if the characteristics are averaged over all involved tones, then voice type (high or low) can be predicted without any error, and only with at most two instruments (French horn and Marimba) severe classification problems appear, French horn not being a problem when using all characteristics for classification. Thus, there are small problems with predicting the register of individual tones, but on averages the instruments can be identified as high or low nearly without problems, with the exception of at least Marimba in its Bass version.

References

BROCKWELL, P.J. and DAVIS, R.A. (1991): *Time Series: Theory and Methods*. Springer, New York.

GÜTTNER, J. (2001): *Klassifikation von Gesangsdarbietungen*. Diploma Thesis, Fachbereich Statistik, Universität Dortmund, Germany.

HALL, D.E. (1997). *Musikalische Akustik: Ein Handbuch*. Schott, Mainz

KLEBER, B. (2002): *Evaluation von Stimmqualität in westlichem, klassischen Gesang*. Diploma Thesis, Fachbereich Psychologie, Universität Konstanz, Germany

LIGGES, U., WEIHS, C. and HASSE-BECKER, P. (2002): Detection of Locally Stationary Segments in Time Series. In: W. Härdle and B. Rönz (Eds.): *COMPSTAT 2002 - Proceedings in Computational Statistics - 15th Symposium held in Berlin, Germany*. Physika, Heidelberg, 285–290.

McGill University Master Samples. McGill University, Quebec, Canada. URL: `http://www.music.mcgill.ca/resources/mums/html/index.htm`

MICHIE, D., SPIEGELHALTER, D.J. and TAYLOR, C.C. (Eds.) (1994): *Machine Learning, Neural and Statistical Classfication*. Ellis Horwood, New York

REUTER, C. (1996): *Die auditive Diskrimination von Orchesterinstrumenten - Verschmelzung und Heraushörbarkeit von Instrumentalklangfarben im Ensembleslespiel*. Peter Lang, Frankfurt/M.

REUTER, C. (2002): *Klangfarbe und Instrumentation - Geschichte - Ursachen - Wirkung*. Peter Lang, Frankfurt/M.

SCHEMINZKY, F. (1943): *Die Welt des Schalls*. Salzburg.

THERNEAU, T.M. and ATKINSON, E.J. (1997): *An Introduction to Recursive Partitioning Using the RPART Routines*. Technical Report, Mayo Foundation.

VOIGT, W. and REUTER, C. (1998): About the timbre quality in case of the Thereminvox. *Proceedings of the Russian Conference on Musicology: Organology, Petersburg, 18–20*.

WEIHS, C., BERGHOFF, S., HASSE-BECKER, P. and LIGGES, U. (2001): Assessment of Purity of Intonation in Singing Presentations by Discriminant Analysis. In: J. Kunert and G. Trenkler (Eds.): *Mathematical Statistics and Biometrical Applications*. Josef Eul, Köln, 395–410.

WEIHS, C. and LIGGES, U. (2003a): Automatic transcription of singing performances. *Bulletin of the International Statistical institute, 54th Session, Proceedings, Volume LX, 507–510*.

WEIHS, C. and LIGGES, U. (2003b): Voice Prints as a Tool for Automatic Classification of Vocal Performance. In: R. Kopiez, A.C. Lehmann, I. Wolther and C. Wolf (Eds.): *Proceedings of the 5th Triennial ESCOM Conference*. Hanover University of Music and Drama, Germany, 8-13 September 2003, 332–335.

Classification of Processes by the Lyapunov Exponent

Anja M. Busse

University of Dortmund*
Department of Statistics
44221 Dortmund, Germany

Abstract. This paper deals with the problem of the discrimination between well-predictable and not-well-predictable time series. One criterion for the separation is given by the size of the Lyapunov exponent, which was originally defined for deterministic systems. However, the Lyapunov exponent can also be analyzed and used for stochastic time series. Experimental results illustrate the classification between well-predictable and not-well-predictable time series.

1 Introduction

For the description and the analysis of time series it is useful to initially introduce a coarse classification in order to be able to choose the most appropriate tools for the more detailed analysis.

One important classification is to discriminate between well-predictable and not-well-predictable processes. Information about the predictability of a process facilitates e.g. a sensible choice of the forecasting window. In the case of chaotic time series the prediction accuracy can decrease considerably already after only a few time-steps in contrast to a stationary stochastic process (Abarbanel (1996), Casdagli (1991)).

In addition, in the analysis of stochastic processes there often is the problem that only one time series is available and no previous knowledge about the temporal-functional relationship is given.

Despite these restrictions a formal identification of predictable time series can be achieved by analyzing the Lyapunov spectrum or the largest Lyapunov exponent of the time series (this is often just referred to as the Lyapunov exponent). Originally, the Lyapunov exponent was defined for non-stochastic, deterministic systems. Anyhow, the concept behind the Lyapunov exponent can be embedded into a statistical framework.

The remainder of this paper is organized as follows. After an introduction of the Lyapunov exponent (Sec 2) we will show that it can be used as a criterion to discriminate between well-predictable and not-well predictable time series (Sec 4). Experimental results of a *BTA*-deep-hole drilling process

* This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

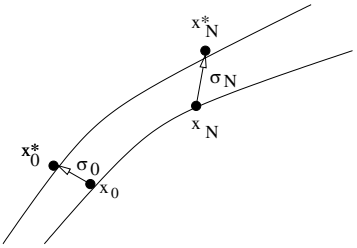


Fig. 1. Two trajectories are regarded over time in order to observe the convergence or divergence of a process.

illustrate the method of separation by the Lyapunov exponent (Sec 4). A conclusion is drawn in Sec. 5.

2 Lyapunov exponent

One possibility to distinguish between well-predictable and not-well-predictable time series is given by the computation of the largest Lyapunov exponent (often briefly called the Lyapunov exponent). This was originally defined for non-stochastic, deterministic processes. However, the Lyapunov exponent can also be analyzed and used for the stochastic case.

Firstly, it will be introduced for deterministic processes. The dynamics of deterministic processes is defined by

$$x_{t+1} = f_t(x_0) = f(x_t) \ , \tag{1}$$

with initial point or initial state $x_0 \in \mathbb{R}^k$, x_t describes the state at time t. The functional relationship is described by f and it is assumed that f is differentiable everywhere. Hence, the dynamics is entirely deterministic.

The Lyapunov exponent describes the divergence of two different trajectories. This can be motivated as follows:

In Fig. 1 the behavior of two nearby trajectories is shown. The starting point x_0^* is "nearby" but displaced from x_0. Furthermore, the trajectories follow the same functional relationship. The distance between x_0 and x_0^* is given by

$$\Delta_0 = |x_0^* - x_0|. \tag{2}$$

Hence, the distance after one iteration can be approximated by applying the first order Taylor expansion as follows:

$$\Delta_1 = |x_1^* - x_1| = |f(x_0^*) - f(x_0)| \approx |f'(x_0)| \cdot |x_0^* - x_0|. \tag{3}$$

After N iterations the distance between the trajectories arises from using the chain rule:

$$\Delta_N = |x_N^* - x_N| \approx \prod_{i=0}^{N-1} |f'(x_i)| \cdot \Delta_0 \tag{4}$$

Thus, we are interested in diverging or converging of the trajectories after N iterations in comparison to the beginning. This is estimated by an expansion rate. Obviously, the expansion rate of the trajectories can be expressed by

$$\frac{\Delta_N}{\Delta_0} \approx \prod_{i=0}^{N-1} |f'(x_i)| = e^{N \cdot \lambda_N(x_0)} \ , \tag{5}$$

where λ_N is the characteristic value dependent on time N and x_0. This expansion rate illustrates the behavior of the trajectories after N iterations in dependence of Δ_0 and x_0.

The consideration of the asymptotic behavior for $N \to \infty$ yields the definition for the Lyapunov exponent of deterministic processes:

$$\lambda(x_0) := \lim_{N \to \infty} \frac{1}{N} \sum_{i=0}^{N-1} \ln |f'(x_i)|. \tag{6}$$

It is the long time consideration of the average logarithmic derivation after N equals infinity many iterations. The Lyapunov exponent measures the asymptotic average logarithmic expansion rate along two trajectories.

The derivative f' of the function f is often unknown. It has to be evaluated from the given observation series. Various approaches for the calculation of λ have been suggested in the literature (for more details see for example Sano and Sawada (1985), Kantz and Schreiber (1997)).

If stochastic processes are considered, two cases have to be distinguished separately: The random effect is additive in the functional equation and the random effect is not necessarily additive.

First the case with an additive noise is considered. The dynamics of stochastic processes with an additive random effect is defined by

$$X_{t+1} = f(X_t) + \epsilon_t. \tag{7}$$

By transforming

$$X_{t+1} = g(X_t, \epsilon_t), \quad \text{with} \quad g(X_t, \epsilon_t) = f(X_t) + \epsilon_t \tag{8}$$

we obtain the same derivatives of g and f so that the definition of the Lyapunov exponent for stochastic processes with an additive noise is directly derived from the deterministic case. The function g is inserted in the definition of the Lyapunov exponent for deterministic processes and the definition for stochastic processes with an additive noise is obtained:

$$\lambda(X_0) := \lim_{N \to \infty} \frac{1}{N} \sum_{i=0}^{N-1} \ln || \frac{d}{dX_t} g(X_t(x_i), \epsilon_t)|. \tag{9}$$

However, an additive noise can not always be justified because this assumption is too restrictive with regard to possible model classes. Thus, the

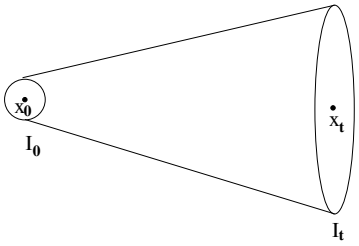


Fig. 2. Information loss of the information area I_0 in comparison to the information area I_t

general case is considered. The dynamics of stochastic processes with a non-necessarily additive noise is defined by

$$X_{t+1} = h(X_t, \epsilon_t). \tag{10}$$

The Lyapunov exponent can be naturally generalized as:

$$\tilde{\lambda}(x_0) := \lim_{N\to\infty} \frac{1}{N} \sum t = 0N - 1[\ln |\frac{d}{dX_t} h(X_t(x_0), \epsilon_t)|] \tag{11}$$

In any case, as an estimator of the Lyapunov exponent

$$\hat{\lambda} = \frac{1}{N} \sum_{i=0}^{N-1} \ln |\frac{d}{dX_t} h(X_t, \epsilon_t)|. \tag{12}$$

is taken. For details about the Lyapunov exponent for deterministic and stochastic processes see Busse et al. (2001), Busse (2003), and Busse and Weihs (2004).

3 Well-predictable and not-well-predictable processes

The knowlegde about the quality of prediction of processes is an important property for the interpretation of the predicted results. The greater the information loss in a multi-step-forcasting the greater the decrease in the quality of prediction. Thus, it is interesting to know a measure of information loss for avoiding possible misinterpretations. The Lyapunov exponent can be interpreted as an expansion rate with a direct context to the information loss over time.

If we assume that the true starting point x_0 of a time series is possibly displaced by an ϵ, we know only the information area about the starting point we do not know the proper position of x_0. After t-time steps the time series is in the information area at time t, I_t and after $t+1$-time steps in the information area I_{t+1} (Fig. 2). If the information area is small, we have more information about the true position of the data point in contrast to a greater

information area (Beck (1993)). As an adequate measure of information the information content b_n of a true position of a data point in an information area I_n of the volume Δ_n is given by:

$$b_n := \ln \frac{1}{\Delta_n} = -\ln(\Delta_n). \tag{13}$$

The connection to the volume of an information area is given by

$$\Delta_n = \exp(-b_n). \tag{14}$$

It can be characterized by the distance between two trajectories of a process at time n. For the evaluation of the quality of prediction we are interested in the information loss from one time to the next. For this the difference of two information contents before and after an iteration step are determined. Thus, the information loss IV about the true position of a data point in one iteration step is given by

$$IV = b_n - b_{n+1} = \ln \Delta_{n+1} - \ln \Delta_n \approx \ln |f'(x_n)|, \tag{15}$$

with $\Delta_{n+1} \approx |f'(x_n)| \cdot \Delta_n$. If the difference is positive, IV describes an information increase, whereas an information loss is given, if I_n is less than I_{n+1}.

The information loss is the logarithmic first derivative of the functional relationship of a process, so that the Lyapunov exponent can be used for the description of the average information loss:

$$\lambda(X_0) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \ln |f'(X_i)| \approx \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} (b_i - b_{i+1}). \tag{16}$$

In contrast to the traditional classification of time series we do not use the given data points but the possible position areas like *k-means clustering* (Hastie et al. (2001)).

The classification of both deterministic and stochastic processes by the Lyapunov exponent is given by:

- $\lambda(x_0) < 0 \Leftrightarrow \Delta_N < \Delta_0 \Rightarrow$ good predictability
  The information about the true position of the data increases due to the reduction of the information area. Consequently, we get a good predictability.
- $\lambda(x_0) \approx 0 \Leftrightarrow \Delta_N \approx \Delta_0 \Rightarrow$ predictability like a random walk
  Here, the information content levels off. We have neither information loss nor information increase.
- $\lambda(x_0) > 0 \Leftrightarrow \Delta_N > \Delta_0 \Rightarrow$ bad predictability
  The information loss about the true position of the data increases over time due to the information area increases. Consequently, we get a bad predictability.

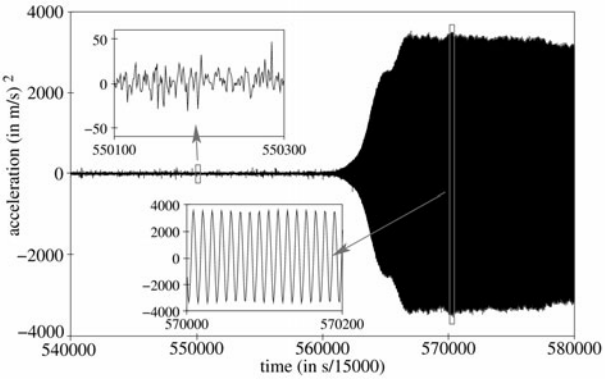


Fig. 3. Acceleration of the drill head in the non-chatter (left) and the chatter (right) area

4 Experimental results

The Lyapunov exponent achieved a distinction between well-predictable and not-well-predictable time series. We applied this classification to a real-world-problem. The aim was to analyze a *BTA*-deep-hole drilling process and to control at best working conditions (VDI (1974)). *BTA*-deep-hole drilling is used to produce holes with a high length-to-diameter-ratio. But the slenderness of the tool can yield unwanted states, like chatter. This should be avoided, because chatter generates surface discontinuity at the workpiece, noise exposure and increases the wear of cutting edges substantially.

For the analysis we were given a time series of acceleration data with different types of the process. First the non-chatter area with a weakly periodical part and the chatter area with a strongly periodical part. The transition between these areas appears funnel shaped.

The aim was to identify the chatter early to avoid the possible consequences. For this we characterize the transition in time windows of length 1024 data points. We chose the Lyapunov exponent because its ability to distinguish between good and bad predictability makes it possible to estimate the starting point of the transition. For every time window in the transition area the Lyapunov exponent of the given time series was evaluated. For this, equation (6) and equation (12) respectively is estimated by the approach of Kantz and Schreiber (1997). In order to identify the "true" transition it is important whether the Lyapunov exponent is less than 0, that is to classify a good forecast-property or whether λ is greater than 0.

For the interpretation of the resultsnote that amplitude increase is in the time window 559313–560335. The identification of the change between the non-chatter and the chatter area occured one and a half drill rotations earlier than the real amplitude increase (see Table 1, 556240–558287).

Table 1. Lyapunov exponent in time windows in the transition area.

time windows (data points)	Lyapunov exponent	classification decision
550096–551119	0.004	> 0
551120–552143	0.009	> 0
552144–553167	0.008	> 0
553168–554191	0.019	> 0
554192–555215	0.016	> 0
555216–556239	0.012	> 0
556240–557263	0.003	> 0
557264–558287	-0.004	< 0
558288–559312	-0.003	< 0
559313–560335	-0.004	< 0
560336–561359	-0.004	< 0
561360–562383	-0.013	< 0

The distinction between well predictable and not-well predictable processes by the Lyapunov exponent was applied with good results to various time series. For more details about other applications see for example Busse (2003).

5 Conclusion

We analyzed the Lyapunov exponent in the context of the separation between well-predictable and not-well-predictable processes. Such a classification seems useful since it would facilitate a more detailed analysis of the underlying process with respect to the choice of the appropriate tools. In this work the Lyapunov exponent was suggested for separation. This criterion describes the asymptotical average logarithmic expansion of the model derivative.

It was shown that the Lyapunov exponent can be used for the evaluation of predictability. The Lyapunov exponent as a classification criterion can be used without knowledge stochastics of the process and without knowledge about the temporal-functional relationship, only using the given time series.

In addition, different areas of a *BTA*-deep-hole drilling process were classified by the Lyapunov exponent. Detection of the transition to chatter was possible substantially earlier than the rise in acceleration was visible.

References

ABARBANEL, H. D. I. (1996): *Analysis of Observed Chaotic Data*. Institute for Nonlinear Science. Springer Verlag, New York.

ARNOLD, V. I. and AVEZ, A. (1968): *Ergodic problems of classical mechanics*. W. A. Benjamin, New York.

BECK, C. and SCHLÖGL, F. (1993): *Thermodynamics of chaotic systems*. Cambridge University Press, Cambridge.

Busse, A. M., STEUER, D. and WEIHS, C. (2001): An Approach for the Determination of Predictable Time Series. *Technischer Bericht 12, SFB 475, Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany*.

BUSSE, A. M. (2003): *Klassifikation von Datenreihen mit Hilfe des Lyapunov–Exponenten*. Dissertation Fachbereich Statistik. Universität Dortmund. URL http://eldorado.uni-dortmund.de:8080/FB5/ls7/forschung/2003/Busse.

BUSSE, A. M. and WEIHS, C. (2004): Lyapunov exponent for stochastic time series. *Technischer Bericht 37, SFB 475, Fachbereich Statistik, Universität Dortmund, 44221 Dortmund, Germany*.

CASDAGLI, M. (1991): Chaos and deterministic versus stochastic non-linear modeling. *J. R. Statist. Soc. B, 54, 303–328*.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, New York.

KANTZ, H. and SCHREIBER, T. (1997): *Nonlinear time series analysis*. Cambridge University Press, Cambridge.

SANO, M. and SAWADA, Y. (1985): Measurements of the lyapunov spectrum from a chaotic time series. *Physical Review Letters, 55, 1082–1085*.

STOUT, W. F. (1974): *Almost sure convergence*. Academic Press, New York.

VDI-Richtlinie 3210 (1974, Juni). Tiefbohrverfahren. VDI, Düsseldorf.

Desirability to Characterize Process Capability

Jutta Jessenberger and Claus Weihs*

University of Dortmund**
Department of Statistics
44221 Dortmund, Germany

Abstract. Over the past few years continuously new process capability indices have been developed, most of them with the aim to add some feature missed in former process capability indices. Thus, for nearly any thinkable situation now a special index exists which makes choosing a certain index as difficult as interpreting and comparing index values correctly.
In this paper we propose the use of the expected value of a certain type of function, the so-called desirability function, to assess the capability of a process. The resulting index may be used analogously to the classical indices related to C_p, but can be adapted to nearly any process and any specification. It even allows a comparison between different processes regardless of their distribution and may be extended straightforwardly to multivariate scenarios. Furthermore, its properties compare favorably to the properties of the "classical" indices.

1 Introduction

The amount of indices developed in recent years has lead to insecurity among the practitioners which index to use. The deficits of the classical indices related to the C_p-family have been extensively discussed (Kotz and Johnson (1993), Jessenberger (1999)) and have led to the continuous development of new indices which are custom-designed to eliminate these deficits one after the other. A special problem arises with the interpretation of all these index values. Most often the index values are associated with the percentage of conforming or non-conforming (NC) product. However, this is only true if implicit assumptions are valid. A C_p-value of 1 will only indicate a percentage of NC product of 0.27% if and only if the specification is symmetric, the process is normally distributed, the mean equals the specification midpoint and the process is under statistical control. In practice, the validity of these assumptions often is not verified. If one additionally takes into account that most capability values are estimates rather than true values and that the estimators of nearly all process capability indices are biased it is understandable that some authors propose to stop using process capability indices at all

** The work of Claus Weihs has been partly supported by the Collaborative Research Centre "Reduction of Complexity in Multivariate Data Structures" (SFB 475) of the German Research Foundation (DFG). The simulations were run on personal computers using the software S-PLUS 4.3 (Statistical Sciences Inc., Seattle). The simulation programs are available through the first author.

* e-mail: weihs@statistik.uni-dortmund.de

(Pignatiello and Ramberg (1993)).
However, in this paper we will propose another index which overcomes the above-mentioned problems and has the properties needed to assess the quality of a process: ease of interpretation, validity for all specification types, flexibility with respect to process distributions, and existence of a good estimator.
In the following we will present the new index for uni- and multivariate processes, and compare its properties with the main classical indices in the bivariate case. As all classical indices are dependent on the validity of the normal assumption we will only discuss the new index for normal processes although the extension for other distributions is straightforward. For quality characteristics X we will assume a univariate or multivariate normal distribution (with p dimensions) with mean (vector) μ and variance σ^2 / covariance matrix Σ , denoted by $X \sim \mathrm{N}(\mu, \sigma^2)$ and $X \sim \mathrm{Np}(\mu, \Sigma)$, respectively. In the multivariate case often a process ellipsoid is used to characterize the process properties. This is defined to be the ellipsoid which contains a certain percentage (usually 99.73%) of the process distribution. In the univariate case the ellipsoid collapses to an interval containing the desired percentage of the process distribution.
For the univariate case, specifications consist of a target value T and lower and/or upper specification limits (LSL, USL). If the target value lies on the midpoint $m := (LSL+USL)/2$ of the specification interval, the specification is called a (two-sided) symmetric specification, else a (two-sided) asymmetric specification. If either the lower or the upper specification limit is infinite while still retaining the nominal optimal target value, the specification is called one-sided. Multivariate specification is described by the Cartesian product of the univariate specifications and denoted by (M1). Frequently, ellipsoids are used as multivariate specification regions which are typically the largest-volume ellipsoids completely contained in (M1), denoted by (M2).
In this paper we will develop uni- and multivariate indices, but restrict ourselves to the discussion of the multivariate case because of space restrictions. For comparison we will use the most common classical multivariate indices. A multivariate analogue of the univariate C_p-Index (Taam et al. (1993)) is given by:

$$\mathrm{MVC}_p := \frac{\text{vol(max. vol. ellipsoid in specification)}}{\text{vol(process ellipsoid)}} = \left(\frac{|A|}{|\Sigma|}\right)^{1/2} \left(\frac{1}{\chi^2_{p;0.9973}}\right)^{p/2}.$$

$\chi^2_{p;0.9973}$ denotes the 99.73%-quantile of the χ^2-distribution with p degrees of freedom and $A = diag(d_1^2, ..., d_p^2)$, $d_j := (USL_j - LSL_j)/2, j = 1, ..., p$, is the specification matrix defining the specification ellipsoid (M2) given by $\{x|(x-m)'A^{-1}(x-m) \leq 1\}$, $m := (m_1...m_p)'$, $m_i := (USL_i + LSL_i)/2$.
The multivariate analogue of the C_{pm}-Index (Taam et al. (1993)) additionally includes the Mahalanobis distance between the mean and the target vector to measure a possible deviation of the mean from the target:

$$\begin{aligned}\mathrm{MVC}_{pm} :&= \frac{\text{vol(max. vol. ellipsoid in specification)}}{\text{vol}((x-T)'\Sigma_T^{-1}(x-T) \le \chi^2_{p;0.9973})} \\ &\quad \text{where} \quad \Sigma_T := E[(X-T)(X-T)'] = \Sigma + (\mu - T)(\mu - T)' \\ &= \left(\frac{|A|}{|\Sigma|}\right)^{1/2} \left(\frac{1}{\chi^2_{p;0.9973}}\right)^{p/2} \Big/ \sqrt{1 + (\mu - T)'\Sigma^{-1}(\mu - T)}\,.\end{aligned}$$

2 Combining capability and desirability - the indices EDU and EDM

Desirability indices were invented in experimental design to summarize several response variables and thus identify the direction of optimization (Derringer and Suich (1980)). Typically, several - possibly contradicting - response variables have to be optimized simultaneously where each response can be modeled as a (different) function of a common set of predictors. The aim is to have each response approach as much as possible their target optimum value while at the same time ensuring that the overall result still is unacceptable if only one of the responses attains an unacceptable value.
Derringer and Suich (1980) propose to transform the responses to so-called "desirability values" between 0 and 1, which takes the value 1 if the quality characteristic attains the target value and decreases if it deviates from the target. Undesirable values have the desirability 0. Typically, for a two-sided specification with target T and lower and upper specification limits LSL and USL one would choose a desirability function as follows:

$$\begin{aligned}&D_{r;s} : \mathrm{IR} \to [0,1], \\ &\qquad x \mapsto D_{r;s}(x) := \begin{cases} (x - LSL)^r/(T - LSL)^r & \text{, for } x \in [LSL, T] \\ (USL - x)^s/(USL - T)^s & \text{, for } x \in [T, USL] \\ 0 & \text{, else.} \end{cases}\end{aligned}$$

where $r, s \in \mathrm{IR}$ are suitably chosen constants to reflect how rapidly a deviation from the target becomes undesirable. For one-sided specifications the idea is easily extended. Usually in one-sided specifications there exists a point beyond which desirability improves only marginally and thus is defined to be constant 1.
The desirability index is then defined as the geometric mean of the desirability functions in each dimension (cf. Harrington (1965), and Derringer and Suich (1980)). In this paper we will use a different approach and define the indices EDU and EDM as the expected desirability for a given process. The EDU-Index (expected desirability, univariate) is defined as:

$$EDU := E(D(X)).$$

Analogously, the multivariate index EDM (expected desirability, multivariate) is given as

$$EDM := E(D_{MV}(X)),$$

where $D(x)$ and $D_{MV}(x)$ are suitable univariate and multivariate desirability functions.
This construction has the advantage that the EDU and EDM index values in principle can be calculated regardless of the distribution of the process and regardless of the specification, as long as the expectation over the desirability function exists.
Through the custom-designed desirability function the practitioner gains flexibility as any shape and structure of the specification region can be modeled through this function. Taking the expectation even allows differently distributed processes to be compared directly with each other whereas for the classical indices normality must hold.
The explicit form of the EDU-Index with a linear desirability function D_1 for a normal distribution is given in the following (cf. Jessenberger (1999)) as an example.
Let $X \sim \mathrm{N}(\mu, \sigma^2)$ with density function f and distribution function F, (LSL, T, USL) a two-sided specification and D_1 the linear desirability function $(r = s = 1)$.
Let $\delta := (\mu - T)/d$, $\eta := \sigma/d$, $\beta := (T - m)/d$, $d := (USL - LSL)/2$, $m := (USL + LSL)/2$, $a := (-(\delta + \beta) - 1)/\eta$, $b := (-(\delta + \beta) + 1)/\eta$. Then EDU is given as:

$$\begin{aligned} EDU = \delta & \left[\frac{2}{1-\beta^2} \Phi\left(\frac{-\delta}{\eta}\right) - \frac{\Phi(a)}{1+\beta} - \frac{\Phi(b)}{1-\beta} \right] \\ & - \eta \left[\frac{2}{1-\beta^2} \varphi\left(\frac{-\delta}{\eta}\right) - \frac{\varphi(a)}{1+\beta} - \frac{\varphi(b)}{1-\beta} \right] - \Phi(a) + \Phi(b) \end{aligned}$$

where φ and Φ denote the standard normal density and distribution function, respectively.

The multivariate desirability function D_{MV} is defined as follows:

$$D_{MV}(x_1, x_2, ..., x_p) := \min(D_1(x_1), D_2(x_2), ..., D_p(x_p)),$$

where $D_i(x_i)$ are desirability functions for $X_i, i = 1, ..., p$.
This also shows an obvious way of finding the distribution of D_{MV}. Thus, D_{MV} is defined as a non-standard desirability index in that the univariate desirability functions are joined via a minimum function and not via the more usual geometrical mean (cp. Kim and Lin (2000)). For the explicit expression of EDM in the bivariate case with linear desirability functions and normality see Jessenberger (1999).

In this paper we will concentrate on the comparison of EDM with the classical indices.

3 Discussion

Bivariate normal processes will be used to illustrate the performance of the EDM-Index. Table 1 gives the variances $\sigma_1^2 = \sigma_2^2$, covariances σ_{12} and correlations ρ of the examined processes A, B, C and D. The two quality characteristics of the processes A and C are highly correlated, whereas the variables for processes B and D are uncorrelated.

Table 1. Example processes

Process	A	B	C	D
$\sigma_1^2 = \sigma_2^2$	50	15	15	10
σ_{12}	49	0	14	0
ρ	0.98	0	0.9333	0

For both quality characteristics a symmetric univariate specification of $(LSL, T, USL) = (35, 50, 65)$ is assumed. The Cartesian product of the univariate specifications will be denoted by (M1) = (35,50,65)×(35,50,65), the specification given by the largest-volume ellipsoid will be denoted by (M2). Furthermore, the behavior for on-target and off-target processes will be examined. For on-target processes μ equals the target value: $\mu = T = (50, 50)'$, for off-target processes the process mean is moved into the direction of the bottom-left corner of the specification area: $\mu = (40, 40)' \neq T$. Figure 1 shows process and specification ellipses for on- and off-target comparison. Table 2

Table 2. Example processes

Process	$1 - q$ $(\mu = T)$	MVC_p	EDM	$1 - q$ $(\mu = (40, 40)')$	MVC_{pm}	EDM
A	0.175	1.912	0.595	0.171	1.100	**0.338**
B	0.466	1.268	0.709	0.340	0.335	0.211
C	0.612	**3.532**	0.760	**0.513**	**1.257**	0.311
D	**0.751**	1.902	**0.762**	0.470	0.415	0.224

shows the percentage of conforming product and the process capability index (PCI) values for MVC_p, MVC_{pm} and EDM for the four example processes, the highest values indicating the best processes are marked in bold type. For all indices except EDM the processes C or D are the best. This result seems to be intuitively sensible because these are the processes with the smallest variation. For MVC_p and MVC_{pm} even the ranking of the processes A to D is the same: The best process is the process with the smallest variation and highest correlation (process C) and the worst process is the process with largest variation and without correlation (process B). The reason for the good performance of process C is that the volume of the corresponding process ellipsoid is much smaller than the specification ellipsoid. However, especially in the case of process A (ranked second for both classical indices) this ignores

the fact that a high percentage of the product produced with process A lies outside the specification limits and is thus not or only partially fit for use.
The process ranking according to the EDM index is different. For the on-target case the ranking of the processes by the EDM index is analogous to the percentage of conforming product 1-q. Thus process D is the best and process A the worst process according to EDM. For the off-target case the situation changes. The highly correlated processes A and C are preferred to the processes without correlation (B and D). It is intuitively clear that this is because the deviation of the process means coincides with the direction of correlation for processes A and C so that a larger percentage of the distribution is close to the target. With the same argument A is preferred to C. Overall it can be said that the EDM index - per definition -prefers processes that are "on average close to target".

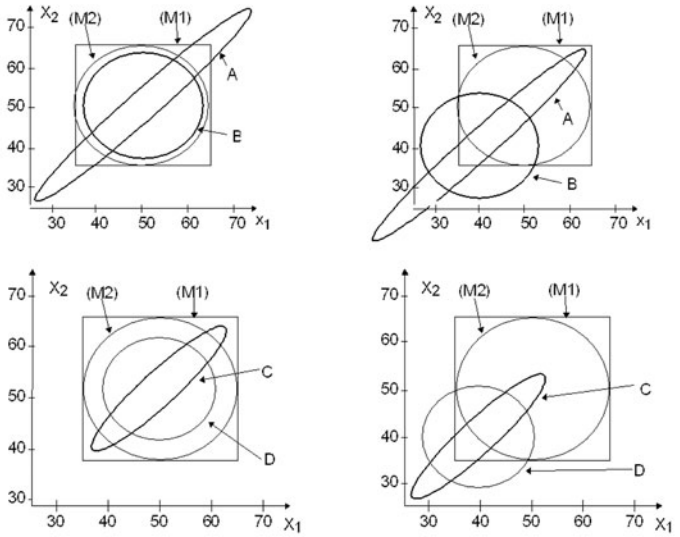


Fig. 1. Specification (M1), process and specification ellipses (M2) for bivariate processes

4 Estimation

Let us concentrate again on the bivariate case. Let $\delta := D(\mu - T)$, $H := D\Sigma D$ and $\beta := D(T - m)$, where $D := \text{diag}(1/d_1, 1/d_2)$, $m := (m_1\, m_2)'$, $m_i := (USL_i + LSL_i)/2$ and $d_i := (USL_i - LSL_i)/2$, $i = 1, 2$. With this, MVC_p, MVC_{pm} may be written as functions of δ, η or H, respectively, EDU was expressed in analogous terms above.
For the estimation of MVC_p and MVC_{pm} estimates of the transformed mean and variation are inserted into the functional form and the resulting value

is used as an estimate for each index. For the estimation of the EDM-Index two approaches are considered. If the functional form of EDM is known, it is possible to insert estimates instead of the unknown distribution parameters expectation and variation (plug-in estimator). Secondly, estimates can be achieved by the average of individually determined desirability values:
Let X, $X_i := (X_{1i}, X_{2i})' \sim \mathrm{N}_2(\delta, H)$ and $X, X_1, ..., X_n$ independently identically distributed variables. Further, let $\hat{\delta} := (\hat{\delta}_1, \hat{\delta}_2)'$ with
$\hat{\delta}_j := (X_{j1}, \ldots, X_{jn})/n$, $j = 1, 2$, and $\hat{\eta}_j^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_{ji} - \hat{\delta}_j)^2$,
$\hat{\eta}_{12}^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_{1i} - \hat{\delta}_1)(X_{2i} - \hat{\delta}_2), \quad \hat{\rho} = \hat{\delta}_{12}/(\hat{\eta}_1 \hat{\eta}_2)$.
Then two possible estimators of the EDM-Index are given as:

1. $\hat{\mathrm{EDM}} := \mathrm{EDM}(\hat{\delta}_1, \hat{\delta}_2, \hat{\eta}_1, \hat{\eta}_2, \hat{\rho})$ and
2. $\hat{\mathrm{DMV}} := \frac{1}{n} \sum_{i=1}^{n} D_{MV}(X_{1i}, X_{2i})$.

From the functional form of the estimator it is clear that it is unbiased and asymptotically normally distributed.

5 Simulation

A common criterion to compare the performance of estimators is the mean squared error MSE. However, the MSE is heavily scale-dependent. As the values of the "classical" indices are not bounded from above as is EDU/EDM, to compare the performance of the estimators for EDU/EDM with the estimators for the "classical" indices it is necessary to standardize the MSE by the magnitude of the estimated quantity:
Let $\mathrm{MSE}(\hat{\theta})$ denote the mean squared error of the estimator $\hat{\theta}$ for a statistic $\theta > 0$ or $\theta < 0$. Then the standardized MSE, $\mathrm{MSE}_{st}(\hat{\theta})$, is defined as $\mathrm{MSE}_{st}(\hat{\theta}) := \mathrm{MSE}(\hat{\theta})/\theta^2$.
For the bivariate case combinations of the values -0.5, 0, 0.5 for β_1, β_2, δ_1, δ_2 and ρ and combinations of the values 0.1 and 1.1 for η_1 and η_2 have been examined. The number of random variates used for estimating the location and variation parameters is n = 50, the number of repetitions N = 1000 which was sufficient for a good precision of the simulation (cf. Jessenberger (1999)). Due to space restrictions we give summary results rather than all detailed results which can be found in the above-mentioned literature. $\hat{\mathrm{EDM}}$ and $\hat{\mathrm{DMV}}$ were shown to be better than $\hat{\mathrm{MVC}}_{pm}$ in terms of the maximal values of MSE_{st}. Moreover, the MSE_{st} for $\hat{\mathrm{EDM}}$ and $\hat{\mathrm{DMV}}$ are maximal for small values of the distribution parameters. If the variation and correlation increases the mean standardized error decreases and even the maximum MSE_{st} is reduced by half or more. In contrast, for $\hat{\mathrm{MVC}}_{pm}$ the maximum value of MSE_{st} may be attained throughout all considered combinations of η_1, η_2 and ρ. Only the spread of the values decreases. With regard to the comparison between $\hat{\mathrm{EDM}}$

and $\hat{\text{DMV}}$, the former has smaller MSE_{st} than $\hat{\text{DMV}}$ for all simulations. Thus overall, $\hat{\text{EDM}}$ is the best estimator among the estimators considered.

6 Conclusion

In this paper we have presented a new index for assessing process capability which is based on the expected value of desirability functions. These desirability functions assign a "desirability value" to each value a quality/process characteristic may take. An average desirability of a process may then be used as a measure for process capability.
The proposed approach is feasible for any given specification and distribution and allows a wide range of processes to be compared directly. A comparison of the newly proposed EDM index with the classical multivariate analoga of C_p and C_{pm} shows that the new index compares favorably. Moreover, in choosing different desirability functions EDU/EDM-indices offer a good chance to reflect virtually every specification region as long as the corresponding expectation exists. Furthermore, regardless of underlying process distribution the interpretation of the index values is always the same so that processes following different distributions may be compared directly. Simulation studies show that an obvious estimator for EDM exhibits equally good or better behavior than the usual estimators for the classical indices.

References

DERRINGER, G.C. and SUICH, R. (1980): Simultaneous optimization of several response variables. *Journal of Quality Technology, 12, 337-45.214-219.*

HARRINGTON JR., E.C. (1965): *The desirability function. Industrial Quality Control, 21, 494–498.*

JESSENBERGER, J. (1999): *Prozeßfähigkeitsindizes in der Qualitätssicherung.* Ph.D. Thesis, Department of Statistics, University of Dortmund.

KIM K.-J. and LIN, D.K.J. (2000): Simultaneous optimization of mechanical properties of steel by maximizing desirability functions. *Applied Statistics, 49(3), 311–326.*

KOTZ, S., and JOHNSON, N.L. (1993): *Process Capability Indices.* Chapman & Hall, London.

PIGNATELLO, J.J. and RAMBERG, J.S. (1993): Process Capability Indices: Just say "No!". *ASQC Quality Congress Transactions, 92–104*, Boston.

TAAM, W., SUBBAIAH, P. and LIDDY, J.W. (1993): A note on multivariate capability indices. *Journal of Applied Statistics, 20(3), 229–351.*

Application and Use of Multivariate Control Charts in a BTA Deep Hole Drilling Process

Amor Messaoud, Winfried Theis, Claus Weihs, and Franz Hering

University of Dortmund*, Department of Statistics, 44221 Dortmund, Germany

Abstract. Deep hole drilling methods are used for producing holes with a high length-to-diameter ratio, good surface finish and straightness. The process is subject to dynamic disturbances usually classified as either chatter vibration or spiralling. In this paper, we will focus on the application and use of multivariate control charts to monitor the process in order to detect chatter vibrations. The results showed that chatter is detected and some alarm signals occur at time points which can be connected to physical changes of the process.

1 Introduction

Deep hole drilling methods are used for producing holes with a high length-to-diameter ratio, good surface finish and straightness. For drilling holes with a diameter of 20 mm and above, the BTA (Boring and Trepanning Association) deep hole machining principle is usually employed. The process is subject to dynamic disturbances usually classified as either chatter vibration or spiralling. Chatter leads to excessive wear of the cutting edges of the tool and may also damage the boring walls. Spiralling damages the workpiece severely. The defect of form and surface quality constitutes a significant impairment of the workpiece. As the deep hole drilling process is often used during the last production phases of expensive workpieces, process reliability is of primary importance and hence disturbances should be avoided. For this reason, process monitoring is necessary to detect dynamic disturbances.

In this work, we will focus on chatter which is dominated by single frequencies, mostly related to the rotational eigenfrequencies of the boring bar. Therefore, we propose to monitor the amplitude of the relevant frequencies in order to detect chatter vibration as early as possible. In practice, it is necessary to monitor several relevant frequencies because the process is subject to different kind of chatter (i. e., chatter at the beginning of the drilling process, high and low frequency chatter). The first idea is to monitor each relevant frequency separately, using a proposed univariate control chart. This strategy is discussed in section 2. Another solution is to use a multivariate control chart to monitor several relevant frequencies simultaneously, which is

* The work of Winfried Theis and Claus Weihs has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

investigated in section 3. In section 4, the different control charts are applied to real data.

2 Monitoring the process using multiple Residual Shewhart control charts

Weinert et al. (2002) used the van der Pol equation to describe the transition from stable operation to chatter in one frequency

$$\frac{d^2M(t)}{dt^2} + h(t)(b^2 - M(t)^2)\frac{dM(t)}{dt} + w^2M(t) = W(t), \tag{1}$$

where $t \in [0, \infty)$, $M(t)$ is the drilling torque, $b \in \mathbb{R}$, the frequency $w \in$ [200,2500], $h(t) : \mathbb{R} \to \mathbb{R}$ is an integrable function and $W(t)$ is a white noise process. Theis (2004) described the main features of the variation of the amplitudes of the relevant frequencies, using a logistic function. He showed that his approximation is directly connected to the proposed model. In fact, he considered $M(t)$ as a harmonic process

$$M(t) = R(t)cos(w + \phi),$$

where ϕ is the corresponding phase. He showed that

$$2\frac{dR(t)}{dt} + h(t)R(t)\left(b^2 - \frac{R(t)^2}{2}\right) = \frac{W(t)}{w}. \tag{2}$$

is the amplitude-equation for the differential equation in (1) if there is only one frequency present in the process. From equation (2), the observed variation in amplitude of the relevant frequencies may be described by

$$R_t = (1 + a_t)R_{t-1} - a_tb_tR_{t-1}^3 + \varepsilon_t, \tag{3}$$

where a_t and b_t are time varying parameters and ε_t is normally distributed with mean 0 and variance σ_ε^2. Messaoud et al. (2004a) used the autoregressive part of equation (3) to monitor the variation of the amplitude of the relevant frequencies of the process using residual control charts. We showed that the variation in amplitude of the relevant frequencies of the process can be approximated by the autoregressive AR(1) model when the process is stable and that the nonlinear term $-a_tb_tR_{t-1}^3$ is not important before chatter.

For the monitoring procedure, the AR(1) model is used to calculate the residuals. The idea behind residual control charts is if the AR(1) model fits the data well, the residuals will be approximately uncorrelated. Then, traditional control charts, such as Shewhart chart can be applied to the residuals. A window of the m recent observations is used to estimate parameters a, β and σ_ϵ of the linear regression model

$$R_t = \beta + (1 + a)R_{t-1} + \epsilon_t,$$

where β is the mean of the autoregressive process. Note that β is included because there is a general shift in the amplitudes after depth 35 mm due to a change in the physical conditions of the process, see section 4.2. The residuals are calculated using

$$e_t = R_t - (1 + \hat{a}_{t-1})R_{t-1} - \hat{\beta}_{t-1}, \tag{4}$$

where $\hat{a}_{t-1}$ and $\hat{\beta}_{t-1}$ are estimates of the regression parameters a and β at time $t-1$. The lower and upper control limits LCL and UCL, respectively are given by

$$LCL = -k\sigma_{\epsilon,t-1} \text{ and } UCL = k\sigma_{\epsilon,t-1},$$

where $\sigma_{\epsilon,t-1}$ is the estimated standard deviation of the residuals at time $t-1$ and k is a constant. The residual Shewhart control chart operates by plotting residuals e_t given by equation (4). It signals that the process is out-of-control when e_t is outside UCL or LCL.

In order to monitor several relevant frequencies, the residual Shewhart is used to monitor the variation in amplitude of each relevant frequency, w, separately. The resulting monitoring strategy signals an out-of-control condition when any univariate control chart produces an out-of-control signal.

3 Monitoring the process using multivariate control charts

3.1 Data depth

Data depth measures how deep (or central) a given point $\mathbf{X} \in \mathbb{R}^d$ is with respect to (w.r.t.) a probability distribution F or w.r.t. a given data cloud $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$. There are several measurements for the depth of the observations, such as Mahalanobis depth, the simplicial depth, half-space depth, and the majority depth of Singh, see Liu et al. (1999). In this work, the Mahalanobis depth and simplicial depth are considered.

1. The Mahalanobis depth (MD_F) of a given point $\mathbf{X} \in \mathbb{R}^d$ w.r.t. F is defined to be

$$MD_F(\mathbf{X}) = \frac{1}{1 + (\mathbf{X} - \mu_F)' \Sigma_F^{-1} (\mathbf{X} - \mu_F)},$$

where μ_F and Σ_F are the mean vector and dispersion matrix of F, respectively. The sample version of MD_F is obtained by replacing μ_F and Σ_F with their sample estimates. In fact, how deep $\mathbf{X}$ is w.r.t. F is measured by how small its quadratic distance is to the mean.

2. The simplicial depth (SD_F) (Liu (1990)) of a given point $\mathbf{X} \in \mathbb{R}^d$ w.r.t. F is defined to be

$$SD_F(\mathbf{X}) = P_F\{\mathbf{X} \in s[\mathbf{Y}_1, \ldots, \mathbf{Y}_{d+1}]\},$$

where $s[\mathbf{Y}_1, \ldots, \mathbf{Y}_{d+1}]$ is a d-dimensional simplex whose vertices are random observations $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_{d+1}\}$ from F. The sample simplicial depth $SD_{F_m}(\mathbf{X})$ is defined to be

$$SD_{F_m}(\mathbf{X}) = \binom{m}{d+1}^{-1} \sum_{1 \le i_1 < \cdots < i_{d+1} \le m} I\left(\mathbf{X} \in s[\mathbf{Y}_{i_1}, \ldots, \mathbf{Y}_{i_{d+1}}]\right),$$

where $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ is a random sample from F, F_m denotes the empirical distribution of $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ and $I(.)$ is the indicator function. For example, the bivariate $SD_{F_m}(\mathbf{X})$ relative to $\{\mathbf{Y}_1, \ldots, \mathbf{Y}_m\}$ is equal to the proportion of closed triangles with vertices $\mathbf{Y}_i$, $\mathbf{Y}_j$, $\mathbf{Y}_k$ that contain $\mathbf{X}$, $1 \le i < j < k \le m$. Liu (1990) showed that if F is absolutely continuous, then as $m \longrightarrow \infty$, SD_{F_m} converges uniformly and strongly to $SD_F(\mathbf{X})$ and that $SD_F(\mathbf{X})$ is affine invariant.

3.2 A control chart based on sequential rank of data depth measures

Liu (1995) was the first who used the concept of data depth to construct a nonparametric control chart for monitoring processes of multivariate quality measurements. Messaoud et al. (2004b) considered an EWMA chart based on sequential ranks of data depth measures to monitor multivariate processes. The proposed chart is a generalization of the nonparametric EWMA for individual observations proposed by Hackl and Ledolter (1992).

For this chart, the depth of $\mathbf{X}_t$ is calculated w.r.t. a reference sample considered as the $m > 1$ most recent observations taken from the process $\mathbf{X}_{t-m+1}, \mathbf{X}_{t-m+2}, \ldots, \mathbf{X}_t$. That is, this sample will be used to decide whether or not the process is still in control at time t.

The sequential rank S_t^* is the rank of $D_m(\mathbf{X}_t)$ among $D_m(\mathbf{X}_{t-m})$, $\ldots$, $D_m(\mathbf{X}_{t-1})$. That is,

$$S_t^* = 1 + \sum_{i=t-m}^{t-1} I(D_m(\mathbf{X}_t) > D_m(\mathbf{X}_i)),$$

where $I(.)$ is the indicator function. For tied observations, the authors used the midrank, see Gibbons and Chakraborti (1992). In fact, the simplicial depth is a discrete measure and ties may occur. Especially, there always exist at least $(d+1)$ extreme points that share the minimum simplicial depth of

$(d+1)/m$, see Stoumbos and Reynolds (2001). The standardized sequential rank $S_t^{(m)}$ is defined as

$$S_t^{(m)} = \frac{2}{m}\left(S_t^* - \frac{m+1}{2}\right).$$

The control statistic T_t is the exponentially weighted moving averages (EWMA) of standardized ranks, computed as follows

$$T_t = \min\{B, (1-\lambda)T_{t-1} + \lambda S_t^{(m)}\},$$

$t = 1, 2, \ldots$, where $B > 0$ is a reflection boundary, T_0 is a starting value, usually set equal to zero, and $0 < \lambda < 1$ is a smoothing parameter. The reflection boundary is included to prevent the EWMA from drifting to the upper side indefinitely. The process is considered in-control as long as $T_t > h$, where $h < 0$ is a lower control limit. In fact, we consider a lower one sided EWMA chart because $S_t^{(m)}$ is "higher the better". For more details, see Hackl and Ledolter (1992) and Messaoud et al. (2004b).

4 Application

The proposed monitoring procedures are used to jointly monitor the amplitudes of frequencies 234 Hz and 703 Hz, which are among the eigenfrequencies of the boring bar, in an experiment with feed $f = 0.185$ mm, cutting speed $v_c = 90$ m/min and amount of oil $\dot{V}_{oil} = 300$ l/min. For more details, see Weinert et al. (2002).

4.1 Choice of the control charts parameters

Traditionally, a reference sample of 100-200 observations is used in SPC applications, see Montgomery (1996). In this work, the $m = 100$ recent observations $\mathbf{R}_{t-m}, \ldots, \mathbf{R}_{t-1}$, where $\mathbf{R}_t = (R_{t,234}, R_{t,703})'$ are used to estimate the parameters of the two AR(1) models and to calculate the residuals. Furthermore, they are used to calculate the data depth. $R_{t,234}$ and $R_{t,703}$ are the actual amplitudes of frequencies 234 Hz and 703 Hz, respectively. A larger sample cannot be used because the monitoring procedures should start before depth 35 mm (observation 120). In fact, chatter may be observed after that depth because the guiding pads of the BTA tool leave the starting bush, see section 4.2.

Usually, the performance of control charts are evaluated by the average run length (ARL). The run length is defined as the number of observations that are needed to exceed the control limit for the first time. The ARL should be large when the process is statistically in-control (in-control ARL) and

small when a shift has occurred (out-of-control ARL).

The parameters of the different control charts are selected so that all the charts have the same in-control ARL equal to 370. This choice should not give a lot of false alarm signals because all control charts are applied to 900 observations. A value $k = 3.205$ is used for the two residual Shewhart control charts. Note that the probability that the two charts generate a false alarm is given by

$$P_0 = 1 - (1 - p_1)(1 - p_2),$$

where p_i, $i = 1, 2$, is the probability that the ith chart produces a false alarm. The resulting in-control ARL is equal to $1/P_0$. For this formulation it is assumed that $\mathbf{R}_{234}$ and $\mathbf{R}_{703}$ are mutually independent, which may be not the case in practice.

For the EWMA chart, we used $B = -h$. Typical values of λ are in the range of $0.1 < \lambda < 0.3$, see Hackl and Ledolter (1992). In this work, we used $\lambda = 0.1$, 0.2 and 0.3. The corresponding values for h are respectively -0.314, -0.475 and -0.591. Messaoud et al. (2004b) used an integral equation to approximate the in-control ARL. The simplicial depth is computed using the FORTRAN algorithm developed by Rousseeuw and Ruts (1992).

4.2 Results

Table 1 shows the results for depth $\leq$ 270 mm. The EWMA charts based on MD_F produces more out-of-control signals than the EWMA charts based on SD_F. This is due to the sensitivity of the MD_F measure to the extreme values. Table 1 shows that all control charts signal at $32 \leq$ depth ≤ 35 mm. In fact, it is known that approximately at depth=35 mm the guiding pads of the BTA tool leave the starting bush, which induces a change in the dynamics of the process. From previous experiments, the process has been observed to either stay stable or start with chatter vibration. A great number of out of control signals occur at $35 \leq$ depth ≤ 45 mm. Indeed, the new physical state of the process is represented in the reference sample after depth 45 mm.

All control charts signal at depth $110 \leq$ depth ≤ 120 mm and it is known that depth 110 mm is approximately the position where the tool enters the bore hole completely. Theis (2004) noted that this might lead to changes in the dynamic process because the boring bar is slightly thinner than the tool and therefore the pressures in the hole may change. The important out-of-control signals are produced at $250 \leq$ depth ≤ 255 mm. Messaoud et al. (2004a) showed that a change occurred in the process at depth=252.19 mm and they concluded that this change may indicate the presence of chatter or that chatter will start in a few seconds. Therefore, in this experiment chatter may be avoided if corrective actions are taken after these signals.

Table 1. Out of control signals of the different control charts applied to the amplitude of frequencies 234 Hz and 703 Hz (m=100)

Hole Depth (mm)	Observation number	Residual Shewharts	EWMA $\lambda = 0.1$		$\lambda = 0.2$		$\lambda = 0.3$	
			MD_F	SD_F	MD_F	SD_F	MD_F	SD_F
≤32	≤107	0	0	0	0	0	0	0
32-35	108-117	2	1	1	3	1	3	1
35-45	118-150	9	29	27	21	15	13	6
45-70	151-249	2	1	0	0	0	0	0
70-110	250-366	2	9	5	3	1	1	0
110-125	370-416	1	9	10	4	4	1	1
125-200	417-665	3	3	0	2	0	2	2
200-250	666-832	6	7	8	3	2	2	1
250-255	833-849	1	4	3	4	2	4	2
255-260	850-865	0	8	7	3	3	1	1
260-270	866-898	1	4	2	0	0	0	0
Total		27	75	63	43	28	27	14

4.3 Discussion

In this experiment, the EWMA chart with λ=0.3 is the best, and should be chosen among the three EWMA charts considered in this work. Indeed, only 14 out-of-control signals are produced and all changes of the physical conditions of the process are detected. In practice, a procedure to choose the smoothing parameter λ is required.

For the process adjustment, once the EWMA chart has produced a signal, a procedure to estimate the shift magnitude and to identify the time point at which the shift occurred is required, see Messaoud et al. (2004b).
Moreover, the future research should focus on the out-of-control interpretation. In fact, when the control chart indicates an out-of-control condition, it is important to determine which frequency, or combination of frequencies, of the multivariate process caused the process to go out-of-control. In practice, the identification of the type of chatter (i.e., chatter at the beginning of the drilling process, low-high frequency chatter) will usually make it easier for engineers to adjust the process.

5 Conclusion

A main objective of this work is to investigate whether multivariate control charts can be used to monitor the drilling process. The results showed that the different control charts can detect chatter and that some out-of-control signals are related to changing physical conditions of the process (i.e., guiding

pads leave the starting bush, the tool is completely in the hole).
Multiple residual Shewhart charts assume independence and normality of the residuals, see Messaoud et al. (2004a), and in practice it is difficult to interpret multiple control charts. Multivariate control charts based on data depth are "distribution-free" control charts and are easy to visualize and interpret.

Acknowledgements

The authors would like to thank Prof. Regina Liu and two anonymous referees for their helpful comments.

References

GIBBONS, J. D. and CHAKRABORTI, S. (1992): *Nonparametric Statistical Inference*, 3rd ed. Marcel Dekker, New York.

HACKL, P. and LEDOLTER, J. (1992): A New Nonparametric Quality Control Technique. *Communications in Statistics-Simulation and Computation, 21, 423–443.*

LIU, Y. R. (1990): On a notion of data depth based on random simplices. *The Annals of Statistics, 18, 405–414.*

LIU, Y. R. (1995): Control Charts for multivariate Processes. *Journal of the American Statistical Association, 90, 1380–1387.*

LIU, Y. R., PARELIUS, J. M. and SINGH, K. (1999): Multivariate analysis by data depth: Descriptive Statistics, Graphics and Inference. *The Annals of Statistics, 27, 783–858.*

MESSAOUD, A., THEIS, W., WEIHS, C. and HERING, F. (2004a): Monitoring the BTA Deep Hole Drilling Process Using Residual Control Charts. *Technical Report 60/2004 of SFB 475, University of Dortmund.*

MESSAOUD, A., WEIHS, C. and HERING, F. (2004b): A Nonparametric Multivariate Control Chart Based on Data Depth. *Technical Report 61/2004 of SFB 475, University of Dortmund.*

MONTGOMERY, D. C. (1996): *Introduction to Statistical Quality Control*, 3rd ed. John Wiley and Sons, New York.

ROUSSEEUW, P. and RUTS, I. (1996): AS 307: bivariate location depth. *Applied Statistics, 45, 516–526.*

STOUMBOS, Z. G., JONES, L. A., WOODALL, W. H. and REYNOLDS Jr, M. R. (2001): On Shewhart-Type Nonparametric Multivariate Control Charts Based on Data Depth. In: H. J. Lenz and P. Th. Wilrich: *Frontiers in Statistical Quality Control 6*, Springer, Heidelberg, 207–227.

THEIS, W. (2004): Modelling Varying Amplitudes. *PhD dissertation, Department of Statistics, University of Dortmund.*
URL http://eldorado.uni-dormund.de:8080/FB5/ls7/forschung/2004/Theis

WEINERT, K., WEBBER, O., HÜSKEN, M., MEHNEN, J. and THEIS, W. (2002): Analysis and prediction of dynamic disturbances of the BTA deep hole drilling process. *Proceedings of the* 3^{rd} *CIRP International Seminar on Intelligent Computation in Manufacturing Engineering.*

Determination of Relevant Frequencies and Modeling Varying Amplitudes of Harmonic Processes

Winfried Theis and Claus Weihs

University of Dortmund*
Department of Statistics
44221 Dortmund, Germany

Abstract. When a process is dominated by few important frequencies the observations of this process can be modelled by a harmonic process (Bloomfield (2000)). If the amplitudes of these dominating frequencies vary over time their dominance may not be apparent during the whole process.

To discriminate between frequencies relevant for such a process we determine the distribution of the periodogram ordinates, and use this distribution to derive a procedure to assess the relevance of the frequencies. This procedure uses the standardized median (Gather and Schultze (1999)) to determine the variance of the error process. In a simulation study we show that this procedure is very efficient even under difficult conditions such as a low signal-to-noise ratio or AR(1) disturbances. Furthermore, we show that the necessary transformation to estimate the amplitudes from periodogram ordinates leads to a good normality approximation which makes it especially easy to model the development of the amplitudes from these estimates.

1 Introduction

Many processes dominated by few frequencies with varying amplitudes are well-known, e.g. music, resonance, etc.. When such a non-stationary process is observed in a noisy environment or the oscillating part of the process is obscured by an inherent stochastic process it becomes of interest to determine the really relevant frequencies. We encountered such a difficulty when investigating the BTA deep-hole drilling process and one process disturbance – called chatter – observed in this process. It turned out that chatter can be described by specific eigen-frequencies of the drilling tool bar and the development of the amplitudes of these frequencies (Weinert et al. (2002)). As long as the process stays stable the harmonic process is obscured by the noise in the process which led to the question how to determine the relevant frequencies from such data and how to model the time development of the amplitudes on these frequencies.

* This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475.

In this paper we first determine the distribution of periodogram ordinates of a harmonic process with only a few relevant frequencies, show how to incorporate this distribution to find the relevant frequencies and that there exists a normality approximation which readily facilitated constructing a model for time varying amplitudes. Finally we demonstrate the practical value of the procedure by results from an extensive simulation study.

2 Determination of the distribution of periodogram ordinates

Gallant et al. (1974) consider analysis of variance (ANOVA) models on periodograms. Their argument – based on a Taylor series extension of the distribution function – is to transform the observed ordinates with $g(x) = x^{\frac{1}{4}}$ to increase the convergence of the χ^2_2-distributed measurements to a normal distribution and thereby make a common ANOVA sensible in this situation.

The periodogram ordinate at frequency f equals n times the squared absolute value of the Fourier-transform F of the time series y_t at frequency f, that is

$$I[y_t](f) = n|F[y_t](f)|^2,$$

where n is the number of observations in the series.

If y_t is a Gaussian process with distribution $\mathcal{N}(0, \sigma^2)$, $F[y_t](f)$ as a linear transformation of y_t has again a normal distribution. $|F[y_t](f)|^2 = (\mathrm{Re}(F[y_t](f))^2 + (\mathrm{Im}(F[y_t](f))^2$ is therefore χ^2 distributed with 2 degrees of freedom, which equals an exponential distribution (cf. e.g. Fisz (1970)) with $E(|F[y_t](f)|^2) = 2\sigma_y^2$. On the basis of this argument and using the fact that the Fourier-transform is a linear operator it follows that periodogram ordinates of AR(p) processes are χ^2_{2p}-distributed.

When the amplitudes at the relevant frequencies f_k, $k = 1, \ldots, K$, of a harmonic process are influenced by some input variables $\boldsymbol{x}$ and possibly time t, it is of interest to investigate the form of this influence. So the following model is considered:

$$H_t(\boldsymbol{x}) = \sum_{k=1}^{K} h_k(\boldsymbol{x}, t) \cos 2\pi(f_k t + \varphi) + \varepsilon_t, \tag{1}$$

for $t \in \{0, \ldots, n-1\}$ and $K \ll n$. The functions of the amplitudes of the relevant frequencies are possibly time-dependent. Since only discrete time is considered, they are defined by $h_k : \mathbb{R}^d \times \mathbb{N} \to [0, \infty)$. For h_k only the existence of a Fourier-transform is assumed.

When all h_k are time constant it is clear that the expected value of the periodogram ordinates at the relevant frequencies is

$$E\big(I_{H_t(\boldsymbol{x})}(f)\big) = n(|e^{i\pi\varphi}|\, h_k(\boldsymbol{x})^2 + 2\sigma_\varepsilon^2) \text{ for } f = f_k,\ k = 1, \ldots, K. \tag{2}$$

Note that the phase is of no interest in this model because it contributes only a constant factor in the complex Fourier transform equal to $e^{i\pi\varphi}$ of which the absolute value is 1. So the phase does not contribute to the estimates described above.

If the amplitudes are slowly time-varying (i.e. slower than the smallest estimated frequency), the corresponding model in frequency domain is in terms of the complex Fourier-transform:

$$F[H_t(\boldsymbol{x})](f) = \begin{cases} F[\varepsilon](f) + B_f & \text{for } f \neq f_k, \\ F[h_k(\boldsymbol{x},t)\cos 2\pi(f_k t + \varphi)](f) + F[\varepsilon](f) & \text{for } f = f_k \end{cases}, \quad (3)$$

where $k = 1, \ldots, K$, and $B_f \neq 0$ is only true for frequencies near to one of f_k , $k = 1, \ldots, K$ and possible harmonics.

Again a result on the distribution of the periodogram ordinates is readily gained by the same arguments as above: they are χ^2-distributed. It is only close to the relevant frequencies that you get non-central χ^2-distribution with non-centrality parameter $\nu = n(h_k(\boldsymbol{x})^2 + 2\sigma_\varepsilon^2)$.

A more general determination of the distribution of periodogram ordinates can be found in Wittwer (1986). In her paper G. Wittwer determines the moment generating function and the general properties of the distribution of the periodogram ordinates for stationary sequences.

3 Regression models on periodogram ordinates

3.1 Modelling varying amplitudes

The periodogram is only able to estimate the amplitudes of Fourier frequencies, so it is of interest to know what happens when the relevant frequencies are Fourier frequencies. When the amplitudes are varying over time, we want to estimate the form of this variation. This is done by dividing the time series into sections of equal length and calculating the periodogram on these sections. Then the estimates of the amplitudes on each relevant frequency are used as objective in a – linear or nonlinear – regression to fit a proposed functional form. It can be easily proved that a linear trend in the amplitudes is transformed into a linear trend in the periodogram ordinates. When calculating the fourier transformations it turns out that using the periodogram to estimate a function of the amplitudes over time possibly underestimates the values of the function (cf. Theis (2004)).

When f_k is a non-Fourier frequency the finite Fourier transform introduces additional non-zero terms to the periodogram because it only considers Fourier frequencies. This comes from the fact that $e^{i2\pi(f_k - f)}$ is not only non-zero at the nearest Fourier frequencies but also in a neighbourhood. This has to be taken into account when deciding how many significant appearances of a frequency in an experiment are necessary to make that frequency a relevant frequency.

3.2 Estimating the variance of ε (σ_ε^2)

The Fourier-transform of a harmonic process with a small number of relevant frequencies K compared to the number of observations n can be viewed as a sample from a χ^2-distribution contaminated by some non-central χ^2-distributed observations, where the distributions have the same degrees of freedom. As remarked before the expected value of the majority of observations is $2\sigma_\varepsilon^2$, i.e. proportional to the variance of the disturbance process. A robust estimator of the expected value of this distribution is thus proportional to an estimator for the variance of the disturbance process with known proportionality factor.

Since it is assumed that in a regression situation the error processes are independent between experiments and identically distributed over all experiments, the following procedure looks promising:

1. Estimate the periodogram $I[H_t(\boldsymbol{x_l})]$ for all input values $\boldsymbol{x_l}$, $l \in \{1, \ldots, L\}$
2. Merge all $I[H_t(\boldsymbol{x_l})](f)$ into one sample
3. Calculate a robust estimator for the expected value of $I[H_t(\boldsymbol{x_l})](f)$, e.g. the standardized median $med_{st.}(X) = \frac{1}{\log(2)} med(X)$ on the merged sample

Step 2 enlarges the database for the robust estimate, because it is assumed, that the observations with different input values are independent and the realisations of $I_{H_t(\boldsymbol{x_l})}(f)$ for different Fourier frequencies are independent due to the orthogonality relations of the Fourier transform. If $K \ll n$ and $\frac{K}{n}$ is lower than breakdown point of the robust estimator, which equals $\frac{1}{2}$ for the standardized median (Gather and Schulze (1999)), one gets an estimator – in the case of Gaussian white noise – for $2\sigma_\varepsilon^2$.

Transforming Periodogram Ordinates

Given that the goal of the regression on periodogram ordinates is to estimate the influences on the amplitudes, the observations have to be transformed in the following way to get an estimator for $h_k(\boldsymbol{x})$ (cf. (2)):

$$\hat{h_k}(\boldsymbol{x}) = \sqrt{\frac{I[H_t(\boldsymbol{x})](f_k) - 2n\sigma_\varepsilon^2}{n}},$$

Johnson et al. (1994) state that this square-root of the non-central χ^2-distributed variable is a normal approximation. It depends on the value of the non-centrality parameter, which in return depends here on the value of the functions h_k , $k = 1, \ldots, K$, and the number of observations. The impact of this approximation is tested in a simulation study.

4 Simulation study on time-varying amplitudes

4.1 Design considerations

We chose a full factorial 2^7 design to compare the effects on the Normality assumption, the frequency detection, and the goodness of fits of the following

influences: the signal-to-noise ratio, the number of frequencies, number of observations, Fourier or non-Fourier frequencies, and the distance between the relevant frequencies. Additionally the effect of AR(1)-disturbances was checked.

For the influences the following values for treatment low, high, respectively were chosen: Number of frequencies (1 / 5), Fourier Frequencies (no/yes), Distance of frequencies δ_f (3 / 10 Fourier frequencies), Length of series (2560 / 102400), and Signal-to-Noise ratio (1.1 / 100). The frequencies considered are $\frac{5+\delta_f k-1}{n}$ with, in the case of non-Fourier frequencies, addition of $\frac{1}{\sqrt{2}}$ in the numerator.

The following functions were chosen as 'true' models for the variation of the amplitudes:

$$h_{lin}(t) = 2 + 0.001t \quad \text{or} \quad h_{nonlin}(t) = 2 + \frac{2}{(1 + \exp\left(\frac{m-t}{d}\right))} \tag{4}$$

The parameters m, d in equation (4) are changed for each frequency if 5 frequencies are included in the model. This is done by setting $m = 5\,l$ or $m_i = (2+i)\,l$ and $d = l$ or $d_i = \frac{l}{i}$ where $i = 1, \ldots, 5$. These different values for the parameters in the nonlinear function were chosen to test whether differing functions on the frequencies can be found in the data.

The choice of these functions had the following reasons: the first slow linear trend may be useful as an approximation for a slow nonlinear trend in the amplitude. Generally it can be assumed that amplitudes have an upper bound because oscillating systems break down when the amplitude becomes too large. This is the reason for the chosen logistic function. The inclusion of a mean intercept of 2 in both cases is done to ensure a true harmonic process right from the start of the observations.

For all settings 100 repetitions were evaluated. The function `nls` from **R** (R Core Team (2003)) was used to fit the nonlinear models. Since it is well known that nonlinear regressions tend to fail with some starting values, ten randomly chosen starting values were tested and the first successful set was used for the fit.

4.2 Results

First we checked whether the procedure to find the relevant frequencies was influenced by the time varying amplitudes, or the AR(1) disturbances. Both influences did not show an effect on the performance of the method in the sense that the correct frequencies are always found. This becomes obvious from the histograms of the found relevant frequencies in Figure 1.

The left panel in Figure 1 shows the results on relevant frequencies for experiments with the high level of observations, 10240, and high signal-to-noise ratio of 100 and five non-Fourier frequencies with a distance of ten Fourier frequencies in the simulated model that is, the true values of the frequencies are: `0.00557 0.0251 0.03487 0.04463 0.0544`. That there are more

frequencies found than just the highest peaks, is due to leakage (Bloomfield (2000)) and does not present a serious problem since it is easily possible to narrow the relevant frequencies by e.g. using higher significance levels or adding a further step where the peak(s) of the amplitudes of the found relevant frequencies is determined.

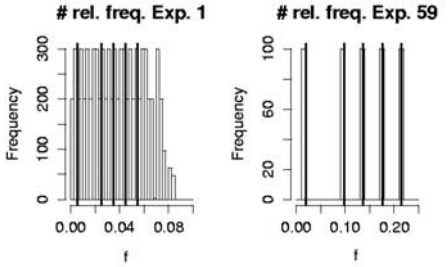


Fig. 1. The histograms of the found relevant frequencies f show clearly that at least the true frequencies (black vertical lines) are found and only a few others besides.

The right panel in Figure 1 gives an impression of the more difficult situation with only 2560 observations and AR(1) disturbances but Fourier frequencies. From this panels it is clear that at least the true frequencies are found by the method for the detection of relevant frequencies. In this case the true frequencies are: `0.0195 0.0977 0.1367 0.1758 0.2148`

The proposed normality approximation was checked for appropriateness. First we applied a Shapiro-Wilk test (Shapiro et al. (1968)) to the observations. For each true relevant frequency and each of the ten observations the 100 repetitions were collected and tested for normality on the 5%−level. The test rejected the hypothesis only in 4.86% of the cases for $h_{lin.}$, and in 5.82% of the cases for $h_{nonlin.}$. The number of rejections of normality of the observations for the normal disturbances is slightly higher than with AR(1) disturbances (linear case: 5.21% vs. 4.51%; nonlinear case: 6.04% vs. 5.63%). This was expected by the theoretical model because the goodness of the approximation is influenced by the number of stochastic components, i.e. the order of the disturbance process and the value of the non-centrality parameter. No assignable pattern was found in the rejections.

The distribution of the parameter estimates was also investigated. In the linear case the parameters displayed an even greater degree of normality. The Shapiro-Wilk test rejected only in 3.47% of the situations. In the nonlinear case it cannot be expected to find normality in the parameters. It is hard to define a distribution for the parameters in nonlinear regression, only when a linear approximation approach is chosen as fitting procedure normality is expected (cf. e.g. Ratkowski (1990, p. 20)).

Figures 2 gives an insight into the goodness of the fits of the functions on the truely relevant frequencies. In all cases it was obvious that the observations lie below the values of the true functions and therefore the fitted functions underestimate the true values as well.

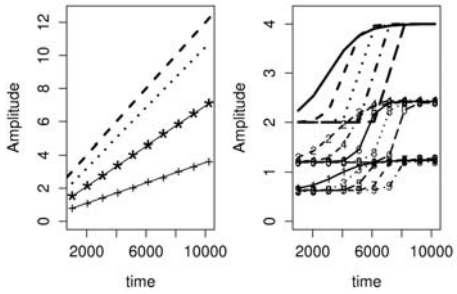


Fig. 2. Left: fitted linear functions (true: dashed line) on varying amplitudes, with AR(1) disturbances on a non-Fourier frequency, therefore two lines on the neighbouring Fourier frequencies. Right: non-linear functions of five non-Fourier frequencies, marked by numbers. True functions are again in the upper half of the graphic.

The left panel of Figure 2 shows a fit for the case of non-Fourier frequency and linear time dependence of the amplitude. This graphic gives the impression that the underestimation may be cured by summing over neighbouring frequencies in an appropriate way. This is emphasized by the dotted line which is the sum of the fitted values.

The right panel in Figure 2 underlines the previous impression as well. Furthermore, it is obvious that the general form of the influences on the amplitudes is found even if they are different for the different frequencies. This is also not influenced by the number of observations or the kind of disturbances. All fits show that the general fit of the regressions is very good which was also found when checking for the goodness of fit over all situations in the simulation study.

Studying the effect of the varying amplitudes on the performance of the proposed variance estimator, a slight overestimation of the true standard deviation σ occured. The two most important influences on the difference between the true and the estimated σ are the signal-to-noise ratio followed by the number of observations. It turns out that a high signal-to-noise ratio also leads to better estimates of the standard deviation of the disturbance term. Of course a higher number of observations leads to a better estimation since then there are more observations following the distribution of the Fourier transformation of the AR(1) or white noise normally distributed disturbances.

5 Conclusions

We introduced a method for the identification of relevant frequencies of a harmonic process with error processes based on the normal distribution. The crucial idea for this method is to look at the estimates of the periodogram ordinates as a contaminated sample of a χ^2 distribution and use this to get an estimate of the variance of the error process. Additionally we showed that the necessary transformation of the periodogram ordinates to get an estimator for the amplitude leads to a normal approximation. Finally, we established the fact that the linearity of the Fourier transformation makes it possible to evaluate time trends in the amplitude by regression methods.

Our simulation proved all theoretical results to work even in difficult situations, i.e. low signal-to-noise ratio, non-Fourier frequencies and differing influences on the relevant frequencies. The only significant drawback of the method is the underestimation of the true amplitudes which may be tackled by summing over an approprate neighbourhood of the found relevant frequencies.

References

BLOOMFIELD, P. (2000): *Fourier Analysis of Time Series*, 2nd ed. Wiley, New York.

FISZ, M. (1970): *Wahrscheinlichkeitsrechnung und Mathematische Statistik*. VEB Deutscher Verlag der Wissenschaften.

GALLANT, A. R., GERIG, T. M. and EVANS, J. W. (1974) Time Series Realizations Obtained According to An Experimental Design. *JASA, 69, 639–645.*

GATHER, U. and SCHULTZE V. (1999): Robust estimation of scale of an exponential distribution. *Statistica Neerlandica, 53(3), 327–341.*

JOHNSON, N. L., KOTZ, S. and BALAKRISHNAN, N. (1994) *Continuous univariate distributions*, volume 1,2. Wiley, New York.

R DEVELOPMENT CORE TEAM (2003): R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria, URL: http://www.R-project.org.

SHAPIRO, S.S., WILK, M.B. and CHEN, M.J. (1968): A comparative study of various tests for normality. *Journal of the American Statistical Association, 63, 1343–1372.*

THEIS, W. (2004): *Modelling Varying Amplitudes*, Dissertation, Fachbereich Statistik, Universität Dortmund, http://eldorado.uni-dortmund.de:8080/FB5/ls7/forschung/2004/Theis.

WEINERT, K., WEBBER, O., HÜSKEN, M., MEHNEN, J. and THEIS, W. (2002): Analysis and Prediction of Dynamic Disturbances of the BTA Deep Hole Drilling Process. In Teti, R., editor, *Proceedings of the* 3^{rd} *CIRP International Seminar on Intelligent Computation in Manufacturing Engineering.*

WITTWER, G. (1986): On the Distribution of the Periodogram for Stationary Random Sequences. *statistics, 17(2), 201–219.*

Part IV

Contest: Social Milieus in Dortmund

Introduction to the Contest "Social Milieus in Dortmund"

Ernst-Otto Sommerer[1] and Claus Weihs[2]

[1] Bureau of Statistics and Elections Dortmund 44137 Dortmund, Germany
[2] University of Dortmund Department of Statistics 44221 Dortmund, Germany

Abstract. Goal and data of the contest "Social Milieus in Dortmund" are introduced.

1 Contest goal and data

Many cities have problems characterizing the social milieus of the different city districts. A correct classification of these milieus enables the cities to monitor changes and be able to react, if a milieu change of a district might be undesirable, or to detect positive developments. These information are valuable for example in determining the value of buildings and for the development of micro locations.

Therefore, the goal of this contest is to develop a statistical model and criteria for the classification of the social milieus of the city, using the following variables provided by the city of Dortmund (for a full list see table 1):

- population structure,
- unemployment,
- no. of employees,
- no. of welfare recipients,
- motoring, and
- buildings.

Participants were even allowed to add other relevant data. The data is available for 170 statistical sub-districts of Dortmund.

Table 1. Variables available for the contest provided by the city of Dortmund

Shortcut	Description
Alosaus	Unemployed foreigners
Alosdeut	Unemployed Germans
Alosf	Unemployed women
Alosins	Unemployed overall
Alosm	Unemployed men
Anhänger	Trailers
Auflieger	Semitrailers
BJ01bis18	Year of construction 1901 to 1918

BJ19bis48	Year of construction 1919 to 1948
BJ49bis57	Year of construction 1949 to 1957
BJ58bis62	Year of construction 1958 to 1962
BJ63bis72	Year of construction 1963 to 1972
BJ73bis82	Year of construction 1973 to 1982
BJ83bis92	Year of construction 1983 to 1992
BJ93bis01	Year of construction 1993 to 2001
BJbis1900	Year of construction until 1900
Bus	Buses
Dreirad	Tricycles
F26bis35	Women 26 to 35 years
F36bis45	Women 36 to 45 years
F46bis55	Women 46 to 55 years
F55bis65	Women 55 to 65 years
Fins	Women overall
Flächeha	Area in ha
Fu25	Women under 25 years
Fü66	Women 66 years and older
Gebäufläche	Rebuilt building with housing space - living area
Gebäuins	Rebuilt building with housing space - overall
Gebäusonsteinh	Rebuilt building with housing space - other residential units
Gebäuwohn	Rebuilt building with housing space - apartments
GebBest10fläche	R.b.w.l.s.[1] 10 and more apartments - living area
GebBest10ins	R.b.w.l.s.[1] 10 and more apartments - overall
GebBest10raum	R.b.w.l.s.[1] 10 and more apartments - rooms
GebBest10wohnung	R.b.w.l.s.[1] 10 and more apartments - apartments
GebBest1u2fläche	R.b.w.l.s.[1] 1 and 2 apartments - living area
GebBest1u2raum	R.b.w.l.s.[1] 1 and 2 apartments - rooms
GebBest1u2wohnung	R.b.w.l.s.[1] 1 and 2 apartments - apartments
GebBest1und2ins	R.b.w.l.s.[1] 1 and 2 apartments - overall
GebBest3fläche	R.b.w.l.s.[1] 3 and more apartments - living area
GebBest3ins	R.b.w.l.s.[1] 3 and more apartments - overall
GebBest3raum	R.b.w.l.s.[1] 3 and more apartments - rooms
GebBest3wohnung	R.b.w.l.s.[1] 3 and more apartments - apartments
GebBestfläche	R.b.w.l.s.[1] - living area
GebBestins	R.b.w.l.s.[1] - overall
GebBestraum	R.b.w.l.s.[1] - rooms
GebBestwohnung	R.b.w.l.s.[1] - apartments
Gebm	Births male
GebSterbBilm	Birth/death balance male
GebSterbBilw	Birth/death balance female
GebSterbBilzus	Birth/death balance both
Gebw	Births female
Gebzus	Births both
GenNeu10fläche	R.b.w.l.s.[1] 10 and more apartments - living area - building permits

[1] Residential building with living space

GenNeu10ins	R.b.w.l.s.[1] 10 and more apartments - overall - building permits
GenNeu10raum	R.b.w.l.s.[1] 10 and more apartments - rooms - building permits
GenNeu10wohnung	R.b.w.l.s.[1] 10 and more apartments - apartments - building permits
GenNeu1u2fläche	R.b.w.l.s.[1] 1 and 2 apartments - living area - building permits
GenNeu1u2raum	R.b.w.l.s.[1] 1 and 2 apartments - rooms - building permits
GenNeu1u2wohnung	R.b.w.l.s.[1] 1 and 2 apartments - apartments - building permits
GenNeu1und2ins	R.b.w.l.s.[1] 1 and 2 apartments - overall - building permits
GenNeu3fläche	R.b.w.l.s.[1] 3 and more apartments - living area - building permits
GenNeu3ins	R.b.w.l.s.[1] 3 and more apartments - overall - building permits
GenNeu3raum	R.b.w.l.s.[1] 3 and more apartments - rooms - building permits
GenNeu3wohnung	R.b.w.l.s.[1] 3 and more apartments - apartments - building permits
GenNeufläche	R.b.w.l.s.[1] living area - building permits
GenNeuins	R.b.w.l.s.[1] Overall - building permits
GenNeuraum	R.b.w.l.s.[1] Rooms - building permits
GenNeuwohnung	R.b.w.l.s.[1] apartments - building permits
GenUm10fläche	R.b.w.l.s.[1] 10 and more apartments - living area - rebuilding permits
GenUM10ins	R.b.w.l.s.[1] 10 and more apartments - overall - rebuilding permits
GenUm10raum	R.b.w.l.s.[1] 10 and more apartments - rooms - rebuilding permits
GenUm10wohnung	R.b.w.l.s.[1] 10 and more apartments - apartments - rebuilding permits
GenUm1u2fläche	R.b.w.l.s.[1] 1 and 2 apartments - living area - rebuilding permits
GenUm1u2wohnung	R.b.w.l.s.[1] 1 and 2 apartments - rooms - rebuilding permits
GenUm3fläche	R.b.w.l.s.[1] 1 and 2 apartments - apartments - rebuilding permits
GenUm3ins	R.b.w.l.s.[1] 1 and 2 apartments - overall - rebuilding permits
GenUm3raum	R.b.w.l.s.[1] 3 and more apartments - living area - rebuilding permits
GenUm3wohnung	R.b.w.l.s.[1] 3 and more apartments - overall - rebuilding permits
GenUmfläche	R.b.w.l.s.[1] 3 and more apartments - rooms - rebuilding permits

[1] Residential building with living space

GenUmraum	R.b.w.l.s.[1] 3 and more apartments - apartments - rebuilding permits
GenUmt1u2raum	R.b.w.l.s.[1] living area - rebuilding permits
GenUmt1und2ins	R.b.w.l.s.[1] Overall - rebuilding permits
GenUmtins	R.b.w.l.s.[1] Rooms - rebuilding permits
GenUmwohnung	R.b.w.l.s.[1] apartments - rebuilding permits
HBWins	Overall population classified by site of major apartment (PMA)
HBWins0bis6	Overall PMA younger than 6 years
HBWins10bis13	Overall PMA 10 to 12 years
HBWins13bis16	Overall PMA 13 to 15 years
HBWins16bis18	Overall PMA 16 to 17 years
HBWins18bis26	Overall PMA 18 to 25 years
HBWins26bis30	Overall PMA 26 to 29 years
HBWins30bis40	Overall PMA 30 to 39 years
HBWins40bis50	Overall PMA 40 to 49 years
HBWins50bis60	Overall PMA 50 to 59 years
HBWins5bis6	Overall PMA older than 5 and younger than 6 years
HBWins60bis65	Overall PMA 60 to 64 years
HBWins6bis10	Overall PMA 6 to 9 years
HBWinsü65	Overall PMA 65 years and older
Hh10K	Households with 10 children
Hh11K	Households with 11 children
Hh1K	Households with 1 child
Hh2K	Households with 2 children
Hh3K	Households with 3 children
Hh4K	Households with 4 children
Hh5K	Households with 5 children
Hh6K	Households with 6 children
Hh7K	Households with 7 children
Hh8K	Households with 8 children
Hh9K	Households with 9 children
Hhins	Households overall
HWBins	Overall population classified by site of major apartment (PMA)
HWBins0bis1	Overall PMA under 1 year
HWBins1bis2	Overall PMA older than 1 and younger than 2 years
HWBins2bis3	Overall PMA older than 2 and younger than 3 years
HWBins3bis4	Overall PMA older than 3 and younger than 4 years
HWBins4bis5	Overall PMA older than 4 and younger than 5 years
HWBinsA	Overall PMA foreigners
HWBinsaF	Overall PMA foreigners women
HWBinsaM	Overall PMA foreigners men
HWBinsD	Overall PMA Germans
HWBinsdF	Overall PMA German women
HWBinsdM	Overall PMA German men
HWBinsF	Overall PMA women

[1] Residential building with living space

HWBinsM	Overall PMA men
ID	Running identification number
InstdtUmBilm	Balance of moves within city male
InstdtUmBilw	Balance of moves within city female
InstdtUmBilzus	Balance of moves within city both
InstdtUmFortm	Move−outs within city male
InstdtUmFortw	Move−outs within city female
InstdtUmFortzus	Move−outs within city both
InstdtUmZum	Move−ins within city male
InstdtUmZuw	Move−ins within city female
InstdtUmZuzus	Move−ins within city both
Kins	Children overall
Kombi	Estate cars
LKW	Trucks
M26bis36	Men 26 to 35 years
M36bis45	Men 36 to 45 years
M46bis55	Men 46 to 55 years
M55bis65	Men 55 to 65 years
Mins	Men overall
Motorrad	Motorcycles
Mu25	Men under 25 years
Mü66	Men 66 years and older
NameRaum	Name of area unit
NumBezRaum	Identification number of area unit
ohneAng	Vehicles without specifications
PKW	Cars
Sonder	Special−purpose vehicles
sonstGfläche	Rebuilt other building with living space - living area
sonstGins	Rebuilt other building with living space - overall
sonstGsonsteinh	Rebuilt other building with living space - other residential units
sonstGwohn	Rebuilt other building with living space - apartments
SozempfausF	Welfare recipients - foreigners women
SozempfausM	Welfare recipients - foreigners men
SozempfdF	Welfare recipients - Germans women
SozempfdM	Welfare recipients - Germans men
sozvpflBeschAus	Subjects to social insurance contribution - foreigners
sozvpflBeschDeut	Subjects to social insurance contribution - Germans
sozvpflBeschF	Subjects to social insurance contribution - women
sozvpflBeschins	Subjects to social insurance contribution - overall
sozvpflBeschM	Subjects to social insurance contribution - men
Sterbm	Deaths male
Sterbw	Deaths female
Sterbzus	Deaths both
Wanbilm	Migration balance male
Wanbilw	Migration balance female
Wanbilzus	Migration balance both
WanFortm	Emigration male

WanFortw	Emigration female
WanFortzus	Emigration both
WanUmGesbilm	Migration/moves overall balance male
WanUmGesbilw	Migration/moves overall balance female
WanUmGesbilzus	Migration/moves overall balance both
WanZum	Immigration male
WanZuw	Immigration female
WanZuzus	Immigration both
Wgeb10fläche	New r.b.w.l.s.[1] 10 and more apartments - living area
Wgeb10ins	New r.b.w.l.s.[1] 10 and more apartments - overall
Wgeb10raum	New r.b.w.l.s.[1] 10 and more apartments - rooms
Wgeb10wohnung	New r.b.w.l.s.[1] 10 and more apartments - apartments
Wgeb1u2fläche	New r.b.w.l.s.[1] 1 and 2 apartments - living area
Wgeb1u2ins	New r.b.w.l.s.[1] 1 and 2 apartments - overall
Wgeb1u2raum	New r.b.w.l.s.[1] 1 and 2 apartments - rooms
Wgeb1u2wohnung	New r.b.w.l.s.[1] 1 and 2 apartments - apartments
Wgeb3fläche	New r.b.w.l.s.[1] 3 and more apartments - living area
Wgeb3ins	New r.b.w.l.s.[1] 3 and more apartments - overall
Wgeb3raum	New r.b.w.l.s.[1] 3 and more apartments - rooms
Wgeb3wohnung	New r.b.w.l.s.[1] 3 and more apartments - apartments
Wgebäufläche	Rebuilt r.b.w.l.s.[1] - living area
WGebäuins	Rebuilt r.b.w.l.s.[1] - overall
WGebäusonsteinh	Rebuilt r.b.w.l.s.[1] - other residential units
WGebäuwohn	Rebuilt r.b.w.l.s.[1] - apartments
Wgebfläche	New r.b.w.l.s.[1] - living area
Wgebins	New r.b.w.l.s.[1] - overall
Wgebraum	New r.b.w.l.s.[1] - rooms
Wgebwohn	New r.b.w.l.s.[1] - apartments
Wheimfläche	Dormitories - living area
Wheimins	Dormitories - overall
Wheimsonsteinh	Dormitories - other residential units
Wheimwohn	Dormitories - apartments
Zugmaschine	Tractors

The map of Dortmund together with the Identification number of the area unit is shown in figure 1.

Steps suggested for the data analysis include:

1. Clustering of sub-districts (unsupervised learning). The geographical relationship of the clusters should be taken into account (see the map of Dortmund). Spatial smoothing may be necessary.
2. Interpretation of the clusters and assignment of milieu types.
3. Deriving a classification rule for the milieu types.
4. Characterization of classes (i.e. milieu types) by the measured variables.

Other steps leading to comparable results were welcome.

[1] residential building with living space

Fig. 1. Map of Dortmund

Application of a Genetic Algorithm to Variable Selection in Fuzzy Clustering

Christian Röver and Gero Szepannek

Fachbereich Statistik,
Universität Dortmund,
44221 Dortmund, Germany
roever@statistik.uni-dortmund.de
szepannek@statistik.uni-dortmund.de

Abstract. In order to group the observations of a data set into a given number of clusters, an 'optimal' subset out of a greater number of explanatory variables is to be selected. The problem is approached by maximizing a quality measure under certain restrictions that are supposed to keep the subset most representative of the whole data. The restrictions may either be set manually, or generated from the data. A genetic optimization algorithm is developed to solve this problem.
The procedure is then applied to a data set describing features of sub-districts of the city of Dortmund, Germany, to detect different social milieus and investigate the variables making up the differences between these.

1 The problem

Before the observations are clustered, the data need to be reduced. A reduction is necessary to

1. avoid overfitting,
2. exclude noise and redundant variables and
3. keep the data perceptible and interpretable.

To achieve these goals, we would like to use a subset of the original variables rather than, for example, linear combinations (like principal components) that are harder to interpret.
To determine an 'optimal' subset of variables, some measure of cluster quality needs to be optimized; this measure should return comparable values regardless of the number or scale of variables in the subset. Also, some restrictions should be met to make sure that, for example, the subset has more than one element and, in some sense, most data features are reflected in the subset.

2 Tackling the problem

We focused on *fuzzy clustering methods*, that is, methods that do not assign fixed clusters to each observation, but that return posterior probabilities of

cluster membership instead. These methods are often a more appropriate approach to clustering problems.
Validity measures are then computed from the membership matrix that is yielded by clustering with a specific variable set, and thus independently from the underlying variables themselves. In particular, they do not depend directly on the number or scales of variables. Assessment of clusterings with different variable sets can then be based on such measures.

Basing variable selection on the membership matrix alone still may lead to in some sense 'optimal', but still useless solutions. The final variable set may consist of a single, or some highly correlated variables, for example.
Instead, we try to keep the selected subset of variables as representative as possible of the complete data set. In order to achieve this, we are introducing subgroups of variables that have to be represented in the selected subset. These subgroups are either arranged 'by hand' (groups of variables with similar meaning or representing a certain aspect of the data set) or automatically (groups of correlated variables).

The selection itself is then performed by a genetic algorithm that can pretty easily be adapted to handle a parameter space of this kind (that is, a restricted space with varying dimension).

All computations will be performed using R, a free software for data analysis (Ihaka and Gentleman (1996)).

3 Methods

3.1 Fuzzy clustering

Usually, a clustering procedure returns specific assignments of clusters to all observations. Fuzzy clustering methods instead are those methods, that for each observation provide indices measuring the potential affiliation to *all* of the clusters.
The result of a fuzzy clustering then is a $(N \times k)$ membership matrix U, whith u_{ij} denoting the probability that observation i belongs to cluster j; or in other words: each row of U corresponds to one observation (i) and is the distribution of membership over clusters $1, \ldots, k$. An example with 3 clusters:

$$U = \begin{pmatrix} 0.95 & 0.02 & 0.03 \\ 0.50 & 0.30 & 0.20 \\ \vdots & \vdots & \vdots \end{pmatrix}$$

Both observations would be assigned to cluster 1, while the second assignment is not as clear as the first one.
We considered two different clustering methods, the `cmeans`-procedure from

the `e1071` package, which is a fuzzy version of the known k-means clustering, and the `EMclust`-procedure from the package `mclust`. '`EMclust`' fits a Gaussian mixture model with k components to the data; in this case the components have the same covariance structure and differ by their means and a-priori-probabilities. The data is then clustered by assigning each observation to one of the k mixture components. We eventually decided in favour of the second method, mostly for interpretability reasons: while k-means-clustering by its nature carves the data into sphere-shaped clusters, model-based clustering is able to handle clusters with covariance structures even different to spheric shapes and does not require (and depend on) normalization (Fraley and Raftery (2002)).

3.2 Measuring the clustering quality

Let U be a $(N \times k)$ membership matrix, as above. All the u_{ij} should be close to one or zero, so the clustering yields distinct assignments. A measure of this feature is the *classification entropy*:

$$\mathrm{CE}(U) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}(u_{ij} \cdot \log_2 u_{ij})$$

$CE(U)$ is zero, if all elements of U are either 1 or 0 (a most crisp partitioning), and takes its maximum if all of them are $= \frac{1}{k}$ (the fuzziest partitioning) (Hall (1999)).
The classification entropy allows comparison of clusterings based on different variable sets with different numbers of variables, but is sensitive to the number of clusters (k), so this quantity needs to be fixed beforehand.

While a most fuzzy clustering pretty obviously is a bad clustering, a crisp clustering does not necessarily have to be a good clustering. A variable subset leading to low entropy may in some sense not represent the data appropriately. In order to force some structure into the subset selection process, the concept of subgroups is introduced in the following section. This allows for the injection of expert knowledge or of further information on the data (correlations) into the procecedure.

3.3 Defining subgroups of variables

Subgroups can be defined manually, or they are constructed systematically as groups of *correlated variables*. These subgroups are generated by agglomerative hierarchical clustering (Kaufman and Rousseeuw (1990)); the *variables* are clustered, and to do so, the 'distance' between two variables X_1 and X_2 is defined as:

$$d(X_1, X_2) = 1 - |\mathrm{Cor}(X_1, X_2)|$$

Thus, variables with a high (absolute) correlation are 'close' to each other, while uncorrelated variables are 'farther' from each other. With this definition, the correlation matrix can directly be transformed into a distance matrix, which is the only basis needed for the clustering. Using either *complete* or *single linkage* yields groups of variables with different interpretations:

complete linkage: the (absolute) correlation of variables from the same group is bounded below.
single linkage: the (absolute) correlation of variables from different groups is bounded above.

In both cases, the groups may be interpreted as variable sets with some common source of variability; and by picking variables from different groups, the intention is to cover these different sources.

3.4 Genetic optimization algorithms

Optimization problems, in general, are problems of finding the minimum of some function $f : \mathcal{M} \to \mathbb{R}$ that projects from some space $\mathcal{M}$ to the real line. Genetic algorithms are stochastic optimization algorithms to solve these kinds of problems by making use of evolutionary principles as known from biology, namely *mutation*, *recombination* and *selection ('survival of the fittest')*.
In nature, the fitness of individuals depends on their genes. Individuals with a greater fitness have a greater chance of survival and also the wider range of mating partners to choose between. 'New' individuals are generated by

mutation: single genes of an individual are changed, or
recombination: two genomes are combined to a new one.

Again, new individuals have to compete with the current population for partners and survival. The competitiveness of each individual is determined by its fitness.
Analogously, in genetic algorithms the goal function to be optimized corresponds to the fitness, and the individuals are parameter sets for the function. To start the algorithm, a starting population is generated. Then, generation by generation, the population is multiplied by mutating and breeding individuals and only the 'fittest' ones (as judged by the goal function) survive until the next generation. At some point, the procedure stops and the (so far) best parameter set is returned.
An advantage of genetic algorithms is, that the parameter space ($\mathcal{M}$) can literally be *any* space, as long as the mutation- and recombination procedures can be defined reasonably. Restrictions are implemented pretty easily as well (Goldberg (1989)).

3.5 Implementation

Parameters to be defined beforehand are: the subgroups of variables, the minimum numbers of representatives for each group that have to be in a variable subset (≥ 0), the (total) maximum number of variables in a subset, the 'population size' and the number of generations.
A 'genome' (an 'individual') is a vector of variable indices; its minimum length depends on the sum of minimum numbers for each variable group, and the maximum length is defined explicitly. The population is made up by a set of these individuals. The fitness of each individual is determined by clustering the data using the corresponding subset of variables and then computing the classification entropy as the measure of clustering quality that is achieved with this subset (the smaller the entropy, the greater the fitness).
In the beginning, a random starting population (of the given size) is created, and the fitness of each individual is determined. In each generation, individuals are mutated (a new individual is generated by changing, adding or deleting single indices from a given individual), and pairs of individuals are crossed (a new set of indices is selected from the union of parental indices). The chance of being mutated or crossed is proportional to the individuals' fitness. In each step (creation of starting population, mutation, recombination) it is made sure that the resulting individuals comply with the restrictions (given by the pre-defined subsets and referring minimum numbers). After each generation, the population is cut down again to the former population size (fittest individuals are kept).
After a given number of generations the fittest individual (the best subset of variables) is returned.

4 Applying the procedure

4.1 The Dortmund data

The data consisted of 170 observations of 200 variables, referring to 170 subdistricts of the city. All variables were total numbers (of inhabitants, females, births,...), so in order to make them comparable across districts, we first constructed normalized variables like '*fraction of female inhabitants*', '*birth rate*' and so on. The result was a set of 57 variables describing features like

i. age distribution
ii. births, deaths, migration
iii. motoring
iv. buildings, housing
v. employment, welfare
vi. some of the above broken down by sex or citizen/alien status

12 out of the 170 observations were considered as outliers; they showed extreme values in some variables, and by checking the corresponding district

on a city map, one could see that these were either extremely sparsely populated or contained some special feature like a boarding school, an old people's home, etc. These were then ignored in the further analysis.
The four groups that we considered should be represented are described by points i, ii, iv and v of the above enumeration. The remaining variables form a group that does not necessarily have to appear.
Grouping the variables by correlations in this case resulted in either huge numbers of subgroups, most of which containing only one variable, or the respective lower/upper correlation bound would be of insignificant order, leading to rather meaningless groupings. So we eventually dropped the automatic grouping approach and only used the subgroups arranged by variable meanings.
Each of the 4 groups should be represented by 1 variable in the final variable subset. In order to keep the data comprehensible, we set the maximum number of variables to 6. That forces the algorithm to choose 1 variable from each of the 4 groups, the remaining variables can then be picked arbitrarily.
Another quantity to be defined beforehand is the number of clusters. After some data exploration, repeated application of the procedure for different values and inspection of the resulting clusterings we found that the different city districts indicated the presence of 4 clusters that repeatedly showed up with a variety of variable sets.

4.2 Results

The 'optimal' set of variables, with respect to the clustering quality measure and restrictions, that we found, is shown in Table 1.

Table 1. Clustering variables and their means.

		Cluster			
Variable	**Group**	**1**	**2**	**3**	**4**
fraction of population of age 60–65	i.	0.057	0.065	0.064	0.083
moves to district per inhabitant	ii.	0.075	0.054	0.035	0.025
apartments per house	iv.	7.831	5.331	3.367	2.524
people per apartment	iv.	1.877	1.676	2.216	2.029
fraction of welfare recipients	v.	0.129	0.031	0.066	0.023
fraction of aliens of employed people	vi.	0.274	0.073	0.086	0.032

Figure 1 displays the distribution of the clusters across the city map. Clusters 1 and 2 roughly cover the city center, subdividing it into north (1) and south (2), while clusters 3 and 4 cover the remaining suburbs (roughly northwest and southeast).
The greatest differences are between clusters 1 (center north) and 4 (southeast suburbs). Cluster 1 has a low fraction of older inhabitants, great fractions of

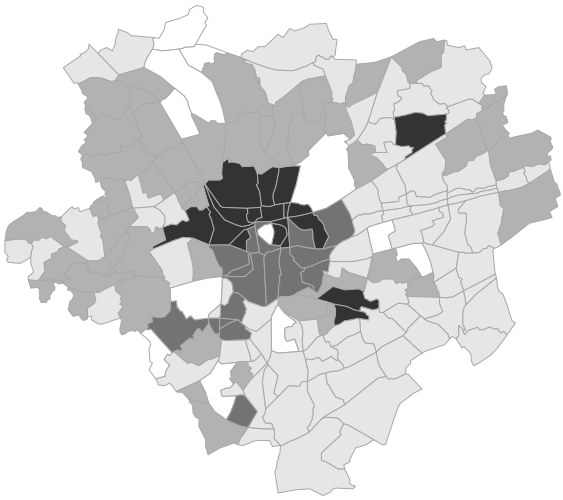


Fig. 1. Map of Dortmund showing the 4 clusters (from Cluster 1=darkgrey to Cluster 4=lightgrey; white districts are the outliers that were omitted).

aliens and welfare recipients, more migration and many apartments per house while cluster 4 takes the opposite extreme values. Clusters 2 and 3 are both more or less between these two extremes and differ by their buildings/housing structure: cluster 2 (center south) has more apartments per house and the fewest people per apartment while cluster 3 (northwest suburbs) has the most people per apartment.

4.3 Comparing the results

Clustering the data by *all* variables instead of a subset leads to pretty similar maps, for both the traditional k-means-algorithm and EM clustering based on gaussian mixture models.
Differences become evident when it comes to interpretation. When clustering with all variables, the different variable types (as indicated in the table in section 4.1) are weighted by the number of variables in each of the groups, which are rather random. In contrast, in the approach presented here these proportions are set manually. Also, a selection of necessary variables and elimination of noise variables does not take place. Using only a subset of variables, clusters can thus be easily characterized by the distribution of the (far fewer) variables that were actually used.

5 Summary

The variable selection problem was approached by introducing a quality measure for clusterings and certain restrictions to retain as many information as possible from the complete data set in the variable subset. The optimal variable selection was then performed by a genetic optimization algorithm.
For the Dortmund data, the attempt to define variable subgroups based on correlations proved to be impractical, so the variables were only grouped manually by their respective meanings. Data exploration suggested the presence of 4 clusters. The application of the developed procedure resulted in a plausible set of discriminating variables and a reasonable distribution of the clustered districts across the map. While actual clustering results are similar to those of traditional methods, the necessary data was reduced to a minimum on which to focus any possible further investigation.

References

FRALEY, C. and RAFTERY, A.E. (2002): `mclust`: Software for model-based clustering, density estimation and discriminant analysis. *Technical Report, Department of Statistics, University of Washington.*
See `http://www.stat.washington.edu/mclust`.

GOLDBERG, D.E. (1989): *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Boston.

HALL, M.A. (1999): Correlation-based feature subset selection for machine learning. *PhD thesis, Department of computer science, University of Waikato.*

IHAKA, R. and GENTLEMAN, R. (1996): R: A language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics, 5 (3), 299-314.*
See also `http//:www.r-project.org`

KAUFMAN, L. and ROUSSEEUW, P.J. (1990): *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York.

Annealed k-Means Clustering and Decision Trees

Christin Schäfer and Julian Laub

Fraunhofer Institute FIRST, Intelligent Data Analysis Group
Kekuléstr.7, 12489 Berlin, Germany.

Abstract. This paper describes a contribution to the GfKl 2004 Contest. The contest task is to cluster, classify and interpret the 170 districts of the city of Dortmund with respect to their 'social milieux'. A data set containing 204 variables measured for every district is given.

We apply annealed k-means clustering to the preprocessed contest data. Superparamagnetic clustering is used to foster insight into the natural partitions of the data. A stable and interpretable solution is obtained with $k = 3$ clusters, dividing Dortmund into three social milieux. A decision tree is deduced from this cluster solution and is used for interpretation and rule generation. The tree offers the possibility to monitor and predict future assessments. To gain information about cluster solutions with $k > 3$ a stability analysis based on a resampling approach is performed resulting in further interesting insights.

1 Introduction

Unsupervised grouping or *clustering* aims at extracting hidden structure from data (see e.g. Duda et al. (2001)). However, because of the absence of labels giving a ground truth, clustering is an inherently ill-defined problem. There is no natural measure of goodness, or cost function. The cluster solution must be evaluated and interpreted by the experimenter.

Furthermore, with regard to the present problem, the notion of the very goal, clustering social milieux, is ill defined too, thus impairing the necessary intuition of the experimenter judging the clustering result.

Dortmund is subdivided into 170 districts. Since the districts are administrative subdivisions and do not occur as abrupt, physical disruptions (akin the former Wall of Berlin) we assume smooth continuity at the respective borders of the (theoretical) surface spanned by the measured data. Therefore the short-range geographical relationship among the districts is not further taken into account. As to the long-range correlation between districts, they are expected to vanish quickly beyond a given district as we assume that Dortmund is a very heterogenous city.

In section 2 the preprocessing and reduction of the data is described. The subsequent clustering procedure is delineated in section 3. From the resulting cluster solution a decision tree is derived in section 4 and the interpretation of the result is given in section 5. Finally section 6 presents an outlook to solutions with a larger number of clusters.

Table 1. Remaining variables after preprocessing. The abbreviations in the brackets are used in text and figures.

Unemployed Germans (`Alosdeut`)
Migration balance both (`Wanbilzus`)
Balance of moves within city both (`InStdtUmBilzus`)
Car (`PKW`)
Bus (`Bus`)
Tricycle (`Dreirad`)
Children overall (`Kins`)
Men overall (`Mins`)
Women overall (`Fins`)
Subjects to social insurance contribution Germans (`sozvpflBeschDeut`)

2 Preprocessing

The data set of the GfKl 2004 Contest is provided by the *Amt für Statistik und Wahlen* of Dortmund and consists of 204 variables for the 170 sub-districts of the city measured in 2002. The data set offers a description of Dortmund from the view of official social statistics. The variables themselves form semantic groups like population structure, unemployment, number of employees, welfare recipients etc. All variables are measured in absolute frequencies, only the area of districts is measured in square meters.

The variables forming a semantic group exhibit high mutual correlations. For the purpose of dimension reduction we eliminate those groups with intragroup correlation > 0.7 and use one single hand-chosen group member as representative. For example, the semantic group 'unemployment' is represented through `Alosdeut`. This preprocessing step reduces noticeably the number of variables. The only semantic group which cannot be represented by a single variable is the 'motoring'. The variables `HWB` and `Flaeche` are kept aside and used for normalization only. The remaining variables are analyzed with respect to the discrimination information they contain: a variable with measurements all at the same level for every district is useless for clustering and classification. Only those variables are kept, which are informative. The final set of variables is depicted in table 1.

3 Clustering

3.1 Annealed k-means

One of the most popular clustering methods is k-means clustering. It derives a set of k prototype vectors which quantize the data set with minimal quadratic quantization error. However, the iterative optimization of the cost function, starting from a random initial distribution of k vectors, is prone to local minima – especially in the case where the dimensionality p of the data is of

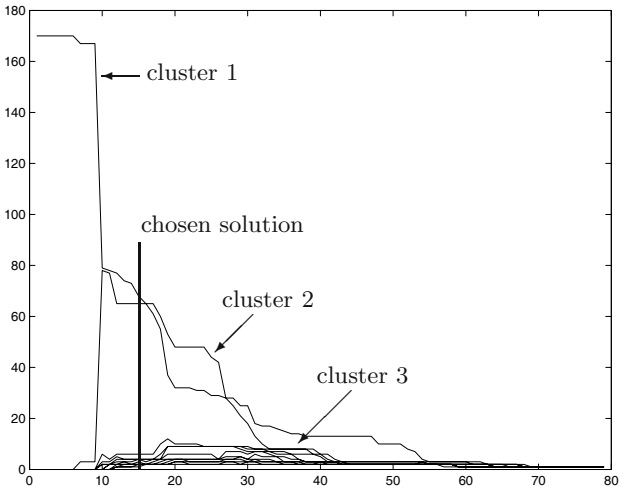


Fig. 1. Cluster size and number (y-axis) versus temperature (x-axis).

the same order or even larger than the number of samples n – resulting in poor stability. A common workaround consists of using *deterministic annealing* for optimization, which grant better optima. See e.g. Hofmann and Buhmann (1997) for annealing techniques in unsupervised learning.

3.2 Learning about k

Solutions for a large number k of clusters pose serious problems for interpretation. Therefore, in order to assess methodological validity of our technique and gain intuition for the data, we first will choose $k = 3$.

This choice is far from being arbitrary. It can be justified by an analysis using *super-paramagnetic clustering* (Blatt et al. (1996)). Super-paramagnetic clustering is based upon the physical analogy of Pott spins with k states. Initially one starts from n points at zero temperature with identical spin. By successively increasing the temperature, the points break into separate domains ('clusters') of different spin. Finally, at high temperature, every point forms a single cluster.

Figure 1 shows the size of the ten largest clusters as the temperature increases. At the 6th time-step ($t = 6$) the n-data block splits into two clusters. Further splits occurs at $t = 9$, creating beside the two large clusters a myriad of very small clusters, which we may reunite into a third, bulk-cluster. As temperature increases, the bulk-cluster increases at the expense of the two main clusters. In spite of this decrease, we see that the overall structure of three clusters is preserved over a wide temperature range. It is only above $t = 30$ that the second cluster becomes so small that it is no longer distinguishable from the bulk-cluster. At $t = 55$ this happens for the first cluster.

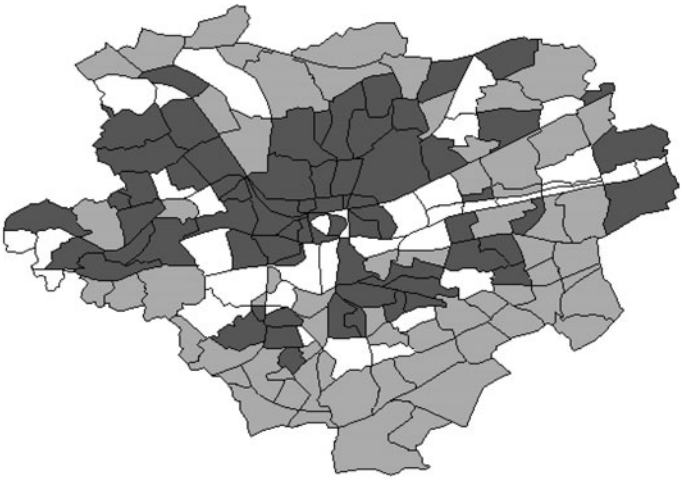


Fig. 2. Visualization of the super-paramagnetic clustering solution. The black areas corresponds to cluster 1 in figure 1, the grey areas refers to cluster 2 and therefore, cluster 3 consists of the white districts.

The cluster solution is obtained by choosing a temperature in a range in which the solution does not change much. $k = 3$ thus appears to be a very sensible choice, however, leaving us quite a range to choose a solution from. We choose $t = 15$ – indicated in figure 1 by a solid line – yielding the clustering depicted in figure 2.

3.3 Solution

The cluster solution given through the super-paramagnetic clustering can be seen as a byproduct of the model selection. However our aim is to use annealed k-means clustering, that is k-means clustering optimized by deterministic annealing. Consequentially we present the cluster solution for $k = 3$ for annealed k-means clustering in figure 3. The overlap of the two solutions is remarkable and corroborates the validity of our procedure, as well as the underlying stability of the result. The sizes of the three clusters for the super-paramagnetic solution are 68, 65 and 37, and for the annealed k-means solution 53, 79, 38. Note that migration of districts between clusters from one solution to the other cannot be summarized in an easy way.

4 Classification

The next step is to deduce a classification rule from the initially unsupervised problem. This is done by using the cluster-assignments induced by annealed k-means clustering for $k = 3$ to learn a decision tree. We derive a classification

Fig. 3. Visualization of the solution obtained through annealed k-means clustering. Similarly as in figure 2, we label the black areas as cluster 1, the grey one as cluster 2, and the white one as cluster 3.

tree by using the well known algorithm *C4.5* of Quinlan (1993). For the problem the use of decision trees offers several advantages: first, the rules generated to build the tree are helpful for labeling, describing and interpreting the clusters, especially for the interpretation with respect to 'social milieu'. A second advantage of decision trees is, that monitoring as well as predicting the 'social milieu' of a district, can be done easily. Figure 4 shows the decision tree derived by C4.5. One has to keep in mind that the depicted variables are only representatives of their semantic group. The variable `PKW` contains the largest amount of information with respect to the given classification. Splitting the districts using their values of `PKW` separates the cluster 2 from the others. Noting that the leaves at the second level of the tree cover more than 90% of the districts (depicted as rectangles in figure 4). Hence, to separate cluster 2 from cluster 1 and 3 mainly only one split of `PKW` is needed. To distinguish between cluster 1 and 3 only one further split by `Kins` must be conducted. The elongated sub-tree at the right hand side of the tree illustrates the cost to classify the remaining 10% of districts.

Figure 5 shows the classification solution if one uses only the information of the first two splits, combining the remaining not yet classified districts to a fourth cluster. This cluster is depicted by white areas. One may criticize that the derived rules for the classification of the three clusters are rough and simple. It turns out that even if one uses for learning the whole data set instead of the reduced one, the resulting set of rules will be the same, both in numbers as in semantic structure.

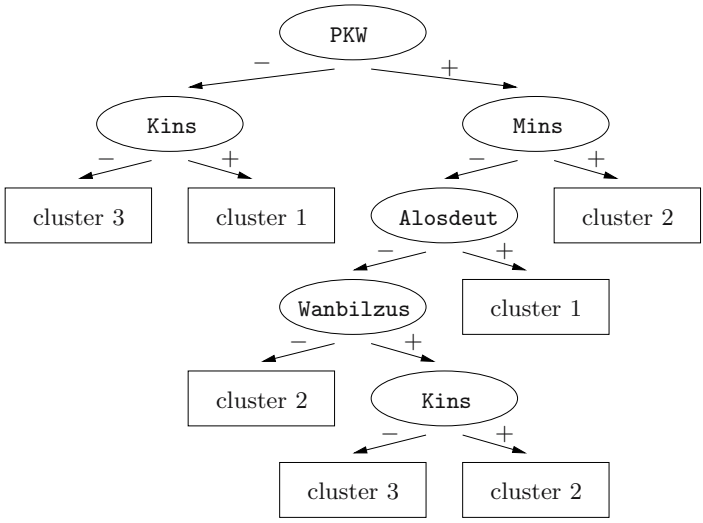


Fig. 4. Classification Tree for the clusters from annealed k-mean clustering. 'Cluster 1' refers to the black cluster in figure 3, 'cluster 3' is given through the white cluster and therefore 'cluster 2' is depicted by the grey cluster.

5 Interpretation

The results provided by the cluster solutions in figures 2, 3, and 5 and the decision tree in figure 4 allow for the following interpretation: It is possible to cluster the 170 sub-districts of Dortmund into 3 clusters. The resulting cluster solution is very stable. The most important variable is `PKW`. Cluster 2 is characterized through a high level of `PKW`. From the figures 2, 3, and 5 we derived a label for cluster 2, that is 'Dörferrunde', ring of small villages. In fact, cluster 2 contains the outer parts of Dortmund, especially in the south, that are villages suburbanized to Dortmund. The 'Dörferrunde' may suffer from an underdeveloped public transportation system and therefore cars are important in daily life. This explanation approach may be valid for districts like Salingen (district number 643) and Kruckel (662). On the other hand, a large number of cars may indicate prosperity of the habitants, which prefer to live in nice houses with garden. This explanation may be true for districts like Höchsten (541), Syburg (551) and parts of Aplerbeck (e.g. 416, 417, 451, 452). Summarizing, we associate carefully a 'social milieu' to cluster 2: young families, countryside middle class and upper class.

In contrast, cluster 1 and 3 are characterized by a lower level of `PKW` and separated by their level of the variable `Kins`. Cluster 1 is characterized by the higher level of `Kins` and when incorporating the geographical information it can be denoted as 'north city' and 'harbor'. One can conclude that cluster 1 is a living area for workers. A district falls into cluster 3 if both `PKW` and `Kins` are realized on their lower levels. Therefore, this cluster can be assumed to

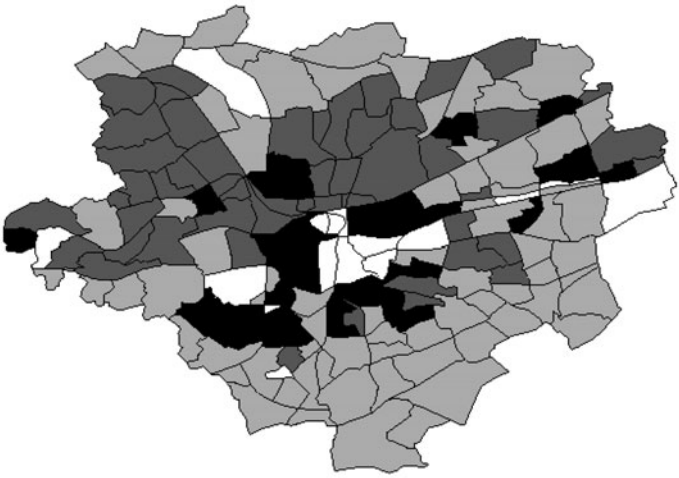


Fig. 5. White areas show districts which are not yet classified through the first two levels of the tree, whereas the colored clusters are defined through three easy rules containing only two constraints.[cluster 1 (gray), cluster 2 (light gray), cluster 3 (black)]

be real urban districts with a population consisting of elderly people, young couples without kids and young singles. From figure 3 it is noticeable that the cluster is mainly build from the southern part of the inner city and the region around the university. More precisely, Universität (642), Eichlinghofen (641), Hombruch (651), Westfalenhalle (11) and Westfalendamm (81,83), Ruhrallee West (91) and Ost (92) are members of cluster 3. To summarize we may denote the 'social milieu' of cluster 3 with urban middle class and students.

6 Outlook

The preceding section highlights that the derived cluster solution can be interpreted in a meaningful way. But inside the interpretation-attempts it becomes obvious that the multifariousness of a society and therefore of 'social milieux' can not be comprised with only 3 main groups. Thus, we are interested in cluster solutions with $k > 3$ classes. The question arises which k to choose. A criterion for a stability based choice of k is given in Roth et al. (2002). It is based upon evaluating over a certain range of k an instability index related to the comparison of two cluster solutions obtained by resampling the initial data set. The k for which this instability index is (locally) minimal is stable, and hence, for lack of extrinsic criteria, is a suitable choice.

For the contest data there are three potentially interesting local minima: $k = 9$, $k = 12$ and $k = 22$. Figure 6 shows the situation for $k = 9$. Note that the assignment of colors inside the figure is arbitrary. Therefore, one has to resist the temptation to interpret clusters with similar colors as similar

Fig. 6. Solution obtained for 9 clusters.

with respect to their 'social milieu'. However, this gives evidence of the issue of interpretation for large k's. This problem of interpretation increases with the number of clusters k. Another drawback of this kind of visualization is that within the black-to-white coloring scheme the clusters become indistinguishable for large k. Therefore we do not show the solutions for $k = 12$ and $k = 22$. The resulting decision trees can not be properly depicted on a paper of size Din A4. On request, the decision trees for $k = 9$, $k = 12$ and $k = 22$ can be obtained from the authors. Notice, that especially the tree for $k = 22$ is a refinement of the solution for $k = 3$ shown in figure 4.

Acknowledgement The authors want to thank Klaus-Robert Müller for valuable discussions and suggestions. We gratefully acknowledge the grants # MU 987/1-1, # BU 914/4-1 and # JA 379/13-2 from the Deutsche Forschungsgemeinschaft, # FKZ 01-SC40A from the Bundesministerium für Bildung und Wissenschaft, as well as PASCAL Network of Excellence (EU # 506778).

References

BLATT, M., WISEMAN, S. and DOMANY, E. (1996): Super-parametric clustering of data. *Physical Review Letters, 76.*

DUDA, R.O., HART, P.E. and STORK, D.G. (2001): *Pattern classification.* John Wiley & Sons, second edition.

HOFMANN, T. and BUHMANN, J. (1997): Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19 (19), 1–14.*

QUINLAN, J.R. (1993): *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo.

ROTH, V., BRAUN, M., LANGE, T. and BUHMANN, J. (2002): A Resampling Approach to Cluster Validation. *Computational Statistics–COMPSTAT'02, 123–128.*

Correspondence Clustering of Dortmund City Districts

Stefanie Scheid

Department for Computational Molecular Biology,
Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany

Abstract. We combine correspondence analysis (CA) and K-means clustering to divide Dortmund's districts into groups that are associated to particular variables and thus represent a social cluster. CA visualizes associations between rows and columns of a frequency matrix and can be used for dimension reduction. Based on the first three dimensions after CA mapping we find a stable partition into five clusters. We further identify variables that are highly associated with the cluster centroids and thus represent a cluster's social condition.

1 Introduction

The city of Dortmund is regarded as a highly heterogeneous urban area. The 170 districts comprise the city center, the ancient Dortmund, and incorporated suburbs that are nowadays either connected to the city center or maintained as village-like entities. The city covers industrial areas as well as fields and forests.

The City of Dortmund provides data of 200 variables for each district. The question is whether the city's heterogeneity is reflected in the data and whether it is possible to classify the districts with respect to their social environment. Since the definition of "social environment" is wide, we do not only regard typical social variables like amount of unemployed adults. By defining general preselection rules based on correlation and variability, we arrive at a set of variables that does not only represent social but general residential information.

The reduced data set is submitted to correspondence analysis (CA). CA is a useful tool to visualize associations between rows and columns of a frequency matrix, that is between districts and variables. Row and column vectors are simultaneously mapped into a space where the similar direction of vectors reflects their association. Proximity of districts alone reflects close relationship with respect to similar associated variables. We apply K-means clustering after mapping to CA space. We derive a stable cluster pattern of districts that can be interpreted with regard to the corresponding variables.

Both CA and K-means clustering are standard classification techniques. However, it is possible to analyse the given data set with various classification tools. The solution presented in this paper was chosen because of the convenient visualization and interpretation features of CA and the availability of

Table 1. List of remaining variables after variable selection.

Alosaus	Alosdeut	BJ01bis19	BJ19bis48
BJ49bis57	BJ58bis62	BJ63bis72	BJ73bis82
BJ83bis92	BJ93bis01	BJbis1900	Fins
GebBest10ins	GebBest1und2ins	GebBest3ins	Gebzus
GenNeuins	Hhins	HWBinsA	HWBinsD
InstdtUmZuzus	Kins	LKW	Mins
PKW	Sozempfzus	sozvpflBeschAus	sozvpflBeschDeut
Sterbzus	WanZuzus	WGebäuins	Zugmaschine

K-means as a standard tool with regard to future application by the City of Dortmund.

2 Material and methods

Material. The data set consists of 170 districts covering Dortmund and 200 variables collected in 2002. The variables give complete inventory of population (German, foreign, births, deaths, movements), unemployment, social welfare, buildings (stock, construction, covered area) and motor vehicles. All variables are measured as absolute frequencies except those representing areas. Area variables are given in square meters, the total area is given in hectares. We exclude the variables "total population" (`HBWins`) and "total area" (`Flächeha`) and use them for scaling. The data set is accompanied by geographical information. For each district a planar polygonal representation is given to account for spatial relations.

Variable selection and scaling. Many variables are related to each other by linear combination or high correlation. We scan variables belonging to one topic separately. For each group of variables with mutual correlation coefficients exceeding ± 0.7, one representative is selected manually. If possible, the variable containing a grand total is preferred. The four variables regarding welfare recipients are merged into one variable (`Sozempfzus`). The remaining variables are scaled by their median absolute deviation. Variables with median absolute deviation of zero are removed since these do not contain a considerable amount of information. The preselection process results in 32 variables shown in Table 1.

The remaining variables are measured in absolute frequencies scaled by median absolute deviation per variable. Since the city districts differ with respect to population and area, the variables are not comparable between districts. To correct for district densities, for example small areas with high population or large areas with low population, each district is scaled by its density, that is overall district population (`HBWins`) divided by overall district area (`Flächeha`). After variable selection and scaling, the remaining data matrix contains informative variables on comparable scale.

Correspondence analysis. Given a matrix of absolute frequencies, we can compute the χ^2 test statistic for homogeneity. Similar to principal component analysis, CA provides the mapping of variables into a lower space

while preserving a considerable percentage of χ^2. CA maps row and column vectors simultaneously into the same space. A row and a column vector are positively associated if they point to the same direction, that is they share a small angle. The more remote they are from the origin, the more associated they are. Row and column vectors pointing to opposite directions can be interpreted as negatively associated. In the following, we recall the main concepts of CA. For detailed theory and further concepts see for example Mirkin (1996, chap. 2.3.3) and Nakayama (2001).

Let $\mathbf{N}$ be a matrix of absolute frequencies with r rows, c columns, entries n_{ij} and grand total n. Hence, $\mathbf{F} = n^{-1}\mathbf{N}$ is the matrix of corresponding relative frequencies. We further define $\mathbf{D}_r$ as the $r \times r$ diagonal matrix of mean row profile $(n_{1.}/n, \ldots, n_{r.}/n)$ and $\mathbf{D}_c$ as the $c \times c$ diagonal matrix of mean column profile $(n_{.1}/n, \ldots, n_{.c}/n)$, where $n_{i.}$ and $n_{.j}$ are row and column sums of $\mathbf{N}$. The distance between two columns j and j' of $\mathbf{N}$ can be measured in χ^2 distance d^2 as

$$d^2(j, j') = \sum_{i=1}^{r} \frac{n}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}} \right)^2 .$$

The χ^2 distance is related to the Euclidean distance: If we apply the transformation $\tilde{\mathbf{F}} = \mathbf{D}_r^{-1/2}\mathbf{F}\mathbf{D}_c^{-1}$, then $d(j, j')$ is the Euclidean distance between columns j and j' of $\tilde{\mathbf{F}}$.

The objective of CA is to find a mapping of the columns of $\mathbf{N}$ from space $\mathbb{R}^r$ to a lower dimensional space $\mathbb{R}^m$ with $m < r$. The relative positions of two columns in χ^2 distance before mapping and Euclidean distance after mapping are preserved. A suitable mapping is found by singular value decomposition of the matrix $\mathbf{S} = \mathbf{D}_r^{-1/2}\mathbf{F}\mathbf{D}_c^{-1/2} - \mathbf{D}_r^{1/2}\mathbf{1}_r\mathbf{1}_c^T\mathbf{D}_c^{1/2}$. Singular value decomposition decomposes $\mathbf{S}$ into $\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{W}^T$. The matrices $\mathbf{V}$ and $\mathbf{W}$ contain left and right singular vectors respectively. Matrix $\mathbf{\Lambda}$ is the diagonal matrix of positive singular values $\lambda_1 \geq \cdots \geq \lambda_k > 0$ with $k \leq \min(r, c)$. The columns of $\mathbf{N}$ can be presented in space $\mathbb{R}^k$ as columns of $\mathbf{C} = (\lambda_1\mathbf{D}_c^{-1/2}w_1, \ldots, \lambda_k\mathbf{D}_c^{-1/2}w_k)$, where w_s denotes the sth column of $\mathbf{W}$. Applying the same considerations to the rows of $\mathbf{N}$ it follows that rows can be presented as columns of $\mathbf{R} = (\lambda_1\mathbf{D}_r^{-1/2}v_1, \ldots, \lambda_k\mathbf{D}_r^{-1/2}v_k)$, where v_s denotes the sth column of $\mathbf{V}$.

The total sum of squares of $\mathbf{S}$ is equal to the χ^2 statistic of $\mathbf{N}$ divided by n. The value χ^2/n is called *inertia*, a term that interprets relative frequency as mass. Due to singular value decomposition, the total inertia is equal to the sum of squared singular values of $\mathbf{S}$. The proportion of inertia explained by the first singular value is then $\lambda_1^2/\sum_{s=1}^{k}\lambda_s^2$. Regarding a chosen proportion of explained inertia, we represent the data in a lower dimensional space $\mathbb{R}^m$ by only considering the first m columns of $\mathbf{C}$ and $\mathbf{R}$.

K-means clustering. The K-means algorithm of Hartigan and Wong (1979) divides data points into K clusters. Initially, K data points are chosen

randomly as cluster centroids. In an iterative process, the algorithm assigns a data point to a cluster if this minimizes the within-cluster sum of squares of Euclidean distances to the corresponding cluster centroid. After reallocation of data points, cluster centroids are updated by replacing them with the centroid of within-cluster points. The iteration stops in a local minimum where no further reallocation of data points by this rule reduces the within-cluster sum of squares.

The number of clusters K is chosen such that the resulting partition of data points into clusters $C_1, \ldots, C_K$ is optimal with respect to a cluster validation index. A suitable index can be derived from silhouette scores, Rousseeuw (1987). For data point x_i assigned to cluster C_k, the average Euclidean distance a_i between point x_i and its within-cluster points $x_j \in C_k, j \neq i$, is defined as $a_i = |C_k|^{-1} \sum_{x_j \in C_k} \|x_i - x_j\|_2$, where $\|\cdot\|_2$ denotes the Euclidean norm. For all clusters not containing point x_i, the average Euclidean distance between x_i and within-cluster points is computed. The minimum b_i of these values, that is the average distance of x_i to its nearest neighboring cluster, is defined as $b_i = \min_{l \neq k} |C_l|^{-1} \sum_{x_j \in C_l} \|x_i - x_j\|_2$. The silhouette score s_i of point x_i is then given as $s_i = (b_i - a_i)/(\max\{a_i, b_i\})$ and ranges from -1 to 1. A score near 1 supports the allocation of x_i to an appropriate cluster. A point allocated to an inadequate cluster has a score near -1. A general validation index for a specific K-means partition is given by the global silhouette score, that is silhouette scores averaged over all points. The optimal number of clusters is then chosen from the partition with highest observed global silhouette score.

Two runs of K-means can lead to different partitions if the initial cluster centroids were chosen with random seeds. The process is considered to be stable if a reasonable percentage of several runs with K random centroids results in identical partitions. Therefore, we do not only consider the global silhouette score as a criterion for an optimal number of clusters but also the stability of the process under reruns. A process is stable if the resulting partitions are highly concentrated, that is, a small number of non-identical partitions is observed with high frequency when several runs are conducted. We allow a difference in one point to call two partitions identical. Given l non-identical partitions with observed relative frequencies $\hat{\pi}_j$, $j = 1, \ldots, l$, we measure concentration via normalized entropy $\mathrm{NE}_K = -\sum_{j=1}^{l} \hat{\pi}_j \cdot \log \hat{\pi}_j / \log l$. A low value of normalized entropy corresponds to high concentration. We base the final decision about the optimal number of clusters on a two-step combination of global silhouette scores and normalized entropy.

The K-means clustering is performed on the transformed district points after application of CA. District points that cluster together are considered to be associated to the same set of variable points, thus allowing for the interpretation that district points within a cluster share a common frequency profile on a set of variables. For data analysis, we use the statistical software

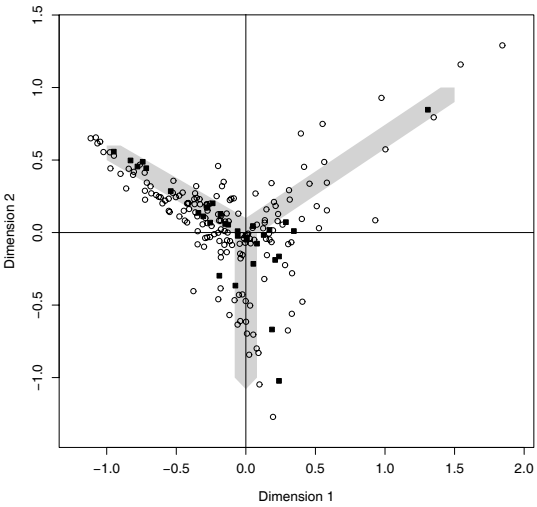


Fig. 1. Biplot of districts (circles) and variables (squares) on first two dimensions after CA mapping. Background arrows denote main directions of association.

R, version 1.7.1, with incorporated functions `kmeans` and `silhouette`, Ihaka and Gentleman (1996). Source code is available on request.

3 Results

After variable selection, the scaled data matrix containing 170 districts and 32 variables is submitted to CA. Figure 1 shows the two-dimensional biplot of districts and variables after mapping. Districts and variables that point to the same direction are positively associated. From this low dimensional plot we can guess at least three clusters of districts which are associated with the same variables. Many districts scatter around the origin and are not associated with a particular variable.

We apply K-means clustering on the districts after CA transformation. As each additional CA dimension explains a smaller percentage of inertia than those before, we first maximize the global silhouette score with respect to the number of CA dimensions m and the number of clusters K. For $K = 2, \ldots, 10$ and $m = 3, \ldots, 11$, we conduct 1 000 partitions each and compute the average global silhouette score for each combination. The maximal average score is reached for partitions with 4 clusters on 3 dimensions followed closely by 5 clusters on 3 dimensions. The latter has a three times lower standard error. Regarding the combination of mean score and standard error we choose 5 clusters on 3 dimensions which explain 58.5% of total inertia.

The 5-means clustering on 3 dimensions is optimal regarding its concentration. We conduct 10 000 runs for each combination of dimensions and clusters as above. The minimal normalized entropy is reached for 5 clusters

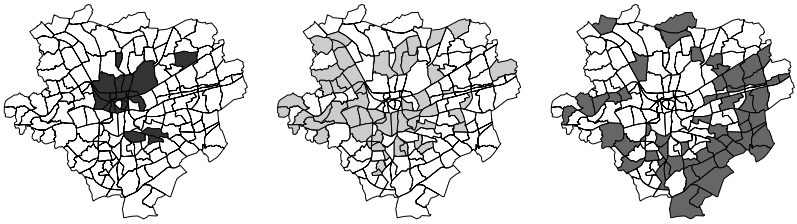


Fig. 2. Maps of inner city cluster (21 districts), western circle (57 districts) and eastern circle (59 districts).

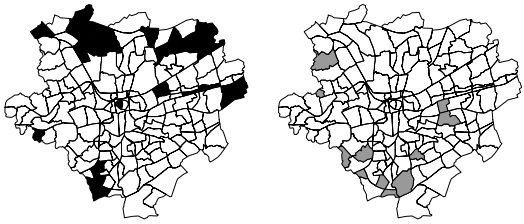


Fig. 3. Maps of northern suburbs (22 districts) and small cluster (11 districts).

and 3 dimensions where the algorithm converges in 69% of all runs to the cluster pattern shown in Figures 2 and 3. Of all partitions with $K = 5$ and $m = 3$ found in 10 000 runs, the presented pattern has not the highest global silhouette score. However, only two other partitions were found with convergence rates higher than 3% (12% and 11%), and those partitions with global silhouette scores higher than the presented one were found in a total of only 5% of all runs.

In addition, a randomization test for spatial correlation is performed to test whether neighboring districts in CA are also near in reality. A spatial correlation test compares two distance matrices corresponding to the same variables by correlation coefficient and assigns a randomization p-value, Manly (2001, chap. 9). One distance matrix contains the mutual Euclidean distances based on the first three CA dimensions. The entries of the second matrix are 1 if two districts are real neighbors, that is, if they share at least one point in their polygonal representation, and 0 if not. The two matrices are weakly negatively correlated with a correlation coefficient of -0.10 but show a significant p-value < 0.0001 based on 10 000 randomizations. Thus, clustering on CA mapping corresponds to the simple neighboring relationship of districts.

Dortmund is divided into a small, two mediate and two large clusters of districts. As suggested by the spatial correlation analysis, most members of a cluster are geographically neighboring. The inner city districts cluster together and are surrounded by the two large clusters. The latter span a circle of outer city districts, accumulating the western and eastern districts respectively. A fourth cluster contains mostly northern rural suburbs but also

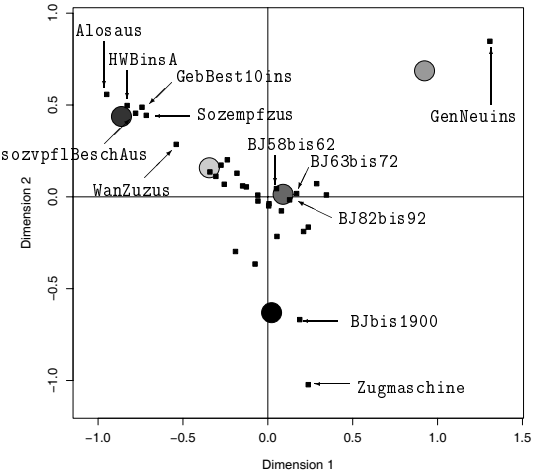


Fig. 4. Two-dimensional biplot of cluster centroids (circles) and variables (squares). Colors correspond to clusters in Figures 2 and 3. Labeled variables are positively associated to nearest cluster centroid based on first three CA dimensions.

eastern industrial areas and a part of the city center. The smallest cluster is scattered over Dortmund with emphasis on southern districts. Figures 2 and 3 show maps of each cluster versus the others.

Finally, we identify those variables that are highly associated with the cluster centroids. In three-dimensional space we compute the lengths of variable vectors and the angles between centroids and variables. The smaller the angle and the further apart from the origin the variable point is, the higher is the association to the corresponding cluster. We select variables with an angle smaller than 0.3 radian and a vector length greater than 0.4.

The analysis shows that all clusters are associated with one or more variables except the western circle which is simply not associated with any variable. Figure 4 shows a biplot similar to Figure 1 but with districts represented by their cluster centroids. Variables that are highly associated with the nearest cluster centroid are labeled.

Six variables are positively associated to the inner city cluster: Social welfare recipients (`Sozempfzus`), unemployed foreigners (`Alosaus`), foreign population (`HWBinsA`), foreigners subjected to social insurance contribution (`sozvpflBeschAus`), immigration (`WanZuzus`) and residential building stock with 10 and more apartments (`GebBest10ins`). Districts of the inner city cluster are separable from other districts with respect to similar high relative frequencies of these variables. For interpreting these variables in a social context one has to keep in mind that all remaining variables were selected as representatives of their group.

The eastern circle is associated to building stock variables spanning the sixties and eighties as years of construction (`BJ58bis62`, `BJ63bis72` and `BJ82bis92`). Districts in this cluster have similar high relative frequencies

of postwar building stock. The cluster of northern suburbs shows high frequencies of very old houses (`BJbis1900`) and tractors (`Zugmaschine`).

The small cluster is positively associated with the variable `GenNeuins`, that is total building permits for residential buildings with living space. The interpretation of building permits with respect to social environment is difficult: Large numbers of new buildings either represent development areas that usually attract young families or represent dense housing areas of lower social level.

4 Conclusion

After variable selection and preprocessing, the remaining data set was submitted to correspondence analysis. Examination of the first two dimensions after mapping already suggested that the districts do not form a homogeneous entity. Considering the first three new dimensions, a stable partition into five clusters was found: The inner city and adjacent districts, a western and an eastern circle, a cluster of northern suburbs and industrial areas and a small cluster scattered over the outer city area.

For four clusters we identified highly associated variables which represent population structure and building stocks. However, the interpretation regarding the social environment is restricted to the available variables. The western city cluster is not particularly associated with any variable and represents the cluster of remaining districts after separation of other clusters.

The analysis reflects Dortmund's social conditions in 2002. To monitor changes in social or residential conditions, next year's data can be submitted to the same procedures with reduced variables as in 2002. Few changes in 2003 will probably result in a similar partition, whereas dramatic changes may lead to another number of clusters and other associated variables.

References

HARTIGAN, J. A. and WONG, M. A. (1979): Algorithm AS 136: a k-means clustering algorithm, *Applied Statistics, 28, 100–108*.

IHAKA, R. and GENTLEMAN, R. (1996): R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics, 5(3), 299–314*. Software: http://www.r-project.org/.

MANLY, B.F.J. (2001): *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Chapman & Hall, London.

MIRKIN, B. (1996): *Mathematical Classification and Clustering*, Kluwer Academic Publishers, Dordrecht.

NAKAYAMA, T. (2001): Tests for redundancy of some variables in correspondence analysis, *Hiroshima Mathematical Journal, 31, 1–34*.

ROUSSEEUW, P.J. (1987): Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics, 20, 53–65*.

Keywords

Authors