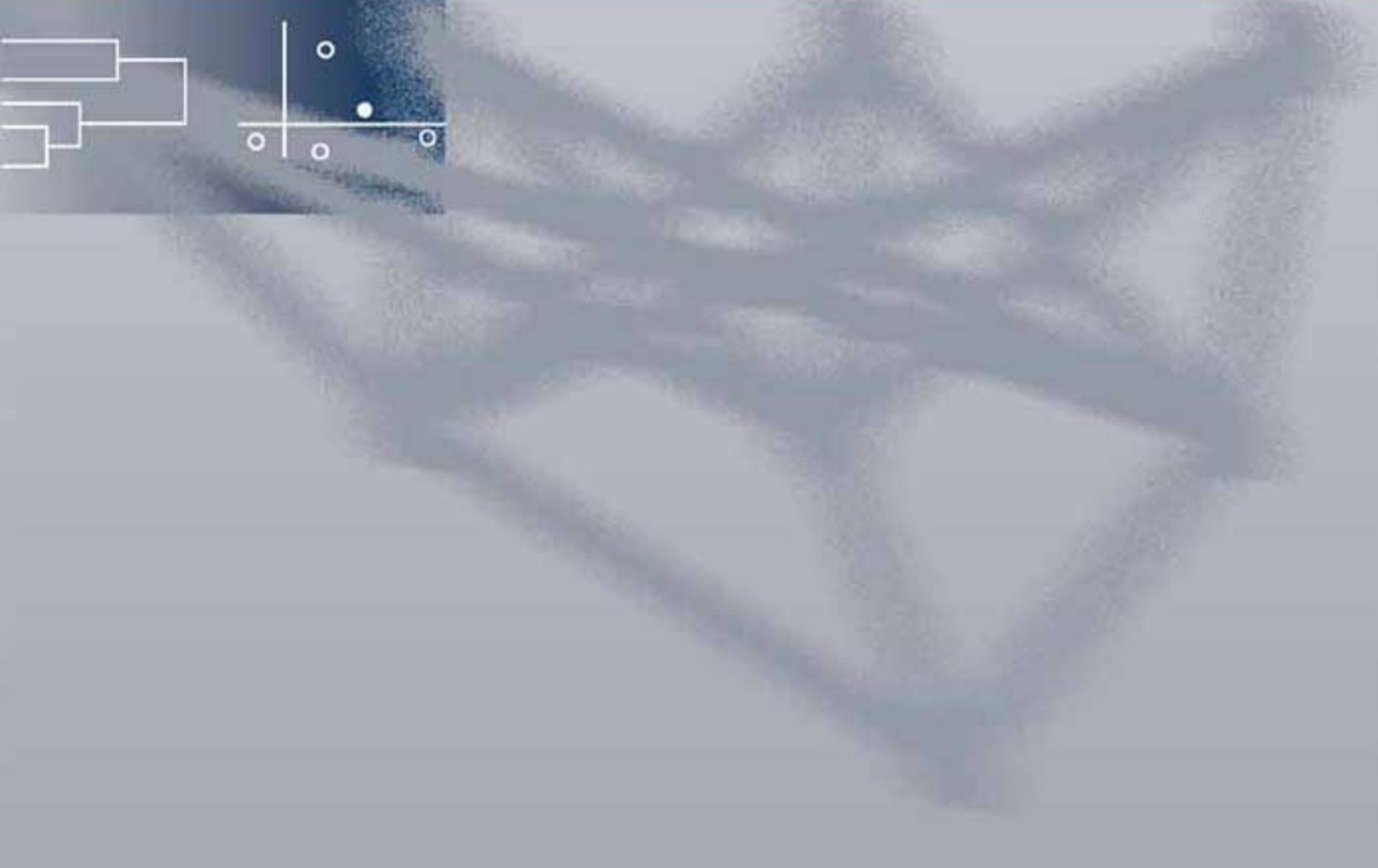


**STUDIES IN CLASSIFICATION,
DATA ANALYSIS,
AND KNOWLEDGE ORGANIZATION**

D. Baier
K.-D. Wernecke
Editors

Innovations in Classification, Data Science, and Information Systems



Studies in Classification, Data Analysis, and Knowledge Organization

Managing Editors

H.-H. Bock, Aachen
W. Gaul, Karlsruhe
M. Vichi, Rome

Editorial Board

Ph. Arabie, Newark
D. Baier, Cottbus
F. Critchley, Milton Keynes
R. Decker, Bielefeld
E. Diday, Paris
M. Greenacre, Barcelona
C. Lauro, Naples
J. Meulman, Leiden
P. Monari, Bologna
S. Nishisato, Toronto
N. Ohsumi, Tokyo
O. Opitz, Augsburg
G. Ritter, Passau
M. Schader, Mannheim
C. Weihs, Dortmund

Titles in the Series

- H.-H. Bock and P. Ihm (Eds.)
Classification, Data Analysis,
and Knowledge Organization. 1991
(out of print)
- M. Schader (Ed.)
Analyzing and Modeling Data
and Knowledge. 1992
- O. Opitz, B. Lausen, and R. Klar (Eds.)
Information and Classification. 1993
(out of print)
- 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 Millennium.
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, 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

Daniel Baier
Klaus-Dieter Wernecke
Editors

Innovations in Classification, Data Science, and Information Systems

Proceedings of the 27th Annual Conference
of the Gesellschaft für Klassifikation e.V., Brandenburg
University of Technology, Cottbus, March 12–14, 2003

With 143 Figures and 111 Tables

Prof. Dr. Daniel Baier
Chair of Marketing and Innovation Management
Institute of Business Administration and Economics
Brandenburg University of Technology Cottbus
Konrad-Wachsmann-Allee 1
03046 Cottbus
Germany
daniel.baier@tu-cottbus.de

Prof. Dr. Klaus-Dieter Wernecke
Department of Medical Biometrics
Charité Virchow-Klinikum
Humboldt University Berlin
13344 Berlin
Germany
klaus-dieter.wernecke@charite.de

ISBN 3-540-23221-4 Springer-Verlag Berlin Heidelberg New York

Library of Congress Control Number: 2004114682

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer · Part of Springer Science+Business Media

springeronline.com

© Springer-Verlag Berlin · Heidelberg 2005
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Softcover-Design: Erich Kirchner, Heidelberg

SPIN 11326427 43/3130/DK - 5 4 3 2 1 0 - Printed on acid-free paper

Preface

This volume contains revised versions of selected papers presented during the 27th Annual Conference of the Gesellschaft für Klassifikation (GfKl), the German Classification Society. The conference was held at the Brandenburg University of Technology (BTU) Cottbus, Germany, in March 2003. Klaus-Dieter Wernecke chaired the program committee, Daniel Baier was the local organizer. Krzysztof Jajuga and Andrzej Sokolowski and their colleagues in Sekcja Klasyfikacji i Analizy Danych (SKAD), the Polish Classification Society, provided strong support during all phases of the conference.

The program committee was able to select 124 talks for 36 sessions. Additionally, it was possible to recruit 19 notable and internationally renowned invited speakers for plenary and semi-plenary talks on their current research work regarding the conference topic "Innovations in Classification, Data Science, and Information Systems" or, respectively, on the GfKl members' general fields of interest "Classification, Data Analysis, and Knowledge Organization". Thus, the conference, which was traditionally designed as an interdisciplinary event, again provided a large number of scientists and experts from Germany and abroad with an attractive forum for discussions and the mutual exchange of knowledge.

Besides on traditional subjects, the talks in the different sections focused on topics such as Methods of Data Analysis for Business Administration and Economics as well as Medicine and Health Services. This suggested the presentation of the papers of the volume in the following eight chapters:

- Discrimination and Clustering,
- Probability Models and Statistical Methods,
- Pattern Recognition and Computational Learning,
- Time Series Analysis,
- Marketing, Retailing, and Marketing Research,
- Finance, Capital Markets, and Risk Management,
- Production, Logistics, and Controlling,
- Medicine and Health Services.

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

- BTU Cottbus,
- Chair of Marketing and Innovation Management, BTU Cottbus,
- Holiday Inn Hotel, Cottbus,
- MTU Maintenance Berlin-Brandenburg GmbH, Ludwigsfelde,
- Scicon Scientific Consulting GmbH, Karlsruhe,
- Sparkasse Spree-Neiße, Cottbus,

- Synergy Microwave Europe GmbH & Co. KG, München,
- Volkswagen AG, Wolfsburg, and
- various producers of Scottish single malt whisky

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

Additionally, we wish to express our gratitude towards 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 Dr. Alexandra Rese who made an excellent job in organizing the refereeing process and preparing this volume. We also wish to thank Michael Brusch and his GfKI-2003 team for perfectly organizing the conference and helping to prepare the final program. In this context, special thanks are given to Jörg Swienty, Nadja Schütz, Matthias Kaiser, Christoph Schauenburg, and other members of the Chair of Marketing and Innovation Management, BTU Cottbus.

Finally, we want to thank Dr. Martina Bihm of Springer-Verlag, Heidelberg, for her support and dedication to the production of this volume.

Cottbus and Berlin, September 2004

*Daniel Baier
Klaus-Dieter Wernecke*

Contents

Part I. Discrimination and Clustering

A New Agglomerative 2-3 Hierarchical Clustering Algorithm	3
<i>Sergiu Chelcea, Patrice Bertrand, Brigitte Trousse</i>	
Symbolic Classifier with Convex Hull Based Dissimilarity Function	11
<i>Francisco de A.T. de Carvalho, Simith T. D'Oliveira Júnior</i>	
Two-Mode Cluster Analysis via Hierarchical Bayes	19
<i>Wayne S. DeSarbo, Duncan K. H. Fong, John Liechty</i>	
On Application of a Certain Classification Procedure to Mean Value Estimation Under Double Sampling for Nonresponse	30
<i>Wojciech Gamrot</i>	
Regression Clustering with Redescending M-Estimators	38
<i>Tim Garlipp, Christine H. Müller</i>	
ClusCorr98 - Adaptive Clustering, Multivariate Visualization, and Validation of Results	46
<i>Hans-Joachim Mucha, Hans-Georg Bartel</i>	
Stratification Before Discriminant Analysis: A Must?	54
<i>Jean-Paul Rasson, Jean-Yves Pirçon, François Roland</i>	
An Exchange Algorithm for Two-Mode Cluster Analysis	62
<i>Manfred Schwaiger, Raimund Rix</i>	
Model-Based Cluster Analysis Applied to Flow Cytometry Data	69
<i>Ute Simon, Hans-Joachim Mucha, Rainer Brüggemann</i>	
On Stratification Using Auxiliary Variables and Discriminant Method	77
<i>Marcin Skibicki</i>	
Measuring Distances Between Variables by Mutual Information	81
<i>Ralf Steuer, Carsten O. Daub, Joachim Selbig, Jürgen Kurths</i>	
Pareto Density Estimation: A Density Estimation for Knowledge Discovery	91
<i>Alfred Ultsch</i>	

Part II. Probability Models and Statistical Methods

Modelling the Claim Count with Poisson Regression and Negative Binomial Regression	103
<i>Bartłomiej Bartoszewicz</i>	
Chemical Balance Weighing Design with Different Variances of Errors	111
<i>Bronisław Ceranka, Małgorzata Graczyk</i>	
Combination of Regression Trees and Logistic Regression to Analyse Animal Management and Disease Data	120
<i>Susanne Dahms</i>	
Robustness of ML Estimators of Location-Scale Mixtures	128
<i>Christian Hennig</i>	
On the Modification of the David-Hellwig Test	138
<i>Grzegorz Konczak</i>	
Simultaneous Selection of Variables and Smoothing Parameters in Additive Models	146
<i>Rüdiger Krause, Gerhard Tutz</i>	
Multiple Change Points and Alternating Segments in Binary Trials with Dependence	154
<i>Joachim Krauth</i>	
Outlier Identification Rules for Generalized Linear Models ...	165
<i>Sonja Kuhnt, Jörg Pawlitschko</i>	
Dynamic Clustering with Non-Quadratic Adaptive Distances for Interval-Type Data.....	173
<i>Renata M. C. R. de Souza, Francisco de A. T. de Carvalho</i>	
Partial Moments and Negative Moments in Ordering Asymmetric Distributions	181
<i>Grazyna Trzpiot</i>	

Part III. Pattern Recognition and Computational Learning

Classification of Method Fragments Using a Reference Meta Model	191
<i>Werner Esswein, Andreas Gehlert</i>	

Finding Metabolic Pathways in Decision Forests	199
<i>André Flöter, Joachim Selbig, Torsten Schaub</i>	
Randomization in Aggregated Classification Trees	207
<i>Eugeniusz Gatnar</i>	
Data Mining – The Polish Experience	217
<i>Eugeniusz Gatnar, Dorota Rozmus</i>	
Extracting Continuous Relevant Features	224
<i>Amir Globerson, Gal Chechik, Naftali Tishby</i>	
Individual Rationality Versus Group Rationality in Statistical Modelling Issues	239
<i>Daniel Kosiorowski</i>	
Mining Promising Qualification Patterns	249
<i>Ralf Wagner</i>	

Part IV. Time Series Analysis

Partial Correlation Graphs and Dynamic Latent Variables for Physiological Time Series	259
<i>Roland Fried, Vanessa Didelez, Vivian Lanius</i>	
Bootstrap Resampling Tests for Quantized Time Series	267
<i>Jacek Leśkow, Cyprian Wronka</i>	
Imputation Strategies for Missing Data in Environmental Time Series for an Unlucky Situation	275
<i>Daria Mendola</i>	
Prediction of Notes from Vocal Time Series: An Overview	283
<i>Claus Weihs, Uwe Ligges, Ursula Garczarek</i>	
Parsimonious Segmentation of Time Series by Potts Models . .	295
<i>Gerhard Winkler, Angela Kempe, Volkmar Liebscher, Olaf Wittich</i>	

Part V. Marketing, Retailing, and Marketing Research

Application of Discrete Choice Methods in Consumer Preference Analysis	305
<i>Andrzej Bąk, Aneta Rybicka</i>	
Competition Analysis in Marketing Using Rank Ordered Data	313
<i>Reinhold Decker, Antonia Hermelbracht</i>	

Handling Missing Values in Marketing Research Using SOM .	322
<i>Mariusz Grabowski</i>	
Applicability of Customer Churn Forecasts in a Non-Contractual Setting .	330
<i>Jörg Hopmann, Anke Thede</i>	
A Gravity-Based Multidimensional Unfolding Model for Preference Data .	338
<i>Tadashi Imaizumi</i>	
Customer Relationship Management in the Telecommunications and Utilities Markets .	346
<i>Robert Katona, Daniel Baier</i>	
Strengths and Weaknesses of Support Vector Machines Within Marketing Data Analysis .	355
<i>Katharina Monien, Reinhold Decker</i>	
Classification of Career-Lifestyle Patterns of Women .	363
<i>Miki Nakai</i>	
Joint Space Model for Multidimensional Scaling of Two-Mode Three-Way Asymmetric Proximities .	371
<i>Akinori Okada, Tadashi Imaizumi</i>	
Structural Model of Product Meaning Using Means-End Approach .	379
<i>Adam Sagan</i>	
The Concept of Chains as a Tool for MSA Contributing to the International Market Segmentation .	388
<i>Elżbieta Sobczak</i>	
Statistical Analysis of Innovative Activity .	396
<i>Marek Szajt</i>	
The Prospects of Electronic Commerce: The Case of the Food Industry .	406
<i>Ludwig Theeuwesen</i>	
<hr/>	
Part VI. Finance, Capital Markets, and Risk Management	
Macroeconomic Factors and Stock Returns in Germany .	419
<i>Wolfgang Bessler, Heiko Opfer</i>	

Application of Classification Methods to the Evaluation of Polish Insurance Companies	427
<i>Marta Borda, Patrycja Kowalczyk-Lizak</i>	
Analytic Hierarchy Process – Applications in Banking	435
<i>Czesław Domański, Jarosław Kondrasiuk</i>	
Tail Dependence in Multivariate Data – Review of Some Problems	446
<i>Krzysztof Jajuga</i>	
The Stock Market Performance of German Family Firms	454
<i>Jan Kuklinski, Felix Lowinski, Dirk Schiereck, Peter Jaskiewicz</i>	
Testing of Warrants Market Efficiency on the Warsaw Stock Exchange – Classical Approach	461
<i>Agnieszka Majewska, Sebastian Majewski</i>	
Group Opinion Structure: The Ideal Structures, their Relevance, and Effective Use	471
<i>Jan W. Owsiński</i>	
Volatility Forecasts and Value at Risk Evaluation for the MSCI North America Index	482
<i>Momtchil Pajarliev, Wolfgang Polasek</i>	
Selected Methods of Credibility Theory and its Application to Calculating Insurance Premium in Heterogeneous Insurance Portfolios	490
<i>Wanda Ronka-Chmielowiec, Ewa Poprawska</i>	
Support Vector Machines for Credit Scoring: Extension to Non Standard Cases	498
<i>Klaus B. Schebesch, Ralf Stecking</i>	
Discovery of Risk-Return Efficient Structures in Middle-Market Credit Portfolios	506
<i>Frank Schlottmann, Detlef Seese</i>	
Approximation of Distributions of Treasury Bill Yields and Interbank Rates by Means of α -stable and Hyperbolic Distributions	515
<i>Witold Szczepaniak</i>	
Stability of Selected Linear Ranking Methods – An Attempt of Evaluation for the Polish Stock Market	523
<i>Waldemar Tarczyński, Małgorzata Łuniewska</i>	

Part VII. Production, Logistics, and Controlling

A Two-Phase Grammar-Based Genetic Algorithm for a Workshop Scheduling Problem	535
<i>Andreas Geyer-Schulz, Anke Thede</i>	
Classification and Representation of Suppliers Using Principle Component Analysis.....	544
<i>Rainer Lasch, Christian G. Janker</i>	
A Knowledge Based Approach for Holistic Decision Support in Manufacturing Systems	552
<i>Uwe Meinberg, Jens Jakobza</i>	
Intelligent Fashion Interfaces – Questions to New Challenges of Classifying	559
<i>Astrid Ullsperger</i>	
Full Factorial Design, Taguchi Design or Genetic Algorithms – Teaching Different Approaches to Design of Experiments	567
<i>Ralf Woll, Carina Burkhard</i>	

Part VIII. Medicine and Health Services

Requirement-Driven Assessment of Restructuring Measures in Hospitals	577
<i>Werner Esswein, Torsten Sommer</i>	
Analyzing Protein Data with the Generative Topographic Mapping Approach	585
<i>Isabelle M. Grimenstein, Wolfgang Urfer</i>	
How Can Data from German Cancer Registries Be Used for Research Purposes?	593
<i>Alexander Katalinic</i>	
Probabilistic Record Linkage of Anonymous Cancer Registry Records	599
<i>Martin Meyer, Martin Radespiel-Tröger, Christine Vogel</i>	
An Empirical Study Evaluating the Organization and Costs of Hospital Management	605
<i>Karin Wolf-Ostermann, Markus Lüngen, Helmut Mieth, Karl W. Lauterbach</i>	
Index	613

Part I

Discrimination and Clustering

A New Agglomerative 2-3 Hierarchical Clustering Algorithm

Sergiu Chelcea¹, Patrice Bertrand^{2,3}, and Brigitte Trousse¹

¹ INRIA, AxIS Research Group, BP 93, 06902 Sophia-Antipolis Cedex, France

² GET-ENST Bretagne, IASC, Technopôle Brest-Iroise
CS 83818, 29238 BREST Cedex, France

³ INRIA, Axis Research Group, BP 105, 78 153 Le Chesnay Cedex, France

Abstract. We studied a new general clustering procedure, that we call here Agglomerative 2-3 Hierarchical Clustering (2-3 AHC), which was proposed in Bertrand (2002a, 2002b). The three main contributions of this paper are: first, the theoretical study has led to reduce the complexity of the algorithm from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 \log n)$. Secondly, we proposed a new 2-3 AHC algorithm that simplifies the one proposed in 2002 (its principle is closer to the principle of the classical AHC). Finally, we proposed a first implementation of a 2-3 AHC algorithm.

1 Motivations

Our motivation concerns the use of clustering techniques for user profiling and case indexing inside a Case-Based Reasoning framework (Jaczynski (1998)). It is in this context that we studied a new clustering strategy, called Agglomerative 2-3 Hierarchical Clustering (2-3 AHC). This strategy was recently proposed in Bertrand (2002a, 2002b) to generalize and to make more flexible the Agglomerative Hierarchical Clustering method (AHC).

Section 2 briefly presents the concept of 2-3 hierarchy together with the 2-3 AHC algorithm introduced in Bertrand (2002a). Section 3 derives a new 2-3 AHC algorithm while proposing to integrate the refinement step into the merging step. Before concluding in Section 5, Section 4 presents the complexity analysis, the implementation and some tests.

2 The 2-3 hierarchies and the 2-3 AHC algorithm

The following definitions and results of this section were established in Bertrand (2002a), in order to extend the framework of hierarchies¹.

In this text, we denote as E an arbitrary set of n objects to be clustered, and we suppose that E is described by a *dissimilarity*, say δ , i.e. $\delta(x, y)$ indicates the degree of dissimilarity between two arbitrary objects x and y .

¹ For the usual definitions in classification the reader can refer to (Gordon 1999).

2.1 2-3 Hierarchies

We consider a collection \mathcal{C} of nonempty subsets of E , often called *clusters* in the rest of the text. If $X, Y \in \mathcal{C}$ satisfy $X \cap Y \neq \emptyset$, $X \not\subseteq Y$ and $Y \not\subseteq X$, then it will be said that X *properly intersects* Y . A *successor* of $X \in \mathcal{C}$ is any largest cluster, say X' , that is strictly contained in X . If X' is a successor of X , then X is said to be a *predecessor* of X' (see Figure 1). The collection \mathcal{C} is said to be *weakly indexed* by a map $f : \mathcal{C} \rightarrow \mathbf{R}^+$ if $X \subset Y$ implies $f(X) \leq f(Y)$ and if $f(X) = f(Y)$ with $X \subset Y$, implies that X is equal to the intersection of its predecessors. We recall also that \mathcal{C} is said to be *indexed* by f if $X \subset Y$ implies $f(X) < f(Y)$. A *2-3 hierarchy* on E is a collection \mathcal{C} which contains E and its singletons, which is closed under nonempty intersections, and such that each element of \mathcal{C} properly intersects no more than one other element of \mathcal{C} . A small example of a 2-3 hierarchy is presented in Figure 1a.

A 2-3 hierarchy on E is a family of intervals of at least a linear order defined on E . This property allows to represent graphically a 2-3 hierarchy as a pyramidal classification (cf. Figure 1). According to Theorem 3.3 in Bertrand (2002a), any 2-3 hierarchy on E has a maximum size of $\lfloor \frac{3}{2}(n-1) \rfloor$, excluding the singletons. In the following, we will say that two clusters X and Y are *noncomparable* if $X \not\subseteq Y$ and $Y \not\subseteq X$, and that a cluster X is *maximal* if $\nexists Z \in \mathcal{C}$ such that $X \subset Z$.

2.2 From AHC to 2-3 AHC

We first recall that the principle of AHC is to merge repeatedly two clusters until the cluster E is formed, the initial clusters being all the singletons. Each cluster is merged only once, and two clusters can be merged if they are closest - in the sense of a chosen *aggregation link*, denoted μ , and called simply *link*. Usual links are *single link*, *complete link*, *average link* and *Ward link*. When two clusters X and Y are merged, the link $\mu(X, Y)$ between these two clusters can be interpreted as a measurement, denoted $f(X \cup Y)$, of the degree of heterogeneity of $X \cup Y$. In addition, if we set $f(X) = 0$ for $|X| = 1$, the so defined map f on the set of clusters is not necessarily a weak index in the sense of Section 2.1, so that a refinement step (removing of certain clusters) is performed, in order that f becomes a weak index.

The 2-3 AHC algorithm below (Bertrand (2002a)) extends the AHC.

Algorithm of the 2-3 AHC (Bertrand (2002a)):

1. **Initialization:** $i = 0$; The set of clusters and the set of candidate² clusters \mathcal{M}_i coincide with the set of singletons of E .
2. **Merge:** $i = i + 1$; Merge a pair $\{X_i, Y_i\}$ such that $\mu(X_i, Y_i) \leq \mu(X, Y)$, among the pairs $\{X, Y\} \subseteq \mathcal{M}_{i-1}$, which are noncomparable and satisfy α or β :
 - (α) X and Y are maximal, and X (resp. Y) is the only cluster that may properly intersect Y (resp. X).

- (β) One of X or Y is maximal, and the other admits a single predecessor Z . No cluster is properly intersected by X , Y or Z .
3. **Update:** $\mathcal{M}_i \leftarrow \mathcal{M}_{i-1} \cup \{X_i \cup Y_i\}$, from which we eliminate any cluster strictly included in at least a cluster of \mathcal{M}_{i-1} and in $X_i \cup Y_i$. Update μ by using an extension of Lance and Williams Formula. Update f by using $f(X_i \cup Y_i) = \max\{f(X_i), f(Y_i), \mu(X_i, Y_i)\}$.
 4. **Ending test:** repeat steps 2 et 3, until the cluster E is created.
 5. **Refinement:** remove some clusters so that f is a weak index.

It has been proved in Bertrand (2002a) that for any choice of μ , this algorithm converges in at most $O(n^3)$, that after each step of the algorithm, the set of created clusters (completed by E) is a 2-3 hierarchy (cf. Bertrand (2002a), Proposition 5.4), and that the final structure is weakly indexed.

3 Proposition of a new 2-3 AHC algorithm

We present here a new 2-3 AHC algorithm derived from the previous one and based on the ideas presented in the following two subsections. Besides a simpler formulation (cf. Fact 34), the interest of this new algorithm (cf. Section 3.2) is two-fold: first, its principle is more similar to the principle of the AHC algorithm (cf. Fact 33) and second, we will see that the integration of the refinement phase into the merging phase (cf. Fact 35), allows to reduce the complexity of the algorithm (cf. Section 4).

3.1 Modifying the update and the merging steps

We begin with a reformulation of the update of candidates set \mathcal{M}_i (Step 3).

Proposition 31 *In the 2-3 AHC algorithm, we can, without changing the results of the merging, choose \mathcal{M}_i (step 3) in the following way: \mathcal{M}_i equals $\mathcal{M}_{i-1} \cup \{X_i \cup Y_i\}$, from which we eliminate every successor of X_i or Y_i , and also the two clusters X_i and Y_i , if $X_i \cap Y_i \neq \emptyset$ or the merging of X_i and Y_i is of type β .*

Proof: In the initial algorithm, like in the new formulation, \mathcal{M}_i is equal to $\mathcal{M}_{i-1} \cup \{X_i \cup Y_i\}$, deprived of certain clusters included in $X_i \cup Y_i$. It is thus enough to compare the two ways of defining \mathcal{M}_i only for the clusters of \mathcal{M}_{i-1} which are included in $X_i \cup Y_i$. We first examine the successors of X_i or of Y_i . In the initial algorithm, they don't belong to \mathcal{M}_i , because they are included in X_i or Y_i , and in $X_i \cup Y_i$. It is also clearly the case in the new formulation. In addition, in both ways of choosing \mathcal{M}_i , if a cluster W is included in one of the successors of X_i (resp. Y_i), then W does not belong to \mathcal{M}_{i-1} , because W

² X is candidate if $\exists Y \in \mathcal{C}$ such that X and Y are noncomparable, and their merging satisfy the 2-3 hierarchy definition (conditions α and β below).

was already eliminated from $\mathcal{M}_{i'}$ with $i' \leq i - 1$ (we use the same arguments as for the elimination of the successors of X_i or Y_i , but to a stage previous to the formation of $X_i \cup Y_i$). Since X_i and Y_i are the only successors of $X_i \cup Y_i$, these are thus the only clusters left to examine, in order to determine if the choice of \mathcal{M}_i varies according to the two formulations for choosing \mathcal{M}_i .

There are only three possible cases according to whether the merging of X_i and Y_i , is (a) of the type α with $X_i \cap Y_i = \emptyset$, (b) of the type α with $X_i \cap Y_i \neq \emptyset$, and (c) of the type β .

Case (a): α merging of X_i and Y_i , with $X_i \cap Y_i = \emptyset$. In this case, $X_i \cup Y_i$ is the only cluster containing X_i (resp. Y_i), because X_i (resp. Y_i) was maximal before the creation of $X_i \cup Y_i$. Thus neither X_i nor Y_i are removed from \mathcal{M}_i in the initial algorithm, and also in the new formulation. It results that the two formulations are equivalent here.

Case (b): α merging of X_i and Y_i , with $X_i \cap Y_i \neq \emptyset$. Using the same argument as in case (a), we deduce that neither X_i nor Y_i are removed from \mathcal{M}_i in the initial algorithm. On the other hand, X_i and Y_i do not belong to \mathcal{M}_i , if the new formulation is used. However according to the initial algorithm, neither X_i nor Y_i will be aggregate during a later merging of this algorithm. Indeed on the one hand, none of the clusters X_i and Y_i can be used for a β type merging, because X_i and Y_i properly intersect each other. On the other hand, none of the clusters X_i and Y_i can be used for an α merging, because X_i and Y_i are not maximal any more. Thus, the pairs of clusters that can be merged are the same in the two approaches.

Case (c): β merging of X_i and Y_i . Let us suppose - without any loss of generality - that Z is the (only) predecessor of X_i . Thus $X_i \notin \mathcal{M}_i$ in the initial algorithm, but $Y_i \in \mathcal{M}_i$ because Y_i is included in only one cluster ($X_i \cup Y_i$). On the other hand, X_i and Y_i do not belong to \mathcal{M}_i , if the new formulation is used. However according to the initial algorithm, Y_i will not be aggregate during a later merging of the algorithm. Indeed, Y_i has a single predecessor $X_i \cup Y_i$ but $X_i \cup Y_i$ properly intersects Z (because Z strictly contains X_i but is disjoint of Y_i). Thus Y_i cannot be used for a β type merging, nor for an α type one. Thus, again the pairs of clusters that can be merged are the same in the two approaches, which finally proves that the new way of choosing \mathcal{M}_i does not change the possibilities of merging at each iteration. \square

The following property highlights the need of adding a merging step, that we call *intermediate merging* step, at the end of each β merging.

Proposition 32 *If the merging of the i^{th} step of the algorithm is of type β , then the cluster $X_i \cup Y_i$ formed at this stage, will necessarily be merged with the predecessor of X_i or Y_i , in a later step of the algorithm.*

Proof: Let us suppose - without any loss of generality - that Z is the (only) predecessor of X_i , before the β merging of X_i and Y_i . Let us place at the end of the β merging. Clearly $X_i \cup Y_i$ is maximal and $X_i \cup Y_i \in \mathcal{M}_i$.

Suppose that Z is not maximal, then $X_i \subset Z \subset Z'$, which implies that X_i has been eliminated from $\mathcal{M}_{i'}$ ($i' < i$) no later than during the update following

the creation of Z' : this contradicts $X_i \in \mathcal{M}_{i-1}$. Thus Z is maximal, and so $Z \in \mathcal{M}_i$, because a maximal cluster cannot be eliminated from any \mathcal{M}_j ($j \leq i$). It results that the clusters $X_i \cup Y_i$ and Z belonging to \mathcal{M}_i , are maximal and properly intersect themselves. Thus they can be merged together in an α merging. Moreover, cluster $X_i \cup Y_i$ (resp. cluster Z) can be merged together only with cluster Z (resp. cluster $X_i \cup Y_i$) according to algorithm conditions. Assume that these two clusters are not merged together. Then we would merge together two other clusters A and B . These clusters A and B cannot be neither successors of X_i or of Y_i , nor X_i or Y_i themselves by Proposition 31. Moreover, A and B cannot be Z or its successors, since Z already properly intersects $X_i \cup Y_i$. Thus A and B would be included in $E - (X_i \cup Y_i \cup Z)$. Otherwise, the algorithm ends only when cluster E is created and we known that it ends (cf. Bertrand 2002a). However E cannot be created as long as only clusters included in $E - (X_i \cup Y_i \cup Z)$ are merged, so as long as the merging of $X_i \cup Y_i$ and Z is not performed, which completes the proof. \square

3.2 New 2-3 AHC algorithm integrating the refinement step

We begin with three facts before presenting our new 2-3 AHC algorithm:

Fact 33 If at the end of any β merging of X_i and Y_i (i unspecified), we decide, following the Proposition 32, to merge $X_i \cup Y_i$ with the predecessor Z (of X_i or Y_i), then at the end of the so modified step 2, no cluster properly intersects a maximal cluster. In other words, *at the end of each modified step 2, the maximal clusters form a partition of E* , which underlines a strong analogy with the AHC algorithm characterized by this property.

Fact 34 For each i , the set \mathcal{M}_i represents all the maximal clusters plus their successors when these successors are disjoint. This is a direct consequence of Proposition 31 and to the fact that each merging creates a maximal cluster. It results (taking into account the significant remark according to which the maximal clusters are disjoint) that one reformulates the (α) and (β) conditions in the following way, where $X, Y \in \mathcal{M}_{i-1}$: (α) “ X and Y are maximal”, (β) “only one of the clusters X and Y is maximal”.



Fig. 1a.

Refinement example

Fig. 1b.

Fact 35 The refinement step can be integrated into the merging step, in order to obtain a weak indexing f . For this, each time we create a cluster $X \cup Y$, we compare $f(X \cup Y)$ with $f(X)$ and $f(Y)$. If $f(X \cup Y) = f(X)$ (resp. $f(X \cup Y) = f(Y)$), we remove X (resp. Y), provided that $X \cup Y$ is the only predecessor of X (resp. Y). This last case is illustrated in the example

from Figure 1 where $f(X) < f(Y) = f(X \cup Y)$: Y must then be eliminated from the structure.

New 2-3 AHC algorithm (see also Chelcea et al. (2002)):

1. **Initialization:** The candidate clusters set, \mathcal{M}_0 , is the set of singletons of E . Let $i = 0$.
2. a) **Merge:** Let $i = i + 1$; Merge two clusters X_i and Y_i which are closest (in the sense of μ) among the pairs from \mathcal{M}_{i-1} , which are noncomparable and such that at least one of them is maximal;
b) **Intermediate Merge:** If Z is a predecessor of the cluster X_i or Y_i such that $Z \neq X_i \cup Y_i$, then merge Z and $X_i \cup Y_i$, and eliminate from \mathcal{M}_i these two clusters and their successors.
3. **Refinement:** Eliminate any cluster $W \in \{X_i, Y_i, X_i \cup Y_i, Z\}$ such that W has one predecessor, W' , and such that $f(W) = f(W')$.
4. **Update:** Update \mathcal{M}_i by adding the last formed cluster and eliminating the successors of the merged clusters and also the merged clusters if they properly intersect each other.
Update μ and f .
5. **Ending test:** Repeat steps 2-4 until E is a cluster.

Concerning this new algorithm, we may notice that facts 33 and 34 imply that the clusters generated by the new merging step 2, form a 2-3 hierarchy. The integration of the refinement step inside the loop defined by steps 2-5, ensures that the clustering structure is weakly indexed by f , whereas it is clear that the deletion of some clusters having only one predecessor, does not change the property for the generated clusters to form a 2-3 hierarchy.

4 Complexity analysis and tests

4.1 Specifications

With the aim to specify and implement the new 2-3 AHC algorithm, we need to choose a link μ . In order to compare two non disjoint clusters, the definition of μ must extend the classical definitions of link used for disjoint clusters. Here we will use $\mu(X, Y) = \min\{\delta(x, y) : x \in X - Y, y \in Y - X\}$, together with an extension of the Lance and Williams formula.

In order to store and manage the matrix containing the link values between clusters, which is the most time expensive operation, we propose to use an *ordered tree structure* that puts in correspondence these values and the pairs of candidate clusters. The purpose is to search among all candidate cluster pairs for merging, the one that minimise a/several criteria/criterions.

We use three criterions in order to choose the merging pair: (1) *Minimal link*, since we search two closest clusters, (2) *Minimal cardinality*, meaning the number of elements of the clusters to be merged, when we have multiple pairs at a minimal link and (3) *Minimal lexicographical order* on the clusters

identifiers, when the two first criteria are satisfied by several pairs. Therefore, we have on the first level of the structure the ordered link values, on the second the ordered cardinalities of the pairs situated at the same link between clusters and on the third the lexicographically ordered identifiers.

4.2 Complexity analysis

The complexity of the **Initialization** (step 1) is larger than in Bertrand (2002a): $\mathcal{O}(n^2 \log n)$. The other steps are repeated n times and in the worst case the operations complexity will be reduced to $\mathcal{O}(n \log n)$ instead of $\mathcal{O}(n^2)$.

As follows we will analyze the complexity of the steps 2-4, which are repeated until the cluster E is created, that's at most $\lfloor \frac{3}{2}(n - 1) \rfloor$ times. In the **Merging** step (Step 2.a), we first retrieve the pair that minimise our criteria, in $\mathcal{O}(1)$, and we create the new cluster $X_i \cup Y_i$ also in $\mathcal{O}(1)$. If one of the merged clusters has another predecessor, we perform an **Intermediate merge** (Step 2.b) with the same complexity as the one before. Thus the whole complexity of the step 2 is $\mathcal{O}(n)$.

In the **Refinement** step (Step 3), we will eliminate from the structure the clusters found on the same level with their predecessors and we will update the *predecessor*, *successor* links between the remaining clusters, which is done in $\mathcal{O}(n)$, since a cluster can have at most $\lfloor \frac{3}{2}(n - 1) \rfloor$ successors.

In the **Update** step (Step 4) we first update \mathcal{M}_i in $\mathcal{O}(n)$ since adding the new formed cluster is constant and since a cluster can have at most n successors to eliminate from \mathcal{M}_i . In the μ update we eliminate from the structure the pairs containing at least a cluster to be eliminated. Since a pair is eliminated in $\mathcal{O}(\log n)$ and we have at most $\lfloor \frac{3}{2}(n - 1) \rfloor$ clusters, we have here an $\mathcal{O}(n \log n)$ complexity. Then, the links between the new formed cluster and the rest of the candidates are computed, each in $\mathcal{O}(n)$, and inserted into the matrix, in $\mathcal{O}(\log n)$ each. Therefore, the complexity of step 4 is $\mathcal{O}(n \log n)$.

Thus, the total worst case complexity is then reduced to $\mathcal{O}(n^2 \log n) + n \times \mathcal{O}(n \log n) = \mathcal{O}(n^2 \log n)$.

4.3 Implementation and tests

We designed an object-oriented model of the algorithm, which was implemented in Java, and integrated into the CBR*Tools framework (Jaczynski (1998)). We begun to test this algorithm as an indexing method in a CBR application for car insurance, based on a database³ usually used in CBR. Then we carried out a series of tests on random generated data. Figure 2 indicates the execution times of our algorithm compared to the AHC algorithm depending on the size n of the uniformly generated data set E . Figure 3 shows the convergence of the ratio $\frac{\text{Execution time}(n)}{\mathcal{O}(n^2 \log n)}$ for the AHC and our 2-3 AHC algorithm, which confirms the theoretical complexity analysis.

³ <ftp://ftp.ics.uci.edu/pub/ml-repos/machine-learning-databases/autos>

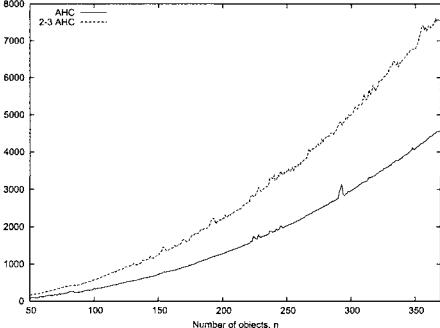


Fig. 2. Execution times

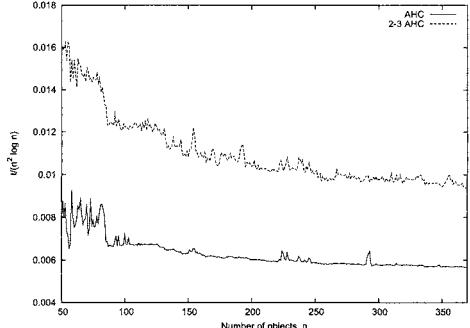


Fig. 3. Complexity validation

5 Conclusions and future work

The originality of this work is based on the four following points: (1) a new 2-3 AHC clustering algorithm, which simplifies the one proposed in 2002 (its principle is closer to the principle of the classical AHC), (2) a complexity reduction of the 2-3 AHC algorithm from $\mathcal{O}(n^3)$ to $\mathcal{O}(n^2 \log n)$, where n represents the number of objects to cluster, (3) a first object-oriented design and implementation of such an algorithm (in Java) and its integration in CBR*Tools, a Case-Based Reasoning framework and (4) an experimental validation of the algorithm complexity on simulated data.

Our current and future work concerns the following topics: (1) study of the quality of the 2-3 AHC compared with AHC and other classification methods and (2) study of the relevance of this new algorithm in the context of Web Usage Mining.

References

- BERTRAND, P. (2002a): *Set systems for which each set properly intersects at most one other set - Application to pyramidal clustering*. Cahier du Ceremadé numéro 0202, Ceremadé, Université Paris-9, France.
- BERTRAND, P. (2002b): *Les 2-3 hiérarchies : une structure de classification pyramidale parcimonieuse*. Actes du IX ème Congrès de la Société Francophone de Classification. 16-18 September, Toulouse, France.
- CHELCEA, S., BERTRAND, P., and TROUSSE, B. (2002): *Theoretical study of a new 2-3 hierarchical clustering algorithm*. Symbolic and Numeric Algorithms for Scientific Computing, 9-12 Octobre, Timisoara, Romania.
- GORDON, A.D. (1999): *Classification*. 2nd ed., Chapman and Hall, London.
- JACZYNSKI, M. (1998): *Scheme and Object-Oriented Framework for case Indexing By Behavioural Situations : Application in Assisted Web Browsing*. Doctorat Thesis of the University of Sophia-Antipolis (in french), December, France.

Symbolic Classifier with Convex Hull Based Dissimilarity Function

Francisco de A.T. de Carvalho and Simith T. D'Oliveira Júnior

Centro de Informática - UFPE,
Av. Prof. Luiz Freire, s/n - Cidade Universitária,
CEP - 50740-540 - Recife - PE - Brasil
email: {stdj,fatc}@cin.ufpe.br

Abstract. This work presents a new symbolic classifier based on a region oriented approach. At the end of the learning step, each class is described by a region (or a set of regions) in \mathbb{R}^p defined by the convex hull of the objects belonging to this class. In the allocation step, the assignment of a new object to a class is based on a dissimilarity matching function that compares the class description (a region or a set of regions) with a point in \mathbb{R}^p . This approach aims to reduce the over-generalization that is produced when each class is described by a region (or a set of regions) defined by the hyper-cube formed by the objects belonging to this class. It then seeks to improve the classifier performance. In order to show its usefulness, this approach was applied to a study of simulated SAR images.

1 Introduction

New approaches have been recently proposed to discover knowledge and summarize the information stored in large data sets. Symbolic Data Analysis (SDA) is a new domain related to multivariate analysis, pattern recognition, databases and artificial intelligence. It is concerned with the generalization of classical exploratory data analysis and statistical methods (visualization, factorial analysis, regression, clustering methods, classification, etc.) into symbolic data (Bock and Diday (2000)). Symbolic data are more complex than the standard data because they contain internal variations and are structured.

In Ichino et al. (1996), a symbolic classifier was introduced as a region-oriented approach. The learning step uses an approximation of the Mutual Neighborhood Graph (MNG) and a symbolic operator (join) to furnish the symbolic description of each class. In the classification step, the allocation of an individual to a class is based on a matching function that compares the description of the individual with the symbolic description of the class. In Souza et al. (1999) and De Carvalho et al. (2000), another MNG approximation was proposed to reduce the learning step complexity without losing the classifier performance in terms of prediction accuracy. In the allocation step, alternative similarity and dissimilarity functions have been used to assign an individual to a class.

This work presents a new symbolic classifier based on a region-oriented approach. At the end of the learning step, each class is described by a region

(or a set of regions) in \Re^p defined by the convex hull formed by the objects belonging to this class. This is obtained through a suitable approximation of a Mutual Neighborhood Graph (MNG). In the allocation step, the assignment of a new object to a class is based on a dissimilarity matching function that compares the class description (a region or a set of regions) with a point in \Re^p . This approach aims to reduce the over-generalization that is produced when each class is described by a region (or a set of regions) in \Re^p defined by the hyper-cube formed by the objects belonging to this class. It then seeks to improve the classifier performance. In order to show its usefulness, this approach was applied to a study of simulated SAR images.

2 Symbolic data

In this paper, we are concerned with symbolic data that are represented by quantitative feature vectors. More general symbolic data type can be found in Bock and Diday (2000). Let $\Omega = \{\omega_1, \dots, \omega_n\}$ be a set of n individuals described by p quantitative features $X_j (j = 1, \dots, p)$. Each individual $\omega_i (i = 1, \dots, n)$ is represented by a quantitative feature vector $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, where x_{ij} is a *quantitative feature value*. A quantitative feature value may be either a continuous value (e.g., $x_{ij} = 1.80$ meters in height) or an interval value (e.g., $x_{ij} = [0.2]$ hours, the duration of a student evaluation).

Example. A segment (set of pixels) described by the grey level average and standard deviation calculated from its set of pixels may be represented by the continuous feature vector $\mathbf{x} = (50, 7.5)$. The description of a group of segments may be represented by the interval feature vector $\mathbf{y} = ([120.68, 190.53], [0.36, 0.65])$, where the grey level average and standard deviation calculated from the set of pixels of each segment takes values in the interval $[120.68, 190.53]$ and in the interval $[0.36, 0.65]$, respectively.

2.1 Regions

Let $C_k = \{\omega_{k1}, \dots, \omega_{kN_k}\}$, $k = 1, \dots, m$, be a class of individuals with $C_k \cap C_{k'} = \emptyset$ if $k \neq k'$ and $\cup_{k=1}^m C_k = \Omega$. The individual ω_{kl} , $l = 1, \dots, N_k$, is represented by the continuous feature vector $\mathbf{x}_{kl} = (x_{kl1}, \dots, x_{klp})$.

A symbolic description of the class C_k can be obtained by using the join operator (Ichino et al. (1996)).

Definition 1. The join between the continuous feature vectors \mathbf{x}_{kl} ($l = 1, \dots, N_k$) is an interval feature vector defined as $\mathbf{y}_k = \mathbf{x}_{k1} \oplus \dots \oplus \mathbf{x}_{kN_k} = (x_{k11} \oplus \dots \oplus x_{kN_k1}, \dots, x_{k1j} \oplus \dots \oplus x_{kN_kj}, \dots, x_{k1p} \oplus \dots \oplus x_{kN_kp})$, where $x_{k1j} \oplus \dots \oplus x_{kN_kj} = [\min\{x_{k1j}, \dots, x_{kN_kj}\}, \max\{x_{k1j}, \dots, x_{kN_kj}\}]$.

We can associate two regions in \Re^p to each class C_k : one spanned by the join of its elements and another spanned by the convex hull of its elements.

Definition 2. The *J-region* associated to class C_k is a region in \Re^p that is spanned by the join of the objects belonging to class C_k . It is defined as

$R_J(C_k) = \{\mathbf{x} \in \Re^p : \min\{x_{k1j}, \dots, x_{kN_k j}\} \leq x_j \leq \max\{x_{k1j}, \dots, x_{kN_k j}\}, j = 1, \dots, p\}$. The volume associated to the hyper-cube defined by $R_J(C_k)$ is $\pi(R_J(C_k))$.

Definition 3. The *H-region* associated to class C_k is a region in \Re^p that is spanned by the convex hull formed by the objects belonging to class C_k . It is defined as $R_H(C_k) = \{\mathbf{x} = (x_1, \dots, x_j, \dots, x_p) \in \Re^p : \mathbf{x}$ is inside the envelop of the convex hull defined by the continuous feature vectors $\mathbf{x}_{kl} = (x_{kl1}, \dots, x_{klp}), l = 1, \dots, N_k\}$. The volume associated to the internal points within the convex hull envelop defined by $R_H(C_k)$ is $\pi(R_H(C_k))$.

2.2 Graph concepts

The *mutual neighborhood graph (MNG)* (Ichino et al. (1996)) yields information on interclass structure.

Definition 4. The objects belonging to class C_k are each *mutual neighbors* (Ichino et al. (1996)) if $\forall \omega_{k'l} \in C_{k'} (k' \in \{1, \dots, m\}, k' \neq k), \mathbf{x}_{k'l} \notin R_J(C_k) (l = 1, \dots, N_{k'})$. In such a case, the MNG of C_k against $\overline{C}_k = \bigcup_{\substack{k'=1 \\ k' \neq k}}^m C_{k'}$, which is constructed by joining all pairs of objects that are mutual neighbors, is a complete graph. If the objects belonging to class C_k are not each mutual neighbors, we look for all the subsets of C_k where the elements are each mutual neighbors and which are a *maximal clique* in the MNG. In such a case, the MNG is not a complete graph. We can associate a *J-region* to each of these subsets of C_k and calculate the volume of the corresponding hyper-cube it defines.

In this paper we introduce an additional definition to the MNG.

Definition 5. The objects belonging to class C_k are each *mutual neighbors* if $\forall \omega_{k'l} \in C_{k'}, k' \in \{1, \dots, m\}, k' \neq k, \mathbf{x}_{k'l} \notin R_H(C_k) (l = 1, \dots, N_{k'})$. The MNG of C_k against $\overline{C}_k = \bigcup_{\substack{k'=1 \\ k' \neq k}}^m C_{k'}$ defined in this way is also a complete graph. If the objects belonging to class C_k are not each mutual neighbors, again we look for all the subsets of C_k where the elements are each mutual neighbors and which are a maximal clique in the MNG. We can then associate *H-region* to each of these subsets of C_k and calculate the volume of the corresponding convex-hull it defines.

3 Symbolic classifier

This section introduces the learning and allocation steps of the symbolic classifier presented in this paper.

3.1 Learning step

The idea of this step is to learn the regions associated to each class so as to allow the classification of a new individual into a class through the comparison of the class description (regions) with a point in \Re_p according to a dissimilarity matching function.

We have two basic remarks concerning this step. The first is that a difficulty arises when the objects belonging to a class C_k are not each mutual neighbors. In such a case, we look for all the subsets of C_k where its elements are each mutual neighbors and which are a *maximal clique* in the MNG (which is not a complete graph in such a case). However, it is well known that the computational complexity in time to find all cliques on a graph is exponential. It is then necessary to construct an *approximation* of the MNG.

The second remark concerns what kind of region (*J-region* or *H-region*) is suitable for describing a class C_k . Figure 1 illustrates the description of a class by a *J-region* and by a *H-region*. It is clear that the representation based on a *J-region* (see Ichino et al. (1996), Souza et al. (1999), De Carvalho et al. (2000)) over-generalizes the class description given by a *H-region*. For this reason, the latter option will be used in this paper.

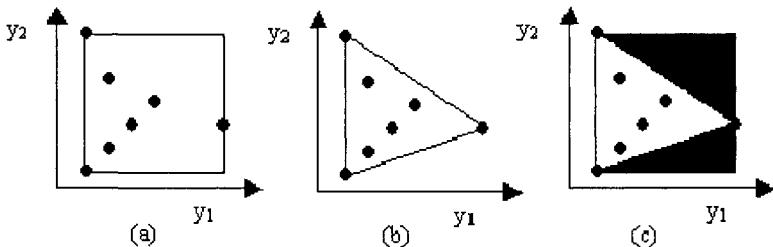


Fig. 1. (a) *J-region*, (b) *H-region*, (c) Over-generalization

The construction of the MNG for the classes C_k ($k = 1, \dots, m$) and the representation of each class by a *H-region* (or by a set of *H-regions*) is accomplished in the following way:

For $k = 1, \dots, m$ do

- 1 Find the the region $R_H(C_k)$ (according to *definition 3*) associated to class C_k and verify if the objects belonging to this class are each mutual neighbors according to *definition 5*
- 2 If so, construct the MNG (which is a complete graph) and stop.
- 3 If this is not the case, (MNG approximation) do the following:
 - 3.1 choose an object of C_k as a seed according to the lexicographic order of these objects in C_k ; do $t = 1$ and put the seed in C_k^t ; remove the seed from C_k
 - 3.2 add the next object of C_k (according to the lexicographic order) to C_k^t if all the objects belonging now to C_k^t each remain mutual neighbors according to *definition 5*; if this is true, remove this object from C_k
 - 3.3 repeat step 2) for all remaining objects in C_k

3.4 Find the region $R_H(C_k^t)$ (according to *definition 3*) associated to C_k^t

3.5 if $C_k \neq \emptyset$, do $t = t + 1$ and repeat steps 3.1 to 3.4) until $C_k = \emptyset$

4 construct the MNG (which is now not a complete graph) and stop.

At the end of this algorithm the subsets $C_k^1, \dots, C_k^{n_k}$ of class C_k are computed and the description of this class is obtained by the *H-regions* $R_H(C_k^1), \dots, R_H(C_k^{n_k})$.

As an example in the case of two classes, Figure 3 shows a) the complete Mutual Neighborhood Graph and the class descriptions based on *J-regions* (Ichino et al. (1996)), b) The MNG approximation and the class descriptions based on *J-regions* (Souza et al. (1999), De Carvalho et al. (2000)) and c) the MNG approximation and the class descriptions based on *H-regions* (the approach presented in this paper).

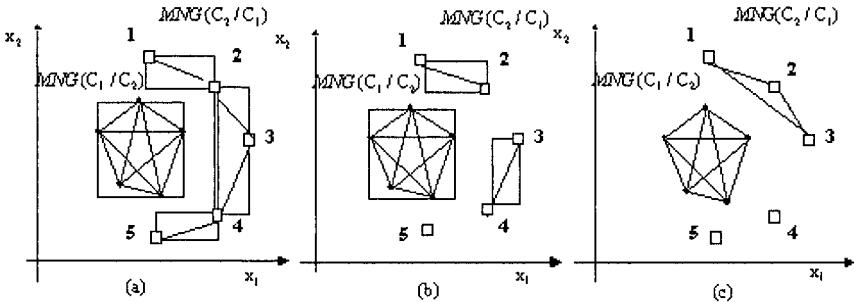


Fig. 2. (a) Complete MNG and *J-regions*, (b) MNG approximation and *J-regions*, (c) MNG approximation and *H-regions*

3.2 Allocation step

In the allocation step, a new object ω is compared with each class C_k and a dissimilarity score is computed according to a suitable matching function. Then, the minimal dissimilarity score is sought out and we assign the object ω to the class that corresponds to this minimal score.

Let ω be a new object to be assigned to a class C_k that is described by a continuous feature vector $\mathbf{x} = (x_1, \dots, x_p)$. Remember that the subsets $C_k^1, \dots, C_k^{n_k}$ of C_k are computed from the learning step.

The *classification rule* is defined as following: ω is affected to the class C_k if

$$\delta(\omega, C_k) \leq \delta(\omega, C_h), \forall h \in \{1, \dots, m\} \quad (1)$$

where $\delta(\omega, C_h) = \min\{\delta(\omega, C_h^1), \dots, \delta(\omega, C_h^{n_h})\}$.

In this paper, the dissimilarity matching function δ is defined as

$$\delta(\omega, C_h^s) = \frac{\pi(R_H(C_h^s \cup \{\omega\})) - \pi(R_H(C_h^s))}{\pi(R_H(C_h^s \cup \{\omega\}))}, \quad s = 1, \dots, n_h \quad (2)$$

4 Monte Carlo experience

In order to show the usefulness of the method proposed in this paper, a special kind of SAR simulated image is classified in this section.

4.1 SAR simulated images

Synthetic Aperture Radar (SAR) is a system that possesses its own illumination and produces images with a high capacity for discriminating objects. It uses coherent radiation, generating images with speckle noise. SAR data display random behaviour that is usually explained by a multiplicative model (Frery et al. (1997)). This model considers that the observed return signal Z is a random variable defined as the product of two other random variables: X (the terrain backscatter) and Y (the speckle noise).

The process for obtaining simulated images consists in creating classes of idealized images (a phantom), and then associating a particular distribution to each class.

Different kinds of detection (intensity or amplitude format) and types of regions can be modelled by different distributions associated to the return signal. The homogeneous (e.g. agricultural fields), heterogeneous (e.g. primary forest) and extremely heterogeneous (e.g. urban areas) region types are considered in this work. According to Frery et al. (1997), we assume that the return signal in the amplitude case has the square root of a Gamma distribution, the K-Amplitude distribution and the G0-Amplitude distribution in homogeneous, heterogeneous and extremely heterogeneous areas, respectively.

Two situations of images are considered ranging in classification from moderate to greatly difficult. We generate the distribution associated to each class in each situation by using an algorithm for generating gamma variables.

The *Lee filter* (Lee (1981)) was applied to the data before segmentation in order to decrease the speckle noise effect. The segmentation was obtained using the region growing technique (Jain (1988)), based on the t-student test (at the 5% significance level) for the merging of regions.

Each segment (set of pixels) is described by two features (gray level average and standard deviation calculated from the segment set of pixels). The convex hull of a set of points (segments) in \mathbb{R}^2 is defined as the minimal convex polygon encompassing these points. A number of algorithms have been developed to construct a convex hull from a given set of points. We have chosen the Graham scan algorithm (O'Rourke (1998)) because, it has the minimal time complexity ($O(n \log n)$, n being the cardinality of the set) among the thus far algorithms when applied to points in \mathbb{R}^2 .

4.2 Experimental evaluation

The evaluation of the approach presented in this paper (named here *H-region approach*, where class representation, MNG approximation and dissimilarity matching function are based on *H-regions*, is performed based on prediction accuracy, in comparison with the approach where class representation, MNG approximation and dissimilarity matching function are based on the *J-regions* (named here *J-region approach*).

The Monte Carlo experience was performed for images of sizes 64×64 , 128×128 and 256×256 , taking into consideration situations 1 and 2. 100 replications were obtained with identical statistical properties and the prediction accuracy, speed and storage were calculated.

The prediction accuracy of the classifier was measured through the error rate of classification obtained from the test set. The estimated error rate of classification corresponds to the average of the error rates found for these replications.

The comparison according to the average of the error rate was achieved by a paired Student's t-test at the significance level of 5%. Table 3 shows the average error rate, suitable (null and alternative) hypothesis and the observed values of the test statistics for various sizes and the two image situations. In this table, the test statistics follow a Student's t distribution with 99 degrees of freedom, and μ_1 and μ_2 are, respectively, the average error rate for the *H-region approach* and the *J-region approach*.

From Table 3, we can conclude that in all cases (size and image situation) the average error rate for the *H-region approach* is lower than that for the the *J-region approach*. Also, the test statistics shows that the *H-region approach* outperforms the *J-region approach*.

SAR images	<i>H-region</i> Approach	<i>J-region</i> Approach	$H_0 : \mu_2 \geq \mu_1$
			$H_1 : \mu_2 < \mu_1$
64×64 situation 1	5.78	8.29	-5.19
64×64 situation 2	24.83	24.92	-0.15
128×128 situation 1	2.68	3.42	-5.03
128×128 situation 2	16.52	16.89	-1.45
256×256 situation 1	1.39	1.87	-8.38
256×256 situation 2	13.67	14.34	-4.57

Table 1. Comparison between the classifiers according to the average error rate.

5 Conclusion

A new symbolic classifier based on a region-oriented approach is presented in this paper. At the end of the learning step, each class is described by a region

(or a set of regions) in \Re^p defined by the convex hull formed by the objects belonging to this class, which is obtained through a suitable approximation of a Mutual Neighborhood Graph (MNG). This approach aims to reduce the over-generalization that is produced when each class is described by a region (or a set of regions) in \Re^p defined by the hyper-cube formed by the objects belonging to this class. It then seeks to improve the classifier performance.

In order to show its usefulness, this approach was applied in the study of simulated SAR images presenting situations ranging in classification from "moderately easy" to "greatly difficult". The input (segments of images) is a set of continuous feature vectors. To assign a segment to a region, a dissimilarity matching function, comparing the class description (a region or a set of regions) with a point in \Re^p , was introduced.

The evaluation of the approach presented in this paper (called the *H-region approach*) was based on prediction accuracy as measured through the error rate of classification obtained from the test set in comparison with the *J-region approach*. This measurement was accomplished in the framework of a Monte Carlo experience. The results showed that, concerning the prediction accuracy, the *H-region approach* outperforms the *J-region approach*. Future work must also consider the speed and storage performance of the *H-region approach* in comparison with the *J-region approach*.

Acknowledgments: The authors would like to thank CNPq (Brazilian Agency) for its financial support.

References

- BOCK, H.-H. and DIDAY, E. (2000): *Analysis of Symbolic Data*. Springer, Heidelberg.
- DE CARVALHO, F.A.T., ANSELMO, C.A.F., and SOUZA, R.M.C.R. (2000): Symbolic approach to classify large data sets, In: H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, and M. Schader (Eds.): *Data Analysis, Classification, and Related Methods*. Springer, Berlin, 375–380.
- FRERY, A.C., MUELER, H.J., YANASSE, C.C.F., and SANT'ANA, S.J.S. (1997): A model for extremely heterogeneous clutter. *IEEE Transactions on Geoscience and Remote Sensing*, 1, 648–659.
- ICHINO, M., YAGUCHI, H., and DIDAY, E. (1996): A fuzzy symbolic pattern classifier In: E. Diday, Y. Lechevallier, and O. Opitz (Eds.): *Ordinal and Symbolic Data Analysis*. Springer, Berlin, 92–102.
- JAIN, A.K. (1988): *Fundamentals of Digital Image Processing*. Prentice Hall International Editions, Englewood Cliffs.
- LEE, J.S. (1981): Speckle analysis and smoothing of synthetic aperture radar images. *Computer Graphics and Image Processing*, 17, 24–32.
- O'Rourke, J. (1998): Computational Geometry in C (Second Edition), Cambridge University Press, New York.
- SOUZA, R.M.C.R., DE CARVALHO, F.A.T., and FRERY, A.C. (1999): Symbolic approach to SAR image classification. *IEEE 1999 International Geoscience and Remote Sensing Symposium*, Hamburg, 1318–1320.

Two-Mode Cluster Analysis via Hierarchical Bayes

Wayne S. DeSarbo, Duncan K. H. Fong, and John Liechty

Marketing Dept., Smeal College of Business, Pennsylvania State University,
University Park, PA, USA 16802

Abstract. This manuscript introduces a new Bayesian finite mixture methodology for the joint clustering of row and column stimuli/objects associated with two-mode asymmetric proximity, dominance, or profile data. That is, common clusters are derived which partition both the row and column stimuli/objects simultaneously into the same derived set of clusters. In this manner, interrelationships between both sets of entities (rows and columns) are easily ascertained. We describe the technical details of the proposed two-mode clustering methodology including its Bayesian mixture formulation and a Bayes factor heuristic for model selection. Lastly, a marketing application is provided examining consumer preferences for various brands of luxury automobiles.

1 Introduction

Two-mode cluster analysis involves the simultaneous and joint amalgamation of both the row and column objects contained in a two-mode data matrix. Examples of such two-mode data include: asymmetric two-mode proximity data (e.g., confusions data), two-way dominance data (e.g., subjects eliciting preferences or choices with respect to different column objects), two-way profile data (e.g., objective quantitative features or attributes for a set of designated objects), etc. A number of psychometric and classification related procedures for the clustering of such two-mode data have been published over the past few decades (see DeSarbo, Fong, Liechty, and Saxton, 2003 for an excellent literature review on two-mode clustering).

Bayesian approaches to traditional *one-mode* cluster analysis began with the seminal work of Binder (1978) who described a general class of normal mixture models and introduced various ingredients of Bayesian approaches to classification, clustering, and discrimination into this finite mixture framework. Later, work on Bayesian estimation of finite mixture models for classification via posterior simulation followed by Gilks, Oldfield, and Rutherford (1989), Diebolt and Robert (1994), Gelman and King (1990), Verdinelli and Wasserman (1991), Evans, Guttman and Olkin (1992). Lavine and West (1992) extended Binder's (1978) work by applying an iterative resampling approach to Monte Carlo inference, Gibbs sampling, to this same mixture framework, stressing the ease with which such analyses may be performed in more general settings. Their Bayesian framework allowed for the generalization to several normal mixture components having different covariance

matrices, the computation of exact posterior classification probabilities for observed data as well as for future cases to be classified, and the extraction of posterior distributions for these probabilities that allow for the assessment of uncertainties in classification. Again, with respect to clustering, all of this Bayesian research dealt with traditional, one-mode clustering via the use of model based clustering with finite mixtures.

This manuscript proposes a new Bayesian methodology for *two-mode* clustering. Given empirical two-mode data as input (e.g., asymmetric proximities, preferences, choices, attributes, etc.), the proposed methodology derives discrete groupings of row and column objects jointly in the same derived clusters. We present a stochastic mixture model in a Bayesian context, and use Bayes factors for determining the appropriate number of clusters. The next section of the paper presents the technical aspects of the proposed methodology. Then, a marketing application involving consumer preferences is provided which illustrates the proposed Bayesian procedure.

2 The Bayesian two-mode clustering methodology

Here we propose a new Bayesian two-mode clustering methodology that simultaneously allocates the row and column objects of two-mode data into joint clusters. We estimate these different clustering solutions using a hierarchical Bayes mixture model where Beta densities are used to illustrate the methodology. We begin the specification of the *proposed* model by defining the appropriate membership random variables and formulating the likelihood function. Next, we specify priors and derive the resulting full conditional distributions that are necessary for the implementation of a Monte Carlo Markov chain algorithm that we devise for estimation.

In our Bayesian mixture model framework, we define two sets of membership random variables to identify cluster membership for each observed data point. To help understand how clusters are defined, consider two-way preference/dominance data where each row represents a consumer and each column represents a brand/product. For a group/cluster of consumers, we assume that there are a collection of brands that are preferred by these consumers and another collection of brands that are not preferred. All of the brands assigned to a cluster are preferred by the consumers in that cluster, and the remaining brands are not preferred. Stated in terms of rows and columns, observed values in the cells from a row will tend to be larger for columns that are assigned to the same cluster as the row and smaller for columns not in the same cluster. We assume that observations from the preferred brands follow one probability distribution and observations from the non-preferred brands follow another probability distribution. DeSarbo, Fong, Liechty, and Saxton (2003) describe a number of alternative two-mode clustering specifications.

2.1 The likelihood function

Since our application involves the analysis of two-way, two-mode preference/dominance data in marketing where the preferences of consumers concerning different brands of luxury automobiles is examined, we will hereby assume that the row objects are consumers, and the column objects are brands of luxury automobiles in this product class. Let:

- $i = 1, \dots, N$ consumers;
- $j = 1, \dots, J$ brands of luxury automobiles;
- y_{ij} = the preference or usage of brand j by consumer i ;
- $\mathbf{y} = [y_{ij}]$ be the data matrix;
- $k = 1, \dots, K$ clusters or market segments;
- sc_i be a membership variable identifying the cluster for consumer i ;
- sp_j be a membership variable identifying the cluster for brand j ;
- $\mathbf{sc} = (sc_1, \dots, sc_N)^T$ and $\mathbf{sp} = (sp_1, \dots, sp_J)^T$.

Conditional on \mathbf{sc} and \mathbf{sp} , the likelihood function is given by:

$$\begin{aligned} L(y|\Theta_{1k}, \Theta_{2k}, k = 1, \dots, K, \mathbf{sc}, \mathbf{sp}) = \\ I\{\mathbf{sp} \in S_p^*, \mathbf{sc} \in S_c^*\} \prod_{i,j} \left[\sum_{k=1}^K I\{sc_i = k\} I\{sp_j = k\} f(y_{ij}|\Theta_{2k}) \right] (1) \\ + I\{sp_j \neq k\} f(y_{ij}|\Theta_{1k})], \end{aligned}$$

where $I\{\cdot\}$ is the indicator function, $f(\cdot|\Theta_h)$ is an appropriate probability density with parameter Θ_h , and $S_c^*(S_p^*)$ is the set of all possible groupings of consumers (brands) that are allowed. Since $0 \leq y_{ij} \leq 1$ in our marketing example, we assume a Beta density:

$$f(y_{ij}|a_h, b_h) = \frac{\Gamma(a_h + b_h)}{\Gamma(a_h)\Gamma(b_h)} (y_{ij})^{a_h-1} (1-y_{ij})^{b_h-1}, a_h > 0, b_h > 0. \quad (2)$$

Note that, for this specification, $\Theta_{1k} = (a_{1k}, b_{1k})$ and $\Theta_{2k} = (a_{2k}, b_{2k})$ in (1). The sets S_c^* and S_p^* are used to reflect two basic types of restrictions. The first type of restriction originates from the actual model specification. Here, each brand must be assigned to one cluster, and each cluster must have at least one brand assigned to that cluster. The second type of restriction is imposed directly by the researcher, either to incorporate prior beliefs about the particular application at hand or to facilitate the implementation of the method in practice. For example, because of cost/revenue implications, the researcher may require that a certain minimum percentage of the consumers in the sample be assigned to each cluster (we use a 10% restriction in the application to follow). Finally, for identification purposes and in order to ensure that the brands assigned to cluster k are preferred over the brands not assigned to cluster k , we require:

$$\frac{a_{2k}}{(a_{2k} + b_{2k})} > \frac{a_{1k}}{(a_{1k} + b_{1k})}. \quad (3)$$

2.2 Prior distributions

Here we specify prior distributions for the unknown Beta parameters and the membership variables. For all i and j , let:

$$\phi_{ck} = Pr(sc_i = k) \quad \text{and} \quad \phi_{pk} = Pr(sp_j = k), k = 1, \dots, K. \quad (4a)$$

We assume *a priori* that:

$$Pr(\mathbf{sc} = (k_1, \dots, k_N)^T) = \prod_{i=1}^N \phi_{ck_i} \quad \text{and} \quad (4b)$$

$$Pr(\mathbf{sp} = (k_1', \dots, k_J')^T) = \prod_{j=1}^J \phi_{pk'_j}, 1 \leq k_i, k_j' \leq K.$$

Thus, the prior distributions of \mathbf{sc} and \mathbf{sp} are specified when values of ϕ_{ck} and ϕ_{pk} are given. If we denote $w_{ik} = I\{sc_i = k\}, i = 1, \dots, N, k = 1, \dots, K$, then (w_{i1}, \dots, w_{iK}) follows a multinomial distribution $Mu(\phi_{c1}, \dots, \phi_{cK}; 1)$. We assume, *a priori*, that each individual and product has equal chance of being in each state; i.e., we let $\phi_{ck} = \frac{1}{K}$ and $\phi_{pk} = \frac{1}{K}$, for $k = 1, \dots, K$, to represent vague prior information. Finally, we assume that the parameters of the Beta densities are independent and that they follow exponential prior densities with means reflecting the identification restrictions in (3):

$$f(a_{rk}) = \lambda_{rk} \exp(-\lambda_{rk} a_{rk}) \quad \text{and} \quad f(b_{rk}) = \gamma_{rk} \exp(-\gamma_{rk} b_{rk}) \quad (5)$$

for $r = 1, 2$, where:

$$\frac{E[a_{2k}]}{E[a_{2k}] + E[b_{2k}]} = \frac{1/\lambda_{2k}}{1/\lambda_{2k} + 1/\gamma_{2k}} > \frac{1/\lambda_{1k}}{1/\lambda_{1k} + 1/\gamma_{1k}} = \frac{E[a_{1k}]}{E[a_{1k}] + E[b_{1k}]} \quad (6)$$

This particular specification is selected in order to accommodate the positivity of a_{rk} and b_{rk} , as well as to conveniently handle the necessary identification restrictions discussed earlier. In order to set the parameters for the exponential distributions of a_{rk} and b_{rk} , we consider the first order, *a priori* approximations to the mean and variance of the accompanying Beta density in (2):

$$mean_{rk} = \frac{E[a_{rk}]}{E[a_{rk}] + E[b_{rk}]} \quad \text{and} \quad (7)$$

$$variance_{rk} = \frac{E[a_{rk}]E[b_{rk}]}{(E[a_{rk}] + E[b_{rk}])^2(E[a_{rk}] + E[b_{rk}] + 1)}.$$

Since

$$\text{mean}_{rk} = (1 - \text{mean}_{rk}) > \text{variance}_{rk} > 0, \quad (8)$$

we let the variance be a percentage of the maximum variance as given in (8), or $\text{variance}_{rk} = \text{scale} \times \text{mean}_{rk}(1 - \text{mean}_{rk})$, where $1 > \text{scale} > 0$. Based on specified mean and scale values, we solve (7) to find $E[a_{rk}]$ and $E[b_{rk}]$. Since $E[a_{rk}] = 1/\lambda_{rk}$ and $E[b_{rk}] = 1/\gamma_{rk}$, λ_{rk} and γ_{rk} are determined accordingly. Because $B(1, 1)$ is the Uniform density with mean 0.5, we assume that $\text{mean}_{2k} = 0.51$ and $\text{mean}_{1k} = 0.49$, for all k , and employ appropriate scale values to represent vague prior information. Note that the scale value should not be set too close to 1 or the prior distributions for a_{rk} and b_{rk} will concentrate around zero which lead to subjective Beta priors favoring small (close to zero) and large (close to 1) values. Observing that the variance of the Uniform distribution is one third of the product of its mean and one minus mean, we use scale values of 0.1 and 0.5 in the prior specification. Two different scale values are used to allow us to check prior sensitivity of our results.

2.3 Full conditional distributions

Let $a = (a_{11}, \dots, a_{2K})^T$ and $b = (b_{11}, \dots, b_{2K})^T$. Conditional on the data \mathbf{y} , the full conditional distributions come from the joint posterior distribution of $(\mathbf{a}, \mathbf{b}, \mathbf{sc}, \mathbf{sp})$, which is proportional to the product of Likelihood (1) and the prior distributions (4a), (4b) and (5), i.e.,

$$L(\mathbf{y}|\mathbf{a}, \mathbf{b}, \mathbf{sc}, \mathbf{sp}) \cdot \prod_{i=1}^N [\sum_{k=1}^K \phi_{ck}^{I\{sc_i=k\}}] \prod_{j=1}^J [\sum_{k=1}^K \phi_{pk}^{I\{sp_j=k\}}] \prod_{k=1}^K \prod_{r=1}^2 f(a_{rk})f(b_{rk}). \quad (9)$$

Note that the normalizing constant of the posterior distribution is needed in the computation of Bayes factor for model comparison, but it is not required in the Markov chain Monte Carlo (MCMC) algorithm to obtain posterior estimates of the quantities of interest. Let $\mathbf{a}_{(rk)}$ be the vector of \mathbf{a} without the component a_{rk} , $\mathbf{b}_{(rk)}$ be the vector of \mathbf{b} without the component b_{rk} , $r = 1, 2$, $\mathbf{sc}_{(i)}$ be the vector of \mathbf{sc} without the component sc_i , and $\mathbf{sp}_{(j)}$ be the vector of \mathbf{sp} without the component sp_j . The full conditional distributions are as follows:

$$Pr(sp_j = k|\mathbf{y}, \mathbf{a}, \mathbf{b}, \mathbf{sc}, \mathbf{sp}_{(j)}) = \frac{L(\mathbf{y}|\mathbf{a}, \mathbf{b}, \mathbf{sc}, \mathbf{sp}_{(j)}, sp_j = k)\phi_{pk}}{\sum_l L(\mathbf{y}|\mathbf{a}, \mathbf{b}, \mathbf{sc}, \mathbf{sp}_{(j)}, sp_j = l)\phi_{pl}}, \quad (10)$$

$$k = 1, \dots, K; j = 1, \dots, J$$

$$Pr(sc_i = k | \mathbf{y}, \mathbf{a}, \mathbf{b}, \mathbf{sc}_{(i)}, \mathbf{sp}) = \frac{L(\mathbf{y} | \mathbf{a}, \mathbf{b}, \mathbf{sc}_{(i)}, \mathbf{sp}, sc_i = k) \phi_{ck}}{\sum_l (L(\mathbf{y} | \mathbf{a}, \mathbf{b}, \mathbf{sc}_{(i)}, \mathbf{sp}, sc_i = l) \phi_{cl})}, \quad (11)$$

$$k = 1, \dots, K; i = 1, \dots, N$$

$$f(a_{1k} | \mathbf{y}, \mathbf{a}_{(1k)}, \mathbf{b}, \mathbf{sc}, \mathbf{sp}) \quad (12)$$

$$\propto e^{-\lambda_{1k} a_{1k}} \prod_{ij} \left(\frac{\Gamma(a_{1k} + b_{1k})}{\Gamma(a_{1k}) \Gamma(b_{1k})} (y_{ij})^{a_{1k}-1} \right)^{I\{sc_i=k\} I\{sp_j \neq k\}}$$

$$f(a_{2k} | \mathbf{y}, \mathbf{a}_{(2k)}, \mathbf{b}, \mathbf{sc}, \mathbf{sp}) \quad (13)$$

$$\propto e^{-\lambda_{2k} a_{2k}} \prod_{ij} \left(\frac{\Gamma(a_{2k} + b_{2k})}{\Gamma(a_{2k}) \Gamma(b_{2k})} (y_{ij})^{a_{2k}-1} \right)^{I\{sc_i=k\} I\{sp_j = k\}}$$

$$f(b_{1k} | \mathbf{y}, \mathbf{a}, \mathbf{b}_{(1k)}, \mathbf{sc}, \mathbf{sp}) \quad (14)$$

$$\propto e^{-\gamma_{1k} b_{1k}} \prod_{ij} \left(\frac{\Gamma(a_{1k} + b_{1k})}{\Gamma(a_{1k}) \Gamma(b_{1k})} (1 - y_{ij})^{b_{1k}-1} \right)^{I\{sc_i=k\} I\{sp_j \neq k\}}$$

$$f(b_{2k} | \mathbf{y}, \mathbf{a}, \mathbf{b}_{(2k)}, \mathbf{sc}, \mathbf{sp}) \quad (15)$$

$$\propto e^{-\gamma_{2k} b_{2k}} \prod_{ij} \left(\frac{\Gamma(a_{2k} + b_{2k})}{\Gamma(a_{2k}) \Gamma(b_{2k})} (1 - y_{ij})^{b_{2k}-1} \right)^{I\{sc_i=k\} I\{sp_j = k\}}$$

2.4 The Markov chain Monte Carlo algorithm

Given these derived full conditional distributions, we employ the Markov chain Monte Carlo (MCMC) algorithm by drawing random deviates iteratively and recursively from the distributions to obtain an approximate random sample from the joint posterior distribution. Based on these draws, one may use the sample mean and sample variance to estimate the corresponding posterior mean and variance of various quantities of interest (see Gilks, Richardson, and Spiegelhalter (1996) for a more elaborate discussion of MCMC methods). Because the full conditional densities for the parameters of the Beta density are non-standard, we use a version of the Metropolis Hastings (MH) algorithm to generate random deviates from expressions (12) to (15). This requires one to sample from an approximating standard probability distribution (proposal distribution) and to use a criterion to decide whether a sampled value should be accepted or rejected. If the draw from the proposal distribution is not accepted, then the current value will remain unchanged. We first describe the general version of our MH algorithm, and then provide the proposal distributions used to generate random deviates of a_{1k} , a_{2k} , b_{1k} and b_{2k} :

1. Given the current parameter value x^s at the s-th iteration, draw a random deviate x from a proposal distribution,

$$x \sim Q(x|x^s), \quad (16)$$

where \sim means "drawn from" and $Q(x|x^s)$ is the conditional proposal density.

2. Calculate the following acceptance probability:

$$a(x^s, x) = \min \left\{ \frac{Q(x^s|x)f(x)}{Q(x|x^s)f(x^s)}, 1 \right\}, \quad (17)$$

where $f(\cdot)$ is the density to be sampled from.

3. With probability $a(x^s, x)$, let $x^{s+1} = x$. Otherwise, set $x^{s+1} = x^s$.

For the parameters of the Beta densities, a_{1k} , a_{2k} , b_{1k} and b_{2k} , we used the MH algorithm with a truncated Gamma as the proposal distribution, where the non-truncated version of the Gamma density has a mean equal to the current parameter value and a variance that can be used as a tuning parameter. After generating a_{1k} , a_{2k} , b_{1k} and b_{2k} , we check whether the restrictions in (3) are satisfied. If the restrictions are not satisfied, a new set of random deviates are then generated.

2.5 Model selection: The Bayes factor

The Bayes factor (BF) is a well developed and frequently used tool for model selection in most Bayesian formulations, which naturally accounts for both the explanatory power and the complexity of competing models. Several authors have discussed the advantages and appropriate uses of the Bayes factor (cf. Berger 1985 and O'Hagan 1994). In addition to calculating the BF, we also calculate the squared residual between the observed data and the values predicted by the model being estimated.

For two competing models (M_1 and M_2), the Bayes factor:

$$B_{21} = \frac{\text{Posterior odds}_{21}}{\text{Prior odds}_{21}} = \frac{pr(y|M_2)}{pr(y|M_1)} \quad (18)$$

is defined as the ratio of the posterior odds of the two models over the prior odds of the two models. Using Bayes theorem, this ratio is equal to the ratio of the probability of the data conditioned on model 2 over the probability of the data conditioned on model 1. To calculate a BF, we need to calculate the marginal probability of the data conditioned on each model (cf. Equation 18). In general, this can be a rather challenging problem. Kass and Raftery (1995) offer a survey of methods that can be used to calculate a BF. In this paper, we modify the fourth and final sampling based estimator proposed by Newton and Raftery (1994) to calculate the BF. The Bayes factor gives clear evidence for choosing one model over another. Under unity prior odds, if $B_{21} > 1$, then the posterior odds favor model 2; when $B_{21} < 1$, the posterior odds favor model 1. Obviously, the strength of support for one model versus another depends on the size of the BF. Jeffreys (1961) suggests that a Bayes factor greater than 10 gives support for choosing model 2 over

model 1, and that a Bayes factor of greater than 100 gives strong support for choosing model 2 over model 1. Kass and Raftery (1995) offer an extended guide, based on Jeffreys suggestions, for interpreting a BF:

BF	Evidence for Model 2
$1 \leq BF < 3$	Weak
$3 \leq BF < 20$	Positive
$20 \leq BF < 150$	Strong
$150 \leq BF$	Overwhelming

The naïve or base line model utilized in our BF calculation in this paper assumes that every cell has the same estimate and so the likelihood function is given by:

$$L(\mathbf{y}|\bar{a}, \bar{b}) = \prod_{i,j} f(y_{ij}|\bar{a}, \bar{b}). \quad (19)$$

where $f(\cdot|\bar{a}, \bar{b})$ is the Beta density in (2).

3 Application: Luxury automobile preferences

3.1 The Study

DeSarbo and Jedidi (1995) report of a study sponsored by a major U.S. automobile manufacturer where personal interviews were conducted with some N=240 consumers in various geographical areas. These consumers were initially screened in automobile clinics as to stating that they were intending to purchase a replacement luxury automobile within the next six months. One section of the questionnaire asked these intenders to render preference judgments (actually degree of purchase consideration) for each of some ten different nameplates or makes of luxury automobiles specified by the manufacturer as brands thought to compete in the same market at that time (based on previous research). The ten nameplates tested were:

Lincoln Continental	Cadillac Seville
Buick Riviera	Oldsmobile Ninety-Eight
Lincoln Town Car	Mercedes 300E
BMW 325i	Volvo 740
Jaguar XJ6	Acura Legend

Unfortunately, we were not able to have access to the demographic and psychographic batteries of variables also collected for this sample of consumer intenders.

3.2 Proposed model results

Table 1 presents the Bayes Factors for model selection in comparing the 2,3,4, and 5 segment solutions. As can be seen, the four-segment solution appears to dominate in terms of both the Bayes Factor and the R-square statistic. Tables 2–5 present the resulting solutions for these analyses in sequence. This is done to examine the evolution of the solutions in relationship to analyses performed with fewer segments. For example, the two-segment solution clearly separates the foreign nameplates from the domestic ones. When one moves to the three-segment solution, the domestic segment in the two-segment solution has now been split into GM vs. Ford segments. The four-segment solution (which is selected by the Bayes factor as most parsimonious) further splits the GM segment in the three-segment solution into Cadillac vs. Olds/Buick segments. Thus, the increasing complexity of higher segment solutions can be easily tracked as one moves from lower to higher segmented solutions.

Model	BF (vs. Base line model)	R^2
2 segments	$\exp\{308\}$	0.251
3 segments	$\exp\{396\}$	0.314
4 segments	$\exp\{444\}$	0.335
5 segments	$\exp\{407\}$	0.314

Table 1. Bayes Factor results for the luxury car application

Tables 2–5 also present the size of the segments in terms of the distribution of the sample allocated to each segment. Actually, a matrix of posterior probabilities is calculated by the proposed methodology, and we instead summarize the aggregate totals by segment. Had demographic or psychographic data been available, one could perform posterior analyses to examine the specific nature of these segments in terms of both automobile types and customer types. For the four-segment solution, the size of the segments are 22.0%, 23.4%, 26.2%, and 28.4% respectively for Segment 1 (Foreign), Segment 2 (Cadillac), Segment 3 (Ford), and Segment 4 (Olds/Buick). It is interesting to note that increasing differentiation appears to occur on the domestic side of these ten nameplates. Evidently, this particular sample of intenders perceives more differentiation with the set of domestic automobiles vs. the foreign ones in the study.

Seg	% In	Cont.	Seville	Rivera	Olds98	T C	Merc	BMW	Volvo	Jag	Legend
1	46.9	0	0	0	0	0	1	1	1	1	1
2	53.1	1	1	1	1	1	0	0	0	0	0

Table 2. Two segment solution (domestic vs. foreign)

Seg	% In	Cont.	Seville	Rivera	Olds98	T C	Merc	BMW	Volvo	Jag	Legend
1	24.9	0	0	0	0	0	1	1	1	1	1
2	34.9	1	0	0	0	1	0	0	0	0	0
3	40.2	0	1	1	1	0	0	0	0	0	0

Table 3. Three segment solution (Ford vs. GM vs. Foreign)

Seg	% In	Cont.	Seville	Rivera	Olds98	T C	Merc	BMW	Volvo	Jag	Legend
1	22.0	0	0	0	0	0	1	1	1	1	1
2	23.4	0	1	0	0	0	0	0	0	0	0
3	26.2	1	0	0	0	1	0	0	0	0	0
4	28.4	0	0	1	1	0	0	0	0	0	0

Table 4. Four segment solution (Ford vs. Cadillac vs. GM(-Cadillac) vs. Foreign)

Seg	% In	Cont.	Seville	Rivera	Olds98	T C	Merc	BMW	Volvo	Jag	Legend
1	17.2	0	0	1	1	0	0	0	0	0	0
2	19.1	0	0	0	0	0	1	1	1	1	1
3	20.3	1	0	0	0	0	0	0	0	0	0
4	21.1	0	0	0	0	1	0	0	0	0	0
5	22.3	0	1	0	0	0	0	0	0	0	0

Table 5. Five segment solution (Continental vs. Cadillac vs. GM(-Cadillac) vs. Town Car vs. Foreign)

References

- BERGER, J. (1985): *Statistical Decision Theory and Bayesian Analysis*. Springer Verlag, New York, NY.
- BERNARDO, J.M. and SMITH, A.F.M. (1994): *Bayesian Theory*. John Wiley & Sons Ltd., Chichester.
- BINDER, D.A. (1978): Bayesian Cluster Analysis. *Biometrika*, 65, 31–38.
- DESARBO, W.S. and JEDIDI, K. (1995): The Spatial Representation of Heterogeneous Consideration Sets. *Marketing Science*, 14, 326–342.
- DESARBO, W.S., FONG, D.K.H., LIECHTY, J., and SAXTON, M.K. (forthcoming): A Hierarchical Bayesian Procedure for Two-Mode Cluster Analysis. *Psychometrika*.
- DIEBOLT, J. and ROBERT, C. (1994): Estimation of Finite Mixture Distributions Through Bayesian Sampling. *Journal of the Royal Statistical Society*, 56, 163–175.
- EVANS, M., GUTTMAN, I., and OLKIN, I. (1992): Numerical Aspects in Estimating the Parameters of a Mixture of Normal Distributions. *Journal of Computational and Graphical Statistics*, 1, 351–365.
- GELMAN, A. and KING, G. (1990): Estimating the Electoral Consequence of Legislative Redirecting. *Journal of the American Statistical Association*, 85, 274–282.

- GILKS, W.R., OLDFIELD, L., and RUTHERFORD, A. (1989): Bayesian Approaches to Mixtures. In: W. Knapp, B. Dorken, W.R. Gilks, and S.F. Schlossman (Eds.): *Levoctye Typing IV*. Oxford University Press, London, 6–12.
- GILKS, W.R., RICHARDSON, S., and SPIEGELHALTER, D.J. (1996): *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
- JEFFREYS, H. (1961): *Theory of Probability*. 3rd ed., Oxford University Press, London.
- KASS, R. and RAFTERY, A.E. (1995): Bayes Factors. *Journal of the American Statistical Association*, 90, 40–60.
- LAVINE, M. and WEST, M. (1992): A Bayesian Method of Classification and Discrimination. *Canadian Journal of Statistics*, 20, 451–461.
- NEWTON, M.A. and RAFTERY, A.E. (1994): Approximate Bayesian Inference by the Weighted Likelihood Bootstrap (with Discussion). *Journal of the Royal Statistical Society, Series B*, 56, 3–48.
- O'HAGAN, A. (1994): *Kendall's Advanced Theory of Statistics: Volume 2b Bayesian Inference*. John Wiley and Sons, New York, NY.
- VERDINELLI, I. and WASSERMAN, L. (1991): Bayesian Analysis of Outlier Problems Using the Gibbs Sampler. *Statistics and Computing*, 1, 105–177.

On Application of a Certain Classification Procedure to Mean Value Estimation Under Double Sampling for Nonresponse

Wojciech Gamrot

Department of Statistics,
Katowice University of Economics, ul. Bogucicka 14, 40-226 Katowice, Poland

Abstract. One of the techniques commonly used to reduce nonresponse bias is the two-phase sampling scheme. According to this scheme, when deterministic nonresponse appears, estimates of respondent and nonrespondent stratum means are weighted by sample respondent and nonrespondent fractions which estimate unknown shares of both strata in the whole population. In this paper, an alternative method of estimating these shares by using auxiliary information is considered and the application of the Ho-Kashyap (1965) classification procedure to mean value estimation is discussed. Some simulation results are presented.

1 Introduction

Let us assume that the mean value \bar{Y} of some characteristic Y in the population U of the size N is to be estimated and that nonresponse mechanism is a deterministic one. Therefore, the population can be divided into two non-overlapping strata U_1 and U_2 , of unknown sizes N_1 and N_2 respectively, such that population units belonging to U_1 always co-operate if contacted whereas units from U_2 always refuse to provide answers. Let us denote $W_1 = N_1/N$ and $W_2 = N_2/N$.

In the first phase of the survey a simple random sample s of size n is drawn without replacement from the population, according to the sampling design:

$$P_1(s) = \binom{N}{n}^{-1}. \quad (1)$$

The sample s is partitioned into two disjoint random sets $s_1 \subset U_1$ of respondents and $s_2 \subset U_2$ of nonrespondents with sizes $0 \leq n_1 \leq n$ and $0 \leq n_2 \leq n$ such that $s_1 \cup s_2 = s$, $s_1 \cap s_2 = \emptyset$ and that $n_1 + n_2 = n$. Sizes of both subsets are random, but observable variables having a hypergeometric distribution. After contacting all the units included in the sample, still nothing is known about the mean value of the variable Y in stratum U_2 , as the units from the set s_2 do not respond. To overcome this problem, a subsample u of size $n_u = cn_2$ (where $0 < c < 1$) is drawn without replacement from among the units of set s_2 , with conditional probability:

$$P_2(u|n_2) = \binom{n_2}{n_u}^{-1}. \quad (2)$$

All units in the subsample are re-contacted and it is assumed that data collection procedures used in the second phase guarantee obtaining responses for each unit. Let us define

$$\bar{y}_1 = \frac{1}{n_1} \sum_{i \in s_1} y_i, \quad \bar{y}_u = \frac{1}{n_u} \sum_{i \in u} y_i, \quad (3)$$

and consider the following statistic (see Wywial (2001)):

$$\bar{y}_s(\alpha) = \alpha \bar{y}_1 + (1 - \alpha) \bar{y}_u. \quad (4)$$

When $\alpha = n_1/n$ the statistic above takes the well-known form:

$$\bar{y}_w = \frac{n_1}{n} \bar{y}_1 + \frac{n_2}{n} \bar{y}_u. \quad (5)$$

Its variance is given by the expression (see Särndal et al. (1992)):

$$V(\bar{y}_w) = \frac{N-n}{Nn} S^2 + \frac{W_2}{n} \left(\frac{1-c}{c} \right) S_2^2, \quad (6)$$

where S^2 is the variance of the characteristic under study in the population U and S_2^2 is its variance in the stratum U_2 . In our simulation study below, this estimator will be a termed standard estimator and denoted by the symbol S. Let us assume that α is constant. Wywial (2001) derives an approximate expression (for large n) for the variance of $\bar{y}_s(\alpha)$:

$$V(\bar{y}_s(\alpha)) \approx \frac{1}{n} \left(\frac{\alpha^2}{W_1} S_1^2 + \frac{(1-\alpha)^2}{c(1-W_1)} S_2^2 \right) \quad (7)$$

and for its bias:

$$B(\bar{y}_s(\alpha)) = (W_1 - \alpha)(\bar{Y}_1 - \bar{Y}_2), \quad (8)$$

where \bar{Y}_1 is the mean value of the characteristic under study in the stratum U_1 and \bar{Y}_2 is its mean value in the stratum U_2 . Consequently, for large n the mean square error of this estimator may be expressed as:

$$MSE(\bar{y}_s(\alpha)) \approx \frac{1}{n} \left(\frac{\alpha^2}{W_1} S_1^2 + \frac{(1-\alpha)^2}{c(1-W_1)} S_2^2 \right) + (W_1 - \alpha)^2 (\bar{Y}_1 - \bar{Y}_2)^2 \quad (9)$$

and it takes its minimum value if:

$$\alpha = \frac{z}{S_1^2 + z} \quad (10)$$

where:

$$z = nW_1 \left(\frac{S_2^2}{cn(1-W_1)} + (\bar{Y}_1 - \bar{Y}_2)^2 \right). \quad (11)$$

However, the expression above can not be used to establish the optimum value of the constant α because stratum means \bar{Y}_1 and \bar{Y}_2 , as well as fractions W_1 and W_2 are unknown.

When values of k auxiliary variables x_{i1}, \dots, x_{ik} are observed for any i -th population unit, Wywial (2001) suggests to apply some discrimination methods to establish the value of the weight α as close as possible to the population respondent fraction W_1 . According to this proposition the population is divided into two classes (subsets) U'_1 and U'_2 , using classification algorithms. The division is aimed at obtaining the classes that are as close (similar) as possible to the strata U_1 and U_2 respectively. These classes can be treated as some kind of estimates of actual strata, and their sizes N'_1 and N'_2 can be used as estimates of unknown stratum sizes N_1 and N_2 . Finally the weight may be set to $\alpha = N'_1/N$.

The classes U'_1 and U'_2 may (and usually will) differ from the original strata, but resulting errors in estimating stratum sizes may be lower than errors occurring when estimating stratum sizes on the basis of the initial sample respondent fraction, according to the standard two-phase estimation strategy. Under this approach α is a random variable, because it is evaluated on the basis of sample realizations. The expression (9) describes only the conditional MSE of the estimator $\bar{y}_s(\alpha)$, without taking into account the variability of respondent fraction estimates.

A discussion of the application of several classification methods to mean value estimation under nonresponse, and their comparison by Monte Carlo simulation are given by Gamrot (2002). In this paper another classification method is considered as a means to assess the weight α .

2 Assessment of the weight α using the Ho-Kashyap algorithm

In order to divide the population into subsets U'_1 and U'_2 , let us consider a linear discriminant function of the form:

$$g(x_{i1} \dots x_{ik}) = w_0 + \sum_{i=1}^k w_i x_{ik}, \quad (12)$$

where w_1, \dots, w_k are weights corresponding to each auxiliary variable and w_0 is a constant termed *threshold weight*. Any i -th unit will be classified as belonging to the stratum U'_1 when $g(x_{i1}, \dots, x_{ik}) > 0$ and to the stratum U'_2 when $g(x_{i1}, \dots, x_{ik}) < 0$ (we neglect here the unlikely case of $g(x_{i1}, \dots, x_{ik})$ being exactly equal to zero). Admitting that $x_{j0} = 1$ for $j = 1, \dots, N$, and denoting $\mathbf{x}_i = [x_{i0}, x_{i1}, \dots, x_{ik}]$ and $\mathbf{a} = [w_0, w_1, \dots, w_k]^T$ we can express the discrimination function above in equivalent, but simpler form:

$$g(\mathbf{x}_i) = \mathbf{x}_i \mathbf{a} \quad (13)$$

It is desired to find such a vector a that $g(\mathbf{x}_i) > 0$ for any unit belonging to U_1 and $g(\mathbf{x}_i) < 0$ for any unit belonging to U_2 . Obviously, this requirement is unverifiable because the stratum membership is unknown for non-sampled units. Moreover, such a vector of weights may simply not exist. Instead we will try to find some parameter vector \mathbf{a} for which the discrimination function properly classifies units from the sample s into subsets s_1 and s_2 . We will then expect that this function properly classifies non-sampled units or at least that misclassification probability is relatively low. Let us introduce the matrix:

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{10} \dots x_{1k} \\ \vdots \quad \ddots \quad \vdots \\ x_{n0} \dots x_{nk} \end{bmatrix} \quad (14)$$

containing the observations of auxiliary variables for units included in the sample s . As indicated by Jajuga (1990), in order to determine the parameter vector \mathbf{a} we can transform the matrix \mathbf{X} by multiplying all the rows corresponding to sample nonrespondents by (-1). Hence we will introduce another matrix:

$$\mathbf{Z} = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} z_{10} \dots z_{1k} \\ \vdots \quad \ddots \quad \vdots \\ z_{n0} \dots z_{nk} \end{bmatrix} \quad (15)$$

where for any $i = 1, \dots, n$:

$$\mathbf{z}_i = \begin{cases} \mathbf{x}_i & \text{if the } i\text{-th unit responds} \\ -\mathbf{x}_i & \text{otherwise} \end{cases} \quad (16)$$

and for any $i = 1, \dots, n, j = 0, \dots, k$:

$$z_{ij} = \begin{cases} x_{ij} & \text{if the } i\text{-th unit responds} \\ -x_{ij} & \text{otherwise} \end{cases} \quad (17)$$

Consequently, if any vector \mathbf{a} properly classifies the observed units, it must satisfy the condition:

$$\mathbf{Z}\mathbf{a} > \mathbf{0}. \quad (18)$$

If there exists any vector \mathbf{a} satisfying this condition, then the set of n points in k -dimensional space, given by matrix \mathbf{Z} and representing the training sample s is said to be *separable* (see Duda et al. (2001)). For convenience we will call the matrix \mathbf{Z} *separable* if there exists some \mathbf{a} satisfying (18) and *nonseparable* otherwise. As indicated by Jajuga (1990), for separable matrix \mathbf{Z} , there may exist many solution vectors \mathbf{a}_i satisfying this condition. This suggests the need for an additional requirement to choose one of the possible solutions. Another reason to introduce such a requirement is to enable the use of gradient search (gradient descent) methods to find the solution. Such additional criterion function $J(\mathbf{a})$ should take its minimum value for some weight vector satisfying

(18). In order to construct the criterion function let us consider the problem of solving the equation:

$$\mathbf{Z}\mathbf{a} = \mathbf{b} \quad (19)$$

where $\mathbf{b} = [b_1 \dots b_n]^T$ is a vector of arbitrarily chosen positive constants called *margin vector*. Usually $n > k$ and the system of equations given above has no solutions (\mathbf{a} is overdetermined). Instead, we can find some vector \mathbf{a} , that minimizes the length of an error vector $\mathbf{e} = \mathbf{Z}\mathbf{a} - \mathbf{b}$. This is equivalent to minimization of a criterion function:

$$J(\mathbf{a}) = \|\mathbf{Z}\mathbf{a} - \mathbf{b}\|^2 = \sum_{i=1}^n (\mathbf{z}_i \mathbf{a} - b_i)^2. \quad (20)$$

The gradient of $J(\mathbf{a})$ with respect to \mathbf{a} takes the form $\nabla J(\mathbf{a}) = 2\mathbf{Z}^T(\mathbf{Z}\mathbf{a} - \mathbf{b})$ and it is equal to zero when:

$$\mathbf{a} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{b}. \quad (21)$$

Consequently, for a predefined margin vector \mathbf{b} , and nonsingular $\mathbf{Z}^T \mathbf{Z}$ the corresponding optimum vector \mathbf{a} is given by (21). However, the optimal \mathbf{b} is unknown. To overcome this, an iterative procedure has been proposed by Ho and Kashyap (1965). The procedure aims at finding solution vectors \mathbf{a}_x and $\mathbf{b}_x > 0$ simultaneously.

At the beginning the vector \mathbf{b}_0 is initialized with arbitrary, but positive constants, which allows to compute the vector \mathbf{a}_0 according to formula (21). In subsequent steps of this procedure the approximations \mathbf{a}_i and \mathbf{b}_i of solution vectors \mathbf{a}_x and \mathbf{b}_x are generated by repeatedly executing the following steps:

1. $\mathbf{e}_i := \mathbf{Z}\mathbf{a}_i - \mathbf{b}_i$
2. $\mathbf{e}_i^+ := 0.5(\mathbf{e}_i + |\mathbf{e}_i|)$
3. $\mathbf{b}_{i+1} := \mathbf{b}_i + 2t\mathbf{e}_i^+$
4. $\mathbf{a}_{i+1} := (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{b}_{i+1}$

Let us assume that $0 < t < 1$. It has been shown by Duda et al. (2001), that if the matrix \mathbf{Z} is separable then either $\|\mathbf{e}_j\| = 0$ for some j or the sequence $\|\mathbf{e}_1\|^2, \|\mathbf{e}_2\|^2, \dots$ converges to zero. In the case of \mathbf{Z} being nonseparable the sequence $\|\mathbf{e}_1\|^2, \|\mathbf{e}_2\|^2, \dots$ converges to some constant greater than zero.

The original Ho-Kashyap procedure terminates, when all the elements of error vector \mathbf{e}_i in the i -th iteration are close enough to zero (which means that a solution vector \mathbf{a}_x was found) or when all the elements of this vector are negative, which provides evidence that \mathbf{Z} is nonseparable. For the purpose of mean value estimation under nonresponse we will be interested in determining the vectors \mathbf{a}_x and \mathbf{b}_x no matter if \mathbf{Z} is separable or not (if it is not, then some points may be misclassified, but the procedure should keep the extent of misclassification modest). Consequently, we will terminate computations, if for some step the improvement in the value of $J(\mathbf{a})$ is smaller than some predefined small value Δe_{stop} . Hence, the slightly modified procedure is:

1. INITIALIZE $\mathbf{b}_0 > 0$, $\mathbf{a}_0 = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{b}_0$, $\mathbf{e}_0 = \mathbf{z}\mathbf{a}_0 - \mathbf{b}_0$, t , Δe_{stop}
2. $\mathbf{e}_i^+ := 0.5(\mathbf{e}_i + |\mathbf{e}_i|)$
3. $\mathbf{b}_{i+1} := \mathbf{b}_i + 2t\mathbf{e}_i^+$
4. $\mathbf{a}_{i+1} := (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{b}_{i+1}$
5. $\mathbf{e}_{i+1} := \mathbf{Z}\mathbf{a}_{i+1} - \mathbf{b}_{i+1}$
6. IF $\|\mathbf{e}_i - \mathbf{e}_{i+1}\| > \Delta e_{stop}$ THEN GO TO STEP 2
7. RETURN a_{i+1}, b_{i+1}

After the vector \mathbf{a}_x is determined, the discrimination function is evaluated for each non-sampled unit, which is then classified as belonging to the class U'_1 if $g(\mathbf{x}_i) > 0$ or to the class U'_2 if $g(\mathbf{x}_i) < 0$. Then the ratio N'_1/N is used as an estimate of population respondent fraction W_1 which allows to compute the population mean value estimator according to expression (4). In the sequel this mean value estimator using Ho-Kashyap discrimination procedure will be denoted by the symbol H-K.

3 Monte Carlo simulation results

A simulation study was performed to assess the properties of the proposed estimator. The goal of the simulation was to investigate how the bias and the MSE of H-K estimator depend on the initial sample size n , and to compare them with the properties of the standard estimator S. The experiments were carried out by repeatedly generating the values of the variable under study and three auxiliary variables for every population unit, using a pseudo-random number generator. Consequently, several complete populations of pseudo-random numbers were generated. It was assumed that within-stratum distribution of all variables is multivariate normal. Parameters of the generator differed between respondent and nonrespondent stratum. Consequently, the probability distribution of population characteristics was equivalent to the mixture of adequate within-stratum probability distributions. For the stratum U_1 mean values of the variable under study and all the auxiliary variables were set to 0, whereas for the stratum U_2 all these mean values were set to 2. All variables were uncorrelated within strata and their within-stratum standard deviations were set to one. Stratum sizes were assumed to be constant in successive populations and equal to $N_1 = 600$ and $N_2 = 400$, resulting in a population respondent fraction being equal to 60%.

From each population generated this way, several samples and appropriate subsamples, were repeatedly drawn. The size of a subsample was always equal to 30% of nonrespondent number n_2 observed in the first-phase sample. For each sample-subsample pair the realizations of the H-K estimator and the standard estimator were computed. By this way an empirical distribution of mean value estimates was obtained for each population. On the basis of each distribution the bias and the MSE of estimates were computed. Then the bias and the MSE of the estimator were computed by averaging biases and MSE's over all populations. The simulation was executed for $n = 40, 50, \dots, 200$. For

each n from this range a total of 100 populations were generated and 100 samples were drawn from each population. The observed relative accuracy of the strategies (proportion of the MSE of H-K estimator to the MSE of standard estimator S) as a function of initial sample size n is presented on the Figure 1.

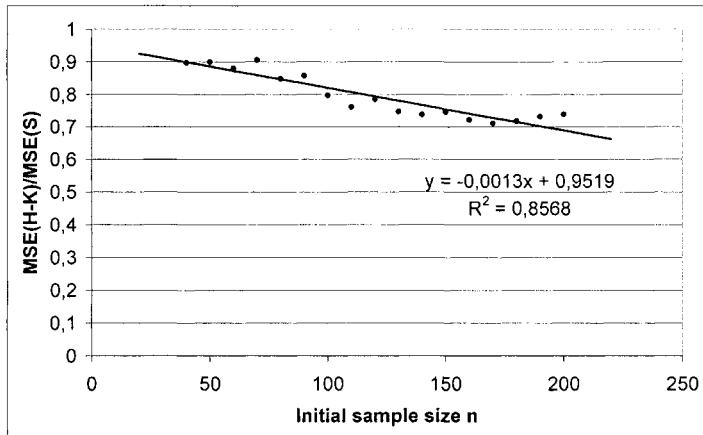


Fig. 1. The MSE of the H-K estimator, relative to the MSE of the standard estimator as a function of initial sample size n .

As it can be seen on the graph, for any value of n for which the simulations were executed, the observed relative accuracy is below one. This means that the H-K estimator has a lower MSE than the standard estimator S. If the initial sample size n increases, then the advantage of H-K estimator also increases. To illustrate this the linear regression function was fitted to the observed relative MSE data using least squares method, resulting in determination coefficient $r_{xy}^2 = 0.8568$.

Figure 2 shows the proportion of squared bias to the MSE of the H-K estimator, plotted as a function of the initial sample size n . The growth of this ratio with increasing n is mainly due to the decrease of MSE in the denominator. However, it should be emphasized that the ratio does not exceed 0.3% for any value of n . Low absolute values of this ratio mean, that the squared bias is lower by more than two orders of magnitude than MSE of population mean estimates. Consequently the bias of the H-K estimator may be treated as negligible, and that the estimation error is mainly due to the non-systematic variability of mean value estimates.

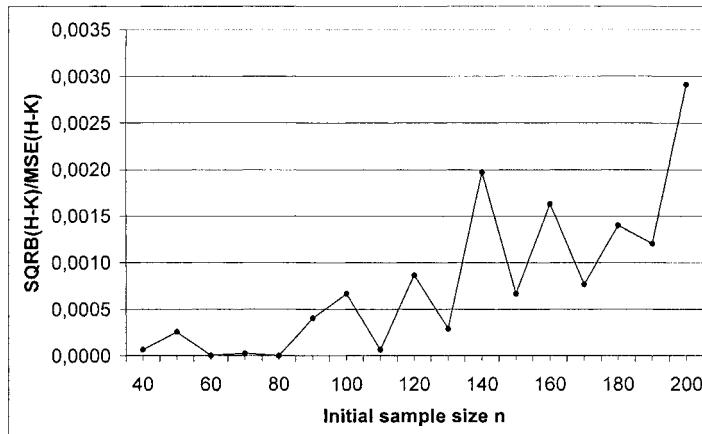


Fig. 2. The dependence between initial sample size n and the ratio of squared bias to the overall MSE of the H-K estimator.

4 Conclusions

The results presented above support the hypothesis that the use of discrimination methods may increase the accuracy of population mean estimates. The observed contribution of the bias in the MSE of the presented estimator was very modest, which suggests, that in practice this estimator may be treated as nearly unbiased. Its advantage is also the ability to achieve alternative estimates on the basis of the same two-phase sample as in the standard, non-discrimination-based strategy, whenever the auxiliary information is available.

References

- DUDA, R.O., HART, P.E., and STORK, D.G. (2001): *Pattern Classification*. Wiley, New York.
- GAMROT, W. (2002): On Application of Some Discrimination Methods to Mean Value Estimation in the Presence of Nonresponse. In: J. Wywial (Ed.): *Metoda reprezentacyjna w Badaniach Ekonomiczno-Społecznych*, Katowice, 37–50.
- HO, Y.C. and KASHYAP, R.L. (1965): An algorithm for linear inequalities and its applications. *IEEE Transactions on Electronic Computers*, 14, 683–688.
- JAJUGA, K. (1990): *Statystyczna teoria rozpoznawania obrazów*. PWN, Warszawa.
- SÄRNDAL, C.E., SWENSSON, B., and WRETMAN, J. (1997): *Model Assisted Survey Sampling*. Springer, New York.
- WYWIAL, J. (2001): On Estimation of Population Mean in the Case When Nonrespondents Are Present. *Prace Naukowe AE Wrocław*, 8, 906, 13–21.

Regression Clustering with Redescending M-Estimators

Tim Garlipp and Christine H. Müller

Universität Oldenburg, Fachbereich 6 Mathematik
Postfach 2503, D - 26111 Oldenburg, Germany

Abstract. We use the local maxima of a redescending M-estimator to identify clusters, a method proposed already by Morgenthaler (1990) for finding regression clusters. We work out the method not only for classical regression but also for orthogonal regression and multivariate locations and give consistency results for all three cases. The approach of orthogonal regression is applied to the identification of edges in noisy images.

1 Introduction

For independently and identically distributed random variables Y_1, \dots, Y_N , the (location-) M-estimator is defined as (global or some local) maximum of

$$H_N(y) = \sum_{n=1}^N \rho(Y_n - y).$$

If ρ' is strictly monotone the objective function is unimodal, so that the maximum is unique. For example with $\rho(y) = -y^2$, the maximum is attained at the mean of the observations. But using score functions with redescending derivatives, $H_n(y)$ can have several local maxima, what especially has the disadvantage that computation of the M-estimator is more complicated. But since these local maxima correspond to substructures in the data, they can be used for clustering.

Section 2 motivates this approach in the case of location clustering. In Section 3 it is applied to clustering regression data in the case of classical vertical regression (Section 3.1) and in the case of orthogonal regression, which has several advantages (Section 3.2). In Section 4 the orthogonal regression method is used for identifying edges in noisy images.

All proofs can be found in Müller and Garlipp (2003).

2 Clustering with redescending M-estimators

Let $y_N = (y_{1N}, \dots, y_{NN})$ be a realization of independently and identically distributed random variables $Y_{nN} \in \mathbb{R}^k$ following a distribution with density

h . The positions of the local maxima of the density h are considered as true cluster center points and are denoted by \mathcal{M} , i.e.

$$\mathcal{M} := \{\mu \in \mathbb{R}^k; h(\mu) \text{ has local maximum at } \mu\}.$$

If the distribution of Y_{nN} is a mixture of distributions with unimodal densities, for example $Y_{nN} = \mu_l + E_{nN}$ with probability γ_l ($\sum \gamma_l = 1$) and E_{nN} has density f_l with maximum at 0, then the local maxima of the density h are attained at the maxima μ_l of the densities $f_l(\cdot - \mu_l)$ only if the supports of the $f_l(\cdot - \mu_l)$ do not overlap, what in general is not the case. Nevertheless, to define the true cluster center points via the maxima of the density h is more general, since the densities within the clusters are not known in practice and this definition is even appropriate for the general situation, where no assumptions for the model are made and only a general density h is used. Hence the aim is to estimate the positions of the local maxima of h .

Having the result that kernel density estimates are consistent estimates of the density h (see e.g. Silverman (1986)), we estimate the local maxima of h and thus the center points by the local maxima of the estimated density given by the kernel estimator. In Theorem 1 we show the consistency of this estimator under some regularity conditions. A kernel density estimator for $h(\mu)$ is given by

$$H_N(\mu, y_N) := \frac{1}{N} \sum_{n=1}^N \frac{1}{s_N^k} \rho \left(\frac{y_{nN} - \mu}{s_N} \right),$$

where $\mu \in \mathbb{R}^k$, $\rho : \mathbb{R}^k \rightarrow \mathbb{R}^+$ is the kernel function and $s_N \in \mathbb{R}^+ \setminus \{0\}$ is the bandwidth. If s_N converges to zero, then $H_N(\mu, y_N)$ converges to $h(\mu)$ in probability under some regularity conditions. Hence, the local maxima of $H_N(\cdot, y_N)$, which also can be considered as M-estimates with respect to the objective function $H_N(\cdot, y_N)$, can be used as estimate for the set \mathcal{M} .

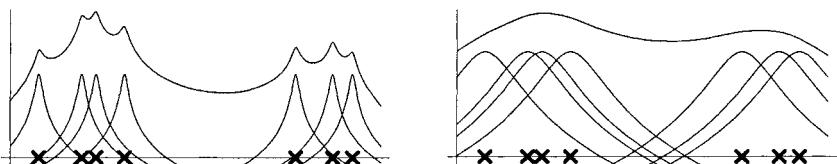


Fig. 1. Some one dimensional observations with corresponding score functions and their sum (objective function) with small (left) and large (right) scale parameter.

Usually ρ will be a unimodal density. Hence, if the scale parameter s_N is small enough and the distance between the y_{nN} are large enough, every y_{nN} is a local maximum. But usually there is so much overlap of the $\rho \left(\frac{1}{s_N} (y_{nN} - \mu) \right)$ that none of the y_{nN} is a local maximum (Figure 1). However, searching the local maxima in increasing direction starting at any y_{nN}

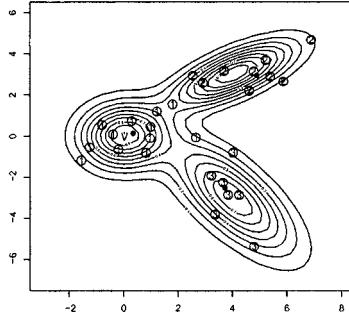


Fig. 2. Contour plot of the density of the mixture of three two dimensional normal distributions with generated observations and estimated cluster center points.

should provide the relevant maxima. This is an approach used also by Chu et al. (1998) for constructing corner preserving M-smoother for image reconstruction. The consistency of these M-smoothers even at jumps was shown by Hillebrand and Müller (2001). A similar proof can be used here for the consistency of the set

$$\mathcal{M}_N(y_N) := \{\mu \in \mathbb{R}^k; H_N(\mu, y_N) \text{ has local maximum at } \mu\}$$

which is the estimate of the set \mathcal{M} of the positions of the true local maxima. The local maxima of $H_N(\cdot, y_N)$ can be found by Newton Raphson method starting at any y_{nN} with $n = 1, \dots, N$.

To avoid problems like "bump hunting" (see e.g. Donoho (1988)), we need not only pointwise convergence of $H_N(\mu, y_N)$ to $h(\mu)$ but additional assumptions to achieve the consistency of the set $\mathcal{M}_N(y_N)$ for the set \mathcal{M} . One is the uniform convergence which can be achieved by intersecting $\mathcal{M}_N(y_N)$ with a compact subset of \mathbb{R}^k . Appropriate compact subsets are given by

$$\Theta_\eta := \left\{ \mu \in \mathbb{R}^k; h(\mu) \geq \frac{1}{\eta} \right\} \text{ with } \eta \in \mathbb{N}.$$

Then, with $\lambda_{\max} h''(\mu)$ denoting the maximum eigenvalue of $h''(\mu)$, we have

Theorem 1. *If $\min\{|\lambda_{\max} h''(\mu)|; \mu \in \mathcal{M}_0\} > 0$, then there exists $\eta_0 \in \mathbb{N}$ so that for all $\eta \geq \eta_0, \epsilon > 0, \delta > 0$ there exists an $N_0 \in \mathbb{N}$ with*

$$P\left(\mathcal{M}_N(Y_N) \cap \Theta_\eta \subset \mathcal{U}_\delta(\mathcal{M}) \text{ and } \mathcal{M} \subset \mathcal{U}_\delta(\mathcal{M}_N(Y_N) \cap \Theta_\eta)\right) > 1 - \epsilon$$

for all $N \geq N_0$, where $\mathcal{U}_\delta(\mathcal{M}) := \{\mu \in \mathbb{R}^k; \text{there exists a } \mu_0 \in \mathcal{M} \text{ with } \|\mu - \mu_0\| < \delta\}$ and $\mathcal{M}_0 := \{\mu \in \mathbb{R}^k; h'(\mu) = 0 \text{ and } h(\mu) > 0\}$.

Figure 2 shows a contour plot of the density $h(\mu)$ of a mixture of three two dimensional normal distributions with parameters $\mu_1 = (0, 0)^\top$, $\Sigma_1 = \begin{pmatrix} 10 \\ 01 \end{pmatrix}$, $\mu_2 = (4, 3)^\top$, $\Sigma_2 = \begin{pmatrix} 21 \\ 11 \end{pmatrix}$, $\mu_3 = (4, -3)^\top$, $\Sigma_3 = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}$ and $\gamma_1 = \gamma_2 = 0.36, \gamma_3 = 0.28$ with 28 generated observations and the three estimated local maxima (black dots).

3 Clustering of regression data

3.1 Vertical regression

Regard a mixture of L regression models with different parameter vectors β_l . Then we have observations $z_N = (z_{1N}, \dots, z_{NN})$, which are realizations of independently and identically distributed random variables $Z_{nN} := (X_{nN}^\top, Y_{nN})^\top$, with

$$Y_{nN} = X_{nN}^\top \beta_l + E_{nN}$$

if the n 'th observation is coming from the l 'th cluster.

In the case of $L = 1$, the M-estimator for the regression parameter β is defined as a maximum point of the objective function

$$H_N(\beta, z_N) := \frac{1}{N} \sum_{n=1}^N \frac{1}{s_N} \rho \left(\frac{y_{nN} - x_{nN}^\top \beta}{s_N} \right),$$

where $\rho : \mathbb{R} \rightarrow \mathbb{R}^+$ is the score function and $s_N \in \mathbb{R}^+ \setminus \{0\}$ is a scale parameter (see e.g. Huber (1973, 1981), Hampel et al. (1986)).

If ρ is not convex, what means that the derivative of ρ is redescending, then $H_N(\cdot, z_N)$ has several local maxima. As Morgenthaler (1990), Hennig (1997, 2003), and Chen et al. (2001) already proposed, these can be used for finding regression clusters. Under some regularity conditions for $s_N \rightarrow 0$ and ρ , we have then

$$H_N(\beta, Z_N) \xrightarrow{N \rightarrow \infty} h(\beta) := \sum_{l=1}^L \gamma_l \int f(x^\top (\beta - \beta_l)) G_l(dx)$$

in probability for all $\beta \in \mathbb{R}^p$, where G_l is the distribution of X_{nN} coming from the l 'th cluster and f denotes the density function of the distribution of E_{nN} . Again $\gamma_l > 0$ denotes the probability that the n 'th observation is coming from the l 'th cluster and $\sum_{l=1}^L \gamma_l = 1$ holds. The function h plays now the same role as the density h in multivariate density estimation.

Under enough separation the local maxima of h are attained at β_1, \dots, β_L . Hence as in the multivariate case we regard the positions of the local maxima of h as the true parameter vectors which shall be estimated. Let \mathcal{M} be the set of the positions of these local maxima, i.e.

$$\mathcal{M} := \{\beta \in \mathbb{R}^p; h(\beta) \text{ has local maximum at } \beta\},$$

which can be estimated by

$$M_N(z_N) := \{\beta \in \mathbb{R}^p; H_N(\beta, z_N) \text{ has local maximum at } \beta\}.$$

The local maxima of $H_N(\cdot, z_N)$ can be found by Newton Raphson method starting at any hyperplane through $(x_{n_1 N}^\top, y_{n_1 N}), \dots, (x_{n_p N}^\top, y_{n_p N})$ with $\{n_1, \dots, n_p\} \subset \{1, \dots, N\}$.

As in the multivariate case, $\mathcal{M}_N(z_N)$ is a consistent estimator for \mathcal{M} if it is intersected with a compact subset, which is here

$$\Theta_\eta := \left\{ \beta \in \mathbb{R}^p; h(\beta) \geq \frac{1}{\eta} \right\} \text{ with } \eta \in \mathbb{N}.$$

However, here the compactness of Θ_η is not always satisfied. In particular, it is not satisfied if one of the distributions G_l is discrete so that regression experiments with repetitions at finite design points are excluded.

Hence, with $\mathcal{U}_\delta(\mathcal{M})$ and \mathcal{M}_0 as in Theorem 1 we have the

Theorem 2. *If Θ_η is compact for all $\eta \in \mathbb{N}$ and $\min\{|\lambda_{\max} h''(\beta)|; \beta \in \mathcal{M}_0\} > 0$, then there exists $\eta_0 \in \mathbb{N}$ so that for all $\eta \geq \eta_0, \epsilon > 0, \delta > 0$ there exists an $N_0 \in \mathbb{N}$ with*

$$P\left(\mathcal{M}_N(Z_N) \cap \Theta_\eta \subset \mathcal{U}_\delta(\mathcal{M}) \text{ and } \mathcal{M} \subset \mathcal{U}_\delta(\mathcal{M}_N(Z_N) \cap \Theta_\eta)\right) > 1 - \epsilon$$

for all $N \geq N_0$.

3.2 Orthogonal regression

For orthogonal regression usually an error-in-variable model is assumed. Considering a mixture of L regressions with parameters $(a_l^\top, b_l) \in S_1 \times \mathbb{R}$, ($S_1 = \{a \in \mathbb{R}^p : \|a\| = 1\}$), this means that we have observations $z_N = (z_{1N}, \dots, z_{NN})$, which are realizations of independent and identically distributed random variables $Z_{nN} := (V_{nN}^\top, W_{nN})^\top$, with

$$(V_{nN}^\top, W_{nN}) = (X_{nN}^\top, Y_{nN}) + (E_{1nN}^\top, E_{2nN})$$

for $n = 1, \dots, N$, where $(X_{nN}^\top, Y_{nN}), E_{1nN}, E_{2nN}$ are independent, $X_{nN}, V_{nN}, E_{1nN} \in \mathbb{R}^{p-1}$, $Y_{nN}, W_{nN}, E_{2nN} \in \mathbb{R}$, and

$$a_l^\top \begin{pmatrix} X_{nN} \\ Y_{nN} \end{pmatrix} = b_l \text{ almost surely,}$$

for Z_{nN} coming from the l -th regression.

In the case of $L = 1$, an M-estimator for (a, b) was proposed by Zamar (1989) and extends the orthogonal least squares regression estimator. It is defined as a maximum point of the objective function

$$H_N(a, b, z_N) := \frac{1}{N} \sum_{n=1}^N \frac{1}{s_N} \rho \left(\frac{a^\top z_{nN} - b}{s_N} \right),$$

where $\rho : \mathbb{R} \rightarrow \mathbb{R}^+$ is the score function and $s_N \in \mathbb{R}^+ \setminus \{0\}$ is a scale parameter.

For finding regression clusters, redescending M-estimators for orthogonal regression were also proposed by Chen et al. (2001). Under some regularity conditions for $s_N \rightarrow 0$ and ρ , we have then

$$H_N(a, b, Z_N) \xrightarrow{N \rightarrow \infty} h(a, b)$$

in probability for all $(a^\top, b) \in S_1 \times \mathbb{R}$, where $h(a, b) = f_{a^\top Z_{nN}}(b)$ is the density

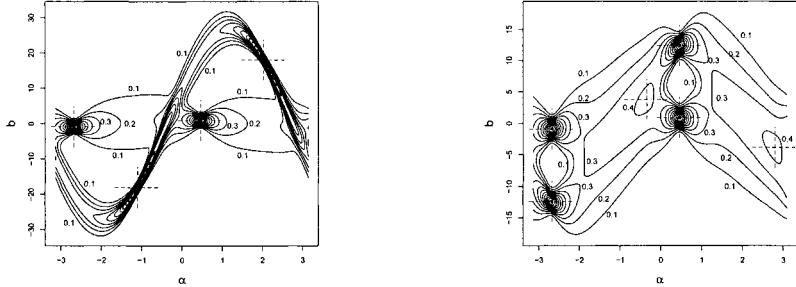


Fig. 3. Contour plot of the limit function $h(a, b)$ for a mixture of two nonparallel (left, $(\alpha_1, b_1) = (0.46, 0.9)$, $(\alpha_2, b_2) = (-1.11, -18)$) and two parallel (right, $(\alpha_1, b_1) = (0.46, 0.9)$, $(\alpha_2, b_2) = (0.46, 12.4)$) regression lines.

of the distribution on $a^\top Z_{nN}$. Note that, in opposite to classical vertical regression, the function $h(a, b)$ again is a density and shows therefore more relations to the function h in the multivariate case of Section 2. Moreover, as in Section 3.1, h is independent of ρ .

If the regression hyperplanes given by (a_l^\top, b_l) are enough separated, then $h(a, b)$ will have local maxima at $(a, b) = (a_l, b_l)$.

See for example Figure 3 for the two-dimensional case with $a_l = (\cos(\alpha_l), \sin(\alpha_l))^\top$. Note that the symmetry in Figure 3 is caused by the π -periodicity of the parameter α . Hence it turns out for orthogonal regression that, for clusters around nonparallel lines, only two local maxima appear where, for clusters around two parallel lines, a third local maximum with a rather small height appears. Figure 4 shows a simulation for both cases. Here, with the used scale parameter $s_N = 2$, the objective function $H_n(a, b, z_n)$ has a third local maximum also in the case of nonparallel lines but again with a smaller height.

The aim is now to estimate the local maxima of $h(a, b)$, or more precisely, the set

$$\mathcal{M} := \{(a^\top, b) \in S_1 \times \mathbb{R}; h(a, b) \text{ has local maximum at } (a^\top, b)\}.$$

As for classical vertical regression, if the derivative of ρ is redescending, then $H_N(a, b, z_N)$ has several local maxima so that we define

$$\begin{aligned} M_N(z_N) &:= \\ &\{(a^\top, b) \in S_1 \times \mathbb{R}; H_N(a, b, z_N) \text{ has local maximum at } (a^\top, b)\}. \end{aligned} \tag{1}$$

The local maxima of $H_N(\cdot, z_N)$ can be found as for vertical regression (see Section 3.1).

As before, the consistency of $M_N(z_N)$ can be shown only if $M_N(z_N)$ is intersected with the set

$$\Theta_\eta := \left\{ (a^\top, b) \in S_1 \times \mathbb{R}; h(a, b) \geq \frac{1}{\eta} \right\}.$$

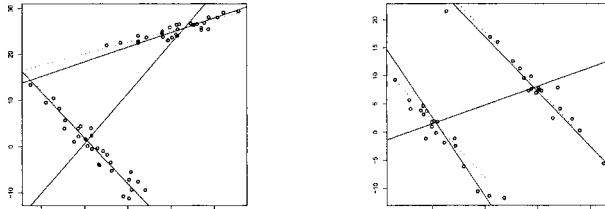


Fig. 4. True (dashed) and estimated (solid) regression lines in the case of two nonparallel and two parallel regression lines.

Since a is lying in the compact set S_1 and $h(a, \cdot)$ is a density function, the compactness of Θ_η holds here for all distributions of the regressor X_{nN} . Hence, orthogonal regression is also in this sense superior to classical vertical regression where a restriction on the distribution of X_{nN} is necessary to ensure the compactness of Θ_η (see Section 3.1).

With $\mathcal{U}_\delta(\mathcal{M})$ and \mathcal{M}_0 as in Theorem 1 we have the

Theorem 3. *If $\min\{|\lambda_{\max} h''(a, b)|; (a^\top, b) \in \mathcal{M}_0\} > 0$, then there exists $\eta_0 \in \mathbb{N}$ so that for all $\eta \geq \eta_0, \epsilon > 0, \delta > 0$ there exists an $N_0 \in \mathbb{N}$ with*

$$P\left(\mathcal{M}_N(Z_N) \cap \Theta_\eta \subset \mathcal{U}_\delta(\mathcal{M}) \text{ and } \mathcal{M} \subset \mathcal{U}_\delta(\mathcal{M}_N(Z_N) \cap \Theta_\eta)\right) > 1 - \epsilon$$

for all $N \geq N_0$.

4 Edge identification

As an application of the orthogonal regression cluster method, we use it to detect edges in noisy images (see Figure 5.A). We first use a generalized version of the Rotational Density Kernel Estimator (RDKE) introduced by Qiu (1997) to estimate those pixels, which may belong to one of the edges, which correspond to the regression lines in our model. Then, these points are used as observations z_{nN} .

We choose the RDKE-method because it does not only estimate the points lying on the edges like other methods do, but also the direction of the jump curve at these points. This provides canonical start values for the Newton Raphson method, namely the lines given by the estimated points and directions, which we used instead of those given by any two observations (see the remark after (1) in Section 3.2). Applying a multiple test based on the RDKE-method, we found 2199 points, which could belong to one of the edges (see Figure 5.B). For details see Müller and Garlipp (2003). On these points, we apply the orthogonal regression estimator with the density of the standard normal distribution as score function ρ . The scale parameter s_N is chosen with respect to the window size of the RDKE method (for details, see again Müller and Garlipp (2003)). For deciding which of the seven found center lines (Figure 5.C) belong to the true clusters, we used the absolute height of the local maxima. The result is shown in Figure 5.D.

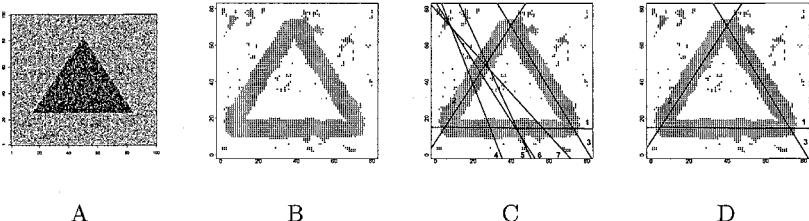


Fig. 5. Original image with 100×100 pixels, overlayed by normal distributed noise (A); Estimated jump points, respectively observations $z_{n,2199}$ (B); Observations z_{2199} with the estimated cluster lines $M_{2199}(z_{2199})$ (C); Observations with the three center lines with the largest maxima (D).

References

- CHEN, H., MEER, P., and TYLER, D.E. (2001): Robust regression for data with multiple structures. *Computer Vision and Pattern Recognition Conference, Kauai, Hawaii, December 2001, vol. I*, 1069–1075.
- CHU, C.K., GLAD, I.K., GODTLIEBSEN, F., and MARRON, J.S. (1998): Edge-preserving smoothers for image processing. *Journal of The American Statistical Association*, 93, 526–541.
- DONOHO, D.L. (1988): One-sided inference about functionals of a density. *Annals of Statistics*, 16, 1390–1420.
- HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., and STAHEL, W.A. (1986): *Robust Statistics - The Approach Based on Influence Functions*. John Wiley, New York.
- HENNIG, C. (1997): Fixed Point Clusters and Their Relation to Stochastic Models. In: R. Klar and O. Opitz (Eds.): *Classification and knowledge organisation*. Springer, Berlin, 20–28.
- HENNIG, C. (2003): Clusters, outliers, and regression: Fixed point clusters. *Journal of Multivariate Analysis* 86/1, 183–212.
- HILLEBRAND, M. and MÜLLER, CH.H. (2001): On consistency of redescending M-kernel smoothers. *Submitted*.
- HUBER, P.J. (1973): Robust regression: Asymptotics, conjectures, and Monte Carlo. *Annals of Statistics*, 1, 799–821.
- HUBER, P.J. (1981): *Robust Statistics*. John Wiley, New York.
- MORGENTHALER, S. (1990): Fitting redescending M-estimators in regression. In: H.D. Lawrence and S. Arthur (Eds.): *Robust Regression*. Dekker, New York, 105–128.
- MÜLLER, CH.H. and GARLIIPP, T. (2003): Simple consistent cluster methods based on redescending M-estimators with an application to edge identification in images. *Revised version for: Journal of Multivariate Analysis*.
- QIU, P. (1997): Nonparametric estimation of jump surface. *The Indian Journal of Statistics*, 59, Series A, 268–294.
- SILVERMAN, B.W. (1986): *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London.
- ZAMAR, R.H. (1989): Robust estimation in the errors-in-variables model. *Biometrika*, 76, 149–160.

ClusCorr98 - Adaptive Clustering, Multivariate Visualization, and Validation of Results

Hans-Joachim Mucha¹ and Hans-Georg Bartel²

¹ Weierstraß-Institute of Applied Analysis and Stochastic,
D-10117 Berlin, Germany

² Institut für Chemie, Humboldt-Universität zu Berlin,
D-12489 Berlin, Germany

Abstract. An overview over a new release of the statistical software ClusCorr98 will be given. The emphasis of this software lies on an extended collection of exploratory and model-based clustering techniques with in-built validation via resampling. Using special weights of observations leads to well-known resampling techniques. By doing so, the appropriate number of clusters can be validated. As an illustration of an interesting feature of ClusCorr98, a general validation of results of hierarchical clustering based on the adjusted Rand index is recommended. It is applied to demographical data from economics. Here the stability of each cluster can be assessed additionally.

1 Introduction

Cluster analysis aims at finding interesting structures directly from the data without using any background knowledge. In the following exhaustive partitions $P(I, K)$ of the set of I objects (observations) into K non-empty clusters (subsets, groups) C_k are considered. The clusters are pair-wise disjointed. On the other hand, a hierarchy is a sequence of nested partitions. This is usually the result of a hierarchical cluster analysis. Here the validation of hierarchies based on their partitions will be recommended.

Both well-known model-based and heuristic clustering techniques are part of the statistical software ClusCorr98. At most one will set up new statistical hypotheses about the data. On the other hand clustering should result at least in practical useful partitions or hierarchies. More details and many more references can be found in the monographs of Jain and Dubes (1988), Kaufman and Rousseeuw (1990), Mucha (1992), and Gordon (1999).

Beside the most general Gaussian model for clustering, two simple models will be considered in a generalised form using weighted observations. They lead to the sum-of-squares and logarithmic sum-of-squares criterion. Both criteria can be formulated in an equivalent fashion using pair-wise distances between observations. Here the "decomposition" of the criterion into the part of weights of observations and the fixed part of pair-wise distances becomes obvious. This independence of distances from weights is important in view of

the automatic validation technique that will be applied here to demographical data from economics.

2 Some features of ClusCorr98

Model-based and heuristic clustering techniques are part of the statistical software ClusCorr98. Moreover we offer multivariate visualization techniques like principal components analysis (PCA). ClusCorr98 uses the Excel spreadsheet environment and its database connectivity. The programming language is Visual Basic for Applications (VBA).

A typical feature is the use of both weights of variables and weights of observations in the algorithms. For example, adaptive cluster analysis is based on special weights of variables that are used in the distance measures. It is recommended to use these adaptive distances in clustering as well as for multivariate visualization techniques. In that way one can find usually much more stable solutions (Mucha (1992), Mucha et al. (2002a)). Especially in case of models like sum of squares or logarithmic sum of squares the performance can be improved when some of the variables do not reflect a cluster structure. These adaptive weights should also be used in projection techniques like PCA.

Clustering techniques based on special weights of observations can deal with both huge data sets and outliers. Core-based clustering is one example (Mucha et al. (2003)). A core is a dense region in the high-dimensional space that can be represented by its most typical observation, by its centroid or, more generally, by assigning weight functions to the observations. In the case of representatives of cores, one has to weight them proportional to the cardinality of the cores. In the case of outliers, one has to downweight them in order to reduce their influence (Mucha et al. (2002b), Mucha et al. (2003)). Almost all techniques of high dimensional data visualization (multivariate graphics, projection techniques) can also take into account weighted observations.

ClusCorr98 offers some automatic validation techniques that can be considered also as a so-called in-built validation of the number of clusters. In the case of hierarchical cluster analysis, this in-built validation can be used for the investigation of the stability of each cluster.

3 Model-based Gaussian clustering

Concerning model-based clustering the papers of Banfield and Raftery (1993) and Fraley (1996) give 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 when the covariance matrix Σ_k of each cluster k is allowed to vary completely. Then the log-likelihood is maximized

whenever the partition $P(I, K)$ minimizes

$$Y_K = \sum_{k=1}^K n_k \log \left| \frac{\mathbf{W}_k}{n_k} \right|. \quad (1)$$

Herein $\mathbf{W}_k = \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T$ is the sample cross-product matrix for the k -th cluster C_k , and $\bar{\mathbf{x}}_k$ 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 $\Sigma_k = \lambda \mathbf{I}$, the well-known sum-of-squares criterion has to be minimized. It can be written in the following equivalent form without explicit specification of cluster centres (centroids) $\bar{\mathbf{x}}_k$

$$\sum_{k=1}^K \text{tr}(\mathbf{W}_k) = \sum_{k=1}^K 1/n_k \sum_{i \in C_k} \sum_{l \in C_k, l > i} d_{il}, \quad (2)$$

where d_{il} is the pair-wise squared Euclidean distance between the observations i and l . When the covariance matrix of each cluster is constrained to be $\Sigma_k = \lambda_k \mathbf{I}$, the logarithmic sum-of-squares criterion or its equivalent formulation

$$\sum_{k=1}^K n_k \log \text{tr}(\mathbf{W}_k/n_k) = \sum_{k=1}^K n_k \log \left(\sum_{i \in C_k} \sum_{l \in C_k, l > i} \frac{1}{n_k^2} d_{il} \right), \quad (3)$$

has to be minimized. Another benefit of clustering based on pair-wise distances is the more general meaning of distances. For example, going via distances allows cluster analysis of mixed data (Gower (1971)).

4 Weighted observations

The right part of the above given criterion (2) can be generalized by using nonnegative weights of observations m_i to

$$V_K = \sum_{k=1}^K \frac{1}{M_k} \sum_{i \in C_k} m_i \sum_{l \in C_k, l > i} m_l d_{il}, \quad (4)$$

where $M_k = \sum_{i \in C_k} m_i$ denotes the weight of cluster C_k ($M_k > 0$). From (3) the generalized logarithmic sum-of-squares criterion can be derived:

$$\sum_{k=1}^K M_k \log \left(\sum_{i \in C_k} \sum_{l \in C_k, l > i} \frac{m_i m_l}{M_k^2} d_{il} \right). \quad (5)$$

5 Resampling techniques and weights of observations

The criteria of section 3 for bootstrap samples and subsets are equivalent to the choices of weights given in the equations. The weights m_i can be used for resampling purposes. Once a distance matrix is figured out it will be fixed during simulations. One has to change the weights only. For example, the well-known bootstrap resampling technique gives the rules to choose the following random 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 for bootstrap-resampling with replacement. Now let's focus on effective simulations based on pair-wise distances that will include some of the well-known model-based clustering algorithms (*K-means*, *Log-K-means*, *Ward*, and *LogWard*, see Mucha et al. (2003)). The stability of every hierarchical cluster analysis method based on pair-wise distances can be investigated with the following approach of assigning these special weights to the observations:

$$m_i = \begin{cases} 1 & \text{if observation } i \text{ is drawn randomly} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

Here $p = \sum_i m_i$ gives the number of drawn objects. Thus $p = I/2$ means drawing half of the total sample at random. This resampling technique is without replication. 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 one of several options, they can be allocated after clustering into the partitions and hierarchies according to their distance values. Usually this will be done by k nearest neighbour classification.

6 In-built validation of results of clustering

Partitions are the basic results of cluster analysis that cover also hierarchies. The latter can be comprehended as a set of partitions. Therefore comparing any two partitions P and Q becomes a basic and general purpose tool for a validation of cluster analysis results (Table 1). This in-built validation is an automatic technique with default values for the parameters of simulations. The following defaults are in use: the interval for the number of clusters is [2, 9], the number of replicates is 250, and the number of neighbours used for clustering of the supplementary objects is 3. Of course, these values are recommendations only, and they can be changed by the user.

6.1 Rand's measure for comparing partitions

The key approach for comparing partitions is based on comparison of pairs of objects concerning their class membership (Rand (1971)). In order to compare two partitions $P(I, K)$ of the set of I objects into K clusters and $Q(I, L)$ of I observations into L clusters, respectively, the Rand similarity index R^* can be applied: $R^* = (a + d)/\binom{I}{2}$. Here the quantities a and d count the pairwise matches that can be seen as agreements (correspondence, see Table 1). Equivalently, the Rand index R^* can be expressed by using a contingency table that can be 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 tables of pairs of observations concerning their cluster membership in 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		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. It has the important advantage that the stability of every single cluster can be investigated additionally. 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) recommend the adjusted Rand index R based under the assumption of the generalized hypergeometric model:

$$R = \frac{\sum_{k=1}^K \sum_{l=1}^L \binom{n_{kl}}{2} - [\sum_{k=1}^K \binom{n_{k+}}{2} \sum_{l=1}^L \binom{n_{+l}}{2}] / \binom{I}{2}}{1/2[\sum_{k=1}^K \binom{n_{k+}}{2} + \sum_{l=1}^L \binom{n_{+l}}{2}] - [\sum_{k=1}^K \binom{n_{k+}}{2} \sum_{l=1}^L \binom{n_{+l}}{2}] / \binom{I}{2}}. \quad (7)$$

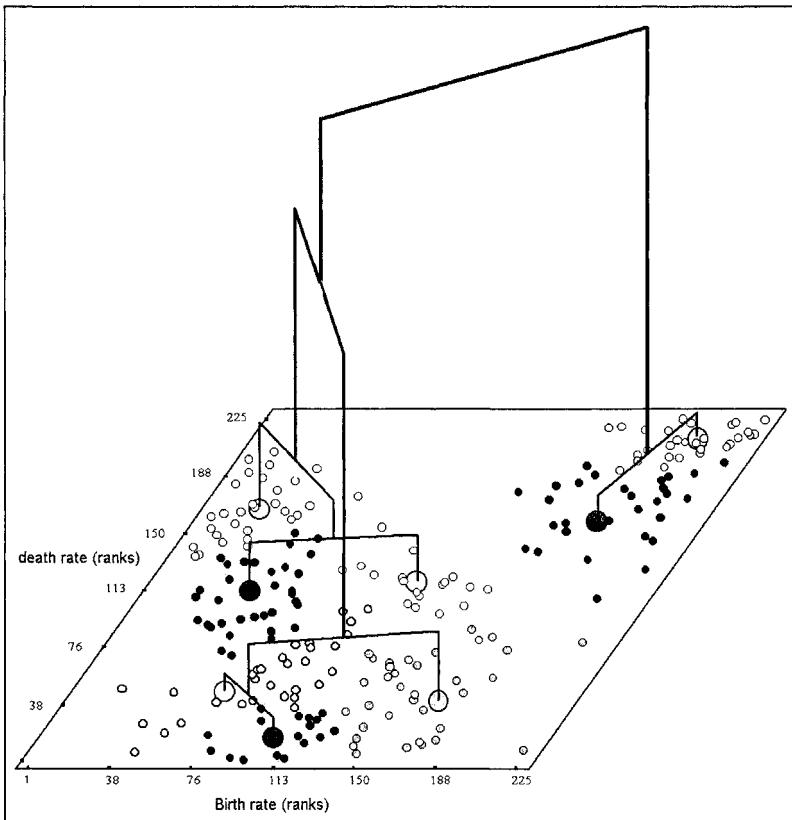


Fig. 1. Graphical presentation of *Ward's* hierarchical cluster analysis.

This measure is more suitable for decision about the number of clusters K because it takes the value 0 when the index R^* equals its expected value for each $k, k = 2, 3, \dots, K$.

7 Simulation studies

Hierarchical clustering gives a single unique solution (hierarchy). This is in opposition to some iterative method like k-means. Here the data matrix \mathbf{X} under investigation consists of 225 observations (countries) and the two variables birth rate and death rate in 1999. These variables are part of the population statistics that is published by CIA World Factbook (1999). Instead of absolute values their ranks are used here.

Figure 1 shows the (unique) result of hierarchical clustering by *Ward's* minimum variance method. The stability of this result was investigated by random sampling as proposed by (6) with $p = 1/2$ (i.e. random weighting

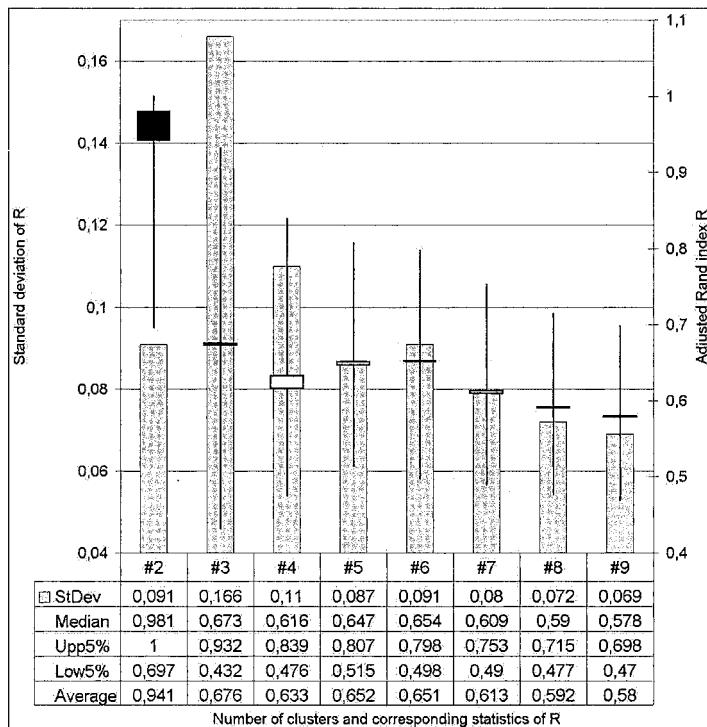


Fig. 2. Summary of simulation results of clustering by *Ward's* method.

of observations). In doing so, 200 such replicates were clustered by *Ward's* method. Herein each supplementary observation is assigned after clustering into each partition of the hierarchy according to the three nearest neighbours rule. Figure 2 shows some statistics of comparing the unique result with each of the other 200 clustering results by (7). 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-whisker-plots are assigned to other statistics of R (Median, Average, upper and lower 5 percent quantile). The median of R for $K = 2$ is nearby to its theoretical maximum value 1. That means, the two cluster solution is a stable one. It can be confirmed in a high degree for almost all samples. For more than two clusters the median (or average) of the adjusted Rand values becomes much lower. Therefore the number of cluster $K = 2$ is most likely. Cluster 1 has 58 countries and is located at the upper right hand corner of Figure 1. Cluster 2 consists of 167 countries. Moreover, the corresponding 200 (2×2)-contingency tables can be summarized by maximum correspondence to this unique partition into 2 clusters. As a result cluster 1 seems to be more stable than cluster 2 because it counts only a sum of 173 discrepancies in comparison to 199 discrepancies of cluster 2.

8 Conclusions

This paper is meant as an overview over some interesting features of Clus-Corr98. Especially the feature of the in-built validation of results of clustering is emphasized here. The presented automatic validation technique is a most general one that can assess the results of any hierarchical clustering. It is based on the use of special weights that lead to well-known resampling techniques. However there are some open questions like the dependence of the results of validation on the number p of active objects in relation to all I objects. Obviously there is also a dependence on the classification method that is used for the supplementary objects.

References

- BANFIELD, J.D. and RAFTERY, A.E. (1993): Model-Based Gaussian and non-Gaussian Clustering. *Biometrics*, 49, 803–821.
- CIA World Factbook (1999): Population by Country. <http://www.geographic.org>.
- FRALEY, C. (1996): Algorithms for model-based Gaussian Hierarchical Clustering. *Technical Report*, 311. Department of Statistics, University of Washington, Seattle.
- GORDON, A.D. (1999): *Classification*. Chapman & Hall/CRC, London.
- GOWER, J.C. (1971): A General Coefficient of Similarity and some of its Properties. *Biometrics*, 27, 857–874.
- HUBERT, L.J. 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.
- 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., BARTEL, H.-G., and DOLATA, J. (2002a): Exploring Roman Brick and Tile by Cluster Analysis with Validation of Results. In: W. Gaul and G. Ritter (Eds.): *Classification, Automation, and New Media*. Springer, Heidelberg, 471–478.
- MUCHA, H.-J., BARTEL, H.-G., and DOLATA, J. (2003): Core-based Clustering Techniques. In: M. Schader, W. Gaul, and M. Vichi (Eds.): *Between Data Science and Applied Data Analysis*. Springer, Berlin, 74–82.
- MUCHA, H.-J., SIMON, U., and BRÜGGEMANN, R. (2002b): 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.
- WARD, J.H. (1963): Hierarchical Grouping Methods to Optimise an Objective Function. *JASA*, 58, 235–244.

Stratification Before Discriminant Analysis: A Must?

Jean-Paul Rasson, Jean-Yves Pirçon, and François Roland

Department of Mathematics, University of Namur, 8, Rempart de la Vierge,
B-5000 Namur, Belgium

Abstract. It could be said as a tautology that, if we want to make a discriminant analysis between two or more populations and if we are able to divide these populations and training sets into some homogeneous subsets, it will be more efficient to make it on each of these subsets and then to combine the results. This can be done using one or two variables highly correlated with the one we want to predict. Our point of view will be a bit different: we will use a classification tree on all the available variables. We will first recall the first attempt (presented at IFCS2002 in Krakow). This one allowed us to obtain on an example of prediction of failure of the enterprises a gain of 5% of well classified data, using, after and before stratification, the classical Fisher's linear discriminant rule or the logistic regression. We intend to present a new method, still a classification tree, but with a multivariate criterion and in an agglomerative way. We compare both methods. In the same conditions and with the same data set, the gain is as high as 20%! Results will obviously also be presented when the methods are applied to test sets. Finally, we will conclude.

1 Method 1: Clustering tree

Generally, the interpretation of results for clustering methods is difficult. There are however some criteria where interpretation is easy: for example, discrimination trees in discriminant analysis. Discrimination trees (Breiman et al. (1984)) have the great advantage of being easily interpretable since only one variable is selected at each stage. It is for this principal reason that we wanted to apply this method to clustering.

Within the framework of clustering, the assignment of a class to each leave has no sense. On the other hand, the general principle of discrimination trees is preserved. Note that the first monothetic methods for clustering appeared in 1959 (Williams and Lambert (1959)) and recent works can be viewed in Chavent (1997). For our first method, the original contribution lies in the way of cutting a node. Indeed, the cut is based on the assumption that points distributions can be modelled by nonhomogeneous Poisson processes on the unknown domains (D_1, D_2) whose intensity is estimated either by the kernel method or by histograms. The cut is made in order to find the domains D_1 and D_2 which maximize the likelihood function: $F_{D_1, D_2}(\underline{x}) = \frac{1}{(\rho(D_1) + \rho(D_2))^n} \cdot \prod_{i=1}^n \mathbb{I}_{D_1 \cup D_2}(x_i) \cdot q(x_i)$ where $\underline{x} = (x_1, x_2, \dots, x_n)$ with $x_i \in$

$\mathbb{R}^d, i = 1, \dots, n$ are the observations, $q(\cdot)$ is the estimated intensity of the process, $\rho(D_j) = \int_{D_j} q(x) dx$ ($j = 1, 2$) is the integrated intensity and $\mathbb{I}_{\bullet}(\cdot)$ is the indicating function.

Consequently, the solution of the maximum likelihood corresponds to 2 disjoint convex domains D_1 and D_2 containing all the points and for which the sum of their integrated intensities is minimal ; ie. which maximize the integrated intensity of the gap between the two sets D_1 and D_2 .

As for classification, the best model is not necessarily the one obtained after the construction of the tree. Indeed, we obtain an “*overfitted*” tree. Consequently, once the tree is built, we seek to simplify its structure in order to reduce this problem, using a pruning method. For clustering, we have also found such method.

At this stage of research, clustering trees can treat only quantitative variables since our cut criterion is not yet adapted for other types of variables. Moreover, no missing data are accepted.

1.1 Intensity estimation by kernels

Kernels are in fact a generalization of histograms (Daudin et al. (1988)). This estimator, called **kernel estimator**, is a sum of bumps centered on the observations and is defined by $\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$, where h , called **smoothing parameter**, is the width of the window and K is the kernel function which satisfies: $\int_{-\infty}^{+\infty} K(x)dx = 1$, K symmetric and continuous.

Silverman (1986, page 137) shows which interests could bring the research of modes for clustering. He distinguishes concepts of bumps and modes by the following definitions: a **mode** in a density f is a local maximum, while a **bump** is characterized by an interval $[a, b]$ in such a way that the density f is concave on this interval but not on a larger one.

As Bock (1989) announces it, majority of usual densities don't have multiple bumps and therefore, the presence of more than one bump in a density indicates a superposition of densities. Moreover, Silverman (1986, page 139) tells that for the density estimation by the kernel method, we can expect, for very great values of h , that the estimation \hat{f} of the density is unimodal. On the other hand, as h decreases, the number of modes increases. This behavior is described mathematically as “*the number of modes is a decreasing function of the width of the window h* ”. This is guaranteed only for certain kernels, such the normal kernel. Consequently, during the intensity estimation for the nonhomogeneous Poisson process, we use the kernel method with the normal kernel. Moreover, it is the kernel generally used for quantitative variables. It is defined by $K_N(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$.

Since we use the normal kernel, there is a breaking value h_{crit} of the smoothing parameter for which the estimation changes from unimodality to multimodality (Silverman (1986, page 140)). Generally, there is an interval

of possible values h . We choose the greatest value of h for which the number of modes is the smallest and strictly larger than 1. This value corresponds to the value of transition from multimodality to unimodality. This property is especially important for the pruning and is used by Silverman for testing the unimodality.

1.2 Cut criterion

When we cut a node, we suppose to have 2 convex disjoint groups whose observed points are generated by a nonhomogeneous Poisson process. We try to determine the optimal cut, using the rule of the maximum likelihood.

Our cut criterion is thus the following : at each node, we start by considering the intensity of the points that lie in the node, either by the histograms, or by the kernels (see 1.1). The method of density estimation is not very important. The estimation by wavelets has also been tested and the results are similar for the cutting. Once the intensity is estimated, we separate the initial domain into 2 convex disjoint domains for which the likelihood function is maximum; i.e. for which the integrated surface in the gap is the biggest one. Proceeding in this way, variable by variable, we are able to select the variable which generates the greatest likelihood function integrated onto the maximal gap.

Thanks to this cut criterion, we have a divisive hierarchical method of clustering.

1.3 Pruning

The pruning procedure makes again the assumption that the data follow a nonhomogeneous Poisson process. It is based on a hypothesis test called **Gap Test** which tests the hypothesis H_0 : “*the $n = n_1 + n_2$ points are realizations of the Poisson process over the domain D ; D unknown*”, against the hypothesis H_1 : “ *n_1 points are realizations of the Poisson process over the domain D_1 and n_2 points are realizations over the domain D_2 with $D_1 \cap D_2 = \emptyset$; D_1, D_2 unknowns*”. In other words, the null assumption involves one class (*bad cut*) whereas the alternative assumption gives two different classes (*good cut*).

The pruning method built here crosses the tree branch by branch from its root to its end in order to index the good cuts (Gap Test satisfied) and the bad cuts (Gap Test non-satisfied). The ends of the branches for which there are only bad cuts are pruned.

Given that the Gap test is valid for data following a homogeneous Poisson process, we make a change of variables of the type $\tau(t) = \int_{x_1}^t q(x)dx$ where x_1 is the smallest point of the realization of the nonhomogeneous Poisson process and $q(x)$ is the intensity of the Poisson process. The realization of the nonhomogeneous Poisson process (with the order statistics (x_1, \dots, x_n)) is so transformed in a realization of a homogeneous Poisson process $((\tau_1, \dots, \tau_n)$

where $\tau_i = \tau(x_i)$). The integrated intensity of the largest gap between the domains D_1 and D_2 (found in section 1.2) is noted by N_1 .

Under the null assumption H_0 , the random variable $\frac{nN_1}{\tau_n} - \ln n$ has a distribution identical to the distribution of the extreme value (of density function $f(x) = e^{-x} e^{-e^{-x}}$) when n tends towards infinity. Consequently, the critical area is $\{x | m' = \frac{nN_1}{\tau_n} - \ln N \geq K_\alpha = -\ln(-\ln(1-\alpha))\}$. α is the level of the test.

For the pruning, we always estimate the intensity by kernels. Indeed, Silverman (1986) shows that kernels have special properties interesting for the Gap test. The property of modalities (as explained in section 1.1) is important to see the formation of two groups.

2 Method 2: An agglomerative classification method

As first attempt, we want to take a criterion which is the dual of the first method. In other terms, this second method is an agglomerative method with non parametric density estimation. The principle is the following: first, we measure the area sustended by the density between 2 points (or groups of points) on each axe. Then, we gather 2 points (or groups) which are the closest in this sense on one axe. The danger is to gather 2 points (or groups) which are obviously in different groups.

Naturally, if we really work with a model in dimension d , the real criterion for the divisive method, e.g. between two convex clusters (maximum likelihood criterion) is : find the two clusters for which the difference of the hypervolumes sustended by the density between the global convex hulls of the two clusters is the biggest. For the agglomerative method, this difference should be the smallest. This causes computational problems.

But, if all the sustended areas (on each axis) between the respective co-ordinates of the two points are small, then the hypervolume in dimension d is small. This implication is not reversible.

Thus for 2 points $x_i = (x_{i1}, \dots, x_{id})$ and $x_j = (x_{j1}, \dots, x_{jd})$, we define our dissimilarity as $diss(x_i, x_j) = \max_{1 \leq k \leq d} \left| \int_{x_{ik}}^{x_{jk}} \hat{f}_k(x) dx \right|$ where $\hat{f}_k(\cdot)$ is an estimation of the density function for the variable k .

We gather x_s and x_t ($1 \leq s \leq t \leq n$) if $diss(x_s, x_t) = \min_{1 \leq s \leq t \leq n} diss(x_i, x_j)$.

In order to estimate the density, we use the normal kernel : $\hat{f}_k(x) = \frac{1}{nh} \sum_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{|x - x_i|}{h}}$, where $h = 1,06 \cdot \min(\sigma, \frac{R}{1,34}) \cdot n^{-0,2}$, (Silverman (1986, page 48)) (σ is the standard deviation and R the interquartile range).

Once the first two points are gathered, they form a class C and we have to update the dissimilarity matrix.

To do this, we choose these two possibilities:

- the single link method where $diss(C, x_i) = \min_{x \in C} diss(x, x_i)$, $x_i \notin C$,
- the complete link method where $d(C, x_i) = \max_{x \in C} d(x, x_i)$, $x_i \notin C$.

Then we merge the two “closest” classes and so on until we reach the number of classes (maybe 1).

2.1 Discriminant criterion associated

When other points have to be attributed to one of the classes (e.g. points of a test set), we use the following rule:

Suppose we have n points x_i ($1 \leq i \leq n$) divided into m clusters C_j ($1 \leq j \leq m$) and we want to affect a new point x to one of the classes, we use $diss(C, x) = \min_{x_i \in C} diss(x_i, x)$ and x is affected to C_s if $diss(C_s, x) = \min_{1 \leq j \leq m} diss(C_j, x)$.

3 Comparison of the two methods

As we said previously, the first method is based on a descending hierarchy whereas the second method is based on an ascending hierarchy. It influences the execution time of the two methods as well as the storage cost.

The ascending algorithm is faster than the descending one. If at one stage, the partition comprises K clusters, the descending algorithm requires the comparison of the $2^{n-1} - 1$ possible divisions of the n points into two clusters whereas the ascending algorithm only requires the comparison of $\frac{n(n-1)}{2}$ fusions. The fact that the first method is monothetic implies that the cost is a bit less than $2^{n-1} - 1$ but still greater than the second method.

The first algorithm also estimates the density at each level of the tree : we estimate the densities for each node whereas the second algorithm estimates the densities only one time at the beginning.

On the other hand, with the second method, we have to compute and to store a dissimilarity matrix. Even if we only store the lower triangular part and even if we update the dimension of the matrix at each step, the storage cost is at least $\frac{n^2}{2}$. For the two methods, we have also a storage cost due to density estimation and the building of the tree.

A second point on which the two methods are different lies in the fact that the first method is monothetic and the second polythetic. This means that the first method works variable by variable whereas the second one treats all the variables at a time. This implies both advantages and disagreements. The first method has the merit to provide very easily interpretable results in terms of the starting variables. It also automatically operates the variables selection. Unfortunately, that implies that the cuts are carried out only perpendicularly to the axes. That does not occur with the second method : the clusters can

lay in space in any orientation. That has however a price to pay and that is made with the detriment of the interpretation of the results.

We can also associate a discriminant criterion with the two methods. As we obtain a tree with the first method, we can throw new points in it and observe in which leaf they are assigned, following the successive cuts. We present the discriminant criterion associated with the second method in the section 2.

4 Some examples

Three examples are presented. The first example consists in the results obtained with the two methods on the well-known Ruspini data set. Then, we present two examples with financial data coming from the Public Federal Office and the SES.

First example : Ruspini data set (Fig.1, Fig.2). For both methods, the good classes are found. Indeed, for the first method (Fig.1), the discontinued lines are pruned cuts. Therefore, the classes are determined by the continued lines. For the second method (Fig.2), the found classes are detectable by the different symbols.

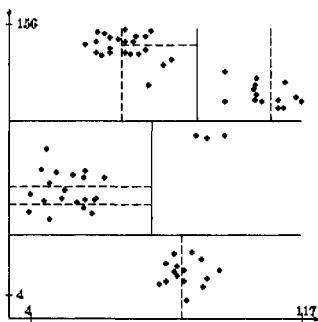


Fig. 1. Clusters obtained on Ruspini data (Method 1, Kernel)

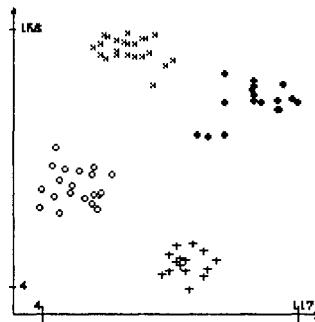


Fig. 2. Clusters obtained on Ruspini data (Method 2, Complete Link)

Second example : Financial data. The objective is the following: starting from financial data of the companies and variables such as the size and the branch of industry, we try to find homogeneous classes and to characterize these groups starting from selected variables according to their order of financial importance. Traditional methods don't allow to find coherent groups. The seven used variables are: credit, sales turnover, employment, added value, requirement in working capital, go back to constitution, and remuneration and social contributions. The set contains 945 firms.

After the construction of homogeneous groups, we make a Fisher's discriminant analysis with the size variable in accordance with the European standards. There are four possible sizes (TPE, PE, ME and GE¹). We compare results when we apply Fisher's discriminant analysis on the whole data set and when we apply it on each formed groups and then sum the results. To apply Fisher's discriminant analysis, we must have only two classes. Thus, on one hand, class 1 consists of TPE, PE and ME (662 enterprises) and class 2 consists of GE (242 enterprises). On the other hand, class 1 consists of PE, ME and GE (432 enterprises) and class 2 consists of TPE (472 enterprises). The results are summarised in the Table Tab.1.

Example 1			Example 2		
Class 1 : TPE, PE and ME			Class 1 : PE, ME and GE		
Class 2 : GE			Class 2 : TPE		
Fisher on the whole data set			Fisher on the whole data set		
	Class 1	Class 2		Class 1	Class 2
Class 1	610	52	Class 1	250	182
Class 2	81	161	Class 2	36	436
85.28 % in the good class			75.88 % in the good class		
Fisher on groups formed by method 1 (kernel) + sum of results			Fisher on groups formed by method 1 (kernel) + sum of results		
	Class 1	Class 2		Class 1	Class 2
Class 1	655	7	Class 1	389	43
Class 2	55	187	Class 2	0	472
93.14 % in the good class			95.24 % in the good class		
Fisher on groups formed by method 2 (complete link) + sum of results			Fisher on groups formed by method 2 (complete link) + sum of results		
	Class 1	Class 2		Class 1	Class 2
Class 1	650	12	Class 1	427	5
Class 2	87	155	Class 2	0	472
89.05 % in the good class			99.67 % in the good class		

Table 1. Results of Fisher discriminant analysis on the whole data set, on groups formed by method 1 and on groups formed by method 2. The results are for both possibilities of constituted classes.

We can see that groups formed by both methods constitute a good stratification of the data set. Indeed, the gain in good classification is up to 20 %.

Third example: Failed Enterprises. This example is similar to the preceding one but more interesting because we have a test set. The goal is to find the failed enterprises. As for the preceding example, we cluster data by our methods and we apply then the fisher's discrimination to all formed groups. In the table Tab 2, we can see that both methods are stable and tend to have a classification rate equal for the two classes.

¹ very small enterprises, small enterprises, middle size enterprises and big enterprises

Training set			Test set		
2000 firms Failed firms is class 2			721 firms Failed firms is class 2		
Fisher on the whole data set			Fisher on the whole data set		
	Class 1	Class 2		Class 1	Class 2
Class 1	822	404	Class 1	434	188
Class 2	380	384	Class 2	52	47
60.80 % in the good class			66.71 % in the good class		
Fisher on groups formed by method 1 (kernel) + sum of results			Fisher on groups formed by method 1 (kernel) + sum of results		
	Class 1	Class 2		Class 1	Class 2
Class 1	866	360	Class 1	429	193
Class 2	231	543	Class 2	45	54
70.45 % in the good class			66.99 % in the good class		
Fisher on groups formed by method 2 (complete link) + sum of results			Fisher on groups formed by method 2 (complete link) + sum of results		
	Class 1	Class 2		Class 1	Class 2
Class 1	995	231	Class 1	448	174
Class 2	175	599	Class 2	42	57
79.70 % in the good class			70.04 % in the good class		

Table 2. Results of Fisher discriminant analysis on the whole data set, on groups formed by method 1 and on groups formed by method 2. The results are for detection of failed enterprises.

5 Conclusion

Two different methods are proposed. They are tested on different examples. The first example shows that methods operate correctly. Indeed, they find the homogeneous classes. The other examples show that it is interesting to make a stratification before making a discriminant analysis. Indeed, the percentages of good classification increase when the discriminant analysis is practised on the different groups found by the clustering.

References

- BOCK, H.H. (1989): Probabilistic Aspects in Cluster Analysis. *Conceptual and Numerical Analysis of Data*, 12-44.
- BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., and STONE, C.J. (1984): *Classification and Regression Trees*. Belmont, Wadsworth.
- CHAVENT, M. (1997): *Analyse des données symboliques. Une méthode divisive de classification*. PhD thesis. Universit de Paris IX, Dauphine.
- DAUDIN, J-J., MASSON, J-P., TOMASSONE, R., and DANZART, M. (1988): *Discrimination et classement*. Masson, Paris.
- SILVERMAN, B.W. (1986): *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.
- WILLIAMS, W.T. and LAMBERT, J.M. (1959): Multivariate Method in Plant Ecology. *Journal of Ecology*, 47, 83-101.

An Exchange Algorithm for Two-Mode Cluster Analysis

Manfred Schwaiger¹ and Raimund Rix²

¹ Institute for Corporate Development and Organization, Munich School of Management, Ludwig-Maximilians-University, D-80539 Munich, Germany

² Accenture GmbH, Maximilianstr. 35, D-80539 Munich, Germany

Abstract. A comprehensive simulation study recently has shown that, in order to identify best two-mode classifications, the user may apply different algorithms and select the result yielding the lowest squared centroid distance measurement (SCD). Knowing the outperformer among several goodness-of-fit measures creates the premises to develop an exchange algorithm for two-mode classifications. This paper presents the algorithm and discusses significance in gain of precision based on a large Monte Carlo Simulation study.

1 On the absence of exchange algorithms in two-mode cluster analysis

Within two-mode cluster analysis row and column elements of a data matrix are clustered simultaneously. As input, a data matrix $\mathbf{X} = (x_{ij})_{(n \times m)}$ is used, which is created by assigning numerical values x_{ij} to the elements $O_i \in O$, $A_j \in A$ of the cartesian product $O \times A$ ($O = \{O_1, \dots, O_n\}$, $A = \{A_1, \dots, A_m\}$). Applications in economic and social sciences are described in DeSarbo (1982), De Soete (1984), Both and Gaul (1986), Eckes (1993 and 1995), a comprehensive survey on two-mode cluster analysis and its applications is given by Schwaiger (1997) and Rix (2003).

When using hierarchical two-mode procedures we may encounter the same problems that are well described in the one-mode case: Once joined, elements have to remain in the same cluster even if during the fusion process assigning them to another cluster would increase a given goodness-of-fit criterion. Therefore, in one-mode cluster analysis exchange algorithms like CLUDIA or KMEANS (just to name the most renowned ones) were developed (Opitz 1980, p. 87ff.). These procedures can be described as follows: Starting with an original classification $\mathcal{K}^0 = \{C_1^0, \dots, C_k^0\}$ showing k clusters we consider a special object $O_i \in O$. The original cluster containing O_i is denoted by \hat{C}_t^0 . In case \hat{C}_t^0 contains more than one object, all remaining clusters C_l^0 are evaluated by checking whether moving O_i from \hat{C}_t^0 to C_l^0 ($l \neq t$) would decrease the goodness-of-fit index b and thus improve the classification. In case there are more exchange possibilities available fulfilling this condition one selects that move that yields the largest decrease in b .

Formally, depending on the classification index b the optimization problem is specified by (Hartung and Elpelt (1986), p. 466):

$$\min_{O_i \in \mathcal{O}} \min_{C_i^0 \in \mathcal{K}^0} b(\mathcal{K}) : \mathcal{K} = \{C_1, \dots, C_k\},$$

$$\text{where } C_\mu = \begin{cases} C_\mu^0 - O_i: & C_\mu^0 = \hat{C}_t^0, \quad |\hat{C}_t^0| > 1, \quad C_t^0 \neq \hat{C}_t^0 \\ C_\mu^0 \cup O_i: & C_\mu^0 = C_t^0, \quad |\hat{C}_t^0| > 1, \quad C_t^0 \neq \hat{C}_t^0 \\ C_\mu^0 & : \text{else} \end{cases} \quad \mu = 1, \dots, k.$$

If the resulting criterion value $b(\mathcal{K})$ is lower than $b(\mathcal{K}^0)$, we accept the new classification \mathcal{K}^1 . Analogously one proceeds with \mathcal{K}^1 as (new) original classification and calculates \mathcal{K}^2 . One obtains a finite sequence of classifications $\mathcal{K}^0, \mathcal{K}^1, \mathcal{K}^2, \dots$ with monotonously decreasing goodness-of-fit indices $b(\mathcal{K}^0) > b(\mathcal{K}^1) > b(\mathcal{K}^2) > \dots$. The procedure is terminated in loop ν , if for the first time $b(\mathcal{K}^\nu) = b(\mathcal{K}^{\nu+1})$. The result $b(\mathcal{K}^\nu)$ ist locally optimal (Opitz (1980), p. 88).

Up to now there was, unlike in one-mode classification, no consensus about which goodness-of-fit measure b to use, so there has been no basis for developing an exchange algorithm yet.

2 An exchange algorithm to improve SCD

Rix (2003) has performed a large Monte Carlo simulation study in order to evaluate two-mode hierarchical algorithms and corresponding goodness-of-fit measures. By using the Adjusted Rand Index (ARI) as an outer goodness-of-fit criterion he shows that using the squared centroid distance (SCD) for b yields best two-mode classification results, i.e. these results that are most similar to the given artificial classification he uses to build the simulation.

The squared centroid distance coefficient was developed by Schwaiger (1997, p. 122) and improved by Unterreitmeier and Schwaiger (2002). Its application is based on the assumption that the user considers specific clusters as homogenous, i.e. the elements within a cluster are considered to be equivalent at least to some extent. In the one-mode case the centroid is defined as "central point" of the objects of a cluster in the attribute space:

$$\bar{x}_r := (\bar{x}_{1r}, \dots, \bar{x}_{mr}) \quad \text{with} \quad \bar{x}_{jr} = \frac{1}{n_r} \sum_{O_{i'} \in C_r} x_{i'j}. \quad (1)$$

In the two-mode case we have to distinguish two types of centroids. The *object specific* centroid of a two-mode cluster Clusters C_r can be described as above, whereas the *attribute specific* centroid of a cluster C_r is corresponding to a "central point" of attributes in the space of objects:

$$\bar{x}_{jr}^O = \frac{1}{n_r} \sum_{O_{i'} \in C_r} x_{i'j}, \quad \bar{x}_{ir}^A = \frac{1}{m_r} \sum_{A_{j'} \in C_r} x_{ij'}. \quad (2)$$

Looking at a classification result the user may intuitively presume that, within a certain two-mode cluster, all objects can be substituted by the object-specific centroid and all attributes by the attribute-specific centroid. To quantify the error resulting from this simplification, SCD compares x_{ij} to \bar{x}_{jr}^O with respect to the objects in cluster C_r and x_{ij} to \bar{x}_{ir}^A with respect to the attributes.

Formally, we calculate

$$\xi_{ij}^O = \bar{x}_{jr}^O, \quad O_i \in C_r \quad \text{and} \quad \xi_{ij}^A = \bar{x}_{ir}^A, \quad A_j \in C_r. \quad (3)$$

The auxiliary SCD coefficient is given by

$$\widehat{\text{SCD}} = \sum_{i=1}^n \sum_{j=1}^m (\xi_{ij}^O - x_{ij})^2 + \sum_{i=1}^n \sum_{j=1}^m (\xi_{ij}^A - x_{ij})^2.$$

We normalize then $\widehat{\text{SCD}}$ by dividing it by the obtainable maximum, that will be achieved if all objects and attributes form one two-mode cluster. The corresponding formulas are given by:

$$\xi_j^O = \frac{1}{n} \sum_{i'=1}^n x_{i'j}, \quad \xi_i^A = \frac{1}{m} \sum_{j'=1}^m x_{ij'}, \quad (4)$$

$$\widehat{\text{SCD}}_{\max} = \sum_{i=1}^n \sum_{j=1}^m (\xi_j^O - x_{ij})^2 + \sum_{i=1}^n \sum_{j=1}^m (\xi_i^A - x_{ij})^2 \quad \text{and} \quad (5)$$

$$\text{SCD} = \frac{\widehat{\text{SCD}}}{\widehat{\text{SCD}}_{\max}} \in [0, 1]. \quad (6)$$

Smaller SCD values indicate more homogenous classifications.

Having built a goal function that is improving (minimizing) SCD, we can now design an exchange algorithm. Its flow chart is given in Figure 1.

The two-mode exchange algorithm is very similar to one-mode procedures: it checks for all objects and for all attributes in turn, which relocation yields maximum decrease of SCD. This exchange is processed, the resulting classification becomes the starting point for a new loop. The iteration process is terminated if an improvement of SCD is no longer possible.

3 Simulation

3.1 Simulation set-up

To evaluate the performance of the two-mode hierarchical cluster analysis algorithms and the goodness-of-fit measures we performed Monte Carlo simulations covering a wide spectrum of data constellations. The simulation program builds artificial classifications and measures, to which degree the algorithms were able to recover the original classifications.

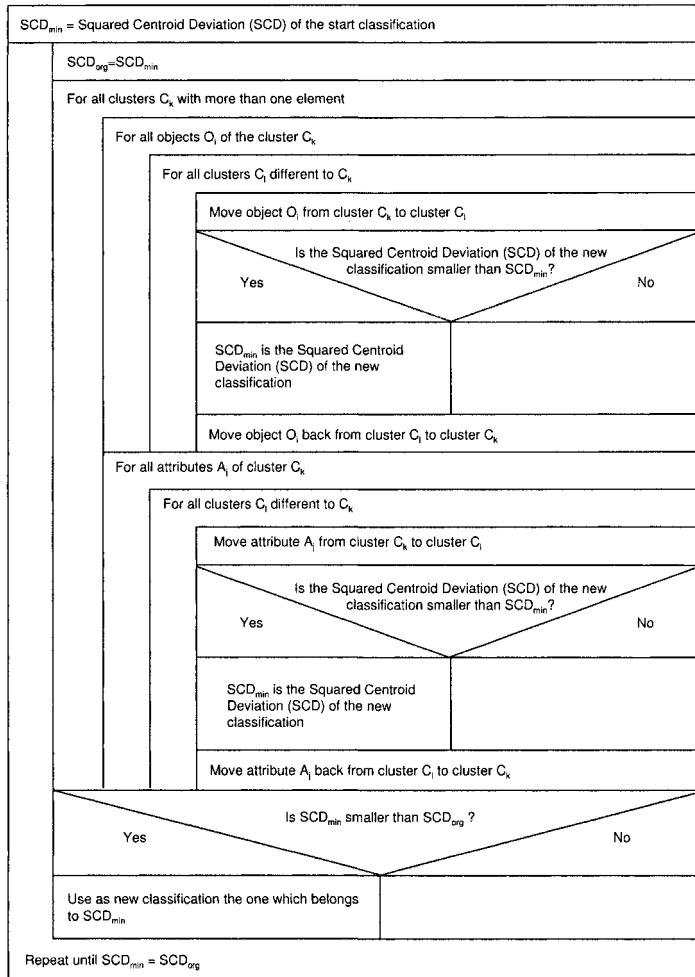


Fig. 1. Flow chart of the exchange algorithm.

To test the effects of an additional application of the exchange algorithm we carried out simulations using artificial data matrices. The creation of these data bases is thoroughly described in Rix (2003, p. 75ff.). Due to space restrictions we can only give a brief overview on the most important settings. In order to create "ideal" two-mode classifications (whose reproduction by different hierarchical algorithms was then evaluated) the following requirements had to be fulfilled:

- the objects O_i of a cluster C_k should be similar to each other, so the values x_{ij} of an attribute A_j must be similar for all objects $O_i \in C_k$.

- the attributes A_j of a cluster C_l should be similar to each other as well, so every attribute $A_j \in C_l$ has similar values x_{ij} for one object O_i .
- all objects of a cluster have to be strongly associated to the attributes of this cluster.

Similarity of values was set by specifying a Gaussian distribution around the mean value with a standard deviation tied to the width of the cluster, i.e. the bandwidth of attribute values for all objects assigned to a specific cluster. Moreover, we met the condition, that these values must not exceed the cluster width.

In a second step, perturbations were applied to get a more realistic classification and data matrix:

- the number of objects and the number of attributes are set randomly according to a Gaussian distribution around an average cluster size with a standard deviation, which is proportional to the average cluster size
- besides two-mode clusters one-mode clusters (containing objects or attributes only) were introduced
- errors in attribute values that may occur due to a lack of reliability or data entry mistakes were considered within the simulation through an injection of random values into the data matrix
- solitary objects and solitary attributes were introduced.

For the evaluation of the result of an algorithm, the Adjusted Rand Index by Hubert and Arabie (1985, p. 198) is used as an external goodness-of-recovery measure. It indicates how similar the results of an algorithm are to the original classification. With

N Number of clusters,

ν_{kl} Number of elements, which are simultaneously part of cluster C_k of the original classification and of cluster \tilde{C}_l of the result of the algorithm,

$\nu_{k\cdot}$ Number of elements of cluster C_k of the original classification,

$\nu_{\cdot l}$ Number of elements of cluster \tilde{C}_l of the result of the algorithm and

ν Total number of elements,

the Adjusted Rand Index if defined as (Hubert/Arabie 1985, p. 198)

$$\text{ARI} = \frac{\sum_{k=1}^N \sum_{l=1}^N \binom{\nu_{kl}}{2} - \sum_{k=1}^N \binom{\nu_{k\cdot}}{2} \sum_{l=1}^N \binom{\nu_{\cdot l}}{2} / \binom{\nu}{2}}{\frac{1}{2} \left(\sum_{k=1}^N \binom{\nu_{k\cdot}}{2} + \sum_{l=1}^N \binom{\nu_{\cdot l}}{2} \right) - \sum_{k=1}^N \binom{\nu_{k\cdot}}{2} \sum_{l=1}^N \binom{\nu_{\cdot l}}{2} / \binom{\nu}{2}}. \quad (7)$$

The Adjusted Rand Index has the value 1 for a perfect recovery of the original classification and an average value 0 for random classifications which are independent from the parameters of the ideal classification. Therefore, it is possible to compare the ARI values of results with different parameters of the ideal classification.

For the measurement of the performance of the algorithms, the average goodness-of-recovery value of 10 000 simulation cycles is used.

3.2 Results

To identify the potential of the exchange algorithm we processed 10 000 simulations as described, where all possible perturbations were set at random presence. Hence, we allowed for every possible interaction and provided conditions most close to real applications. We applied different two-mode hierarchical algorithms, which are described in Schwaiger (1997, p. 102ff.). The reader may find a comprehensive list of references there, too. We use the following abbreviations in Figure 2: Centroid Effect Method (CEM), Missing Value Single Linkage (MV SL), Average Linkage (MV AL) and Complete Linkage (MV CL) algorithms, and ESOCLUS Single, Average and Complete Linkage (ESOCLUS *L) using block-specific maxima and block specific medians (indicated by a P) to weight distances in the grand matrix. The results of the exchange algorithm are denoted by "relocate". The goodness-of-fit measures used are Cophenetic Correlation Coefficient (CCC), Theil's Inequality Coefficient (TIC) and Rix-Schwaiger-Coefficient (RSC). A thorough description of these measures is given by Rix (2003, p. 57ff.). We took that classification as starting point for the exchange algorithm that showed the lowest SCD value. The grey bars in Figure 2 show the average outer goodness if we had used other inner goodness-of-fit measures to identify the best result.

The classification resulting after termination of the exchange algorithm was compared to the artificial classification that served as data base for the corresponding simulation run by means of the Adjusted Rand Index. Thus we could calculate the outer goodness-of-fit of the exchange solution, and we used the arithmetic mean over 10 000 simulation runs to quantify the mean outer goodness-of-fit, that we compared in Figure 2 to the mean outer goodness-of-fit of the hierarchical procedures (without exchange algorithm).

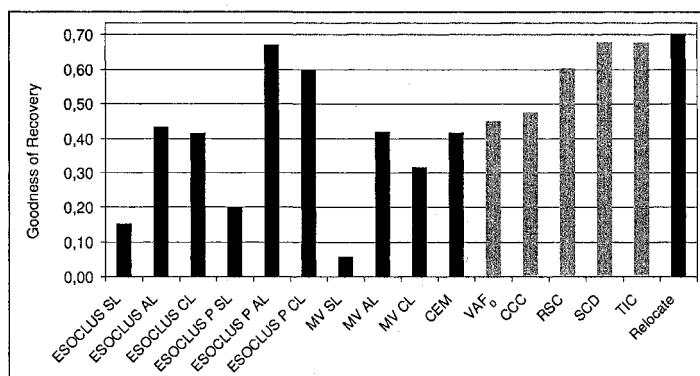


Fig. 2. Mean outer goodness-of-fit measures.

4 Summary and recommendations

In the large simulation study of Rix (2003) SCD was identified as premium goodness-of-fit measure for two-mode classification results. Using SCD allows to develop an exchange algorithm that minimizes the sum of squared distances between the elements of the two-mode clusters and the corresponding centroids. As we can see in Figure 2, the application of an exchange algorithm improves results of hierarchical two-mode algorithms in a considerable way. Hence, we recommend to use the described exchange algorithm successively to any two-mode clustering procedure, may it be hierarchical or non-hierarchical. Moreover, the number of elements (objects and attributes) that have to be relocated, can be considered as an indicator of the stability of the initial classification.

References

- BOTH, M. and GAUL, W. (1986): Ein Vergleich zweimodaler Clusteranalyseverfahren. *Methods of Operations Research*, 57, 593–605.
- DE SOETE, G. (1984): Ultrametric Tree Representations of Incomplete Dissimilarity Data. *Journal of Classification*, 1, 235–242.
- DESARBO, W.S. (1982): GENNCLUS: New Models for General Nonhierarchical Clustering Analysis. *Psychometrika*, 47, 449–475.
- ECKES, T. (1993): Multimodale Clusteranalyse: Konzepte, Modelle, Anwendungen. In: L. Montada (Ed.): *Bericht ueber den 38. Kongress der Deutschen Gesellschaft fuer Psychologie in Trier*, 2, Goettingen, 166–176.
- ECKES, T. (1995): Recent Developments in Multimode Clustering. In: W. Gaul and D. Pfeiffer (Eds.): *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis and Knowledge Organisation*. Berlin, Heidelberg, 151–158.
- HARTUNG, J. and ELPELT, B. (1986): *Multivariate Statistik – Lehr- und Handbuch der angewandten Statistik*. München, Wien.
- HUBERT, L. and ARABIE, P. (1985): Comparing Partitions. *Journal of Classification*, 2, 193–218.
- OPITZ, O. (1980): *Numerische Taxonomie*. Stuttgart, New York.
- RIX, R. (2003): *Zweimodale hierarchische Clusteranalyse*. Wiesbaden.
- SCHWAIGER, M. (1997): *Multivariate Werbewirkungskontrolle, Konzepte zur Auswertung von Werbetests*. Reihe neue betriebswirtschaftliche Forschung, 231, Wiesbaden.
- UNTERREITMEIER, A. and SCHWAIGER, M. (2002): Goodness of Fit Measures for Two-Mode Cluster Analyses. In: W. Gaul and G. Ritter (Eds.): *Classification, Automation, and New Media. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation*. Berlin, Heidelberg, 401–408.

Model-Based Cluster Analysis Applied to Flow Cytometry Data

Ute Simon¹, Hans-Joachim Mucha², and Rainer Brüggemann¹

¹ Leibniz-Institute of Freshwater Ecology and Inland Fisheries,
D-12587 Berlin, Germany

² Weierstraß-Institute of Applied Analysis and Stochastic,
D-10117 Berlin, Germany

Abstract. Flow cytometry is an appropriate technique for the investigation and monitoring of phytoplankton (algae), providing quick, semi-automatic single-cell analysis. However, to use flow cytometry in phytoplankton research routinely, an objective and automated i.e. computer-supported data analysis is demanded. For this reason, in a pilot study a sequence of different steps of cluster analysis has been developed, including model-based and hierarchical clustering, as well as the concept of cores and weighting of observations and parameters. A successful application of the method is demonstrated for a snapshot of a sample of Lake Müggelsee in Berlin (Germany).

1 Introduction

Flow cytometry provides quick, semi-automatic single-cell analysis, based on optical characteristics such as scattered light and fluorescence. As algal cells contain several auto-fluorescent pigments such as chlorophyll or carotenoides, cytometry turns out to be an appropriate technique for phytoplankton investigation and monitoring (e.g. Hofstraat et al. (1994)). Previous studies have shown, that the pigment-composition enables to discriminate algae on the level of taxonomic classes and sometimes even species (e.g. Chisolm et al. (1988)). Beyond this, ataxonomic approaches such as biomass-size-spectra based on flow cytometry data has been used for the ecological evaluation of the integrity of freshwater systems (Steinberg and Brüggemann (1998)). However, to investigate phytoplankton by flow cytometry, there are some challenges to meet. Each particle is characterised by up to eight optical parameters (variables) and about 30000 particles in one sample are not unusual. Thus, one has to face large matrices for data analysis and the objective and automated identification and quantification of algae-groups (classes) with similar optical characteristics is urgently necessary. Even though there are only few variables, the problem of clustering flow cytometry data from phytoplankton is not trivial, as the number of classes is not known, the size of the classes might differ extremely and the assumption of normality, especially concerning their symmetry, is violated obviously in some degree. There is a number of observations equal to or very near to the detection limits of the

Parameter	Parameter (short cut)	Excitation wavelength [λ]	Detected- emission [λ]	Description
Front scatter	FSC	633nm (red)	>610nm	Size of the cells
Side scatter	SSC	633nm (red)	>610nm	Structure of the cell-surface
Fluorescence 1	FL1	633nm (red)	>665nm	Pigment chlorophyll a
Fluorescence 3	FL3	532nm (green)	>665nm	Pigment chlorophyll a
Fluorescence 4	FL3	532nm (green)	575nm	Pigment phycoerithrin

Table 1. Configuration of the flow cytometer

variables surrounded by comparatively sparse regions. Thus, the densities are sometimes far from symmetry because of truncation.

Some working groups are using neural networks for data analysis (e.g. Boddy et al. (2000)). This is a very promising approach because of its data-mining facility. However, the practical application might be restricted by the need to train the net. As we tend to investigate and evaluate ecosystems of different ecological characteristics, such as trophic states, we prefer a method which requires as little a priori interpretation of the data as possible. To define classes of algal cells with similar optical characteristics qualitatively and quantitatively, we developed a sequence of different steps of cluster algorithms, including model-based and hierarchical clustering as well as the concept of cluster cores and weighting of observations and parameters respectively (Mucha et al. (2002)). In a pilot study an application of the method is demonstrated for a snapshot sample of Lake Müggelsee in Berlin (Germany). The comparison with microscopic counts verifies the result of our iterative scheme of the model-based Gaussian clustering.

2 Flow cytometry

Flow cytometry is a laser-optical technique, providing single cell analysis. For a schematic view of a flow cytometer see for example Mucha et al. (2002). The cells are passed, one by one, through the intersect of two lasers and signals of scattered light and fluorescence of each cell is measured and stored in a computer. In our example five optical parameters are detected, each in a differentiation of 1024 channels. In Tab. 1 the main characteristics are summarised.

3 Cluster-analysis

The requirements of analysing data from flow cytometry translated in a mathematical language can be formulated as follows: Find for the $(I \times J)$ -matrix the number of classes of similar optical characteristics and describe them by Gaussian functions (I : observations, here algal cells and other particles; J :

variables, here optical parameters). The most general model-based Gaussian clustering is when the covariance matrix \mathbf{W}_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

$$V_K = \sum_{k=1}^K n_k \log \left| \frac{\mathbf{W}_k}{n_k} \right|. \quad (1)$$

Herein $\mathbf{W}_k = \sum_{i \in C_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T$ is the sample cross-product matrix for the k -th cluster C_k , and $\bar{\mathbf{x}}_k$ is the usual maximum likelihood estimate of expectation values in cluster C_k . This criterion is obtained by taking advantage of the monotone log-function with regard to the multivariate normal densities that are used in the general classification maximum likelihood approach (Fraley and Raftery (2002)). Usually all observations have the same weight. The principle of weighting the observations is a key idea for handling cores (representatives) and outliers. In the case of outliers one has to downweight them in some way in order to reduce their influence. In the case of representatives of cores, one can weight them proportional to the cardinality of cores. Some details about model-based clustering using weighted observations are given by Mucha et al. (2002).

Since for the sample of Lake Müggelsee we do not know, how many classes are to be expected, and as these classes might differ in size and shape extremely, 4 different worksteps has been used sequently. They constitute the following, briefly described, scheme.

Step 1: Clustering of the original data with the K-means method (Macqueen (1967)). The aim is to differentiate between observations for further analysis (algae) and those observations which are of no interest, such as dead algal cells or detritus (any kind of dead organic matter).

To obtain a better match of the data to the Gaussian model, the ratio of FL1/FL3 is calculated and used as a new variable (FL1/3), replacing the original variables of FL1 and FL3. Thus for the subsequent analysis only four variables, the FSC, SSC, FL1/3 and FL4 are to be used.

Step 2: Hierarchical clustering with the Ward-method (Ward (1963)). The aim is to get a good initial solution for a more complex Gaussian model used in the following and to find an appropriate number of clusters for further analysis. The hierarchical clustering was performed by a randomly drawn subset of all observations under investigation.

Step 3: Application of the determinant criterion. As the initial solution the result of the hierarchical cluster analysis (2. step) is used. By doing so, we obtain a certain number of well separated clusters, and as in the case of flow cytometry data expected, an additional cluster with a quite flat density. This cluster collects all observations from sparse regions, which cannot be neglected or handled as an own cluster for ecological reasons. To identify outliers we used a simulation with the K-mean method again.

Step 4: Final model based Gaussian clustering with weighted observations. To avoid a strong influence of the outliers on the final result of the clustering with a determinant criterion, they are downweighted. Since distance values are still available, the outliers are added to the nearby cluster.

For a more detailed description of the worksteps described above see Mucha et al. (2002).

4 Results

The sample of water of Lake Müggelsee contains 21778 observations in total. The K-means clustering (step 1) was carried out with a number of nine initial clusters. By biological expert knowledge, four of the nine clusters are identified as non-interesting observations, and are removed from the matrix. Thus, for further analysis only 8786 observations remain. To find an appropriate number of clusters for further analysis and to get a good initial solution for a more complex Gaussian model, in the following steps a simulation study of hierarchical clustering has been used. By optical inspection of the sum of squares criterion, the number of six clusters was chosen. Using the simulation with the K-means method (step 3), 588 of the 8786 observations under investigation has been identified as outliers. In the final clustering with the determinant criterion these outliers are downweighted in order to reduce or even negate their influence on the result of the clustering (step 4). The final result is depicted in Fig. 1. There are six well separated clusters. Two groups consist of phycoerithrin-containing algae (open squares and grey rhombs respectively), which can be clearly identified in the variable FL4 (Fig. 2). The diagonal band of algal cells in the centre (grey dots in Fig. 1) is known to represent green-algae (*Chlorophytes*). In the diagonal band above (black dots) algae containing yellow and brownish pigments respectively, here diatoms containing carotenoides, are identified as one group. In the diagonal below the green-algae, there are two groups of algae containing phycocyanin, a blue pigment (open triangle and black bars respectively). The differentiation of these two groups is caused by their cell-sizes, the variable FSC and is not shown here.

To verify the result of the model based Gaussian clustering from an ecological point of view, the result is compared with microscopic counting and determination of species. Since the microscopic determination of the algae is based on taxonomic classifications and those of flow cytometry on the pigment composition, only three different groups can be compared directly: 1. Diatoms, 2. algae containing the pigments phycoerithrin and phycocyanin respectively and 3. algae dominated by the pigment of chlorophyll (green algae). Only in the latter group of the green algae, the counts of both methods differ significantly. For the total cell count and the 1. and 2. group respectively, the quantities are in good agreement.

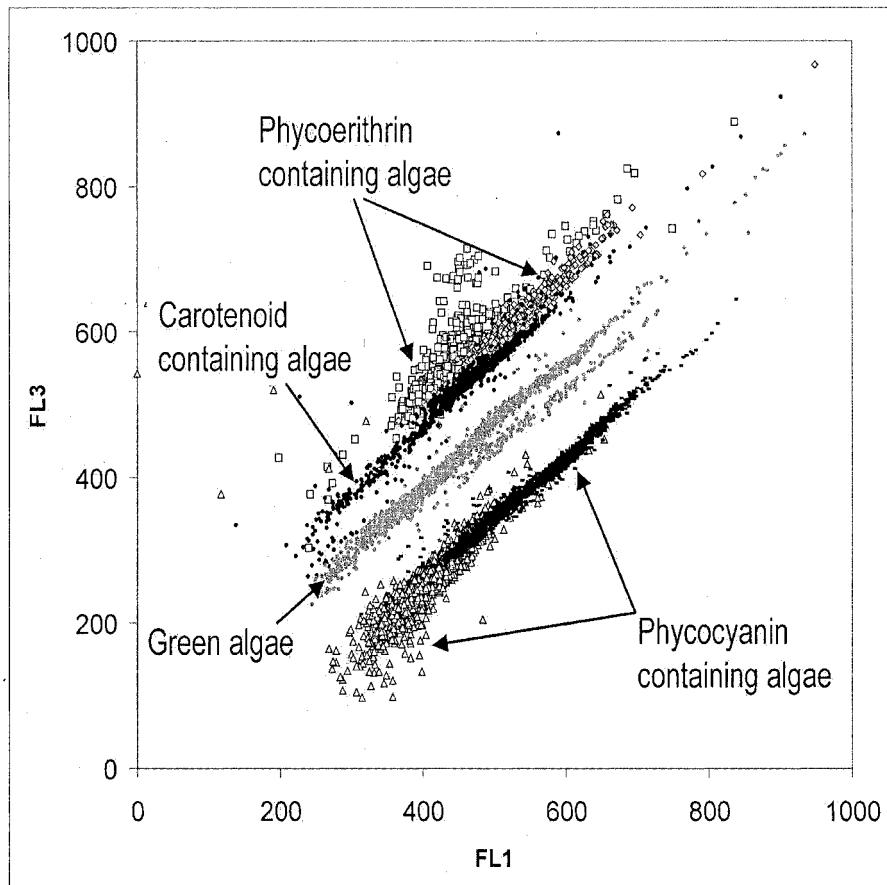


Fig. 1. Final result of model-based Gaussian clustering of 8256 observations. Sample of Lake Müggelsee (9th of July 2002). Green algae: grey dots; phycocyanin containing algae: open triangle and black bars respectively; carotenoid containing algae: black dots; phycoerithrin containing algae: open squares and grey rhombus respectively.

5 Discussion

The analysis of the snapshot of Lake Müggelsee proved, that the application of a sequence of different methods of cluster analysis to flow cytometry data has been successful. However, there are some topics to discuss.

The first crucial step of the whole procedure is the choice of the number of initial clusters to carry out step 1, the differentiation between algal-cells and other non-interesting observations. To assure, that by the cluster analysis with K-means all classes of ecological interest will be identified, the reasonable high number of nine classes has been chosen. The removal of classes

containing non-interesting observations was done by expert knowledge and is another critical step. To a certain degree, the decision is rather based on experience than on objective criteria. To objectify the differentiation between algal cells and other observations, in the future we will test the potency of a new calculated parameter, the ratio of the cell size to its pigment content (FSC/FL1). Algal cells should be identifiable as dense clouds of data points, whereas all other particles occur in a long tail with a declining ratio of FSC/FL1.

In step 2 the hierarchical Ward approach was carried out in order to find out an appropriate number of clusters for further clustering and a good initial solution for the more general Gaussian clustering. In the example of the sample of Lake Müggelsee the sum of square distance turned out to be a sensitive criterion, since the number of six clusters was rather sufficient. However, from an ecological point of view the two diagonal bands in the class of the green algae could be separated. When other samples has been analysed the sum of square criterion seem to be not appropriate.

In step 3, outliers have been identified in order to almost negate their influence on the final clustering with the determinant criterion and to allot them to a nearby cluster. This turned out to deliver a very satisfactory result, not only from the mathematical point of view but also from that of ecology. Even though in the example discussed here, there has been only 6,7% outliers, we know that in other samples the number can rise up to about 20-30%. Since the abundance of the algae is of ecological importance, these observations should not be neglected. The good agreement of the model based cluster analysis with microscopic quantification and qualification seem to confirm this opinion. Only in the group of the green algae there are significant quantitative differences. They might be explained with the difficulties in flow cytometry to distinguish between picoplanktic algae (up to $2\mu\text{m}$ of size) and larger cells. There is a high probability, that the higher number of green alga detected with flow cytometry is caused by small cells, which has not been counted in the microscope.

The description of the algal pigment-groups as Gaussian functions (Fig. 2) seem to be a promising approach to get precise mathematical descriptors. They should enable the detailed investigation of spatial and temporal changes in the pigment-composition of the phytoplankton community, which supposed to be sensitive to changes of the ecological state of the system. We think this might be a promising way to derive indicators, setting from them a multivariate problem, deduce a partial order ranking and finding finally a ranking probability distribution. By this different ecological states can be evaluated.

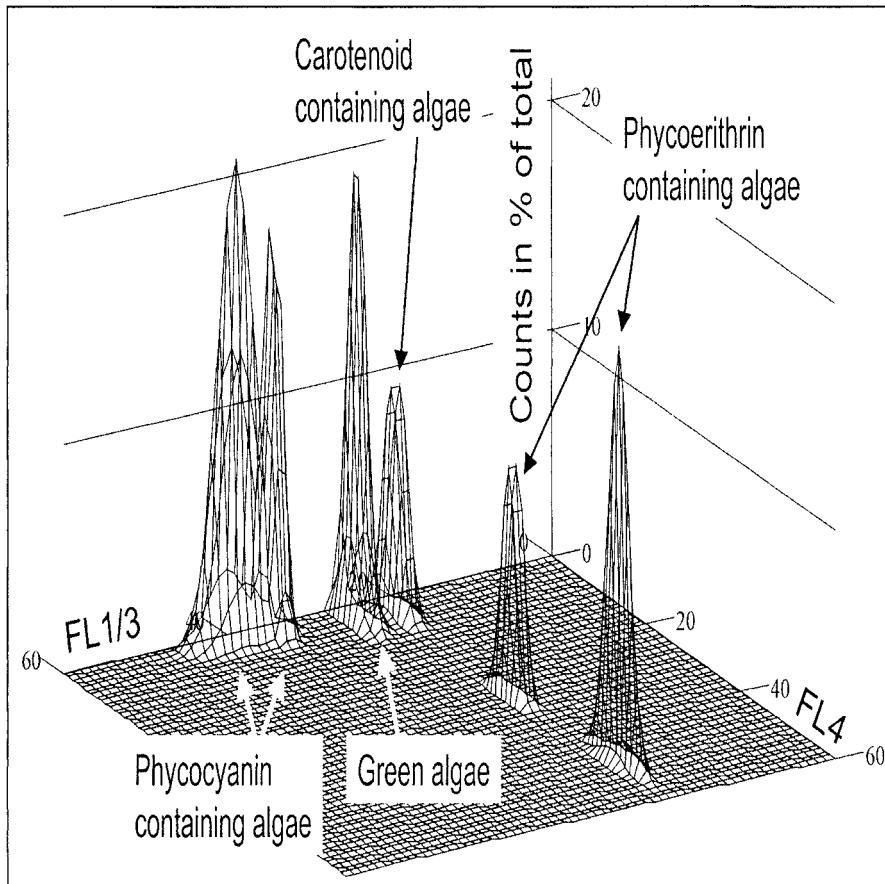


Fig. 2. Final result of the model-based clustering depicted as Gaussian-functions.

6 Conclusions

In general the pilot study to analysis flow cytometry data by a sequence of cluster analysis deliver satisfactory results. These results establish a benchmark that will be used in future investigations with other cluster analysis models. However, there are two main topics to improve. First, the differentiation between algal cells and observations which are supposed to be not interesting for further analysis has to be objectified. The calculation of the new parameter of the ratio of cell-size (FSC) to the pigment content of the cells (FL1) might be promising. Its potency has to be investigated in further studies. Second, a more sensitive criterion to find out an appropriate number of initial clusters for the application of the determinant criterion has to be developed. There are two main way to do so. Beside simulation studies based on the determinant criterion itself (rather than on the sum of square

criterion that is used here, because it is not the appropriate model) some variants of the Bayesian Information Criterion (BIC) should be used (Fraley and Raftery (2002)). Here the stability of the BIC against violation of normality (symmetry) has to be investigated beforehand. Once the number of clusters is determined by using either simulations or BIC, in a further step these hypotheses could be tested. The sum of square criterion turned out to be not suitable, as it does not enable the detection of an optimal number of classes from an ecological point of view.

It seems as if a true progress into insights of the ecology of lakes, which might be possible by flow cytometry, needs in future an extensive cooperation between ecologists and mathematician. The paper shows just the beginning of that faithful synthesis.

References

- BODDY, L., MORRIS, C.W., WILKINS, M.F., AL-HADDAD, L., TARRAN, G.A., JONKER, R.R., and BURKILL, P.H. (2000): Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. *Marine Ecology Progress Series*, 195, 47–59.
- CHISOLM, S.W., OLSON, R.J., ZETTLER, E.R., GOERICKE, R., WATERBURY, J.B., and WELSHMEYER, N.A. (1988): A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature*, 334, 340–343.
- FRALEY, C. and RAFTERY, A.E. (2002): Model-based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97, 458, 611–631.
- HOFSTRAAT, J.W., ZEIJL VAN, W.J.M., VREEZE DE, M.E.J., PEETERS, J.C.H., PEPERZAK, L., COLIJN, F., and RADEMAKER, T.W.M. (1994): Phytoplankton monitoring by flow cytometry. *Journal of Plankton Research* 16 (9), 1197–1224.
- MACQUEEN, J.B. (1967): Some Methods for Classification and Analysis of Multivariate Observations. In: L. Lecam and J. Neyman (Eds.): *Proc. 5th Berkeley Symp. Math. Statist. Prob.*, Vol. 1. Univ. California Press, Berkeley, 281–297.
- 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/>.
- STEINBERG, C.E.W. and BRÜGGEMANN, R. (1998): Integrity of limnic ecosystems. In: J.A. Van de Kraats (Eds.): *Let the Fish Speak: The Quality of Aquatic Ecosystems as an Indicator for Sustainable Water Management*. EURAQUA: Fourth Technical Report, Koblenz, 89–101.
- WARD, J.H. (1963): Hierarchical Grouping Methods to Optimise an Objective Function. *JASA*, 58, 235–244.

On Stratification Using Auxiliary Variables and Discriminant Method

Marcin Skibicki

Department of Statistics,
Katowice University of Economics, ul. Bogucicka 14, 40-226 Katowice, Poland

Abstract. Let U be a fixed population of size N from which a sample S of size n be drawn. We assume that values of variable Y are observed in the sample S and mean estimation is a goal of the survey. Moreover, let us assume that values of auxiliary variables X_1, \dots, X_m are known in the whole population. With fixed n , the mean should be estimated with an error as low as possible. Thus, the sample S selection method and an unbiased estimator with possible low variance should be established. A sampling strategy using auxiliary variables data may be an alternative to simply random sampling. The method proposed below depends on the selection of a preliminary sample S_p of size n_p ($n_p < n$) and next stratification of the remainder population and selection of a stratified sample S_s with size n_s .

1 Sampling strategy

We assume that the preliminary sample S_p is a simple sample drawn without replacement from population U . Let $U_s = U \setminus S_p$ be the part of population not included in S_p . The set U_s , after observation of S_p , will be divided into H (fixed number) strata U_{sh} of size N_{sh} . From the stratum U_{sh} a simple sample S_{sh} of size n_{sh} is drawn without replacement. We obtain the sample $S_c = \bigcup_{h=1}^H S_{sh}$.

As an estimator of the variable Y mean from the sample $S = S_p \cup S_s$ we take formula of the form:

$$\bar{y}_s = \frac{n_p}{N} \bar{y}_{S_p} + \frac{N - n_p}{N} \sum_{h=1}^H w_{sh} \bar{y}_{S_{sh}}, \quad (1)$$

where \bar{y}_{S_p} is the mean from S_p , $\bar{y}_{S_{sh}}$ is the mean from S_{sh} and w_{sh} is the fraction of elements in the stratum U_{sh} . This estimator is unbiased with variance:

$$\begin{aligned} D^2(\bar{y}_s) &= D_p^2(E_s(\bar{y}_s|S_p)) + E_p(D_s^2(\bar{y}_s|S_p)) \\ &= D_p^2\left(\frac{n_p}{N} \bar{y}_{S_p} + \frac{N - n_p}{N} \bar{y}_{U_s}\right) \end{aligned}$$

$$\begin{aligned}
& + E_p \left(\frac{(N - n_p)^2}{N^2} \sum_{h=1}^H \frac{(N_{sh} - n_{sh})}{N_{sh} n_{sh}} w_{sh}^2 V_{sh} \right) \\
& = E_p \left(\sum_{h=1}^H \frac{(N_{sh} - n_{sh}) N_{sh}}{N^2 n_{sh}} V_{sh} \right), \tag{2}
\end{aligned}$$

where $V_{sh} = \frac{1}{N_{sh}-1} \sum_{k \in U_{sh}} (y_k - \bar{y}_{U_{sh}})^2$ is the variance of variable Y in stratum U_{sh} .

To obtain a low value of the variance $D^2(\bar{y}_s)$, the set U_s must be divided into strata with low values of the variances V_{sh} . After observation of the preliminary sample S_p , we can divide this sample into subsets corresponding with the hypothetical strata U_{sh} . The K-means method of searching optimum division may be used. Next, applying one of the discriminant analysis methods, we can identify membership of the U_s elements to these subsets on the basis of auxiliary variables X_1, \dots, X_m values. In this way, strata U_{sh} will be obtained.

Thus the procedure of sample selection is as follows:

1. A preliminary sample S_p is drawn and observed.
2. Sample S_p is divided into H subsets on the basis of variable Y . Denote this partition by $\pi_{\text{opt}}(S_p)$.
3. Partition $\pi_{\text{opt}}(S_p)$ is treated as the model partition and set U_s is discriminated into U_{sh} . Denote this partition by $\pi_{\text{discr}}(U_s)$.
4. A stratified sample S_s is drawn.

If there is no assumption about the distribution of the vector $[Y, X_1, \dots, X_m]^T$ in the population U and thereby about the distribution in the strata of $\pi_{\text{opt}}(S_p)$, we can use one of the nonparametric discriminant methods, e.g. the recursive partitioning method (see Gatnar (2000)).

In point 4 of the procedure, sizes of samples S_{sh} must be established. Simply, they may be calculated proportional to the strata sizes, i.e.

$$n_{sh} = \left\lfloor n_s \frac{N_{sh}}{N - n_p} \right\rfloor.$$

2 Stratification with assumption of continuous distribution

Let us assume that the vector $[Y, X_1, \dots, X_m]^T$ has in population U a continuous distribution with density $f(y, x_1, \dots, x_m)$. From properties of the k-means method follows that $\pi_{\text{opt}}(S_p)$ is a partition into separated intervals of variable Y values. Let $y_{[h]}$, for $h = 1, \dots, H$, denotes the top limits of these intervals (for notation simplification we take $y_{[0]} = -\infty$ and $y_{[H]} = +\infty$).

Then the marginal distribution of Y in h -th stratum is a truncated distribution with a density function of the form:

$$f_{Yh}(y) = \frac{1}{F_Y(y_{[h]}) - F_Y(y_{[h-1]})} f_Y(y), \quad (3)$$

where f_Y and F_Y is the density and distribution function of Y adequately. Density of the marginal distribution of $[X_1, \dots, X_m]^T$ in h -th stratum is of the form:

$$f_{Xh}(x_1, \dots, x_m) = \frac{1}{F_Y(y_{[h]}) - F_Y(y_{[h-1]})} \int_{-\infty}^{\infty} I_{(y_{[h-1]}, y_{[h]})} f_Y(y, x_1, \dots, x_m) dy. \quad (4)$$

If there is a fixed partition $\pi_{\text{opt}}(S_p)$ into intervals and distribution (4) is known, we can calculate discriminant indicators:

$$\begin{aligned} C_h &= (F_Y(y_{[h]}) - F_Y(y_{[h-1]})) f_{Xh}(x_1, \dots, x_m) = \\ &= \int_{-\infty}^{\infty} I_{(y_{[h-1]}, y_{[h]})} f_Y(y, x_1, \dots, x_m) dy. \end{aligned} \quad (5)$$

Let us assume that the vector $[Y, X_1, \dots, X_m]^T$ has $m+1$ dimensional normal distribution. Then optimum values of $y_{[h]}$ are quantiles of the marginal normal distribution. These quantiles for $H = 2, 3, \dots$, were calculated by Cox (1957) and Dalenius (1951). But, because the parameters of the Y distribution are unknown, $y_{[h]}$ (proper quantiles) must be estimated on the basis of sample S_p . The estimator of r -th quantile is as follows:

$$\hat{y}_{(r)} = \hat{F}^{-1}(r),$$

where \hat{F}^{-1} is an estimator of distribution function of variable Y . Sampling variance of this estimator is of the form:

$$D^2(\hat{y}_{(r)}) = \left(1 - \frac{n_p}{N}\right) \frac{r_s(1 - r_s)}{n_p - 1},$$

where r_s is the fraction of elements in the sample that have values of Y not greater than $\hat{y}_{(r)}$.

3 Example

Let us take vector $[Y, X_1, X_2]$ jointly normal distributed in a population of size $N = 100\,000$. Also we take sample S_p size $n_p = 1\,000$, number of strata $H = 5$ and the variances $\sigma_Y = \sigma_{X_1} = \sigma_{X_2} = 100$. The partitions are formed using the above described methods. Distribution of the vector $[X_1, X_2]$ is treated as known. The $y_{[h]}$ are estimated and the discriminant indicators (5) are used. Table 1 shows the results at different sizes of samples,

and at different correlations ρ between variables. The $\bar{D}_s^2(\bar{y}_s | S_p)$ are averages of the variances $D_s^2(\bar{y}_s | S_p)$ for 1000 repetitions of population generation and sample S_p drawing. For comparison, variance of the mean estimator for simple sampling without replacement ($n = 1000$) takes value 0,99.

Size of S_p	Size of S_s	$[\rho_{YX_1}, \rho_{YX_2}, \rho_{X_1X_2}]$	$\bar{D}_s^2(\bar{y}_s S_p)$
700	300	[0.3, 0.5, 0.0]	2.419
		[0.5, 0.7, 0.0]	1.291
		[0.7, 0.9, 0.0]	0.493
600	400	[0.3, 0.5, 0.0]	1.813
		[0.5, 0.7, 0.0]	0.983
		[0.7, 0.9, 0.0]	0.368
500	500	[0.3, 0.5, 0.0]	1.442
		[0.5, 0.7, 0.0]	0.796
		[0.7, 0.9, 0.0]	0.301
400	600	[0.3, 0.5, 0.0]	1.194
		[0.5, 0.7, 0.0]	0.663
		[0.7, 0.9, 0.0]	0.251
300	700	[0.3, 0.5, 0.0]	1.033
		[0.5, 0.7, 0.0]	0.575
		[0.7, 0.9, 0.0]	0.221
200	800	[0.3, 0.5, 0.0]	0.957
		[0.5, 0.7, 0.0]	0.527
		[0.7, 0.9, 0.0]	0.199

Table 1. Results of simulations

References

- COX, D.R. (1957): Note On Grouping. *J. Am. Stat. Assoc.*, 52, 543–547.
- DALENIUS, T. (1951): Problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, 34, 133–148.
- GATNAR, E. (2001): *Nonparametric Method of Discrimination and regression* (in Polish), PWN, Warszawa.
- HARTIGAN, J.A. (1975): *Clustering Algorithms*. John Wiley, New York.
- HUBERTY, C.J. (1994): *Applied discriminant analysis*. Wiley and Sons, New York.
- SARNDAL, C.A., SWENSSON, B., and WRETMAN, J. (1992): *Model Assisted Survey Sampling*. Springer-Verlag, New York.
- WYWIAL, J. (1998): Estimation of population average on the basis of strata formed by means discrimination functions. *Statistics in Transition*, 4, 5, 903–913.
- WYWIAL, J. (2002): On Stratification of Population on the Basis of Auxiliary Variable and the Selected Sample. *Acta Universitatis Lodzienensis, Folia Oeconomica* 156, 83–90.

Measuring Distances Between Variables by Mutual Information

Ralf Steuer¹, Carsten O. Daub², Joachim Selbig² and Jürgen Kurths¹

¹ University of Potsdam, Nonlinear Dynamics Group, Am Neuen Palais 10,
D-14469 Potsdam, Germany

² Max-Planck Institute for Molecular Plant Physiology, Am Mühlenberg 1,
D-14476 Golm, Germany

Abstract. Information theoretic concepts, such as the mutual information, provide a general framework to detect and evaluate dependencies between variables. In this work, we describe and review several aspects of the mutual information as a measure of 'distance' between variables. Giving a brief overview over the mathematical background, including its recent generalization in the sense of Tsallis, our emphasis will be the numerical estimation of these quantities from finite datasets. The described concepts will be exemplified using large-scale gene expression data and compared to the results obtained from other measures, such as the Pearson Correlation.

1 Introduction

The detection of relationships between two or more variables is of central interest in many areas of science. Among the most recent examples are the latest breakthroughs in molecular biology, leading to the ability to quantify whole-genome mRNA abundance in large-scale experiments (Schena et al. (1995), Brazma and Vilo (2000)). With these data at hand, the need for new approaches to unravel the functional relationships implicit in these datasets has become apparent (Eisen et al. (1998), D'haeseleer (2000)). In this context, information theoretic ideas, such as the mutual information, have been suggested to extend traditional analysis, with examples ranging from the clustering of expression (Michaels et al. (1998), Butte and Kohane (2000), Somogyi et al. (2001)) and cDNA-fingerprinting data (Herwig et al. (1999)) to reverse engineering (Liang et al. (1998)). Also, apart from these fields, the mutual information is widely utilized in diverse scientific disciplines, such as physics (Fraser and Swinney (1986)), data and sequence analysis (Grosse et al. (2000), Herzel and Grosse (1997)), among various others.

In this work, we describe and review several aspects of the mutual information as a measure of 'distance' between variables. We will not aim at the mathematical subtleties of information theory, but focus on the practical applicability of the described concepts. The paper is organized as follows: Starting with a brief review of some information-theoretic ideas, we provide two approaches leading to the definition of the mutual information. In the

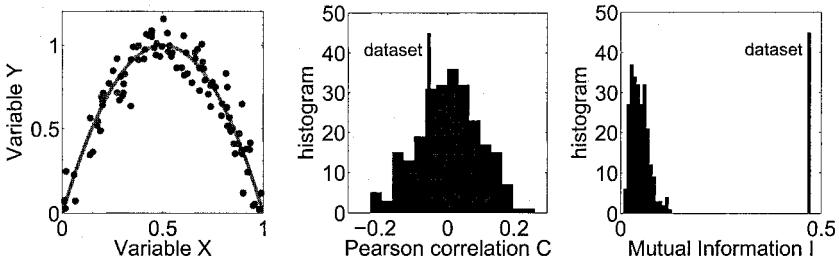


Fig. 1. *Left:* A hypothetical dependency between two variables X and Y (100 data-points). *Center:* The Pearson Correlation fails to detect any significant correlation, with respect to an ensemble of shuffled counterparts (shown as a histogram). *Right:* The mutual information clearly indicates that these two variables are not independent.

subsequent section a recently proposed generalization is discussed. The next part is devoted to an overview on algorithms to estimate the mutual information from finite datasets, ranging from simple partition-based estimation to more complex schemes. In the last part, we focus on the application of these concepts in the analysis of large-scale gene expression data.

2 The mutual information

The mutual information provides a general criterion for statistical independence between random variables. In contrast to other measures, such as the Euclidean distance or the Pearson correlation, it is not restricted to linear dependencies, but is able to quantify arbitrary functional relationships between two (or more) variables. An illustrative example is given in Fig. 1. The mathematical definitions are given in the following.

The Shannon entropy

For a random variable A with a finite set of M_A possible states $\{a_1, a_2, \dots, a_{M_A}\}$, the Shannon entropy $H(A)$ is defined as (Shannon (1948))

$$H(A) = - \sum_{i=1}^{M_A} p(a_i) \log p(a_i) \quad (1)$$

with $p(a_i)$ denoting the probability of the state a_i . As already pointed out by Fraser and Swinney (1986), $H(A)$ can be described as the ‘quantity of surprise you should feel upon reading the result of a measurement’. That is, if the outcome of the measurement is completely determined to be a_l ($p(a_l) = 1$ and $p(a_i) = 0$ for $i \neq l$), the Shannon entropy is zero. On the other hand, $H(A)$ is maximal if all possible outcomes a_i are equiprobable.

For two variables A and B the joint entropy $H(A, B)$ is defined analogously and obeys the relation

$$H(A, B) = - \sum_{i=1}^{M_A} \sum_{j=1}^{M_B} p(a_i, b_j) \log p(a_i, b_j) \leq H(A) + H(B) \quad (2)$$

with the equality only if and only if both systems are statistically independent, i.e. $p(a_i, b_j) = p(a_i) p(b_j)$ for all i, j . The *mutual information* or *transinformation* $I(A, B)$ between two variables A and B can then be defined as (Shannon (1948), Cover and Thomas (1991))

$$I(A, B) := H(A) + H(B) - H(A, B) \geq 0 \quad (3)$$

Note that the mutual information $I(A, B)$ can never be larger than any of the individual entropies.

$$I(A, B) \leq \min\{H(A), H(B)\} \quad (4)$$

The Kullback entropy

A different approach to the mutual information is given by the *Kullback entropy* or *relative entropy* $K(p|p^0)$ (Cover and Thomas (1991)). If $p = (p_1, \dots, p_M)$ and $p^0 = (p_1^0, \dots, p_M^0)$ are two probability distributions on the same set $\{1, \dots, M\}$, it is defined by

$$K(p|p^0) := \sum_{i=1}^M p_i \log \frac{p_i}{p_i^0} \geq 0 \quad (5)$$

Even if the Kullback entropy is not symmetric and thus not a distance in the mathematical sense, it can be interpreted as a measure of distance between the distributions p and p^0 . In particular, $K(p|p^0)$ gives the *information gain* when replacing an initial probability distribution p^0 by a final distribution p . In our case, the probability distribution p^0 is given by the hypothesis of statistical independence $p^0(a_i, b_j) = p(a_i) p(b_j)$, whereas the distribution p is given by the actual joint probability distribution $p(a_i, b_j)$. Thus

$$K(p|p^0) = \sum_{i=1}^{M_A} \sum_{j=1}^{M_B} p(a_i, b_j) \log \frac{p(a_i, b_j)}{p(a_i) p(b_j)} = I(A, B) \quad (6)$$

which is identified to be the mutual information $I(A, B)$, as given in Eq. (3). $I(A, B)$ thus establishes a measure of 'distance' between the hypothesis of statistical independence and the actual joint probability distribution.

An approximation for weak dependencies

Of particular interest is the asymptotic form of the mutual information for weak dependencies. Let $\epsilon_i = p_i - p_i^0$ denote the deviations of p and p^0 . A Taylor expansion of Eq. (5) in powers of ϵ yields (Cover and Thomas (1991)):

$$K = \frac{1}{2} \sum_{i=1}^M \frac{\epsilon_i^2}{p_i^0} + \mathcal{O}(\epsilon^3) \approx \frac{1}{2} \sum_{i=1}^M \frac{(p_i - p_i^0)^2}{p_i^0} \quad (7)$$

Equation (7) corresponds to the chi-square statistics, which provides the common test for differences between (discrete) distributions (Press et al. (1992)).

3 The generalized mutual information

As already indicated in Eq. (7), the mutual information does not stand isolated, but is closely related to other measures. This becomes even more apparent, if we consider a generalized Kullback entropy, as proposed by Tsallis (1998). Recall the definition of the generalized nonextensive Tsallis entropy $H_q(A)$ of order q (Curado and Tsallis (1991), Cover and Thomas (1991)).

$$H_q(A) = -\frac{1}{q-1} \sum p(a_i) [p(a_i)^{q-1} - 1] \quad (8)$$

Equation (8) indeed represents a proper generalization of Eq. (1), which is recovered in the limit $q \rightarrow 1$. Along these lines, it is possible to obtain a generalized form K_q of Eq. (5) (Tsallis (1998)):

$$K_q = \frac{1}{q-1} \sum p_i \left[\left(\frac{p_i}{p_i^0} \right)^{q-1} - 1 \right] \quad q > 0 \quad (9)$$

For $q > 0$ the generalized Kullback entropy K_q is always greater than zero, and zero if and only if $p_i = p_i^0$. It is straightforward to verify that in the limit $q \rightarrow 1$ Eq. (9) corresponds to the traditional Kullback entropy:

$$\lim_{q \rightarrow 1} K_q = \sum p_i \log \frac{p_i}{p_i^0} = K(p|p^0) \quad (10)$$

Further, for $q = 2$ we get

$$K_2 = \sum p_i \left[\left(\frac{p_i}{p_i^0} \right) - 1 \right] = \sum \frac{(p_i - p_i^0)^2}{p_i^0} \quad (11)$$

which is the chi-square statistics Eq. (7). Thus, if we again identify p with a joint probability distribution and p^0 with the product of the marginal distributions, Eq. (9) gives a proper generalization $I_q(A, B)$ of the mutual information $I(A, B)$ for evaluating dependencies between variables A and B .

4 Numerical estimation from finite data

In most practical applications the probability distributions used in the definitions of the mutual information are unknown and have to be estimated from finite data. In the following and in Section 5, an overview on several algorithms, as presented in (Steuer et al. (2002)), is supplemented with some additional aspects.

The simple algorithm

Suppose that the experimental data consists of two (or more) random variables X, Y, \dots which have been jointly measured at several times $k = 1, \dots, N$, yielding the sample $(x_1, y_1, \dots), \dots, (x_N, y_N, \dots)$. The most widely used approach to estimate the mutual information is then based on binning the data into M discrete and equally sized intervals a_1, a_2, \dots, a_M and b_1, b_2, \dots, b_M . An indicator function Θ_{ij} counts the number of datapoints within each bin $a_i \times b_j$ and the corresponding probabilities are estimated by the relative frequencies of occurrence.

$$\hat{p}(a_i, b_j) = \frac{1}{N} \sum_k \Theta_{ij}(x_k, y_k) \quad (12)$$

Here $\Theta_{ij}(x_k, y_k) = 1$ if $x_k \in a_i$ and $y_k \in b_j$ and zero otherwise. The estimated marginal probabilities are $\hat{p}(a_i) = \sum_{j=1}^M \hat{p}(a_i, b_j)$.

It should be emphasized that, provided the number of datapoints is sufficiently high, this simple approach already gives reliable results and is frequently used in the literature (e.g. Butte and Kohane (2000)).

However, if only a moderate amount of datapoints is available, the result is substantially affected by finite-size effects. For finite N , the mutual information gets systematically overestimated (Herzel et al. (1994)).

$$\langle I^{\text{observed}} \rangle = I^{\text{true}} + \frac{(M-1)^2}{2N} + \mathcal{O}\left(\frac{1}{N^2}\right) \quad (13)$$

An example is given in Fig. 2. We used $N = 100$ datapoints in $[0, 1]^2$, partitioned into $M = 6$ bins on each axis. Since we used two uncorrelated variables, we have $I^{\text{true}} = 0$, while the predicted deviation according to Eq. (13) is $\langle I^{\text{observed}} \rangle \approx 0.125$ in good agreement with the numerical results shown in Fig. 2. Note that Eq. (13) also sets a limit on the minimal number of datapoints required for a numerical estimation of $I(A, B)$. As a rule of thumb, each 'bin' in Fig. 2a should have a chance to appear (at least) three times, thus $N_{\min} \geq 3M^2$ (Herzel et al. (1994)).

The variance of the mutual information: Not only the systematic deviations, but also the variance of the mutual information can be approximated analytically (Herzel and Grosse (1995, 1997)).

$$\sigma^2(I) = \frac{1}{N} \left[\sum p(a_i, b_j) \log^2 \frac{p(a_i, b_j)}{p(a_i)p(b_j)} - I^2 \right] + \mathcal{O}\left(\frac{1}{N^2}\right) \quad (14)$$

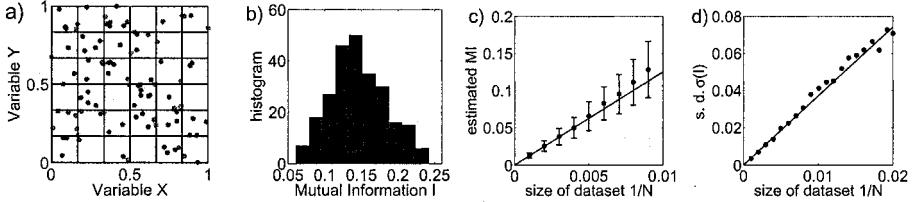


Fig. 2. The mutual information for finite data. *a)* $N = 100$ equidistributed random datapoints, partitioned into $M = 6$ bins on each axis. *b)* A histogram of the estimated mutual information (300 realizations). The mean value is $\langle I^{\text{observed}} \rangle = 0.15 \pm 0.04$ in good agreement with Eq. (13). *c)* The average estimated mutual information as a function of the (inverse) size of the dataset with errorbars denoting the standard deviation. The solid line corresponds to the theoretical prediction of Eq. (13). *d)* The standard deviation σ of the mutual information as a function of the (inverse) size of the dataset. The solid line shows a least squares fit.

Note that the $1/N$ -term vanishes for statistically independent events. In this case the leading term is of order $1/N^2$ for the variance, thus $1/N$ for the standard deviation. This is shown in Fig. 2d.

5 Kernel density estimation

While most algorithms found in the literature rely on partitioning the data into discrete bins, more sophisticated alternatives are available for the continuous form of the mutual information (Moon et al. (1995)).

$$I(X, Y) = \int_x \int_y f(x, y) \log \frac{f(x, y)}{f(x) f(y)} dx dy \quad (15)$$

To evaluate Eq. (15) from finite data, we have to estimate the probability density $f(x, y)$. The method of choice for this task is a kernel density estimator $\hat{f}(\mathbf{x})$ (Silverman (1986)),

$$\hat{f}(\mathbf{x}) := \frac{1}{N h^d} \sum_{i=1}^N \mathcal{K}\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \quad \mathbf{x} \in \mathbb{R}^d \quad (16)$$

where $\mathbf{x} = (x, y, z, \dots)^T$ denotes a point from \mathbb{R}^d and \mathbf{x}_i is one of the d -dimensional vector of observations. For simplicity we restrict ourselves to the multivariate Gaussian kernel function, for other possible choices see (Silverman (1986)).

$$\mathcal{K}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2} \mathbf{x}^T \mathbf{x}\right) \quad (17)$$

Heuristically Eqs. (16) and (17) may be understood as placing little Gaussian ‘bumps’ at the position of each observation, as illustrated in Fig. 3.

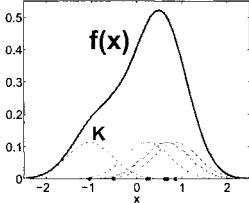


Fig. 3. A one-dimensional kernel density estimator using a Gaussian kernel may be explained as placing Gaussian ‘bumps’ at the position of each observation x_i . The estimator according to Eq. (16) is then given by the sum of these bumps (Silverman (1986)).

The only free parameter in this procedure is the *smoothing parameter* or *bandwidth* h . As a tradeoff between computational effort and performance, Silverman (1986) suggests an optimal bandwidth that minimizes the mean integrated squared error for Gaussian distributions.

$$h_{\text{opt}} \approx \sigma \left(\frac{4}{d+2} \right)^{1/(d+4)} N^{-1/(d+4)} \quad (18)$$

Previous studies revealed that the mutual information is not too sensitive to a particular choice of h and that in most cases Eq. (18) gives reasonable results (Steuer et al. (2002)).

Once we have obtained an estimate of the probability densities, an evaluation of Eq. (16) may proceed with standard techniques of numerical integration. However, since this may put high demands on computational power, we can ask for a substantial simplification. Note that the mutual information Eq. (15) represents an *average* over the xy -plane. Thus, provided that our dataset is itself a faithful sample of the underlying distribution, we can estimate Eq. (15) as follows:

$$I(X, Y) = \left\langle \log \frac{f(x, y)}{f(x)f(y)} \right\rangle \approx \frac{1}{N} \sum_{i=1}^N \log \left[\frac{\hat{f}(x_i, y_i)}{\hat{f}(x_i)\hat{f}(y_i)} \right] =: \hat{I}(X, Y) \quad (19)$$

In the following, Eq. (19) will be used to exemplify the application on experimental data.

6 Application on experimental data

We will now apply the above-described concepts on large-scale gene expression data, obtained from cDNA microarrays measurements. For this kind of data, one of the prevailing methods of analysis is the clustering of genes into groups of ‘similar’ expression patterns (D’haeseleer et al. (2000)). However, in most cases, ‘similarity’ is restricted to linear measures, such as the Euclidean distance or the Pearson correlation C , defined as

$$C = \frac{\Gamma_{xy}}{\sqrt{\Gamma_{xx}\Gamma_{yy}}} \quad \Gamma_{xy} = \langle xy \rangle - \langle x \rangle \langle y \rangle \quad (20)$$

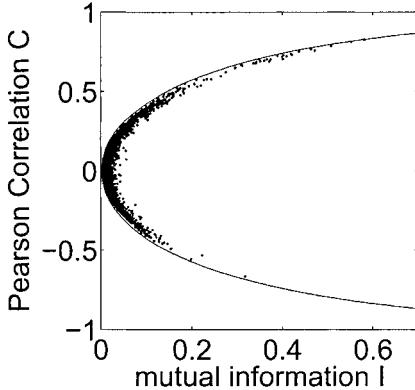


Fig. 4. The mutual information, estimated according to Eq. (19), versus the Pearson correlation for large-scale expression data. The pair-wise Pearson correlation C and the mutual information I was numerically estimated for all 300 experimental conditions. Each dot corresponds to a tuple (\hat{I}, \hat{C}) . The solid line is the theoretical relationship given by Eq. (23). All data were rank-ordered prior to the analysis.

with Γ denoting the covariance of the data. Thus, it remains crucial to verify that this restriction does not miss a substantial fraction of the correlations contained in the data.

To this end, we compute both, the Pearson correlation as well as the mutual information for a publicly available dataset of gene expression, corresponding to up to 300 diverse mutations and chemical treatments in yeast (Hughes et al. (2000)). In extension to a previous study (Steuer et al. (2002)), here we focus on the pair-wise correlations between experimental conditions, each consisting of data for more than 6000 genes.

If no *nonlinear* dependencies are present within the data, we expect a one-to-one relationship between the (absolute value of the) Pearson correlation and the mutual information. In particular, if we assume that the data were drawn from a multivariate Gaussian distribution

$$p(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det\Gamma}} \exp\left(-\frac{1}{2}\mathbf{x}^T\Gamma^{-1}\mathbf{x}\right) \quad \text{with } \mathbf{x} = (x, y)^T \quad (21)$$

and are thus fully characterized by their covariance matrix Γ (and hence their Pearson correlation), we can calculate this expected relationship explicitly. Inserting Eq. (21) into the continuous definition of the Shannon entropy yields

$$H(X) = \frac{1}{2} [1 + \log(2\pi\Gamma_{xx})] \quad \text{and} \quad H(X, Y) = 1 + \log[2\pi\sqrt{\det\Gamma}] \quad (22)$$

Thus the expected mutual information I is

$$I = -\log[\sqrt{1 - C^2}] \quad (23)$$

Figure 4 shows a numerical comparison between the mutual information and the Pearson correlation for the dataset described in (Hughes et al. (2000)). As can be observed, we detect no genuinely *nonlinear* dependencies (high mutual information, but vanishing Pearson correlation) within this dataset. Similar results are obtained if we consider the pair-wise dependencies between genes, instead of experimental conditions (Steuer et al. (2002)).

7 Discussion and conclusions

Clustering of co-expressed genes has become one of the major techniques to extract putative functional relationships from large-scale expression data. To assess the similarity of expression patterns, most algorithms make use of linear measures, such as the Pearson correlation or the Euclidean distance. However, such a restriction might miss a substantial fraction of potential relationships. Consequently, the use of alternative measures, such as the mutual information, has been suggested in the literature (e.g. Michaels et al. (1998)).

In this work, we have provided a brief review of current numerical methods and presented a systematic comparison between the mutual information and the Pearson correlation. No genuinely nonlinear relationships were detected between the analyzed dataset. Note that this is by no means trivial: While most nonlinear dependencies, like the one depicted in Fig. 1, can be easily classified by eye, clustering large-scale expression data usually involves several millions of pair-wise similarities. It is thus necessary to validate the appropriateness of distance measures by computational means. Since linear measures offer several advantages in terms of numerical effort and datapoints required for reliable estimation, they are often used on an ad-hoc basis without any further justification (Brazma and Vilo (2000)). Also, due to a limited number of datapoints, most available datasets do not allow for a comparison between more elaborate distance measures. Taking one of the largest available datasets as an example, our findings suggest that the Pearson correlation is sufficient to extract all pair-wise correlations. While, of course, general conclusions cannot be drawn from the analysis of one single dataset alone, our results still indicate that for the clustering of co-expressed genes the restriction to linear measures is appropriate when the number of datapoints is small.

The authors would like to thank C. Zhou, A. Floeter, U. Schwarz (*all University Potsdam*), and W. Ebeling (*HU-Berlin*) for fruitful discussion. R. S. acknowledges financial support by the HSP-N grant of the the state of Brandenburg.

References

- BRAZMA, A. and VILO J. (2000): Gene expression data analysis. *FEBS Letters*, 480, 17–24.
- BUTTE, A.J. and KOHANE, I.S. (2000): Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 5, 415–426.
- COVER, T.M. and THOMAS, J.A. (1991): *Elements of Information Theory*. John Wiley, New York.
- CURADO, E.M.F. and TSALLIS, C. (1991): Generalized statistical mechanics: Connection with thermodynamics. *J. Phys. A*, 24, L69.
- D'HAESELEER, P., LIANG, S., and SOMOGYI, R. (2000): Genetic network inference: From co-expression clustering to reverse engineering. *Bioinformatics*, 16 (8), 707–726.

- EISEN, M.B., SPELLMAN, P.T., BROWN, P.O., and BOTSTEIN, D. (1998): Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, *95*, 14863–14868.
- FRASER, A.M. and SWINNEY, H.L. (1986): Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, *33* (2), 2318–2321.
- GROSSE, I., HERZEL, H., BULDYREV, S.V., and STANLEY, H.E. (2000): Species independence of mutual information in coding and noncoding DNA. *Phys. Rev. E*, *61* (5), 5624–5629.
- HERWIG, R., POUSTKA, A.J., MUELLER, C., BULL, C., LEHRACH, H., and O'BRIAN, J. (1999): Large-scale clustering of cDNA-fingerprinting data. *Genome Research*, *9* (11), 1093–1105.
- HERZEL, H. and GROSSE, I. (1995): Measuring correlations in symbols sequences. *Physica A*, *216*, 518–542.
- HERZEL, H. and GROSSE, I. (1997): Correlations in DNA sequences: The role of protein coding segments. *Phys. Rev. E*, *55* (1), 800–810.
- HERZEL, H., SCHMITT, A.O., and EBELING, W. (1994): Finite sample effects in sequence analysis. *Chaos, Solitons & Fractals*, *4* (1), 97–113.
- HUGHES, T.R. et al. (2000): Functional discovery via a compendium of expression profiles. *Cell*, *102*, 109–126.
- LIANG, S., FUHRMAN, S., and SOMOGYI, R. (1998): Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pacific Symposium on Biocomputing*, *3*, 18–29.
- MICHAELS, G.S., CARR, D.B., ASKENAZI, M., FUHRMAN, S., WEN, X., and SOMOGYI, R. (1998): Cluster analysis and data visualization of large-scale gene expression data. *Pacific Symposium on Biocomputing*, *3*, 42–53.
- MOON, Y., RAJAGOPALAN, B., and LALL, U. (1995): Estimation of mutual information using kernel density estimators. *Phys. Rev. E*, *52* (3), 2318–2321.
- PRESS, W.H., TEUKOLSKY, S.A., VETTERLING, W.T., and FLANNERY, B.P. (1992): *Numerical Recipes in C*. Second edition, Cambridge University Press, Cambridge.
- SCHENA, M., SHALON, D., DAVIS, R.W., and BROWN, P.O. (1995): Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*, 467–470.
- SHANNON, C.E. (1948): A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423, ibid. 623–656.
- SILVERMAN, B.W. (1986): *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.
- SOMOGYI, R., FUHRMAN, S., and WEN, X. (2001): Genetic network inference in computational models and applications to large-scale gene expression data. In: J. M. Bower and H. Bolouri (Eds.): *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, Cambridge, 129–157.
- STEUER, R., KURTHS, J., DAUB, C.O., WEISE, J., and SELBIG, J. (2002): The mutual information: Detecting and evaluating dependencies between variables. *Bioinformatics*, *18* (Suppl. 2), 231–240.
- TSALLIS, C. (1998): Generalized entropy-based criterion for consistent testing. *Phys. Rev. E*, *58* (2), 1442–1445.

Pareto Density Estimation: A Density Estimation for Knowledge Discovery

Alfred Ultsch

Databionics Research Group,
University of Marburg, D-35032 Marburg, Germany

Abstract. Pareto Density Estimation (PDE) as defined in this work is a method for the estimation of probability density functions using hyperspheres. The radius of the hyperspheres is derived from optimizing information while minimizing set size. It is shown, that PDE is a very good estimate for data containing clusters of Gaussian structure. The behavior of the method is demonstrated with respect to cluster overlap, number of clusters, different variances in different clusters and application to high dimensional data. For high dimensional data PDE is found to be appropriate for the purpose of cluster analysis. The method is tested successfully on a difficult high dimensional real world problem: stock picking in falling markets.

1 Introduction

Density based clustering algorithms have drawn much attention in the last years within the context of knowledge discovery in databases (Ester et al. (1996), Xu et al. (1998), Hinneburg and Keim (1998)). All these algorithms rely on methods to estimate the probability density function from the observed data. Methods for density estimation have been studied intensively in mathematics and statistics. Density estimation using the number of points within a hypersphere of a fixed radius around each given data point is used in many of the density clustering algorithms. In this paper we propose a radius for hypersphere density estimation that is optimal in an information theoretic sense. Information optimization calls for a radius such that the hyperspheres contain a maximum of information using minimal volume. Consequence of this approach is that a radius is optimal, if a hypersphere contains in the average about 20% of the data. This gives more than 80% of the possible information any subset of data can have. Since these results coincide with the widely known rule of thumb called "Pareto's 80/20 law" we decided to call this radius the Pareto radius. The hypersphere density estimation method using this radius is called Pareto Density Estimation (PDE). We show in this paper that PDE is an optimal hypersphere density estimation method for data with a mixture of Gaussians as probability density function. It turns out that the method is valid even when the clusters overlap to certain degrees and when the inner cluster variances differ. Furthermore the method scales

appropriately with dimensionality of the data set. PDE is tested on a difficult real world problem: the selection of winning stocks in a falling market.

2 Methods for density estimation

Density estimation means the construction of an estimate of the true probability density function from the observed data. Methods for density estimation have been studied intensively. See Scott (1992) for an overview. Most density-estimators are based upon one or more of the following techniques: finite mixture models, variable kernel estimates, uniform kernel estimates. Finite mixture models attempt to find a superposition of parameterized functions, typically Gaussians which best account for the sample data. The method can in principle model any shape of cluster, and works best when the data's probability density can be described as a mixture of Gaussians. With kernel based approaches the true probability density function is estimated using local approximations. The local approximations are parameterized such that only data points within a certain distance of the point under consideration have an influence on the shape of the kernel function. This is called (band-) width or radius of the kernel (Scott (1992)). Variable kernel methods adjust the radius of the kernel. Uniform kernel algorithms use a fixed global radius. A special case of uniform kernel estimates is hypersphere density estimation. The number of points within a hypersphere around each data point is used for the density estimation at the center of the hypersphere. Uniform kernel estimates can approximate the true probability up to any desired degree of accuracy, if the true probability is known (Devroye and Lugosi (1996, 1997)). Tuning the bandwidth for optimal variable kernel estimation is computationally expensive and proven to be a computational hard task (Devroye and Lugosi (2000)). This is a clear disadvantage of such methods for large data sets. This is one of the reasons why uniform kernel methods have become popular within the context of knowledge discovery in (large) databases (KDD). Clustering methods as used for KDD usually require the definition of a (dis-)similarity measure between two data points. Density estimation within the context of KDD should therefore use the benefits of a given distance measure in particular for high dimensional data. Regions in data space with high density values are good candidates for clusters if these regions are surrounded by substantially less dense or even empty regions. A suitable estimate for KDD needs therefore to be precise in dense regions and less precise in almost empty regions. Data points in very low density regions are most likely outliers, i.e., are of no concern for clustering. All these requirements make hypersphere density estimation with a global radius a good candidate for density estimation for clustering in databases. For clustering the selection of an appropriate radius should be based on the distribution of the data distances.

3 Pareto density estimation

Let S be a subset of a set of n points with $|S| = s$ the number of elements in S . Then $p = s/n$ is the relative size of the set. If there is an equal probability that an arbitrary point x is observed, p is the probability $p = p(x \in S)$. Information theory calculates the entropy or (partial) information using p . Scaled to the range $[0, 1]$, the information of a set is calculated as $I(S) = -e p \ln(p)$. To find an optimal set size, define the unrealized potential $URP(S)$ of a set as the Euclidian distance from the ideal point, i.e. an empty set producing 100% of information. This definition of $URP(S)$ leads to: $URP(S) = \sqrt{p^2 + (1 + e p \ln(p))^2}$. Minimizing the unrealized potential results in an optimal set size of $p_u = 20.13\%$. This set size produces 88% of the maximum information. For details see (Ultsch (2001)). The optimality of this set at about (20%, 80%) might be the reason behind the so called Pareto 80/20 law, which is empirically found in many domains (Ultsch (2001)). Subsets or volumes which contain in the average p_u data points are optimal in the sense that they give as much information as possible with a minimal set size. Define the neighborhood number $NN(x, r)$ as the number of input data points within a hypersphere (neighborhood) with radius r around a point x in data space. Even if the input is drawn from a Normal distribution, the neighborhood numbers are not normally distributed. The Pareto Radius r_p of a data set is a radius such that for all data points the median of $NN(x, r)$ equals $p_u d$, with $p_u = 0.2013$ and d the number of data points in the data set. Searching among the distance percentiles of the data is a useful way to limit the effort to approximate the Pareto Radius in practical applications. Let $pc(p)$ denote the p -th percentile of the distances between two different points in the data set. The Pareto Percentile p_{par} is that percentile of all distances which is closest to the Pareto Radius i.e. $p_{par} = \text{argmin}(|pc(p) - rp|)$, $\forall p \in 1, \dots, 99$.

4 Adjusting for intra/inter cluster distances

The detection of the cluster structure in an input data set using distance structures is only possible, if most of the data distances within a cluster are smaller than the distances measured between data from different clusters. Let v denote the ratio of intra cluster distances to inter distances. If this ratio is known a priori the neighborhood radius for the estimation of a suitable data density can be adapted. The Pareto percentile within a cluster can be calculated as $p = p_{par} * v$. To estimate v for unknown number of clusters and sizes, experiments for a wide range of cluster numbers and sizes were performed. Data set size was set to $d = 1000$ points. For the number of clusters k within the range $[1, 40]$ the relative size p_i of cluster i was randomly drawn from a normal distribution $N(m, s)$ with mean $m = k^{-1}$, and $s = 10$. The variance was chosen so large to generate in particular very uneven cluster sizes. Each cluster was required to consist of at least one single data point.

For each of the cluster numbers k , 10.000 cases of cluster sizes were generated and the ratio v was calculated. The mean values of $\bar{v}(k)$ for each number of clusters k are shown with a circle in Figure 1. The 95% confidence interval for v was calculated. This is the interval in which the values for v can be found with at most an error probability of 5%. Figure 1 shows $\bar{v}(k)$ and the 95% confidence interval versus the cluster number k . For details see Ultsch (2003).

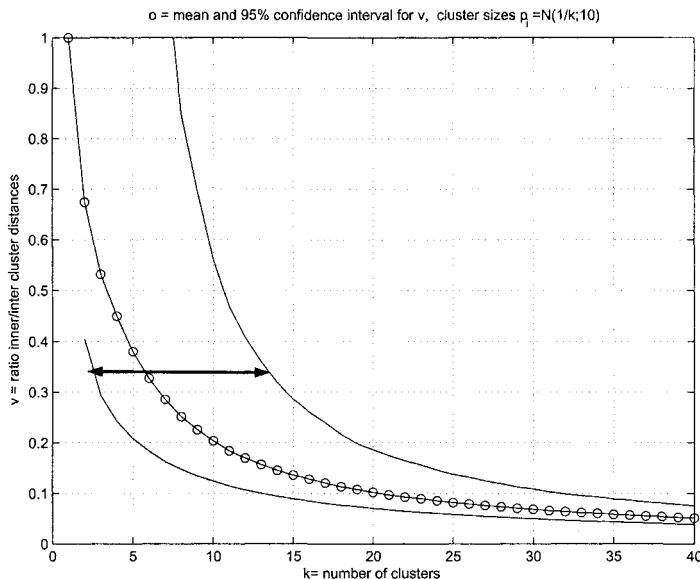


Fig. 1. Ratio of intra/inter cluster distances

Different experimental setting for variance s in the range of $s \in [1, 30]$ and set sizes $d \in [50, 5000]$ produced results that were within pen point size equal to the results of Figure 1. So we conclude that $\bar{v}(k)$ is a robust estimation for an initial guess of the intra/inter distance ratio for data mining on typical input data sets containing about k clusters. If the number of clusters is not known, an initial value of $v = 0.33$ is a meaningful starting point for data mining (see the arrows in Figure 1). This value of v can be typically found in data sets containing from 3 to about 13 clusters. If the number of clusters is known to be k , v can be taken as $\bar{v}(k)$. If there are only one or two clusters in the data set $v_{\text{est}} = 0.7$ can be used. In case the minimum number of clusters in the data set is known, the lower of the 95% confidence interval boundaries is a good choice for v . If k is large ($k > 40$), the empirical Pareto Radius converges to the 1-percentile $pc(1)$ of the data distances.

5 Pareto probability density estimation

For one dimensional data the PDE can be scaled such that it's integral is one. The trapezoidal method on $(x_i, \text{NN}(x_i, r_p))$ can be used as the scaling factor. This leads to a probability density estimation PPDE. To measure the quality of the probability density estimation the mean of the sum of squared errors (MSSE) is used. Error is the difference between PPDE and the true probability density. Two sets containing 500 data points with $N1(0,1)$ and $N2(20,1)$ were generated. The union of these sets represents data points with two clusters of distance 20. As Pareto radius the 18th percentile of all distances was found in all experiments. For 200 such sets the density estimation using hyperspheres with the 1 ... 99 percentiles of distances were measured. Figure 2 shows the MSSE +/- the standard deviation of these experiments. Minimum error was encountered at the Pareto radius, i.e. the 18th percentile.

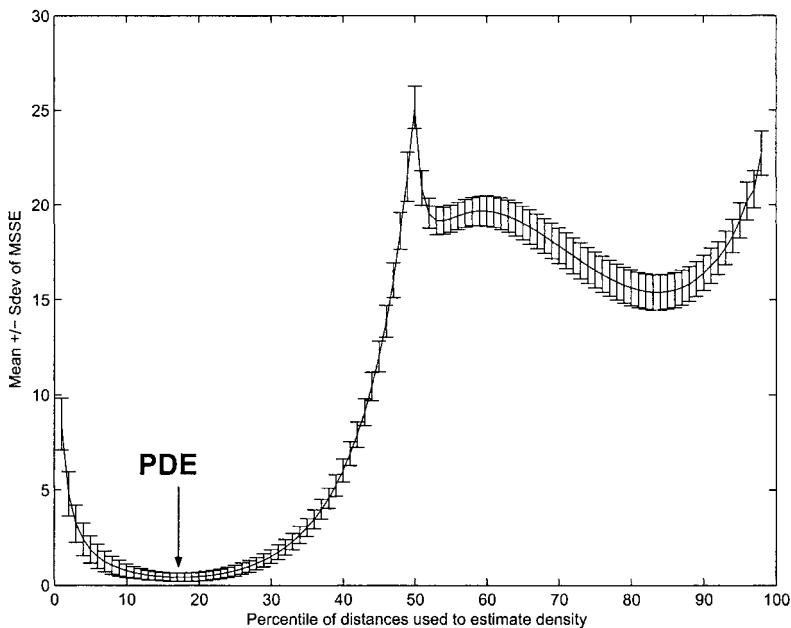


Fig. 2. Quality of density estimation using all distance percentiles

6 Robustness of the method

In this chapter we investigate the properties of PDE with regard to number of clusters, higher dimensionality and cluster overlap. The first question is

how well PDE generalizes to more than two clusters. For one to 50 clusters with a distance of the cluster centers of 20, 100 Experiments with 1000 data points were performed. A t-test was used to test the null hypothesis that Pareto Probability Density Estimation is different from the best radius to measure density. It turned out that for a 2% error level the null hypothesis - PPDE being different from true density - could be rejected for all these cluster numbers. The next question concerns the overlapping of clusters. In the two-cluster data set as described in the last chapter, the distances between the clusters were decreased stepwise form 20 to 1. For non overlapping clusters the Pareto radius is best. A t-test with alpha level of 5% rejects the hypothesis that the Pareto Radius is different from the best radius starting with a distance of 2.7 between the cluster centers. This corresponds to about 20% of common points. The distribution of distances is strongly influenced by inner cluster variances. In order to investigate the dependency of PDE on inner cluster variance 1000 data sets from a distribution with $N(0,1)$ and $N(20,s^2)$ were generated. For s^2 in the range of [0.2, 20], the best distance estimation was compared with PPDE. The experiments were repeated 100 times. It turned out the MSSE for PPDE differs less than 5% compared to the mean plus standard deviation of the best of the hypersphere density estimation. For small values of s^2 , i.e. $s^2 < 0.1$, PDE overestimates the true density. The next question concerns the dimensionality of the data. 100 experiments were performed with a two dimensional data set with 200 MMI distributed data points.

7 Stock picking: a difficult real world problem

Selecting stocks for a portfolio that have a high potential for rising stock prices is in general a difficult problem(O'Neil (1995)). Today there are more than 7000 stocks traded at the US stock exchanges. The problem is to pick between six and fifteen stocks that have a high potential for rising prizes (Maranjian (2002)). Each stock is characterized by 15 variables, from the company's fundamental data and the stock's performance (Deboeck and Ultsch (2002)). A stock is classified as a winner when the stocks price rises more than 10% within a period of 60 marked days compared to the mean price level of the 10th to 15th market day following the beginning of a quarter. It is classified a loser if the price falls more than 5%. For the first quarter (q1) of 2002 (January 1st to March 31st) the percentages of winners within the Pareto spheres of all data points were calculated. Using a binomial with 0.4% alpha level one single Pareto sphere contained a significantly high winner percentage. This Pareto sphere consisted of 15 stocks with 14 winners. The same Pareto sphere was used to select a portfolio in the second quarter (q2) of 2002 (April 1st to July 1st). Within this Pareto sphere were 11 stocks with 7 (64%) winners. An investment of equal value in each stock of this Pareto portfolio starting April 10th 2002 gave a steady increase in the total

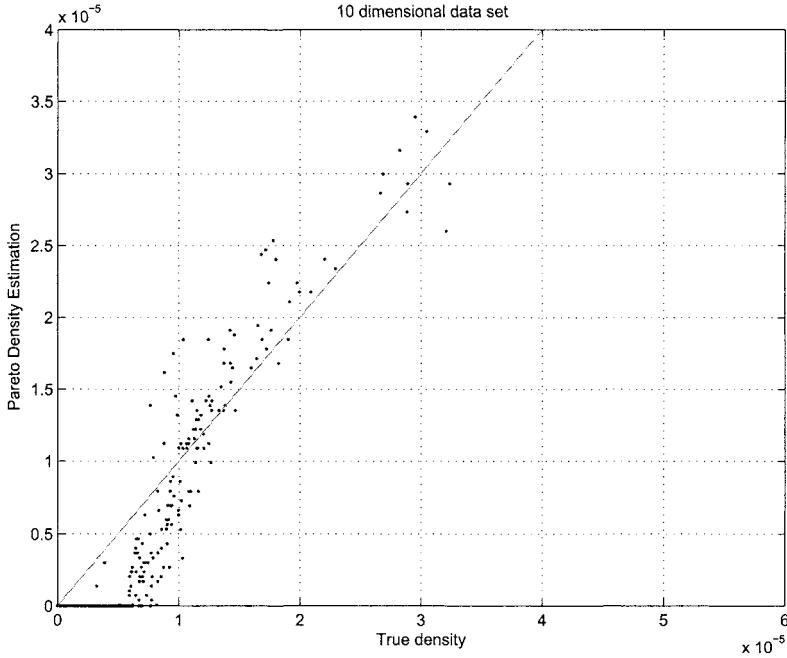


Fig. 3. PDE vs. true density for 10 dimensional data

value of the portfolio up to 108% until May 3rd. Here after the portfolio's value has a mean of 106% with a range of 104% to 108%. At the first of July the Portfolio closed with 5% gain compared to the -12% loss of the S&P500 index. Figure 4 shows the total portfolio value compared to the development of the S&P 500 index.

8 Discussion

Density based cluster algorithms for large high dimensional data as used for KDD impose special requirements on density estimation. For this kind of application efficiency of the density estimation is very important. It is well known that variable kernel methods estimate the density function with more precision than uniform density methods (Hall (1992)). The tuning of the radius according to the true density has, however, been proven to be intractable for large data sets (Devroye and Lugosi (2000)). The main disadvantage of uniform kernel methods is the overestimation of the true density in low density regions. With the focus on clustering thin regions are typically of no concern. They may even be regarded to contain "outliers". Fixing a global radius for density estimation has the advantage that the density estimation can be efficiently calculated using the number of points within the hypersphere. In practical applications clusters are often not well separated

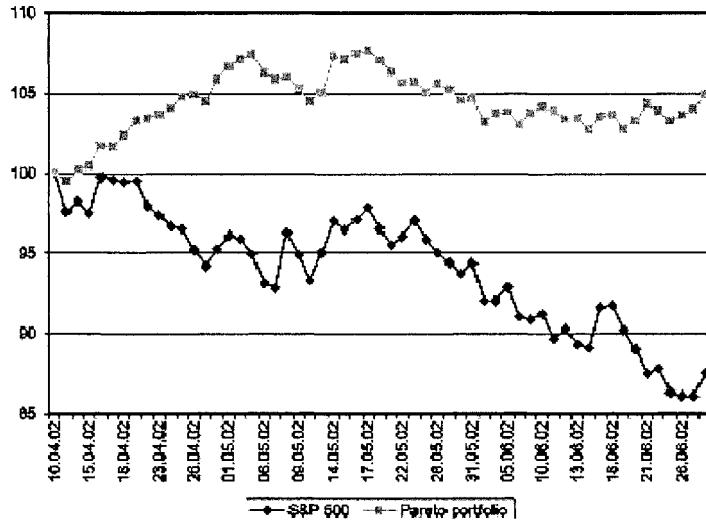


Fig. 4. PDE selected Portfolio vs. S&P 500 stock market index

but overlap to a certain extend. We showed here that PDE is optimal up to 20% of overlapping points of two clusters. This is a very high overlap. It is questionable that data with this percentage of overlap would be considered to be from different clusters. Therefore PDE seems to work even for a reasonable overlap in clusters. Since a distance measure is required for clustering, the density estimation should also make use of this measure. PDE is measured using the 18 percentile of the distances. For large data sets it might be infeasible to calculate the distance percentiles by first calculating the $O(n^2)$ distances between all data points. The URP function is rather flat around the minimum p_u (see Ultsch (2001)). In Ultsch (2001) it is demonstrated that set sizes in the range 16% to 24% give more than 90% of the information obtained using the optimal p_u . Furthermore the findings on Gaussian inner cluster structure presented in this work showed also, that the exact value of the Pareto radius is not really critical for these cluster structures. We expect therefore that a computational cheap estimation for the Pareto radius is sufficiently good enough for PDE. Sampling techniques combined with hashing methods may be used for an efficient calculation of the Pareto radius. The practicability of the method was demonstrated on a large database of high dimensionality ($d=15$). The search for a portfolio containing winning stocks is by itself difficult. In the quarter from which the Pareto sphere was constructed the prior winner probability was 40%. In the test quarter this probability dropped to 22%. The performance of indices showed also that the market situation became more difficult in the test quarter. Finding a portfolio with more than 44% winners would have been significant for a binomial

model with a 2% alpha level. The Pareto portfolio containing 64% winners surpassed this substantially.

9 Conclusion

The aim to discover new and useful knowledge in large sets of data has brought up the requirements for efficient cluster algorithms. Density based clustering methods have been proposed for this. Density estimation using hyperspheres with a global radius are a simple and efficient way to estimate data density. In this paper the radius for such density estimation is optimized according to an information theoretic criterion. The radius is adjusted such that in the average information optimal subsets are used for density calculations. This density estimation is called Pareto Density Estimation(PDE). PDE is a robust against a wide range of inner cluster variance, cluster numbers and dimensionality. This allows an efficient implementation of density based clustering algorithms for clustering in large databases with high dimensional data. PDE was tested on a difficult real world problem: the selection of stocks in a falling market. This problem consisted of a large number of high dimensional data. Although the marked situation deteriorated from the period when the parameters for PDE were constructed, the predicted portfolio substantially outperformed a broad marked index.

References

- DEBOECK, G.J. and ULTSCH, A. (2002): Picking Stocks with Emergent Self-Organizing Value Maps. In: M. Novak (Ed.): *Neural Networks World*, 10, 1-2, 203–216.
- DEVROYE, L. and LUGOSI, G. (1996): A universally acceptable smoothing factor for kernel density estimation. *Annals of Statistics*, 24, 2499–2512.
- DEVROYE, L. and LUGOSI, G. (1997): Non-asymptotic universal smoothing factors kernel complexity and Yatracos classes. *Annals of Stat.*, 25, 2626–2637.
- DEVROYE, L. and LUGOSI, G. (2000): Variable kernel estimates: on the impossibility of tuning the parameters. In: E. Giné and D. Mason (Eds.): *High-Dimensional Probability*. Springer-Verlag, New York.
- ESTER, M., KRIEGEL, H.-P., and SANDER, J. (1996): A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proc. 2nd Int. Conf. On Knowledge Discovery and Data Mining.
- HALL, P. (1992): On global properties of variable bandwidth density estimators. *Annals of Statistics*, 20, 762–778.
- HINNEBURG, A. and KEIM , D.A. (1998): An Efficient Approach to Clustering in Large Multimedia Databases with Noise, Proc. 4th Int.Conf. on Knowledge Discovery and Data Mining.
- MARANJIAN, S. (2002): The Best Number of Stocks, The Motley Fool, 26.
- O’NEIL, W.J. (1995): *How to make money in stocks*. Mc Gaw Hill, New York.
- SCOTT, D.W. (1992): Multivariate Density Estimation. Wiley-Interscience, New York.

- ULTSCH, A. (2001): Eine Begründung der Pareto 80/20 Regel und Grenzwerte für die ABC-Analyse, Technical Report Nr. 30, Department of Computer Science, University of Marburg.
- ULTSCH, A. (2003): Optimal density estimation in data containing clusters of unknown structure, Technical Report Nr. 34, Department of Computer Science, University of Marburg.
- XU, X., ESTER, M., KRIEGEL, H.-P., and SANDER, J. (1998): Distribution-Based Clustering Algorithm for Mining in Large Spatial Databases, Proc. Conf. on Data Engineering, 324–331.

Part II

Probability Models and Statistical Methods

Modelling the Claim Count with Poisson Regression and Negative Binomial Regression

Bartłomiej Bartoszewicz

Department of Econometrics,
Wrocław University of Economics,
ul. Komandorska 118/120, 53-345 Wrocław, Poland

Abstract. It is of interest for an insurance company to model the claim count or the claim severity in the presence of covariates or covariate factors describing the policyholders. In this paper the Poisson regression is presented (in the context of generalized linear models) and fitted to car insurance claims data. Since there are symptoms of over-dispersion in the data, the Poisson distribution of the response variable is replaced with the negative binomial distribution and another model is fitted to the number of claims. Finally, a method of testing the significance of the differences between groups of policyholders is shown and applied to the negative binomial regression.

1 Introduction

It is of interest for an insurance company to model the claim count (the number of claims arising from a block of policies) or the claim severity (expected claim size) when additional information about policyholders is available. Such information might take the form of continuous variables or classifying factors (variables taking only several values) which are likely to affect the claim count or claim severity. The goal here is to quantify and analyse the relationship between the risk (measured by one of variables related to claims indicated above) and covariates describing policyholders. Questions like “Are bigger cars more likely to be damaged in an accident?” or “Do young males cause significantly more accidents than young females?” arise and need an answer.

A statistical tool suitable for modelling and analysing such relationships is known under the label of generalized linear models. These models were widely used in actuarial practice, e.g. McCullagh and Nelder (1989) gave an example of modelling an average claim size in automobile insurance, they also modelled damage incidents to cargo-carrying vessels. Renshaw (1994) applied GLMs to both the claim count and the claim severity. Silva (1989) used these models to the Portuguese motor claims data. Mildenhall (1999) showed in detail how GLM correspond to classification rates with minimum bias. Holler et al. (1999) implemented GLMs to calculate classification relativities and applied it to credit insurance. Murphy et al. (2000) used GLMs for estimating the risk premium and price elasticity components of the customer value models. Guiahi (2001) discussed the interaction of parametric

loss distributions, deductibles, policy limits and rating variables in the context of fitting distributions to losses. The concept of generalized linear models will be briefly introduced in the second section of this paper. In the next section the car insurance claims data used here as an example will be presented. This section will also show the parameter estimates of the Poisson regression fitted to the claims data and symptoms of over-dispersion. In the fourth section the negative binomial regression will be introduced. Advantages as well as disadvantages of the model will be indicated. The parameter estimates and indicators of correctly estimated variance will also be given. Finally, the method of testing statistical significance of differences among groups of policyholders using an alternative design matrix and the well known likelihood ratio test will be demonstrated.

2 Generalized linear models

The term “generalized linear models” is due to Nelder and Wedderburn (1972). These two authors showed, how a number of well known statistical models share several properties, such as linearity or a method of calculating parameter estimates. Generalized linear models include as special cases classical linear regression, logit and probit models, analysis of variance, multinomial response models and log-linear models.

Basic assumption of GLMs is that the observations of the response variable are independent (or at least uncorrelated). Second assumption is that there is only one error term in the model¹. What makes GLMs more flexible than classical regression analysis is that the assumption of a constant variance of errors across all observations is replaced by that the variance is now a known function of the expected value. GLMs consist of three components (see McCullagh and Nelder (1989)):

1. **The systematic component** - the linear predictor of the form:

$$\eta = X\beta \quad (1)$$

The linear predictor is a product of the covariates in X and a vector of the unknown parameters β .

2. **The random component** - the response variable Y_i has a distribution in the exponential family of the form:

$$f_Y(y; \theta, \phi) = \exp \left(\frac{y\theta - b(\theta)}{a(\theta)} + c(y, \theta) \right) \quad (2)$$

$$E(Y_i) = \mu = b'(\theta), \text{var}(Y_i) = b''(\theta)a(\phi)$$

¹ Generalized Linear Mixed Models incorporate an additional random term in the linear predictor, but GLMMs are outside the GLM framework.

where $a(\cdot)$, $b(\cdot)$, $c(\cdot)$ are some specific functions, ϕ is a known dispersion parameter assumed constant for all Y_i , θ is a canonical parameter and $b'(\theta)$, $b''(\theta)$ are derivatives with respect to θ .

The most important distributions in the exponential family are the normal, gamma and inverse Gaussian for continuous cases and two discrete distributions: binomial and the Poisson.

3. **The link function** is of the form:

$$\eta = g(\mu) \Leftrightarrow \mu = g^{-1}(\eta) \quad (3)$$

where $g(\cdot)$ and its inverse $g^{-1}(\cdot)$ are monotonic differentiable functions. The link function is a connection between the random component and the systematic component. It can be thought of as the way the covariates (through the linear predictor η) govern the behaviour of the response variable through its expected value μ .

McCullagh and Nelder (1989) argue that the link function should be chosen in order to make the effect of covariates additive on an appropriate scale.

The maximum likelihood estimates for β in the linear predictor can be obtained by iterative weighted least squares (see McCullagh and Nelder (1989)). This property of GLMs makes the estimation process easy to implement in any statistical package.

3 Car insurance claims data

Generalized linear models described above will be applied to car insurance claims data taken from McCullagh and Nelder (1989), section 8.4.1, page 296. The data consist of the number claims for damage to policyholder's car in Britain, 1975 in each of 96 groups of policyholders. Every group is defined by appropriate levels of the three classifying factors:

1. policyholder's age (PA) taking 8 levels (17–20, 21–24, 25–29, 30–34, 35–39, 40–49, 50–59, 60+);
2. car group (CG) taking 4 levels (A, B, C, D);
3. car age (CA) taking 3 levels (0–3, 4–7, 8+).

The data does not include the number of policies in each group, but this doesn't affect the conclusions since the goal of this paper is to present methods of modelling and analysing this kind of relationships, not the exact results of the analysis of this particular data set. Of course, the insurance company has information on the number of exposure units in each block of policies. In order to conduct a real-life analysis this information should be incorporated in the model as an offset.

Since the number of claims takes non-negative integer values, the Poisson distribution for the response variable seems to be right. The random

component is given by the probability density function:

$$P(Y = y_0) = \frac{e^{-\mu} \mu^{y_0}}{y_0!} \quad (4)$$

$$\text{E}(Y) = \text{var}(Y) = \mu$$

The linear predictor η is a product of matrix X containing 13 dummy variables representing the three rating factors (no interactions between factors are included) and vector β of 13 unknown parameters.

The link function is

$$\eta = \log(\mu) \Leftrightarrow \mu = \exp(\eta). \quad (5)$$

$\hat{\beta}$	(1)	(2)	(3)	(4)	(5)	(6)
1.76 1.42 2.35 2.52 2.58 3.22 3.00 2.64 1.12 0.76 -0.17 -0.15 -1.19	1.76	1.98	1.98	1.87	group "0"	PA: 17–20; CG: A; CA: 0–3
	1.42	1.44	1.44	1.45		21–24
	2.35	2.40	0.95	1.09		25–29
	2.52	2.56	0.16	⁻²	policy-holder's age (PA)	30–34
	2.58	2.61	0.06	⁻²		35–39
	3.22	3.24	0.63	0.61		40–49
	3.00	3.01	-0.23	⁻²		50–59
	2.64	2.75	-0.26	-0.38		60+
	1.12	0.91	0.91	0.92	car	B
	0.76	0.44	-0.47	-0.46	group (CG)	C
	-0.17	-0.53	-0.97	-0.96		D
	-0.15	-0.17	-0.17	⁻²	car	4–7
	-1.19	-1.28	-1.11	-1.20	age (CA)	8+

Table 1. Estimates of β for Poisson regression and negative binomial regression.

The parameter estimates are given in Table 1, column (1). The first parameter corresponds to the basic group (PA 17–20, CG A, CA 0–3). Other parameters correspond to differences for other levels of every factor to the basic group. The log likelihood is 696.11. The dispersion parameter estimate is 10.196. This indicates the strong over-dispersion - the true variance significantly exceeds the expected value.

Another method of detecting over-dispersion is to calculate for each observation the "tail probability" of the form:

$$\hat{P}_{tail,i} = \begin{cases} \hat{P}(Y \leq y_i) & \text{for } y_i \leq \hat{\mu}_i \\ \hat{P}(Y \geq y_i) & \text{for } y_i < \hat{\mu}_i \end{cases} \quad (6)$$

which can be interpreted as an estimated probability of observing the number of claims that is more distant from the fitted expected value than actually

² Parameter restricted to 0.

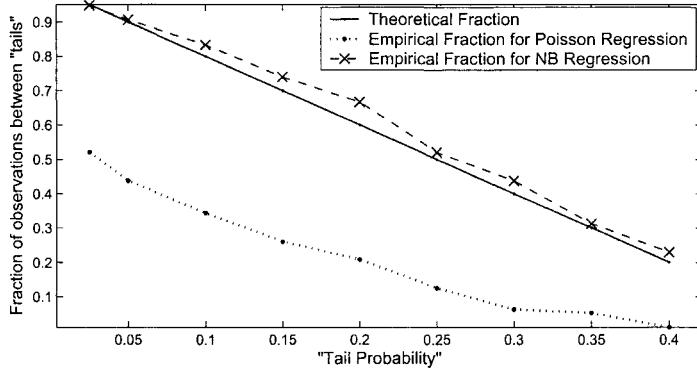


Fig. 1. Theoretical and empirical fractions of observations between “tails” for Poisson regression and negative binomial regression.

observed. Leaving out all the observations with the “tail” lower than, say, 5%, should result in getting about 90% of observations between “tails”. Figure 1 shows the theoretical (solid line) and empirical (dotted line) fractions of observations between the “tails” for Poisson regression, given the “tail probability” from 0.025 to 0.4. For the 5% “tails” there are about 45% of observations between “tails” instead of 90%. The figure clearly indicates strong over-dispersion.

4 Negative binomial regression

In order to relax the assumption of variance equal to μ the Poisson distribution will be replaced with more a flexible distribution. It seems reasonable to use the negative binomial distribution:

$$P(Y = y_0) = \frac{\frac{1}{\gamma} \left(\frac{1}{\gamma} + 1\right) \cdots \left(\frac{1}{\gamma} + y_0 - 1\right)}{y_0!} \left(\frac{\mu\gamma}{1 + \mu\gamma}\right)^{y_0} \left(\frac{1}{1 + \mu\gamma}\right)^{1/\gamma} \quad (7)$$

$$\text{E}(Y) = \mu, \text{var}(Y) = \mu + \gamma\mu^2$$

where γ is an unknown parameter assumed constant for all observations.

The linear predictor η and the log link function (5) used for the Poisson regression remain unchanged.

It is possible to use negative binomial distribution within the GLM framework (see McCullagh and Nelder (1989)) with the canonical link and variance:

$$\eta = \log \left(\frac{\mu}{\mu + k} \right), \text{var}(Y) = \mu + \frac{1}{k}\mu^2,$$

but the link function depends on parameter k , which has to be assumed known. The linear predictor as a function of the constant parameter of vari-

ance function makes the use of the model problematic. It is possible to estimate k along with β from the data (as shown in this paper), but this takes the negative binomial distribution outside the exponential family and thus takes the model outside GLM framework.

Negative binomial regression has several advantages over Poisson regression. First, the claim count is much more likely to have a negative binomial distribution than a Poisson distribution. The true variance of the number of claims in automobile insurance portfolios is usually higher than the expected value, mainly because of the heterogeneity of policyholders. An additional parameter included in variance makes the model much more flexible. Although γ is constant across all cells, it is estimated from the sample and shifts the variance to appropriate level.

Unfortunately, there are also disadvantages of the negative binomial regression. Since the model with log link is not included in the GLM framework, parameter estimates need to be obtained by numerical maximization of the log likelihood function which contains terms depending only on β or γ as well as terms depending on β and γ together. The normal equations for the negative binomial regression are non-linear and cannot be solved directly.

Table 1, column (2) gives parameter estimates for negative binomial regression. Again, the first element of $\hat{\beta}$ corresponds to basic group (PA 17–20, CG A, CA 0–3) whereas other parameter estimates correspond to other levels of the classifying factors. The estimate for γ is $\hat{\gamma} = 0.1341$. The log likelihood is 426.71 which is by about 270 better than for the Poisson regression. Introduction of additional parameter γ has significantly improved the fit. Figure 1 shows the theoretical (solid line) and empirical (dashed line) fractions of observations between the tails for the negative binomial regression. The two lines of theoretical and empirical fractions are close to each other which indicates that the variance is estimated correctly.

5 Testing statistical significance of differences among groups of policies

The fifth section of the paper shows a method of testing statistical significance of differences among groups of policyholders. The method is based on the well known likelihood ratio test and an alternative way of constructing design matrix of covariate factors.

The “classical” design matrix (without interactions between factors) consists of one column filled with “1”s and several columns filled with “0”s and “1”s. The first column corresponds to the intercept which represents the level of the response variable for an arbitrarily chosen basic group. Other columns correspond to parameters indicating the difference of the response variable for a particular group to the basic group. Table 3 shows an alternative design matrix of covariate factors (without interactions) for a simple case of two factors A and B taking 3 levels each. Boxes indicate “0”s converted into

basic group (A1, B1), β_0	A2, β_1	A3, β_2	B2, β_3	B3, β_4
1	0	0	0	0
1	1	0	0	0
1		1	0	0
1	0	0	1	0
1	1	0	1	0
1		1	1	0
1	0	0		1
1	1	0		1
1		1		1

Table 2. Likelihood ratio test statistics and p-value for restrictions on single parameters to 0 for negative binomial regression.

Restriction on	β_4	β_3	β_6	β_7	β_{11}	β_5
Likelihood Ratio test statistics	0.1	1.0	2.2	2.7	2.9	14.5
p-value	0.723	0.320	0.141	0.101	0.089	< 0.001

Table 3. An alternative design matrix for two classifying factors.

“1”s. This operation changes the interpretation of parameters - now each parameter (except for the intercept) is a difference to the previous level of appropriate factor: β_2 refers to the difference between A1 and A3 in the classical design whereas in the alternative design β_2 refers to difference between A2 and A3. Thus, the difference from A1 to A3 is expressed as $\beta_1 + \beta_2$. Testing statistical significance of a parameter is asking whether the difference between two particular levels of a factor is significant.

The method described above was applied to the negative binomial regression. Table 1, column (3) gives parameter estimates with the alternative design matrix ($\hat{\eta} = 0.1341$). Table 5 gives the likelihood ratio test statistics and p-values for models with one parameter restricted to 0. There were 12 parameters tested (the intercept doesn't need to be tested) and 6 of them with the highest p-value are given in Table 5. The next step was to test restrictions on pairs of parameters, then groups of three, four and five parameters. The final model is the model with restrictions $\beta_3 = \beta_4 = \beta_6 = \beta_{11} = 0$ (parameter estimates are given in Table 1, column (4), $\hat{\eta} = 0.1473$). Likelihood ratio test statistics is 6.88 and has the χ^2 distribution with 4 degrees of freedom which yields a p-value of 0.142. The chosen model was the best model in terms of p-value among all models with 4 restrictions. The best model with five restrictions ($\beta_3 = \beta_4 = \beta_6 = \beta_7 = \beta_{11} = 0$) has the test statistics 13.41 and p-value 0.020 (χ^2 distribution with 5 degrees of freedom).

The testing procedure shown above led to the conclusion that policyholders aged 25–29, 30–34 and 35–39 do not show significant differences and the groups were merged without losing accuracy of the fitted expected values of the number of claims. Restriction on β_6 results in merging drivers aged 40–49

and 50–59 into one block. Restriction on β_{11} means that the car age group 4–7 can be merged with the previous car age group which in this case is the basic group (CA 0–3).

The same procedure was applied to the Poisson regression. In this case only β_4 is statistically insignificant (LR test statistics 2.54, p-value 0.111), for other parameters the p-value is less than 0.001. This is caused by overdispersion and leads to false conclusions about merging groups of policies.

6 Conclusions

The alternative design matrix of covariate factors together with the likelihood ratio test make it very easy to test statistical significance of differences between groups of objects of interest when groups are defined by classifying factors. The procedure can be used whenever modelling claim count or claim severity. It can be applied to generalized linear models as well as to other useful models not included in GLM framework.

The negative binomial regression with log link might be of interest when modelling claim count in the presence of covariates. Although the estimation of parameters is a burden, the model provides flexible adjustment for variance of the response variable, especially in comparison to the Poisson regression.

Acknowledgements: The author is grateful to two anonymous referees for their remarks and comments that led to significant improvements.

References

- GUIAHI, F. (2001): Fitting to Loss Distributions with Emphasis on Rating Variables. *Casualty Actuarial Society Forum*, Winter 2001, 133–174.
- HOLLER, K.D., SOMMER, D., and TRAHAIR, G. (1999): Something Old, Something New in Classification Ratemaking With a Novel Use of GLMs for Credit Insurance. *Casualty Actuarial Society Forum*, Winter 1999, 31–84.
- MURPHY, K.P., BROCKMAN, M.J., and LEE, P.K.W. (2000): Using Generalized Linear Models to Build Dynamic Pricing Systems for Personal Lines Insurance. *Casualty Actuarial Society Forum*, Winter 2000, 107–140.
- MCCULLAGH, P. and NELDER, J.A. (1989): *Generalized Linear Models*. 2nd Edition, Chapman and Hall, London.
- MILDENHALL, S. (1999): A Systematic Relationship Between Minimum Bias and Generalized Linear Models. *PCAS LXXXVI*, 393–487.
- NELDER, J.A. and VERRALL, R.J. (1997): Credibility Theory and Generalized Linear Models. *ASTIN Bulletin*, 27, 71–82.
- NELDER, J.A. and WEDDERBURN, R.W.M. (1972): Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, 135, 370–384.
- RENSHAW, A.E. (1994): Modelling the Claims Process in the Presence of Covariates. *ASTIN Bulletin*, 24, 265–286.
- SILVA, A.J.E. (1989): An Application of Generalized Linear Models to Portuguese Motor Insurance. *Proceedings XXI ASTIN Colloquium*, 633.

Chemical Balance Weighing Design with Different Variances of Errors

Bronisław Ceranka, Małgorzata Graczyk

Department of Mathematical and Statistical Methods,
Agricultural University of Poznań,
ul. Wojska Polskiego 28, 60-637 Poznań, Poland
e-mail: bronicer@owl.au.poznan.pl
e-mail: magra@owl.au.poznan.pl

Abstract. The paper is studying the estimation problem of individual weights of objects using a chemical balance weighing design under the restriction on the number of times in which each object is weighed. We assume that errors are uncorrelated with different variances. The necessary and sufficient condition under which the lower bound of variance of each of the estimated weights is attained is given. For a new construction method of the optimum chemical balance weighing design we use the incidence matrices of the balanced incomplete block designs and the ternary balanced block designs.

1 Introduction

Let us consider the class $\Phi_{n \times p, m}(-1, 0, 1)$ of the $n \times p$ matrices \mathbf{X} with elements equal to $-1, 0$ or 1 , where m is the maximum number of elements equal to -1 and 1 in each column of the matrix \mathbf{X} . The matrices belonging to this class are the design matrices of the chemical balance weighing designs. We can write a suitable model in the form

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \quad (1)$$

where \mathbf{y} is a $n \times 1$ random vector of the recorded results of weighings, \mathbf{w} is a $p \times 1$ column vector representing unknown weights of objects and \mathbf{e} is a $n \times 1$ random vector of errors. We assume that $E(\mathbf{e}) = \mathbf{0}_n$ and the errors are uncorrelated and they have different variances, i.e. $\text{Var}(\mathbf{e}) = \sigma^2 \mathbf{G}$, where $\mathbf{0}_n$ is the $n \times 1$ column vector of zeros, \mathbf{G} is a $n \times n$ positive definite diagonal matrix of known elements.

For estimating individual unknown weights of the objects we can use the normal equations

$$\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\hat{\mathbf{w}} = \mathbf{X}'\mathbf{G}^{-1}\mathbf{y}, \quad (2)$$

where $\hat{\mathbf{w}}$ is the vector of the weights estimated by the least squares method.

The chemical balance weighing design is singular or nonsingular depending on whether the matrix $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}$ is singular or nonsingular, respectively. It is obvious that because of the assumption connected with the matrix \mathbf{G} , the

matrix $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}$ is nonsingular if and only if the matrix $\mathbf{X}'\mathbf{X}$ is nonsingular, i.e. if and only if \mathbf{X} is of full column rank ($= p$).

If $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}$ is nonsingular, the least squares estimator of \mathbf{w} is given in the form

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{G}^{-1}\mathbf{y} \quad (3)$$

and the variance - covariance matrix of $\hat{\mathbf{w}}$ is given by formula

$$\text{Var}(\hat{\mathbf{w}}) = \sigma^2(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X})^{-1}. \quad (4)$$

In the case $\mathbf{G} = \mathbf{I}_n$, Hotelling (1944) has studied some problems connected with chemical balance weighing designs. He has shown that for the chemical balance weighing design the minimum attainable variance for each of the estimated weights is σ^2/n . He proved the theorem that each of the variances of the estimated weights attains the lower bound if and only if $\mathbf{X}'\mathbf{X} = n\mathbf{I}_p$. This design is called the optimum chemical balance weighing design. It implies that for the optimum chemical balance weighing design the elements of matrix \mathbf{X} are -1 and 1 , only. In this case, several methods for constructing the optimum chemical balance weighing designs are available in Raghavarao (1971) and Banerjee (1975).

Katulska (1989) has shown that the minimum attainable variance for each of the estimated weights for a chemical balance weighing design with positive definite diagonal variance - covariance matrix of errors $\sigma^2\mathbf{G}$ is $\sigma^2/\text{tr}(\mathbf{G}^{-1})$. In the same paper she has shown that each of the variance of the estimated weights attains the minimum if and only if $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X} = \text{tr}(\mathbf{G}^{-1})\mathbf{I}_p$. This condition implies that elements of the matrix \mathbf{X} of an optimum chemical balance weighing design are equal -1 or 1 only. In this case some methods of construction of the optimum chemical balance weighing designs are given in the literature.

2 Variance limit of estimated weights

Let us suppose, there are t kinds of uncorrelated chemical balances and they are with different precision, one with usual precision, the other with higher and one with the highest. n_1, n_2, \dots, n_t are numbers of times in which the respectively balance is used. In this case the variance - covariance matrix of errors is $\sigma^2\mathbf{G}$ where the matrix \mathbf{G} is given by

$$\mathbf{G} = \begin{bmatrix} \frac{1}{a_1}\mathbf{I}_{n_1} & \mathbf{0}_{n_1}\mathbf{0}'_{n_2} & \dots & \mathbf{0}_{n_1}\mathbf{0}'_{n_t} \\ \mathbf{0}_{n_2}\mathbf{0}'_{n_1} & \frac{1}{a_2}\mathbf{I}_{n_2} & \dots & \mathbf{0}_{n_2}\mathbf{0}'_{n_t} \\ \dots & \dots & \dots & \dots \\ \mathbf{0}_{n_t}\mathbf{0}'_{n_1} & \mathbf{0}_{n_t}\mathbf{0}'_{n_2} & \dots & \frac{1}{a_t}\mathbf{I}_{n_t} \end{bmatrix} \quad (5)$$

where $\sum_{h=1}^t n_h = n$, $a_h > 0$ and \mathbf{I}_{n_h} is the $n_h \times n_h$ identity matrix, $h = 1, 2, \dots, t$. In other words, we determine the optimality in the class

$\Phi_{n \times p, m}(-1, 0, 1)$ with over restriction for a given a_1, a_2, \dots, a_t and n_1, n_2, \dots, n_t . Suppose further that the matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ is partitioned correspondingly to the matrix \mathbf{G} , i.e.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_t \end{bmatrix}. \quad (6)$$

Ceranka and Graczyk (2003) proved the following theorems

Theorem 2.1 For any nonsingular chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by the form in (2.2) and with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where the matrix \mathbf{G} is given by (2.1), the variance of \hat{w}_j for any j , $j = 1, 2, \dots, p$, cannot be less than σ^2/q , where $q = \sum_{h=1}^t a_h m_h$, $m_h = \max\{m_{h_1}, m_{h_2}, \dots, m_{h_p}\}$.

Definition 2.1 A nonsingular chemical balance weighing design is said to be optimal for estimating individual weights of objects if the variances of their estimators attain the lower bound given in the theorem 2.1, i.e., if

$$Var(\hat{w}_j) = \frac{\sigma^2}{q}, \quad j = 1, 2, \dots, p.$$

Theorem 2.2 For any positive definite diagonal matrix \mathbf{G} given by the form in (2.1) and any matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by the form (2.2), under a nonsingular chemical balance weighing design, each of the variances of the estimated weights attains the minimum if and only if

$$\mathbf{X}' \mathbf{G}^{-1} \mathbf{X} = q \mathbf{I}_p. \quad (7)$$

For $\mathbf{G} = \mathbf{I}_n$, Theorems 2.1 and 2.2 were originally proved in Ceranka and Graczyk (2001).

It is obvious that we have many interesting possibilities of patterns of \mathbf{G} . The construction of the optimum chemical balance weighing design with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$ for each of the forms of \mathbf{G} must be investigated separately.

Now, we assume that the matrix \mathbf{G} is given in the simplest form $\mathbf{G} \neq \mathbf{I}_n$, i.e.

$$\mathbf{G} = \begin{bmatrix} \frac{1}{a} \mathbf{I}_{n_1} & \mathbf{0}_{n_1} \mathbf{0}'_{n_2} \\ \mathbf{0}_{n_2} \mathbf{0}'_{n_1} & \mathbf{I}_{n_2} \end{bmatrix}, \quad (8)$$

where $a > 0$, $n = n_1 + n_2$, n_1 and n_2 are the numbers of times in which the respectively balance is used. Let the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ be partitioned corresponding to the matrix \mathbf{G} , i.e.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}. \quad (9)$$

Thus, it follows from Theorem 2.2 that the chemical balance weighing design with $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by (2.5) and the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is of the form (2.4), is optimal if and only if

$$a\mathbf{X}'_1\mathbf{X}_1 + \mathbf{X}'_2\mathbf{X}_2 = (am_1 + m_2)\mathbf{I}_p. \quad (10)$$

In the next sections we will construct the matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ of the optimum chemical balance weighing design with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is given by (2.4). It is based on the incidence matrices of balanced incomplete block designs and ternary balanced block designs.

3 Balanced block designs

A balanced incomplete block design is an arrangement of v treatments in b blocks, each of size k , in such a way, that each treatment occurs at most once in each block, occurs in exactly r blocks and every pair of treatments occurs together in exactly λ blocks. The integers v, b, r, k, λ are called the parameters of the balanced incomplete block design. Let \mathbf{N} be the incidence matrix of balanced incomplete block design. It is straightforward to verify that

$$\begin{aligned} vr &= bk, \\ \lambda(v-1) &= r(k-1), \\ \mathbf{NN}' &= (r-\lambda)\mathbf{I}_v + \lambda\mathbf{1}_v\mathbf{1}'_v, \end{aligned}$$

where $\mathbf{1}_v$ is the $v \times 1$ vector of units.

A ternary balanced block design is defined as the design consisting of b blocks, each of size k , chosen from a set of objects of size v , in such a way that each of the v treatments occurs r times altogether and 0, 1 or 2 times in each block, (2 appears at least once) and each of the distinct pairs appears λ times. Any ternary balanced block design is regular, that is, each treatment occurs alone in ρ_1 blocks and is repeated two times in ρ_2 blocks, where ρ_1 and ρ_2 are constant for the design. Let \mathbf{N} be the incidence matrix of the ternary balanced block design. It is straightforward to verify that

$$\begin{aligned} vr &= bk, \\ r &= \rho_1 + 2\rho_2, \\ \lambda(v-1) &= \rho_1(k-1) + 2\rho_2(k-2) = r(k-1) - 2\rho_2, \\ \mathbf{NN}' &= (\rho_1 + 4\rho_2 - \lambda)\mathbf{I}_v + \lambda\mathbf{1}_v\mathbf{1}'_v = (r + 2\rho_2 - \lambda)\mathbf{I}_v + \lambda\mathbf{1}_v\mathbf{1}'_v. \end{aligned}$$

4 Construction of the design matrix

Let \mathbf{N}_1 be the incidence matrix of the balanced incomplete block design with parameters $v, b_1, r_1, k_1, \lambda_1$ and \mathbf{N}_2 be the incidence matrix of the ternary balanced block design with parameters $v, b_2, r_2, k_2, \lambda_2, \rho_{12}, \rho_{22}$. From the matrices \mathbf{N}_1 and \mathbf{N}_2 we build the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ of the chemical balance weighing design in the form (2.5). It is given by formula

$$\mathbf{X} = \begin{bmatrix} 2\mathbf{N}'_1 - \mathbf{1}_{b_1}\mathbf{1}'_v \\ \mathbf{N}'_2 - \mathbf{1}_{b_2}\mathbf{1}'_v \end{bmatrix}. \quad (11)$$

In this design we can measure weights of p objects. Each object is weighed $b_1 + b_2 - \rho_{12}$ times in $n = b_1 + b_2$ measurement operations.

Lemma 4.1 A chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by formula (4.1) and with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is of the form (2.4), is nonsingular if and only if

$$2k_1 \neq k_2$$

or

$$2k_1 = k_2 \neq v.$$

Proof: Because \mathbf{G} is a positive definite diagonal matrix then the chemical balance weighing design is nonsingular if and only if the matrix $\mathbf{X}'\mathbf{X}$ is non-singular. Hence for the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ in the form (4.1) we have

$$\mathbf{X}'\mathbf{X} = [4(r_1 - \lambda_1) + r_2 + 2\rho_{22} - \lambda_2]\mathbf{I}_v + [b_1 - 4(r_1 - \lambda_1) + b_2 + \lambda_2 - 2r_2]\mathbf{1}_v\mathbf{1}'_v.$$

$$\text{Thus } \det(\mathbf{X}'\mathbf{X}) = [4(r_1 - \lambda_1) + r_2 + 2\rho_{22} - \lambda_2]^{v-1}.$$

$$\cdot \frac{1}{k_1 k_2} [v^2(r_1 k_2 + r_2 k_1) - 2vk_1 k_2(2r_1 + r_2) + k_1 k_2(4r_1 k_1 + r_2 k_2)].$$

It is obvious, that when $4(r_1 - \lambda_1) + r_2 + 2\rho_{22} - \lambda_2 > 0$ then $\det(\mathbf{X}'\mathbf{X}) \neq 0$ if and only if $2k_1 \neq k_2$ or $2k_1 = k_2 \neq v$. Hence the thesis.

Theorem 4.1 Any nonsingular chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by (4.1) and with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is of the form (2.4), is optimal for estimation unknown measurements of objects if and only if

$$a[b_1 - 4(r_1 - \lambda_1)] + [b_2 + \lambda_2 - 2r_2] = 0. \quad (12)$$

Proof. It is the consequence of the relation $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X} = q\mathbf{I}_p$ and equality

$$\mathbf{X}'\mathbf{G}^{-1}\mathbf{X} = [4a(r_1 - \lambda_1) + r_2 + 2\rho_{22} - \lambda_2]\mathbf{I}_v + [ab_1 - 4a(r_1 - \lambda_1) + b_2 + \lambda_2 - 2r_2]\mathbf{1}_v\mathbf{1}'_v.$$

If the chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by (4.1) with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is of the form (2.4), is optimal then

$$\text{Var}(\hat{w}_j) = \frac{\sigma^2}{ab_1 + b_2 - \rho_{12}}, \quad j = 1, 2, \dots, p.$$

Let us consider the equality (4.2). If the parameters of the balanced incomplete block design satisfy the equality $b_1 - 4(r_1 - \lambda_1) = 0$ then they belong

to the A - family given in Raghavarao (1971). Parameters of the ternary balanced block design satisfying the equality $b_2 + \lambda_2 - 2r_2 = 0$ are given in Billington and Robinson (1983). Thus we receive

Theorem 4.2 If the parameters of the balanced incomplete block design satisfy the equality $b_1 - 4(r_1 - \lambda_1) = 0$ and the parameters of the ternary balanced block design satisfy the equality $b_2 = 2r_2 - \lambda_2$ then the optimum chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by (4.1) with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is of the form (2.4), exists for each $a > 0$.

Corollary 4.1 The existence of the balanced incomplete block design and the ternary balanced block design with parameters

- (i) $v = b_1 = 4g^2, r_1 = k_1 = g(2g + 1), \lambda_1 = g(g + 1)$ and $v = 4g^2, b_2 = 4g^2s, r_2 = 2s(2g^2 - 1), k_2 = 2(2g^2 - 1), \lambda_2 = 4s(g^2 - 1), \rho_{12} = 4s(g^2 - 1), \rho_{22} = s$, where $g = 2, 3, \dots, s = 1, 2, \dots,$
- (ii) $v = b_1 = 4g^2, r_1 = k_1 = g(2g - 1), \lambda_1 = g(g - 1)$ and $v = 4g^2, b_2 = 4g^2s, r_2 = 2s(2g^2 - 1), k_2 = 2(2g^2 - 1), \lambda_2 = 4s(g^2 - 1), \rho_{12} = 4s(g^2 - 1), \rho_{22} = s$, where $g = 2, 3, \dots, s = 1, 2, \dots,$
- (iii) $v = 4g^2, b_1 = 4gt, r_1 = t(2g - 1), k_1 = g(2g - 1), \lambda_1 = t(g - 1)$ and $v = 4g^2, b_2 = 4g^2s, r_2 = 2s(2g^2 - 1), k_2 = 2(2g^2 - 1), \lambda_2 = 4s(g^2 - 1), \rho_{12} = 4s(g^2 - 1), \rho_{22} = s$, where $g, t = 2, 3, \dots, s = 1, 2, \dots, t \geq g,$
- (iv) $v = (2g + 1)^2, b_1 = 4t(2g + 1), r_1 = 4gt, k_1 = g(2g + 1), \lambda_1 = t(2g - 1)$ and $v = (2g + 1)^2, b_2 = s(2g + 1)^2, r_2 = s(4g^2 + 4g - 1), k_2 = 4g^2 + 4g - 1, \lambda_2 = s(4g^2 + 4g - 3), \rho_{12} = s(4g^2 + 4g - 3), \rho_{22} = s$, where $g, t = 2, 3, \dots, s = 1, 2, \dots, 4t \geq 2g + 1,$
- (v) $v = b_1 = 4g^2, r_1 = k_1 = g(2g + 1), \lambda_1 = g(g + 1)$ and $v = 4g^2, b_2 = 4g^2s, r_2 = s(4g^2 - 3), k_2 = g^2 - 3, \lambda_2 = 2s(2g^2 - 3), \rho_{12} = s(4g^2 - 9), \rho_{22} = 3s$, where $g = 2, 3, \dots, s = 1, 2, \dots,$
- (vi) $v = b_1 = 4g^2, r_1 = k_1 = g(2g - 1), \lambda_1 = g(g - 1)$ and $v = 4g^2, b_2 = 4g^2s, r_2 = s(4g^2 - 3), k_2 = 4g^2 - 3, \lambda_2 = 2s(2g^2 - 3), \rho_{12} = s(4g^2 - 9), \rho_{22} = 3s$, where $g = 2, 3, \dots, s = 1, 2, \dots,$
- (vii) $v = 4g^2, b_1 = 4gt, r_1 = t(2g - 1), k_1 = g(2g - 1), \lambda_1 = t(g - 1)$ and $v = 4g^2, b_2 = 4g^2s, r_2 = s(4g^2 - 3), k_2 = 4g^2 - 3, \lambda_2 = 2s(2g^2 - 3), \rho_{12} = s(4g^2 - 9), \rho_{22} = 3s$ where $g, t = 2, 3, \dots, s = 1, 2, \dots, t \geq g,$
- (viii) $v = (2g + 1)^2, b_1 = 4t(2g + 1), r_1 = 4gt, k_1 = g(2g + 1), \lambda_1 = t(2g - 1)$ and $v = (2g + 1)^2, b_2 = s(2g + 1)^2, r_2 = 2s(2g^2 + 2g - 1), k_2 = 2(2g^2 + 2g - 1), \lambda_2 = s(4g^2 + 4g - 5), \rho_{12} = 4s(g^2 + g - 2), \rho_{22} = 3s$, where $g, t = 2, 3, \dots, s = 1, 2, \dots, 4t \geq 2g + 1,$
- (ix) $v = b_1 = 4g^2, r_1 = k_1 = g(2g + 1), \lambda_1 = g(g + 1)$ and $v = 4g^2, b_2 = 4g^2s, r_2 = 4s(g^2 - 1), k_2 = 4(g^2 - 1), \lambda_2 = 4s(g^2 - 2), \rho_{12} = 4s(g^2 - 4), \rho_{22} = 6s$, where $g = 3, 4, \dots, s = 1, 2, \dots,$
- (x) $v = b_1 = 4g^2, r_1 = k_1 = g(2g - 1), \lambda_1 = g(g - 1)$ and $v = 4g^2, b_2 = 4g^2s, r_2 = 4s(g^2 - 1), k_2 = 4(g^2 - 1), \lambda_2 = 4s(2g^2 - 2), \rho_{12} = 4s(g^2 - 4), \rho_{22} = 6s$, where $g = 3, 4, \dots, s = 1, 2, \dots,$

- (xi) $v = 4g^2, b_1 = 4gt, r_1 = t(2g - 1), k_1 = g(2g - 1), \lambda_1 = t(g - 1)$ and $v = 4g^2, b_2 = 4g^2s, r_2 = 4s(g^2 - 1), k_2 = 4(g^2 - 1), \lambda_2 = 4s(g^2 - 2), \rho_{12} = 4s(g^2 - 4), \rho_{22} = 6s$, where $g, t = 3, 4, \dots, s = 1, 2, \dots, t \geq g$,
- (xii) $v = (2g + 1)^2, b_1 = 4t(2g + 1), r_1 = 2gt, k_1 = g(2g + 1), \lambda_1 = t(2g - 1)$ and $v = (2g + 1)^2, b_2 = s(2g + 1)^2, r_2 = s(4g^2 + 4g - 3), k_2 = 4g^2 + 4g - 3, \lambda_2 = s(4g^2 + 4g - 7), \rho_{12} = s(4g^2 + 4g - 15), \rho_{22} = 6s$, where g is odd integer, $g, t = 2, 3, \dots, s = 1, 2, \dots, 4t \geq 2g + 1$,
- (xiii) $v = b_1 = 36g^2, r_1 = k_1 = 3g(6g + 1), \lambda_1 = 3g(3g + 1)$ and $v = 36g^2, b_2 = 12g^2s, r_2 = s(12g^2 - 1), k_2 = 3(12g^2 - 1), \lambda_2 = 2s(6g^2 - 1), \rho_{12} = 3s(4g^2 - 1), \rho_{22} = s$, where $g = 1, 2, \dots, s = 4, 5, \dots$,
- (xiv) $v = b_1 = 36g^2, r_1 = k_1 = 3g(6g - 1), \lambda_1 = 3g(3g - 1)$ and $v = 36g^2, b_2 = 12g^2s, r_2 = s(12g^2 - 1), k_2 = 3(12g^2 - 1), \lambda_2 = 2s(6g^2 - 1), \rho_{12} = 3s(4g^2 - 1), \rho_{22} = s$, where $g = 1, 2, \dots, s = 4, 5, \dots$,
- (xv) $v = 144g^2, b_1 = 24gt, r_1 = t(12g - 1), k_1 = 6g(12g - 1), \lambda_1 = t(6g - 1)$ and $v = 144g^2, b_2 = 48gs, r_2 = s(48g - 1), k_2 = 3(48g - 1), \lambda_2 = 2s(24g - 1), \rho_{12} = 3s(16g - 1), \rho_{22} = s$, where $s = 4, 5, \dots, g = 1, 2, \dots, t = 6, 7, \dots, t \geq 6g$

implies the existence of the optimum chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by (4.1) and with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is of the form (2.4), for each $a > 0$.

The matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ of the optimum chemical balance weighing design should be constructed according to the form of the matrix $\sigma^2 \mathbf{G}$, where \mathbf{G} is given by (2.4), or in other words, according to the parameter a . Thus for different values of a we have to take different parameters of balanced incomplete block design and ternary balanced block design.

Corollary 4.2 Let $a = \frac{1}{2}$. The existence of the balanced incomplete block design and the ternary balanced block design with parameters

- (i) $v = 5, b_1 = 10, r_1 = 4, k_1 = 2, \lambda_1 = 1$ and $v = 5, b_2 = 5(s + 2), r_2 = 3(s + 2), k_2 = 3, \lambda_2 = s + 3, \rho_{12} = s + 6, \rho_{22} = s$,
- (ii) $v = 9, b_1 = 18, r_1 = 8, k_1 = 4, \lambda_1 = 5$ and $v = 9, b_2 = 3(s + 4), r_2 = 2(s + 4), k_2 = 6, \lambda_2 = s + 5, \rho_{12} = 8, \rho_{22} = s$,
- (iii) $v = 12, b_1 = 22, r_1 = 11, k_1 = 6, \lambda_1 = 5$ and $v = 12, b_2 = 3(2u + 5), r_2 = 2(2u + 5), k_2 = 8, \lambda_2 = 2(u + 3), \rho_{12} = 6 - 2u, \rho_{22} = 3u + 2$
- (iv) $v = 15, b_1 = 42, r_1 = 14, k_1 = 5, \lambda_1 = 4$ and $v = 15, b_2 = s + 14, r_2 = s + 14, k_2 = 15, \lambda_2 = s + 13, \rho_{12} = s, \rho_{22} = 7$,

where $s = 1, 2, \dots, u = 0, 1, 2$, implies the existence of the optimum chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by (4.1) and with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is of the form (2.4).

Corollary 4.3 Let $a = \frac{3}{2}$. The existence of the balanced incomplete block design and the ternary balanced block design with parameters

- (i) $v = 9, b_1 = 18, r_1 = 8, k_1 = 4, \lambda_1 = 3$ and $v = 9, b_2 = 27, r_2 = 15, k_2 = 5, \lambda_2 = 6, \rho_{12} = 13, \rho_{22} = 6$,
- (ii) $v = 12, b_1 = 22, r_1 = 11, k_1 = 6, \lambda_1 = 5$ and $v = 12, b_2 = 14, r_2 = 7, k_2 = 6, \lambda_2 = 3, \rho_{12} = 5, \rho_{22} = 1$

implies the existence of the optimum chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by (4.1) and with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is of the form (2.4).

Corollary 4.4 Let $a = 2$. The existence of the balanced incomplete block design and the ternary balanced block design with parameters

- (i) $v = 7, b_1 = 7, r_1 = 3, k_1 = 3, \lambda_1 = 1$ and $v = 7, b_2 = 21, r_2 = 12, k_2 = 4, \lambda_2 = 5, \rho_{12} = 6, \rho_{22} = 3$,
- (ii) $v = 7, b_1 = 21, r_1 = 6, k_1 = 2, \lambda_1 = 1$ and $v = 7, b_2 = s + 13, r_2 = s + 13, k_2 = 7, \lambda_2 = s + 11, \rho_{12} = s + 1, \rho_{22} = 6$,
- (iii) $v = 12, b_1 = 33, r_1 = 11, k_1 = 4, \lambda_1 = 3$ and $v = 12, b_2 = s + 22, r_2 = s + 22, k_2 = 12, \lambda_2 = s + 20, \rho_{12} = s, \rho_{22} = 11$,
- (iv) $v = 13, b_1 = 22, r_1 = 4, k_1 = 4, \lambda_1 = 1$ and $v = 13, b_2 = s + 25, r_2 = s + 25, k_2 = 13, \lambda_2 = s + 23, \rho_{12} = s + 1, \rho_{22} = 12$

implies the existence of the optimum chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by (4.1) and with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is of the form (2.4).

Corollary 4.5 Let $a = 3$. The existence of the balanced incomplete block design and the ternary balanced block design with parameters

- (i) $v = 11, b_1 = 11, r_1 = 5, k_1 = 5, \lambda_1 = 2$ and $v = 11, b_2 = 22, r_2 = 12, k_2 = 6, \lambda_2 = 5, \rho_{12} = 2, \rho_{22} = 5$,
- (ii) $v = 15, b_1 = 15, r_1 = 7, k_1 = 7, \lambda_1 = 3$ and $v = 15, b_2 = 25, r_2 = 15, k_2 = 9, \lambda_2 = 8, \rho_{12} = 7, \rho_{22} = 4$

implies the existence of the optimum chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ given by (4.1) and with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is of the form (2.4).

Let us suppose that in the matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ we replace the incidence matrices, i.e. \mathbf{N}_1 is the incidence matrix of the ternary balanced block design with parameters $v, b_1, r_1, k_1, \lambda_1, \rho_{11}, \rho_{21}$ and \mathbf{N}_2 is the incidence matrix of the balanced incomplete block design with parameters $v, b_2, r_2, k_2, \lambda_2$. Thus we get

$$\mathbf{X} = \begin{bmatrix} \mathbf{N}'_1 - \mathbf{1}_{b_1} \mathbf{1}'_v \\ 2\mathbf{N}'_2 - \mathbf{1}_{b_2} \mathbf{1}'_v \end{bmatrix}. \quad (13)$$

A chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ of the form (4.3) with the variance - covariance matrix of

errors $\sigma^2 \mathbf{G}$, where

$\mathbf{G} = \begin{bmatrix} a^* \mathbf{I}_{n_1} & \mathbf{0}_{n_1} \mathbf{0}'_{n_2} \\ \mathbf{0}_{n_2} \mathbf{0}'_{n_1} & \mathbf{I}_{n_2} \end{bmatrix}$ is optimal if and only if

$$a^*[b_1 + \lambda_1 - 2r_1] + [b_2 - 4(r_2 - \lambda_2)] = 0, \quad (14)$$

where $a^* = \frac{1}{a}$. It implies

Corollary 4.6 An optimum chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ of the form (4.1) with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where \mathbf{G} is given by (2.4), exists if and only if there exists the optimum chemical balance weighing design with the design matrix $\mathbf{X} \in \Phi_{n \times p, m}(-1, 0, 1)$ of the form (4.3) with the variance - covariance matrix of errors $\sigma^2 \mathbf{G}$, where $\mathbf{G} = \begin{bmatrix} a \mathbf{I}_{n_1} & \mathbf{0}_{n_1} \mathbf{0}'_{n_2} \\ \mathbf{0}_{n_2} \mathbf{0}'_{n_1} & \mathbf{I}_{n_2} \end{bmatrix}$.

References

- BANERJEE, K.S. (1975): *Weighing Designs for Chemistry, Medicine, Economics, Operations Research, Statistics*. Marcel Dekker Inc., New York.
- BILLINGTON, E.J. and ROBINSON, P.J. (1983): A list of balanced ternary block designs with $r \leq 15$ and some necessary existence conditions. *Ars Combinatoria*, 16, 235–258.
- CERANKA, B. and GRACZYK, M. (2001): Optimum chemical balance weighing designs under the restriction on weighings. *Discussiones Mathematicae - Probability and Statistics* 21, 111–121.
- CERANKA, B. and GRACZYK, M. (2003): Optimum chemical balance weighing designs. *Tatra Mountains Math. Publ.*, 29, 1–9.
- HOTELLING, H. (1944): Some improvements in weighing and other experimental techniques. *Ann. Math. Stat.*, 15, 297–305.
- KATULSKA, K. (1989): Optimum chemical balance weighing designs with non-homogeneity of the variances of errors. *J. Japan Statist. Soc.*, 19, 95–101.
- RAGHAVARAO, D. (1971): *Constructions and Combinatorial Problems in Designs of Experiments*. John Wiley Inc., New York.

Combination of Regression Trees and Logistic Regression to Analyse Animal Management and Disease Data

Susanne Dahms

Institut für Biometrie und Informationsverarbeitung,
FB Veterinärmedizin, FU Berlin, Oertzenweg 19b, D-14163 Berlin, Germany

Abstract. Integrated quality control measures have been discussed in farm animal husbandry and veterinary medicine for some time in search of concepts to link information on management factors with disease data gained by veterinary meat inspection at slaughter. For this purpose an exploratory modelling strategy has been developed that combines a characterization of variance structures, the generation of regression trees to explore association structures in available animal management and disease data, and logistic regression used to quantify systematic effects of farm management factors.

1 Background and data

For some years now quality control measures have been discussed in farm animal husbandry and veterinary medicine trying to establish an intensified integration of meat production from stable to table. Better understanding of relationships between on-farm circumstances and meat quality, together with an exchange of data between farms and slaughterhouses, may help to improve farm practices as well as meat safety (Berends et al. (1996)).

Several fields of veterinary medicine are involved in these discussions. Herd-health management deals with veterinary action to cure or prevent animal diseases, and with consultation concerning conditions under which farm animals are raised and fattened. On the other hand, a major task of meat hygiene is post-mortem inspection of slaughtered animals in order to sort out carcasses not suitable for human consumption. Then, in search of concepts how to link information on farm management factors and disease data gained by meat inspection, epidemiology is involved. In this context epidemiology takes a population-based approach and comprises methods to quantify dependencies between animal husbandry conditions and occurrence of disease, which require statistical analyses and association modelling techniques.

From a statistical point of view there are several possibilities to define **study units**: single animals, day-to-day deliveries of animals to slaughterhouses, fattening batches, i.e. groups of animals fattened together on a farm under the same conditions, or farms as such, i.e. the entirety of animals fattened on the same farm during a specified period of time regardless of different fattening batches they belong to.

All these choices have their pros and cons. For instance, farm management factors affect units like single animals or fattening batches. Slaughterhouse data, however, commonly refer to single animals or day-to-day deliveries. The latter are usually not identical with fattening batches which are sent in several deliveries and maybe even to different slaughterhouses. Therefore completeness of data is a problem and its lack suggests a selection bias. Interpretation of single animal data, on the other hand, is complicated by individual biases due to different meat inspectors and their experience. Problems like reliability of lesion recording and diagnostic sensitivity should be considered.

Farm management factors are recorded as variables representing hygienic or immunologic impacts on animal health. Other information available on-farm refer to occurrence of diseases or to animal performance, like weight which can serve as an indicator for health problems. During slaughter performance data are recorded as well. However, in this context **disease data**, as for instance occurrence of inflammatory lesions on carcasses, are more relevant. Depending on the choice of study units, disease data can be recorded and analysed as dichotomous variables with regard to individual animals (lesion found or not) or as prevalences (relative frequencies) in groups of animals.

2 Model construction — a multistage process

The development of strategies pursued in the model building process has to consider several aspects (Hand (1997)).

First, it is useful to clarify which main purpose the model being developed should serve. Is it prediction of disease occurrence or factor evaluation?

If it is prediction, emphasis will be put on reliability of, say, predicted prevalences, providing slaughterhouses with the information which lesions are to be expected at meat inspection. This does not necessarily need predictors with a sensible biological interpretation that could be understood as factors directly causing or promoting disease.

However, if the purpose of modelling is factor evaluation, emphasis will be put on ‘operational’ models that enhance understanding of dependencies and give practical advice for action on farm. This needs predictors and model representations that inform herd-health managers on which conditions should be changed in order to prevent animal health disorders and lesions on slaughter animals.

A second aspect guiding model construction is how much prior knowledge about the problem is already available (Hand (1997)). In practice a large number and diversity of farm management variables has to be considered as possibly related to disease occurrence, some of them interacting with each other. Therefore, three stages of model construction have been worked out addressing structures in the data that become important with increasing knowledge about the problem (Dahms (2000)):

- To identify those study units that are mainly affected by influencing variables the **variation structure** in terms of components of variance or deviance is characterized with regard to farm or fattening batch effects. The result can guide the decision with which type of study units analysis should proceed and may thus narrow down the choice of influencing factors.
- Then, in order to select factors to be included in a preliminary model the **association structure** between factors and disease occurrence is explored.
- Following factor selection **quantification of systematic factor effects** is pursued to describe and distinguish their impacts more precisely.

Last but not least, technical aspects have to be considered that may limit the available selection of modelling techniques, like types of data, scales on which variables are measured, or the occurrence of randomly or non-randomly missing values.

Variation structure

Suppose the target variable is measured as lesion prevalences for units like fattening batches, preferably, or deliveries. The variation structure can be characterized by estimated variance components if we can assume an approximative normal distribution, at least after some appropriate transformation. The underlying model for (transformed) lesion prevalences,

$$Y_{ij} = \beta_0 + b_i + e_{ij}$$

consists of a constant value β_0 , a random farm effect b_i , and a random residual effect e_{ij} that describes the remaining variation between fattening groups within farms. The estimated variance components $\text{Var}(b_i)$ and $\text{Var}(e_{ij})$ in proportion to each other point out whether variation of prevalences is mainly due to factor variation between fattening batches or to factors that differ mainly between farms.

Association structure

Subsequently the generation of classification or regression trees for (transformed) lesion prevalences serves to explore their associations with farm management factors (see Breiman et al. (1984), or Clark and Pregibon (1992)). A regression tree with its resemblance to decision trees can be understood as an ‘operational’ model. It describes relationships between target variable and predictors in a way easily understood by users like herd-health managers.

The regression tree is generated by recursive binary partitioning of the study group, trying to split up subgroups as homogeneous as possible with regard to the target variable. Groups are partitioned according to management factors that may affect prevalences on those levels (batches or farms) worked out before. At each partitioning step the split maximizing deviance reduction ΔD for the target variable is chosen, with

$$\Delta D = D(\hat{\mu}; y) - \left[\sum_L D(\hat{\mu}_L; y_i) + \sum_R D(\hat{\mu}_R; y_i) \right],$$

provided the group to be split up has a given minimal deviance and sub-groups contain at least a given number of study units. Here, $D(\hat{\mu}; y)$ denotes the deviance in the parent group which would be reduced by the sum of deviances in the resulting left (L) and right (R) sub-groups.

Accordingly goodness-of-fit is measured by tree residual deviance:

$$D = \sum_j D(\hat{\mu}_j; y_j) = \sum_j \sum_i D(\hat{\mu}_j; y_{ji}),$$

or mean residual deviance $D/(n - J)$, respectively, where J is the number of end groups split up in the partitioning process.

In case of regression trees the deviance is estimated as a weighted sum of squared deviations from the expected mean value μ_j :

$$D(\hat{\mu}_j; y_j) = \sum_i w_{ij} \cdot (y_{ij} - \hat{\mu}_j)^2.$$

Expected means μ_j are estimated by the arithmetic means of observations in sub- or end-groups.

Maximization of deviance reduction has been chosen from the variety of split criteria discussed in the literature to follow principles of maximum-likelihood estimation. With regard to the modelling techniques used for quantifying systematic factor effects this choice ensures a consistent treatment of target variables and comparability of results. Based on the generated tree those factors yielding major contributions to deviance reduction are selected for inclusion in a preliminary model.

Quantification of systematic factor effects

To quantify systematic factor effects generalized linear models are estimated (see McCullagh and Nelder (1989), or Fahrmeir and Tutz (1994)). With regard to lesion prevalence as the target variable a linear model is chosen for transformed values, or a logistic regression model for prevalences as such.

In case of logistic regression prevalences are modelled as relative frequencies Y_i/n_i assuming a binomial distribution for the counts Y_i with parameters n_i and π_i denoting the probability for lesion occurrence.

The logit-transformation is used as the link-function yielding the linear predictor:

$$\eta = \log \left(\frac{\pi}{1 - \pi} \right) = X\beta,$$

where logit-transformed lesion probabilities π_i are explained by a linear combination of influencing factors comprised by X . Based on the logit-link, estimates for odds-ratios can be derived from estimated factor effects as $\exp(\hat{\beta})$.

3 Slaughter pigs — an example

Data to illustrate these stages in a model building process can be taken from a German project primarily initiated to investigate alternative meat inspection techniques for slaughter pigs (Fries et al. (1997)). For all farms where pigs were fattened for this study farm management data have been recorded as well as data on lesions found at meat inspection. At three different slaughterhouses a total of about 22 000 pigs from 50 fattening farms were examined during the course of this project. Each farmer always delivered his study pigs to the same slaughterhouse.

The prevalence of callous affections of legs, i.e. lesions caused by reacting tissue due to mechanical impact with signs of inflammation, may serve as an example target variable to illustrate the modelling procedure.

Prevalences referring to fattening batches cannot be analysed as the belonging to such batches was not identified and recorded. Therefore, only day-to-day deliveries of pigs for slaughter or, alternatively, farms comprising the entirety of project animals they delivered can be chosen as study units.

The collection of farm management factors consists of farm characteristics that remained constant over time, like type of pen, type of floor, ventilation system, type of feeding and watering, number of piglet suppliers, cleaning and disinfection system, storage of liquid manure, etc.

Variation structure

Variance components have been calculated with transformed prevalences that have been adjusted for inspection biases as described by Dahms (2000). Such an adjusted prevalence is derived as the difference between lesion prevalence in a single delivery to a slaughterhouse and the total prevalence in all deliveries to that specific slaughterhouse on the same day. Based on all 50 farms involved in the project a number of $n = 689$ day-to-day deliveries with at least 10 pigs from farms with at least 5 deliveries could be used.

The estimation of variance components for random farm and delivery effects within farms yields values of 0.0188 and 0.0165, summing up to a total variance of 0.0348. Only about 52% of the total variance are due to farm effects, and may thus be explained by management factors that vary between farms.

Farm management factors as recorded for this study can provide no explanation for the variation between deliveries. Therefore, further analysis proceeds on a farm basis, using their mean adjusted prevalences to generate

regression trees and using untransformed prevalences in their total supplies of study animals for logistic regression.

Association structure

The regression tree generated with mean adjusted prevalences of callous affections of legs for $n = 50$ farms is shown in Figure 1.

The left subgroup resulting from a split is always characterized by a smaller mean value than the right one, in this case a lower mean of mean adjusted lesion prevalences.

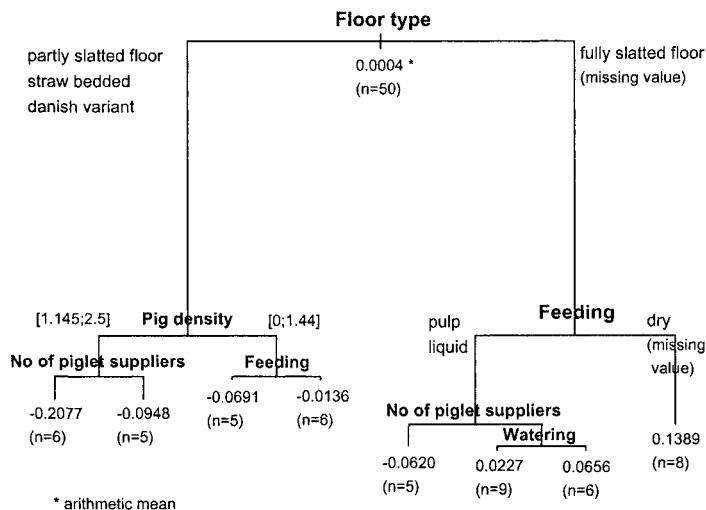


Fig. 1. Regression tree for mean adjusted prevalences of callous affections of legs for $n=50$ farms

In Figure 1 the length of vertical lines describing a split is proportional to the amount of deviance reduction achieved with it. Considering this, two factors are selected for inclusion in a preliminary model: 'floor type' with its different categories describing floor quality and roughness, and 'feeding' with different types of feed as categories.

Quantification of systematic factor effects

To quantify systematic effects of floor type and feeding on untransformed prevalences of callous affections of legs logistic regression and estimation of odds-ratios have been chosen. Table 1 shows resulting estimated odds-ratios and their 95% confidence limits for floor type and feeding categories in re-

lation to a reference group characterized by a fully slatted floor and dry feeding.

Effect	$\widehat{OR} = e^{\beta}$	Confidence Limits
β_2 (partly slatted floor)	0.4727	0.2747 – 0.8134
β_3 (straw bedded)	0.2833	0.1087 – 0.7387
β_4 (Danish variant)	0.2114	0.1370 – 0.3263
β_5 (pulp feeding)	0.8368	0.5507 – 1.2714
β_6 (liquid feeding)	0.9533	0.6345 – 1.4322

Table 1. Estimated odds-ratios (e^{β}) and 95% confidence limits for occurrence of callous affections of legs depending on floor type and feeding (based on $n = 46$ farms) (reference categories: fully slatted floor, dry feeding)

The odds-ratios and their confidence limits for the floor categories stress the conspicuousness of this factor. Chances for lesion occurrence appear to be reduced if there is only a partly slatted floor, if pigs are straw bedded, or if they have even the Danish variant of pen. Feeding types, on the other hand, don't show odds-ratios distinguishable from 1.

These results establish a preliminary working model that should be checked with independent data.

4 Summary

In this paper a multistage procedure is presented that serves to investigate and model relationships between management conditions for farm animals and occurrence of disease or lesions at slaughter. The core of this model construction concept consists of a sequence of three steps:

- characterization of the variation structure, if appropriate by means of variance components estimation,
- selection of farm management factors as predictors by recursive binary partitioning, and
- estimation of systematic factor effects using generalized linear models, for instance logistic regression.

This sequence may be embedded in a modelling cycle that refines recording and/or definition of variables and selection of relevant factors with each passage. In addition it has to be stressed that this modelling process should be understood as an exploratory analysis. Data available for such analyses are usually gained as observational data and biases cannot be precluded. Therefore, the objective can only be to describe association structures and generate hypotheses, where a hypothesis is formalized as a so-called preliminary working model. Repeated analyses with new, independent data would be required to confirm these working models.

There are still many problems to be solved when it comes to procedural details. For instance recording of fattening groups instead of deliveries requires new ideas and systems for animal identification and data acquisition at slaughterhouses. Definition and measurement of farm management factors is another crucial point. These are problems that can not be solved by statistical methods.

There are, however, methodical problems from a statistical point of view as well. For instance, to get a clear picture, lesion prevalences usually need to be adjusted for effects not contributable to farms, like inspection biases or seasonal and climatic effects. However, a completely satisfactory solution has not yet been found. Another problem would arise when there is a possibility to get hold of fattening batches within farms. What would their representation in a regression tree look like? Ways in which recursive binary partitioning could handle such random effects are still to be discussed.

References

- BERENDS, B.R., VAN KNAOPEN, F., and SNIJDERS, J.M.A. (1996): Suggestions for the construction, analysis and use of descriptive epidemiological models for the modernization of meat inspection. *International Journal of Food Microbiology*, 30, 27–36.
- BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., and STONE, C.J. (1984): *Classification and Regression Trees*. Wadsworth & Brooks/Cole, Monterey, California.
- CLARK, L.A. and PREGIBON, D. (1992): Tree-based models. In: J.M. Chambers and T.J. Hastie (Eds.): *Statistical Models in S*. Wadsworth & Brooks/Cole, Pacific Grove, California, 377–420.
- DAHMS, S. (2000): *Bestandsgesundheit und Lebensmittelsicherheit — Beiträge der Biometrie und Epidemiologie*. Habilitationsschrift, Fachbereich Veterinärmedizin, Freie Universität Berlin.
- FAHRMEIR, L. and TUTZ, G. (1994): *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, Berlin.
- FRIES, R., BANDICK, N., and KOBE, A. (1997): *Vergleichende Untersuchungen zur Aussagekraft der amtlichen Schlachttier- und Fleischuntersuchung und einer alternativen Erhebungstechnik an Schlachtschweinen im niederrheinischen Raum*. Forschungsbericht im Auftrag des MURL (NRW), BMG (BgVV) und der Fleischindustrie NRW.
- HAND, D.J. (1997): *Construction and Assessment of Classification Rules*. John Wiley & Sons, Chichester.
- MCCULLAGH, P. and NELDER, J.A. (1989): *Generalized Linear Models*. Chapman and Hall, London.

Robustness of ML Estimators of Location-Scale Mixtures

Christian Hennig

Fachbereich Mathematik - SPST,
Universität Hamburg,
Bundesstr. 50, D-20146 Hamburg, Germany

Abstract. The robustness of ML estimators for mixture models with fixed and estimated number of components s is investigated by the definition and computation of a breakdown point for mixture model parameters and by considering some artificial examples. The ML estimator of the Normal mixture model is compared with the approach of adding a “noise component” (Fraley and Raftery (1998)) and by mixtures of t -distributions (Peel and McLachlan (2000)). It turns out that the estimation of the number of mixture components is crucial for breakdown robustness. To attain robustness for fixed s , the addition of an improper noise component is proposed. A guideline to choose a lower scale bound is given.

1 Introduction

Maximum likelihood (ML)-estimation based on mixtures of Normal distributions (NMML) is a flexible and widely used technique for cluster analysis (see, e.g., Fraley and Raftery (1998)).

Observations x_1, \dots, x_n are modeled as i.i.d. according to the density

$$f_\eta(x) = \sum_{j=1}^s \pi_j f_{a_j, \sigma_j}(x), \text{ where } f_{a, \sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-a}{\sigma}\right), \quad (1)$$

where $\eta = (s, a_1, \dots, a_s, \sigma_1, \dots, \sigma_s, \pi_1, \dots, \pi_s)$ is the parameter vector, the number of components $s \in \mathbb{N}$ may be known or unknown, (a_j, σ_j) pairwise distinct, $a_j \in \mathbb{R}$, $\sigma_j > 0$, $\pi_j > 0$, $j = 1, \dots, s$ and $\sum_{j=1}^s \pi_j = 1$. For the Normal mixture model, $f = \varphi$ is the density of the standard Normal distribution. Often mixtures of multivariate Normals are used, but for the sake of simplicity, I restrict considerations to the case of one-dimensional data in this paper.

As many other ML-techniques based on the Normal distribution, NMML is not robust against gross outliers, at least if the number of components s is treated as fixed: The estimators of the parameters a_1, \dots, a_s are weighted means of the observations where the weights for each observation sum up to one (see Redner and Walker (1984)), which means that at least one of these parameters can get arbitrarily large if a single extreme point is added to a dataset.

There are some ideas to overcome the robustness problems of Normal mixture. The software MCLUST (Fraley and Raftery 1998) allows the addition of a mixture component accounting for “noise”, modeled as a uniform distribution on the convex hull (the range in one dimension, respectively) of the data, i.e., the data is modeled as generated by

$$f_\zeta(x) = \sum_{j=1}^s \pi_j f_{a_j, \sigma_j}(x) + \pi_0 \frac{1(x \in [x_{min}, x_{max}])}{x_{max} - x_{min}}, \quad (2)$$

for given $x_{min}, x_{max} \in \mathbb{R}$, where $\zeta = (s, a_1, \dots, a_s, \sigma_1, \dots, \sigma_s, \pi_0, \pi_1, \dots, \pi_s)$, $\pi_0, \dots, \pi_s > 0$, $\sum_{j=0}^s \pi_j = 1$ and $1(\dots)$ is the indicator function. The corresponding ML procedure will be denoted by NMN in the following.

The software EMMIX (Peel and McLachlan (2000)) can be used to fit a mixture of t -distributions instead of Normals by ML (t_ν MML), i.e., $f = f_\nu$ being the density of the t -distribution with ν degrees of freedom in (1).

Note that the presented theory will hold if f is any continuous density f that is symmetrical about its only mode 0 and that is positive on \mathbb{R} .

There are some alternatives for robust estimation of mixture components, see McLachlan and Peel (2000, p. 222 ff.) and the references given therein.

While a clear gain of stability can be demonstrated for these methods in various examples (see e.g. Banfield and Raftery (1993), McLachlan and Peel (2000, p. 231 ff.)), there is a lack of theoretical justification of their robustness.

In Section 2, I give a formal definition of a breakdown point for estimators of mixture parameters. The breakdown point goes back to Hampel (1971) and measures the smallest amount of contamination that can spoil an estimator completely.

In Section 3.1, some results about the parameter breakdown of the mixture based clustering techniques are given. The number of components s is assumed to be known here. It is shown that for all techniques introduced above r outliers can make $r < s$ mixture components break down.

To attain a better breakdown behavior, I suggest the maximization of a kind of “improper likelihood” in Section 3.2 where “noise” is modeled by an improper uniform distribution on the real line.

In Section 3.3, the case of an estimated number of mixture components s is treated. I consider s as estimated by the maximization of the Bayesian information criterion $BIC(s)$ (Schwarz (1978)):

$$BIC(s) = 2L_{n,s}(\eta_{n,s}) - k \log n, \quad (3)$$

where $L_{n,s}$ denotes the log-likelihood function for n points and s mixture components under one of the models (1) or (2) and $\eta_{n,s}$ is the corresponding ML estimator. k denotes the number of free parameters, i.e., $k = 3s - 1$ for (1) and $k = 3s$ for (2). For alternative methods to estimate s , I refer to Chapter 6 of McLachlan and Peel (2000).

With estimated s , all treated methods are able to isolate gross outliers as new mixture components on their own and are therefore very stable against extreme outliers. Breakdown can happen only because additional points inside the area of the estimated mixture components of the original data can lead to the estimation of a smaller number of components.

An important problem in ML estimation for mixture models is the convergence of the log-likelihood function to ∞ if one of the σ_j^2 converges to 0. In order to get well defined estimators, the log-likelihood function has to be maximized under a restriction on the scale parameters. The simplest possible restriction is $\min_j \sigma_j \geq \sigma_0 > 0$, which is used to obtain the results given below. The choice of σ_0 is discussed in Section 4.

Some examples are given in Section 5. They illustrate that the stability of the methods depends on the scale restriction and the internal stability of the dataset.

2 Breakdown point definitions

The classical meaning of breakdown for finite samples is that an estimator can be driven as far away from its original value as possible by addition of arbitrarily unfortunate points, usually by gross outliers. Donoho and Huber (1983) distinguish this “addition breakdown point” from breakdown by replacement of points. I consider the former definition here.

Breakdown means that estimators that can take values on the whole range of \mathbb{R}^p , can leave every compact set. If the value range of a parameter is bounded, breakdown means that addition of points can take the parameter arbitrarily close to the bound, e.g., a proportion parameter to 0.

A breakdown of an estimator of mixture (or cluster) parameters can be understood in two ways: A situation where at least one of the mixture components explodes is defined as breakdown in Garcia-Escudero and Gordaliza (1999) and Kharin (1996). In contrast to that, Gallegos (2003) defines breakdown in cluster analysis as a situation where *all* clusters explode simultaneously. The definition given here is flexible enough to account for all these situations.

Definition 1. Let $(E_n)_{n \in \mathbb{N}}$ be a sequence of estimators of η in model (1) or of ζ in model (2) on \mathbb{R}^n for fixed $s \in \mathbb{N}$. Let $r \leq s$, $\mathbf{x}_n = (x_1, \dots, x_n)$ be a dataset, where

$$\forall \hat{\eta} = \arg \max_{\eta} L_{n,s}(\eta, \mathbf{x}_n) : \hat{\pi}_j > 0, j = 1, \dots, s. \quad (4)$$

The r -components breakdown point of E_n is defined as

$$B_{r,n}(E_n, \mathbf{x}_n) = \min_g \left\{ \frac{g}{n+g} : \exists j_1 < \dots < j_r \right.$$

$$\forall D = [\pi_{\min}, 1] \times C, \pi_{\min} > 0, C \subset \mathbb{R} \times \mathbb{R}^+ \text{ compact}$$

$$\exists \mathbf{x}_{n+g} = (x_1, \dots, x_{n+g}), \hat{\eta} = E_{n+g}(\mathbf{x}_{n+g}) : (\hat{\pi}_j, \hat{a}_j, \hat{\sigma}_j) \notin D, j = j_1, \dots, j_r \}.$$

The proportions π_j are defined not to break down if they are bounded away from 0, which implies that they are bounded away from 1 if $s > 1$.

In the situation for unknown s , I restrict considerations to the case of 1-components breakdown. Breakdown means that neither of the s mixture components estimated for \mathbf{x}_n vanishes, nor that any of their scale and location parameters explodes to ∞ under addition of points. Further, breakdown of the proportions π_j to 0 is no longer of interest for estimated s according to the BIC, because if some π_j is small enough, s will simply be estimated as being smaller.

Definition 2. Let $(E_n)_{n \in \mathbb{N}}$ be a sequence of estimators of η in model (1) or of ζ in model (2) on \mathbb{R}^n , where $s \in \mathbb{N}$ is unknown and estimated as well. Let $\mathbf{x}_n = (x_1, \dots, x_n)$ be a dataset. Let s be the estimated number of components of $E_n(\mathbf{x}_n)$. The **breakdown point** of E_n is defined as

$$\begin{aligned} B_n(E_n, \mathbf{x}_n) = \min_g \left\{ \frac{g}{n+g} : \forall C \subset \mathbb{R}^s \times (R^+)^s \text{ compact} \right. \\ \exists \mathbf{x}_{n+g} = (x_1, \dots, x_{n+g}), \hat{\eta} = E_{n+g}(\mathbf{x}_{n+g}) : \\ \text{pairwise distinct } j_1, \dots, j_s \text{ do not exist, such that} \\ \left. (\hat{a}_{j_1}, \dots, \hat{a}_{j_s}, \hat{\sigma}_{j_1}, \dots, \hat{\sigma}_{j_s}) \in C \right\}. \end{aligned}$$

This implies especially that breakdown occurs whenever $\hat{s} < s$, \hat{s} being the estimated s for \mathbf{x}_{n+g} .

3 Breakdown results

3.1 Breakdown point for fixed s

Let $r < s$. The contribution of r added points x_{n+1}, \dots, x_{n+r} to the log-likelihood is, for model (1), $\sum_{i=n+1}^r \log \left(\sum_{j=1}^s \pi_j f_{a_j, \sigma_j}(x_i) \right)$. It converges to $-\infty$ if the distances among these r points and between them and the original n points converge to ∞ , and more than $s-r$ mixture components remain in a compact set about the originally estimated mixture. On the other hand, the log-likelihood is bounded from below, if the r additional points are fitted by r mixture components. This means that r additional points make r mixture components break down. The argument holds as well for NMN because the noise density also converges to 0.

Theorem 1. Let $\mathbf{x}_n \in \mathbb{R}^n$, $s > 1$. Let $\eta_{n,s}$ be an ML estimator for model (1) or (2). For $r = 1, \dots, s-1$,

$$B_{r,n}(\eta_{n,s}, \mathbf{x}_n) \leq \frac{r}{n+r}. \quad (5)$$

The proof is given in Hennig (2002), Theorem 4.4. For $r = s$, this remains true for the NMML and NMN, while t_ν MML has a better s -components breakdown point of $\geq \frac{1}{\nu+1}$, see Theorem 4.7 of Hennig (2002).

3.2 An alternative for fixed s

An alternative can be constructed as a modification of NMN. The problem of NMN is that the noise component could be affected by outliers as well, as was shown in the previous section. This can be prevented when the density constant for the noise component is chosen as fixed beforehand, which leads to ML estimation for a mixture where some improper density component is added to catch the noise (NMI). That is, an estimator $\xi_{n,s}$ is defined as the maximizer of

$$L_{n,s}(\xi, \mathbf{x}_n) = \sum_{i=1}^n \log \left(\sum_{j=1}^s \pi_j f_{a_j, \sigma_j}(x_i) + \pi_0 b \right), \quad (6)$$

where $b > 0$. This requires the choice of b . If the objective is cluster analysis and there is a maximum scale σ_{max} , above which a mixture component is no longer accepted as a cluster (compare Section 4), b could be chosen as the density value at the 0.025-quantile of $f_{0,\sigma_{max}}$, so that 95% of the points generated from such a distribution have a larger density value for it than for the noise component. For this estimator the breakdown point depends on the stability of the dataset \mathbf{x}_n . Breakdown can only occur if additional observations allow that the non-outliers can be fitted by fewer than s components, and this means that a relatively good solution for $r < s$ components must exist already for \mathbf{x}_n . This is formalized in (7). Let $L_{n,s} = L_{n,s}(\xi_{n,s}, \mathbf{x}_n)$. I consider only the breakdown of a single mixture component $B_{1,n}(\xi_{n,s}, \mathbf{x}_n)$.

Theorem 2. *Let $\mathbf{x}_n \in IR^n$. Let a_j, σ_j, π_j denote the parameters of $\xi_{n,s}$ and $f_{max} = f(0)/\sigma_0 > b$. If*

$$\begin{aligned} \max_{r < s} L_{n,r} &< \sum_{i=1}^n \log \left(\sum_{j=1}^s \pi_j f_{a_j, \sigma_j}(x_i) + (\pi_0 + \frac{g}{n})b \right) \\ &\quad + g \log(\pi_0 + \frac{g}{n})b + (n+g) \log \frac{n}{n+g} - g \log f_{max}, \end{aligned} \quad (7)$$

then

$$B_{1,n}(\xi_{n,s}, \mathbf{x}_n) > \frac{g}{n+g}. \quad (8)$$

The proof is given in Hennig (2002), Theorem 4.13. The meaning of (7) is illustrated in Section 5.

3.3 Breakdown point for unknown s

The treatment of s as unknown is favorable for robustness against outliers, because outliers can be fitted by additional mixture components. Generally, for large enough outliers the addition of a new mixture component for each outlier yields a better log-likelihood than any essential change of the original

mixture components. That is, gross outliers are almost harmless except that they let the estimated number of components grow.

Breakdown can occur, however, because added points inside the range of the original data, may lead to a preference of a solution with $r < s$ clusters. (9) is a necessary condition to prevent this.

Theorem 3. *Let $\tau_n = (s, \eta_{n,s})$ be a maximizer of the BIC under (1) or (2). If*

$$\min_{r < s} \left[L_{n,s} - L_{n,r} - \frac{1}{2}(5g + 3s - 3r + 2n) \log(n + g) + n \log n \right] > 0, \quad (9)$$

then

$$B_n(\tau_n, \mathbf{x}_n) > \frac{g}{n + g}. \quad (10)$$

The proof is given in Hennig (2002), Theorem 4.15. The meaning of (9) is illustrated in Section 5.

4 Choice of the scale restrictions

In most applications, sufficient prior information to specify the scale restriction constant σ_0 is not available. A common strategy to avoid a sensible specification of these constants in practice is to compute local maximizers of the log-likelihood from initial values which avoid very small values for the sigmas. This, however, avoids the isolation of single points as clusters, which is crucial for good breakdown behavior for estimated s .

Consider s as unknown. A sensible choice of the restriction constant should fulfill two objectives:

1. If a data subset looks like a homogeneous cluster, it should be estimated as one component. No single point of it should form a “one-point-component” with a very small scale. To ensure that, the constant has to be large enough.
2. The constant should be so small that a gross outlier generates a new component instead of being merged with an otherwise homogeneous data subset.

α -outliers (with $\alpha > 0$ but very small) are defined by Davies and Gather (1993) with respect to an underlying model as points from a region of low density, chosen so that the probability of the occurrence of an outlier is $\leq \alpha$. For a standard Normal distribution, for example the points outside $[\Phi^{-1}(\frac{\alpha}{2}), \Phi^{-1}(1 - \frac{\alpha}{2})]$ are the α -outliers, where Φ_{a,σ^2} denotes the cdf of the Normal distribution with parameters a, σ^2 . For $\alpha_n = 1 - (1 - p)^{1/n}$, the probability of the occurrence of at least one α_n -outlier among n i.i.d. points from $\mathcal{N}(0, 1)$ is equal to p .

The strategy is as follows: Choose $p = 0.05$, say, and consider the choice of σ_0 for the NMML with unknown s . The following definition is used to generate reproducible benchmark datasets:

Definition 3. $\Phi_{a,\sigma^2}^{-1}(\frac{1}{n+1}), \dots, \Phi_{a,\sigma^2}^{-1}(\frac{n}{n+1})$ is called a (a, σ^2) -Normal standard dataset (NSD) with n points.

Assume for the moment that at least $n - 1$ points come from a $\mathcal{N}(0, 1)$ distribution. (Denote $c_0 = \sigma_0$ in this particular setup.) c_0 should be chosen so that it is advantageous to isolate an α_n -outlier as its own cluster, but not a non-outlier. This, of course, depends on the non-outlying data. As “calibration benchmark”, form a dataset with n points by adding an α_n -outlier to a $(0, 1)$ -NSD with $n - 1$ points. Choose c_0 so that $\text{BIC}(1) = \text{BIC}(2)$ (this can easily be seen to be uniquely possible). For c_0 small enough, the 2-components solution will consist of one component matching approximately the ML-estimator for the NSD and one component fitting only the outlier. Resulting values are given in Table 1.

The interpretation is as follows: Based on $\sigma_0 = c_0$, a dataset consisting of an $(n - 1)$ -point NSD and an α_n -non-outlier will be estimated as homogeneous, while there will be more than one cluster if the n th point is an outlier. The same holds for an $n - 1$ -point (α, σ^2) -NSD and $\sigma_0 = c_0\sigma$. I suggest the use of $\sigma_0 = c_0\sigma_{max}$, where σ_{max}^2 is the largest variance such that a data subset with this variance can be considered as “cluster” with respect to the given application. This may not look like an advantage, because the need to specify a lower bound σ_0 is only replaced by the need to specify an upper bound σ_{max} . But the upper bound has a clear interpretation which does not refer to an unknown underlying truth. At least if the mixture model is used as a tool for cluster analysis, points of a cluster should belong together in some sense, and with respect to a particular application, it can usually be said that points above a certain variation can no longer be considered as “belonging together”.

A dataset to analyze will usually not have the form “NSD plus outlier”, of course. The clusters in the data will usually be smaller than $n - 1$ points, and they will have a variance smaller than σ_{max}^2 . Assume now that there is a homogeneous data subset of $n_1 < n$ points with variance $\sigma^2 \leq \sigma_{max}^2$. The question arises if an α_{n_1} -outlier, non-outlier, respectively, will be isolated from the cluster in the presence of other clusters elsewhere. σ_0 is calculated on the base of the BIC penalty for 1 vs. 2 clusters with n points. That is, the difference in penalty is $3 \log n$. Table 1 also gives the c_0 -values computed with an NSD of size $n_1 = n/2 - 1$ plus $\alpha_{n/2}$ -outlier and of size $n_1 = n/5 - 1$ plus $\alpha_{n/5}$ -outlier, but again with penalty difference $3 \log n$ to show which restriction constant would be needed to isolate at least $\alpha_{n/2}$ -outliers, $\alpha_{n/5}$ -outliers, respectively, from the homogeneous subset of size n_1 under the assumption that the parameters for the rest of the data remain unaffected. The values coincide satisfactorily with the values computed for n , so that these values look reasonable as well for small homogeneous subsets.

With a variance smaller than σ_{max} , an α -outlier with $\alpha > \alpha_n$ is needed to be isolated from a cluster with a variance smaller than σ_{max} , i.e., the

broad tendency is that components with larger variances are preferred over one-point-components.

n	20	50	100	200	1000
c_0	2.10e-2	4.99e-3	1.66e-3	5.51e-4	4.34e-5
$n_1 = n/2 - 1$	9	24	49	99	499
c_0	2.15e-2	5.25e-3	1.76e-3	5.87e-4	4.57e-5
$n_1 = n/5 - 1$	3	9	19	39	199
c_0	2.25e-2	5.44e-3	1.88e-3	6.35e-4	4.93e-5

Table 1. Minimum scale restriction factor c_0 for Normal mixtures. Note that $\log c_0$ is almost exactly linear in $\log n$, so that further values can easily be obtained by interpolation.

Although the argumentation is only valid for NMML with estimated s , I tentatively suggest to apply the resulting values also for the other methods, because the derivation of analogous strategies for them raises certain difficulties.

5 Examples

Consider a dataset of 50 points, namely a (0,1)-NSD with 25 points combined with a (5,1)-NSD with 25 points. Let $\sigma_{max} = 5 \Rightarrow \sigma_0 = 0.025$, $b = 0.0117$ for NMI. For NMML, t_ν MML with $\nu \geq 1$, NMN and NMI, always components corresponding almost exactly to the two NSDs are optimal under $s = 2$ fixed. How large must an additional outlier be chosen so that the 50 original points fall into only one cluster and the second mixture component fits only the outlier? For NMML, breakdown begins with an additional point at about 15.2 (13.3; values in parentheses are for $\sigma_0 = 0.001$ to demonstrate the dependence of the robustness on σ_0). For t_3 MML, the outlier must lie at about 800 (350), t_1 MML needs the outlier at about 3.8e6 (8e5), and NMN breaks down with an additional point at 3.5e7 (1.5e6). The lower breakdown bound (8) of NMI evaluates to $\frac{2}{52}$. The original components are joined by three outliers at 9.8. While NMN has a smaller breakdown point than NMI, three outliers would need to be larger, namely at 70, to join the original components. If the (5,1)-NSD is replaced by a (50,1)-NSD, the lower breakdown bound of NMI is $\frac{7}{57}$ and experimentally 11 outliers at 100, say, are needed for breakdown. Turning back to the combination of the (0,1)-NSD and the (5,1)-NSD, for $\sigma_0 = 0.001$, the lower breakdown bound reduces to $\frac{1}{52}$, and two outliers at 9.8 suffice to join the original components.

Note that NMN is “practically” robust in the sense that it can cope with more than one large outlier, as long as they are below 3.5e7 and scattered enough. For example, if 7 points 1e3, 5e3, 1e4, 5e4, 1e5, 5e5, 1e6 are added to

the original 50 points, all 7 outliers are classified as noise ($\sigma_0 = 0.025$; the same holds for NMI). To a certain extent this also applies to t_1 MML. The seven additional outliers given above lead to breakdown, while outliers at (100, 200, 500, 1000, 2000, 5000, 10000) do still not join the original components.

With estimated s , (10) gives a lower breakdown bound of $\frac{2}{52}$ for NMML and NMN and $\frac{3}{53}$ for t_1 MML at the original 50 points ($s = 2$ is estimated correctly by all methods). These bounds are rather conservative. Empirically, 13 points equally spaced between 1.8 and 3.2 lead to breakdown by $\hat{s} = 1$ for NMML and NMN. t_1 MML is a bit more stable: the mentioned 13 additional “inliers” lead to the estimation of $\hat{s} = 3$. Extreme outliers always get their own new mixture components. It is interesting that the breakdown point can be driven above $\frac{1}{2}$ by enlarging the separation between the components. For a (0,0.001)-NSD of 25 points and a (100000,0.001)-NSD of 25 points, NMML’s lower breakdown bound is $\frac{58}{108}$. Empirically a breakdown point larger than 0.9 can be reached by much less separation.

Consider as a last example a (0,1)-NSD of 45 points combined with a (5,1)-NSD of 5 points. For fixed $s = 2$, NMN needs an outlier at $2e6$ to join the original two components corresponding to the NSD. t_1 MML interprets the (5,1)-NSD as extreme points belonging to the (0,1)-NSD and isolates outliers down to 7.5 as one-point-components. While this setup may seem to be less stable than the constellation with two clusters of 25 points each, NMML joins an outlier up to 40 with the (5,1)-NSD and NMI breaks down with at least 3 outliers at 11 (compared to 9.8 above) at a lower breakdown bound of $\frac{2}{52}$.

For estimated s , NMML needs 12 points between the components to join them (at a lower breakdown bound of $\frac{2}{52}$), while NMN and t_1 MML estimate the original 50 points as only one regular component, while the (5,1)-NSD is estimated as noise, belonging to the only component, respectively, so that there is no second mixture component which could break down.

Note that the results of this section have been computed by using the EM-algorithm (see, e.g., McLachlan and Peel (2000)) several times with initial configurations chosen by use of prior information about the generation of the data. Not all of the likelihood maxima will be reproduced by default applications of available software.

6 Conclusion

A finite-sample-addition breakdown point for estimators of the parameters of mixture models has been defined for a known and unknown number of mixture components. It has been shown that the ability to estimate the number of mixture components is crucial to attain a satisfactory breakdown point for ML estimators. For fixed s , a better breakdown behaviour can be attained by adding an improper uniform density to the likelihood. Note that the ro-

bustness behaviour for fixed s is relevant in practice, because even if the number of components is estimated, there is usually an upper bound on s for computational reasons. For example, for a dataset of 1000 points, one will often estimate s under the restriction $s \leq 10$, say, while there may be much more than 10 outliers. Therefore, NMI, NMN, or t_1 MML are recommended in spite of the breakdown robustness of the simple NMML under estimated s . However, NMI, NMN and t_ν MML may not recognize mixture components supported by too few points.

Breakdown and robustness in mixture models and cluster analysis do not only depend on the method, but also on the internal stability of the clustering of the dataset.

References

- BANFIELD, J.D. and RAFTERY, A.E. (1993): Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, *49*, 803–821.
- DAVIES, P.L. and GATHER, U. (1993): The identification of multiple outliers. *Journal of the American Statistical Association*, *88*, 782–801.
- DONOHO, D.L. and HUBER, P.J. (1983): The notion of breakdown point. In P.J. Bickel, K. Doksum, and J.L. Hodges jr. (Eds.): *A Festschrift for Erich L. Lehmann*, Wadsworth, Belmont, CA, 157–184.
- FRALEY, C. and RAFTERY, A.E. (1998): How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis. *Computer Journal*, *41*, 578–588.
- GALLEGOS, M.T. (2003): Clustering in the Presence of Outliers. In: M. Schwaiger and O. Opitz (Eds.): *Exploratory Data Analysis in Empirical Research*. Springer, Berlin, 58–66.
- GARCIA-ESCUDERO, L.A. and GORDALIZA, A. (1999): Robustness Properties of k Means and Trimmed k Means. *Journal of the American Statistical Association*, *94*, 956–969.
- HAMEL, F.R. (1971): A General Qualitative Definition of Robustness. *Annals of Mathematical Statistics*, *42*, 1887–1896.
- HENNIG, C. (2002): *Breakdown points for Maximum Likelihood-estimators of location-scale mixtures*, Research Report No. 105, Seminar für Statistik, ETH-Zürich, <ftp://ftp.stat.math.ethz.ch/Research-Reports/105.html>. To appear in *Annals of Statistics*.
- KHARIN, Y. (1996): *Robustness in Statistical Pattern Recognition*, Kluwer Academic Publishers, Dordrecht.
- MCLACHLAN, G.J. and PEEL, D. (2000): *Finite Mixture Models*, Wiley, New York.
- PEEL, D. and MCLACHLAN, G.J. (2000): Robust mixture modeling using the t distribution. *Statistics and Computing*, *10*, 335–344.
- REDNER, R.A. and WALKER, H.F. (1984): Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, *26*, 195–239.
- SCHWARZ, G. (1978): Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

On the Modification of the David-Hellwig Test

Grzegorz Konczak

Department of Statistics,
The Karol Adamiecki University of Economics in Katowice,
ul. Bogucicka 14, 40-226 Katowice, Poland

Abstract. This paper presents a proposal for a nonparametric test which could determine an unknown distribution. The proposal is a modification of the David-Hellwig "empty cells" test. The area of variability of an attribute is divided into m cells. Next we take a sample from a population. Then, we calculate a number of elements for each cell. In the David-Hellwig test, we calculate the value of the statistic $K_n = \text{card}\{j : m_j = 0\}$ which counts empty cells. In the proposed modification, we calculate the value of several statistics, which are based on the number of cells which contain $0, 1, \dots, k$ (k is the parameter of this test) elements.

1 Introduction

The goodness-of-fit tests are one of the most often used nonparametric procedures. Among the most often used tests, there are chi-square tests, the Kolmogorov test and runs tests. To verify a hypothesis about the distribution we may use an "empty cells" statistic. In the David-Hellwig test, the area of variability is divided into some cells. If the null hypothesis is true, then probabilities of elements falling into the cells are all equal. The statistics used to test the goodness-of-fit hypothesis is a number of empty cells, that is to say the number of these areas in which no element has been found. This value is compared with the critical value which is read from statistics tables.

In this paper we have introduced the proposal for a nonparametric test. This is a modification of the David-Hellwig "empty cells" test. The area of variability of a random variable, just like in the David-Hellwig test, is divided into m cells. First, we take a sample of size n from a population. Next, we calculate how many cells are empty or have one, two or more elements. On the basis of this information, we calculate the values of the test statistics. Because the function of probability of the proposed statistics is difficult to find analytically many computer simulations have been made. On the basis of these simulations, the critical values for the proposed test have been found. One of the most interesting possibilities of a practical use of this method is quality control. In the next part, possibilities of using the proposal in statistical procedures in quality control will be presented. An example of using this method in control charts will be described. The proposal may be used to detect out-of-order processes. At the end a comparison of the David-Hellwig test and the proposed modification will be described.

2 The David-Hellwig test

The David-Hellwig test is also called "empty cells" test. The "empty cells" statistic may be used to verify a hypothesis about the form of a distribution. Let us assume that a sample of n elements is taken from a population. To verify the hypothesis, that sample should come from a population with an F_0 distribution, that is to say $H_0 : F = F_0$ whereas the alternative hypothesis is $F \neq F_0$ (eg Domanski and Pruska (2000)), we use the statistics determining the number of cells in which no element of the sample has been found.

In order to test the hypothesis mentioned, we divide the area of variability of the tested variable into m disconnected cells $M_j(j = 1, 2, \dots, m)$, so that the following conditions are fulfilled:

$$\mathbb{X} = \bigcup_{i=1}^m M_j \quad (1)$$

where \mathbb{X} is the area of variability of the variable

$$M_j \cap M_i = \emptyset,$$

for $i \neq j$,

$$P(x \in M_j) = \frac{1}{m}$$

for $j = 1, 2, \dots, m$, if the hypothesis H_0 is true.

To test the null hypothesis David (1950) and Hellwig (1965) used the following statistic:

$$K_n = \text{card}\{j : m_j = 0\} \quad (2)$$

where $m_j(j = 1, 2, \dots, m)$ means the number of elements in the j -th cell. The probability function of the empty cells is known. If we mark the number of empty cells as k and the size of sample as n , the probability of k empty cells can be written down as follows

$$p_k(n, m) = \binom{m}{k} \sum_{r=0}^{m-k} (-1)^r \binom{m-k}{r} \cdot \left(1 - \frac{k+r}{m}\right)^n \quad (3)$$

and the cumulative distribution

$$P_k(n, m) = \sum_{s=0}^k \binom{m}{s} \sum_{r=0}^{m-s} (-1)^r \binom{m-s}{r} \left(1 - \frac{s+r}{m}\right)^n \quad (4)$$

For the assumed significance level the rejection region is denoted as

$$K = \{k : k \geq K_{n,\alpha}\} \quad (5)$$

where $K_{n,\alpha}$ is taken from the tables (eg. David (1950), Hellwig (1965)).

David - Hellwig test	Modification
$k_0 = 0$	$k_0 = 0 \text{ and } k_1 = 5$
$k_0 = 4$	$k_0 = 4 \text{ and } k_1 = 0$
$k_0 = 3$	$k_0 = 3 \text{ and } k_1 = 1$
$k_0 = 1$	$k_0 = 1 \text{ and } k_1 = 3$
$k_0 = 3$	$k_0 = 3 \text{ and } k_1 = 0$

Fig. 1. The graphic comparison of the David-Hellwig test with the modification ($n = m = 5$)

3 The modification of the test

Let us consider the case of a partition of the area of variability into cells whose number is not equal to the sample size. If the number of the cells is smaller than the sample size, the number of empty cells will often be 0 even in the case when the hypothesis H_0 is false. In this case, we should count the cells containing 1, 2 or more elements. When we have counted these cells we will have more information from a sample than in the David-Hellwig test, where we only count empty cells.

To test the goodness-of-fit hypothesis, we use a set of statistics connected with the number of cells which contain $0, 1, \dots, n$ elements.

$$K_0 = \text{card}\{j : m_j = 0\}$$

$$K_1 = \text{card}\{j : m_j = 1\}$$

.....

$$K_n = \text{card}\{j : m_j = n\}$$

In the examples presented we will concentrate on the case when we count the number of cells in which 0 or 1 element from the sample have been found.

Example 1.

Let us assume that the area of variability is divided into m cells. We take an n -element sample, where $n \geq m$. We take two statistics k_0 and k_1 , which denote the numbers of empty and one-element cells. These K_0 and K_1 statistics can take the values adequately $0, 1, \dots, m-1$ and $0, 1, \dots, m$, adequately. A comparison of the David-Hellwig test with the proposed modification is schematically presented in Figure 1.

For 5 hypothetical distributions of 5 elements in 5 cells, the numbers of empty cells for the David-Hellwig test and numbers of empty and one-element cells for the proposed modification have been calculated. We can easily see that events of occurrence of empty cells and one-element cells are mutually dependent. For the case $m = n = 5$, we have the following implications:

$$k_0 = 0 \Rightarrow k_1 = 5$$

$$k_0 = 4 \Rightarrow k_1 = 0$$

$$k_0 = 1 \Rightarrow k_1 = 3$$

$$k_1 = 3 \Rightarrow k_0 = 1$$

From the last two conditions, we have

$$k_0 = 1 \Leftrightarrow k_1 = 3.$$

We have also

$$k_0 = 2 \Rightarrow (k_1 = 1 \text{ or } k_1 = 2)$$

$$k_0 = 3 \Rightarrow (k_1 = 1 \text{ or } k_1 = 0).$$

4 The rejection region

The scheme of construction of the rejection region for the proposed modification is presented below:

1. We specify a significance level.
2. For every possible $(k_0, k_1, \dots, k_n)_i$ we find the probability p_i (the probability can be calculated analytically or found by means of using a computer simulation).
3. Put $(k_0, k_1, \dots, k_n)_i$ in such an order that the $p_{(i)}$ are from smallest to largest.
4. We add $(k_0, k_1, \dots, k_n)_{(i)}$ to the rejection region K while

$$p_{(1)} + p_{(2)} + \dots + p_{(i)} \leq \alpha \quad (6)$$

If we have two sequences: $(k_0, k_1, \dots, k_n)_i$ and $(k_0, k_1, \dots, k_n)_j$ for whom we have $p_i = p_j$ we add both of them or none of them to the rejection region taking (6) into account. This guarantees us that the rejection region is unique.

We mark the probabilities of all possible occurrences of empty and one-element cells to n -element cells $(k_0, k_1, \dots, k_n)_{(i)}$ as $p_{(i)}$. Then, we find

$$i^* = \max_{\sum_{j=1}^i p_{(j)} \leq \alpha} i \quad (7)$$

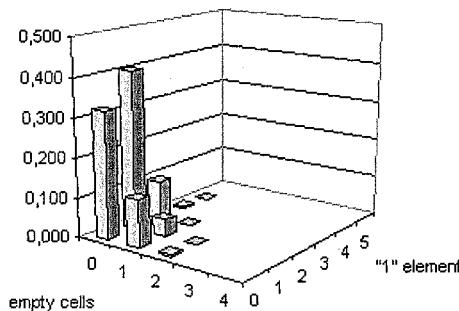
The rejection region can be written down in the following way:

$$K = \bigcup_{j=1}^{i^*} (k_0, k_1, \dots, k_n)_{(j)} \quad (8)$$

In the general case, the rejection region may be not convex, as well as in the David-Hellwig test, where we reject the null hypothesis if we have too few or too many empty cells.

Example 2.

Let us assume that we have a sample of size 15 and the area of variability is divided into 5 cells. First, we calculate the probabilities p_i for every pair $(k_0, k_1)_i$. The probabilities are shown in Table 1 and they are presented in Figure 2 as well as in Table 1. Then, we put these pairs in such an order that the probabilities are from smallest to largest (see Table 2). We also calculate cumulative sums of probabilities $p_{(i)}$. Using these cumulative probabilities, we can construct the rejection region for a given significance level. The rejection region for the significance level 0.1 is presented in Figure 3.



k_0	k_1						Sum
	0	1	2	3	4	5	
0	0.322	0.404	0.099	0.004	0.000	X	0.82907
1	0.121	0.043	0.002	0	X	X	0.16630
2	0.004	0.000	0	X	X	X	0.00462
3	0	0	X	X	X	X	0
4	0	X	X	X	X	X	0
Sum	0.447	0.448	0.101	0.004	0.000	0.000	1

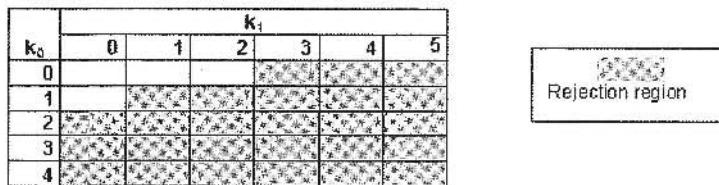
X denotes impossible events

Table 1. The distribution function of empty and one element cells ($n = 15, m = 5$)

The rejection region can be written down as follows:

$$K = \{(k_0, k_1) : \{0, 1, 2, 3, 4\} \times \{0, 1, 2, 3, 4\} - \{(0, 0), (0, 1), (0, 2), (1, 0)\}\} \quad (9)$$

$(k_0, k_1)_i$	(4,0),(3,0) (3,1),(2,2),(2,1) (1,3),(0,4)	(1,2)	(0,3)	(2,0)	(1,1)	(0,2)	(1,0)	(0,0)	(0,1)
p_i	0.000	0.002	0.004	0.004	0.043	0.099	0.121	0.322	0.404
$Cump_i$	0.000	0.002	0.006	0.010	0.053	0.152	0.273	0.595	1.000

Table 2. The scheme of designing the rejection region for $n = 15$ and $m = 5$ Fig. 2. The graphic display of the critical area for significance level $\alpha = 0.1$ and $n=15$

In Figure 3 a critical area for the considered case ($m = 5$ and $n = 15$) is schematically introduced. In the classic form of the David-Hellwig goodness-of-fit test, we reject the null hypothesis when the number of empty cells is equal or greater than 2. In the proposed modification we reject the null hypothesis in the same case and, additionally, when the number of one-element cells is equal or greater than 3 and when $k_0 = 1$ and $(k_1 = 1 \text{ or } k_1 = 2)$.

5 The results of computer simulations

The results of computer simulations are presented in this part. Using these results of computer simulations, the rejection regions have been found. The result for the David-Hellwig test and for the proposed modification have been compared in the case of detecting out-of-order processes in quality control procedures.

Sample size (n)	α								
	0.1			0.05			0.01		
	k_0	k_1	(k_0, k_1)	k_0	k_1	(k_0, k_1)	k_0	k_1	(k_0, k_1)
15	2	3	(1,1), (1,2)	2	3	(1,2)	2	4	(1,2), (1,3)
16	2	3	(1,1), (1,2)	2	3	(1,1), (1,2)	2	3	(1,2)
17	2	2	(1,1)	2	3	(1,1), (1,2)	2	3	(1,2)
18	2	2	(1,1)	2	2	(1,1)	2	3	(1,2)
19	1	2	-	2	2	(1,1)	2	3	(1,1), (1,2)
20	1	2	-	2	2	(1,1)	2	3	(1,1), (1,2)

Table 3. Critical values - the results of computer simulations

100,000 samples, from normal distribution, of size n have been generated (eg Brandt (1997)). The expected value of this distribution was 100 and standard deviations equals 5. The sample size was 15 to 20. The samples were generated in such a way that each element of the sample falls into a given cell with the same probability 0.2 (since we have 5 cells). The frequencies of events for every pair (k_0, k_1) have been found. The Table 3 presents the critical values k_0 and k_1 for 0.1, 0.05 and 0.01. In the case $n = 15$ and $\alpha = 0.1$ we reject the null hypothesis (see Table 3 and Figure 3) if $k_0 \geq 2$ or if $k_1 \geq 3$ or if $(k_0, k_1) \in \{(1, 1), (1, 2)\}$.

6 The comparison of the David-Hellwig test with its modification

In the nineteen twenties, W. A. Shewhart introduced control charts. The control charts are used to monitor technological processes. In a classical view, if a process is running correctly there are no signals on control charts. If the process is out of order, the probabilities of occurrence of empty cells or k element cells are changing. So, we can use the David-Hellwig test and its modification to monitor processes.

Let us consider the fact that the observed random variable has a normal distribution with a mean equal 100 and a standard deviation equal 5. We take a sample of size 15. The area of variability of X is divided into 5 such cells, that the probabilities that an element falls into a cell are the same for every cell. Using quantiles of a normal distribution, we have the following cells:

$$M_1 = (-\infty; -0.842)$$

$$M_2 = (-0.842; -0.253)$$

$$M_3 = (-0.253; 0.253)$$

$$M_4 = (0.253; 0.842)$$

$$M_5 = (0.842; \infty)$$

If the process runs correctly, then

$$P(x \in M_1) = P(x \in M_2) = \dots = P(x \in M_5)$$

In the David-Hellwig test case, we reject the null hypothesis for the significance level 0.01 if the number of empty cells is greater or equal to 2. In the modification, we define the rejection region on the basis of the information from Table 3 and we reject the null hypothesis if k_0 is greater or equal to 2 or k_1 is greater or equal to 4 or if k_0 equals 1 and k_1 equals 2 or 3.

The results of computer simulations for some out-of-order processes are presented in Table 4. Three kinds of process shifts have been generated.

The first one is connected with means shift, the second one with a standard deviation shift and the third one with the shift of both mean and standard deviation.

n	$X : N(\mu + \sigma, \sigma)$		$X : N(\mu, 1.1\sigma)$		$X : N(\mu + 0.1\sigma, 1.1\sigma)$	
	David-Hellwig	Modification	David-Hellwig	Modification	David-Hellwig	Modification
15	0.382	0.695	0.011	0.021	0.009	0.139
16	0.364	0.676	0.007	0.106	0.006	0.107
17	0.323	0.635	0.004	0.075	0.004	0.076
18	0.285	0.598	0.002	0.055	0.003	0.058
19	0.260	0.816	0.001	0.055	0.001	0.054
20	0.227	0.792	0.001	0.040	0.001	0.036

Table 4. The probabilities of occurrence of out-of-order signals for the David-Hellwig test and its modification

We can notice that the proposed modification gives signals of irregular processes more often than the David-Hellwig test (see the results in Table 4).

7 Concluding remarks

The David-Hellwig test is a well-known goodness-of-fit test. This test uses only part of the information, which we obtain from the sample. The proposed modification enables us to use more information from the sample than the David-Hellwig test. Using numbers of one-element cells, two-element cells etc., we can get more precise information about the tested distribution. Series of computer simulations have been made. As result of these simulations, we have got the rejection regions for various sample sizes and significance levels. If the null hypothesis is false, then the modification more often leads to the rejection of this hypothesis.

The proposed test is easy to use and has a readable graphic view. This test can be used to monitor processes in quality control and wherever the David-Hellwig goodness-of-fit test is used.

References

- BRANDT, S. (1997): *Statistical and Computational Methods in Data Analysis*. Springer Verlag, New York.
- DAVID, F.N. (1950): *Order Statistics*. J. Wiley & Sons Inc., New York.
- DOMANSKI, Cz. and PRUSKA, K. (2000): *Nieklassyczne metody statystyczne*, PWE, Warszawa.
- HELLWIG, Z. (1965): Test zgodnosci dla malej proby. *Przeglad Statystyczny*, 12, 99–112.

Simultaneous Selection of Variables and Smoothing Parameters in Additive Models

Rüdiger Krause and Gerhard Tutz

Institut für Statistik,
Ludwig-Maximilians Universität München, Akademiestr.1, D-80799 München,
Germany

Abstract. For additive models of the type $y = f_1(x_1) + \dots + f_p(x_p) + \epsilon$ where $f_j, j = 1, \dots, p$, have unspecified functional form the problem of variable selection is strongly connected to the choice of the amount of smoothing used for components f_j . In this paper we propose the simultaneous choice of variables and smoothing parameters based on genetic algorithms. Common genetic algorithms have to be modified since inclusion of variables and smoothing have to be coded separately but are linked in the search for optimal solutions. The basic tool for fitting the additive model is the expansion in B-splines. This approach allows for direct estimates which is essential for the method to work.

1 Introduction

The problem of variable selection (or subset selection) arises when the relationship between a response variable and a subset of potential explanatory variables is to be modelled, but there is substantial uncertainty about the relevance of the variables. In many statistical applications (e.g. analysis of gene expression data) there are large sets of explanatory variables which contain many redundant or irrelevant variables. Hence these applications dependent on approaches of variable selection. Beside variable selection we are also interested in appropriate estimation of the additive terms. In this paper we choose the approach of using a large number of basis functions with penalization of the coefficients. The danger of overfitting, resulting in wiggly estimated curves, is avoided by introducing a penalty term, characterized by a smoothing parameter λ . The smoothing parameter controls the influence of the penalty term and hence the smoothness of the estimated function. A large parameter value yields smooth estimates (e.g. $\lambda \rightarrow \infty$ leads to a linear estimator). In contrast, a small parameter value yields wiggly estimated curves (the extreme case is an interpolation of data for $\lambda = 0$). To prevent over- and underfitting, respectively, of data accurate choice of the smoothing parameter is essential.

Many software packages have separate tools for variable selection and appropriate smoothing parameter choice which are applied successively. In this paper we propose simultaneous selection of variables and smoothing parameters by application of genetic algorithms (e.g. Goldberg (1989)).

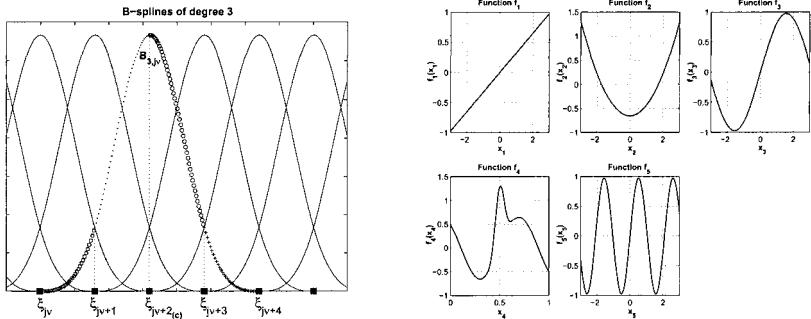


Fig. 1. The left panel shows B-splines of degree 3 for equally spaced knots. For one B-spline the different polynomials are exemplarily plotted. The right panel shows the five original functions (with effect) of the additive model used in the simulation study of section 5.

2 Additive model and expansion in a B-spline basis

A popular approach which assumes some structure in the predictor space is the additive model (Hastie and Tibshirani (1990)). For observations $(y_i, \mathbf{x}_i), i = 1, \dots, n$, where each \mathbf{x}_i is a vector of p components $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$. The response variable y_i depends on \mathbf{x}_i by $y_i = \beta_0 + \sum_{j=1}^p f_j(x_{ij}) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Thus the additive model replaces the problem of estimating a function f of a p -dimensional variable \mathbf{x}_i by estimating p separate one-dimensional functions f_j . The advantage of the additive model is its potential as a data analytic tool: since each variable is represented separately one can plot the p coordinate functions separately and thus evaluate the roles of the single predictors.

An approach which allows flexible representations of the functions f_j is the *expansion in basis functions*. The function e.g. f_j is represented as $f_j(x_{ij}) = \sum_{\nu=1}^{K_j} \beta_{j\nu} \phi_{j\nu}(x_{ij})$, where the $\beta_{j\nu}$ are unknown coefficients and $\{\phi_{j\nu}(x_{ij}), \nu = 1, \dots, K_j, K_j \geq 1\}$ is a set of basis functions. Each basis function $\phi_{j\nu}(x_{ij})$ is characterized by a knot $\xi_{j\nu}$ which is from the range of the j th covariate. As basis functions we use B-splines of degree 3 or order 4, respectively (Figure 1). These so called cubic B-splines are generated by four polynomials of degree 3 which are joint at the inner knots ($\xi_{j,\nu+1}$ and $\xi_{j,\nu+2}$). In this case the first and the second derivatives are equal at the joining points. A detailed presentation about B-splines is given in de Boor (1978).

3 Estimation with penalized shrinkage

For the additive model the parameters are estimated by minimizing the *penalized residual sum of squares (pRSS)*

$$\min_{\beta} \left\{ \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \sum_{\nu=1}^{K_j} \beta_{j\nu} \Phi_{j\nu}(x_{ij}))^2 + \tau(\{\lambda_j\}) \right\} \quad (1)$$

where $\tau(\{\lambda_j\}) = \sum_{j=1}^p \lambda_j \sum_{\nu=d+1}^{K_j} (\Delta^d \beta_{j,\nu})^2$ denotes the penalty term and $\lambda_j \geq 0$, $j = 1, \dots, p$, are smoothing parameters that control the amount of shrinkage: the larger the values of λ_j , the larger the amount of shrinkage. The penalization is the same as in Eilers and Marx (1996). In (1) the expression $\Delta^d \beta_{j,\nu+1}, d = 1, 2, \dots$ denotes the d th difference, e.g. the 2th difference has the form $\Delta^2 \beta_{\nu+1} = \Delta^1(\beta_{\nu+1} - \beta_\nu) = (\beta_{\nu+1} - 2\beta_\nu + \beta_{\nu-1})$. It can be shown (Krause and Tutz (2003)) that the estimator $\hat{\beta}(\Lambda)$ which minimizes (1) has the form $\hat{\beta}(\Lambda) = (\mathbf{B}^T \mathbf{B} + \mathbf{D}^T \Lambda \mathbf{D})^{-1} \mathbf{B}^T \mathbf{y}$, where \mathbf{B} is a design matrix of dimension $n \times [(K_1 - 1) + \dots + (K_p - 1)] + 1$, \mathbf{D} is a $[(K_1 - d) + \dots + (K_p - d)] + 1 \times [(K_1 - 1) + \dots + (K_p - 1)] + 1$ -penalization matrix and $\Lambda = \text{diag}(0, \lambda_1, \dots, \lambda_p)$ is a smoothing matrix of dimension $[(K_1 - d) + \dots + (K_p - d)] + 1 \times [(K_1 - d) + \dots + (K_p - d)] + 1$.

The performance of the penalized estimate strongly depends on the choice of the smoothing parameters $\lambda_j, j = 1, \dots, p$. A criterion with favourable properties has been proposed by Hurvich and Simonoff (1998). It is given by

$$AIC_{imp} = \log \left[\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right] + 1 + 2 \cdot \frac{[tr(H) + 1]}{n - tr(H) - 2} \quad (2)$$

where $\mathbf{H} = \mathbf{B}(\mathbf{B}^T \mathbf{B} + \mathbf{D}^T \Lambda \mathbf{D})^{-1} \mathbf{B}^T$ is the hat matrix. The smoothing parameters have to be chosen such that the criterion becomes minimal.

4 Selection of variables and smoothing parameters by genetic algorithms

Genetic Algorithms (Goldberg (1989)) are originally based on Darwin's evolution theory which refers to the principle that better adapted (fitter) individuals win against their competitors under equal external conditions. For some background on the biological processes of genetics and the origin of the terminology see e.g. Mitchell (1996).

The smallest units linked to relevant information of a genetic algorithm are called *genes*. The genes are either single units or short blocks of adjacent units and the information is coded in form of numbers, characters, or other symbols. In our case of simultaneous optimization of variable selection and smoothing parameter choice we use two different codings: 0-1- respectively real-value coding. Usually several genes are arranged in a linear succession which is called *string*.

Given the metrical variables x_1, \dots, x_p the coding of the inclusion of variables is given by

$$\delta_i = \begin{cases} 1 & \text{if variable } x_i \text{ is included} \\ 0 & \text{else} \end{cases}, \quad i = 1, \dots, p.$$

The parameters λ_i denote the real-coded smoothing parameters for the variables x_i . In the present problem a string is a combination of a 0-1 string $\delta = (\delta_1, \dots, \delta_p)^T$ and a real-valued string $\lambda = (\lambda_1, \dots, \lambda_p)^T$, i.e. (δ, λ) . The genetic algorithm always uses several strings as a potential solution of an optimization problem. This collection of strings is called *population*. If we apply operators to strings we generate a population with new different strings. This new population of strings is called *offspring*. The function which has to be optimized by the genetic algorithm is called fitness function (short *fitness*). Here the fitness bases on the criterion in (2).

Crossover of indicator variables

Suppose we have two 0-1 strings $\delta = (\delta_1 \dots \delta_i \dots \delta_k)$ and $\bar{\delta} = (\bar{\delta}_1 \dots \bar{\delta}_i \dots \bar{\delta}_k)$ with indicator variables. Based on a freely chosen crossover probability p_c the crossover procedure yields two offspring δ' and $\bar{\delta}'$ with

$$\begin{aligned}\delta'_i &= 0 \wedge \bar{\delta}'_i = 0 \quad \text{if } \delta_i + \bar{\delta}_i = 0 \\ \delta'_i &= 1 \wedge \bar{\delta}'_i = 1 \quad \text{if } \delta_i + \bar{\delta}_i = 2 \\ \delta'_i &= 0 \wedge \bar{\delta}'_i = 1 \quad \text{if } \delta_i + \bar{\delta}_i = 1 \text{ and } \tau_i < 0.5 \\ \delta'_i &= 1 \wedge \bar{\delta}'_i = 0 \quad \text{if } \delta_i + \bar{\delta}_i = 1 \text{ and } \tau_i \geq 0.5\end{aligned}$$

and $\tau_i \in [0, 1], i = 1, \dots, k$ are uniformly distributed random numbers and p_c determines which strings of the population are selected for crossover. A string is used for crossover operation if $r_i < p_c$ holds, with a random number $r_i \in [0, 1], i = 1, \dots, popsize$ (population size). In the crossover process we need couples of strings and thus it is necessary to select an even number of parent strings. We denote this crossover procedure as *SelCross*-procedure. During crossover process smoothing parameters are only determined for the new indicator variables δ'_i and $\bar{\delta}'_i$ which are selected (i.e. $\delta'_i = 1, \bar{\delta}'_i = 1$).

Crossover of smoothing parameters

Here we shortly sketch a crossover operator, called *improved arithmetical crossover (IAC)* (see Krause and Tutz (2003)). Suppose we have two strings $\lambda = (\lambda_1 \dots \lambda_i \dots \lambda_k)$ and $\bar{\lambda} = (\bar{\lambda}_1 \dots \bar{\lambda}_i \dots \bar{\lambda}_k)$ with values in an interval $[l_{lo}, l_{up}]$ with lower limit l_{lo} and upper limit l_{up} . The IAC operator yields three offspring λ' , λ'' and λ''' . Each gene of the first offspring λ'_i is located in the parents' interval $[\lambda_i, \bar{\lambda}_i]$. The other children are randomly positioned left and right outside the interval $[\lambda_i, \bar{\lambda}_i]$. Every string only takes values within the range between l_{lo} and l_{up} .

$$\begin{aligned}\lambda'_i &= a\lambda_i + (1 - a)\bar{\lambda}_i \quad i = 1, \dots, k \\ \lambda''_i &= \bar{\lambda}_i + b_1(l_{up} - \bar{\lambda}_i), \quad i = 1, \dots, k \\ \lambda'''_i &= \lambda_i - b_2(\lambda_i - l_{lo}), \quad i = 1, \dots, k\end{aligned}$$

where $a \in [0, 1]$ can be chosen constant or variable over the number of iterations. The parameters $b_i \in [0, 1], i = 1, 2$, are uniformly distributed random numbers. Every string takes values in the default interval $[l_{lo}, l_{up}]$. The IAC operator is again specified by the crossover probability p_c .

This operator type has been constructed for the case that the whole population of strings consists of the same number of variables and hence the fitness only depends on the different smoothing parameter values. But here two strings containing different combinations of variables generally yield different fitness values. Now we modify the IAC-operator in such a way that for each of the two parent strings, taking part in the crossover process, the fitness of the three offspring is calculated separately. Then the best offspring of each group replaces the appropriate parent string. We denote the modified crossover operator as *modified improved arithmetical crossover (modIAC)*.

Mutation of indicator variables

The genes of a parent string $\delta = (\delta_1 \dots \delta_i \dots \delta_k)$ are mutated yielding a new string $\tilde{\delta} = (\tilde{\delta}_1 \dots \tilde{\delta}_i \dots \tilde{\delta}_k)$. Thereby for every gene of a string we generate a random number $r_{gene} \in [0, 1]$ and compare r_{gene} with a default mutation probability p_m . If $r_{gene} < p_m$ the gene mutates, i.e. it changes its value from 0 to 1 or vice versa. We denote this mutation procedure of indicator variables as *SelMut*-procedure. For the case that a mutated gene $\tilde{\delta}_i = 1$ but the parent gene $\delta_i = 0$ there is no smoothing parameter λ_i available. The corresponding λ_i is randomly chosen from the default smoothing parameter interval.

Mutation of smoothing parameters

For each parent indicator gene $\delta_i = 1$ the smoothing parameter λ_i is mutated by the *non-uniform mutation* operator presented by Michalewicz (1996). A randomly selected gene λ_i mutates in

$$\tilde{\lambda}'_i = \begin{cases} \lambda_i + (l_{up} - \lambda_i)(1 - r^{(1 - \frac{\tau}{T})^b}) & \text{if } \tau = 0 \\ \lambda_i - (\lambda_i - l_{lo})(1 - r^{(1 - \frac{\tau}{T})^b}) & \text{if } \tau = 1 \end{cases} . \quad (3)$$

Here τ is a random number which may have a value of zero or one, $r \in [0, 1]$ is an uniform random number, T is the maximum number of iterations and b is a user-dependent system parameter which determines the degree of non-uniformity. For further details see Michalewicz (1996).

The essential difference to traditional genetic algorithms is the linkage between an indicator variable δ_i and a real-valued smoothing parameter λ_i . Here we present a combined genetic algorithm for variable selection with simultaneous smoothing parameter choice. Therefore the operators described above are integrated in a selection procedure. Here we use the *modified selection procedure (modSP)* which is presented in Krause and Tutz (2003) in detail. The structure of this procedure is illustrated in Figure 2.

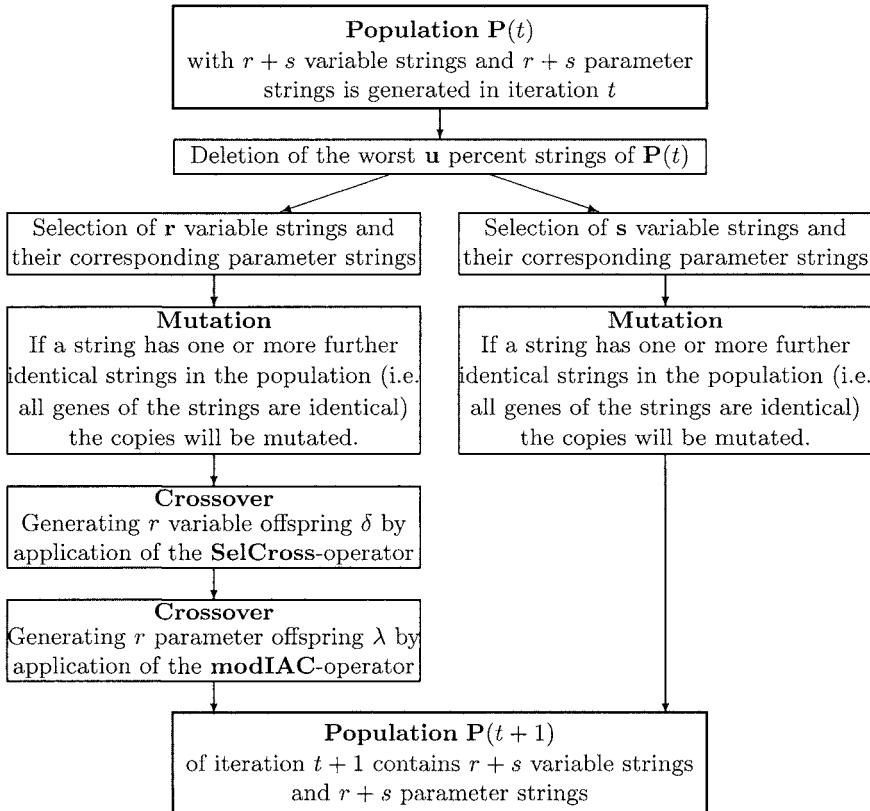


Fig. 2. Structure of a genetic algorithm for simultaneous selection of variables and smoothing parameters given as a flowchart.

5 Simulation study

Here we choose an additive model, consisting of 10 functions $f_j(x_{ij}), j = 1, \dots, 10$ where five functions have no effect, i.e. $f_j(x_{ij}) = 0$. The curves of the remaining functions are shown in Figure 1. We simulate 200 data sets, each one consists of 100 (respectively 200) independent and uniform distributed observations with $\sigma = 0.1$ (respectively $\sigma = 0.2$). For estimation the single functions $f_j(x_{ij})$ are expanded in 15 cubic B-spline basis functions. As penalty we use the third difference of adjacent coefficients. The smoothing parameters are chosen from the interval $[10^{-4}, 10^4]$. The default parameters of the used genetic algorithm are: $popsize = 48$ strings, $p_c = 0.4$, $p_m = 0.25$, deletion of $u = 60$ percent of the worst strings, selection of $r = 28$ and $s = 20$ strings, $\nu = 0.5$, $T = 1000$ and $b = 1$.

For comparison we have chosen three alternative methods: the software package S-Plus offers a restricted possibility of variable selection and simultaneous function estimation based on the AIC-criterion. By the stepwise se-

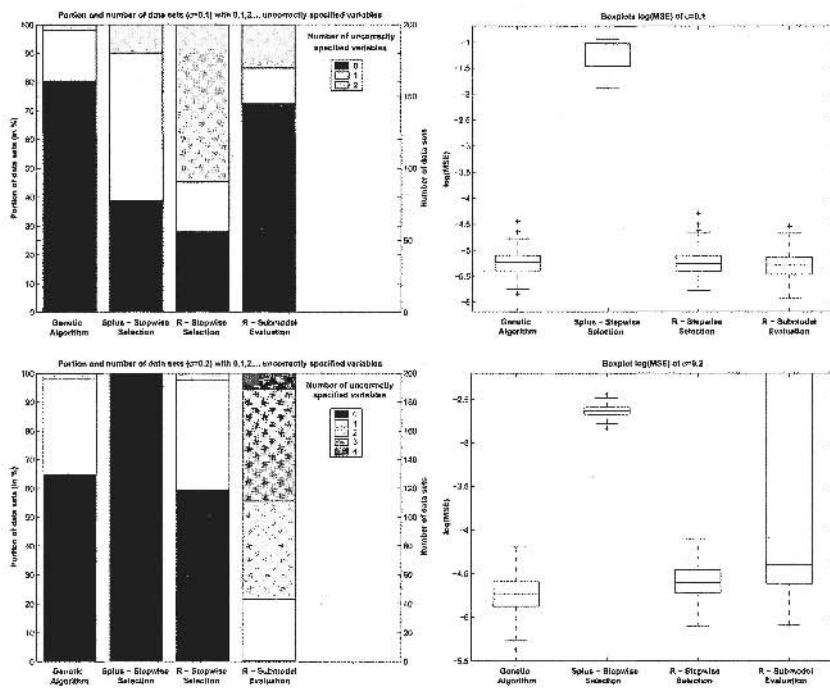


Fig. 3. The figures on the left side show the portion respectively the number of data sets with 0, 1, ... incorrectly specified variables. The figures on the right side show the appropriate boxplots of $\log(MSE)$, where $MSE(f) = \frac{1}{n} \sum_{i=1}^n (f(x_i) - \hat{f}(x_i))^2$, for the estimation of the additive function described in the text.

lection procedure **step** each covariate can be dropped or integrated in the model as a linear term respectively as a cubic smoothing spline with a default smoothing parameter (with degrees of freedom $df = 2, 6, 10, 14$). The software package R offers two approaches to variable selection. The function **stepAIC**, implemented in the package **MASS**, chooses a model by the AIC-criterion in a stepwise algorithm. Each covariate can be dropped or integrated in the current model as a linear term or as a polynomial up to degree 4. The package **wle** evaluates the AIC-criterion for each submodel and selects the model with the lowest AIC-value. Because of the immense computational cost the current model only linear terms and polynomials up to degree 2 can be used. After variable selection the R-package **mgcv** (Wood (2001)) yields an automatic smoothing parameter selection.

Figure 3 shows the results of variable selection and prediction accuracy for both cases: in case of $\sigma = 0.1$ the genetic algorithm yields the best results. In 80% of the data sets the number of incorrectly specified variables is zero. All other methods lead to larger rates of incorrectly specified variables. The prediction accuracy shows no obvious differences between the genetic algorithm and the R-programs. However S-Plus has substantial poorer per-

formance compared with the other methods. In the case of $\sigma = 0.2$ the S-Plus approach performs best in the selection of variables but its estimation results obviously lie below the other results. While genetic algorithm and the stepwise selection in R yield similar performance the submodel evaluation in R shows peculiar results: some data sets lead to good estimators whereas other ones yield unuseable estimators. Thus the procedure seems rather unstable.

6 Conclusions

This paper presents a new automatic procedure for simultaneous variable selection and smoothing parameter choice based on genetic algorithms. The simulation study shows that the genetic algorithm provides a reliable tool for variable selection and function estimation which has comparable or superior performance. Even if the S-Plus program yields better performance in variable selection its very poor function estimation compensates the good results. The differences in the performance between the approaches should be much more obvious for data sets which consist of larger sets of variables. In this paper we restricted ourselves to models with additive structure and uniformly distributed observations. In future work the genetic algorithm will be analysed in models with categorical and metrical variables as well as its extension to models with interactions. We are also interested in a generalization to non-normal distributed response.

References

- DE BOOR, C. (1978): *A Practical Guide to Splines*. Springer, New York, Heidelberg, Berlin.
- EILERS, P.H.C. and MARX, B.D. (1996): Flexible Smoothing with B-splines and Penalties. *Stat. Science*, 11(2), 89–121.
- GOLDBERG, D.E. (1989): *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.
- HASTIE, T. and TIBSHIRANI, R. (1990): *Generalized Additive Models*. Chapman and Hall, London.
- HURVICH, C.M. and SIMONOFF, J.S. (1998): Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B*, 60(2), 271–293.
- KRAUSE, R. and TUTZ, G. (2003): Additive Modeling with Penalized Regression Splines and Genetic Algorithms. *Discussion Paper Nr. 312, SFB 386, Ludwig-Maximilians-Universität München*.
- MICHALEWICZ, Z. (1996): *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, Berlin.
- MITCHELL, M. (1996): *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, Massachusetts
- RUPPERT, D. and CAROLL, R. (2000): Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics*, 42(2), 205–223.
- WOOD, S. (2001): mgcv: GAMs and Generalized Ridge Regression for R. *Rnews*, 1(2), 20–25.

Multiple Change Points and Alternating Segments in Binary Trials with Dependence

Joachim Krauth

Department of Psychology,
University of Düsseldorf, D-40225 Düsseldorf, Germany

Abstract. In Krauth (2003) we derived modified maximum likelihood estimates to identify change points and changed segments in Bernoulli trials with dependence. Here, we extend these results to the situation of multiple change points in an alternating-segments model (Halpern (2000)) and to a more general multiple change-points model. Both situations are of interest, e.g., in molecular biology when analyzing DNA sequences.

1 Introduction

In Krauth (1999, 2000) we considered tests and estimates for detecting and locating change points and changed segments in Bernoulli sequences. Here, the assumption of independence was crucial. Avery and Henderson (1999) observed that this assumption may be violated if DNA sequences or rainfall data are considered. The problem of estimating parameters in stationary Bernoulli sequences with dependence was studied by many authors (e.g. by Klotz (1973), Devore (1976), Price (1976), Lindqvist (1978), Moore (1979), Kim and Bai (1980), Budescu (1985)). It was found that the derivation of explicit expressions for the maximum likelihood estimates is not possible. Therefore, Devore (1976) considered instead a modified likelihood function where certain terms in the full likelihood were neglected and derived explicit modified maximum likelihood estimates. Following an argument by Billingsley (1961, pp. 4–5) this is reasonable in case of long sequences.

The estimates derived by the authors above cannot be used for the change-point and changed-segment problems, because here the assumption of stationarity no longer holds. Therefore, we derived in Krauth (2003) modified maximum likelihood estimates for these problems and applied the results to five binary sequences derived from a nucleotide sequence which is 1,200 nt in length and was reported by Robb et al. (1998, Fig. 1).

One obvious problem when applying the estimates above to real data is the model assumption that only one change point or only one changed segment, respectively, exists in the sequence. In most cases, this assumption is difficult to justify. In the situation with independent trials several authors, e.g. Fu and Curnow (1990) for Bernoulli variables or Hawkins (2001) in a more general context for exponential families derived estimates for multiple change-points models. By Venter and Steel (1996) and Hawkins (2001) certain

heuristics are proposed by which it is tried to identify the number of change points by means of a hierarchical sequence of statistical tests. However, the problem of the determination of the number of change points which has also been addressed e.g. by Fu and Curnow (1990) and Halpern (2000) has not found a satisfactory answer up to now.

An interesting special case of the multiple change-points model, which will also be considered here, has been introduced by Halpern (2000). In this alternating-segments model it is assumed that an observed binary sequence with values 0 and 1 is composed of two sorts of segments, alternating with each other, that differ in the probability that a given position is a 1. According to Halpern (2000) this model is of interest if we want to determine whether a given strain of a virus is likely to have arisen as the result of recombination between two donor strains. By recombination is meant the creation of a genetic chimera which may result if the genetic material of two genetically distinct viruses comes into physical contact producing a new virus with a genetic sequence consisting of alternating segments of DNA (or RNA) corresponding to pieces of the original viruses.

2 Estimates for the multiple change-points model

We consider a sequence of $n(n \geq 4)$ random variables $X_1, \dots, X_n \in \{0, 1\}$ and $m \in \{1, \dots, [\frac{n}{2} - 1]\}$ possible change points τ_1, \dots, τ_m with $0 < \tau_1 < \tau_2 < \dots < \tau_m < n$. In addition we define $\tau_0 = 0, \tau_{m+1} = n$. With m change points we have $(m + 1)$ segments and we assume that each segment has at least length 2. We define

$$\begin{aligned} P(X_i = 1) &= 1 - P(X_i = 0) = \pi_{j+1} \text{ for } \tau_j + 1 \leq i \leq \tau_{j+1}, j = 0, 1, \dots, m; \\ &\quad \tau_j \in \{\tau_{j-1} + 2, \dots, n - 2(m - j + 1)\}, \\ &\quad j = 1, 2, \dots, m, 0 < \pi_1, \pi_2, \dots, \pi_m, \pi_{m+1} < 1. \end{aligned}$$

For $m = 1$ the model with one change point (cf. Krauth (2003)) results. Further, we define first-order transition probabilities $\pi_{st,i} = P(X_i = t | X_{i-1} = s)$ for $i = 2, \dots, n; s, t \in \{0, 1\}$ and assume their stationarity within each segment. The following reparameterization is considered:

For $\tau_j + 1 < i \leq \tau_{j+1}, j = 0, 1, \dots, m$:

$$\begin{aligned} \pi_{11}(j+1) &= \pi_{11,i} = \lambda_{j+1}, \quad \pi_{10}(j+1) = \pi_{10,i} = 1 - \lambda_{j+1}, \\ \pi_{01}(j+1) &= \pi_{01,i} = \frac{(1 - \lambda_{j+1})\pi_{j+1}}{1 - \pi_{j+1}}, \\ \pi_{00}(j+1) &= \pi_{00,i} = \frac{1 - 2\pi_{j+1} + \lambda_{j+1}\pi_{j+1}}{1 - \pi_{j+1}}, \end{aligned}$$

for $i = \tau_j + 1, j = 1, \dots, m$:

$$\begin{aligned}\pi_{11}^*(\tau_j) &= \pi_{11, \tau_j+1} = \lambda_{(j)}, \\ \pi_{10}^*(\tau_j) &= \pi_{10, \tau_j+1} = 1 - \lambda_{(j)}, \\ \pi_{01}^*(\tau_j) &= \pi_{01, \tau_j+1} = \frac{\pi_{j+1} - \lambda_{(j)}\pi_j}{1 - \pi_j}, \\ \pi_{00}^*(\tau_j) &= \pi_{00, \tau_j+1} = \frac{1 - \pi_j - \pi_{j+1} + \lambda_{(j)}\pi_j}{1 - \pi_j}.\end{aligned}$$

From this, we derive the correlations

$$\begin{aligned}\rho_{k+1} &= \rho[X_i, X_{i+1}] = \frac{\lambda_{k+1} - \pi_{k+1}}{1 - \pi_{k+1}}, \\ \rho[X_i, X_j] &= \rho_{k+1}^{j-i} \text{ for } \tau_k + 1 \leq i < j \leq \tau_{k+1}, k = 0, 1, \dots, m, \\ \rho[X_{\tau_j}, X_{\tau_j+1}] &= \frac{(\lambda_{(j)} - \pi_{j+1})\pi_j}{\sqrt{\pi_j(1 - \pi_j)\pi_{j+1}(1 - \pi_{j+1})}} \text{ for } j = 1, \dots, m.\end{aligned}$$

Estimates of the correlation coefficient are a convenient measure of dependence if we are not sure that it is really necessary to consider a possible dependence of the sequence. In sequences without a change point this parameter was introduced by Lindqvist (1978).

In particular, if X_1, \dots, X_n are independent we have $\lambda_1 = \pi_1; \lambda_{j+1} = \lambda_{(j)} = \pi_{j+1}, j = 1, \dots, m$ and the multiple change-points model for Bernoulli sequences results. If we assume for the moment that τ_1, \dots, τ_m are fixed and known and if we define

$$\begin{aligned}n_{11}^j &= \sum_{i=2}^j x_{i-1}x_i, & n_{10}^j &= \sum_{i=2}^j x_{i-1}(1-x_i), \\ n_{01}^j &= \sum_{i=2}^j (1-x_{i-1})x_i, & n_{00}^j &= \sum_{i=2}^j (1-x_{i-1})(1-x_i)\end{aligned}$$

for $j = 2, \dots, n$, the modified likelihood function is given by

$$\begin{aligned}L_M(\pi_1, \dots, \pi_{m+1}, \lambda_1, \dots, \lambda_{m+1}, \tau_1, \dots, \tau_m) \\ = \{ \pi_{11}(1)^{n_{11}^{\tau_1}} (1 - \pi_{11}(1))^{n_{10}^{\tau_1}} (1 - \pi_{00}(1))^{n_{01}^{\tau_1}} \pi_{00}(1)^{n_{00}^{\tau_1}} \} \\ \prod_{j=2}^{m+1} \{ \pi_{11}(j)^{n_{11}^{\tau_j} - n_{11}^{\tau_{j-1}} - x_{\tau_{j-1}}x_{\tau_{j-1}+1}} \\ (1 - \pi_{11}(j))^{n_{10}^{\tau_j} - n_{10}^{\tau_{j-1}} - x_{\tau_{j-1}}(1 - x_{\tau_{j-1}+1})} \\ (1 - \pi_{00}(j))^{n_{01}^{\tau_j} - n_{01}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}})x_{\tau_{j-1}+1}} \\ \pi_{00}(j)^{n_{00}^{\tau_j} - n_{00}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}})(1 - x_{\tau_{j-1}+1})} \}.\end{aligned}$$

Here, the modification of the likelihood function consists in neglecting the distribution of X_1 and the transitions from X_{τ_j} to $X_{\tau_j+1}, j = 1, \dots, m$. This is in analogy to Devore (1976) and is justified for large values of n by an argument by Billingsley (1961, pp. 4–5).

The modified MLE's derived from L_M are given by

$$\begin{aligned}\hat{\lambda}_j &= \hat{\pi}_{11}(j), j = 1, \dots, m + 1, \\ \hat{\pi}_j &= \frac{1 - \hat{\pi}_{00}(j)}{2 - \hat{\pi}_{00}(j) - \hat{\pi}_{11}(j)}, j = 1, \dots, m + 1, \\ \hat{\rho}_j &= \hat{\pi}_{11}(j) + \hat{\pi}_{00}(j) - 1, j = 1, \dots, m + 1,\end{aligned}$$

where

$$\begin{aligned}\hat{\pi}_{11}(1) &= \frac{n_{11}^{\tau_1}}{n_{11}^{\tau_1} + n_{10}^{\tau_1}}, \quad \hat{\pi}_{00}(1) = \frac{n_{00}^{\tau_1}}{n_{00}^{\tau_1} + n_{01}^{\tau_1}}, \\ \hat{\pi}_{11}(j) &= \frac{n_{11}^{\tau_j} - n_{11}^{\tau_{j-1}} - x_{\tau_{j-1}}x_{\tau_{j-1}+1}}{n_{11}^{\tau_j} - n_{11}^{\tau_{j-1}} - x_{\tau_{j-1}}x_{\tau_{j-1}+1} + n_{10}^{\tau_j} - n_{10}^{\tau_{j-1}} - x_{\tau_{j-1}}(1 - x_{\tau_{j-1}+1})}, \\ j &= 2, \dots, m + 1, \\ \hat{\pi}_{00}(j) &= \frac{n_{00}^{\tau_j} - n_{00}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}})(1 - x_{\tau_{j-1}+1})}{n_{00}^{\tau_j} - n_{00}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}})(1 - x_{\tau_{j-1}+1}) + n_{01}^{\tau_j} - n_{01}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}})x_{\tau_{j-1}+1}}, \\ j &= 2, \dots, m + 1.\end{aligned}$$

Now, we drop the assumption that τ_1, \dots, τ_m are fixed and known. The modified MLE's of the change points τ_1, \dots, τ_m are those values of τ_1, \dots, τ_m for which $L_M(\hat{\pi}_1, \dots, \hat{\pi}_{m+1}, \hat{\lambda}_1, \dots, \hat{\lambda}_{m+1}, \tau_1, \dots, \tau_m)$ is maximum for $\tau_j \in \{\tau_{j-1} + 2, \dots, n - 2(m - j + 1)\}, j = 1, 2, \dots, m$.

It should be noted that our estimates of change points may be biased if higher-order dependencies are present. Further, for short sequences it may be necessary to modify the estimates given above in such a way that all estimates are well defined, and the estimates of τ_1, \dots, τ_m may not be unique.

3 Estimates for the alternating-segments model

Under the same conditions as in the preceding chapter, we define now

$$P(X_i = 1) = 1 - P(X_i = 0)$$

$$= \begin{cases} \pi_1 & \text{for } \tau_j + 1 \leq i \leq \tau_{j+1}, j = 0, 2, \dots, 2[\frac{m}{2}] \\ \pi_2 & \text{for } \tau_j + 1 \leq i \leq \tau_{j+1}, j = 1, 3, \dots, 2[\frac{m+1}{2}] - 1 \end{cases},$$

$$\tau_j \epsilon \{ \tau_{j-1} + 2, \dots, n - 2(m-j+1) \}, j = 1, 2, \dots, m, 0 < \pi_1, \pi_2 < 1.$$

For $m = 1$ the one change-point and for $m = 2$ the changed-segment model result (cf. Krauth (2003)).

Further, we define first-order transition probabilities $\pi_{st,i} = P(X_i = t | X_{i-1} = s)$ for $i = 2, \dots, n; s, t \in \{0, 1\}$ and assume their stationarity within each segment. The following reparameterization is considered:

For $\tau_j + 1 < i \leq \tau_{j+1}, j = 0, 2, \dots, 2[\frac{m}{2}]$ in case of $k = 1$ and for $\tau_{j+1} < i \leq \tau_{j+1}, j = 1, 3, \dots, 2[\frac{m+1}{2}] - 1$ in case of $k = 2$:

$$\begin{aligned}\pi_{11}(k) &= \pi_{11,i} = \lambda_k, \quad \pi_{10}(1) = \pi_{10,i} = 1 - \lambda_k, \\ \pi_{01}(k) &= \pi_{01,i} = \frac{(1 - \lambda_k)\pi_k}{1 - \pi_k}, \quad \pi_{00}(k) = \pi_{00,i} = \frac{1 - 2\pi_k + \lambda_k\pi_k}{1 - \pi_k},\end{aligned}$$

for $i = \tau_j + 1, j = 1, 3, \dots, 2[\frac{m+1}{2}] - 1$ in case of $k = 1$ and for $i = \tau_j + 1, j = 2, 4, \dots, 2[\frac{m}{2}]$ in case of $k = 2$:

$$\begin{aligned}\pi_{11}^*(\tau_k) &= \pi_{11,\tau_j+1} = \lambda_{(k)}, \quad \pi_{10}^*(\tau_k) = \pi_{10,\tau_j+1} = 1 - \lambda_{(k)}, \\ \pi_{01}^*(\tau_k) &= \pi_{01,\tau_j+1} = \frac{\pi_{3-k} - \lambda_{(k)}\pi_k}{1 - \pi_k}, \\ \pi_{00}^*(\tau_k) &= \pi_{00,\tau_j+1} = \frac{1 - \pi_1 - \pi_2 + \lambda_{(k)}\pi_k}{1 - \pi_k}.\end{aligned}$$

From this, we derive the correlations

$$\rho_k = \rho[X_i, X_{i+1}] = \frac{\lambda_k - \pi_k}{1 - \pi_k}, \quad \rho[X_i, X_j] = \rho_k^{j-i}$$

for $\tau_j + 1 \leq i < j \leq \tau_{j+1}, j = 0, 2, \dots, 2[\frac{m}{2}]$ in case of $k = 1$ and for $\tau_j + 1 \leq i < j \leq \tau_{j+1}, j = 1, 3, \dots, 2[\frac{m+1}{2}] - 1$ in case of $k = 2$,

$$\rho[X_{\tau_j}, X_{\tau_{j+1}}] = \frac{(\lambda_{(k)} - \pi_{3-k})\pi_k}{\sqrt{\pi_1(1 - \pi_1)\pi_2(1 - \pi_2)}}$$

for $j = 1, 3, \dots, 2[\frac{m+1}{2}] - 1$ in case of $k = 1$ and for $j = 2, 4, \dots, 2[\frac{m}{2}]$ in case of $k = 2$.

In particular, if X_1, \dots, X_n are independent, we have $\lambda_1 = \lambda_{(2)} = \pi_1, \lambda_2 = \lambda_{(1)} = \pi_2$ and the alternating-segments model for Bernoulli sequences results.

If we assume for the moment that τ_1, \dots, τ_m are fixed and known, and if we define $n_{11}^j, n_{10}^j, n_{01}^j$, and n_{00}^j as above, the modified likelihood function is given by

$$L_M(\pi_1, \pi_2, \lambda_1, \lambda_2, \tau_1, \dots, \tau_m)$$

$$\begin{aligned}
&= \{\pi_{11}(1)^{n_{11}^{\tau_1}} (1 - \pi_{11}(1))^{n_{10}^{\tau_1}} (1 - \pi_{00}(1))^{n_{01}^{\tau_1}} \pi_{00}(1)^{n_{00}^{\tau_1}}\} \\
&\quad \prod_{j \in \{3, 5, \dots, 2[\frac{m}{2}] + 1\}} \{\pi_{11}(1)^{n_{11}^{\tau_j} - n_{11}^{\tau_{j-1}} - x_{\tau_{j-1}} x_{\tau_{j-1}+1}} \\
&\quad (1 - \pi_{11}(1))^{n_{10}^{\tau_j} - n_{10}^{\tau_{j-1}} - x_{\tau_{j-1}} (1 - x_{\tau_{j-1}+1})} \\
&\quad (1 - \pi_{00}(1))^{n_{01}^{\tau_j} - n_{01}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}}) x_{\tau_{j-1}+1}} \\
&\quad \pi_{00}(1)^{n_{00}^{\tau_j} - n_{00}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}}) (1 - x_{\tau_{j-1}+1})}\} \\
&\quad \prod_{j \in \{2, 4, \dots, 2[\frac{m+1}{2}\}]}} \{\pi_{11}(2)^{n_{11}^{\tau_j} - n_{11}^{\tau_{j-1}} - x_{\tau_{j-1}} x_{\tau_{j-1}+1}} \\
&\quad (1 - \pi_{11}(2))^{n_{10}^{\tau_j} - n_{10}^{\tau_{j-1}} - x_{\tau_{j-1}} (1 - x_{\tau_{j-1}+1})} \\
&\quad (1 - \pi_{00}(2))^{n_{01}^{\tau_j} - n_{01}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}}) x_{\tau_{j-1}+1}} \\
&\quad \pi_{00}(2)^{n_{00}^{\tau_j} - n_{00}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}}) (1 - x_{\tau_{j-1}+1})}\},
\end{aligned}$$

where the modification consists again in neglecting certain terms.

The modified MLE's derived from L_M are given by $\hat{\lambda}_i = \hat{\pi}_{11}(i)$, $\hat{\pi}_i = \frac{1 - \hat{\pi}_{00}(i)}{2 - \hat{\pi}_{00}(i) - \hat{\pi}_{11}(i)}$, $\hat{\rho}_i = \hat{\pi}_{11}(i) + \hat{\pi}_{00}(i) - 1$, $i = 1, 2$ where

$$\hat{\pi}_{11}(1) =$$

$$\frac{n_{11}^{\tau_1} + \sum_{j \in \{3, 5, \dots, 2[\frac{m}{2}] + 1\}} (n_{11}^{\tau_j} - n_{11}^{\tau_{j-1}} - x_{\tau_{j-1}} x_{\tau_{j-1}+1})}{n_{11}^{\tau_1} + n_{10}^{\tau_1} + \sum_{j \in \{3, 5, \dots, 2[\frac{m}{2}] + 1\}} (n_{11}^{\tau_j} - n_{11}^{\tau_{j-1}} - x_{\tau_{j-1}} x_{\tau_{j-1}+1} + n_{10}^{\tau_j} - n_{10}^{\tau_{j-1}} - x_{\tau_{j-1}} (1 - x_{\tau_{j-1}+1}))},$$

$$\hat{\pi}_{00}(1) =$$

$$\begin{aligned}
&\{n_{00}^{\tau_1} + \sum_{j \in \{3, 5, \dots, 2[\frac{m}{2}] + 1\}} (n_{00}^{\tau_j} - n_{00}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}}) (1 - x_{\tau_{j-1}+1}))\} / \\
&\{n_{00}^{\tau_1} + n_{01}^{\tau_1} + \sum_{j \in \{3, 5, \dots, 2[\frac{m}{2}] + 1\}} (n_{00}^{\tau_j} - n_{00}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}}) (1 - x_{\tau_{j-1}+1})) \\
&\quad + n_{01}^{\tau_j} - n_{01}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}}) x_{\tau_{j-1}+1})\},
\end{aligned}$$

$$\hat{\pi}_{11}(2) = \frac{\sum_{j \in \{2, 4, \dots, 2[\frac{m+1}{2}\]}\} (n_{11}^{\tau_j} - n_{11}^{\tau_{j-1}} - x_{\tau_{j-1}} x_{\tau_{j-1}+1})}{\sum_{j \in \{2, 4, \dots, 2[\frac{m+1}{2}\]}\} (n_{11}^{\tau_j} - n_{11}^{\tau_{j-1}} - x_{\tau_{j-1}} x_{\tau_{j-1}+1} + n_{10}^{\tau_j} - n_{10}^{\tau_{j-1}} - x_{\tau_{j-1}} (1 - x_{\tau_{j-1}+1}))},$$

$$\hat{\pi}_{00}(2) =$$

$$\frac{\sum_{j \in \{2, 4, \dots, 2[\frac{m+1}{2}\]}\} (n_{00}^{\tau_j} - n_{00}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}}) (1 - x_{\tau_{j-1}+1}))}{\sum_{j \in \{2, 4, \dots, 2[\frac{m+1}{2}\]}\} (n_{00}^{\tau_j} - n_{00}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}}) (1 - x_{\tau_{j-1}+1}) + n_{01}^{\tau_j} - n_{01}^{\tau_{j-1}} - (1 - x_{\tau_{j-1}}) x_{\tau_{j-1}+1})}.$$

Now we drop the assumption that τ_1, \dots, τ_m are fixed and known. The modified MLE's of the change points τ_1, \dots, τ_m are those values of τ_1, \dots, τ_m

for which $L_M(\hat{\pi}_1, \hat{\pi}_2, \hat{\lambda}_1, \hat{\lambda}_2, \tau_1, \dots, \tau_m)$ is maximum for $\tau_j \in \{\tau_{j-1} + 2, \dots, n - 2(m-j+1)\}$, $j = 1, 2, \dots, m$. The remarks at the end of Section 2 are also valid for this section.

4 Example

In Robb et al. (1998, Fig. 1) a nucleotide sequence is reported which 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 (2003) 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. For this sequence we identified $m = 1$ to $m = 4$ change points as described above. The results are given in Table 1. Further, we identified for the same sequence $m = 1$ to $m = 4$ change points for the alternating-segments model as given in Table 2. Obviously, both kinds of analysis must yield the same results for $m = 1$. The estimates of the correlations indicate that for $m \geq 2$ the assumption of a certain dependence of the trials cannot be ruled out for this sequence.

$m = 1$	$lL_M = -665.207$	$\hat{\tau}_1 = 865$	
	$\hat{\pi}_1 = .215$	$\hat{\pi}_2 = .356$	
	$\hat{\lambda}_1 = .274$	$\hat{\lambda}_2 = .353$	
	$\hat{\rho}_1 = .075$	$\hat{\rho}_2 = -.005$	
$m = 2$	$lL_M = -653.954$	$\hat{\tau}_1 = 1005$	$\hat{\tau}_2 = 1041$
	$\hat{\pi}_1 = .229$	$\hat{\pi}_2 = .543$	$\hat{\pi}_3 = .361$
	$\hat{\lambda}_1 = .287$	$\hat{\lambda}_2 = .211$	$\hat{\lambda}_3 = .411$
	$\hat{\rho}_1 = .075$	$\hat{\rho}_2 = -.727$	$\hat{\rho}_3 = .077$
$m = 3$	$lL_M = -648.459$	$\hat{\tau}_1 = 1005$	$\hat{\tau}_2 = 1018$
	$\hat{\pi}_1 = .229$	$\hat{\pi}_2 = .500$	$\hat{\pi}_3 = .591$
	$\hat{\lambda}_1 = .287$	$\hat{\lambda}_2 = .000$	$\hat{\lambda}_3 = .308$
	$\hat{\rho}_1 = .075$	$\hat{\rho}_2 = -1.000$	$\hat{\rho}_3 = -.692$
$m = 4$	$lL_M = -639.858$	$\hat{\tau}_1 = 799$	$\hat{\tau}_2 = 853$
	$\hat{\pi}_1 = .228$	$\hat{\pi}_2 = .019$	$\hat{\pi}_3 = .298$
	$\hat{\lambda}_1 = .280$	$\hat{\lambda}_2 = .000$	$\hat{\lambda}_3 = .326$
	$\hat{\rho}_1 = .068$	$\hat{\rho}_2 = -.020$	$\hat{\rho}_3 = .040$
$m^* = 4$	$lL_M = -641.632$	$\hat{\tau}_1 = 799$	$\hat{\tau}_2 = 838$
	$\hat{\pi}_1 = .228$	$\hat{\pi}_2 = .000$	$\hat{\pi}_3 = .277$
	$\hat{\lambda}_1 = .280$	$\hat{\lambda}_2 = .000$	$\hat{\lambda}_3 = .319$
	$\hat{\rho}_1 = .068$	$\hat{\rho}_2 = .000$	$\hat{\rho}_3 = .059$

Table 1. Multiple change-points analysis for the binary sequence with $A = 1$ for $m = 1$ to $m = 4$ change points, where lL_M denotes the modified loglikelihood and where the results for $m^* = 4$ are based on a simulation with 10^9 replications

$m = 1$	$lL_M = -665.207$	$\hat{\tau}_1 = 865$
	$\hat{\pi}_1 = .215$	$\hat{\lambda}_1 = .274$
	$\hat{\pi}_2 = .356$	$\hat{\lambda}_2 = .353$
	$\hat{\rho}_1 = .075$	$\hat{\rho}_2 = -.005$
$m = 2$	$lL_M = -659.015$	$\hat{\tau}_1 = 1005$
	$\hat{\tau}_2 = 1041$	$\hat{\pi}_1 = .247$
	$\hat{\pi}_2 = .543$	$\hat{\lambda}_1 = .311$
	$\hat{\rho}_1 = .085$	$\hat{\lambda}_2 = .211$
	$\hat{\rho}_2 = -.727$	
$m = 3$	$lL_M = -656.645$	$\hat{\tau}_1 = 11$
	$\hat{\tau}_2 = 1005$	$\hat{\tau}_3 = 1041$
	$\hat{\pi}_1 = .511$	$\hat{\lambda}_1 = .174$
	$\hat{\pi}_2 = .246$	$\hat{\rho}_1 = -.690$
	$\hat{\lambda}_2 = .316$	$\hat{\rho}_2 = .092$
$m = 4$	$lL_M = -651.860$	$\hat{\tau}_1 = 1005$
	$\hat{\tau}_2 = 1041$	$\hat{\tau}_3 = 1081$
	$\hat{\tau}_4 = 1096$	$\hat{\pi}_1 = .244$
	$\hat{\pi}_2 = .531$	$\hat{\lambda}_1 = .315$
	$\hat{\rho}_1 = .094$	$\hat{\lambda}_2 = .192$
	$\hat{\rho}_2 = -.721$	
$m^* = 4$	$lL_M = 652.616$	$\hat{\tau}_1 = 1005$
	$\hat{\tau}_2 = 1041$	$\hat{\tau}_3 = 1086$
	$\hat{\tau}_4 = 1095$	$\hat{\pi}_1 = .245$
	$\hat{\pi}_2 = .548$	$\hat{\lambda}_1 = .313$
	$\hat{\rho}_1 = .090$	$\hat{\lambda}_2 = .217$
	$\hat{\rho}_2 = -.733$	

Table 2. Alternating-segments analysis for the binary sequence with $A = 1$ for $m = 1$ to $m = 4$ change points, where lL_M denotes the modified loglikelihood and where the results for $m^* = 4$ are based on a simulation with 10^9 replications

In Tables 1 and 2 the estimates $(\hat{\tau}_1, \dots, \hat{\tau}_m)$ of the m change points indicate where the probability of a 1 becomes different, i.e. where a new segment begins while the estimates $(\hat{\pi}_1, \dots, \hat{\pi}_{m+1})$ of the probability of a 1 in the $(m + 1)$ segments reveal the extent to which the segments differ.

The estimates were computed by considering a total enumeration of all possible m -tuples of change points (τ_1, \dots, τ_m) . This procedure is no longer feasible if the length of the chain (n) and the number of change points (m) are too large. In our example, with $n = 1,200$, we were able by optimizing the used algorithm to perform the analysis up to $m = 5$ for the first model and up to $m = 4$ for the second model before computing time became too large.

A first alternative is a simulation based on a random generation of tuples (τ_1, \dots, τ_m) of possible change points. Results for $m = 4$ are depicted in Tables 1 and 2. It is obvious that the number of replications has to be very high to get a good approximation. We used 10^9 replications and the results illustrate that a number of 10^4 or 10^6 replications which is typical for other contexts would be too small here. Of course, the total number of random numbers required is much larger than 10^9 and the pseudorandom number generators which are used in most applications will not give satisfactory results here. E.g., the generator implemented in DELPHI 5 proved to be insufficient for the present problem because it produced multiples of suboptimal solutions. However, we obtained satisfactory results with the Mersenne Twister by Matsumoto and Nishimura (1998) which can be recommended for this kind of extended simulations.

Another useful approach is to generate first a good initial set of m potential change points by a simulation and then to improve this set by using a kind of discrete steepest descent algorithm. For this we choose a small integer as an initial step size and consider the 3^m m -tuples resulting from the initial set by adding or subtracting the step size from each potential change point or by leaving it unchanged. That new set which corresponds to the highest value of the loglikelihood is selected as a new initial set and the procedure is repeated. Here, it is either possible to diminish the discrete step size during the iteration or to use from the beginning the smallest possible step size given by 1. If no improvement is observed at a certain step the iteration procedure finds an end and at least a local maximum has been found.

5 Discussion

We derived estimates for change points and changed segments in the case of a Bernoulli sequence with a first-order Markov dependence. It is certainly disappointing that we did not present significance tests as we did in Krauth (1999, 2000) for the case of independent Bernoulli trials. Other tests for this situation were proposed by Avery and Henderson (1999) and Halpern (1999, 2000). The difference is of course that we consider here the dependent case. However, even in this case tests and confidence intervals were derived, e.g. in Johnson and Klotz (1974), Crow (1979), and Bedrick and Aragon (1989). In contrast to the situation here, those authors considered only the stationary case and no tests or confidence intervals were derived for change points. Thus, it has to be admitted that there are still some unsolved problems.

Several authors have proposed very different methods for DNA sequence segmentation. An overview of various statistical approaches is given by Braun and Müller (1998). Of all these methods there seems to be nowadays one method which 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 using the EM algorithm. As noted in Braun and Müller (1998) this approach assumes independent data, requires large data sets and uses the EM algorithm which may fail to find the global optimum. In Liu et al. (1999) other problems with this approach are indicated, e.g. the inherent assumption that duplications 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.

In view of these arguments we believe that it is worthwhile to extend the present approach in that way that at least approximate tests for change points are derived in the future.

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.
- BEDRICK, E.J. and ARAGON, J. (1989): Approximate confidence intervals for the parameters of a stationary binary Markov chain. *Technometrics*, *31*, 437–448.
- BILLINGSLEY, P. (1961): *Statistical inference for Markov processes*. The University of Chicago Press, Chicago, London.
- BRAUN, J.V. and MÜLLER, H.G. (1998): Statistical methods for DNA sequence segmentation. *Statistical Science*, *13*, 142–162.
- BUDESCU, D.V. (1985): Analysis of dichotomous variables in the presence of serial dependence. *Psychological Bulletin*, *97*, 547–561.
- CHURCHILL, G.A. (1989): Stochastic models for heterogeneous DNA sequences. *Bulletin of Mathematical Biology*, *51*, 79–94.
- CROW, E.L. (1979): Approximate confidence intervals for a proportion from Markov dependent trials. *Communications in Statistics - Simulation and Computation*, *B8*, 1–24.
- DEVORE, J.L. (1976): A note on the estimation of parameters in a Bernoulli model with dependence. *Annals of Statistics*, *4*, 990–992.
- FU, Y.X. and CURNOW, R.N. (1990): Maximum likelihood estimation of multiple change points. *Biometrika*, *77*, 563–573.
- HALPERN, A.L. (1999): Minimally selected p and other tests for a single abrupt changepoint 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.
- HAWKINS, D.M. (2001): Fitting multiple change-point models to data. *Computational Statistics & Data Analysis*, *37*, 323–341.
- JOHNSON, C.A. and KLOTZ, J.H. (1974): The atom probe and Markov chain statistics of clustering. *Technometrics*, *16*, 483–493.
- KIM, S. and BAI, D.S. (1980): On parameter estimation in Bernoulli trials with dependence. *Communications in Statistics - Theory and Methods*, *A9*, 1401–1410.
- KLOTZ, J. (1973): Statistical inference in Bernoulli trials with dependence. *Annals of Statistics*, *1*, 373–379.
- 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 Millennium*. Springer, Heidelberg, 386–395.
- KRAUTH, J. (2003): Change-points in Bernoulli trials with dependence. In: W. Gaul and M. Schader (Eds.): *Between Data Science and Everyday Web Practice*. Springer, Heidelberg.
- LINDQVIST, B. (1978): A note on Bernoulli trials with dependence. *Scandinavian Journal of Statistics*, *5*, 205–208.
- 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.

- MATSUMOTO, M. and NISHIMURA, T. (1998): Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modeling and Computer Simulation*, 8, 3–30.
- MOORE, M. (1979): Alternatives aux estimateurs à vraisemblance maximale dans un modèle de Bernoulli avec dépendance. *Annales des Sciences Mathématiques du Québec*, 3, 119–133.
- PRICE, B. (1976): A note on estimation in Bernoulli trials with dependence. *Communications in Statistics - Theory and Methods*, A5, 661–671.
- 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.
- VENTER, J.H. and STEEL, S.J. (1996): Finding multiple abrupt change points. *Computational Statistics & Data Analysis*, 22, 481–504.

Outlier Identification Rules for Generalized Linear Models

Sonja Kuhnt and Jörg Pawlitschko

Department of Statistics,
University of Dortmund, D-44221 Dortmund, Germany

Abstract. Observations which seem to deviate strongly from the main part of the data may occur in every statistical analysis. These observations, usually labelled as outliers, may cause completely misleading results when using standard methods and may also contain information about special events or dependencies. We discuss outliers in situations where a generalized linear model is assumed as null model for the regular data and introduce rules for their identification. For the special cases of a loglinear Poisson model and a logistic regression model some one-step identifiers based on robust and non-robust estimators are proposed and compared.

1 Introduction

In the statistical analysis of data one often is confronted with observations that “appear to be inconsistent with the remainder of that set of data” (Barnett and Lewis (1994)). Although such “outliers” have been subject of numerous investigations, there is no general accepted formal definition of outlyingness. Most authors, however, agree in that the notion “outlier” is only meaningful in relation to a hypothesized statistical model for the “good” data, the so-called null model. We treat outliers in the sense of Davies and Gather (1993) who define outliers in terms of their position relative to the null model. E.g. for any normal distribution $\mathcal{N}(\mu, \sigma^2)$ and any α , $0 < \alpha < 1$, the corresponding α -outlier region is defined by

$$out(\alpha, \mu, \sigma^2) = \{x: |x - \mu| > \sigma z_{1-\alpha/2}\}$$

where $z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$ -quantile of the standard normal distribution. Any number x is called an α -outlier with respect to $\mathcal{N}(\mu, \sigma^2)$ if $x \in out(\alpha, \mu, \sigma^2)$. Figure 1 shows the 0.05-outlier region of the $\mathcal{N}(0, 1)$ distribution. Any number with an absolute value larger than 1.96 is a 0.05-outlier.

We will extend this general approach to outlyingness to null models for structured data situations such as regression models and contingency tables which are summarized in the broad class of generalized linear models (GLM). This unifying family of models has been introduced by Nelder and Wedderburn (1972) and has had a major influence on statistical modelling in a number of modern applications.

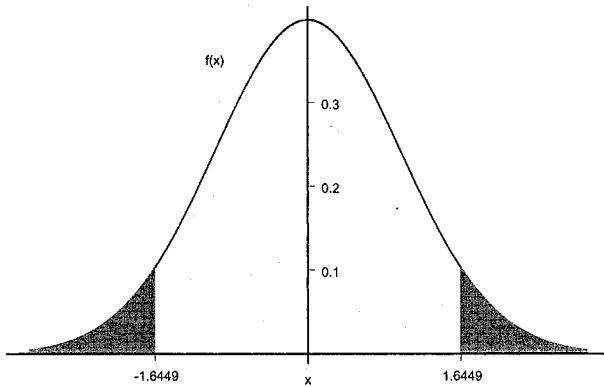


Fig. 1. 0.05-outlier region of the $\mathcal{N}(0, 1)$ -distribution

2 Generalized linear models

Consider the situation where it is of interest to explain a univariate response variable by a set of p fixed or stochastic covariates. Let $(Y_1, X_1), \dots, (Y_n, X_n)$, with $X_i = (X_{i1}, \dots, X_{ip})'$, be a sample of n observations. A generalized linear model (GLM) is characterized by two assumptions:

- *Distributional Assumption:* For each Y_i , $i = 1, \dots, n$, the conditional distribution of Y_i given $X_i = x_i$ belongs to an exponential family with expectation $E(Y_i|X_i = x_i) = \mu_i$ and variance $Var(Y_i|X_i = x_i) = \phi V(\mu_i)$, where V is a known variance function and $\phi > 0$ a common dispersion parameter not depending on i .
- *Structural Assumption:* There exists a so-called link function g , which is a known one-to-one, sufficiently smooth function, and a parameter vector $\beta = (\beta_1, \dots, \beta_p)' \in \mathbb{R}^p$ such that for each i , $i = 1, \dots, n$, the expectation μ_i is related to a linear predictor via

$$g(E(Y_i|X_i = x_i)) = \sum_{j=1}^p x_{ij} \beta_j.$$

The distributional assumption includes for example the families of normal, inverse Gaussian, gamma, Poisson and binomial distributions. The common linear regression model is part of this model class; in this case the responses have a normal distribution, g is the identity function, $\phi = \sigma^2$, and $V(\mu) = 1$.

3 Outliers and identification rules

We consider the distributions P_i of the responses given the covariates, $\{Y_i|X_i = x_i, i = 1, \dots, n\}$ and start by defining α_i -outlier regions for the individual conditional distributions. This definition can be derived from a more general definition of outlier regions given in Gather et al. (2003).

Let \mathcal{P} be an exponential family such that $P_i \in \mathcal{P}$ has density f_i with respect to a dominating measure and has (known) support $supp(P_i)$. Given $\alpha_i \in (0, 1)$ the α_i -outlier region of $P_i \in \mathcal{P}$ is then defined as

$$out(\alpha_i, P_i) = \{x \in supp(P_i): f_i(x) < K(\alpha_i)\} \quad (1)$$

where

$$K(\alpha_i) = \sup\{K > 0: P_i(\{x: f_i(x) < K\}) \leq \alpha_i\}.$$

With $inl(\alpha_i, P_i) = supp(P_i) \setminus out(\alpha_i, P_i)$ we define the corresponding α_i -inlier region of P_i . Each point $x \in out(\alpha_i, P_i)$ is called an α_i -outlier relative to P_i and each $x \in inl(\alpha_i, P_i)$ an α_i -inlier. This definition formalizes the element of ‘‘unlikeliness’’ that is associated with the more informal definition of an outlier cited in the introduction. Furthermore, this definition requires no further properties of a point x to be classified as outlier with respect to P_i than being contained in $out(\alpha_i, P_i)$. Especially outliers need not to come from a special outlier-generating mechanism as it is usually assumed in the literature.

Let now $\{y_i|x_i, i = 1, \dots, n\}$ be a sample that under the null model is assumed to come i.i.d. from a certain GLM. An observed response y_i is then identified as α_i -outlier if it lies in the α_i -outlier region of the corresponding conditional distribution. The levels α_i should be chosen such that under the null model the probability of the occurrence of any outlier in the whole sample does not exceed a given $\tilde{\alpha}$. If equal values of the α_i , $i = 1, \dots, n$, are desired, a natural choice depending on the sample size is given by

$$\alpha_i = 1 - (1 - \tilde{\alpha})^{1/n}. \quad (2)$$

The task of identifying all outliers in a sample $(\mathbf{y}_n|\mathbf{x}_n) = \{y_i|x_i, i = 1, \dots, n\}$ can now be described as the task to find all those y_i which are located in the corresponding outlier region $out(\alpha_i, P_i)$.

Roughly spoken, there are two important types of outlier identification rules, namely rules that proceed in one step and rules that operate step-wise. In the following we focus on the first type of rules. For a GLM, a so-called simultaneous or one-step outlier identifier essentially consists in a set of empirical versions $OI_i(\alpha_i, \mathbf{y}_n|\mathbf{x}_n)$, $i = 1, \dots, n$, of the α -outlier regions $out(\alpha_i, P_i)$, $i = 1, \dots, n$. Each point located in $OI_i(\alpha_i, \mathbf{y}_n|\mathbf{x}_n)$ then is classified as α_i -outlier with respect to the corresponding P_i . The main problem is that the P_i are only partially known. Since the different distributions of the responses are only caused by the different values of the covariates, the P_i

share the same unknown characteristics, namely the parameter vector β and the dispersion parameter ϕ .

To make the performance of different outlier identifiers comparable it is useful to standardize them in an appropriate way. Davies and Gather (1993) suggest two approaches in the i.i.d. case which can be transferred to the more complex setting of a GLM as well. In this case, the first standardization consists in the requirement that under the null model H_0 one has

$$P_{H_0}(Y_i \notin OI_i(\alpha_i, \mathbf{Y}_n | \mathbf{X}_n), i = 1, \dots, n) \geq 1 - \gamma \quad (3)$$

for some $\gamma > 0$ which is often chosen equal to $\tilde{\alpha}$. Their second suggestion leads to the requirement that under the null model one has

$$P_{H_0}(OI_i(\alpha_i, \mathbf{Y}_n | \mathbf{X}_n) \subset out(\alpha_i, P_i), i = 1, \dots, n) \geq 1 - \gamma \quad (4)$$

with γ chosen as in (3). Since both approaches inevitably lead to the laborious task of deriving (or simulating) a large number of normalizing constants we suggest to work without such a type of standardization and to estimate the regions $out(\alpha_i, P_i)$ directly. If $\tilde{\alpha}$ is chosen reasonably small this approach leads to identification rules which are not susceptible to identify too much regular observations as outliers. For estimating the true outlier regions one needs estimators of P_i , $i = 1, \dots, n$, and these are obtained by plugging estimators $\hat{\beta}$ of β and $\hat{\phi}$ of ϕ into the corresponding densities f_i . The classical estimator in GLM is the Maximum Likelihood (ML) estimator which, however, has the disadvantage of being not robust in most cases. For other data situations (see e.g. Davies and Gather (1993), Becker and Gather (1999)) it has been shown that reliable outlier identification rules should be based on robust estimators of the model parameters. Especially, outlier identifiers that are constructed with non-robust estimators are prone to the effects of masking and swamping. Masking occurs if an identification rule fails to identify some outlier although the sample contains two or more apparently outlying observations (which then “mask” themselves). Swamping occurs if some apparent outlier(s) in the sample cause the identification rule to classify a regular observation as outlier as well. These findings lead us to recommend the use of robust estimators for the construction of outlier identifiers in GLM as well.

4 Examples

4.1 Loglinear Poisson models

As a first illustration of outlier identification in GLM, we look at the problem of identifying outlying cells in contingency tables. We concentrate on a 7×8 table from Yick and Lee (1998) containing student enrolment figures from seven community schools in Australia for eight different periods of the year, see Table 1.

93	96	99	99	147	144	87	87
138	141	141	201	189	153	135	114
42	45	42	48	54	48	45	45
63	63	72	66	78	78	82	63
60	60	54	51	51	45	39	36
174	165	156	156	153	150	156	159
78	69	84	78	54	66	78	78

Table 1. Student enrolments data (Yick and Lee, 1998)

The assumed model is that of independence between the row and column classification. The 56 cell counts y_i , $i \in \{1, \dots, 56\}$, are taken to be outcomes of independent Poisson distributed random variables with individual expectations $E(Y_i) = \mu_i = \exp(x'_i \beta)$, where we have the logarithm as link function. The x_i , $i \in \{1, \dots, 56\}$, are defined by the independence assumption and consist only of entries $-1, 0, 1$ if effect coding is used. In case of the Poisson distribution it is not possible to give a simple expression for the outlier region, which is always an upper tail region or the union of an upper and a lower tail region. However, it can easily be derived using the definition. Every α_i -outlier region of a Poisson distribution $Poi(\hat{\mu}_i)$ based on an estimate $\hat{\mu}_i$ can be seen as an “empirical version” of the outlier region $out(\alpha_i, Poi(\mu_i))$, as discussed in Section 3. A one-step outlier identification rule can then be defined by identifying all cell counts lying in the corresponding region $out(\alpha_i, Poi(\hat{\mu}_i))$ as α_i -outliers. With $\tilde{\alpha} = 0.1$ the choice of the individual levels according to (2) leads to $\alpha_i = 1 - (1 - 0.1)^{\frac{1}{56}} = 0.00188$.

The classical estimator for contingency tables is the ML-estimator. Some robust alternatives have been proposed in the last years, including estimates based on the median polish method (Mosteller and Parunak (1985)), L_1 -estimates (Hubert, 1997), minimum Hellinger distances (Kuhnt (2000)), least median of chi-squares residuals and least median of weighted squared residuals (Shane and Simonoff (2001)).

105.20	103.74	105.20	113.48	117.86	111.05	100.98	94.49
149.65	147.57	149.65	161.43	167.67	157.97	143.65	134.41
45.56	44.93	45.56	49.15	51.05	48.09	43.73	40.92
69.76	68.79	69.76	75.25	78.16	73.64	66.96	62.66
48.90	48.22	48.90	52.74	54.78	51.61	46.93	43.92
156.69	154.51	156.69	169.02	175.55	165.40	150.40	140.73
72.23	71.23	72.23	77.92	80.93	76.25	69.33	64.88

Table 2. Maximum likelihood estimates

In case of ML-estimates (Table 2) only observation $y_{53} = 54$ lies in the 0.00118-outlier region of the distribution given by the estimate, $out(0.00118, Poi(80.93)) = \mathbb{N} \setminus \{55, \dots, 110\}$, and is thereby identified as outlier.

We also use median polish estimates, which are the means of the results of two sweeps of median polish on the logarithm of the cell counts once starting

94.34	95.42	99.53	98.47	109.55	101.67	95.32	88.48
138.70	140.26	146.33	144.77	161.06	149.47	140.13	130.09
44.54	45.05	46.99	46.49	51.72	48.00	45.00	41.77
67.17	67.94	70.87	70.11	78.00	72.39	67.86	63.00
46.59	47.13	49.16	48.63	54.10	50.21	47.07	43.70
152.80	154.54	161.20	159.48	177.43	164.66	154.37	143.31
76.38	77.25	80.58	79.72	88.69	82.31	77.17	71.63

Table 3. Median polish estimates

with the rows and once with the columns. Using these estimates, see Table 3, the four observations y_5 , y_6 , y_{12} and y_{53} are identified as 0.00188-outliers. Yick and Lee (1998) obtain the same set of outliers or, depending on the outlier identification procedure used, the set $\{y_5, y_6, y_{12}, y_{13}\}$, which they can explain from subject knowledge. Observations y_5 and y_6 are collected during a period in which a group of transient seasonal fruit picker families have moved near to this school, thus inflating the school's enrolments. y_{12} might show an unexpected high value due to a significant number of people moving into the area for an aboriginal funeral procession, which lasted around three months. Yick and Lee suggest that due to this funeral actually y_{13} might be the outlying observation and y_{53} is judged discordant due to swamping.

4.2 Logistic regression

Consider the case that the responses are binomially distributed according to $Y_i|X_i = x_i \sim Bin(m_i, p_i)$, $i = 1, \dots, n$. We suppose that $p_i = 1/(1 + \exp(-x'_i\beta))$ for some parameter vector β that is, we have a logistic regression model with grouped data: the link function from the structural assumption of the GLM is chosen as the logit function. The corresponding outlier regions can essentially be derived as in the Poisson case. Again, for the construction of a reliable one-step outlier identifier we need a robust estimator of β . For this purpose we may e.g. use the Least Median of Weighted Squares (LMWS) or Least Trimmed Weighted Squares (LTWS) estimators as proposed in Christmann (2001). As an example look at the data in Table 4 which are taken from Myers et al. (2002). These data report the result from a toxicity experiment which has been conducted to investigate the effect of different doses of nicotine on the common fruit fly.

A reasonable model for this set of data is a logistic regression model with

$$p_i = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \ln x_i))}. \quad (5)$$

For the data in Table 4 the ML- and LMWS-estimators of the model parameters are $\hat{\beta}_0^{ML} = 3.1236$, $\hat{\beta}_1^{ML} = 2.1279$, and $\hat{\beta}_0^{LMWS} = 3.3478$, $\hat{\beta}_1^{LMWS} = 2.3047$, respectively. Suppose now that the number of killed insects for the concentration of 0.70 g / 100 cc has been wrongly reported as 5 instead of 50. Then the parameter estimators are given by $\hat{\beta}_0^{ML} = 0.9278$, $\hat{\beta}_1^{ML} = 0.8964$,

x_i Concentration (g / 100 cc)	m_i number of exposed insects	y_i number of killed insects
0.10	47	8
0.15	53	14
0.20	55	24
0.30	52	32
0.50	46	38
0.70	54	50
0.95	52	50

Table 4. Data for the toxicity example

and $\hat{\beta}_0^{LMWS} = 3.3368$, $\hat{\beta}_1^{LMWS} = 2.2989$, respectively. For $\tilde{\alpha} = 0.01$ we now estimate the α_i -inlier regions for the distributions of the number of killed insects, where $\alpha_i = 0.00143$, $i = 1, \dots, 7$, is determined according to condition (2). These estimated inlier regions are contained in Table 5. Here \hat{p}_i^{ML} (\hat{p}_i^{LMWS}) denotes the plug-in estimator of p_i when inserting the ML- (the LMWS-) estimator of β_0 and β_1 into the right hand side of (5). Again, an observation is identified as α_i -outlier if it is not contained in the corresponding estimated α_i -inlier region.

x_i Concentration (g / 100 cc)	$inl(\alpha_i, Bin(m_i, \hat{p}_i^{ML}))$	$inl(\alpha_i, Bin(m_i, \hat{p}_i^{LMWS}))$
0.10	{3, ..., 21}	{0, ..., 14}
0.15	{7, ..., 28}	{5, ..., 25}
0.20	{10, ..., 32}	{12, ..., 34}
0.30	{13, ..., 35}	{22, ..., 43}
0.50	{16, ..., 37}	{31, ..., 45}
0.70	{24, ..., 46}	{43, ..., 54}
0.95	{26, ..., 46}	{45, ..., 52}

Table 5. Estimated inlier regions for the toxicity example

Note that in this example both rules detect the wrongly reported number of killed insects at 0.70 g / 100 cc correctly as 0.0143-outlier. However, the rule based on the ML-estimator also identifies the numbers at 0.50 and 0.95 g / 100 cc as outlying. This is an example of the swamping effect that clearly demonstrates the unreliability of outlier identification rules based on non-robust methods.

5 Outlook

We have discussed outliers with respect to the conditional distribution of the responses given the covariates which are treated as if they are fixed. This is the appropriate approach e.g. for the loglinear Poisson model where the x_i

reflect the structure of a certain contingency table. For a GLM with normal distribution of the responses and the identity as link function (i.e. a linear regression model with normal errors) the distribution of the covariates is often assumed to be multivariate normal. In this case the joint distribution of responses and covariates is multivariate normal as well and the outlier identifiers discussed e.g. in Becker and Gather (1999) can be applied. For other distributions of the responses, especially in the discrete case, it will be more complicated to model the joint distribution of responses and covariates and to derive the corresponding outlier regions.

References

- BARNETT, V. and LEWIS, T. (1994): *Outliers in Statistical Data*. 3rd ed., Wiley, New York.
- BECKER, C. and GATHER, U. (1999): The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association*, 94, 947–955.
- CHRISTMANN, A. (2001): Robust Estimation in Generalized Linear Models. In: J. Kunert and G. Trenkler (Eds.) *Mathematical Statistics with Applications in Biometry: Festschrift in Honour of Siegfried Schach*. Eul-Verlag, Lohmar, 215–230.
- DAVIES, P.L. and GATHER, U. (1993): The Identification of Outliers. *Journal of the American Statistical Association*, 88, 782–792.
- HUBERT, M. (1997): The Breakdown Value of the L_1 Estimator in Contingency Tables. *Statistics and Probability Letters*, 33, 419–425.
- GATHER, U., KUHNT, S., and PAWLITSCHKO, J. (2003): Concepts of Outlyingness for Various Data Structures. In: J.C. Misra (Ed.): *Industrial Mathematics and Statistics*. Narosa Publishing House, New Dehli, 545–585.
- KUHNT, S. (2000): *Ausreißeridentifikation im Loglinearen Poissonmodell für Kontingenztafeln unter Einbeziehung robuster Schätzer*. Dissertation, Department of Statistics, University of Dortmund, Germany.
- MOSTELLER, F. and PARUNAK, A. (1985): Identifying Extreme Cells in a Sizable Contingency Table: Probabilistic and Exploratory Approaches. In: D.C. Hoaglin, F. Mosteller, and J.W. Tukey (Eds.): *Exploring Data Tables, Trends and Shapes*. Wiley, New York, 189–224.
- MYERS, R.H., MONTGOMERY, D.C., and VINING, G.C. (2002): *Generalized Linear Models*. Wiley, New York.
- NELDER, J.A. and WEDDERBURN, R.W.M. (1972): Generalized Linear Models. *Journal of the Royal Statistical Society A*, 134, 370–384.
- SHANE, K.V. and SIMONOFF, J.S. (2001): A Robust Approach to Categorical Data Analysis. *Journal of Computational and Graphical Statistics*, 10, 135–157.
- YICK, J.S. and LEE, A.H. (1998): Unmasking Outliers in Two-Way Contingency Tables. *Computational Statistics & Data Analysis*, 29, 69–79.

Dynamic Clustering with Non-Quadratic Adaptive Distances for Interval-Type Data

Renata M. C. R. de Souza and Francisco de A. T. de Carvalho

Centro de Informática - CIn / UFPE, Av. Prof. Luiz Freire, s/n - Cidade Universitária, CEP: 50740-540 Recife - PE, Brazil, {fatec,rmcrs}@cin.ufpe.br

Abstract. This work presents a dynamic cluster algorithm with non-quadratic adaptive distances for partitioning a set of symbolic objects described by symbolic interval variables. Non-quadratic adaptive distances with one and two components are considered. Experimental results with artificial and real symbolic data demonstrate the performance of the adaptive methods in comparison with the non-adaptive methods. These non-quadratic adaptive methods furnish better quality clusters when compared to the non-adaptive methods.

1 Introduction

Clustering is a widely used technique in such fields as unsupervised pattern recognition and image processing. Its aim is to find similar groups within a given set of items. The items to be clustered are usually represented as a vector of quantitative single values, but with recent advances in the field of databases, it is now common to record interval data. Currently, these kinds of data are widely used in real world applications.

Symbolic Data Analysis (SDA) is a new domain in the area of knowledge discovery. It is related to multivariate analysis and pattern recognition and has provided some suitable tools for extending the standard clustering methods to symbolic objects described by interval variables. A symbolic interval variable takes an interval of \mathbb{R} for an object, where \mathbb{R} is the set of real numbers (Bock and Diday (2000)).

SDA has provided clustering algorithms for interval data: Verde et al. (2001) introduced a dynamic cluster algorithm for interval data considering context-dependent proximity functions. Bock (2001) gives a sequential clustering and updating strategy so as to construct a Self-Organizing Map for visualizing interval-type data. Chavent and Lechevallier (2002) proposed a dynamic cluster algorithm for interval data using a criterion based on Hausdorff distance. However, none of these algorithms uses adaptive distances.

The standard dynamic cluster algorithm (Diday and Simon (1976)) partitions a set of items into k clusters and a corresponding set of prototypes by locally minimizing a criterion that measures the fitting between clusters and their representations (prototypes). The clustering algorithm is performed in two steps: an allocation step to assign individuals to classes according to their

proximity to the prototype, and a representation step where the prototypes are updated according to the assignment of individuals in the allocation step.

The adaptive version of the dynamical cluster method also optimizes a criterion based on a measure of fitting between the clusters and their representations (prototypes), but at each iteration there is a different measure for the comparison of each cluster with its respective representation (Diday and Govaert (1977)). This algorithm also follows the same steps described above.

The main difference between these algorithms occurs in the representation step where the non-quadratic adaptive distances are also updated. The advantage of these adaptive distances lies in the fact that the clustering algorithm is able to find clusters of different shapes and sizes, whereas the dynamical clustering without adaptive distances has a tendency to find spherical clusters.

In this work, we present an adaptive dynamic cluster for interval-type data. The method uses a distance function that is a L_1 Minkowski-like distance for interval-type data. Two adaptive versions of this distance with one and two components are introduced (section 2). In order to show the usefulness of these methods, several symbolic data sets were considered, each with a different degree of clustering difficulty. The evaluation of the results is based on an external validity index in the framework of a Monte Carlo experience. An application with a real symbolic data set is also considered (section 3). Finally, in section 4 final comments and conclusions are given.

2 A dynamic cluster method for interval-type data

In dynamic cluster methods, the definition of the allocation and representation functions are based on a distance function defined in the description space. In our case, a vector of intervals describes each item and, consequently, a distance for interval-type data must be considered. In this section, we present dynamic cluster methods with and without adaptive distances for interval-type data.

2.1 The non-adaptive method

The standard dynamic cluster algorithm partitions a set of items into k clusters C_1, \dots, C_k and a corresponding set of prototypes G_1, \dots, G_k by locally minimizing a criterion that measures the fitting between the clusters and their representations (prototypes). This criterion (noted W_1) is usually defined in the following way:

$$W_1 = \sum_{i=1}^k \sum_{s \in C_i} d(s, G_i) \quad (1)$$

where $d(s, G_i)$ is a dissimilarity measure between an object $s \in C_i$ and the class prototype G_i of C_i .

In De Carvalho and Souza (1999) an interval $[a, b]$ is represented as a point $(a, b) \in \mathbb{R}^2$, where the lower bounds of the intervals are represented in the x-axis and the upper bounds in the y-axis. A L_1 distance between an item $s = ([s_L^1, s_U^1], \dots, [s_L^p, s_U^p])$ and a prototype $G_i = ([G_{iL}^1, G_{iU}^1], \dots, [G_{iL}^p, G_{iU}^p])$ of C_i is computed as:

$$d(s, G_i) = \sum_{j=1}^p |s_L^j - G_{iL}^j| + |s_U^j - G_{iU}^j| \quad (2)$$

Choosing this L_1 distance, the prototype G_i of cluster C_i , which locally minimizes the criterion W_1 , is a vector of intervals, the bounds of which for each variable j , respectively, are the median of the set of lower bounds and the median of the set of upper bounds of the object intervals belonging to the cluster C_i .

The dynamic cluster algorithm with non-adaptive distance contains the following steps:

1. Initialization: selection of an initial partition;
2. Representation step: compute the class prototypes that minimize the adequacy criterion W_1 ;
3. Allocation step: for each object find its nearest prototype and assign it to the class represented by this prototype.

The allocation and representation steps are performed iteratively until convergence, when the adequacy criterion W_1 reaches a stationary value.

2.2 The adaptive methods

The dynamic cluster algorithm with adaptive distances (Diday and Govaert (1977)) also follows the steps above, but with a different distance associated to each cluster. The algorithm looks for a partition in k clusters, its corresponding k prototypes and different k distances associated with the clusters by locally minimizing the criterion

$$W_2 = \sum_{i=1}^k \sum_{s \in C_i} d_i(s, G_i) \quad (3)$$

where $d_i(s, G_i)$ is an adaptive dissimilarity measure between an object $s \in C_i$ and the class prototype G_i of C_i .

According to the intra-class structure of the cluster C_i , we consider here an adaptive distance between an item s and a prototype G_i in two ways.

Adaptive method 1: The distance between an item s and a prototype G_i of C_i is defined by the coefficient λ_i^j considered for each descriptor ($j = 1, \dots, p$):

$$d_i(s, G_i) = \sum_{j=1}^p \lambda_i^j (|s_L^j - G_{iL}^j| + |s_U^j - G_{iU}^j|) \quad (4)$$

These coefficients are obtained by the Lagrange multiplier method. According to Govaert (1975), the method looks for the coefficient $\lambda_i^j (j = 1, \dots, p)$ that satisfies the following restrictions:

1. $\lambda_i^j > 0$
2. $\prod_{j=1}^p \lambda_i^j = 1$

The coefficient λ_i^j that satisfies the restrictions (1) and (2) and minimizes the criterion W_2 is:

$$\lambda_i^j = \frac{[\prod_{h=1}^p (\sum_{s \in C_i} |s_L^h - G_{iL}^h| + |s_U^h - G_{iU}^h|)]^{\frac{1}{p}}}{\sum_{s \in C_i} |s_L^j - G_{iL}^j| + |s_U^j - G_{iU}^j|} \quad (5)$$

Adaptive method 2: In this case, the distance between an item s and a prototype G_i of C_i is defined by the lower bound coefficient λ_{iL}^j and the upper bound coefficient λ_{iU}^j considered for each descriptor ($j = 1, \dots, p$):

$$d_i(s, G_i) = \sum_{j=1}^p \lambda_{iL}^j |s_L^j - G_{iL}^j| + \lambda_{iU}^j |s_U^j - G_{iU}^j| \quad (6)$$

Again, the coefficients λ_{iL}^j and λ_{iU}^j that satisfy the restrictions (1) and (2) and minimize the criterion W_2 given by the equation 3 are also obtained by the Lagrange multiplier method. They are:

$$\lambda_{iL}^j = \frac{[\prod_{h=1}^p (\sum_{s \in C_i} |s_L^h - G_{iL}^h|)]^{\frac{1}{p}}}{\sum_{s \in C_i} |s_L^j - G_{iL}^j|} \quad (7)$$

$$\lambda_{iU}^j = \frac{[\prod_{h=1}^p (\sum_{s \in C_i} |s_U^h - G_{iU}^h|)]^{\frac{1}{p}}}{\sum_{s \in C_i} |s_U^j - G_{iU}^j|} \quad (8)$$

For both adaptive methods the prototype G_i of cluster C_i , which locally minimizes the criterion W_2 , is again the vector of intervals whose bounds for each variable j , respectively, are the median of the set of lower bounds and the median of the set of upper bounds of the object intervals belonging to the cluster C_i . Moreover, these solutions are not unique (Govaert (1975)).

The dynamic cluster algorithm with adaptive distances has the following steps:

1. Initialization: selection of an initial partition;

2. Representation step: compute the class prototypes G_i and coefficient λ_i^j (method 1) or coefficients λ_{iL}^j and λ_{iU}^j (method 2) that minimize the adequacy criterion W_2 ;
3. Allocation step: assign each object to the cluster that has nearest class prototype of this object using a different distance for each cluster.

Again, the allocation and representation steps are performed iteratively until convergence, when the adequacy criterion W_2 reaches a stationary value. The initialization and the allocation step are nearly the same in both the adaptive and non-adaptive dynamic cluster algorithms. The main difference between them occurs in the representation step when the coefficients of the adaptive distances are also updated.

3 Experiments and results

To show the usefulness of these methods, a real interval-type data set and two artificial interval-type data sets with different degrees of clustering difficulty (clusters of different shapes and sizes, linearly non-separable clusters, etc.) are considered. In this section, we present the experiments and results for these interval-type data sets.

3.1 Artificial symbolic data sets

The artificial data sets were constructed according to a uniform distribution in a quantitative bi-dimensional space. Each quantitative data set has 120 points scattered among four clusters of unequal sizes: two clusters with spherical shapes (sizes 8 and 16) and two clusters with ellipsis shapes (sizes 36 and 60). Data set 1 shows well-separated clusters (Figure 1). Data set 2 shows overlapping clusters (Figure 2).

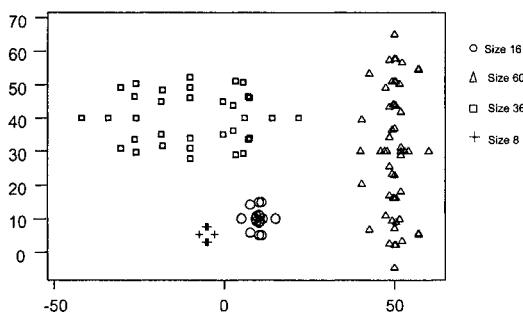


Fig. 1. Seed data set 1 showing well-separated classes

Each point (x_1, x_2) in these data sets is a seed of an interval-type data $([x_1 - \gamma_1, x_1 + \gamma_1], [x_2 - \gamma_2, x_2 + \gamma_2])$, where γ_1 and γ_2 are integer parameters

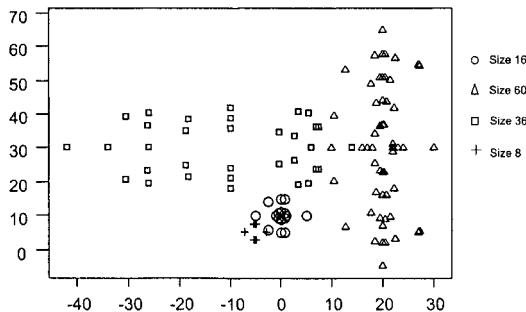


Fig. 2. Seed data set 2 showing overlapping classes

that are randomly selected from a predefined set of values. Five different sets of values were used for selecting the parameters γ_1 and γ_2 .

The evaluation of these methods was performed in the framework of a Monte Carlo experience: 100 replications are considered for each interval data set and the average of the corrected Rand (CR) index among the 100 replications is calculated. This index measures the similarity between an a priori partition and a partition furnished by the clustering algorithm (Hubert and Arabie (1985)). The range of values of the CR index is between -1 and 1. For each iteration a clustering method is run 50 times and the best result is selected according the criterion W_1 (or W_2).

Table 3.1 shows the values of the average CR index according to the different methods and interval-type data sets. The terms Adaptive Method 1 and Adaptive Method 2 signify adaptive methods with one-component and two-components, respectively.

Predefined sets of integer values	Interval-type Data Set 1			Interval-type Data Set 2		
	Adaptive Method 1	Adaptive Method 2	Non-Adaptive Method	Adaptive Method 1	Adaptive Method 2	Non-Adaptive Method
{1, 2}	0.836	0.836	0.634	0.625	0.625	0.396
{1, ..., 4}	0.835	0.835	0.633	0.683	0.688	0.397
{1, ..., 6}	0.837	0.837	0.620	0.709	0.709	0.397
{1, ..., 8}	0.838	0.838	0.611	0.708	0.708	0.395
{1, ..., 10}	0.844	0.846	0.600	0.706	0.705	0.395

Table 1. Comparison between methods according to the correct Rand index

It can be seen from this table that the average CR indices for the adaptive methods are greater or equal to 0.5 in all situations. This is not the case for the non-adaptive method concerning the interval data set 2 (where the average is around 0.396). Also the average CR indices for the adaptive methods are greater than those for the non-adaptive method in all situations.

The comparison between the proposed clustering methods is achieved by a paired Student's t-test at a 5% significance level. Table 2 shows the results

of paired Students' t-tests at a 5% significance level. In this table μ_1 , μ_2 and μ are the mean of the CR index for Adaptive Methods 1, 2 and the non-adaptive method, respectively. These results support the hypothesis that the average performance of the adaptive methods is superior to the non-adaptive method and that the average performance of Adaptive Method 1 is as good as that of Adaptive Method 2.

Predefined sets of integer values	$H_0 : \mu_1 = \mu_2$			$H_0 : \mu_1 = \mu$		
	$H_1 : \mu_1 \neq \mu_2$		Decision	$H_1 : \mu_1 > \mu$		Decision
	Interval-type data set 1	Interval-type data set 2		Interval-type data set 1	Interval-type data set 2	
{1, 2}	-1.00	-1.68	Accept H_0	498.22	85.28	Reject H_0
{1, ..., 4}	-0.54	1.25	Accept H_0	295.69	62.04	Reject H_0
{1, ..., 6}	-0.03	-0.11	Accept H_0	108.22	223.42	Reject H_0
{1, ..., 8}	-0.02	0.26	Accept H_0	108.62	179.38	Reject H_0
{1, ..., 10}	1.42	-0.52	Accept H_0	113.96	196.38	Reject H_0

Table 2. Statistics of paired Student's t-tests comparing the methods

3.2 The fish data set

The fish data set is a symbolic data table with 12 fish species, each species being described by 13 interval symbolic variables and 1 categorical variable. These elements are grouped in four a priori clusters of unequal sizes according to this categorical variable: two clusters (Carnivores and Détritivores) of sizes 4 and two clusters of sizes 2 (Omnivores and Herbivores). Table 3 shows part of the fish data set.

Individuals/labels	Interval Variables				
	Length	Weight	... Intestin/Muscle	Stomach/Muscle	
Ageneiosusbrevifili 1	[1.8 : 7.1]	[2.1 : 7.2]	...	[7.8 : 17.9]	[4.3 : 11.8]
:	:	:	:	:	:
Myleusrubripinis 4	[2.7 : 8.4]	[2.7 : 8.7]	...	[8.2 : 20]	[5.1 : 13.3]

Table 3. Fish Data Set described by 13 interval symbolic variables

Table 4 shows the clusters (individual labels) obtained by the a priori partition according to the categorical variable, Adaptive Methods 1 and 2 and non-adaptive method.

The CR indices according to the clustering results of the Table 4 are, respectively, 0.49, 0.07 and 0.05 for Adaptive Methods 1, 2 and for the non-adaptive method. These results once again show that the performance of the adaptive methods is superior to the non-adaptive method. It is important to note that the performance of Adaptive Method 1 is clearly superior to that of Adaptive Method 2 in this application.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
A priori Partition	1 2 3 4	7 8 9 10	5 6	11 12
Adaptive Method 1	5 6	4 7 8 10	1 2 3	9 11 12
Adaptive Method 2	4 5 8	2 10	1 3 7	6 9 11 12
Non-Adaptive Method	1 4 8	10	5 6 9 11 12	2 3 7

Table 4. Clustering Results for Fish Data Set

4 Conclusions

In this paper, two clustering methods for interval-type data using a dynamic cluster algorithm with non-quadratic adaptive distances were presented. These methods locally optimize an adequacy criterion that measures the fitting between classes and their representatives (prototypes). The results furnished by these clustering methods were evaluated through an external index. The experiments with real and artificial interval-type data sets carried out showed the usefulness of these clustering methods. Moreover, concerning the average CR index, the adaptive methods clearly outperform the non-adaptive method.

Acknowledgments: The authors would like to thank CNPq (Brazilian Agency) for its financial support.

References

- BOCK, H.-H. (2001): Clustering algorithms and kohonen maps for symbolic data. *Proc. ICNCB*, Osaka, 203–215.
- BOCK, H.-H. and DIDAY, E. (2000): *Analysis of Symbolic Data*. Springer, Heidelberg.
- CHAVENT, M. and LECHEVALIER, Y. (2002): Dynamical Clustering Algorithm of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff Distance. In: K. Jajuga, A. Sokolowsky, and H.-H. Bock (Eds.): *Classification, Clustering and Data Analysis*. Springer, Heidelberg, 53–59.
- DE CARVALHO, F.A.T. and SOUZA, R.M.C.R. (1999): New metrics for constrained Boolean Symbolic objects. In: *Studies and Research: Proceedings of the Conference on Knowledge Extraction and Symbolic Data (KESDA '98)*. Office for Official Publications of the European Communities, Luxembourg, 175–187.
- DIDAY, E. and GOVAERT, G. (1977): Classification Automatique avec Distances Adaptatives. *R.A.I.R.O. Informatique Computer Science*, 11 (4), 329–349.
- DIDAY, E. and SIMON, J.J. (1976): Clustering Analysis. In: K.S. Fu (Ed.): *Digital Pattern Recognition*. Springer, Heidelberg, 47–94.
- GOVAERT, G. (1975): *Classification automatique et distances adaptatives*. Thèse de 3ème cycle, Mathématiques appliquée, Université Paris VI.
- HUBERT, L. and ARABIE, P. (1985): Comparing Partitions. *Journal of Classification*, 2, 193–218.
- VERDE, R., DE CARVALHO, F.A.T., and LECHEVALIER, Y. (2001): A Dynamical Clustering Algorithm for symbolic data. In: *Tutorial on Symbolic Data Analysis, Gfkl Conference*, Munich.

Partial Moments and Negative Moments in Ordering Asymmetric Distributions

Grażyna Trzpiot

Department of Statistics,
Katowice University of Economics,
ul. Bogucicka 14, 40-287 Katowice, Poland

Abstract. Moment ordering condition is shown to be necessary for stochastic dominance. In this paper related results of the partial moments and negative moments are presented. The condition for any degree of stochastic dominance, by ordering fractional and negative moments of the distribution, will be shown. We present the sufficient condition for restricted families of distribution functions - a class of asymmetric distributions. Additionally we present a related general measure based on fractional moments, which can be used for complete ordering the set of distributions. The condition applies generally, subject only to the requirement that the moments exist. The result rests on the fact that the negative and the fractional moments of the distribution can be interpreted as constant relative risk aversion utility function.

1 Introduction

Stochastic dominance provides a general set of rules for ranking risky assets. Hadar and Russel (1969) show that first degree stochastic dominance implies all odd moments of the distribution are ordered. Whitmore (1970) shows that an ordering of a combination of the mean and variance is necessary for any degree of stochastic dominance. Considering a more general class of moments, a moment ordering condition is necessary for any degree of stochastic dominance. Related results of the partial moments and negative moments are presented. The condition for any degree of stochastic dominance by ordering fractional and negative moments of the distribution will be shown. We present the sufficient condition for restricted families of distribution functions - a class of asymmetric distributions. Additionally, we present a related general measure based on fractional moments, which can be used for the complete ordering of the set of distributions. The condition applies generally, subject only to the requirement that the moments exist. The results rest on the fact that the negative and the fractional moments of the distribution can be interpreted as a constant relative risk aversion utility function.

2 Stochastic dominance

Let two random variables X and Y have distributions $F(x)$ and $G(x)$. The distributions are assumed to have supports in interval $A = [a, b]$. Let F and

G be the cumulative distributions of two distinct uncertain alternatives X and Y to be compared. X dominates Y by first, second and third stochastic dominance (FSD, SSD, TSD) if and only if

$$H_1(x) = F(x) - G(x) \leq 0 \quad \text{for all } x \in [a, b] \quad (X \text{ FSD } Y) \quad (1)$$

$$H_2(x) = \int_a^x H_1(y)dy \leq 0 \quad \text{for all } x \in [a, b] \quad (X \text{ SSD } Y) \quad (2)$$

$$\begin{aligned} H_3(x) &= \int_a^x H_2(y)dy \leq 0 \quad \text{for all } x \in [a, b] \\ \text{and } E(F(x)) &\geq E(G(x)) \quad (X \text{ TSD } Y) \end{aligned} \quad (3)$$

The relationship between the three stochastic dominance rules can be summarised by the following diagram: $\text{FSD} \Rightarrow \text{SSD} \Rightarrow \text{TSD}$, which means that dominance by FSD implies dominance by SSD and dominance by SSD in turn implies dominance by TSD. For proof of FSD and SSD see Hadar and Russell (1969), Hanoch and Levy (1969) and Rothschild and Stiglitz (1970). The criterion for TSD was suggested by Whitmore (1970). The dominance concepts discussed are based on the theoretical foundation of a utility function.

For two random variables X and Y with distributions F and G , we say that X dominates Y by n th degree stochastic dominance (SD_n) if and only if

$$H_n(x) = \int_a^x H_{n-1}(y)dy \leq 0 \quad \text{for } x \in [a, b] \quad (X \text{ SD}_n Y) \quad (4)$$

3 Partial moments and negative moments

Let X be a random variable with cumulative distribution $F(x)$. If the expected value of X^k

$$EX^k = \int x^k dF(x), \quad k = 1, 2, \dots$$

exists, it is called the k th moment of X .

For absolutely continuous distributions the k th moment becomes

$$EX^k = \int_{-\infty}^{+\infty} x^k f(x)dx$$

where $f(x)$ is the probability density, for (atomic) discrete distributions taking value x_i with probability p_i the k th moment is given by

$$EX^k = \sum_{i=1}^{+\infty} x_i^k p_i.$$

– **Partial moments**

The n th partial (or incomplete) moment of a random variable X over an interval $A \subset R$ is given by

$$E_A X^n = \int_A x^n f(x) dx$$

if X is continuous with density f and by

$$E_A X^n = \sum_{x \in A} x^n p(x)$$

is X is discrete with probability function p .

If $A = R$, then the integral (sum) represents the usual moments of X . Each set A gives rise to partial moments of X .

– **Negative and Fractional Moments**

Let X be a random variable with cumulative distribution $F(x)$. If the expected value of X^k

$$EX^k = \int_A x^k dF(x)$$

exists, it is called the k th moment of X .

Typically k is taken to be a positive and integer, although nothing in the definition requires this. Letting k take any real value permits the existence of the negative and fractional moments.

– **A quantity closely related to the k th moment of X is the generalized mean**

$$E^* X^k = [\int_A x^k dF(x)]^{\frac{1}{k}}.$$

The best known examples are the harmonic mean $E^* X^{-1}$ and the geometric mean, which is the limiting value of $E^* X^k$ as $k \rightarrow 0$. The consideration of these negative and fractional moments leads to a necessary condition for any degree of stochastic dominance.

4 Ordering distribution

Let two random variables X and Y have distributions $F(x)$ and $G(x)$ to be compared. The distributions have supports in interval $A = [a, b]$.

Proposition 1

Provided the moments exist, $X \text{ SD}_k Y$ implies (SD k-degree)

1. $EX^k < EY^k$, for $k < 0$,
2. $EX^k \geq EY^k$, for $k \in [0, 1]$.

Most investigations on necessary conditions for stochastic dominance have focused on positive integer moments. The relative risk aversion measure is $r(x) = -xu''(x)/u'(x)$. Then $u(x) = \text{sgn}(k)x^k (k \neq 0)$ is the utility function with constant relative risk aversion $r(x) = 1 - k$, the utility function is risk averse if $k < 1$. Then the k th moment of X , except possibly for the sign, can be interpreted as an expected utility for a constant relative risk aversion function $EX^k = \text{sgn}(k)E_F(u(x))$. This interpretation makes clear that this is a necessary, but not sufficient, condition. The constant relative risk aversion utility functions are subset of all relevant utility functions.

Second and higher degrees of stochastic dominance are related to expected utility for utility functions that are risk averse. For $k > 1$, a utility function $u(x) = \text{sgn}(k)x^k$ is risk loving. So this necessary condition based on moments above first, including positive integer moments, involve expected utility for a risk loving utility function.

We can notice 1) and 2) by $\text{sgn}(k)[EX^k - EY^k] \geq 0$, for $k \leq 1$. The ordering for a generalized mean follows from the first proposition.

Proposition 2

Provided the moments exist, $X \text{ SD}_k Y$ implies (SD k-degree)

$$E^*X^k \geq E^*Y^k, \text{ for } k \leq 1 \text{ with strict inequality for } k < 1.$$

These results are more general than for a geometric and harmonic mean. This proposition reflects the property of preferences. The generalized mean E^*X^k is a certainty equivalent of F for the constant relative risk aversion utility functions. Then $F \text{ FSD } G$ implies, that F always has a larger certainty equivalent than G . That proposition is a necessary, but not sufficient, condition (the constant relative risk aversion utility functions are subset of all relevant utility functions).

As an application we can show the construction of an efficient set from a given finite set of distributions.

Proposition 2a

1. If exist $k < 1$ such that $E^*X^k < E^*Y^k$ then $\sim F \text{ SD } G$.

2. If for any G implies that exist $k < 1$ such that
 $E^*X^k > E^*Y^k$, then $\sim G$ SD F.

In the empirical analysis the use of negative and fractional moments can significantly reduce the computation required in the empirical stochastic dominance analysis. An important practical advantage of using the negative and fractional moments, or mean order k , is that quantities need to be computed only once for each distribution. A future advantage of the moment based necessary condition is that the sampling distributions of moments are easily obtained.

The generalized welfare measure

Rank dominance is based on the strong Pareto principle and is equivalent to FSD (first stochastic dominance). Generalized Lorenz dominance is based on the strong Pareto principle and is equivalent to SSD (second stochastic dominance). The incomplete ordering provided by both FSD and SSD leaves some ambiguity in ranking distribution.

The generalized welfare measure

$$W = \left[\sum_i (y_i^{1-\varepsilon}/n) \right]^{1/(1-\varepsilon)}, \quad \text{for } \varepsilon \geq 0 \quad \text{and} \quad \varepsilon \neq 1$$

$$W = \left(\prod_i y_i \right)^{1/n} \quad \text{for } \varepsilon = 1$$

where y_i is individual return (income), $i = 1, 2, \dots, n$.

The parameter ε is used to set a relative degree of equity preference to efficiency preference allowing this measure to other standard criteria. For $\varepsilon = 0$ we have that $W = EY$. Ranking using an arithmetic mean represents an extreme case of efficiency preference. The limit $\varepsilon \rightarrow \infty$ is the minimum value in the distribution. This represents an extreme case of equity preference, but remains within the set of minimum bounds for efficiency preference (weak Pareto principle). The geometric mean is the specific case (for $\varepsilon = 1$) in which W responds equally to equal *proportionate* increases of y_i for all i . For $\varepsilon > 1$ W is more sensitive to equal *proportionate* increases of lower return (income). For $\varepsilon < 1$ W is more sensitive to equal *proportionate* increases of higher return (income).

We derive parametric criteria for the optimal efficient set when we have a set of prospects with asymmetric distributions. The following theorem specifies the optimal efficient set for gamma distributions.

Theorem. 1.

Let $F_{\alpha,\beta}$ be a family of gamma distributions with positive parameters α and β the density functions

$$f_{\alpha\beta} = (\alpha^\beta / \Gamma(\beta)) e^{-\alpha x} x^{\beta-1}, \quad x > 0$$

Then:

- (a) $F_{\alpha_1, \beta}$ FSD $F_{\alpha_2, \beta}$ if and only if $\alpha_1 < \alpha_2$
- (a') $F_{\alpha_1, \beta}$ SSD $F_{\alpha_2, \beta}$ if and only if $\alpha_1 < \alpha_2$
- (b) F_{α, β_1} FSD F_{α, β_2} if and only if $\beta_1 > \beta_2$
- (b') F_{α, β_1} SSD F_{α, β_2} if and only if $\beta_1 > \beta_2$
- (c) F_{α_1, β_1} FSD F_{α_1, β_2} if and only if $\alpha_1 \leq \alpha_2$ and $\beta_1 \geq \beta_2$
with at least one strict inequality
- (d) F_{α_1, β_1} SSD F_{α_1, β_2} if and only if $\beta_1/\beta_2 \geq \max(1, \alpha_1/\alpha_2)$
with strict inequality at least when $\alpha_1/\alpha_2 = 1$.

Let the alternative have a gamma distribution with parameters (α, β) then any alternative can be identified by corresponding values (α, β) . According to a) and a') for two alternatives with gamma distributions, which differ only by parameter α , the one with the smaller α is preferable. Similarly according to b) and b') for two alternatives with gamma distributions, which differ only by parameter β , and α is the same for both alternatives, the larger β is preferable. The utility function, on which the preference is based, can be chosen arbitrarily from U_1 or U_2 . As the mean and the variance of alternatives with gamma distributions with parameters (α, β) are respectively $EX = \beta/\alpha$ and $D^2X = \beta/\alpha^2$, it follows that in either case alternatives differing in α or β , the preferred alternative has the larger mean and the larger variance. For any risk averters interpreting a variance as a measure of risk, larger risk is compensated by increasing the mean of the alternative. If we have two alternatives differing in both parameters, then from c) we have that for all non-decreasing utility functions, the preferred alternative should not have either the larger α or smaller β . If one of them has both the larger α and the larger β , then no preference can be established. These conditions can be relaxed if we consider a risk averse utility function (a class of DARA functions). In this case from d), a alternative with the smaller β is never preferred or a alternative with the larger is preferred only when it is compensated by increased β .

Theorem. 2.

Let $F_{\alpha, \beta}$ be a family of beta distribution with positive parameters α and β the density functions

$$f_{\alpha, \beta} = (\Gamma(\alpha + \beta)/\Gamma\alpha\Gamma\beta)x^{\beta-1}(1-x)^{\alpha-1}, \quad 0 < x < 1$$

Then:

- (a) $F_{\alpha_1, \beta}$ FSD $F_{\alpha_2, \beta}$ if and only if $\alpha_1 < \alpha_2$
- (a') $F_{\alpha_1, \beta}$ SSD $F_{\alpha_2, \beta}$ if and only if $\alpha_1 < \alpha_2$
- (b) F_{α, β_1} FSD F_{α, β_2} if and only if $\beta_1 > \beta_2$
- (b') F_{α, β_1} SSD F_{α, β_2} if and only if $\beta_1 > \beta_2$
- (c) F_{α_1, β_1} FSD F_{α_1, β_2} if and only if $\alpha_1 \leq \alpha_2$ and $\beta_1 \geq \beta_2$
with at least one strict inequality
- (d) F_{α_1, β_1} SSD F_{α_1, β_2} if and only if $\beta_1/\beta_2 \geq \max(1, \alpha_1/\alpha_2)$
with strict inequality at least when $\alpha_1/\alpha_2 = 1$.

5 Stochastic dominance rules in portfolio selection

According to the expected utility paradigm, the risk of the investment must be related to the assumed preference of the investor and cannot be objectively defined. Markowitz (1952) method for ranking alternatives was comparing means and variances of two alternatives. Rothschild and Stiglitz (1970) defined risk for two distributions characterised by the same expected return. They used stochastic dominance rules for definition of risk. Rothschild and Stiglitz present several definitions of more risky random variables:

1. Y is more risky than X if $Y = X + Z$ (Y is equal to X plus a noise term Z is a random variable and $E(Z|X) = 0$).
2. Y is riskier than X if and only if every risk-avertor prefers X over Y ($E(u(X)) \geq E(u(Y))$, for all concave u).
3. Y is riskier if it has more weight in the tails than X .
4. Y is riskier if it has greater variance than X .

They analyse these four definitions and conclude that the first three definitions are equivalent. The fourth definition does not necessarily do so. In general, in the Rothschild and Stiglitz framework, the variance cannot serve as an index for risk. We have appropriate investment criteria for the different risk-choice situations. An efficiency criterion is a decision rule for dividing all potential investment alternatives into two mutually exclusive sets: an efficient set and an inefficient set. Firstly, using stochastic dominance tests we reduce the number of investment alternatives by constructing an efficient set of alternatives appropriate for a given class of investors. At the second step, we can make the final choice of alternatives in accordance with particular preferences of the investor. The three stochastic dominance criteria, FSD, SSD and TSD, are optimal in the sense that given the assumptions regarding the investors preferences (describing as a class of utility functions), the application of the corresponding stochastic dominance criterion ensures a

minimal efficient set of investment alternatives. For application of stochastic dominance see Bradley and Lehmann (1988), Levy (1992, 1996), Ogryczak and Ruszczyński (1999), Trzpiot (1998, 2002, 2003).

6 Conclusions

Any degree of stochastic dominance implies that all negative and fractional moments, or equivalently, the means of order k , are ordered. This generalizes the ordering of the geometric mean, harmonic mean and minimum values. The condition applies generally, subject only to the requirement that the moments exist. The results rest on the fact that the negative and fractional moments of the distribution can be interpreted as expected utility for the constant relative risk aversion utility function. The generalized welfare measure illustrates that the use of a special case involves a subjective judgment by the decision-maker of the relative importance of equity preference to efficiency preference.

References

- BRADLEY, M.G. and LEHMANN, D.E. (1988): Instrument Effects and Stochastic Dominance. *Insurance Mathematic And Economics*, 7, 185–191.
- HADAR, J. and RUSSEL, W.K. (1969): Rules for Ordering Uncertain Prospects. *Amer. Economic Rev.*, 59, 25–34.
- HANOCH, G. and LEVY, H. (1969): The Efficiency Analysis of Choices Involving Risk. *Rev. Economic Studies*, 36, 335–346.
- LEVY, H. (1992): Stochastic Dominance and Expected Utility: Survey and Analysis. *Management Science*, 38, 4, 555–593.
- LEVY, H. (1996): Investment Diversification and Investment Specialization and the Assumed Holding Period. *Applied Mathematical Finance*, 3, 117–134.
- MARKOWITZ, H.M. (1952): Portfolio Selection. *Journal of Finance*, 7, 77–91.
- OGRYCZAK, W. and RUSZCZYNSKI, A. (1999): From Stochastic Dominance to Mean Risk Models: Semideviation As Risk Measure. *European Journal Of Operation Research*, 116, 33–50.
- ROTHSCHILD, L.J. and STIGLITZ, J.E. (1970): Increasing risk I. A definition. *Journal of Economic Theory*, 2, 225–243.
- TRZPIOT, G. (1998): Stochastic Dominance Under Ambiguity in Optimal Portfolio Selection: Evidence from the Warsaw Stock Exchange, Short Papers from VI Conference of the International Classification Societies, 311–315, Rome.
- TRZPIOT, G. (2002): Multicriterion analysis based on Marginal and Conditional Stochastic Dominance in Financial Analysis. In: T. Trzaskalik and J. Michnik (Eds.): *Multiple Objective and Goal Programming*. Physica, Heidelberg, 400–412.
- TRZPIOT, G. (2003): Advanced modeling in finance using multivalued stochastic dominance and probabilistic dominance. *Journal of Economics and Management*, KAUe Katowice (in appearance).
- WHITMORE, G.A. (1970): Third Degree Stochastic Dominance. *Amer. Economic Rev.*, 60, 457–459.

Part III

Pattern Recognition and Computational Learning

Classification of Method Fragments Using a Reference Meta Model

Werner Esswein and Andreas Gehlert

Department of Business Management and Economics,
Chair of Business Informatics, esp. Systems Engineering,
Dresden University of Technology,
D-01062 Dresden

Abstract. Modelling Methods were developed to formalise the process of information system development. These methods are often inappropriate for solving specific problems. So, the new discipline situational method engineering was established. The aim of this discipline is to engineer specific modelling methods by combining existing method fragments. In the current approaches, these method fragments are stored independently and unclassified in a method base. Therefore, the administration of a method base is inefficient. Our approach shows how the well established domain of reference modelling can be used for method engineering. We provide a reference meta model suited as a method base. It classifies existing method fragments and stores the relations between them.

1 Motivation

The process of information system (IS) development has been formalised to enhance the quality of software and to forward its development. Different modelling methods were evolved to support this process, especially in its early analysis and design stages.

Harmsen describes a method as "...an integrated collection of procedures, techniques, product description and tools, for effective, efficient and consistent support of the IS engineering process." (Harmsen (1997), p. 11) A method fragment is "a description of an IS engineering method, or any coherent part thereof." (Harmsen (1997), p. 26) According to these definitions a method can be viewed as a composition of method fragments.

The existing modelling methods are often either too unspecific or too extensive for the particular problem at hand (Brinkkemper et al. (1998), p. 381). So, the situational method engineering as part of the method engineering discipline was developed. The aim is the formal controlled, computer-aided construction of situational methods on the basis of existing method fragments (Harmsen (1997), p. 28).

Harmsen describes situational method engineering as a process containing four major steps (see Figure 1). Firstly, the situation of the specific project must be analysed. The method engineer is than, secondly, able to select specific method fragments from an existing method base suited for a specific project. These method fragments are, thirdly, assembled to a cohesive

and consistent method using the pre-defined rules defined by a meta-method (Harmsen (1997), p. 38). The new method is, lastly, used during the execution of the project. Arising problems during the execution of the project can be analysed to enhance the quality of the existing method base (experience accumulation). The method administration executes this task. It is their responsibility to keep the method base consistent and to incorporate practical experiences as well as new theoretical findings.

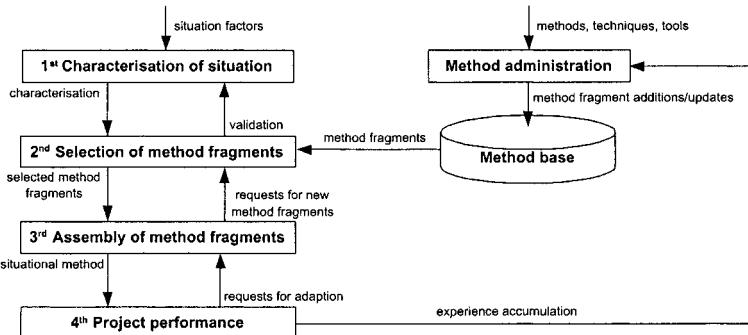


Fig. 1. The process of situational method engineering (Harmsen (1997), p. 31)

One of the key success factors of Harmsen's approach is the method base. To assemble method fragments stored in this method base it must fulfil two criteria. First, the fragments must be classified so that they can be easily retrieved and, second, the method base must provide a mechanism to store the relations between these method fragments. The latter criterion will eliminate the assembly process resulting in a better overall performance of the method engineering process. Current approaches disregard both aspects resulting in inefficient method bases (Harmsen (1997), pp. 208; Brinkkemper et al. (1998), pp. 385).

We will show in this paper how the integration of the reference modelling and the method engineering discipline can improve the administration and usage of method bases. Consequently, we develop a reference model suited as a method base providing a classification of method fragments as well as representing the relations between them.

To show that a reference meta model can be used as a method base it must define all operations for a method base. Harmsen defines the following operations: definition of a new method fragment, update a method fragment, delete a method fragment and extract a method fragment (see Figure 1 and Harmsen (1997), pp. 183).

The paper is structured as follows: The starting point of the description of the reference meta model is the process of constructing reference models (see Section 2). The elements of a reference meta model are deduced from this

process model. It is shown that the elements of the reference meta model can fulfil both, the requirements for a method base as well as the classification of method fragments and their inter-fragment relations. The paper finishes with a critical summary and suggestions for further work.

2 The reference meta model

Rosemann and Schütte define a reference model as model, which describes common knowledge of a problem domain and is useful for the conversion into specific models (Becker and Schütte (1997), p. 428). The components of such reference models can be extracted from their process model described by Rosemann and Schütte (Rosemann and Schütte (1999), pp. 26):

1. Problem definition: definition of the scope of the reference model,
2. Classification schema: schema for classifying the components of the reference schema,
3. Reference schema: modelling of the problem domain,
4. Process model: description of the usage of the reference schema for deriving specific models.

For the description of the reference model of this paper these elements are analysed next in further detail.

2.1 Problem definition

Firstly, the scope of a reference model must be defined (Rosemann and Schütte (1999), p. 28). The suggested reference model is used as a method base for situational method engineering. It defines the existing system engineering method fragments (see section 1).

Cronholm classifies these method fragments according to their layer of granularity in methods, stages, models, diagrams and concepts. Furthermore, he distinguishes between product and process fragments (Cronholm and Agerfalk (1999)).

The reference model presented in this paper resides on the meta-level since it describes models rather than the problem domain itself (see Object Management Group (2002), p. 2-3). According to Rosemann and Schütte the term reference meta model must be used (for classification of reference models see Schütte (1998), p. 71 and Rosemann (1996), p. 22). This reference meta model represents the conceptual product fragments of visual IS modelling techniques (problem definition), their classification (classification schema), the relation between them (reference meta schema) and the description of the use of this reference meta model (process model).

2.2 Classification schema

After the problem definition the classification schema of the reference meta model can now be developed. It provides the uppermost abstraction level of the model. This classification is mapped to the reference meta schema in the next step to structure the model elements in cohesive parts.

Consequently, the classification schema is used to describe the structure of the reference meta schema and eases the administration of the meta schema.

According to the classification schema the reference meta schema is hierarchically decomposed. The concepts in the upper level of this hierarchy are specialised to form more specific concepts in the lower levels. So, concepts residing on different levels are in an abstraction relationship. Furthermore, the abstraction system is poly-hierarchical meaning that a concept on a lower level could be reached from more than one concept on the upper level of this hierarchy (DIN2331 (1980)).

To use this classification schema all elements of the reference meta schema are grouped in four general abstraction layers:

1. The first layer is used to describe general relationships between all method fragments modelled.
2. In the second layer—the view-layer—the fragments are grouped according to the aspect the method-fragment mainly describes (Winter (2000), p. 32).
3. This classification is still not satisfying so that these views are further divided into paradigms. The fragments grouped to a special paradigm follow all together a specific modelling pattern (Winter (2000), p. 34).
4. The last level contains the method fragments themselves.

The first level is used to describe the relations of the view of the second level and, thus, does not provide any classification. It is the starting point of the classification schema (see Figure 2).

The second level is used for the first classification through four different views, the organisation view, the task view, the object view and the process view. Because of the described poly-hierarchical character of the classification schema these views are not orthogonal to each other.

The organisation view groups method fragments for the modelling of organisational systems. Thereby, two aspects can be distinguished. Firstly, the focus can be on the communication between organisational units forming the communication paradigm. Secondly, the structure of organisational units can be modelled using the fragments from the position-structuring paradigm.

Tasks of organisations can be characterised as an execution on one or more objects reflected by the task view. Therefore, the task-structuring paradigm represents all method fragments focussing on the composition and decomposition of tasks.

Method fragments of the object view represent all objects of the system to be modelled as well as their relations. According to the object-relationship

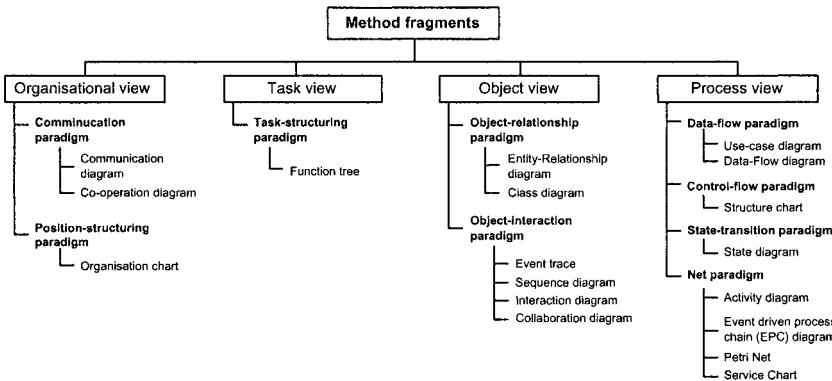


Fig. 2. The classification schema

paradigm these objects are typed to classes. It focusses on the modelling of the static structure of these classes. The object interaction paradigm, in contrast, highlights the description of the interaction between these objects.

The process view describes a process as the logical and temporal sequence of tasks. These can be further distinguished into four different paradigms. The data-flow paradigm illustrates the logical aspects of a process. It focuses on the structure of the processes. To model the dynamic aspects of a system the state-transition paradigm, the data-flow paradigm as well as the net paradigm are suitable. In the state-transition paradigm the aspect of changing in state of each object is described, while in the net paradigm processes are illustrated by successions of events and activities. Method fragments of the control-flow paradigm exclusively concentrate on the modelling of dynamic behaviour caused by regular structures (loops, forks, sequences and parallelism) (Winter (2000), p. 56).

2.3 Construction of the reference meta schema

After the description of the classification schema this schema is used for the construction of the reference meta schema. The classification is carried over to the elements of the meta schema as they are grouped according to it.

A class diagram is used to model the reference meta schema. Concepts of method fragments are represented as classes. Relations between these concepts are modelled as inheritance, association or aggregation. Packages are used to represent the classification schema.

These packages group relevant concepts to form a cohesive sub-model. Each of the packages can be viewed relatively independent from each other so that the administration (add, modify, delete a method fragment) as well as the retrieval are easier to handle. Additionally, because of these groupings, according to the classification schema, the sub-models are less complex and thus easier to handle. Moreover, the abstraction relations between the

elements residing on the higher levels and elements on the lower levels of the hierarchy, eases the understanding of the model by studying the general elements and their relations first and by continuing with the more specific elements on the lower levels afterwards. This is a major advantage over the existing method bases.

As described earlier in this paper the upper package models the general relations between all method fragments (see Figure 3). The concepts object, organisational unit, task and process are modelled. An organisational unit of the organisation view is described as a specific object, which is responsible for specific tasks. Objects provide a solution for these tasks, which are executed by a process.

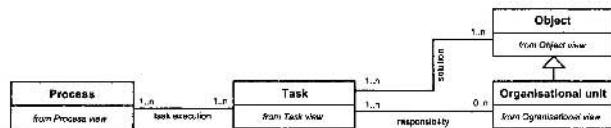


Fig. 3. Reference meta schema: method fragment package

To show the dependencies between the different views all views are depicted in Figure 4. The packages are highlighted.

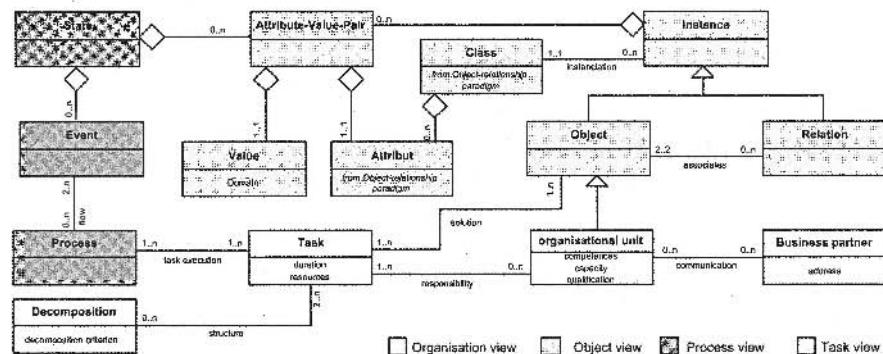


Fig. 4. Reference meta schema: Views

2.4 Process model

After the construction of the reference meta schema the process model can be described. Thereby, the second operation on a method base—the selection and extraction of method fragments—is demonstrated.

Schütte discusses two possibilities to use a reference model (Schütte (1998), p. 218). According to his view, it should be possible to customise the refer-

ence schema first and duplicate it afterwards or to perform the opposite way. The first solution is preferred if similar customisations to the copies of the reference model are needed. In contrast, he prefers the second way, copying before customising, as the modifications of the user can be analysed and used to improve the reference schema. Whereby, the project organisation and the communication infrastructure have to prevent redundant customisations (Schütte (1998), p. 219).

Consequently, the first step of the process model would be the selection of the required elements. Secondly, these elements are copied to a separate meta schema. In the last step, the user can customise this copy according to her needs.

For an easy selection of the required method fragments the classification schema from Subsection 2.2 can be used. Because of the clear structure of the reference meta schema provided by the classification schema the user must only perform three steps to select the method fragments needed. First, the user chooses the relevant view by asking, which perspective is need to be modelled. Second, by selecting of the specific view the user can now choose between different paradigms according to the modelling pattern. After the selection of the paradigm, third, the user can now retrieve the specific method fragment.

Because of the modelled inter-fragment relations the user has the possibility to choose the corresponding fragments as well. The selection of the method fragments finishes with the best match of the concepts offered and the user's specification.

After the selection of the method fragments they are copied into a separate meta schema. The user can now make the modifications to form the specific method fragment needed for the project.

These customisations can be analysed and used for enhancing the quality of the reference meta schema.

With the description of the process model the second operation on a method base is explained (see section 1).

3 Summary

The operations on a method base must be described for the reference model proposed in this paper to show that this kind of model can be used as a method base.

In section 1 the two main operations, administration and selection of method fragments, were analysed. These operations were described in the second section. Since these operations can be defined for the reference meta schema, it fulfils the basic requirements for a method base in the method engineering process.

Furthermore, a classification schema was developed in Subsection 2.2. It was shown that this classification schema not only eases the administration

of the reference meta schema by separating it into different packages but also the selection of method fragments from it. Moreover, the pre-existing inter-fragment relations enhance the selection process because these relations need to be modelled only once.

To prove if this reference meta model can be implemented into practice we will use it to generate specific meta models (method fragments) for an adaptable CASE tool.

Furthermore, an empirical performance evaluation is needed to show if the approach presented in this paper is superior to other method engineering approaches. However, a performance evaluation of IS development methods and techniques has been done very rarely and on very basic concepts only. Even in the well established domain of reference modelling this has not been done so far. Therefore, this should be subject of further research.

References

- BECKER, J. and SCHÜTTE, R. (1997): Referenz-Informationsmodelle für den Handel: Begriff, Nutzen und Empfehlungen für die Gestaltung und unternehmensspezifische Adaption von Referenzmodellen. In: H. Krallmann (Ed.): *Internationale Geschäftsfähigkeit auf der Basis flexibler Organisationsstrukturen und leistungsfähiger Informationssysteme*. 427–447.
- BRINKKEMPER, S., SAEKI, M., and HARMSEN, F. (1998): Assembly Techniques for Method Engineering. In: *Lecture Notes of Computer Science*. No. 1413, Springer-Verlag, 381–400.
- CRONHOLM, S. and AGERFALK, P.J. (1999): On the Concept of Method in Information Systems Development. In: T. Käkölä (Ed.): *Proceedings of the 22nd Information Systems Research In Scandinavia (IRIS 22)*.
- DIN DEUTSCHES INSTITUT FÜR NORMUNG E.V. (NORMENAUS-SCHUSS TERMINOLOGIE) (1980), DIN 2331: *Begriffssysteme und ihre Darstellung*.
- HARMSEN, A.F. (1997): *Situational Method Engineering*. Master's thesis, University of Twente.
- OBJECT MANAGEMENT GROUP (2002): *Meta Object Facility (MOF) Specification*. Version 1.4.
- ROSEMANN, M. (1996): *Komplexitätsmanagement in Prozeßmodellen: methoden-spezifische Gestaltungsempfehlungen für die Informationsmodellierung*. Gabler-Verlag, Wiesbaden.
- ROSEMANN, M. and SCHÜTTE, R. (1999): Multiperspektivische Referenz-modellierung. In: J. Becker, M. Rosemann, and R. Schütte (Eds.): *Referenz-modellierung: State-of-the-Art und Entwicklungsperspektiven*. Physica-Verlag, 22–44.
- SCHÜTTE, R. (1998): *Grundsätze ordnungsmäßiger Referenzmodellierung: Konstruktion konfigurations- und anpassungsfähiger Modelle*. Gabler-Verlag, Wiesbaden.
- WINTER, A. (2000): *Referenz-Metaschema für visuelle Modellierungssprachen*. Deutscher Universitäts-Verlag, Wiesbaden.

Finding Metabolic Pathways in Decision Forests

André Flöter¹, Joachim Selbig², and Torsten Schaub¹

¹ Institut für Informatik, Universität Potsdam, August-Bebel-Str. 89/Hs.4,
D-14482 Potsdam, Germany

² Max-Planck-Institut für molekulare Pflanzenphysiologie, D-14424 Potsdam,
Germany

Abstract. Data of metabolite concentrations in tissue samples is a powerful source of information about metabolic activity in organisms. Several methods have already been reported that allow for inferences about genetic regulatory networks using expression profiling data. Here we adapted these techniques for metabolic concentration data. While somewhat accurate in predicting certain properties these methods encounter problems when dealing with networks containing a high number of vertices (> 50). To circumvent these difficulties, larger data sets are usually reduced à priori by means of preprocessing techniques. Our approach allows to make network inferences using ensembles of decision trees that can handle almost any amount of vertices, and thus to avoid time consuming preprocessing steps. The technique works on a bootstrap principle and heuristically searches for partially correlated relations between all objects. We applied this approach to synthetically generated data as well as on data taken from real experiments.

1 Introduction

The reconstruction of valid generative models from data has been subject of several scientific disciplines. Some of the developed techniques have now been used and enhanced to derive models of genetic networks from gene expression data. Most notable are the works of Dana Pe'er (2001) and Horimoto Katsuhisa (2001). Generated models are hoped to disclose yet unknown coherences, in particular quantifiable dependencies between the expression levels of genes. Such a dependency could be preconceived in future biological research in order to find explanations for the functionality of genes.

With a similar intention we considered those methods for constructing generative models from data of metabolic concentrations produced by gaschromatographic mass spectrometry (GC/MS) (Fiehn (2000)). The derivation of dependencies between the metabolites of the given data set is a desirable capability in the process of finding new or validating known metabolic pathways respectively. Newly discovered dependencies can be used as feasible hypothesis to be examined in further experiments.

Up to now, the reported techniques all have a high computational complexity, which makes it difficult to use them on large data sets, such as expression profiling data. Also, if it is required to include these algorithms in

a loop, e.g. in an automatic analysis, they are unfit for smaller data sets like GC/MS data. This is why we developed a pragmatic approach that adopts some of the heuristics of the decision tree algorithm C4.5 (Quinlan (1993)) and thereby reduces the complexity of the needed calculations. This approach and a sample result will be discussed below.

2 Background and related work

2.1 Visualizing correlation graphs

A very simple approach to extract information from given metabolic concentration data is the visualization of correlations within a data set (Kose (2001)). This method computes Pearson correlation coefficients (Statistica (2002)) between all measured metabolites. The visualization is then given by a graph made of one node for each metabolite. For each correlation coefficient that exceeds a given threshold the respective nodes are linked with an edge. Once all edges are established it is possible to group those metabolites which are interconnected to one another. Groups, having edges between all of their members, are called cliques. It is presumed that cliques contain metabolites which appear in related metabolic pathways.

This approach is based on the Pearson correlation coefficient. It is therefore only possible to detect undirected linear and pairwise correlations with it, thus no information can be drawn about a cause and an effect of a computed correlation. Our approach extends this work by using and calculating a different type of correlation.

2.2 Usefulness of partial correlation

Some illustrative examples of standard statistical textbooks (Statistica (2002)) show that more information can be discovered while searching for dependencies between variables, if partial correlations are computed instead of simple correlation coefficients.

Partial correlation reveals correlations between more than two variables. It is sometimes also called “conditional” correlation, because it searches for correlations between two variables under a certain fixed condition of other variables. With this property, it is potentially possible to establish a cause from an effect (Shipley (2002)). Hence, we can use this feature to attribute directions to the edges of the above graph.

However, there are two drawbacks concerning the computability of partial correlations: First, the calculation has a high complexity, because pairwise correlations have to be calculated of all variables for any considered condition of the other variables. Second, small data sets lead to unreliable results, because there are very few samples left, once the data possessing a certain fixed condition is removed.

Therefore, it would be desirable to have a means of reducing the computational cost and circumvent the problems with sparse data sets.

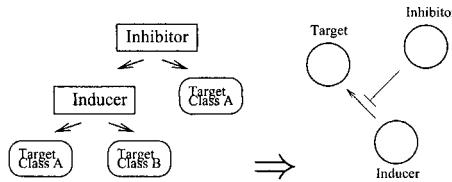


Fig. 1. A two-node tree with a possible interpretation as a hypothesis for a partial correlation.

2.3 Partial correlation and decision trees

Decision trees are a representation of discrete-valued functions. The parameters of the function can either be discrete-valued or real-valued. In practice, these trees are often learned automatically by one of several induction algorithms in order to approximate a target function over a given data set (Mitchell (1997)). The goal of such a function is always to classify a data object (a vector of attributes) into one of several possible target classes. Classifying an object is sometimes also referred to as predicting the value of the object's target variable.

At first glance, inducing a decision tree and calculating (partial) correlation are two completely different things. However, there is a relation between them. A decision tree can be used to predict a discrete target variable. For decision-making the tree uses the values of some other variables of the data set. So, a tree predicts the target variable by the value of other variables. If a correlation between two variables is known with all parameters, it is also possible to predict the value of the one variable by the value of the other. The difference is that a real valued variable can be predicted with the knowledge of a correlation whilst a decision tree can only predict discrete-valued variables.

If we discretize a real-valued target variable in an appropriate way, e.g. intervals of the target variable's domain, the above observation suggests that a decision tree would use those variables for a prediction which are correlating to the target variable. In fact, if there is one strongly correlating variable to the target variable, the induction algorithms will choose it as the only node in a tree for predicting the target, because the entropy regarding the target classes will drop most, if the data is split according to the value of a correlating variable.

Further, if a variable correlates only to the target variable under a certain condition of another variable, that is a partial correlation, the induction algorithms are likely to reflect this by generating a tree with two levels, which tests for the conditional variable in the top node and for the partially correlating variable in the subsequent node(s) as illustrated in Figure 1.

Thus, a decision tree offers a feasible hypothesis for one existing (partial) correlation between the target variable and other variables of the data. Since

decision trees can be grown at a lower computational complexity than direct calculation of partial correlations, they could be used to find them in a faster manner.

3 Algorithm for growing and evaluating trees

3.1 Growing a decision forest

Due to an unlimited number of possible discretizations of a real-valued domain there is potentially a vast number of possible decision trees that can be grown over a real-valued target variable. It would therefore be convenient to have a way of growing a limited set of trees with little redundancy in their indicated correlations. Sets of decision trees are also referred to as decision forests.

We propose an iterative algorithm for the growth of a decision forest that removes each variable that appears as top node of an induced tree from the data set and then induces the next one on the same target. Thereby the maximum number of produced trees is limited to the number of variables in the data set. Since each variable can appear only once as top node in a tree, that is a key variable for the proposed correlation, there will additionally be little redundancy in the set of suggested hypothesis.

Due to the heuristics utilized by C4.5 the algorithm tends to reveal simple correlations in simple trees¹ as long as directly correlating variables are still in the data set. Once those are taken out of the data, C4.5 tries to find combinations of (foremost two) variables that still allow for a precise prediction of the target. These two-node trees primarily test for partially correlating variables. The property of delivering simple trees at first and more complex ones on later iterations will later be used to terminate the computation early.

To run this algorithm we still need to identify a target variable and discretize it in an appropriate way. We suggest to use a variable that indicates the presence of two separate states, because the instances of such a variable can easily be discretized into its two states, leaving only 0 and 1 as classes for the target. To find an eligible target it is helpful to look for bimodally distributed variables as illustrated in Figure 2. This can be done either with statistical means (Silverman (1981), Statistica (2002)), or visually, or through expert knowledge about a specific variable that reflects two states.

A very effective way to reduce the amount of computation in this step is to limit the depth of the decision trees in advance. If the depth is limited to 2 levels, hypothesis for partial correlations will still be found, but the complexity of the calculation is reduced by an entire factor.

¹ normally one-node trees

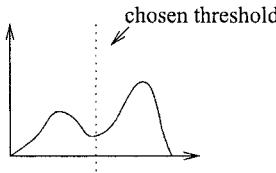


Fig. 2. A bimodally distributed variable being discretized visually.

3.2 Evaluating trees from the forest

To reduce the number of trees to be taken as potential hypothesis for (partial) correlations we introduce an evaluation function. Intuitively, it is desirable to get trees which classify most data objects correctly only with a few decisions. This would mean that the target variable can be predicted on the basis of only a few other variables, proposing that these are in strong (partial) correlation with the target. Whereas a tree needing many decisions for prediction supposedly tests variables that are only weakly (partially) correlated to the target.

With C_i being the set of correctly classified objects in leafs at depth i of a decision tree the function

$$\text{value}(T) := \sum_{i=1}^{d_{\max}} |C_i| \cdot \frac{1}{e^i}$$

delivers a value between 0 and $\frac{1}{e}$ times the number of all data objects. A tree correctly classifying all objects close to the top receives a value close to or above 1, and a value close to 0 is given for trees that need many decisions to make their prediction.

This function can also be used to decrease the number of tree inductions, thus further reducing the computational cost of the growth of the forest. The first trees in the induction process tend to put strongly correlating variables at the top of the tree, because they split the data with least entropy. If there are no strongly correlating variables left, the objects can only be classified after a combination of tests on other variables. The more complicated the tree, the less probably it proposes an existing correlation, and the less is the value of the evaluation function. For this matter the growth of the forest could be stopped when the value of the evaluation function falls short of a certain threshold².

3.3 Assembling a network

At this point we have shown how to efficiently grow a forest and how to select potentially good hypothesis with the help of an evaluation function. We now fit the selected hypothesis into a graph according to a few simple rules:

² Empirically the value 0.5 showed to be a reasonable choice for a threshold.

1. A node is put into the graph for the target attribute.
2. For each one-node³ tree, we insert a node for the test attribute of the tree into the graph and an undirected edge between this node and the node of the target attribute.
3. For each two-node tree, we insert a node into the graph for each node of the tree and a directed edge from the node of the second level of the tree which classifies most objects to the node of the target attribute. A second directed edge is inserted from the node referring to the top node of the tree to the first directed edge. The second edge should have a different styled pointer, because it indicates an inhibition or activation of the influence of the first edge.
4. All nodes in the graph referring to the same variable are merged.

The complete procedure (selecting target, growing forest, assembling graph) can now be rerun with another target variable until all suitable target variables have been used. The graph will grow with each iteration of the procedure.

The algorithm for generating a network needs a data matrix and delivers a hypothesis for a network in form of a graph. Here is a summarized description of it:

1. While suitable target variables are present do
 - (a) Choose a target variable and discretize it.
 - (b) Grow and store a decision tree to predict the target variable.
 - (c) Take the top variable of the generated tree out of the data set.
2. Assemble the network from the stored decision trees.

4 Application on data of metabolic concentrations

We tested the algorithm on metabolic concentration data of potatoes. The potatoes had been grown under two distinct conditions, so some of them could respond to one condition and some to the other. Because of these two situations a few of the concentrations of the metabolites were clearly bimodally distributed, indicating a direct reaction to the two environments.

To examine the output of the algorithm in a biological context, we chose one experimentally unidentified⁴, bimodally distributed metabolite as target, a maximum tree depth of two, and an evaluation threshold of 0.5 for the algorithm. The resulting graph is shown in Figure 3. The numbers in front of each name are the trace and the retention index from the GC/MS experiment. Here they serve exclusively as identifiers; for metabolites with unidentified names, as is the target metabolite in the centre, they are the only reference.

³ We also consider trees that classify a vast majority of all objects correctly with just one node as one-node trees.

⁴ Sometimes, in GC/MS experiments metabolites are clearly detectable and measurable, but they cannot reliably be identified.

The numbers above or below the names are the values of the evaluation function for the corresponding decision trees.

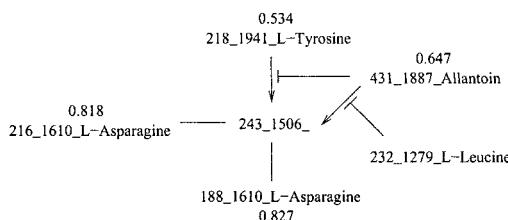


Fig. 3. The output of an iteration of the algorithm for an example metabolite.

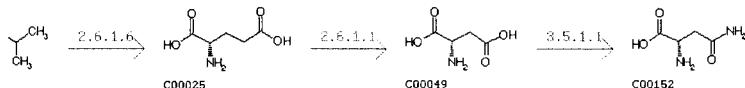


Fig. 4. A pathway of the KEGG database leading from Leucine on the left to Asparagine on the right.

By consulting KEGG we found that all displayed metabolites of the output graph are involved in the amino acid metabolism. The KEGG database (<http://www.kegg.com/>) is a resource comprising large-scale data in Genomics, Proteomics and Metabolomics. In particular, we retrieved the pathway of Figure 4 from Leucine to Asparagine approximately matching an indicated correspondence in the graph. It perambulates the same metabolites as the direct path from Leucine to Asparagine in the graph of Figure 3. The notation and numbers in the figure are taken directly from KEGG. The numbers below the structural depictions are identifiers for the metabolites, the numbers above the arrows indicate the involved enzyme for the reaction. They are discussed in more detail on the KEGG site.

These results confirm an interrelation of the algorithm's output and biological functionality.

5 Discussion

We have introduced an algorithm that produces a partially directed graph visualizing (partial) correlations from data. The output graph can be interpreted as a hypothesis for causalities between the variables of the data set. We have shown in an example that the delivered hypothesis is also consistent with biological knowledge, specifically with metabolisms.

Our approach is driven by the demand to produce hypothesis for metabolic (sub-) networks at a reduced computational cost. The obtained results are encouraging in view of this goal. This allows us to run the algorithm in a loop or a computationally demanding environment. Hence, we hope that the use of our algorithm on metabolic concentration data will allow for quicker and more complex results.

We can only compare our work in terms of computational cost to other works with a similar objective, because the other approaches have only been applied on expression profiling data, yet. But due to the lower complexity of a heuristical search as opposed to an exhaustive search our algorithm does offer an advantage on the computational part.

There are several ideas on how to further improve the algorithm. First, it would be a helpful feature to be able to evaluate the induced hypothesis not only on a statistical measure, but also on a function based on prior knowledge. Second, the discretization of the target variable(s) is mostly done manually at this time. General statistical approaches for finding bimodalities (K-Means-Clustering, Silverman-Test) fail to deliver usable results on sparse data sets. For a fully automatic data analysis, however, it is necessary to automatically find and reliably discretize bimodalities in a biological context.

Acknowledgements: We thank Joachim Kopka, Ralf Steuer, and Jacques Nicolas for their helpful collaboration and comments.

References

- FIEHN, O., KOPKA, J., DÖRMANN, P., ALTMANN, T., TRETHEWEY, R.N., and WILMITZER, L. (2000): Metabolite profiling for plant functional genomics. In: Proc. Natl. Acad. Sci. USA, 95, 14863–14868.
- HORIMOTO, K. and TOH, H. (2001): Automatic System for Inferring a Network from Gene Expression Profiles. *Genome Informatics*, 12, 270–271.
- KOSE, F., WECKWETH, W., LINKE, T., and FIEHN, O. (2001): Visualising plant metabolomic correlation networks using clique-metabolite matrices. *Bioinformatics*, 17, 1198–1208.
- MITCHELL, T. (1997): *Machine Learning*. McGraw-Hill, Boston, USA.
- PE'ER, D., REGEV, A., ELIDAN, G., and FRIEDMAN, N. (2001): Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17, Suppl. 1, 215–S224.
- SHIPLEY, B. (2002): *Cause and Correlation in Biology*. Cambridge University Press, Cambridge, UK.
- SILVERMAN, B.W. (1981): Using Kernel Density Estimates to investigate Multimodality. *Journal of Royal Statistical Society, Series B*, 43, 97–99.
- STATISTICA (2002): *Electronic Textbook StatSoft*.
<http://www.statsoftinc.com/textbook/stathome.html>.
- QUINLAN, J.R. (1993): *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, USA.

Randomization in Aggregated Classification Trees

Eugeniusz Gatnar

Institute of Statistics,
Katowice University of Economics, ul. Bogucicka 14, 40-226 Katowice, Poland

Abstract. Tree-based models are popular and widely used because they are simple, flexible and powerful tools for classification. Unfortunately they are not stable classifiers. Significant improvement of model stability and prediction accuracy can be obtained by aggregation of multiple classification trees. The reduction of classification error is a result of decreasing bias or/and variance of the committee of trees (called also an ensemble or a forest). In this paper we discuss and compare different methods for model aggregation. We also address the problem of finding minimal number of trees sufficient for the forest.

1 Introduction

Much effort is being taken in statistics to improve the prediction accuracy of classification or regression models. Recently instead of improving single models we build hundreds of models and combine them into an ensemble. Then the component models vote for the predicted value of the outcome.

Proposed methods, i.e. bagging, adaptive bagging, and arcing are based on sampling cases from the training set while boosting is deterministic¹ and uses a system of weights for cases and combined models.

Although resampling causes major modification of the distribution of predictors in the training samples, significant improvement of classification accuracy can be also achieved by random selection of variables to training samples or directly to the model.

Recent developments in this field showed that the randomization leads to consistent models while boosted models can overfit for large number of their components.

A classifier C is a function that maps from object descriptions (\mathbf{X}) to class names (\mathbf{Y}):

$$C : \mathbf{X} \rightarrow \mathbf{Y}, \quad (1)$$

and it is found by learning a set of previously classified objects, called „training set” T :

$$T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}, \quad (2)$$

where \mathbf{x}_i is an element from M -dimensional space \mathbf{X} , and y_i is an element from discrete space \mathbf{Y} .

¹ Except stochastic gradient boosting (Friedman, 1999).

The goal of the learning is simple – find a classifier $C(\mathbf{x})$ that gives the lowest prediction error.

2 Classification error

If (x, y) is independent and identically distributed from the (\mathbf{X}, Y) distribution, the prediction error is the expected value of a loss function L :

$$PE(C) = E_{\mathbf{X}, Y} L(y, \hat{C}(\mathbf{x})), \quad (3)$$

where L measures the loss between true class y and the prediction $\hat{C}(\mathbf{x})$.

The mostly use squared loss function:

$$L(y, \hat{C}(\mathbf{x})) = (y - \hat{C}(\mathbf{x}))^2 \quad (4)$$

is suitable for regression (continuous Y), but inappropriate for classification (discrete Y). Therefore the 0-1 loss function is used instead:

$$L(y, \hat{C}(\mathbf{x})) = \begin{cases} 0 & \text{if } y = \hat{C}(\mathbf{x}) \\ 1 & \text{if } y \neq \hat{C}(\mathbf{x}) \end{cases}. \quad (5)$$

The best way to estimate the prediction error PE is to have an independent test set T . In the absence of this set one can use the training set, but the model is likely to overfit, i.e. the classification error is equal 0 for large enough models. Different strategies are used to obtain better a estimation of the error, e.g. cross-validation or bootstrapping.

2.1 Bias-variance decomposition

Very useful in analysis of classification algorithms is the decomposition of the prediction error proposed by Breiman (1996):

$$PE(C) = error(C^B) + bias^2(C) + variance(C), \quad (6)$$

where $error(C^B)$ is the error of the Bayes optimal model (irreducible), $bias$ that is the systematic error measuring how well the classifiers match the true class on average, and $variance$ that measures how much classifiers vary for different training sets.

For the 0-1 loss (5) there are several variants of the decomposition (6), proposed by Dietterich and Kong (1995), Friedman (1996), Breiman (1996b), Tibshirani (1996), and Kohavi and Wolpert (1996).

There is a tradeoff between bias and variance, i.e. the more complex the model C , the lower the bias and the higher the variance.

2.2 Improvement of prediction accuracy of single models

To improve the accuracy of a single tree-based model several methods that reduce the prediction error have been developed (Gatnar (2002)). For example:

- *soft splits* (fuzzy thresholds) proposed by Carter and Catlett (1987) and successfully applied in the C4.5 system,
- *logical combination of splits*,
- *pruning* developed by Breiman et al. (1984) that removes branches of a tree with high variance,
- *shrinking* invented by Hastie and Pregiborn (1991) that is similar to pruning but instead of cutting off subtrees it changes the distribution of the class labels in the leaves of the tree.

3 Model aggregation

Instead of building a single model we can build multiple models (e.g. hundreds): $C_1(\mathbf{x}), C_2(\mathbf{x}), \dots, C_M(\mathbf{x})$ and combine them into one committee or ensemble². There are two general methods to combine trees: merging (multiplicative) that leads to huge models³, and averaging (additive) that is more efficient and frequently used.

In the aggregate $\hat{C}^*(\mathbf{x})$ the component models vote for the predicted class:

$$\hat{C}^*(\mathbf{x}) = \operatorname{argmax}_{y \in Y} \left\{ \sum_{m=1}^M I(\hat{C}_m(\mathbf{x}) = y) \right\}. \quad (7)$$

Several variants of aggregation methods were proposed. They manipulate training cases (random sampling) or predictors (random selection) or values of the y (system of weights) or involve randomness directly.

Combined classifiers have a lower error rate than single models and can give more insight. But their disadvantages are: slow learning (requires large computer memory) and loss of comprehensibility (less structure and huge models).

Some aggregation methods require „weak” classifiers. A „weak” classifier has the error rate bounded by a constant strictly less than 0.5, i.e. its error rate is only slightly better than random guessing. It is also unstable because of its high variance. That is a small change in predictor values which can result in a quite different model⁴.

² When component models are trees the combiner is called a forest.

³ The size of merged trees is proportional to the product of component tree sizes.

⁴ Trees, neural networks and nearest neighbors are examples of weak classifiers.

3.1 Early methods

Perhaps *twicing* (Tukey (1977)) was the first method used for combining models in statistics. But recently *stacking* or stacking generalization (Wolpert (1992)) started again the interest in model aggregation. 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).

Quinlan (1993) implemented in his C4.5 system a method called *windowing* that enlarges the initial random sample drawn from the training set by adding in consecutive steps cases misclassified in the previous step.

3.2 Bagging

Bagging (bootstrap aggregating) was the first aggregation method proposed by Breiman (1996a). It used multiple training bootstrap samples U_1, U_2, \dots, U_M to create classifiers that vote for the final prediction (7).

Breiman proved that bagging reduces the variance but leaves the bias unchanged. Therefore it works better with very large component tree models (unbiased as possible) but they should not be correlated.

There are two variants of bagging. *Adaptive bagging* proposed by Breiman (1999) reduces both bias and variance and works by changing output values y by using „out-of-bag” cases in subsequent steps:

$$y_i^{(s+1)} = y_i^{(s)} - \overline{\hat{C}_m(\mathbf{x}_i)}, \quad (8)$$

where $\overline{\hat{C}_m(\mathbf{x}_i)}$ is the average over predicted values for training samples U_m such that $\mathbf{x}_i \notin U_m$.

Wagging introduced by Bauer and Kohavi (1999) is similar to bagging, but reweights cases instead of sampling them from the training set.

3.3 Boosting

The most accurate classifiers are those combined by *adaptive boosting*. AdaBoost.M1 developed by Freund and Schapire (1997) is deterministic, sequential and works with two systems of weights: assigned to training examples and assigned to component models.

Initially the weights of cases are uniform ($w_i = 1/N$), but in consecutive steps the weights of cases misclassified in the previous step are increased by a factor inversely proportional to the training sample error:

$$w_i^{(s+1)} = w_i^{(s)} e^{a_m I(y_i \neq \hat{C}_m(\mathbf{x}_i))}, \quad (9)$$

where:

$$a_m = \log \left(\frac{1 - e_m}{e_m} \right), \quad (10)$$

and the error rate is:

$$e_m = \frac{\sum_{i=1}^N w_i I(y_i \neq \hat{C}_m(\mathbf{x}_i))}{\sum_{i=1}^N w_i}. \quad (11)$$

The final prediction is:

$$\hat{C}^*(\mathbf{x}) = \operatorname{argmax}_{y \in Y} \left\{ \sum_{m=1}^M a_m I(\hat{C}_m(\mathbf{x}) = y) \right\}. \quad (12)$$

Breiman referred to AdaBoost with trees as „the best classifier in the world”. It is successful because it reduces both the bias and the variance. Boosting requires weak classifiers, e.g. „stumps” or pruned trees. Unfortunately, numerical precision problems occur because for example an error of 0.1 will cause weights to grow by factor of 500.

There are two variants of boosting: *boosting by weighting* (AdaBoost.M1) and *boosting by sampling* (cases are sampled from the training set with probability proportional to their weights).

3.4 Arcing

Because of the apparent success of AdaBoost Breiman (1998) applied a system of weights to his new Arc-x4 algorithm (based on bootstrap sampling) called *Adaptive Resampling and CombinING*. It is also sequential and reduces both bias and variance.

Arc-x4 increases weights of misclassified cases:

$$w_i = 1 + [n(\mathbf{x}_i)]^4, \quad (13)$$

where $n(\mathbf{x}_i)$ is the number of misclassifications of the case \mathbf{x}_i done by classifiers: C_1, C_2, \dots, C_{m-1} .

To compare prediction accuracy of ensembles for different methods we used 5 benchmark datasets (Table 1) from the Machine Learning Repository at the UCI (Blake et al., 1998). Results of the comparisons are presented in

Dataset	Number of cases	Number of cases in test set	Number of predictors	Number of classes
DNA	2000	1186	180	3
Letter	15000	5000	16	26
Satellite	4435	2000	36	6
Soybean	683	68	35	19
Zip-code	7291	2007	256	10

Table 1. Benchmark datasets from UCI Repository.

Table 2. For each dataset an aggregated model has been built containing 100 component trees⁵.

Dataset	Bagging	Arc-x4	AdaBoost	Single tree
DNA	5.2 %	5.0 %	4.4 %	6.4 %
Letter	6.7 %	4.3 %	3.7 %	12.7 %
Satellite	10.2 %	8.9 %	8.7 %	14.7 %
Soybean	6.8 %	5.7 %	5.8 %	8.6 %

Table 2. Classification error for benchmark sets.

4 Randomization

The methods presented so far are based on sampling cases to the training samples (except for AdaBoost) but there are other approaches incorporating randomness directly.

The first method is simple and Ho (1998) called it „random subspaces”. Each tree in the ensemble is fitted to the training sample containing all cases from the training set but with randomly selected features.

Random split selection has two variants. Dietterich and Kong (1995) proposed to select the split at random from among the K best splits, while Breiman (1999) proposed to select at random a small group of predictors and then to find the best of them that would form the split.

Recently Breiman (2001) developed a system involving random sampling called Random Forests (RF). It contains two procedures: RV and RC. In the first one at each node K variables are selected at random and the best split is chosen from among them.

In the other procedure K variables are randomly selected and added with coefficients that are uniform random numbers from $[-1, 1]$. Then L linear combinations are generated and the best one is selected for split.

Breiman also found the upper bound for the prediction error of RF:

$$PE(C^*) \leq \bar{\rho} \frac{1 - s^2}{s^2} \quad (14)$$

that is dependent on average correlation between models ($\bar{\rho}$) and their strength (s).

Random forest is equal or better than AdaBoost in classification accuracy, but it is robust to noise, faster and it does not overfit. Some comparisons are given in Table 3.

⁵ We used the CART software and procedures written for the S-PLUS and R environment (e.g. RPart, RandomForest, etc.).

Dataset	RF-RV	RF-RC	Boosting
DNA	4.2 %	4.3 %	4.4 %
Letter	3.5 %	3.4 %	3.7 %
Satellite	8.6 %	9.1 %	8.7 %
Soybean	5.8 %	5.7 %	5.8 %
Zip-code	6.3 %	6.2 %	6.2 %

Table 3. Classification error for Random Forest and AdaBoost.

5 Limiting number of trees in the forest

Now we turn to the problem of limiting the number of combined classifiers. This problem is important for large data sets because averaging classifiers is the most time consuming process. Most researchers, e.g. Breiman (2001), Dietterich (2000), Amit and Blanchard (2001) reported combining from 50 to 200 trees.

Latinne et al. (2001) used McNemar test to find the limit. For 5 datasets (Table 1) they found the limit M to be $M \ll 200$, but in their experiments they grew up to 200 component trees.

We have built committees of trees for the datasets from Table 1 (boosting used) that contain 300 component trees and observed convergence of their error rates:

$$\left| 1 - \frac{PE(\hat{C}_m^*)}{PE(\hat{C}_{m-1}^*)} \right| \leq 0.01 \quad (15)$$

for 10 consecutive steps.

The number of trees in the forest for that the convergence started is given in Table 4. On average the number of trees required for a forest can

Dataset	Number of trees
DNA	76
Letter	110
Satellite	58
Soybean	87
Zip-code	97

Table 4. The limit for the number of trees in the forest.

be estimated as $M \ll 100$. Figure 1 also shows convergence of the prediction error for the Satellite dataset.

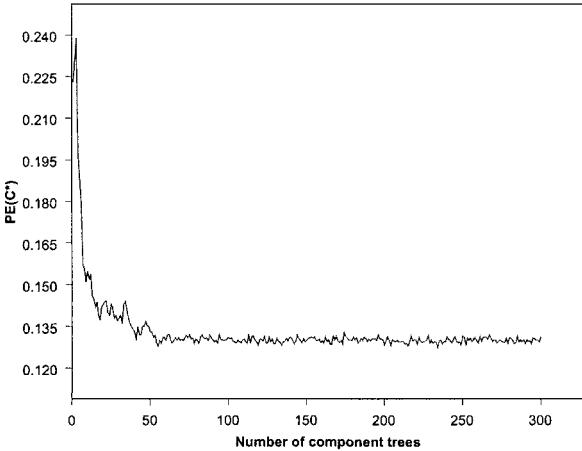


Fig. 1. The prediction error for the Satellite dataset.

6 Consistency

Although random-based methods for building committees of trees are very efficient, the deterministic method (AdaBoost) gives the most accurate ensembles.

But recently some drawbacks of AdaBoost have been reported. First Dietterich (2000) showed that AdaBoost is not robust to noise (incorrect class labels) while bagged trees and random forests are. Then Jiang (2000), and Lugosi and Vyatis (2002) proved that AdaBoost is not consistent. The latter means that its prediction error does not converge to the error of the Bayes optimal model as the number of component trees is increasing:

$$\lim_{m \rightarrow \infty} PE(C_m^*(\mathbf{x})) \rightarrow PE(C^B). \quad (16)$$

In other words: AdaBoost can overfit for large number of component models. Breiman (2002) has an example for an artificial data set where the prediction error starts to increase when the number of trees is greater than 5000.

This overfitting was surprising because of the difference between the proof for the population case where the parameters a_m in (12) are:

$$\sum_m a_m^2 < \infty, \quad (17)$$

and the sample case, when for N examples:

$$a_m \geq \frac{2}{N}. \quad (18)$$

References

- AMIT, Y. and BLANCHARD, 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.
- BREIMAN, L., FRIEDMAN, J., OLSHEN, R., and STONE, C. (1984): *Classification and Regression Trees*, Chapman & Hall/CRC Press, London.
- BREIMAN, L. (1996a): Bagging predictors. *Machine Learning*, 24, 123–140.
- BREIMAN, L. (1996b): Bias, Variance and Arcing Classifiers. *Technical Report*, Statistics Department, University of California, Berkeley.
- 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.
- BREIMAN, L. (2002): Wald Lecture I - Machine Learning. Department of Statistics, University of California, Berkeley.
- CARTER, C. and CATLETT, J. (1987): Assessing Credit Card Applications Using Machine Learning. *IEEE Expert, Fall issue*, 71–79.
- 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 KONG, E. (1995): Machine Learning Bias, Statistical Bias, and Statistical Variance of Decision Tree Algorithms. *Technical Report*, Department of Computer Science, Oregon State University.
- 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.
- FRIEDMAN, J.H. (1996): On Bias, Variance, 0/1-Loss, and The Curse-of-Dimensionality. *Technical Report*, Department of Computer Science, Stanford University.
- FRIEDMAN, J.H. (1999): Stochastic Gradient Boosting. *Technical Report*, Department of Computer Science, Stanford University.
- GATNAR, E. (2002): Tree-based models in statistics: three decades of research. In: K. Jajuga, A. Sokolowski, and H.-H. Bock (Eds.): *Classification, Clustering, and Analysis*. Springer, Berlin, 399–408.
- HASTIE, T. and PREGIBON, D. (1991): Shrinking Trees. *Technical Report*, AT&T Laboratories, Murray Hill.
- HO, T.K. (1998): The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 832–844.
- JIANG, W. (2000): Process Consistency for AdaBoost. *Technical Report 00-05*, Department of Statistics, Northwestern University.
- KOHAVI, R. and WOLPERT, D. (1996): Bias Plus Variance Decomposition for Zero-One Loss Functions. In: L. Saitta (Ed.): *Machine Learning: Proceedings of the XIIIth International Conference*, Morgan Kaufman, 313–321.

- LATINNE, P., DEBEIR, O., and DECAESECKER, Ch. (2001): Limiting the number of trees in random forests, In: J. Kittler and F. Roli (Eds.): *Multiple Classifier System*, LNCS 2096, Springer, Berlin, 178–187.
- LUGOSI, G. and VAYATIS, N. (2002): Statistical Study of Regularized Boosting Methods. *Technical Report*, Department of Economics, Pompeu Fabra University, Barcelona.
- QUINLAN, J.R. (1993): *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo.
- TIBSHIRANI, R. (1996): Regression shrinkage and selection via the lasso. *J.R. Statist. Soc. B*, 58, 267–288.
- TUKEY, J. (1977): *Exploratory Data Analysis*, Addison-Wesly, Reading.
- WOLPERT, D.H. (1992): Stacked Generalization. *Neural Networks*, 5, 241–259.

Data Mining – The Polish Experience

Eugeniusz Gatnar and Dorota Rozmus

Institute of Statistics,
Katowice University of Economics, ul. Bogucicka 14, 40-226 Katowice, Poland

Abstract. Data mining is used to turn data into information and it links several fields including statistics, artificial intelligence, database management, machine learning, pattern recognition and data visualization. It became a field of interest in the early 90s. Since then the interest in data mining has spread worldwide. Also in Poland there are research institutes working in this field, software developers offering specialized software and successful applications of data mining techniques. In this paper we present the development of data mining in Poland.

1 Introduction

The simplest definition says that data mining uses statistical algorithms to discover patterns in data. But, many other definitions can be found in the literature. Friedman (1997) gathered several of them: “Data mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” (Fayyad), “Data Mining is the process of extracting previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions” (Zekulin), “Data Mining is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data” (Ferruzza), “Data mining is the process of discovering advantageous patterns in data” (John).

Most data mining methods are based on concepts from machine learning, pattern recognition and statistics. Their goal is to make predictions or/and to give descriptions. Prediction involves using some attributes to predict unknown values of other features, while description focuses on finding interpretable patterns describing the data.

Data mining uses methods that perform search through the data to find frequently occurring patterns, to detect trends, to produce generalizations about the data, etc. These tools can discover information of various kinds with little guidance from the user.

Although data mining has its origins outside statistics it uses many statistical procedures, for example: classification and regression trees, rule induction, nearest neighbors, clustering methods, feature extraction, data visualization, etc. Gatnar (1997) pointed out that data mining can be seen as the next step beyond exploratory data analysis (EDA).

The academic society and users of data mining software in Poland are familiar with topics associated with neural networks, genetic algorithms, rough

sets, decision trees etc. Perhaps the rough sets theory developed by Pawlak (1992) is the most famous Polish contribution to data mining.

In this paper we would like to present the development of data mining in Poland. We give a short description of research projects that have been done recently in this field in Poland.

Several important Polish developments have been published in international journals and conference proceedings. Especially, in the field of rough sets, e.g. Polkowski and Skowron (2001); Petri nets, e.g. Skowron et al. (2001); evolutionary algorithms, e.g. Michalewicz and Schmidt (2001); action rules, e.g. Ras and Wiczorkowska (2000), etc.

The impact of the Polish contribution can be quantified by number of citations in the NEC citeseer index. For example, we have found 191 citations of the book by Pawlak (1992) and 23 citations of the paper by Polkowski and Skowron (2001).

Also books in Polish on data mining methods have been published recently. For example Gatnar (2001) gave a review of decision tree-based methods, Tadeusiewicz (1995) wrote a comprehensive guide to neural networks, while Witkowska (2000) showed their successful applications in economics, Lasek (2002) presented several applications of data mining techniques in finance, etc.

Several software developers and vendors in Poland are specialized in data mining tools. They sell computer programs both of foreign and Polish origin. We focus on original Polish products.

At last we show some examples of successful use of the data mining methods in practice.

2 Education

In most Polish universities, especially technical universities, there are seminars and courses organised in data mining, neural networks, decision trees, rough sets, evolution algorithms, etc.

Also numerous master theses, PhD theses and habilitation theses are undertaken. Examples are master theses of Reszka (2001) and Mitrowski (2000), and the PhD theses of Stepaniuk (1999) and Szczuka (2000).

It is also worth mentioning the co-operation between Polish academic society and foreign universities in development of data mining tools (e.g. rough sets). Furthermore, universities cooperate with companies developing data mining tools, thanks to which these techniques are used both in science and in practice.

3 Research

Another area we would like to discuss is the scientific research domain. Actually, Polish scientific centers, to name two most important - Institute of

Computer Science and Systems Research Institute of the Polish Academy of Sciences in Warsaw are engaged in developing this area of knowledge.

Activity undertaken in these research centers includes both theoretical background and practical notions. Examples of recently completed projects are: "Knowledge Discovery in Distributed Databases for Intelligent Query Answering" - a project led by prof. Michalewicz in the Institute of Computer Science of Polish Academy of Sciences (1996-1997); "Applications of Rough Sets Theory to Analysis of Conflicts" - a project led by Prof. Witkowska at the Lodz University of Technology (1996-1998), "Application of Artificial Neural Networks in Economic Research" - a project led by Prof. Skowron in the Institute of Computer Science of Polish Academy of Sciences (2000-2001), "Fuzzy Logic in Intelligent Decision Support Systems" - a project led by Prof. Kacprzyk in the Systems Research Institute of Polish Academy of Sciences (1999-2001).

4 Data mining software

As for the practical side of using data mining tools, it can be said that there are numerous companies offering this type of software in Poland, both of Polish and foreign origin.

4.1 Foreign software

Among international firms offering data exploration tools, StatSoft Poland, SPSS Poland, SAS Institute and IBM Poland are worth mentioning.

Statsoft Poland is the only company that sells the software with user interface in Polish. The products offered by the company are: Statistica QC Data Miner PL (for trends exploration, analysing observed schemes and dependencies, forecasting), Neural Networks PL, and Statistica Enterprise (a set of DM tools including five groups of procedures: intersection exploration, general classification, multidimensional exploration with models creator, forecasting and exploration using neural networks).

SPSS Poland sells several products with user interface in English: SPSS Clementine and SPSS Clementine Server (a complex tool for DM needs, which can be applied in business, e-commerce, telecommunication sector, finance, trade and health care system), SPSS Answer Tree and SPSS Answer Tree Server (classification trees), and SPSS Connections (a tool based on neural networks technology, applied in segmentation, building predictive models).

Another famous foreign company offering its products in Poland is the SAS Institute, which sells its solution known as SAS Enterprise Miner, adapted to market segmentation, prediction and marketing analysis based on classification trees and neural networks.

As for IBM Poland, it marketed the following products in Poland: Intelligent Miner for Data 6.1 (which allows quick data extraction of information),

Intelligent Miner for Text (created for data extraction and analysis of text sources such as documents or web sites), and Intelligent Miner Scoring 7.1 (which is characterized by limited modelling function; however, it is ideal for DM appliance with transactional bases).

4.2 Software of Polish origin

Apart from a wide range of foreign products, we learn that Polish companies also have a lot to offer in this area. There are many companies offering solutions either based on originally developed techniques (like GhostMiner), or techniques developed by companies like SAS Institute or Oracle. Table 1 contains all available data mining software developed in Poland. The software details have been taken from from the software documentation and Web pages.

GhostMiner is a result of a close co-operation between FQS Poland consultants and Copernicus University of Torun scientists. This product is mainly applied to solving problems with segmentation and customer profiling, detecting "bad clients" and credit analysis. It is used in banking, telecommunication, distribution, medicine, and technology. Its demonstration version is available for free.

Other products are: Porsenna DSS, Castello OSS and N-Predictor, all of which are based on Oracle tools using decision trees and neural networks.

PORSENNA DSS is mainly applied in the financial sector, and uses neural networks. CASTELLO OSS is designed for telecommunication companies to solve problems connected with customers migration. N-Predictor is a tool used for forecasting and was designed with its own solutions from the area of neural networks.

Another tool is the computer system called NoOne, which is a completely original solution of the Polish company Pentacomp and is applied in business data exploration.

The last two tools: 4eMka and Rose were developed by Predki et al. (1998), the scientists of Poznan University of Technology. Their demonstration versions and manuals are available for free. 4eMka combines advantages of rough sets and dominance relations. It can be used in finance, medicine, geology and pharmacy. Predki and Wilk (1999) describe Rose as modular software system implementing elements of rough sets theory and rule discovery techniques.

5 Applications

Data mining is quickly becoming a necessity, especially for those who must analyse data warehouses containing hundreds of gigabytes or terabytes of information.

Data mining tools offered in Poland found a great variety of users. They have been purchased by: telecommunication firms (deviation detection), fuel

PRODUCT	COMPANY	TECHNIQUES	APPLICATIONS
Ghost Miner	FQS Poland ul. Starowislna 13 31-038 Krakow tel. +48 12 4294345 www.fqspl.com.pl	decision trees neural networks neurofuzzy systems	clients segmentation, customer profiling, detecting bad clients, credit scoring, clinical diagnostics
Porsenna DSS	BMS Software and Consulting ul. Racza 58 32-060 Liszki tel. +48 12 2808036 www.bms.krakow.pl	neural networks, decision trees	building investment portfolios for banks, trust funds, pension funds etc.
Castello OSS	BMS Software and Consulting ul. Racza 58 32-060 Liszki tel. +48 12 2808036 www.bms.krakow.pl	neural networks, decision trees	client service and marketing support for the national telephone operator TPSA
N-Predictor	BMS Software and Consulting ul. Racza 58 32-060 Liszki tel. +48 12 2808036 www.bms.krakow.pl	neural networks	prediction and forecasting
NoOne	Pentacomp ul. Lektykarska 29 01-687 Warszawa tel. +48 22 6393232 www.pentacomp.pl	neural networks	clients segmentation, customer profiling, customer service
4eMka	Laboratory of Intelligent Decision Support Systems, Poznan University of Technology ul. Piotrowo 3A 60-965 Poznan tel. +48 61 6652376 www.idss.cs.put.pl	rough sets, dominance relations	multiple criteria decision support
Rose	Laboratory of Intelligent Decision Support Systems, Poznan University of Technology ul. Piotrowo 3A 60-965 Poznan tel. +48 61 6652376	rough sets, rule discovery	analysis of vast data sets, (with large boundary regions)

Table 1. Data mining software developed in Poland

and energy companies (analysing load patterns), the financial sector (predicting prices at stock exchange), banks (detecting patterns of fraudulent credit card usage), firms from e-commerce sector (identifying buying behaviour patterns), etc.

These tools have also been applied in economic analysis of many kinds (eg. finding associations among demographic characteristics).

Details of some successful applications are presented in Table 2.

AREA	COMPANIES	TASKS
Telecommunication	Polkomtel, P.T.K. Centertel, ERA GSM Mobile Phones	customer migration, customer profiling
Energy and fuel producers	Energy Works Koszalin, Power Station Turow	analysing loading patterns
Financial sector	Polish Stock Exchange, PZU Insurance, Compensa Insurance	predicting share prices, customer profiling, segmentation
Banks	National Bank of Poland, ING BSK Bank, BPH PBK Bank, Bank Pekao SA	detecting patterns of fraudulent credit cards usage
E-commerce sector	Various companies	identifying buying patterns
Medical institutions	Centrum of Child Health in Lodz	identifying successful medical therapies

Table 2. Examples of successful data mining applications in Poland

References

- FRIEDMAN, J. (1997): Data Mining and Statistics: What's the Connection ? *The 29th Symposium on the Interface*, Houston, TX.
- GATNAR, E. (1997): Data Mining and Statistical Data Analysis. *Statistical Revue*, 2, 309–316 (in Polish).
- GATNAR, E. (2001): *The Nonparametric Method of Discrimination and Regression: Decision Trees*. PWN Publishers, Warsaw (in Polish).
- LASEK, M. (2002): *Data Mining: Applications in Bank Clients' Scoring*. Management and Finance Publishers, Warsaw (in Polish).
- MICHALEWICZ, Z. and SCHMIDT, M. (2001): Evolutionary Algorithms. In: *Encyclopedia of Information Systems*. Academic Press, New York.
- MITROWSKI, M. (2000): *Data Exploration Systems*. Master thesis, Warsaw University (in Polish).
- PAWLAK, Z. (1992): *Rough Sets: Theoretical Aspects of Reasoning About Data*. Kluwer, Dordrecht.
- POLKOWSKI, L. and SKOWRON, A. (2001), Rough Set Approach to Computation. *Computational Intelligence* 17, 472–492.

- PREDKI, B., SLOWINSKI, R., STEFANOWSKI, J., SUSMAGA, R., and WILK, S. (1998): ROSE - Software Implementation of the Rough Set Theory. In: L. Polkowski and A. Skowron (Eds.): *Rough Sets and Current Trends in Computing*, Lecture Notes in Artificial Intelligence, vol. 1424, Springer, Berlin, 605–608.
- PREDKI, B. and WILK, S. (1999): Rough Set Based Data Exploration Using ROSE System. In: Z. Ras and A. Skowron (Eds.): *Foundations of Intelligent Systems*, Lecture Notes in Artificial Intelligence, vol. 1609, Springer, Berlin, 172–180.
- RAS, Z. and WIECZORKOWSKA, A. (2000): Mining for action-rules in large decision tables classifying customers. *Advances in Soft Computing*, Physica-Verlag, Heidelberg, 55–64.
- RESZKA, P. (2001): *Knowledge Discovery in Large Datasets*. Master thesis, Warsaw University (in Polish).
- SKOWRON, A., SURAJ, Z., PETERS, J., and RAMANA, S. (2001): Sensor, Filter, and Fusion Models with Rough Petri Nets. *Fundamenta Informaticae* 47, 307–323.
- STEPANIUK, J. (1999): *Knowledge Discovery by Application of Rough Sets Models*. Phd. thesis, Department of Theoretical Foundations of Computer Science, Polish Academy of Sciences (in Polish).
- SZCZUKA, M. (2000): *Symbolic Methods and Neural Networks for Building Classifiers*. Phd. thesis, Warsaw University (in Polish).
- TADEUSIEWICZ, R. (1995): *Neural Networks*. PLJ Publishers, Warsaw (in Polish).
- WITKOWSKA, D. (2000): *Artificial Neural Networks in economic analysis*. Management Publishers, Lodz (in Polish).

Extracting Continuous Relevant Features

Amir Globerson, Gal Chechik, and Naftali Tishby

School of Computer Science and Engineering and
Interdisciplinary Center for Neural Computation,
The Hebrew University,
Jerusalem 91904, Israel

Abstract. The problem of extracting the relevant aspects of data, in face of multiple conflicting structures, is inherent to modeling of complex data. Extracting continuous structures in one random variable that are relevant for another variable has been principally addressed recently via the method of Sufficient dimensionality reduction. However, such auxiliary variables often contain both structures that are relevant and others that are irrelevant for the task in hand. Identifying the relevant structures was shown in the context of clustering to be considerably improved by minimizing the information about another, irrelevant, variable. In this paper we address the problem of extracting continuous relevant structures and derive its formal, as well as algorithmic, solution. Its operation is demonstrated in a synthetic example and in a real world application of face images, showing its superiority over current methods such as oriented principal component analysis.

1 Introduction

A fundamental goal of machine learning is to find regular structures in a given empirical data, and use it to construct predictive or comprehensible models. This general goal, unfortunately, is very ill defined, as many data sets contain alternative, often conflicting, underlying structures. For example, documents may be classified either by subject or by writing style; spoken words can be labeled by their meaning or by the identity of the speaker; proteins can be classified by their structure or function - all are valid alternatives. Which of these alternative structures is “relevant” is often implicit in the problem formulation.

This problem was recently addressed in Chechik and Tishby (2003), by utilizing additional *irrelevant* data, as a source of “side information” that is used to learn the complex structure of the noise in the problem. The goal of the unsupervised learning algorithm is then to find structures which are unique to the empirical data, and do not appear in the side information. However, the particular structure which is used to describe the data is independent of the side information given. For example, in Chechik and Tishby (2003) and Xing et al. (2002) structures are represented in clusters whereas in Mika et al. (2000) and Weinshall (2002) continuous structures are sought.

We have recently introduced an information theoretic notion of continuous structure extraction in contingency tables, which we called Sufficient Dimensionality Reduction (Globerson and Tishby (2003)). This method aims to

find approximate sufficient statistics for a variable X whose expected value provides maximal information about some relevance variable Y . In this work we introduce a variation of this method which takes advantage of side information and thus finds statistics which are maximally informative about the empirical data "and" minimally informative about the side information. It is thus an information theoretic formulation of continuous structure discovery in co-occurrence data.

The search for features (i.e. statistics of empirical data) that carry no information about a parameter, was introduced to statistics through the notion of ancillary statistic (Fisher (1922)). Ancillary statistics are defined as statistics the the parameters are independent of, and are mainly used for estimating the precision of the parameters estimates, rather than their values. *sufficient dimensionality reduction with side information* (SDR-SI) extracts features that are sufficient about Y^+ and in the same time ancillary about Y^- but in a soft manner: the statistics embody a trade-off between carrying information about Y^+ and about Y^- .

2 Problem formulation

To formalize the above ideas consider a scenario where two empirical joint distributions are given for three categorical random variables X , Y^+ and Y^- . The first is the main data, $P(X, Y^+)$ (or p^+), which describes the joint distribution of Y^+ and X . The second is the side data, $P(X, Y^-)$ (or p^-), which is assumed to contain irrelevant structures in the main data. Our goal is to identify features of X that characterize its probabilistic relation to Y^+ but not its relation to Y^- .

We seek a d dimensional continuous feature of X which we denote $\phi(x) : X \rightarrow \mathbb{R}^d$, so that only its expected values $\langle \phi(x) \rangle_{p(x|y^+)}$ characterize the stochastic dependency between X and Y^+ . For example, each document in a corpus of documents may be characterized by the average number of words that belong to a small semantic set (or few sets), each image in a database may be characterized by the average grey level at specific locations, etc.

To evaluate the "goodness" of $\phi(x)$, we use the notion of *measurement information* $I_M[\phi(x), p]$, defined in Globerson and Tishby (2003) and reviewed in the next section. Given this measure for the quality of $\phi(x)$, the goal of relevant feature extraction is to identify features that are maximally informative about Y^+ while minimally informative about Y^- . This dual optimization task can be solved by maximizing the information about Y^+ while constraining the information about Y^- . Formally, the goal of SDR-SI is to find $\phi^*(x)$ such that

$$\phi^*(x) = \arg \max_{\phi(x) : I_M[\phi(x), p^-] < D} I_M[\phi(x), p^+] . \quad (1)$$

with D being a positive constant. This is equivalent to maximizing the variational functional

$$\mathcal{L}[\phi(x)] = I_M[\phi(x), p^+] - \lambda I_M[\phi(x), p^-] \quad (2)$$

over $\phi(x)$, where λ is a Lagrange multiplier corresponding to the value of the constraint D .

3 Information in measurements

Given a joint distribution $p(X, Y)$, we define the *measurement* of $\phi(x)$ as the set of expected values $\langle \phi(x) \rangle_{p(x|y)}$. In order to quantify the information conveyed by such measurements *alone* about Y , we look for the distribution which has the same measurement values as $p(X, Y)$ but contains minimal mutual information between X and Y . This information effectively extracts the dependence between X and Y which can be attributed to knowledge $\phi(x)$ and its expectations.

Formally, denote the set of these distributions by

$$\mathcal{P}(\phi(x), p) \equiv \left\{ \hat{p}(x, y) : \begin{array}{l} \langle \phi(x) \rangle_{\hat{p}(x|y)} = \langle \phi(x) \rangle_{p(x|y)} \\ \hat{p}(x) = p(x) \\ \hat{p}(y) = p(y) \end{array} \right\}. \quad (3)$$

Note that we have added the marginal constraints to this set, since we want to use the relative weights of X and Y in our information measure.

We define the information in the measurement of $\phi(x)$ as

$$I_M[\phi(x), p] \equiv \min_{\hat{p}(x,y) \in \mathcal{P}(\phi(x)), p} I[\hat{p}] \quad (4)$$

where $I[p]$ is the Shannon mutual information of the two variables X and Y with joint distribution $p(X, Y)$ (Shanon (1948), Cover and Thomas (1991))¹

$$I[p(X, Y)] \equiv \sum_{x,y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (5)$$

The variational problem in Equation 2 thus becomes:

$$\begin{aligned} \phi^*(x) &= \arg \max \mathcal{L}[\phi(x)] \\ &= \arg \max_{\phi(x)} \min_{\hat{p}^+ \in \mathcal{P}(\phi, p^+)} I[\hat{p}^+] - \lambda \min_{\hat{p}^- \in \mathcal{P}(\phi, p^-)} I[\hat{p}^-] \end{aligned} \quad (6)$$

¹ We use here the notation $I[p]$ instead of the more common notation $I(X; Y)$ to emphasize that I is a functional of the distribution p .

4 Solution characterization

In order to characterize the solution of the variational problem in Equation 6, we now calculate its gradient and observe its vanishing points.

We start by characterizing the form of distribution $\hat{p}_\phi(X, Y)$ that achieves the minimum of $I_M[\phi(x), p]$ (Equation 4). Since $I[\hat{p}(X, Y)] = H[\hat{p}(X)] + H[\hat{p}(Y)] - H[\hat{p}(X, Y)]$, and the marginals $\hat{p}(x)$, $\hat{p}(y)$ are kept constant by the definition of $\mathcal{P}(\phi(x), p)$, we have $I[\hat{p}(X, Y)] = \text{const} - H[\hat{p}(X, Y)]$. This turns Equation 4 into a problem of entropy maximization under linear constraints

$$\hat{p}_\phi(x, y) = \max_{\hat{p}(x, y) \in \mathcal{P}(\phi(x), p)} H[\hat{p}(x, y)] , \quad (7)$$

whose solutions are known to be of exponential form (Della Pietra et al. (1997))

$$\hat{p}_\phi(x, y) = \frac{1}{Z} \exp(\phi(x) \cdot \psi_\phi(y) + A_\phi(x) + B_\phi(y)) . \quad (8)$$

The $\psi_\phi(y)$, $A_\phi(x)$ and $B_\phi(y)$ are complex functions of $\phi(x)$ that play the role of Lagrange multipliers in the maximum entropy problem derived from Equation 7. While $H[\hat{p}_\phi(x, y)]$ is a complex function of $\phi(x)$, its gradient can be derived analytically using the fact that \hat{p}_ϕ has the exponential form of Equation 8. Appendix A shows that this gradient is

$$\frac{\partial H[\hat{p}_\phi(x, y)]}{\partial \phi(x)} = p(x) (\langle \psi_\phi \rangle_{\hat{p}_\phi(y|x)} - \langle \psi_\phi \rangle_{p(y|x)}) \quad (9)$$

It is now straightforward to calculate the gradient of the functional in Equation 6. Denote by \hat{p}_ϕ^+ and \hat{p}_ϕ^- the information minimizing distribution obtained in $I_M[\phi, p^+]$ and $I_M[\phi, p^-]$, and by ψ_ϕ^+ and ψ_ϕ^- their corresponding Lagrange multipliers. The gradient is then

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \phi(x)} &= p^+(x) \left(\langle \psi_\phi^+ \rangle_{p^+(y^+|x)} - \langle \psi_\phi^+ \rangle_{\hat{p}_\phi^+(y^+|x)} \right) \\ &\quad - \lambda p^-(x) \left(\langle \psi_\phi^- \rangle_{p^-(y^-|x)} - \langle \psi_\phi^- \rangle_{\hat{p}_\phi^-(y^-|x)} \right) \end{aligned} \quad (10)$$

Setting it to zero we obtain the characterization of the extremum point

$$p^+(x) \Delta \langle \psi_\phi^+ \rangle = \lambda p^-(x) \Delta \langle \psi_\phi^- \rangle \quad (11)$$

where $\Delta \langle \psi \rangle$ is the difference in the expectation of ψ taken according to the model and the true distribution.

To obtain some intuition into the last equation consider the following two observations. First, note that maximizing the information $I_M[\phi, p^+]$ requires to minimize the difference between the expectancies of ψ_ϕ^+ , as can be seen when taking $\lambda = 0$. Second, it can be shown than when minimizing $I_M[\phi, p^-]$ some elements of $\phi(x)$ diverge. In these infimum points, $\Delta \langle \psi_\phi^- \rangle$ generally

does not vanish. Taken together, these facts imply that for the $\lambda > 0$ case, the difference $\Delta\langle\psi_\phi^+\rangle$ should generally diverge from zero. This implies, as expected, that the resulting $\phi(x)$ conveys less information than the $\lambda = 0$ solution. The optimal $\phi(x)$ is thus bound to provide an inaccurate model for those aspects of p^+ that also improve the model of p^- .

An additional interesting interpretation of ψ_ϕ^+, ψ_ϕ^- is that they reflect the relative importance of $\phi(x)$ in $\hat{p}_\phi^+, \hat{p}_\phi^-$. This view is prevalent in the boosting literature, where such coefficients function as the weights of the weak learners (see e.g. Lebanon and Lafferty (2002)). When $\lambda = 0$, Equation 11 then requires that \hat{p}_ϕ^+ agrees with p^+ on the expected importance of the features $\phi(x)$. As λ grows this agreement is broken, and the weighting of the weak learners is tilted by the side information.

5 Algorithmic considerations

Unlike the case of $\lambda = 0$ for which an iterative algorithm was described in Globerson and Tishby (2003), the $\lambda > 0$ case poses a special difficulty in developing such an algorithm. One could supposedly proceed by calculating ψ_ϕ^+ , ψ_ϕ^- assuming a constant value of $\phi(x)$ and then calculate the resulting $\phi(x)$ assuming ψ^+ and ψ^- are constant. However, as was shown in Globerson and Tishby (2003) updating ψ_ϕ^- will increase $I_M(\phi(x), p^-)$ thereby decreasing the target function. Thus, such a procedure is not guaranteed to improve the target function. Possibly, an algorithm guaranteed to converge for a limited range of λ values can be devised, as done for IBSI (Chechik and Tishby (2003)), but this remains to be studied.

Fortunately, the analytic characterization of the gradient derived above, allows one to use a gradient ascent algorithm for finding the optimal features $\phi(x)$, for any given value of λ . This requires to calculate a Maximum Entropy distribution on each of its iterations, namely, to calculate numerically the set of Lagrange multipliers $\psi_\phi(y)$, $A_\phi(x)$ and $B_\phi(y)$ which appear in the gradient expression in Equation 9. This convex problem has a single maximum, and well studied algorithms exist for finding Maximum Entropy distributions under linear constraints². These include GIS (Darroch and Ratcliff (1972)), IIS (Della Pietra et al. ((1997)), or gradient based algorithms (see Malouf (2002) for a review of different algorithms and their relative efficiency). In all the results described below we used the GIS algorithm.

6 Relation to other methods

6.1 Sufficient and ancillary statistics

Fisher introduced the notion of a statistic that is sufficient for the estimation of a parameter θ from a sample X (Fisher 1922). This was defined as a

² Note that all the constraints in $\mathcal{P}(\phi(x), p)$ are indeed linear.

statistic $S(X)$ of the sample, that makes θ independent of the sample given the statistic: $p(\theta|X, S(X)) = p(\theta|S(X))$. In information theoretic terms, collapsing a sample to a low dimension statistic can only reduce the information conveyed about the parameter, $I(\theta; S(X)) \leq I(\theta; X)$. However, a sufficient statistic is guaranteed to preserve *all* the information about the parameter $I(\theta; S(X)) = I(\theta; X)$.

Conversely, a statistic $A(X)$ that conveys *no* information about a parameter (i.e. $I(\theta; A(X)) = 0$) is termed *ancillary*. Such statistics usually serve not for estimating of the value of the parameter, but rather to assess the reliability of a parameter's estimate. A known example is the sample variance, which is independent of the expectancy of a normal variable with known variance, but can be used to estimate the variance of the sample mean.

In this statistical context, SDR-SI can be understood as a search for a statistic of X that is simultaneously sufficient about Y^+ and ancillary about Y^- . Since simultaneous sufficiency and ancillarity cannot be usually achieved, SDR-SI looks for soft sufficiency and ancillarity by maximizing an minimizing the corresponding information terms with the tradeoff parameter λ .

6.2 Likelihood ratio maximization

Further intuition into the functional of Equation 6 can be obtained, using the result of Globerson and Tishby (2003) yielding that it equals up to a constant to

$$\mathcal{L} = -D_{KL}[p^+||\hat{p}_\phi^+] + \lambda D_{KL}[p^-||\hat{p}_\phi^-] . \quad (12)$$

where $D_{KL}[p||q] \equiv \sum p_i \log(p_i/q_i)$ is the Kullback-Leibler divergence. When p^+ and p^- share the same marginal distribution $p(x)$, a joint distribution $p(X, Y^+, Y^-)$ can be defined that coincide with the pairs-joint distributions $p^+(X, Y^+)$ and $p(X, Y^-)$. In this case

$$\begin{aligned} \mathcal{L} &= - \sum_{x,y^+,y^-} p(x, y^+, y^-) \log \left(\frac{p^+(x, y^+)}{\hat{p}_\phi^+(x, y^+)} \right) \\ &\quad + \lambda \sum_{x,y^+,y^-} p(x, y^+, y^-) \log \left(\frac{p^-(x, y^-)}{\hat{p}_\phi^-(x, y^-)} \right) \\ &= - \left\langle \log \left(\frac{p^+(x, y^+)}{p^-(x, y^-)^\lambda} \frac{\hat{p}_\phi^-(x, y^-)^\lambda}{\hat{p}_\phi^+(x, y^+)} \right) \right\rangle_{p(x, y^+, y^-)} \\ &= \left\langle \log \left(\frac{\hat{p}_\phi^+(x, y^+)}{\hat{p}_\phi^-(x, y^-)^\lambda} \right) \right\rangle_{p(x, y^+, y^-)} + const \end{aligned} \quad (13)$$

This suggests that in the special case of $\lambda \equiv 1$, SDR-SI operates to maximize the expected log likelihood ratio, between the maximum entropy models \hat{p}_ϕ^+ and \hat{p}_ϕ^- . In the general case of $\lambda > 0$ a weighted log likelihood ratio is obtained. For vanishing λ , the side information is completely ignored and the problem reduces to unconstrained likelihood maximization of the maximum entropy model \hat{p}_ϕ^+ .

6.3 Related methods

The mutual information about a side variable was first presented in the context of lossy compression by Wyner and Ziv (1976). They showed that a lower bound on communication rate for constrained distortion can be achieved when both the encoder and the decoder have access to a side variable W . Basically, this can be achieved by removing some aspects of the source X during compression, that can be retrieved from the knowledge of W . The lower bound on the rate distortion with side information is $I(X; Y) - I(X; W)$. This result motivated Chechik and Tishby (2003), in which clustering is used to compress a source while preserving information about a relevance variable and removing information about a side (irrelevant) variable.

The term “side information” is also used in the more general context, where auxiliary data or additional sources of information are used to enhance learning features of a main data. This idea was mainly studied in the context of spectral analysis. The method of *Oriented Principal COmponent Analysis* (OPCA) (Diamantaras, Kung (1996)) uses a main data set with covariance S_+ and a side data set with covariance S_- to find directions w such that the Signal to Noise Ratio

$$\frac{w^T S_+ w}{w^T S_- w} \quad (14)$$

is maximized. A kernelized version of OPCA was described in Mika et al. (2000). A different method, Relevant Component Analysis (RCA), applies the sphering transformation of the side data to the original data. Note that such methods implicitly assume that both data sets distributed as Gaussians (in the feature space in Mika et al. (2000), in that they only treat second order statistics).

In Xing et al. (2003) the clustering algorithm was given several examples belonging to the same cluster, which it used to learn a distance metric for clustering.

SDR-SI differs from the above methods in that it is a non-linear method for extracting continuous features, which are least informative about the side data. It also introduces a weighting factor λ , which allows one to determine the relative importance of the side data in the feature extraction process.

7 Extensions to multiple variables

The SDR-SI formalism can be extended to the case of multiple relevance and irrelevance variables ($Y^+_{-1}, \dots, Y^+_{-n_+}$) and ($Y^-_{-1}, \dots, Y^-_{-n_-}$), with joint distributions $p_i^+ = p(X, Y_i^+)$ and $p_i^- = p(X, Y_i^-)$. Following a similar constrained optimization problem we write the Lagrange form of the functional

$$\mathcal{L} = \sum_{i=1}^{n^+} \lambda_i I_M[\phi(x), p_i^+] - \sum_{i=1}^{n^-} \lambda_i I_M[\phi(x), p_i^-], \quad (15)$$

which can be maximized as in the two variables case.

8 Applications

We first illustrate the operation of SDR-SI to a synthetic example that demonstrates its main properties. Then, we describe the application of our method to the real life problem of feature extraction for face recognition.

8.1 A synthetic example

To demonstrate the ability of our approach to uncover weak but interesting hidden structures in data, we designed a co-occurrence matrix that contains two competing sub-structures (see figure 1A). The right half of the matrix contains a top-to-bottom gradient, while its left half contains large variance at the middle values of X . The right structure was handcrafted to be stronger in magnitude than the left one.

Projecting X into a one dimensional $\phi(x)$ while preserving information on Y^+ using SDR ($\lambda = 0$), yields a top-to-bottom gradient (Figure 1B). This $\phi(x)$ follows from the strong structure on the right part of 1A.

We now created a second co-occurrence matrix $P(X, Y^-)$ that contains a top-to-bottom structure similar to that of $P(X, Y^+)$. Applying SDR-SI with $\lambda = 1$ on both matrices now successfully ignores the strong top-to-bottom structure in $P(X, Y^+)$ and retrieves the weaker structure that emphasizes the mid values of X (Figure 1D). Importantly, this is done in an unsupervised manner, without explicitly pointing to the strong but irrelevant structure.

Further understanding into the operation of SDR-SI is gained by tracing its output as a function of the tradeoff parameter λ . Figure 2A plots the optimal features $\phi(x)$ extracted for various λ values, revealing a phase transition around a critical value $\lambda = 0.26$. The reason for this behavior is that at the critical λ , the top-to-bottom feature $\phi(x)$ bears larger loss (due to the information $I_M[\phi(x), p^-]$ conveyed on $p(X; Y^-)$) than gain. Figure 2B traces the values of $I_M[\phi(x), p^+]$ and $I_M[\phi(x), p^-]$, again revealing a pronounced phase transition. This discontinuity reflects the removal of “irrelevant” features from $p(X; Y^+)$, and can thus be used to select interesting values of λ .

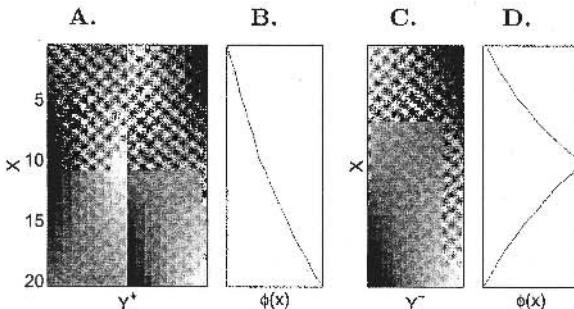


Fig. 1. Demonstration of SDR-SI operation. **A.** A joint distribution $P(X, Y^+)$ that contains two distinct and conflicting structure: one that involves a top-to-bottom gradient, and another that involves a large variance in probability mass for mid values X values. **B.** Extracting a one-dimensional projection identifies the top-to-bottom gradient. **C.** A joint distribution $P(X, Y^-)$ that contains a single structure, similar in nature to the top-to-bottom gradient, the stronger structure of $P(X, Y^+)$. **D.** Extracting a one-dimensional projection using $\lambda = 1$ successfully ignores the top-to-bottom gradient and successfully extracts the weaker structure in $P(X, Y^+)$.

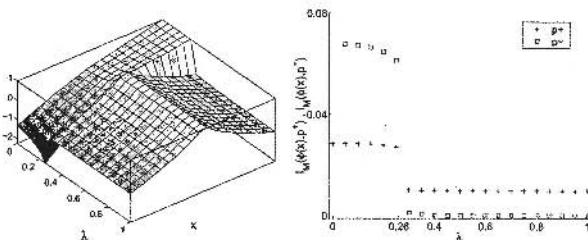


Fig. 2. Operation of SDR-SI on the synthetic example of Figure 1 for various values of λ . **A.** The optimal $\phi(x)$ extracted with SDR-SI. **B.** The information conveyed about Y^+ (crosses) and Y^- (squares), by the optimal $\phi(x)$'s of the left panel. A phase transition around 0.26 is observed both in the information values and the $\phi(x)$'s.

This example was designed for demonstration purposes, thus the irrelevant structures is strongly and cleanly manifested in $P(X; Y^-)$. The next section studies the application of our approach to real data, in which structures are much more covert.

8.2 Face images

The identification of relevant features is important in particular in the field of face recognition, where features are sought to be invariant to various interfering structures, such as face expression, light conditions, or even occlusions due to glasses or scarves. Such nuisance structures are often more pronounced

in the data than the subtle features that are required for face recognition. In this context, given a set of unlabeled face images, side data can be naturally defined by using additional images of other persons or objects under similar conditions.

We tested SDR-SI on the AR database (Martinez and Benavente (1998)), a collection of faces with various face expression and illumination conditions. Each picture was translated into a joint probability matrix, by considering the normalized grey levels of the pixel x in the image y as the probability $p(x|y)$, and setting $p(y)$ uniform.

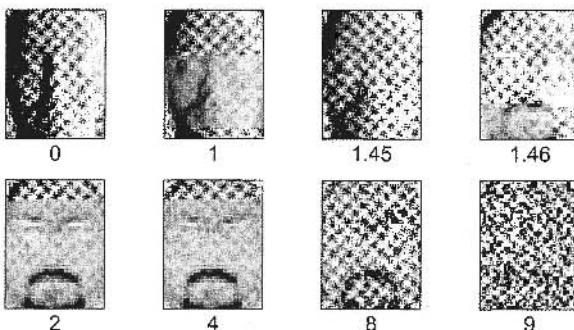


Fig. 3. Extracting a single feature using SDR-SI, as a function of λ . An apparent phase transition is observed around $\lambda = 1.45$. p^+ was created by taking pictures of all men in the database with light either from the right or the left (total of 100 images). p^- was similarly created with 100 female pictures.

To demonstrate the operation of SDR-SI on this data we first trained it to extract a single feature, for various λ values (see details of experiment in caption of figure 3). The obtained $\phi(x)$ are shown in Figure 3. When λ is low (small weight for side information) the main structure captured is the direction of light source (right vs. left). As λ increases the optimal $\phi(x)$ first changes only slightly, but then a phase transition occurs around $\lambda = 1.45$, and a second structure emerges.

This emerging structure should contain features that are highly discriminative between male faces but not between female ones. Indeed, the marked beard that is easily observed characterizes some of the male faces, but is not a good female face discriminator. Similarly, some structures observed around the area of the eyes (somewhat resembling a pair of glasses) reflect changes in the position of the eyes that happened to be more discriminative for faces of men.

The transition at $\lambda = 1.45$ can be well observed when tracing the values of $I_M[\phi(x), p^+]$ and $I_M[\phi(x), p^-]$ as a function of λ (Figure 4), and results from the same reasons discussed in the synthetic example described above.

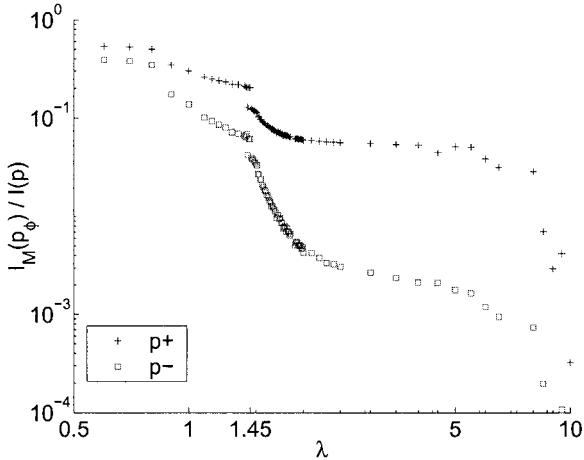


Fig. 4. Normalized information about the main data $I_M[\phi(x), p^+]$ and the side data $I_M[\phi(x), p^-]$, as a function of λ , for the data of Figure 3. Note the phase transition in both information levels for $\lambda = 1.45$.

This result suggest that such information curves can be used to identify “interesting” values of λ for high dimensional complex data.

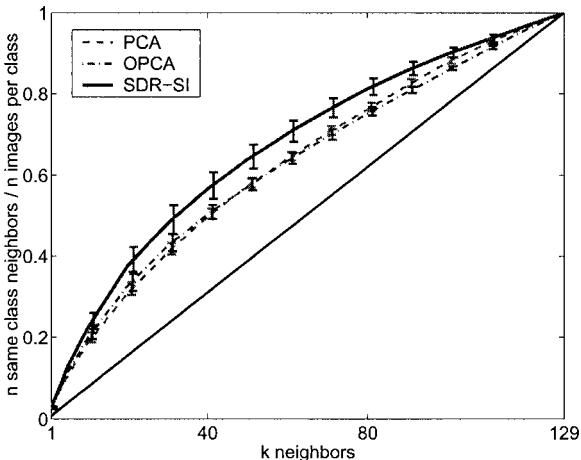


Fig. 5. Mean number of same-class neighbors as a function of neighborhood size for the AR data experiments. The bars denote the standard error of the mean. Averaging was performed over ten cross validation sets. Black line denotes a lower bound on performance, as obtained with random ordering of neighbors.

To test the performance of SDR-SI in a quantitative manner, we used it in a difficult task of face recognition, and compared its performance with

PCA - the most widely used dimensionality reduction method, and OPCA - a method that utilizes the same side data as SDR-SI (Diamantaras and Kung (1996)). To this end, we created a $p^+(X, Y^+)$ with images of five different men, under all the different conditions of face expression and light conditions (a total of 26 images per person). As side data we used all 26 images of another randomly chosen man. The task of clustering these pictures into the five correct sets is hard since the nuisance structures are far more dominant than the relevant structure of inter subject variability.

All methods (PCA,OPCA and SDR-SI) were used for calculating a dimensionality reduction transformation on the images. PCA and OPCA representation were obtained by projecting on the principal components. The low dimensional representation in SDR-SI was obtained by replacing each image y with its expected SDR-SI feature values $\langle \phi(x) \rangle_{p(x|y)}$.

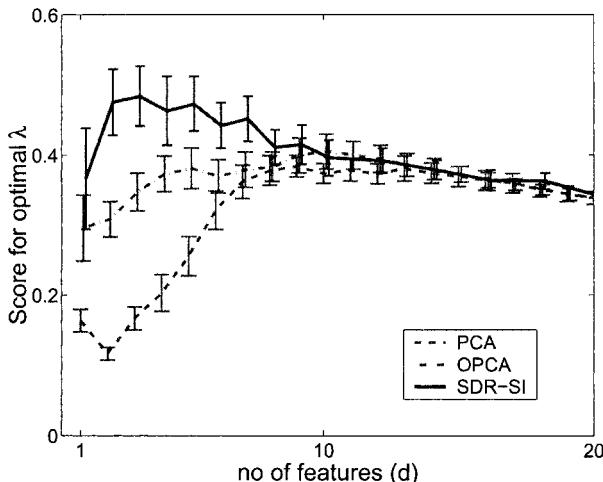


Fig. 6. Performance of SDR-SI compared with PCA and OPCA, as a function of dimensionality d for the AR data experiments. The mean performance over 5 testing sets is reported, and bars denote standard error of the mean over these sets. In SDR-SI, a value of λ was chosen separately for each d , such that it maximizes performance over a (separate) training set. For small number of features SDR-SI is superior to both PCA and OPCA, but their performance coincides for $d > 10$.

To quantify the effectiveness of the reduced representations in preserving person identity, we calculated the number of same-class (same-person) neighbors out of the k nearest neighbors of each image. This was done for all possible values of k and averaged over all images, and yielded *neighborhood curves* as plot in Figure 5³. As a metric for measuring distances between

³ We also evaluated the method by clustering the low dimensional vectors into five groups and comparing the resulting clusters with the true ones. This resulted in

images, we tested both the L2 norm and the Mahalanobis distance in the reduced representation. We report the Mahalanobis results only, since L2 results were considerably worse for PCA. Optimal parameters (dimensionality and λ) for all methods, were chosen to optimize the area under the neighborhood curves, for a training set data. Reported Results were obtained on a separate testing set. This entire procedure was repeated for 10 times on randomly chosen subsets of the database. The resulting average curve is shown in (Figure 5), and SDR-SI clearly out performs all other methods for all k values.

We further compared the performance of the three methods for each pre-defined dimensionality d , to see how performance depends on the number of features used. Figure 6 plots the performance of SDR-SI compared with that of PCA and OPCA, as a function of the dimensionality, showing that SDR-SI dominates PCA and OPCA over all d values. This is more pronounced for lower values of d , which agrees with the intuition that the side-data allows SDR-SI to focus on the more relevant features.

9 Conclusions

We have presented a method for extracting continuous features from co-occurrence data, using side information in the form of additional, irrelevant data, that serves to focus on the relevant features of the main data.

Appendix: Deriving the gradient of the joint entropy

To calculate the gradient of the entropy $H[\hat{p}_\phi(x, y)]$, we first prove some useful properties of the distribution \hat{p}_ϕ . Since \hat{p}_ϕ is in $\mathcal{P}(\phi(x), p)$, it satisfies the margin constraints: $\hat{p}_\phi(x) = \sum_{y'} \hat{p}_\phi(x, y') = p(x)$, $\hat{p}_\phi(y) = \sum_{x'} \hat{p}_\phi(x', y) = p(y)$, as well as the expectation constraints

$$\sum_{x'} \phi(x') (\hat{p}_\phi(x', y) - p(x', y)) = 0 . \quad (16)$$

Deriving the three constraints equations w.r.t. $\phi(x)$ yields

$$\sum_{y'} \frac{\partial \hat{p}_\phi(x, y')}{\partial \phi(x)} = 0; \quad \sum_{x'} \frac{\partial \hat{p}_\phi(x', y)}{\partial \phi(x)} = 0 \quad (17)$$

for the margin constrains, and

$$\hat{p}_\phi(x, y) - p(x, y) + \sum_{x'} \phi(x') \frac{\partial \hat{p}_\phi(x', y)}{\partial \phi(x)} = 0 \quad (18)$$

qualitatively similar result, albeit noisier. We prefer the method presented here since it does not depend on a noisy second phase of clustering.

for the expectation constraints.

The derivative of the entropy can now be written as

$$\begin{aligned}\frac{\partial H[\hat{p}_\phi]}{\partial \phi(x)} &= \sum_{x',y'} \frac{\partial \hat{p}_\phi(x',y')}{\partial \phi(x)} \\ &\quad - \sum_{x',y'} \frac{\partial \hat{p}_\phi(x',y')}{\partial \phi(x)} \log \hat{p}_\phi(x',y') \\ &= \sum_{x',y'} \frac{\partial \hat{p}_\phi(x',y')}{\partial \phi(x)} \log \hat{p}_\phi(x',y')\end{aligned}\tag{19}$$

where the last equality stems from the vanishing derivative of the margin constraints in Equation 17. Plugging in the exponential form of \hat{p}_ϕ from Equation 8, and using Equation 17 again, we have

$$\frac{\partial H[\hat{p}_\phi]}{\partial \phi(x)} = - \sum_{x',y'} \frac{\partial \hat{p}_\phi(x',y')}{\partial \phi(x)} \phi(x') \cdot \psi_\phi(y')\tag{20}$$

Now using Equation 18 for the derivative of the expectation constraints, we have

$$\frac{\partial H[\hat{p}_\phi]}{\partial \phi(x)} = \sum_{y'} \psi_\phi(y') (\hat{p}_\phi(x,y') - p(x,y'))\tag{21}$$

Taking out the margin $p(x)$ we finally obtain

$$\frac{\partial H[\hat{p}_\phi]}{\partial \phi(x)} = p(x) (\langle \psi_\phi \rangle_{\hat{p}_\phi(y|x)} - \langle \psi_\phi \rangle_{p(y|x)})\tag{22}$$

Acknowledgements: A.G. and G.C. are supported by the Israeli ministry of Science, the Eshkol fund.

References

- CHECHIK, G. and TISHBY, N. (2003): Extracting relevant structures with side information. In: S. Becker, S. Thrun, and K. Obermayer (Eds.): *Advances in Neural Information Processing Systems 15*. MIT press, Cambridge, MA.
- COVER, T.M. and THOMAS, J.A. (1991): *The elements of information theory*. Plenum Press, New York.
- DARROCH, J.N. and RATCLIFF, D. (1972): Generalized iterative scaling for log-linear models. *Ann. Math. Statist.*, 43, 1470–1480.
- DELLA PIETRA, S., DELLA PIETRA, V., and LAFFERTY, J.D. (1997): Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4), 380–393.

- DIAMANTARAS, K.I. and KUNG, S.Y. (1996): *Principal Component Neural Networks: Theory and Applications*. John Wiley, New York.
- FISHER, R.A. (1922): On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A*, 222, 309–368.
- GLOBERSON, A. and TISHBY, N. (2003): Sufficient dimensionality reduction. *Journal of Machine Learning Research*, 3, 1307–1331.
- LEBANON, G. and LAFERRTY, J. (2002): Boosting and maximum likelihood for exponential models. In: T.G. Dietterich, S. Becker, and Z. Ghahramani (Eds.): *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, MA.
- MALOUF, R. (2002): A comparison of algorithms for maximum entropy parameter estimation. In: *Sixth Conf. on Natural Language Learning*, 49–55.
- MARTINEZ, A.M. and BENAVENTE, R. (1998): The AR face data base. Technical Report 24, Computer vision Center.
- MIKA, S., Ratsch, G., WESTON, J., SCHOLKOPF, B., SMOLA, A., and MULLER, K. (2000): Invariant feature extraction and classification in kernel space. In: S.A. Solla, T.K. Leen, and K.R. Muller (Eds.): *Advances in Neural Information Processing Systems 12*. MIT Press, Cambridge, MA, 526–532.
- WEINSHALL, D., SHENTAL, N., HERTZ, T., and PAVEL, M. (2002): Adjustment learning and relevant component analysis. In: *7th European Conference of Computer Vision (ECCV 2002), Volume IV*, Lecture Notes on Computer Sciences, 776–792.
- SHANON, C.E. (1948): A mathematical theory of communication. *The Bell systems technical journal*, 27, 379–423, 623–656.
- WYNER, A.D. and ZIV, J. (1976): The rate distortion function for source coding with side information at the decoder. *IEEE Trans. Inform. Theory*, 22(1), 1–10.
- XING, E.P., NG, A.Y., Jordan, M.I., and RUSSELL, S. (2003): Distance metric learning, with applications to clustering with side information. In: S. Becker, S. Thrun, and K. Obermayer (Eds.): *Advances in Neural Information Processing Systems 15*. MIT Press, Cambridge, MA.

Individual Rationality Versus Group Rationality in Statistical Modelling Issues

Daniel Kosiorowski

Department of Statistics,
Cracow University of Economics, ul. Rakowicka 27, 31-510 Kraków, Poland

Abstract. In the paper, an apparatus of stochastic matrix spaces and discrete Markov chains is presented for the purpose of coping with approximating individual choice mechanisms similar to a network of associations. The proposed apparatus makes it possible to investigate dynamical properties of the aggregated choice of a group when interactions exist between individual choice mechanisms within the group. The main point of the paper is to investigate the dynamic aspect of aggregated group choice against interactions between individual choices in terms of group rationality criteria. This investigation leads to the analysis of stability of aggregated group choice and of stationary processes of group choice. An empirical verification and perspectives of further research and applications are proposed.

1 Introduction

Although our knowledge of behaviour laws governing individuals and groups is still poor, today's social sciences offer some terms that could be a good basis for investigating systems consisting of people. One of such terms is *choice* allowing both humans and organizations to be considered as system components. We are able to classify choices by using rationality considered as consistent desire to achieve a goal. The choices made by people are discussed in the first part of this paper. A choice mechanism governing an individual is described as a situation algorithm that depends on time and a group containing the individual. Conditions determining a particular choice include uncertainty. The second part of this paper deals with numerous subjects while focusing on interactions between individual connection grids. Interactions are defined here three types of operation on connection grids. Further an effect of individual's choice on the choice of the group is also analyzed. For aggregated choice made by entire group stability categories are introduced. In this paper the stability categories are used not only to describe the choice process for individuals and groups, but also to forecast a set of most probable choices, interaction types and group structure changes, while considering an effect of the environment. Such an approach is justified by referring to the results of evolutionary psychology that underlines the predominance of some individual behaviour standards for an individual belonging to a group (Buss (2001)). The results of empirical research studies are presented in the next part of

this paper. Finally, the most important results are summarized and further research tasks and possible applications areas are presented.

2 Choice mechanism - connection grid

Let's, for instance, consider a fixed set of choice alternatives: - let's assume four alternatives $V_1 V_2 V_3 V_4$. To introduce a **choice mechanism**, we assume that an individual is able to declare: "If I have chosen alternative V_1 at time t, then I choose V_1 once again at time $t+1$ with probability $p_{V_{11}}$, alternative V_2 with probability $p_{V_{21}}$, alternative V_3 with probability $p_{V_{31}}$, and alternative V_4 with probability $p_{V_{41}}$. Similarly, an individual is able to specify the probability of choosing alternative V_1, \dots, V_4 , when choosing V_2, V_3 or V_4 at time t. A choice mechanism can be presented in the form of a graph or a corresponding matrix:

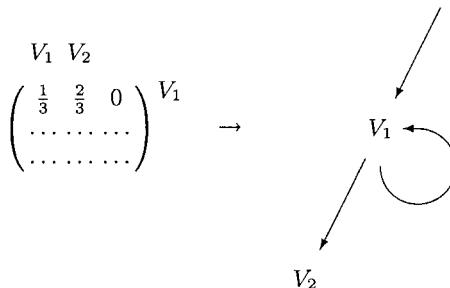


Fig. 1. A graph representing choice mechanism and corresponding matrix

Note: When defining a choice mechanism for humans in the way described above, one can use similarly a homogeneous Markov chain (Feller (1978)).

3 Interaction between subjects

Let's consider two algebras - a ring of square matrices of real numbers, and a linear space of square matrices of elements belonging to a finite field Z_k over this field. Among matrices, we focus our attention on stochastic matrices, i.e. square matrices for which the sum of entries of each row equals one. Operations in and on algebras mentioned above can be used to describe interactions between individuals and the impact of external individuals on the system component (group of people).

For individuals X and Y represented by connection grids we introduce an interaction between them as an operation α of indices a,b defined as follows:

$$\alpha_{a,b} : (X, Y) \longrightarrow (a \circ_{mod(p)} X +_{mod(p)} b \circ_{mod(p)} Y, a \circ_{mod(p)} X +_{mod(p)} b \circ_{mod(p)} Y), \quad (1)$$

where a, b are natural numbers, p is a prime number.

We identify alternatives $V_1 V_2 V_3 V_4$ with numbers 1,2,3,4. The connection grid represented by stochastic matrix we treat as column vector random variable. Each component of this variable takes values 1, 2, 3 or 4 and its probability distribution can be found in the corresponding row of the connection grid.

Since for stochastic matrices of fixed rank k containing real elements $X = [a_{ij}]$, $Y = [b_{ef}]$, the product $Z = [z_{cs}] = X \cdot Y$ is a stochastic matrix, we introduce an interaction between individuals as two-argument operation β

$$\beta : (X, Y) \longrightarrow (X \cdot Y, X \cdot Y), \quad (2)$$

where X, Y are stochastic matrices representing connection grids.

Due to the fact that if X_1, X_2, \dots, X_k are stochastic matrices, then for array of numbers

$$\gamma_1, \gamma_2, \dots, \gamma_k : \forall i \gamma_i \geq 0 \wedge \gamma_1 + \gamma_2 + \dots + \gamma_k = 1 \quad (3)$$

the matrix

$$D = \gamma_1 \cdot X_1 + \dots + \gamma_k \cdot X_k \quad (4)$$

is a stochastic matrix. We introduce an interaction as operation χ of indices $\gamma_1, \gamma_2, \dots, \gamma_k$

$$\chi_{\gamma_1, \dots, \gamma_k} : (X_1, X_2, \dots, X_k) \longrightarrow (\gamma_1 \cdot X_1 + \dots + \gamma_k \cdot X_k, \dots, \gamma_1 \cdot X_1 + \dots + \gamma_k \cdot X_k). \quad (5)$$

An array of weights $\gamma_1, \gamma_2, \dots, \gamma_k$ is interpreted as a measure of mutual influences for individuals X_1, X_2, \dots, X_k .

- The properties of interactions mentioned above differ significantly. The first type can be described as a probability shuffling due to interactions between individuals (Billingsley (1987)). An individual "adjusts" his/her choice mechanism to the mechanism of interacting individual. However, such an "adjustment" is performed within the range of the individual connection grid. The third interaction type results in the weighted averaging of algorithms for a group of interacting individuals. The array of weights indicates an influence of each individual before interacting on a common choice mechanism after interacting (Plonka (2001)). The second interaction type, like the first one, applies to interactions between two individuals, but there is no room for "shuffling" and the sequence of operation plays an important role, as the result of X on Y is usually different from that of Y on X (Wilson (2000)).

4 Problem formulation

4.1 Basic terminology

1. Let's consider a population of s -individuals X_1, \dots, X_s , who make choices among the fixed set of l -alternatives $\Omega = \{V_1, \dots, V_l\}$.
2. Individual makes choices based on a choice mechanism called connection grid as described in the previous section.
3. Let's consider the aggregated choice made at time t by M -individuals belonging to the group:

$$(S^M)_t = X_1 + \dots + X_M \quad (6)$$

i.e. a convolution of M -random variables describing the choice mechanism for an individual at time t .

4. The number of summands at time t is a random variable M^t i.e.

$$P(M^t = x) = (q^t)_i, \quad (7)$$

$x=1,2,\dots,s$, s -individuals forming a group.

4.2 Assumptions

We underline that an individual belongs to a group, thus inducing that his/her actions must fall in line with the group goal.

1. Group goal is a regular, stationary group choice process represented by $(S_M)_t$. The group understands achieving of this goal imprecisely in terms of time series $(E((S^M)_t), Var((S^M)_t))_t$. If the vector of expected value and standard deviation for group choice falls beyond a specified area D , then the group's behaviour is changed so that the characteristics fall again within D .
2. Group goal is a stable process $(S^M)_t$, i.e. resistance to driving distribution function $(S^M)_t$ outside a family of distribution functions Q . The resistance to the external impact from the environment includes resistance to joining new members as well as occurrence of new untypical individual choice mechanisms (Feller (1978)).

4.3 Problem

Within the framework and group goals mentioned above we consider:

1. What type of interactions between individual choice mechanisms among (1),(2),(5) is expected, when the process $(S^M)_t$ deviates from regular run and the group makes attempts to restore regularity ?
2. What type of interaction can assure stability of $(S^M)_t$ under disturbances resulting from external influence on individual choice mechanism or group size changing or joining an individual of untypical choice mechanism ?

5 Discussion

5.1 Regularity condition

Although the random variable distribution function can be derived from known expected value and standard deviation in special cases only, empirical observations indicate that these simple numerical characteristics are very useful for our perception of the world (Buss (2001)). The regularity condition for joint choice process $(S^M)_t$, expressed in terms of expected value and standard deviation leads to the following conclusions: For variables X_1, \dots, X_M , – for which both expected values and variances do exist – the following relationship is fulfilled:

$$Var((S^M)_t) = \sum_{i=1}^M Var(X_i^t) + \sum_{i \neq k, i=1}^M \sum_{k=1}^M Cov(X_i^t, X_k^t) \quad (8)$$

One can easily conclude that to achieve a specified variance level for group choice $(S^M)_t$, the group can force the appropriate variance level for summands, i.e. choice mechanisms of individuals belonging to the group. Variance $(S^M)_t$ can be decreased by reducing the variance of components or decreasing the covariance term - reducing the degree of dependency and appropriate mixing signs and values in covariance term. It should be noted that both the second and third types of interaction "average" and "unify" choice mechanisms, while the first type of interaction "shuffles" these mechanisms. That is why we can expect that if the interest of a group is endangered, the first type of interaction should prevail over the second one that uniforms choice mechanisms for group members. In addition, the first type of interaction enables us to control covariance by "shuffling" distributions X_1, \dots, X_s (Rao Radhakrishna (1982)).

5.2 Stability condition

When considering resistance of a group choice to disturbances, an increase in group size is often indicated as a method to reduce the effect of individuals on group choice (Buss (2001)). Among interactions between individual choice mechanisms, those of the first and the third type seem to be favourable from group point of view. Due to the processes of "averaging" and "unifying", any disturbance is damped by the group and distributed among all members. The interactions of the first type can even intensify disturbances.

5.3 Typicality conditions

Let's consider inequality:

$$P(|(S^M)_t - E((S^M)_t)| \geq \epsilon) \geq P_O \quad (9)$$

meaning that the probability of the distance between the aggregated choice and its expected value exceeding the threshold ϵ is greater or equal to P_O . We interpret it rather in terms of chances of not achieving a typical situation. Using a variant of Chebyshev inequality: for any random variables X, having expected value and variance

$$P(|(S^M)_t - E((S^M)_t)| \geq \epsilon) \leq \frac{D^2((S^M)_t)}{\epsilon^2} \quad (10)$$

we see that estimation connects the chance that a typical situation is not achieved with variance of aggregate choice, thus allowing us to make use of the remarks presented above.

6 Results of empirical research

Empirical research was carried out within the scope of a research project entitled "Statistical Modelling of Customer Behaviour" under direction of Prof. A. Sokolowski. The project included the analysis of customer paths at a shopping center in Krakow meant to identify typical searching methods, i.e. typical choice sequences. The aim of the project is also to estimate how a typical choice path is changed under the influence of such disturbances as promotional activity, impoverishing of local society, changes in place of residence, welfare, type of customer activities. The following areas were distinguished at the shopping center: Entrance, Textiles, Spirits, Appliances, Beverages, Coffee - Confectionery, Canned Food, Meat - Sausages, Cheese - Frozen Food, Fruit and Vegetable, Chemistry- Cosmetics, Office Supplies, Sport, Bank, Electro, Other, Exit. Based on a random sample of 100 customers, the frequency of passages Entrance, Textiles, Textiles, Textiles,..., Electro, Other,..., Exit was recorded. The problem was considered as maximum likelihood multinomial distribution estimation, where particular outcomes were passages between the department store areas [6]. As a result "an average" choice mechanism for customers of the shopping center was established. The results are presented in Table 1.

Having the averaged values of choice mechanisms C_1, \dots, C_t derived from t- repetitions of the test one can consider the most reliable interactions among (1),(2),(5) to indicate the type of interaction that occurs between the customers. *Conditional entropy* for some random experiment as a measure of customer uncertainty what to do next is a useful tool for determining interactions between choice mechanisms, given that customer stays in a specified supermarket area. Conditional entropy for experiment U given x_i if $P(u_j|x_i) = p_{ij}$ is defined by:

$$E(U|x_i) = - \sum_{j=1}^M p_{ij} \ln p_{ij} \quad (11)$$

	ent	text	spi	appl	bev	coff	cann	meat	chee	veg	chem	offi	sport	bank	elec	othe	exi
ent	0	0.13	0.07	0.08	0.05	0.02	0.02	0	0.01	0.04	0.05	0.11	0.28	0	0.05	0.08	0.01
text	0	0.08	0.03	0.32	0.03	0.03	0	0.03	0	0	0.14	0.16	0.14	0	0.03	0	0.03
spi	0	0	0.06	0	0.34	0	0	0	0	0.06	0	0	0	0	0	0	0.56
appl	0	0	0.02	0.06	0.33	0.06	0.02	0	0	0.04	0.2	0.2	0	0	0.04	0	0.2
bev	0	0	0.02	0.05	0.06	0.16	0.02	0.05	0.06	0.16	0.1	0	0	0	0	0	0.32
coff	0	0	0.05	0.05	0.1	0.12	0.12	0.1	0.09	0.08	0.02	0	0	0	0	0	0.27
cann	0	0	0.04	0	0.02	0.26	0.04	0.13	0	0	0	0	0	0.02	0	0	0.47
meat	0	0	0	0.02	0.04	0.12	0.47	0.14	0.04	0.04	0.02	0	0	0	0	0	0.12
chee	0	0	0.03	0	0.03	0.13	0.08	0.46	0.05	0.13	0	0	0	0	0	0	0.1
veg	0	0.02	0	0.03	0.02	0.02	0.02	0.2	0.34	0.09	0.07	0	0	0	0	0	0.03
chem	0	0	0	0.11	0.17	0.11	0.03	0.04	0.03	0.2	0.21	0.04	0	0	0	0	0.04
offi	0.02	0.08	0	0.14	0.03	0	0	0	0	0.09	0.2	0.16	0.05	0.02	0.2	0.02	0
sport	0.02	0.14	0.02	0.04	0.02	0	0	0	0	0.04	0.08	0.34	0.18	0.02	0.08	0.04	0
bank	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
elec	0	0.05	0	0.05	0.03	0	0	0	0.05	0.03	0.16	0.42	0	0	0.21	0	0
othe	0	0.08	0	0	0	0	0	0	0	0	0.08	0.08	0.38	0	0.3	0	0.08
exi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table 1. The shopping center was divided into the following areas: Entrance, Textiles, Spirits, Appliances, Beverages, Coffee - confectionery, Canned Food, Meat - sausages, Cheeses - frozen food, Vegetable-fruit, Chemistry - cosmetics, Office supplies, Sport, Bank, Electro, Other, Exit. Table 1 shows an estimation of the customer connection grid. An intersection of row i and column j indicates the probability that a customer staying in area i moves to area j . The largest probability of each row is shown in bold.

We use the term of conditional entropy to estimate uncertainty, that is the degree of our ignorance about the customer's behaviour in the next step, given that he/she is just at one of 17 areas of the shopping center. To do it we compute the value of conditional entropy for each row of Table 1. The results are listed in Table 2.

ent	cof	off	chem	spo	text	bev	veg	app	chee	elec	meat	other	cann	spi
3.291	3.029	3.002	2.969	2.933	2.871	2.826	2.646	2.62	2.387	2.366	2.364	2.192	2.052	1.462

Table 2. Degree of our ignorance about the customer's behaviour in the next step

For larger values of conditional entropy of individual choice mechanisms we type rather β and $\chi_{\gamma_1, \dots, \gamma_k}$ as interactions that increase choice uncertainty by unifying probability distribution at a cost to interactions of $\alpha_{a,b}$ type.

Research description. Thanks to the kindness of the management of a "Macro Cash & Carry" store located in Krakow, it was possible to conduct a fully anonymous survey in autumn and winter of 2002. During about a two months' period, the security employees, using an industrial cameras system, observed and recorded the route of purchases of randomly selected 100 customers on the plan of the store. The time of the day, as well as the number

of clients (not more, however, than 3 per day) observed at the time was also selected randomly. The results have been arranged in form of a table (see Table 1), where for every section of the store distinguished in this survey, the number of passages to different sections of the market was counted, and each of these numbers was then divided by the total number of passages observed through the considered section. Following the frequent definition of probability - we obtained an estimate of the probability of a customer staying in the section he has found himself in and moving to any of the other sections of the market. It is also possible to look at the problem as an estimate of the density function of a multinomial distribution, where passages through the distinguished sections of the market are the individual results of the experiment - simple counting of passages has in such a case desirable statistical properties.

Attention: In the research we considered statistical space of the product (and not the product of statistical spaces) constructed on the basis of passages through the sections of a store and not on the basis of the sections themselves. This allows us to take into account, among other things, the distance between each of the store's sections. Hence, the passage Canned Fruit - Bank is not independent from passage Spirits - Beverages - as a customer needs some time to cross that distance. Although, the individual records in Table 1 (Bank - Exit) can raise some doubts - we included them, as such were the results of the survey. It seems, however, that it is possible to interpret only a part of the table, most likely because of the small size of the sample group. Some other doubts can be raised by shifting the method of analysis from the conditional distribution to the distribution in the product space. Lets then stress that in the estimation a point of statistical space represents the passing from one section of the store to the other - we estimate the distribution in the whole space - and only then, we consider the conditional distribution, that is the probability that having found himself in certain section of the store a customer will move to another section - as an example we calculate here the conditional entropy from Table 2.

The meaning of research: As a result of the survey (see Table 1) we obtained an estimate of a cumulative construction, called in this analysis "the average mesh of associations". We possess a summary information about sequence of customers' choices made in the observed market. The survey was conducted in a specific socio-economic system surrounding the market including the characteristics of the local environment (competition, accessibility) as well as the broad one (average salary, GNP growth rate, unemployment rate etc.). We wonder how the changes in the local or global environment can alter "the average mesh of associations". Would the activity of competing shopping stores change "the average path of associations" drawn on the plan of the analysed market, and if so - how would it be modified? How does the average path change during the promotional activities? In order to capture

these changes at least qualitatively, we need a simple measure of similarity of the paths referring to the whole of the customer's route and not only to an individual passage. A measure suitable for small samples, we have at our disposal and allowing to interpret changes in the environment. The entropy seems to fulfil these requirements. We suppose that less wealthy customers will be characterized by a more arranged sequence of choices. A promotion of certain article should attract customers - the entropy should then be diminishing. As to the external environment, the entropy allows for energetic interpretations - in the growth phase the economy consisting of consumers almost boils. Conditional entropy (see Table 2) allows for classic interpretations - when we are interested only in a part of the store - the complexity of the estimated distribution function decreases. Let's add, for the advertisement department of the shop, that the higher the entropy of the mesh of associations, the better customers get acquainted with the market's offer. Therefore, even though the cognitive value of the last column of Table 2 may be deemed as low, in practice, let's say during a promotion of chocolate bars in the adjacent section of the store held by attractive hostesses its value should change. Whether it would increase - in effect of customers deciding to learn about the promotion, or decrease - as the customers steered with guilt as to their unhealthy lifestyle would immediately leave the market - remains unanswered.

7 Summary

In economic research of some variables, we often encounter problems with interpretation of some time series. Any attempts to present it in the form of a stochastic process are highly complicated because of our ignorance about the laws of economic processes (Zelias (1997)). The theory of economics provides a framework, while in practice we are still surprised by real processes. In this paper, the choices made by individuals belonging to a larger group and the effect of individual choices on joint choice were discussed. It is considered what type of choice mechanism forces an individual to be a member of the group and how an individual can change the group and how the individual behaviour is influenced by the group. The term of connection grid is introduced in this paper to describe the individual choice mechanism. The three types of interactions between individual choice mechanisms are introduced. An aggregated group choice that involves interactions between individual choice mechanism is considered. It seems that the proposed descriptive and analytic methods along with the conclusions are useful for practical purposes. This research study can be useful for solving the unemployment problem, as searching for optimal behaviour norms is of crucial importance to people who have no job. Group rationality is considered rather from the survival point of view.

References

- BILLINGSLEY, P. (1987): *Prawdopodobienstwo i miara*. PWN, Warszawa.
- BUSS, D.M. (2001): *Psychologia ewolucyjna*. GWP, Gdańsk.
- FELLER, W. (1978): *Wstęp do rachunku prawdopodobienstwa t. 1 i 2*. PWN, Warszawa.
- PLONKA, E. (2001): *Wykłady z algebry wyższej t. 1 i 2*. WPS, Gliwice.
- RAO RADHAKRISHNA, C. (1982): *Modele liniowe statystyki matematycznej*. PWN, Warszawa.
- WILSON, R.J. (2000): *Wprowadzenie do teorii grafów*. PWN, Warszawa.
- ZELIAS, A. (1997): *Teoria prognozy*. PWE, Warszawa.

Mining Promising Qualification Patterns

Ralf Wagner

Department of Business Administration and Marketing,
University of Bielefeld, Universitätsstr. 25, D-33615 Bielefeld, Germany

Abstract. The skills to impart in academic management education are subject of controversial debates. In this paper a web mining approach of learning promising qualification patterns from job openings in the internet by classification of documents with a SOM network is presented. Moreover, the evaluation of clusters by association rules and a new measure for the interestingness of rules are proposed.

1 Introduction

Due to the progress in statistics, data mining, and machine learning there is a multitude of new ideas, algorithms, and techniques one would appreciate to include into the lectures. Thus, additional time and efforts of the students are desirable. In contrast to this, the students are forced to speed up with completing their studies, particularly in Germany, where the examination candidates are found to be older than in other countries. Moreover, the universities and business schools should offer education in a more ‘efficient’ manner. Of course, this development is not restricted to lectures in data analysis, but appears to be valid for a wide range of scientific disciplines. The conflict between the tightening of studies and consideration of new contents directly leads to a selection problem for lectures as well as for students. The results of these selections are patterns of qualifications.

With respect to North American and European MBA programs in business-to-business marketing Narus and Anderson (1998) point out inconsistencies in the offered qualification patterns. Additionally, based on a survey of companies opening positions in marketing research, Achenreiner (2001) worked out that the required skills are merely administrative qualifications (e.g., being familiar with Word, Excel, and Power Point) rather than statistical competence. Weick (2001) emphasizes students’ seeking for best practices without any theoretical underpinnings due to the deficits of guidance. To overcome these deficiencies in this paper

- a data analytical approach of learning useful qualification patterns from the real demand expressed in position openings is offered, and
- a measure of interestingness of specific combinations of qualifications is proposed.

The methodology might be of interest for both, lecturers as well as students. The latter want to focus on qualifications which are increasing their abilities

to reach the preferred employment opportunities. Since the universities and business schools are competing for new students with their programs, lecturers might benefit from the methodology for optimizing their course offers.

The remainder of the paper is structured as follows. In section 2 a data base gained from the internet, preprocessing, and tri-gram coding of the textual data for this specific task are described. The subsequent section 3 is devoted to the clustering of qualification profiles by a SOM-network and the decoding of the vector space model afterwards. In section 4 the evaluation of the resulting map by association rules and the measure of interestingness for specific qualification patterns are described. A brief discussion of results and a final critique are given in section 5.

2 Data base and tri-gram coding

Statistical learning requires a considerable amount of data. For the particular purpose of this study the data should reflect the required qualifications in a variety of combinations from position openings of a wide spread of different companies. This ensures that the manifold of promising qualification patterns of different industries as well as different functions of the individual position in the organizational context are covered by the data. Suitable data meeting these requirements are available from the internet by online recruitment sites. In the following 1,071 position openings for applicants with university degree from the business category of www.jobware.de are used. The analysis of such data has to tackle four challenges:

- Since the position openings are written in subsequent sentences, one has to consider word inflections due to the grammatical construction of German sentences.
- In a data base of considerable size there might occur spelling varieties. Particularly, beginning with the reform of spelling rules for the German language there are at least two correct spellings for many words.
- Since synonyms might be used already within individual documents and additionally in the comparison of different documents, it is desirable to consider the surrounding words by means of the specific context in which words or phrases describing particular qualifications are embedded.
- Of course one has to take into account minor data entry errors, e.g., spelling errors.

Previous to the coding of the documents by tri-grams a preprocessing is done. In the first step all commands of the hypertext markup language are deleted. All non-regular characters such as {+, #, %, &, ä, Ä, ...} are replaced by a blank in a second step. In the third step of preprocessing all remaining letters are transformed into capital letters. Thus, all documents consist of sequences from an alphabet of only 26 letters and blanks after preprocessing. The subsequent data coding is done with the assignment rule given in

equation 1.

$$a = f(a_1) \cdot 27^2 + f(a_2) \cdot 27^1 + f(a_3) \cdot 27^0 \text{ with } \begin{cases} f(a_l) = 0, & \text{if } a_l = ' ' \\ f(a_l) = 1, & \text{if } a_l = 'A' \\ & \vdots \\ f(a_l) = 26, & \text{if } a_l = 'Z' \end{cases} \quad (1)$$

The tri-gram coding is a special case of n -gram coding, where three letters are considered simultaneously. For coding a complete document the three letter window is iteratively shifted over the whole document. In Figure 1 the data coding is exemplarily demonstrated for the word “classification” in the English, respectively American, and the German spelling.

<i>C</i>	$3 \cdot 27^2 + 12 \cdot 27 + 1 \cdot 1 = 2,512$	<i>K</i>	$11 \cdot 27^2 + 12 \cdot 27 + 1 \cdot 1 = 8,344$
<i>L</i>	$12 \cdot 27^2 + 1 \cdot 27 + 19 \cdot 1 = 8,794$	<i>L</i>	$8,794$
<i>A</i>	$1,261$	<i>A</i>	$1,261$
<i>S</i>	$14,373$	<i>S</i>	$14,373$
<i>S</i>	$14,100$	<i>S</i>	$14,100$
<i>I</i>	$6,732$	<i>I</i>	$6,732$
<i>F</i>	$4,620$	<i>F</i>	$4,628$
<i>I</i>	$6,643$	<i>I</i>	$6,859$
<i>C</i>	$2,234$	<i>K</i>	$8,066$
<i>A</i>	758	<i>A</i>	758
<i>T</i>	$14,838$	<i>T</i>	$14,838$
<i>I</i>	$6,980$	<i>I</i>	$6,980$
<i>O</i>	$11,313$	<i>O</i>	$11,313$
<i>N</i>	$10,206$	<i>N</i>	$10,206$
.	.	.	.
.	.	.	.

Fig. 1. Example of data coding

Letters which are divergent in the two languages are italicized in the figure. Since most letters are identical the majority of the resulting tri-grams are identical as well. Thus, the two words will be considered as similar words in a classification task. This holds for spelling varieties as well as word inflections, or minor typing errors.

Because the order of words in the documents is not of interest for the given data mining task, the tri-gram sequences of an individual document d have been mapped into a vector space model, where each document is represented by a binary vector \mathbf{x}_d with $x_{ad} = 1$, if tri-gram a occurs in document d and $x_{ad} = 0$ otherwise. According to the discussions of vector space modeling of document fingerprints by Nürnberg et al. (2003) one main disadvantage is the huge dimensionality of model vectors. A general approach for reducing the dimensionality is introduced with the WINNOWING algorithm by Schleimer et al. (2003). The algorithm basically defines a threshold of the number of occurrences of particular n -grams in the data base for the acceptance of a

n -gram in the vector space. In the study at hand, a tri-gram a was accepted for the vector space only, if $a \in \mathcal{A} = \{a | 100 \leq n(a) \leq 950 \wedge 2 \leq \tilde{n}(a) \leq 12\}$ where $n(a)$ denotes the occurrences of tri-gram a in the whole data base and $\tilde{n}(a)$ denotes the occurrences of tri-gram a in individual document vectors. Thus, the dimensionality of the vector space was reduced from $27^3 = 19,683$ to $|\mathcal{A}| = 1,734$.

3 Clustering of qualification patterns and decoding

In contrast to k -means clustering a nonlinear projection from a higher dimensional feature space into a 2 dimensional space by SOM networks preserves the topological order. A simple one layer self-organizing map ($I \times J$) has been proven to offer most advantages in visualizing large and high dimensional data sets (Rauber et al. (2000)) and, therefore, is used subsequently. The prototypes of the network are represented by $1 \times |\mathcal{A}|$ vectors \mathbf{w}_{ij} . The distance between a document fingerprint \mathbf{x}_d and a prototype \mathbf{w}_{ij} is calculated by the squared Euclidean norm and the winning unit is given by $\mathbf{w}_{ij}^*(\mathbf{x}_d) := \min_{(i,j)} \|\mathbf{x}_d - \mathbf{w}_{ij}\|^2$. Unsupervised learning is done by iterative updating of the weights of the prototypes according to equation 2 for each iteration t .

$$\mathbf{w}_{ij}(t+1) := \mathbf{w}_{ij}(t) + \alpha(t)\beta(\mathbf{w}_{ij}^*(\mathbf{x}_d), t)(\mathbf{x}_d - \mathbf{w}_{ij}(t)) \quad (2)$$

with:

$$\alpha(t) = 0.9 t^{-1} \text{ and } \beta(\mathbf{w}_{ij}^*(\mathbf{x}_d), t) = \exp\left(-\frac{|i - i^*|^2 + |j - j^*|^2}{10 t^{-1}}\right) \quad \forall i, j$$

The decreasing learning rate $\alpha(\cdot)$ and the Gaussian neighborhood function $\beta(\cdot)$ are in line with the recommendations of Kohonen (2001). Document fingerprints \mathbf{x}_d are randomly chosen from the entire data base for the learning task.

The result of this classification are prototype vectors \mathbf{w}_{ij} with $w_{aij} \in [0, 1]$. Thus, one has to consider a gradual affinity of tri-grams to the prototypes instead of the clear representation by binary vectors. Moreover, the aim of the analysis is not a pattern of tri-grams but patterns of whole words representing specific meanings in a human language. For the decoding of prototypes a lookup table of more than 3,200 words v_m in alphabetical order is created and the corresponding binary vectors of tri-grams \mathbf{u}_m are computed. The affinity z_{ijm} of word v_m to prototype vector \mathbf{w}_{ij} is given by

$$z_{ijm} = \mathbf{w}_{ij} \cdot \mathbf{u}'_m \cdot \text{pen}(v_m, \text{language}) \quad \forall i, j, m \quad (3)$$

with a language-dependent penalty term.

$$\text{pen}(v_m, \text{German}) = \left(1 + \frac{\text{length}(v_m)}{4}\right)^{-1} \quad (4)$$

Since long words are translated into more tri-grams than short ones, their corresponding binary vectors have higher sums of elements and, therefore, the scalar products with arbitrary prototype vectors w_{ij} tend to be higher. In order to avoid this artificial affinity we introduce the penalty term $\text{pen}(\cdot)$ for the number of characters word v_m consists of. An adjustment to languages other than German can be made by changing the denominator from 4 to another positive number. In the English language the natural words are shorter and compound words occur rarely. Hence, the denominator should be lower than 4, e.g., 2.5, for the analysis of English documents. Thus, the proposed proceeding for the given text mining task can easily be adjusted to different languages. The following algorithm was used to extract a set \mathcal{P}_{ij} of 30 words with the highest affinity to each prototype w_{ij} from the look up table.

```

1 do for all prototype vectors  $w_{ij}$ 
2   compute  $z_{ijm}$  for all words of the look up table
3    $\mathcal{P}_{ij} \leftarrow \emptyset$ 
4   repeat 30
5     find  $m^*$  with  $\max_m z_{ijm}$  for reference vector  $w_{ij}$ 
6     add  $v_{m^*}$  to  $\mathcal{P}_{ij}$ 
7      $z_{ijm^*} \leftarrow 0$ ,  $z_{ij(m^*-1)} \leftarrow 0$ , and  $z_{ij(m^*+1)} \leftarrow 0$ 
8   end
9 end

```

From line 7 of the algorithm it is obvious, that the words directly preceding or subsequent to a word v_{m^*} are excluded from further consideration as candidates for the set \mathcal{P}_{ij} by assigning the affinity of zero. This has been found to be useful for avoiding trivial results due to word inflections, since these appear subsequently in the alphabetically ordered lookup table.

Figure 2 shows the resulting map after 100,000 iterations in a $I, J = 10$ network. The map consists of 100 units, where two units are highlighted exemplarily. The upper left one is characterized by the requirement of economic knowledge (VOLKSWIRTSCHAFTLICHE, VOLKSWIRTSCHAFTLICHER) and command of English (ENGLISCHKENNTNISSEN, ENGLISCHER). Since the positions are opened by banks, the applicants should be familiar with banking (BANKPARTNERN, BANKWIRTSCHAFTLICHEN). The position holders will be concerned with relations to partners (BANKPARTNERN, CHANNELPARTNERN, PROZESSPARTNERN) and deployment of staff (PERSONALEINSATZ). They should be willing to fill a management position (MANAGERPOSITION). Over all, this pattern appears to be a consistent description of requirements and provides an impression of face validity. The other highlighted unit is chosen because the knowledge of classification (KLASSIFIKATION) is required. As common, additional analytical skills (ANALYSETECHNIKEN, ANALYTISCHE) and the knowledge of statistics (STATISTIK) are necessary. Unexpectedly, the applicants should have knowledge of communication technology, media production and event organization (KOMMUNIKATION).

TIONSTECHNOLOGIE, MEDIENPRODUKTION, MEDIENTECHNIK, VERANSTALTUNGSORGANISATION) as well as innovation (INNOVATION). This demonstrates the ability to extract unattended qualification patterns by the classification of position openings with a SOM network.

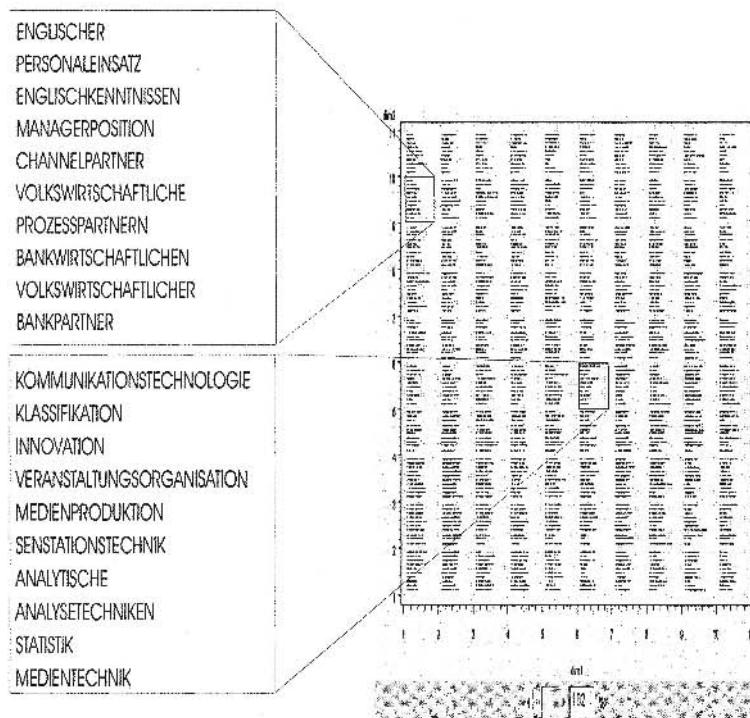


Fig. 2. Map of qualification patterns

A drawback of the proposed method is obvious from Figure 2 as well. The word "PRÄSENTATIONSTECHNIK" is cut because of the replacement of "Ä" with a blank in the preprocessing. Since the words can still be recognized, enlarging the relevant alphabet and thus increasing the dimensionality of the vector space seems not to be justified. Another weak spot is the complexity of the map. Although in Figure 2 each unit is represented by 10 words only, the resulting map consists of 1,000 words. For the evaluation of the map a Java applet was implemented to enable the navigation within the map and the zooming of areas. Nevertheless, the map is too complex to be evaluated by eye-balling only. Instead of graphical finesse, such as colored units, the use of association rules to aid the evaluation of the map is proposed.

4 Evaluation of clusters by association rules

In the given context the rules' antecedent might be interpreted as a set of existing expertise \mathcal{E} . Starting with \mathcal{E} one is interested in a set of promising complementary qualifications \mathcal{C} . Since the number of rules generated from the qualification patterns in the map will grow substantially if the minimum level of support and confidence decreases, an assessment of the rules is desirable.

According to Brin et al. (1997) the interestingness of a rule is a measure of departure from the independence of antecedent and consequent. From a probabilistic point of view the support $sup(\mathcal{E} \cup \mathcal{C})$ is the probability of co-occurrences of both item sets $P(\mathcal{C}, \mathcal{E})$, and the $conf(\mathcal{E} \Rightarrow \mathcal{C})$ might be interpreted as conditional probability $P(\mathcal{C}|\mathcal{E})$. Then the lift, which seems to be the most common measure of interestingness, can be rewritten as $lift(\mathcal{E} \Rightarrow \mathcal{C}) = P(\mathcal{C}, \mathcal{E}) / (P(\mathcal{C}) \cdot P(\mathcal{E}))$. Obviously, the measure suffers from the multiplication in the denominator, if the support of the item sets diverges considerably. This condition is true in the context of mining qualification patterns, because 'basic qualifications', e.g., communication skills, occur more frequently than 'specializations' such as statistical expertise. Thus, the following measure of interestingness for all $\{\mathcal{E}, \mathcal{C} | sup(\mathcal{E} \cup \mathcal{C}) \neq 0\}$ is proposed:

$$int(\mathcal{E} \Rightarrow \mathcal{C}) = \frac{sup(\mathcal{E}) + sup(\mathcal{C}) - 2 \cdot sup(\mathcal{E} \cup \mathcal{C})}{sup(\mathcal{E} \cup \mathcal{C})} + 1 \quad (5)$$

In the numerator of equation 5 the probabilities $P(\mathcal{C}, \neg\mathcal{E})$ and $P(\mathcal{E}, \neg\mathcal{C})$ are summed. Therefore, in the best case the measure has a value equal to 1 by means that the qualifications are only required jointly. For the simplification of interpretation 1 is added to the fraction: If $int(\cdot) = n$ then every n^{th} occurrence of a \mathcal{E} is a co-occurrence with \mathcal{C} . An interesting rule with classification in the antecedent set obtained from the map is

MARKTKOMMUNIKATION & KLASSIFIKATION & ARBEITSORGANISATION \Rightarrow VERANSTALTUNGSSORGANISATION & KOMMUNIKATION

which describes the pattern of 'market communication' & 'classification' & 'organization of work' as antecedent and 'event organization' & 'communication' as consequent. The measure of interestingness for this rule is $int(\cdot) = 1.29$ whereas the $lift(\cdot)$ is 11.11, but there are less interesting rules with a $lift(\cdot) > 11.11$. The least interesting rules with $int(\cdot) = 13.80$ and $lift(\cdot) = .95$ are different qualifications in combination with 'classification' as antecedent and business administration (BETRIEBSWIRTSCHAFTSLEHRE) as consequent only. Obviously, additional knowledge of business administration does not offer a real unique selling proposition to students who already have a set of expertise including classification skills.

5 Discussion and final remarks

In this paper a web mining approach of extracting qualification patterns from job openings in the internet is outlined. Tri-gram coding provided sufficient

precision and, therefore, four or even five-gram coding seems to be unnecessary for the given task. The qualification patterns are represented by a one layer SOM network in a suitable way. The approach differs from other projects, e.g., the WEBSOM project (Lagus et al. (1999)) because the prototypes of the networks are reduced to reflect not a single word but a pattern of different words. Thus, methodical aid for the evaluation of the map is needed, although an algorithm for the recoding of tri-grams into words of natural language is applied. For this purpose, association rules are proposed and a new measure of interestingness is introduced. The measure satisfies the minimum principle discussed by Hilderman and Hamilton (2001) but not the maximum principle. To achieve this one could simply add one to the denominator in equation 5, but this would destroy the intuitive interpretation of the measure. Over all, students as well as lecturers might benefit from extraction of condensed knowledge from the internet for their planning and substitute speculations concerning the relevance of particular qualifications.

References

- ACHENREINER, G. (2001): Market Research in the "Real" World: Are We Teaching Students What They Need To Know? *Marketing Education Review*, 11, 15–25.
- BRIN, S., MOTWANI, R., ULLMAN, J.D., and TSUR, S. (1997): Dynamic Itemset Counting and Implication Rules for Market Basket Data. In: J. Peckham (Ed.): *Proceedings ACM SIGMOD International Conference on Management of Data*. ACM Press, New York, 255–264.
- HILDERMAN, R.J. and HAMILTON H.J. (2001): Evaluation of Interestingness Measures for Ranking Discovered Knowledge. In: D. Cheung, G.J. Williams, and Q. Li (Eds.): *Advances in Knowledge Discovery and Data Mining*. Springer, Berlin, 247–259.
- KOHONEN, T. (2001): *Self-Organizing Maps*. Springer, Berlin.
- LAGUS, K., HONKELA, T., KASKI, S., and KOHONEN, T. (1999): WEBSOM for Textual Data Mining. *Artificial Intelligence Review*, 13, 345–364.
- NARUS, J.A. and ANDERSON, J.C. (1998): Master's Level Education in Business Marketing: Quo Vadis? *Journal of Business-to-Business Marketing*, 5, 75–93.
- NÜRNBERGER, A., KLOSE, A., KRUSE, R., HARTMANN, G., and RICHARDS, M. (2003): Clustering of Document Collections to Support Interactive Text Exploration. In: M. Schwaiger and O. Opitz (Eds.): *Exploratory Data Analysis in Empirical Research*. Springer, Berlin, 257–265.
- RAUBER, A., PARALIC, J., and PAMPALK, E. (2000): Empirical Evaluation of Clustering Algorithms. *Journal of Information and Organizational Sciences*, 24, 195–209.
- SCHLEIMER, S., WILKERSON, D.S., and AIKEN, A. (2003): Winnowing: Local Algorithms for Document Fingerprinting, to appear at the *ACM Special Interest Group on Management of Data*.
- WEICK, K.E. (2001): Gapping the Relevance Bridge: Fashions Meet Fundamentals in Management Research. *British Journal of Management*, 12 (Special Issue), 71–75.

Part IV

Time Series Analysis

Partial Correlation Graphs and Dynamic Latent Variables for Physiological Time Series

Roland Fried¹, Vanessa Didelez², and Vivian Lanius¹

¹ Fachbereich Statistik,
Universität Dortmund, D-44221 Dortmund, Germany

² Department of Statistical Science,
University College London, London, WC1E 6BT, U.K.

Abstract. Latent variable techniques are helpful to reduce high-dimensional time series to a few relevant variables that are easier to model and analyze. An inherent problem is the identifiability of the model and the interpretation of the latent variables. We apply graphical models to find the essential relations in the data and to deduce suitable assumptions leading to meaningful latent variables.

1 Introduction

In high-dimensional time series we may find strong correlations among the observed variables at several time lags. Statistical modelling should appropriately reflect these dependencies. However, complex models involve numerous parameters and require many observations to enable reliable inference. Thus, suitable strategies for dimension reduction provide a useful preliminary step.

A standard approach is to select a subset of the variables and to ignore the others. It is then important to know which and how much information we neglect. Alternatively, techniques like factor and principal component analysis (PCA) allow to extract latent variables describing the correlations among the observed variables and capturing more of their variability than a simple variable selection. However, the extracted variables are typically not easy to interpret although it is often important that they are meaningful.

In order to overcome these difficulties we propose to use partial correlation graphs to learn about the essential relations among the variables. These relations are visualized by a graph, where the variables are represented as vertices and the dependencies among them are shown as edges. Separations in the graph provide information about direct and indirect relations. This can be used to deduce suitable assumptions when applying factor analytic methods.

In this paper, we compare dimension reduction by variable selection, by straight-forward latent variable analysis and by latent variable analysis with restrictions derived from partial correlation graphs. We illustrate these approaches by analyzing physiological time series describing the human hemodynamic system and show that we can extract latent variables that explain

more of the observed variability than a simple variable selection but are still meaningful.

2 Partial correlation graphs and factor analysis

Graphical models visualize and clarify the dependencies among a set of variables (Whittaker (1990), Lauritzen (1996)). A *graph* $G = (V, E)$ consists of a finite set of *vertices* V and a set of *edges* $E \subseteq V \times V$, that are ordered pairs of vertices. It can be visualized by drawing a circle for each vertex and connecting each pair a, b of vertices whenever $(a, b) \in E$ or $(b, a) \in E$ by an edge. We restrict attention to undirected graphs where $(a, b) \in E$ implies $(b, a) \in E$ shown as undirected edge (a simple line) between a and b .

A *path* is a finite sequence of vertices a_0, \dots, a_n , such that there is an edge connecting each pair of subsequent vertices. Subsets $A, B \subset V$ are *separated* by a subset $S \subset V$ if every path from a vertex in A to a vertex in B necessarily includes a vertex in S . A subset $C \subset V$ is called *complete* if all possible edges between pairs of variables in C exist.

Brillinger (1996) and Dahlhaus (2000) introduce partial correlation graphs for multivariate time series. These models focus on the essential linear, possibly time-lagged relations between pairs of component series which persist after eliminating all linear effects of the other variables. Here and in the following we assume that $Y_V = \{Y_V(t), t \in \mathbb{Z}\}$, $V = \{1, \dots, d\}$, is a vector-valued weakly stationary time series with absolutely summable covariance function

$$\gamma_{ab}(h) = \text{Cov}(Y_a(t+h), Y_b(t)), h \in \mathbb{Z}.$$

For $A \subset V$ we denote the subprocess of all variables $a \in A$ by Y_A , and for $a \in V$ we denote the corresponding component process by Y_a .

The *cross-spectrum* between the time series Y_a and Y_b is the Fourier-transform of their covariance function,

$$f_{ab}(\lambda) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma_{ab}(h) \exp(-i\lambda h), \lambda \in [-\pi, \pi].$$

This defines a decomposition of γ_{ab} into periodic functions of frequencies λ . The variables Y_a and Y_b are uncorrelated at all time lags h if $f_{ab}(\lambda)$ equals zero for all frequencies.

In order to distinguish between direct and induced linear relationships between two series Y_a and Y_b , the linear effects of the remaining variables on Y_a and Y_b have to be eliminated. The *partial cross-spectrum* between Y_a and Y_b is defined as the cross-spectrum between the series ϵ_a and ϵ_b ,

$$f_{ab \cdot V \setminus \{a,b\}}(\lambda) = f_{\epsilon_a \epsilon_b}(\lambda),$$

where $\epsilon_a(t)$ and $\epsilon_b(t)$ are the residual series obtained by subtracting all linear influences of $Y_{V \setminus \{a,b\}}$ from $Y_a(t)$ and $Y_b(t)$ respectively (Brillinger (1981)).

Similarly, the (partial) cross-spectrum between two vector time series can be defined. The *partial spectral coherency* is a standardization of the partial cross-spectrum

$$R_{ab \cdot V \setminus \{a,b\}}(\lambda) = \frac{f_{ab \cdot V \setminus \{a,b\}}(\lambda)}{[f_{aa \cdot V \setminus \{a,b\}}(\lambda) f_{bb \cdot V \setminus \{a,b\}}(\lambda)]^{1/2}}. \quad (1)$$

A *partial correlation graph* for a multivariate time series is an undirected graph $G = (V, E)$ with a vertex for each of the components $a \in V$ of the time series, where two vertices a and b are connected by an edge whenever their partial spectral coherency $R_{ab \cdot V \setminus \{a,b\}}(\lambda)$ is not identical to zero for all frequencies λ . A missing edge between a and b indicates that the linear relation between these two variables given the remaining ones is zero, which is denoted by $a \perp b | V \setminus \{a,b\}$. This is known as the *pairwise Markov property*. Under the assumption that the spectral density matrix is regular for all frequencies, Dahlhaus (2000) proves that the pairwise Markov property implies the *global Markov property*, which is a stronger property in general. It states that $A \perp B | S$ for all subsets $A, B, S \subset V$ such that S separates A and B in G .

Dynamic factor analysis allows to model a multivariate time series using a lower dimensional process of latent, i.e. unobserved variables called factors. A general dynamic factor model for an observed multivariate time series Y_V is given by a dynamic regression of Y_V

$$Y_V(t) = \sum_{u \in \mathbb{Z}} \Lambda(u) X(t-u) + e(t) \quad (2)$$

on an unobserved lower dimensional factor process X with an error process e . Here, $\Lambda(u)$ are matrices of unknown parameters called loadings. In this very general form the model is not identifiable, but we nevertheless use it as a starting point, in order to understand the assumptions that can be deduced from partial correlation graphs obtained from empirical data analysis. Brillinger's (1981) dynamic PCA in the frequency domain can be used for fitting model (2) with uncorrelated factors (Forni et al. (2000)). However, the factors extracted in this way are mixtures of all variables because all loadings are distinct from zero. Automatic rotations for improving interpretation, as in the non-dynamic case, are difficult to apply since we need to perform the rotation at each frequency individually. Problems inherent to dynamic PCA are discussed in more detail by Lanius and Gather (2003).

Under the assumption that the spectral density matrix of Y_V is regular at all frequencies, an algorithm has been derived by Fried and Didelez (2003b) to construct the partial correlation graph of Y_V given model (2) with Y_V and e both following a vector autoregressive model (Reinsel (1997)). In case of uncorrelated factors and uncorrelated error processes a pair of observed variables is connected by an edge if and only if both variables have nonzero

loadings for one of the factors. Thus, the resulting graph provides an assistance in identifying the number and types of factors. A complete subset in a partial correlation graph of Y_V can be regarded as generated by a latent factor. However, the identification of such common factors can be obscured by dependencies within the error process or the factors as such dependencies may cause additional edges in the partial correlation graph. Therefore, it seems reasonable to attribute only strong relationships to the factors while the weaker ones are ascribed to errors.

3 Analysis of physiological time series

In the following we analyze multivariate time series from 25 consecutive critically ill patients (9 female, 16 male, mean age 66 years) with extended hemodynamic monitoring requiring pulmonary artery catheterization, acquired on the surgical intensive care unit of the Klinikum Dortmund, a tertiary referral center. The hemodynamic variables heart rate HR, pulse PULS, arterial systolic pressure APS, arterial mean pressure APM, arterial diastolic pressure APD, pulmonary artery systolic pressure PAPS, pulmonary artery mean pressure PAPM, pulmonary artery diastolic pressure PAPD, central venous pressure CVP and blood temperature Temp were stored for each patient in one minute intervals with a standard clinical information system. Hence, 25 ten-variate time series with an average length of about 5200 time points were available for the following analysis.

When using methods for dimension reduction we want to explain as much of the clinically relevant variability in the data as possible, by a reduced set of variables, but not irrelevant artifacts and short-term fluctuations. Therefore we removed outliers for each variable individually using a robust filtering procedure based on the repeated median, which allows to preserve trends as well as systematic shifts in the data (Davies et al. (2003), Fried (2003)).

In order to get a general impression about the relationships between the physiological variables we constructed a partial correlation graph for each patient. We used the program "Spectrum" (Dahlhaus and Eichler (2000)) which estimates the cross-spectra by a nonparametric kernel estimator and allows simultaneous testing of all partial spectral coherencies being zero or not by constructing a sample-size dependent confidence bound. For improving the results we applied a stepwise search strategy based on graph separations described by Fried and Didelez (2003a). This strategy allows to overcome masking of weaker relations by stronger associations, which may occur when estimating all partial linear relations jointly.

We found the essential linear relations revealed by the final partial correlation graphs to match the physiological relations expected by physicians. A typical example of such a graph is shown in Figure 1. Different edge types are used to indicate different strengths of relations as measured by the area below the partial spectral coherencies. For all patients we identified strong

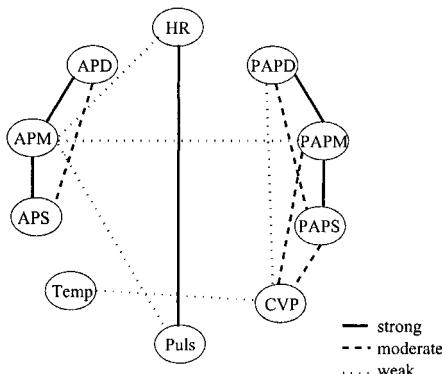


Fig. 1. Typical partial correlation graph for the hemodynamic variables of a patient.

partial correlations among the arterial pressures (APS, APM, APD), among the pulmonary artery pressures (PAPS, PAPM, PAPD) and between heart rate and pulse. The strength of the relation between the systolic and the diastolic pressure was always smaller than between each of these and the corresponding mean pressure. CVP was most strongly related to the pulmonary artery pressures, while the temperature did not show strong relations to any of the other variables. Hence, we can identify the following groups of strongly related variables from the partial correlation graphs: (APS, APM, APD), (PAPS, PAPM, PAPD, CVP), (HR, Puls).

The partitioning of the variables into strongly related subgroups is now used for variable selection. Due to the global Markov property the absence of edges between two groups of variables means that the variables in one of these groups do not contain information on the variables in the other group given the measurements of the separating variables. A variable can be regarded as very informative if it has strong relations to several other variables. Selecting e.g. APM from the strongly related subgroup of arterial pressures and neglecting APD and APSYS for clinical monitoring is therefore meaningful from both, a clinical and a statistical point of view. Applying these principles leads us to select PAPM, APM, HR and Temp.

An alternative approach for dimension reduction is to extract latent variables from the observed time series capturing as much of the total variability as possible. We scale the time series to unit variance and perform a dynamic PCA based on correlations as described by Brillinger (1981). We use four components, which is the minimal number of latent variables suggested by the partial correlation graphs.

For obtaining meaningful latent variables we can extract one component from each group of closely related variables applying dynamic PCA separately to each group. This corresponds to extracting factors as in model (2) with the $\Lambda(u)$ being restricted to be block-matrices. For heart rate and pulse, instead

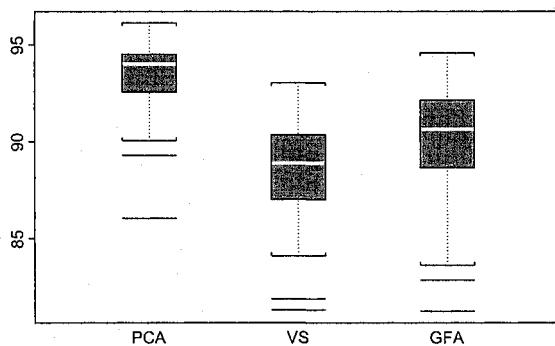


Fig. 2. Boxplots of the total explained variance (in percent): Dynamic PCA (left), variable selection (middle), grouped factor analysis (right).

of extracting a latent variable, we select the heart rate as its measurement is more reliable.

In the following we compare the percentage of variability explained by variable selection, dynamic PCA and grouped factor analysis. This is done via dynamic regression (Brillinger (1981)) of the observed variables on the selected variables and on the extracted components, respectively. Then we investigate the total residual variance as well as the individual residual variance for each variable.

Figure 2 shows that choosing the variables PAPM, APM, HR and Temp indeed explains a large part of the total variability, but less than a dynamic PCA with the same number of components, of course. Performing a grouped factor analysis allows to regain some of this loss while still providing meaningful latent variables. The variable selection explains more than 89% of the total variability for half of the patients whereas the extracted factors do so for about 75% of the patients.

Table 1 shows 5-point summaries of the explained variability for each variable. The factors derived from the groups describe the variables included in the selection very well. The explained variability increases substantially for the variables not captured well by the variable selection, see CVP and APS. When performing a standard dynamic PCA, the percentage of explained variability is at least 87% for 75% of the patients and each of the variables, which is rather high. However, these components are not meaningful to the physician. Thus, extracting latent variables from groups of closely related variables means a compromise between variable selection and factor analysis as we capture more of the total variability and of the variables neglected in the selection still working with interpretable variables.

	Min	25%	50%	75%	Max	Min	25%	50%	75%	Max
	PAPS					PAPM				
PCA	78.3	88.9	92.4	93.8	96.8	92.7	96.3	97.5	98.0	98.5
VS	57.7	72.2	83.1	88.8	94.7	100.0	100.0	100.0	100.0	100.0
GFA	61.5	75.5	84.1	87.8	93.4	85.9	93.9	95.4	96.6	97.9
	PAPD					CVP				
PCA	86.4	92.0	93.7	95.4	98.3	66.0	87.2	89.5	92.8	97.2
VS	70.5	79.5	84.8	89.7	95.2	15.3	46.9	68.0	76.4	92.5
GFA	75.6	83.3	87.5	90.7	96.4	23.6	62.1	80.4	85.2	95.4
	HR					PULS				
PCA	88.8	94.6	95.3	97.1	98.7	88.2	95.1	96.6	97.7	98.8
VS	100.0	100.0	100.0	100.0	100.0	67.3	94.5	97.3	98.7	99.8
GFA	100.0	100.0	100.0	100.0	100.0	68.7	94.4	97.3	98.7	99.8
	APS					APM				
PCA	77.9	89.4	92.7	94.8	96.8	91.6	96.3	97.3	97.7	99.2
VS	55.8	72.8	82.2	86.3	94.6	100.0	100.0	100.0	100.0	100.0
GFA	69.6	78.1	86.4	90.6	95.7	85.1	96.4	97.2	98.1	99.1
	APD					Temp				
PCA	74.1	90.8	93.8	94.7	96.7	38.9	88.4	92.3	95.8	98.6
VS	69.2	84.7	85.9	89.6	94.3	100.0	100.0	100.0	100.0	100.0
GFA	74.7	85.8	89.7	92.8	96.0	100.0	100.0	100.0	100.0	100.0

Table 1. Percentage of variability explained by PCA, by a variable selection (VS) and by a grouped factor analysis (GFA) for each variable.

4 Conclusion

Methods for dimension reduction aim at condensing the information provided by a high-dimensional time series into a few essential variables. Partial correlation graphs are a suitable tool to explore the relations among the observable variables. This information allows an advanced application of dimension reduction techniques. One possibility is to select suitable subsets of important variables from the graphs. Alternatively, we can enhance latent variable techniques. Deducing restrictions on the loading matrices from a graphical model combines variable selection and PCA as the percentage of explained variability is substantially higher than for a variable selection and we obtain meaningful latent variables. In our study the groups of closely related variables obtained from the data analysis agree with the groups anticipated from medical expertise. Therefore, we expect to gain reliable insights also in the relations among other variables, for which we have less background knowledge.

Acknowledgements: The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") and the European Community's Human Potential Programme under contract HPRN-CT-2000-00100 (DYNSTOCH) as well as the helpful comments by two referees are gratefully acknowledged.

References

- BRILLINGER, D.R. (1981): *Time Series. Data Analysis and Theory*. Holden Day, San Francisco.
- BRILLINGER, D.R. (1996): Remarks Concerning Graphical Models For Time Series And Point Processes. *Revista de Econometria*, 16, 1–23.
- DAHLHAUS, R. (2000): Graphical Interaction Models for Multivariate Time Series. *Metrika*, 51, 157–172.
- DAHLHAUS, R. and EICHLER, M. (2000): SPECTRUM. A C program to calculate and test partial spectral coherences. Available via <http://www.statlab.uni-heidelberg.de/projects/>.
- DAVIES, P.L, FRIED, R., and GATHER, U. (2003): Robust Signal Extraction for On-line Monitoring Data. *Journal of Statistical Planning and Inference*, to appear.
- FORNI, M., HALLIN, M., LIPPI, M., and REICHLIN, L. (2000): The Generalized Dynamic Factor Model : Identification and Estimation. *The Review of Economics and Statistics*, 82, 540–554.
- FRIED, R. (2003): Robust Filtering of Time Series with Trends. Technical Report 30/2003, SFB 475, University of Dortmund, Germany.
- FRIED, R. and DIDELEZ, V. (2003a): Decomposability and Selection of Graphical Models for Multivariate Time Series. *Biometrika* 90, 251–267.
- FRIED, R. and DIDELEZ, V. (2003b): Latent Variable Analysis and Partial Correlation Graphs for Multivariate Time Series. Technical Report 6/2003, SFB 475, University of Dortmund, Germany.
- LAURITZEN, S.L. (1996): *Graphical Models*. Clarendon Press, Oxford.
- LANIUS, V. and GATHER, U. (2003): Dimension Reduction for Time Series from Intensive Care. Technical Report 2/2003, SFB 475, University of Dortmund, Germany.
- REINSEL, G.C. (1997): *Elements of Multivariate Time Series Analysis*. Second edition. Springer, New York.
- WHITTAKER, J. (1990): *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

Bootstrap Resampling Tests for Quantized Time Series

Jacek Leśkow¹ and Cyprian Wronka²

¹ Department of Econometrics,
The WSB-NLU Graduate School of Business, 33-300 Nowy Sącz, Poland

² School of Engineering and Physical Sciences,
Heriot-Watt University,
Edinburgh, EH14 4AS, Scotland

Abstract. The aim of this article is to compare the spectral densities of original and quantized time series via bootstrap-based consistency tests. If the quantizing is based on Kohonen's SOM algorithm then the spectral densities of the original and quantized time series are indistinguishable. We illustrate this via studies of bootstrap distributions of tests statistics that compare spectral densities. We obtain that when the quantizing is based on SOM algorithm then the density of the original time series and the spectral density of the quantized one are not significantly different.

1 Introduction

Consider a stationary time series X_1, \dots, X_n with the spectral density function (SDF) f . It is well known (see e.g. Kedem (1994)) that a smoothed periodogram \hat{f}_n provides a reliable estimate of the SDF. Asymptotic distribution of \hat{f}_n enables one to construct simultaneous confidence intervals for the unknown function f . In practice, however, it is not always clear when asymptotic results can be used or, in other language, whether the sample size n is already large enough. This fundamental problem can be solved with the bootstrap resampling techniques that are available for time series. The main idea of the bootstrap is quite simple. It is based on appropriately constructing samples X_1^*, \dots, X_n^* that are similar to the original sample X_1, \dots, X_n . Then, for each such sample X_1^*, \dots, X_n^* one recalculates the initial estimate. In our case we will study \hat{f}_n via its bootstrap version \hat{f}_n^* . This is done by studying the bootstrap sampling distribution of the \hat{f}_n^* obtained from recalculating several times the bootstrap samples X_1^*, \dots, X_n^* . From this sampling distribution one may calculate quantiles that can be later used to construct the bootstrap-based confidence intervals. However, before applying the bootstrap derived confidence intervals one has to check the consistency of the bootstrap. Without going into technical details, consistency of the bootstrap implies convergence of the bootstrap derived finite-sample quantiles to the quantiles of the asymptotic distribution of the considered statistic. More technical details on bootstrap can be found in the monograph of Efron and Tibshirani (1993).

In our article we will briefly review four bootstrap techniques that are available for time series: moving blocks bootstrap (MBB), sieve bootstrap, local bootstrap and subsampling. We treat bootstrap only as a tool, however. The goal of our research is to compare the spectral density of the quantized time series and that of the original one. The motivation comes from the result of Kedem (1994) related to clipping the Gaussian time series. This result establishes a direct link between the binary version of the Gaussian time series and the first order autocorrelation of the original Gaussian series. Our study shows that a similar relationship can be established for a general, stationary time series and its quantized version analyzing spectral densities of both. Given a sufficiently large number of clusters we obtain that the difference between spectral densities of the original time series and the spectral density of the quantized version is not significant.

2 Bootstrap for time series

Moving blocks bootstrap

Let X_1, \dots, X_n be a path of strictly stationary m -dependent process. We will define the *moving blocks bootstrap* scheme (called also the MBB method) in steps.

Step 1. Let us denote the block of length b by $B_i = (X_i, \dots, X_{i+b-1})$, $i = 1, \dots, n - b + 1$.

Step 2. Instead of drawing the bootstrap sample from the original data X_1, \dots, X_n , as it is done in the case of i.i.d. variables, we will draw the sample of blocks B_1, \dots, B_{n-b+1} . The bootstrap sample will contain only $l = [n/b] \cdot b$ elements.

We draw with replacement k ($kb = l \approx n$) blocks B_1^*, \dots, B_k^* from the set $\{B_i; i = 1, \dots, n - b + 1\}$.

Step 3. The bootstrap sample is formed by pasting the blocks B_i^* together. Let us denote the j th component of B_i^* by $Y_{j,i}^*$; then $B_i^* = (Y_{1,i}^*, \dots, Y_{b,i}^*)$ and finally, the bootstrap sample is

$$(X_1^*, \dots, X_l^*) = (Y_{1,1}^*, \dots, Y_{b,1}^*, Y_{1,2}^*, \dots, Y_{b,2}^*, \dots, Y_{1,k}^*, \dots, Y_{b,k}^*).$$

The moving blocks bootstrap also performs well for other weakly dependent processes, not only for m -dependent data. The consistency of this scheme can be shown for the processes of finite fourth moment, possessing the φ -mixing property if the statistic θ is a function of the arithmetic mean. A characterization of class of statistics for which the MBB provides a consistent approximation can be found in Lahiri (1992).

Sieve bootstrap

The sieve bootstrap method (Bühlmann (1998)) is similar to the idea of bootstrap for $AR(p)$ models presented in Efron and Tibshirani (1993). However, in the sieve bootstrap method the parameter p is not assumed to be

fixed. We adjust the $AR(p_n)$ process ($p_n < \infty$) to the data assuming $p_n \rightarrow \infty$, as the sample size $n \rightarrow \infty$. Moreover the process $\{X_n\}$ is not assumed here to be zero-mean and therefore the $AR(p)$ model is adjusted to the centered process $\{X_n - \mu\}$. The mean of the process is estimated by the sample mean and denoted by $\hat{\mu}_n$; the mean of the bootstrap sample will be denoted by $\hat{\mu}_n^*$. Below we present the sieve bootstrap in the form of the algorithm .

Step 1. The mean of the process is estimated by the sample mean $\hat{\mu}_n$. The parameter $p = p(n)$ is chosen by the AIC criterion (see Shumway and Stoffer (2001)) and we estimate the coefficients ϕ_i , $i = 0, 1, \dots, p$, by the Yule-Walker method. In the end, we compute the estimated innovations:

$$\hat{\varepsilon}_n = \sum_{i=0}^p \hat{\phi}_i (\hat{X}_{n-i} - \hat{\mu}_n), \quad n = p+1, \dots, n.$$

Step 2. We assume the innovations to have zero mean hence, as in the previous method, we put $\tilde{\varepsilon}_i = \hat{\varepsilon}_i - \bar{\varepsilon}$, $i = p+1, \dots, n$, where $\bar{\varepsilon} = \frac{1}{(n-p-1)} \sum_{i=p+1}^n \hat{\varepsilon}_i$. Then the bootstrap sample of innovations $\varepsilon_{p+1}^*, \dots, \varepsilon_n^*$ is drawn as in the resampling scheme for i.i.d. variables.

Step 3. The bootstrap sample X_1^*, \dots, X_n^* is the iterative solution of the equations:

$$\hat{\varepsilon}_n^* = \sum_{i=0}^p \hat{\phi}_i (\hat{X}_{n-i}^* - \hat{\mu}_n), \quad n = p+1, \dots, n.$$

where the initial p values of X_i^* , $i = 1, \dots, p$ are equal to X_i , respectively.

The consistency of the sieve bootstrap was proven in Bühlmann (1998) for a class of linear and nonlinear statistics that are functions of the sample mean.

Local bootstrap for periodogram

Another way of dealing with the heterogeneity of periodogram ordinates is the *local bootstrap*. We assume the spectral density to be continuous, thus the values of $f(\lambda)$ do not differ much on small intervals. The bootstrap value of $P_n^*(\lambda_m)$ is drawn from the set of periodogram ordinates in a neighborhood of the frequency λ_m , that is from the set $Q_n(\lambda_m) = \{P_n(\lambda_{m+j}); |j| \leq J_n\}$, where J_n is a fixed positive number depending only on the sample size n . We will define the bootstrap scheme in the following steps.

Step 1. Define the discrete probabilities $p_{J,j}$, $j = 0, \pm 1, \dots, \pm J$, $J = J_n \leq [L/2]$ ($L = [n/2]$), such that $p_{J,j} = p_{J,-j}$, $\sum_{j=-J}^{j=J} p_{J,j} = 1$ and for a random variable X define the distribution $P(X = j) = p_{J,j}$, $j = 0, \pm 1, \dots, \pm J$.

Step 2. Draw the random variables ζ_1, \dots, ζ_L from the distribution defined in step 1.

Step 3. The bootstrap periodogram is defined as

$$\begin{aligned} P_n^*(\lambda_j) &= P_n(\lambda_{j+\zeta_j}) \text{ for } j = 1, 2, \dots, L; \\ P_n^*(\lambda_j) &= P_n^*(-\lambda_j) \text{ for } j = -1, -2, \dots, -L; \\ P_n^*(0) &= 0. \end{aligned}$$

We extend the definition of periodogram to the whole interval $[0, \pi]$ by putting $P_n(\lambda) = P_n(g(\lambda, n))$ for $\lambda \neq 2\pi/n$; $g(\lambda, n)$ denotes the Fourier frequency closest to λ and the smaller one if there are two such frequencies. The consistency of the local bootstrap method was proved in Politis et al. (1992).

A recent method of generating finite-sample distributions via subsampling was presented in Politis et al. (1999). The idea is quite simple and is based on research related to cross validation. In order to apply subsampling technique, the value of the estimator is recalculated on a reduced sample. Next, for consecutive blocks of subsamples the values of estimator are calculated thus generating a subsampling distribution. According to the research presented in Politis et al. (1999) it is relatively straightforward to prove asymptotic validity of such procedures under very broad assumptions - see Theorem 4.2.1 of Politis et al. (1999, p.103). Here, we limit ourselves only to presenting the algorithm.

Subsampling algorithm

Step 1. Assume that $\hat{\theta}_n$ is the estimator based on the full sample X_1, \dots, X_n . We select consecutive blocks of the size b .

Step 2. For each block X_t, \dots, X_{t+b-1} we recalculate the initial estimator and thus we obtain the subsampled value $\hat{\theta}_{n,b,t}$.

Step 3. The sampling distribution of appropriately normalized estimator $\hat{\theta}_{n,b,t}$ is calculated on blocks X_t, \dots, X_{t+b-1} where $t = 1, \dots, n - b + 1$.

The subsampling confidence intervals for the unknown estimated parameter θ are derived as quantiles of the sampling distribution of $\hat{\theta}_{n,b,t}$.

3 Analysis of quantized time series

In this section we would like to present our results related to comparisons between the spectral density of the original time series and the spectral density of its quantized version. The purpose of our work is to check how spectral properties of the time series are influenced by the SOM quantization. We have established that for a sufficiently high number of quantization levels the SOM quantized time series has a spectral density that is not significantly different from the spectral density of continuous valued signal.

For the clarity of the presentation, let us recall the method of constructing the asymptotic simultaneous confidence intervals for the unknown spectral density f corresponding to a stationary time series X_1, \dots, X_n . For the technical details see Shumway and Stoffer (2001).

Fact 1. Let f be an unknown spectral density of the stationary time series X_1, \dots, X_n that is φ -mixing. Then the asymptotic $e^{-2\alpha} \cdot 100\%$ confidence interval for f simultaneous in m points in the frequency domain has the form:

$$\begin{aligned} \exp \left\{ \sqrt{\frac{2m}{n}} \Phi^{-1} \left(\frac{\alpha}{m} \right) \right\} \frac{m}{\pi} \int_{\frac{k\pi}{m}}^{\frac{(k+1)\pi}{m}} \hat{f}_n(\lambda) d\lambda &\leq \\ \leq f((k + \frac{1}{2})h) &\leq \\ \leq \exp \left\{ \sqrt{\frac{2m}{n}} \Phi^{-1} \left(1 - \frac{\alpha}{m} \right) \right\} \frac{m}{\pi} \int_{\frac{k\pi}{m}}^{\frac{(k+1)\pi}{m}} \hat{f}_n(\lambda) d\lambda. \quad (1) \end{aligned}$$

In the formula (1) Φ denotes the standard normal distribution, \hat{f}_n is the estimator of the SDF f , k is the index of the comparison point and $k = 0, 1, \dots, m$ and $h = \frac{\pi}{m}$. Moreover, $m \rightarrow \infty$ and $n/m \rightarrow \infty$, where n is the sample size.

We will modify the classical confidence interval (1) using all of the presented bootstrap methods. The motivation for this is quite fundamental. We know that the confidence interval (1) is asymptotic so it is possible to use it only when the sample size n is large. On the other hand, asymptotic confidence intervals in the analysis of time series may be misleading, as shown in Efron and Tibshirani (1993). Therefore, the bootstrap confidence intervals, based on consistent bootstrap algorithms are more reliable for small and moderate sample sizes.

Let first

$$Y_k^* = \sqrt{\frac{nh}{2\pi}} \log \left[\frac{\int_{kh}^{(k+1)h} \hat{f}_n^*(\lambda) d\lambda}{\int_{kh}^{(k+1)h} \hat{f}_n(\lambda) d\lambda} \right],$$

where \hat{f}_n is the estimator of the SDF and \hat{f}_n^* is a bootstrap replication of \hat{f}_n . Other notation is as in (1). Now let F^* be the distribution function corresponding to the sample $Y_k^*; k = 1, \dots, m$. Then we have the following

Proposition 2. Let the bootstrap confidence interval of the SDF f simultaneous for m points in the frequency domain be given by

$$\begin{aligned} \exp \left\{ \sqrt{\frac{2m}{n}} F^{*-1} \left(\frac{\alpha}{m} \right) \right\} \frac{m}{\pi} \int_{\frac{k\pi}{m}}^{\frac{(k+1)\pi}{m}} \hat{f}(\lambda) d\lambda &\leq \\ \leq f((k + \frac{1}{2})h) &\leq \\ \leq \exp \left\{ \sqrt{\frac{2m}{n}} F^{*-1} \left(1 - \frac{\alpha}{m} \right) \right\} \frac{m}{\pi} \int_{\frac{k\pi}{m}}^{\frac{(k+1)\pi}{m}} \hat{f}(\lambda) d\lambda. \quad (2) \end{aligned}$$

The interval defined in (2) has the limiting $e^{-2\alpha}$ coverage probability.

Proof of this result is straightforward and is based on consistency theorems valid for the bootstrap schemes presented in this article. It is therefore omitted. \square

Proposition 2 enables us to construct a bootstrap-based consistency test for spectral densities. We would like to compare two spectral densities g and f corresponding to the original time series, say X_1, \dots, X_n and the quantized one Y_1, \dots, Y_n , respectively. In our case, $Y_i = SOM(X_i)$, where SOM is the Kohonen's self-organizing map transformation. For the sake of clarity, we will assume that we know the SDF g of X_1, \dots, X_n . We will then estimate the SDF f of the SOM-quantized time series Y_i and compare the significance of the difference between f and g using the bootstrap confidence interval (2) for f . Our method can be also applied in the situation, when the spectral density g is not known.

The consistency test can be formulated as

$$H_0: f = g$$

$$H_A: f \neq g.$$

We will reject the null hypothesis if $g \notin [f_l, f_u]$, where the limits of the confidence interval f_l and f_u are calculated according to (2).

4 Example

In this section we will provide a numerical study related to our bootstrap confidence interval presented in (2). We have generated a sample of 1024 observations from a stationary AR(1) time series with the parameter $\phi = 0.5$ (case of the known SDF g). Then we have derived the SOM version of it with 5 levels of quantization. We can visualize the rejection of H_0 from (3) if the spectral density f does not fall into the band $[g_l, g_u]$. On the picture,

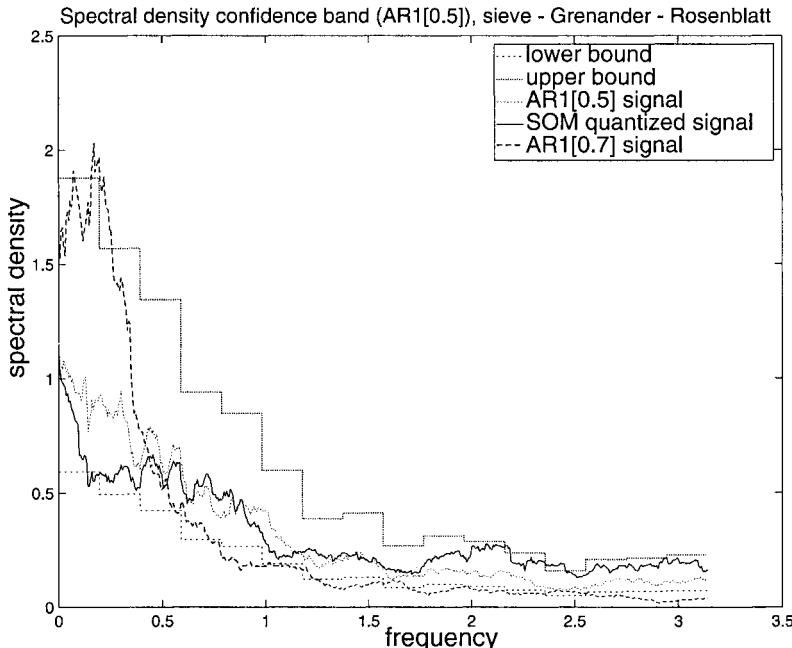


Fig. 1. Comparison of SOM for AR(1) time series, sieve bootstrap.

the lower bound is represented by a dotted stepwise line, the upper bound by a solid stepwise line, the spectral density of the original AR(1) series by a thin solid curve, and the density of the SOM quantized time series by a solid curve. Moreover, as an additional test, a spectral density of a different time series was plotted as a dashed curve.

Observe that in Figure 1 we obtain that SDF of the original AR(1) time series with the parameter $\phi = 0.5$ belongs to the same confidence band as the SDF of the SOM quantized AR(1) with $\phi = 0.5$. On the other hand, the SDF of AR(1) series with $\phi = 0.7$ is for the most part out of the confidence band. This enables us to state that the given SOM quantizer provides new time series with only 5 possible values for which the spectral density is not significantly different from the spectral density of the original time series.

References

- BÜHLMANN, P. (1998): Sieve bootstrap for smoothing in nonstationary time series. *Annals of Statistics*, 26, 1, 48–83.
 EFRON, B. and TIBSHIRANI, R. (1993): *An introduction to the bootstrap*. Chapman and Hall, New York.

- GRENANDER, U. and ROSENBLATT, M. (1957): *Statistical Analysis of Stationary Time Series*. Wiley, New York.
- KEDEM, B. (1994): *Time Series Analysis by Higher Order Crossings*. IEEE Press, Piscataway, NJ.
- KOHONEN, T. (1995): *Self-Organizing Maps*. Springer Series in Information Sciences, Springer, Berlin.
- POLITIS, D.N., ROMANO, J.P., and WOLF, M. (1999): *Subsampling*, Springer, New York.
- SHUMWAY, R. and STOFFER, D. (2001): *Applied Statistical Time Series Analysis*. 2nd edition, Prentice Hall, Englewoods Cliffs.

Imputation Strategies for Missing Data in Environmental Time Series for an Unlucky Situation*

Daria Mendola

Dipartimento di Scienze Statistiche e Matematiche “Silvio Vianelli”,
Università degli Studi di Palermo, I-90128 Palermo, Italy
mendola@dssm.unipa.it

Abstract. After a detailed review of the main specific solutions for treatment of missing data in environmental time series, this paper deals with the unlucky situation in which, in an hourly series, missing data immediately follow an absolutely anomalous period, for which we do not have any similar period to use for imputation. A tentative *multivariate* and *multiple* imputation is put forward and evaluated; it is based on the possibility, typical of environmental time series, to resort to correlations or physical laws that characterize relationships between air pollutants.

1 Introduction

Missing data (MD) are a common feature in database coming from surveys in any field of study; in particular it is an almost endemic problem of high frequency time series. In the literature, “missing data” is a generic term referring to very different situations. For example it is usual to call *indistinctly* MD those arising from a non response in a survey or an interview, and also those arising from a break-down in measurement instruments. This means that one chooses to ignore the relevant distinction between the mechanisms that generated the MD; in particular the former are missing data because of a deliberate will of non response, while the latter are due to external (technical) reasons, presumably due to chance. Anyway, both kinds of MD should be properly called **fully missing** data. On the other hand, numerical or categorical data can be assumed (and treated) as missing even when they are in some way known (measured or collected) but their quality and reliability are not reputed to be adequate: e.g. when readings are below the threshold of detection of the instrument, or when observed values are incoherent with the other collected information. These should be properly called **invalidated** data (ID). It is noteworthy that final users generally do not know the nature of missing data. Classic literature allows to distinguish among missing data according to the nature of the mechanism which generated observed and unobserved information. here, these topics were passed over, referring to the

* Research was supported by MIUR funds PRIN MM13208412_001 (2000) and PRIN 2002134337_002 (2002).

seminal work of Little and Rubin (1987).

In environmental time series, which are the main focus of this paper, one usually deals with physical and chemical measurements. There are no problems like: reticence, reserved information, privacy respect, non-responses as, for example, in social and medical surveys; instead, main sources of missing records in environmental databases are:

- break-down of measurement instruments;
- ordinary maintenance interventions (calibration, quality assurance procedures, ...)
- extraordinary maintenance interventions;
- invalidation of readings (inconsistent or impossible);
- censored data: values under the *limit of detection* (lod).

2 Common strategies to cope with missing data

2.1 Complete-case or available-case analysis

Quite often the approach to missing data problem simply consists of omitting cases and pretend that they never existed. This quite common practice is often damaging and risky: in fact, excluding the units with incomplete observations from the study may imply a heavy loss of information (especially in a multivariate analysis) and can lead to biased results if the incomplete cases differ systematically from the complete cases. Usually we do not have reasons to suppose MD are in any way correctly represented by data we do have. Furthermore, a complete-case or an available-case analysis assume the strong hypothesis of MCAR (in the sense of Rubin (1987)), so it is remarkable to note that excluding cases should not be thought of as “purer” or less “assumption-laden” than imputation.

2.2 State-space models

An attractive alternative is supplied by the use of a state-space framework, which has the ability to treat series that have been observed at unequally spaced time intervals or, which is the same, regular time series with missing observations. Shumway and Stoffer (1982, 2000) and Stoffer (1986) illustrate filtering equations, based on a modified EM algorithm, resulting in maximum likelihood estimation of vectorial ARMA models with missing data. As a brief reference, Jones (1980) provides a state-space framework for calculating exact likelihood function of a ARMA model in (stationary) time series with missing observations, and Gomez and Maravall (1994) provide a new definition of the likelihood of an ARIMA model with missing observations, by means of a state-space representation of the nonstationary (original, i.e. not integrated) data, allowing the use of an ordinary Kalman filter estimation procedure. Anyway, discussing state-space modelling in detail is beyond the scope of this paper, whose main interest relies on imputation procedures, so here we do not linger over this subject.

2.3 Imputation strategies

When the proportion of missing values is relatively small, i.e. less than 5 per cent, *ad hoc* methods such as case deletion or mean substitution may be an appropriate solution. In multivariate settings, however, where one or more variables might be affected by MD, the proportion of cases with missing values can be substantial, and the loss of information too heavy.

Imputation, i.e. substitution, of missing values with some kind of estimate of unobserved values, is a widely adopted strategy to rescue units for which one has at least partial information (especially in a multivariate analysis). Most common procedures for imputation of missing data can be classified as: a) *univariate*, that substantially use information from the distribution of the variable itself; and b) *multivariate*, generally based on regression analyses to estimate what the observations would have been in the light of observed pattern for one or more (possibly) related variables.

A more remarkable distinction is between *single* imputation and *multiple* imputation methods; the former produces a one-to-one substitution, i.e. one imputed value substitutes one missing value; while the latter provides a set of substitutions for each missing value.

Single imputation methods:

Common and easy univariate single imputation methods comprehend: substituting the mean or median of a variable in place of missing values on the same variable; or fitting a function that could well represent the pattern of the variable and use it to impute missing values. Mean substitution, although preserving the variable means, leads to underestimation of the variance-covariance structure, whereas regression methods for imputation tend to inflate observed correlations. Substitution procedures which are considered to be “best estimators” in time series analyses, in case of short length period with MD or ID, include following single imputation methods:

- **persistence:** it is the use of data from the previous time period. Substitutions under the hypothesis of the stability of phenomenon on the same level of the previous period decrease variability of the series and they are reasonable for imputations in short gaps. The error innate in this univariate imputation procedure is increasing with the length of the not observed period, and it is highly correlated to the frequency of time series observations (hourly, weekly, monthly, and so on) and to the variability (instability) of the series;

- **interpolation:** is the joint use of data from the previous and posterior time periods. This univariate imputation procedure is particularly reasonable for diffusion phenomena such as air-water-soil pollution, meteorological measurements, and so on. Even in this case, the goodness of imputations decreases with the increase of the length of the gaps;

- **profiling:** refers to procedures in which MD for one level (site) in a multi-level (multi-site) database are replaced by an estimate based on data from an alternative level or levels (site/s) in the same database, which is

used as a “donor”. This procedure often resort to spatial correlations or time correlations which are characteristic and well evident in environmental time series. In this (single and multivariate) procedure probable errors do not increase with the duration of the missing data.

Multiple imputation methods:

The idea of multiple substitutions and the explanation of the *multiple imputation* (MI) procedure is due to the seminal work of Rubin, organically synthesized in Rubin (1987 and 1996). The technique has been developing in the last two decades, and is now implemented in many statistical software or library (both commercial and free shared), like SODAS, MICE, and many others. MI techniques provide a set of alternative imputed values to be substituted to missing value. The presence of m imputed values for one missing value reflects the uncertainty associated with the missing observation, and allows for providing unbiased estimates for parameters of interest and their variances. MI consists essentially of three steps:

- 1) create m complete datasets by filling in the MD through imputation;
- 2) analyze the m completed datasets;
- 3) combine the results from m analyses to yield fine inference on the parameters of interest.

A crucial matter in MI is however the choice of the imputation model, according to nature of data and of known relationship between variables, i.e. the derivation of an appropriate predictive distribution from which imputations have to be drawn, which has a closed form analytic solution. For MD, none of the parameters in the prior distribution is of interest and all are eventually integrated out, and only the conditional distribution of MD given the observed values is used.

It is noteworthy highlighting that all the above described imputation methods, widely known in the last decades literature, implicitly assume that MD can be assimilated to observed data coming from like situations. In the next section a case-study will be illustrated which represents a very *unlucky* situation: when MD are subsequent to an absolutely anomalous period, for what there is no similar period to use for any tentative imputation. This could be assimilated to an *experiment without replications*.

3 Missing in environmental time series: a real case study

Since 1996 and until 2000 a network of seven monitoring sites for air pollution was established in Palermo. The net is managed by Municipal Environmental Agency which does not perform any imputation on original data, but does a quality control procedure on original readings and proceeds to invalidate incoherent data. As usual in high frequency time series there are many MD due to common reasons above mentioned; now attention is focussed on a very *unlucky* situation: when there are MD after a period of time absolutely

anomalous. In particular, from the 10th to the 16th of December 2000 Palermo hosted the *United Nations Summit on Organized Crime and Drug Smuggling*. Owing to security reasons a wide area in the center of the city was totally inhibited to motor vehicles traffic 6 days 24 hours a day. In that area there is luckily an air monitoring site, so we thought this was a special occasion to evaluate the effect of drastic and lasting (6 days) traffic limitations on concentrations of air pollutants. During the Summit a heavy *temperature inversion* was recorded, characterized by high temperatures (about 15 C° during night in December), a very low humidity percentage and dry wind. This altered the effect of traffic reduction causing an *unexpected* general increase in all pollutants concentrations, due to the reduction of diffusion-chamber for pollutants. This is clearly visible by the peaks of pollution which are recorded for all air pollutants in almost all monitoring sites in the city, and in particular for NO₂, which is represented in Figure 1. However, few hours after the end of traffic restrictions some measurement instruments broke-down, so there was a rather long period of fully MD, especially for NO₂, just during the days of interest for that study on road traffic and air pollution (see Mendola and Lovison (2002)). The period of UN Summit was somewhat unique. In Palermo it was never observed such a long period without new emissions of air pollutants in the city center. Note that road traffic is practically the only source of air pollution in the city. Moreover, this situation is not comparable with others of traffic limitations, e.g. the car-free Sundays situation (Lovison and Mendola (2003)); since it is presumable that the effect of traffic restrictions on pollutant concentrations is not linear but cumulative during such a long period (6 days, about 144 hours of closing). So we face here an "experiment without replications". How can MD be imputed in such a special situation? Dealing with air pollutants, strongly characterized by interactions and sometimes physical dependencies, it seems necessary to resort to *multivariate* imputation methods, so to obtain imputations which are more stable, consistent and efficient which, in some way, could take into account, the interaction structure in data. A decision is made pro *multiple* imputation which is able to account for the uncertainty associated with the missing values. However, despite the availability of a database with more than 45,000 records, the long available time series of air pollutants does not really help, and could even be misleading, due to special nature of the Summit period, as explained above. An obliged choice was to restrict on information on the six days of the Summit. MIs were performed by *MICE* (Multivariate Imputation by Chained Equations) library (van Buuren and Oudshoorn (2000)), using a Bayesian linear regression imputation with normal errors. MD on NO₂ series (which is the most damaged) in the week after the Summit were imputed by using as predictors the hourly readings of CO, SO₂, O₃ and also of meteorological variables (such as temperature and humidity percentage)¹ in the

¹ Not all available air pollutants and meteorological variables were used and this to avoid multicollinearity between predictors. In particular we did not resort

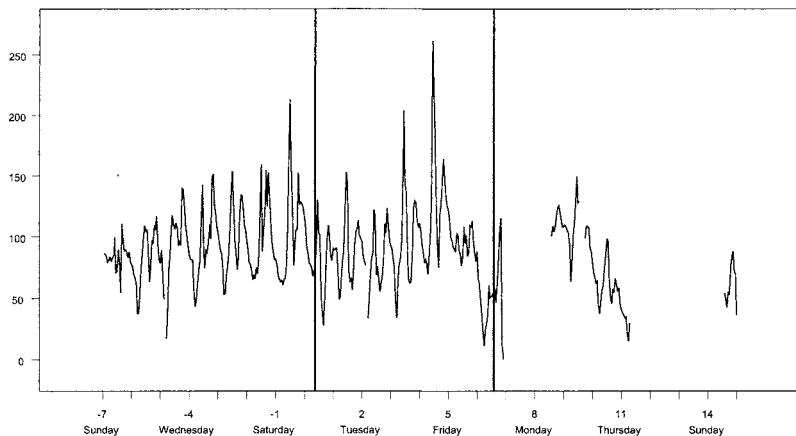


Fig. 1. NO_2 ($\mu\text{g}/\text{m}^3$) three weeks: before-during-after UN Summit

seven days before the end of the Summit. A choice was not to resort to the few observed values of NO_2 in the week after the Summit, while observed values of predictors in the same period were used to improve performances of the imputation procedure. In MI, the imputation problem is to draw from $P(X)$, the unconditional multivariate density of X . Let t denote an iteration counter. Assuming that data are missing at random, one may repeat the following sequence of Gibbs sampler iterations:

for X_1 : draw imputations X_1^{t+1} from $P(X_1 | X_2^t, X_3^t, \dots, X_k^t)$

for X_2 : draw imputations X_2^{t+1} from $P(X_2 | X_1^{t+1}, X_3^t, \dots, X_k^t)$

⋮

⋮

for X_k : draw imputations X_k^{t+1} from $P(X_k | X_1^{t+1}, X_2^{t+1}, \dots, X_{k-1}^{t+1})$,

i.e., condition each time on the most recently drawn values of all other variables.

It is important to note that also in predictor variables there were MD, and so imputation procedure adopts a visiting scheme “from left to right” during with both predictors and predicted series are sequentially updated and imputed (for further details, see van Buuren and Oudshoorn (1999 and 2000)). Note that, in MICE implementation of multiple imputation algorithm, it is *assumed* that a multivariate distribution exists, and that draws from it can be generated by iteratively sampling from the conditional distributions. In this way, the multivariate problem is split into a series of univariate problems. Rubin coined the term “incompatible Gibbs” to refer to these kinds

to particulate matters series which has a different behavior respect to gasiform pollutants, neither to NO_x which measurements are strictly related to that for NO_2 . Moreover readings on wind, solar radiation and pressure were not used because of their strong relationship with Ozone and with CO dynamics.

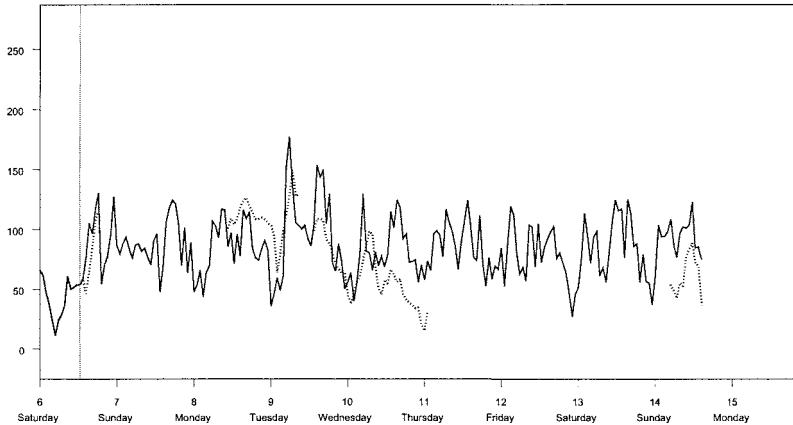


Fig. 2. NO₂: observed (dotted line) vs mean of five imputations (solid line), along 8 days after the end of Summit

of algorithms. Five alternative imputations were performed for each missing value.

3.1 Main results and final considerations

Figure 2 synthesizes results of imputation procedure: it represents hourly observed series (dotted line), which has few observations, and the means of five imputations for each missing value (solid line). Many matrices of predictors were tested; in particular the introduction of meteorological variables allowed to manage possible misleading effect induced by anomalies due to temperature inversion episode (Mendola (2002)). As expected, the knowledge of the readings of other pollutants (which have less MD than the imputed NO₂ series) helped in catching dynamics of NO₂. Imputation procedures seems quite good, despite of the *unlucky* situation: imputations caught, on average, cyclic components of NO₂ series, as can be seen from Figure 2. Mean of the process was substantially preserved: in the three weeks before-during-after UN Summit observed series has mean equal to 86.18 ($\mu\text{g}/\text{m}^3$) and s.d.=34.78, while imputed values (for the eight days after the end of Summit) have mean= 88.32 and s.d.=29.99; it seems quite reasonable. The difference between predicted (imputed) and observed values has a mean of 4.10 (median=0.00) and s.d.= 18.6. However, procedure seems not to be able to fit some of the too fast decays in NO₂ series. This is in some way reasonable, because imputations are founded on data from a period without pollutants emissions. That pattern can well represent the hours immediately after the end of traffic restrictions and becomes less representative, as expected, the most far away from Summit period.

References

- GOMEZ, V. and MARAVALL, A. (1994): Estimation, Prediction, and Interpolation for Nonstationary Series with the Kalman Filter. *JASA*, 89, 426, 611–624.
- HOPKE, P.K., LIU, C., and RUBIN, D.B. (2001): Multiple Imputation for Multivariate Data with Missing and Below-Treshold Measurements: Time-Series Concentrations of Pollutants in the Artic. *Biometrics*, 57, 22–33.
- JONES, R.H. (1980): Maximum Likelihood Fitting of ARMA Models to Time Series With Missing Observations. *Technometrics*, 22, 3, 389–395.
- LITTLE, R.J. and RUBIN, D.B. (1987): *Statistical analysis with Missing Data*. Wiley, New York.
- LOVISON, G. and MENDOLA (2003): Car-free Sundays and Air Pollution. *Conference Book of Abstract of The ISI International Conference on Environmental Statistics and Health, Santiago de Compostela (Spain) July 16–18, 2003*.
- MENDOLA, D. (2002): Road traffic restrictions and air pollution in an urban area. A case study in Palermo. Working Paper GRASPA, n.15, <http://www.graspa.org>.
- MENDOLA, D. and LOVISON, G. (2002): Are car-free days effective. A case study in Palermo using Dynamic Linear Models. *Invited Paper in TIES2002, The Annual Conference of The International Environmetric Society, Conference book of Abstracts*, Genova, June 18–22, 2002.
- MENDOLA, D. and LOVISON, G. (2003): Multivariate Monitoring of Air Pollutants and of Effects of Meterological Conditions. *Proceedings of SIS2003, Annual meeting of Italian Statistical Society*.
- RUBIN, D.B. (1987): *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.
- RUBIN, D.B. (1996): Multiple Imputation After 18+ Years. *JASA*, 91, 434, 473–489.
- SHUMWAY, R.H. and STOFFER, D.S. (1982): An approach to time series smoothing and forecasting using EM algorithm. *J. Time Series Anal.*, 3, 253–264.
- SHUMWAY, R.H. and STOFFER, D.S. (2000): *Time Series Analysis and Its Applications*, Springer, New York.
- STOFFER, D.S. (1986): Estimation and Identification of Space-Time ARMAX Models in the Presence of Missing Data. *JASA*, 81, 395, 762–772.
- VAN BUUREN, S. and OUDSHOORN, C.G.M. (1999): Flexible multivariate imputation by MICE TNO report PG/VGZ/99.054, TNO Prevention and Health, Leiden.
- VAN BUUREN, S. and OUDSHOORN, C.G.M. (2000): Multivariate Imputation by Chained Equations, - MICE V1.0 User's manual. TNO report PG/VGZ/00.038, TNO Prevention and Health, Leiden.

Prediction of Notes from Vocal Time Series: An Overview

Claus Weihs^{1,2}, Uwe Ligges¹, and Ursula Garczarek¹

¹ Fachbereich Statistik, Universität Dortmund, D-44221 Dortmund, Germany

² weihs@statistik.uni-dortmund.de

Abstract. This paper deals with the prediction of notes from vocal time series. Different kinds of classification algorithms using different amounts of background information are described and compared. The results of the methods are presented by an transcription algorithm into musical notes.

1 Introduction

Analogous to speech recognition systems on computers, our aim is the transcription of the correct notes corresponding to a vocal time series. In this paper we describe the main steps and methods on the way from the musical time series to note classification and to transcription to sheet music.

Algorithms for note classification build the center of our theoretical considerations. In Section 4 we compare different kinds of classification algorithms using different amounts of background information for the generation of classification rules producing note predictions. We present segmentation algorithms based on periodograms, pitch estimation, and smoothing of predicted notes (cp. Ligges et al. (2002)). Also we experiment with different amounts of background information corresponding to voice type and expected notes.

As an alternative, we present an algorithm not estimating the pitch but using the whole periodogram for training on parts of the song and testing on the remaining parts. As the classification algorithm we use a radial basis function support vector machine (*RBF SVM*) together with a “Hidden Markov” model as a dynamization mechanism. The parameters of the *RBF SVM* are selected by optimizing the validation set error using experimental design with a quadratic loss function (Garczarek et al. (2003)).

The results of the methods for the experiment described in Section 2 are presented in Section 5 using an automated transcription algorithm into musical notes (Weihs and Ligges (2003)), which is briefly described in Section 3.

2 Data

We base our results on experiments with singing presentations of the classical song “Tochter Zion” by Händel sung by 17 singers, amateurs as well as professionals, to a standardized piano accompaniment played back by headphones.



Fig. 1. Notes of "Tochter Zion"

The singers could choose between two accompaniment versions transposed by a third in order to take into account the different voice types. For more details on the data see Weihs et al. (2001).

Note that "Tochter Zion" has an ABA form (see Figure 1). This will be utilized later.

For our analysis the time series corresponding to the singing presentations are transformed into periodograms since for music the frequencies of the signals are well-known to be decisive. Examples for parts of the involved time series and the corresponding periodograms can be found in Figure 2 (time series) and Figure 3 (periodograms). Note the differences between the amateur and the professional identified to be mainly caused by the presence of vibrato leading to a two mode peak for the professional (cp. the periodograms).

3 Transcription

The whole transcription algorithm is described only briefly in this paper, since its main topic is prediction of notes. For more details on the transcription procedure we refer to Weihs and Ligges (2003).

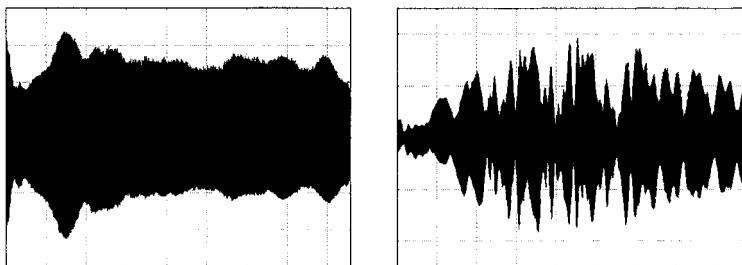


Fig. 2. Waves of syllable “Zi” in “Tochter Zion”

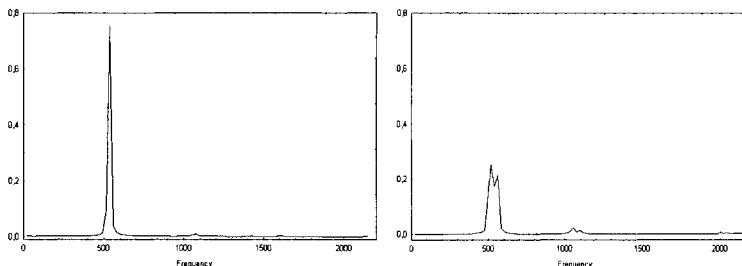


Fig. 3. Periodograms corresponding to waves in Figure 2

Generally, automated transcription starts with a separation step of the singing voice from any other sound (of the same record). In the example in Section 5 we left out any such segmentation step, because sounds were already separated on our records. In the literature, e.g., Hyvärinen et al. (2001) propose *Independent Component Analysis* (ICA) to achieve separation of a polyphonic sound. Von Ameln (2001), however, points out this method’s insufficiency for our experiments.

The next step is segmentation of sound into segments presumably corresponding to notes, silence, or noise, as well as pitch estimation and classification of the corresponding note. This is done by the classification algorithms described in Section 4.

Segmentation is succeeded by a quantization step, where the relative lengths of notes (eighth notes, quarter notes, etc.) are derived from the estimated absolute lengths. In our example this was done manually. In the literature, e.g., the method of Cemgil et al. (2000) can be considered for automatical quantization.

Finally, the data yielded in the first steps have to be transcribed into music notation, so called sheet music. For this purpose we connected the statistical software of our choice, R (Ihaka and Gentleman (1996)), with LilyPond (Nienhuys et al. (2002)) by the interface developed by Preusser et al. (2002). LilyPond can automatically produce sheet music.

4 Note classification

4.1 Segmentation based on pitch estimation

The first algorithm for note classification we describe in this paper is an improved version of the Segmentation algorithm based on Note Classification (*SNC*, for short) described in Ligges et al. (2002). Basically, that algorithm passes through the vocal time series by a window of given size (e.g. $n = 512$ is appropriate for a wave file sampled with 11kHz; we will call each part of the time series covered by one particular window a *section*), estimates the pitch for each section, and classifies the corresponding note, given the frequency of diapason a' is known, or at least has been estimated as well. Since the appearance of vibrato is a big problem for a stable estimate, the classified notes have to be smoothed before segmentation. For a definition of vibrato see Seidner and Wendler (1997). Methods related to vibrato analysis have been developed, e.g., by Rossignol et al. (1999).

For this paper, we improved the SNC algorithm (cp. Ligges et al. (2002)) by replacing the smoother by a doubled *running median* with window width 9. So this approach can be described as follows:

- Estimation of the fundamental frequency of each section of the time series
- Classification using the fundamental frequency note classifier $C_{\lambda_0}^{FF}$ (Ligges et al. (2002))
- Smoothing (doubled running median) of the list of classified tones
- Segmentation, if a change in that smoothed list occurs

Additionally, the algorithm is improved (i.e. misclassification of tones is reduced) by applying a-priori knowledge in the first two steps, i.e. the estimation of the fundamental frequency and the classification procedure:

- *No-Info*: No restriction at all, i.e. all notes can be classified.
- *Voice-Info*: Restrict the notes to those that can theoretically be sung by the voice type (soprano, alto, tenor, bass) of the particular singer. Ranges are taken from Seidner and Wendler (1997).
- *Song-Info*: More specifically, restrict the notes to those that should appear in the particular song the singer performs.

The misclassification rates on the last part of the song “Tochter Zion” (we choose the third part *A* of its *ABA* scheme for comparability with the *RBFSVM* algorithm described below) were calculated for all singers (cp. Table 1, left hand side) for all combinations of

- the three kinds of a priori knowledge, and
- the pure (without any smoother) and smoothed versions of classification.

Consider the (smoothed) list of classified sections, and a list of “correct” results, i.e. a list of manually classified notes we based on the piano accompaniment. Then we can compare these two lists and check whether a tone

is correctly classified with the exception that a tone is allowed to begin two sections too early or too late, respectively. These two section intervals were chosen, because such an inaccuracy (shorter than half an eighth note) would not affect the transcribed sheet music for the given song. Sections with low energy (i.e. low volume because of silence or breathing periods) were omitted from calculations of the misclassification rates.

As expected, the more information is used the better is the classification. Smoothing (of e.g. vibrato effects) additionally improves the misclassification rate. Of course, knowledge about notes of the analysed song might be unrealistic in real world applications, whereas knowledge about the voice type is more realistic. Nevertheless, we used the information about notes in Section 5.

4.2 Radial Basis Function Support Vector Machines (*RBFSVM*)

In Garczarek et al. (2003) we describe methods for note classification and segmentation by learning algorithms without any pitch estimation. The best one uses *Support Vector Machines with Radial Basis Functions with a Hidden Markov Model (RBFSVM-HMM)* combined with a smoothing step.

The data used for these algorithms are periodograms derived in the same way as for the *SNC* algorithm (cp. Section 4.1), but using a window size of 256 observations, which results in 128 Fourier frequencies in the periodogram. Moreover, we selected the most important interval of 40 Fourier frequencies for each particular voice type, since the program is not able to deal with more variables.

The algorithm is described by the following steps. Different prediction rules are built on the learning data that represent different stages of a multi-step learning procedure:

1. Basic quantization of evidence

The evidence on each note $n \in \mathbf{N} := \{1, \dots, N\}$ in the observed periodogram \mathbf{x}_t at time point t , $t=1, \dots, T$, is quantized by the membership functions m on $\mathbf{N} \times \mathbf{x}_t \in \mathbf{X}$ of support vector machines with radial basis functions. The membership functions are scaled such that for any given periodogram they define a probability distribution over the notes: $m(n, \mathbf{x}) \geq 0$ and $\sum_{n=1}^N m(n, \mathbf{x}) = 1$ for all $n \in \mathbf{N}$ and all $\mathbf{x} \in \mathbf{X} \subseteq \mathbb{R}^K$. As a classification algorithm we use support vector machines using radial basis function kernels (cp. Vapnik (1995) and Schölkopf (1998)) as implemented in the Support Vector Machine (SVM) toolbox 2.51 for Matlab by Schwaighofer (2002).

Support vector machines are identifying so called “support vectors” most important for the distinction between the classes being the notes in our case. Radial basis function kernels are local gaussian densities around each data point. With the *RBFSVM* such kernels are only placed on support vectors.

One of the parameters in *RBFSVM* controls the trade off between margin maximization and error minimization. A second parameter defines the

width of the RBF-kernel. A method based on a gradient decent algorithm to adjust these parameters is described in Chapelle et al. (2001). Instead, we choose these parameters by optimizing the validation set error on both of these parameters using experimental design with a quadratic loss function.

For more details we refer to Garczarek et al. (2003).

2. Static Prediction

In a static fashion the note \hat{n}_t^s with the highest current evidence $m(n, \mathbf{x}_t)$ from the periodogram \mathbf{x}_t is predicted:

$$\hat{n}_t^s := \arg \max_{n \in \mathbb{N}} (m(n, \mathbf{x}_t)) \quad (1)$$

3. Estimation of transition probabilities between notes

The transition probabilities between true notes are estimated by the observed frequencies on the learning set.

4. Dynamized Prediction

Dynamic prediction is based on a Hidden Markov Model for the transition between notes (cp. Garczarek et al. (2003)). This model is instantiated with the above transition probabilities and the scaled membership values as emission probabilities. With this model, we estimate the probability $p(n|\mathbf{x}_1, \dots, \mathbf{x}_t)$ that the true note is n given the current observed periodogram \mathbf{x}_t and all observed periodograms $\mathbf{x}_1, \dots, \mathbf{x}_{t-1}$ before. Based on these estimates, the second rule predicts the note with highest estimated probability:

$$\hat{n}_t^d := \arg \max_{n \in \mathbb{N}} (\hat{p}(n|\mathbf{x}_1, \dots, \mathbf{x}_t)) \quad (2)$$

5. New quantization of evidence from predictors

What is left is the definition of the smoother. This is based on evidence on how often a note n is confused with others by some predictor $\hat{n} \neq n$ on the learning set. This is counted in the so-called confusion matrix:

$$\mathbf{C}(\mathbf{L}) = \begin{bmatrix} c_{1,1} & c_{1,2} & \dots & c_{1,K} \\ c_{2,1} & c_{2,2} & \dots & c_{2,K} \\ \vdots & \vdots & \ddots & \vdots \\ c_{K,1} & c_{K,2} & \dots & c_{K,K} \end{bmatrix} \quad (3)$$

with $c_{i,j} := \sum_{t=1}^{T_L} \mathbb{I}_i(\hat{n}_t) \mathbb{I}_j(n_t)$, where $\mathbb{I}_i(\hat{n}_t)$ is the indicator function for $\hat{n}_t = i$. Divided by its sum $c_{i,.} := \sum_{t=1}^{T_L} \mathbb{I}_i(\hat{n}_t)$ each row of $\mathbf{C}(\mathbf{L})$ gives the relative frequencies on the learning set that j was the true note, given that i was predicted. The standardized rows are thus estimators of the conditional probabilities for each note $p(j|\hat{n} = i)$:

$$\hat{p}(j|\hat{n} = i) := \frac{c_{i,j}}{c_{i,.}}$$

That way, $\hat{p}(n|\hat{n}^s)$ and $\hat{p}(n|\hat{n}^d)$, $n, \hat{n}^s, \hat{n}^d \in \mathbf{N}$, quantize the evidence one gets from the predictions \hat{n}^s or \hat{n}^d for the note n .

6. Smoothed Static and Dynamic Prediction

Especially for professional singers vibrato is observed in singing. In order not to mix vibrato with tone changes, a smoothing algorithm with a window size adapted to the individual singer is used. Smoothing at time point t thus uses the evidence for the notes one gets from $\hat{p}(n|\hat{n}_s^s)$ or $\hat{p}(n|\hat{n}_s^d)$ in some window around t :

$$s \in W_t(w) := [max(t - w, 1), min(t + w, T)].$$

To predict, equal weight is given to the evidence at any time point s in this interval:

$$\hat{n}_t^{ss} := \arg \max_{n=1,\dots,N} \left(\sum_{s \in W_t} \hat{p}(n|\hat{n}_s^s) \right) \quad (4)$$

$$\hat{n}_t^{sd} := \arg \max_{n=1,\dots,N} \left(\sum_{s \in W_t} \hat{p}(n|\hat{n}_s^d) \right) \quad (5)$$

The optimal size of $w \in \{1, \dots, L\}$ is determined in terms of the learning error rate. L is some definition of the minimum length of a tone. Here, L is set to 20 which is somewhat less than the length of an eighth note (=24) in the given experiment.

Table 1 on its right hand side shows the misclassification rates for the various variations (*static* vs. *HMM* and *pure* vs. *smoothed*, respectively) of this algorithm on the last part *A* of “Tochter Zion”. The first two parts, *AB*, were used for learning.

Obviously, the main difference of the misclassification rates of the best *SNC* algorithms (“Song-Info” and “smoothing”, cp. Section 4.1) and the best *RBFSVM* algorithms (“HMM” and “smoothing”) is the slightly better maximum misclassification rate of the latter. However, the *RBFSVM-HMM* algorithm can still be improved by modelling silence as for the *SNC* algorithm.

Unfortunately, the prerequisites of *RBFSVM-HMM* are quite strong. For the learning step a singer has to sing each possible note of the song separately, otherwise a segmentation and corresponding “ideal” notes must be given by the supervisor. If arbitrary notes are presented for learning instead of those only appearing in the song, error rates might become worse, because of the bigger number of classes.

Another approach using Hidden Markov Models is described in Cano et al. (1999). Dixon (1996) proposes a method for “multiphonic note identification” that covers the polyphonic case which was not (yet) possible for our methods.

¹ The “silence sections” are a subset of “all sections”.

Singer	SNC								RBFSVM			
	# Sections		No-Info		Voice-Info		Song-Info		static		HMM	
	all	silence ¹	pure	sm.	pure	sm.	pure	sm.	pure	sm.	pure	sm.
B01	837	111	0.42	0.27	0.43	0.29	0.29	0.17	0.17	0.12	0.11	0.08
B03	837	88	0.53	0.52	0.52	0.52	0.36	0.34	0.28	0.24	0.21	0.20
B06	837	138	0.27	0.20	0.27	0.20	0.16	0.11	0.20	0.19	0.14	0.14
B07	837	107	0.60	0.48	0.50	0.33	0.35	0.23	0.29	0.25	0.27	0.28
T03	820	64	0.24	0.13	0.24	0.13	0.14	0.07	0.11	0.09	0.06	0.05
T06	820	60	0.43	0.23	0.43	0.24	0.25	0.14	0.18	0.15	0.10	0.06
T07	815	51	0.31	0.16	0.31	0.16	0.20	0.10	0.18	0.14	0.12	0.10
A02	834	87	0.37	0.28	0.35	0.29	0.16	0.11	0.23	0.20	0.19	0.18
A04	837	104	0.32	0.24	0.29	0.22	0.17	0.14	0.29	0.26	0.20	0.19
A05	837	111	0.14	0.10	0.14	0.10	0.06	0.04	0.13	0.10	0.06	0.05
A09	837	120	0.19	0.12	0.18	0.12	0.12	0.08	0.17	0.12	0.14	0.14
A10	837	119	0.25	0.17	0.23	0.16	0.10	0.05	0.24	0.22	0.11	0.10
S01	819	84	0.29	0.19	0.25	0.17	0.16	0.11	0.21	0.19	0.15	0.15
S02	820	169	0.23	0.18	0.20	0.17	0.12	0.08	0.15	0.12	0.11	0.10
S04	820	99	0.31	0.25	0.27	0.22	0.22	0.19	0.20	0.20	0.13	0.12
S05	820	61	0.46	0.22	0.45	0.20	0.28	0.19	0.15	0.10	0.07	0.06
MIN	815	51	0.14	0.10	0.14	0.10	0.06	0.04	0.11	0.09	0.06	0.05
MED	837	104	0.31	0.20	0.27	0.20	0.16	0.11	0.19	0.17	0.12	0.11
MAX	837	169	0.60	0.52	0.52	0.52	0.36	0.34	0.29	0.26	0.27	0.28

Table 1. Misclassification rates for different kinds of a-priori knowledge and smoothing (sm.) for the *SNC* and the *RBFSVM* algorithms; MIN = minimum, MED = median, MAX = maximum of the corresponding column

5 Example

In this section we present some results of the improved *SNC* algorithm above (cp. Section 4) in detail. Again, we take the last part *A* of the *ABA* scheme of “Tochter Zion” of a professional soprano singer, *S05*, from our experiment mentioned in Section 1.

At first, the pitch of each section of the vocal time series is estimated. The frequency of diapason a' of the piano accompanying the singers was estimated as 443.5 Hz. From these values the corresponding note is estimated for each of the sections.

The raw (i.e. without smoothing) classified sections of the first 2 bars (blank line between bars is inserted manually) are given in Figure 4. The value 0 corresponds to a' , the other integers represent the distance from a' in halftones, thus -2 means g' etc. Silence and quiet noise, respectively, is represented by NA. Singer *S05* has an intensive vibrato, which can easily be seen in the first row, where the classification switches rapidly between 2 (b') and 3 (c'').

NA	2	2	3	2	2	3	3	3	2	3	3	3	3	3	2	2	2	2	2	5	3	2	2
2	2	2	5	3	2	2	2	2	3	2	2	2	2	2	2	2	2	2	NA	NA	NA	NA	NA
-2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	-2	-2	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	1	NA	1	NA	NA	
1	NA	2	3	2	2	2	3	3	3	2	2	2	2	3	3	3	2	2	2	2	3	3	2
3	3	2	2	2	3	3	3	2	2	2	3	3	3	2	2	3	3	3	2	2	2	2	-2
-4	-4	-4	NA	NA	-5	-5	-4	-4	-4	-4	-4	-4	-4	-5	-5	-5	-5	-4	-4	-4	-5	-5	-5
-5	-4	-4	-4	-5	-5	-4	-4	NA	NA	-4	NA	-2											

Fig. 4. Raw section classification of singer S05 for the first 2 bars of the last part of “Tochter Zion”

Having applied the smoother the classified sections are compared to the ideal progression of the song, which is shown in Figure 5. Obviously, the smoothing step does not smooth off the intensive vibrato completely, compare in particular the fourth ideal note (c''). At the bottom of Figure 5 the progression of energy of the voice is plotted, where low energy reflects breathing, silence, and strong consonants (“ch” in “dich” and “jauchze”). We do not count such parts as errors.

In principle, the next step is to apply some quantization as described in Section 3. It was quite easy, though, to derive beat and bars from the piano accompaniment manually. The result of this quantization step is shown in Figure 6, where we took the statistical mode of the classified sections within each eighth note of the beat as an estimator. The data represented in this figure can be transcribed into notes, since the minimal required information (beat, bars, notes, rests, etc.) is available now. For the sheet music we use symbols for *rests* to transcribe silence and low energy noise. The term *rest* is used in the following discussion for low energy parts.

The outcome of applying the transcription method to these data is given in Figure 8, whereas the corresponding ideal sheet music is given in Figure 7. The error rate given in the figures’ captions is calculated as follows (altogether there are 64 eighth notes in all 8 bars):

$$\frac{\# \text{ erroneously classified eighth notes (without counting rests)}}{\# \text{ all eighth notes} - \# \text{ eighth rests}}.$$

In our example, there are 12 erroneously classified eighth notes and 3 eighth rests, i.e. an error rate of $12/61 = 0.20$. Figure 9 shows a transcription based on Voice-Info (cp. Section 4.1). Using this more realistic a-priori information increases the error rate to 0.21. Note that recognizing the structure of the overtones the first note could be easily corrected.

Note that errors and error rate, respectively, are the sum of at least four kinds of errors:

- errors of the transcription algorithm
- errors of the singers’ performances
- errors from inaccurate timings of the singers with respect to the piano

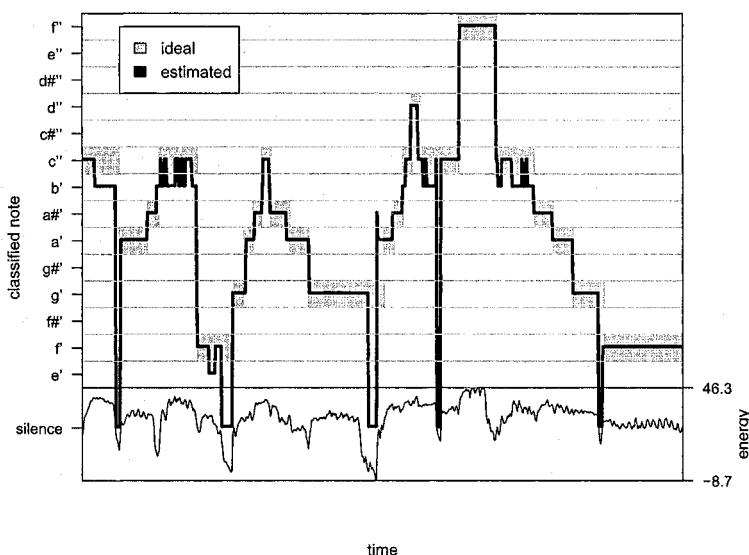


Fig. 5. Estimated vs. ideal notes of singer S05

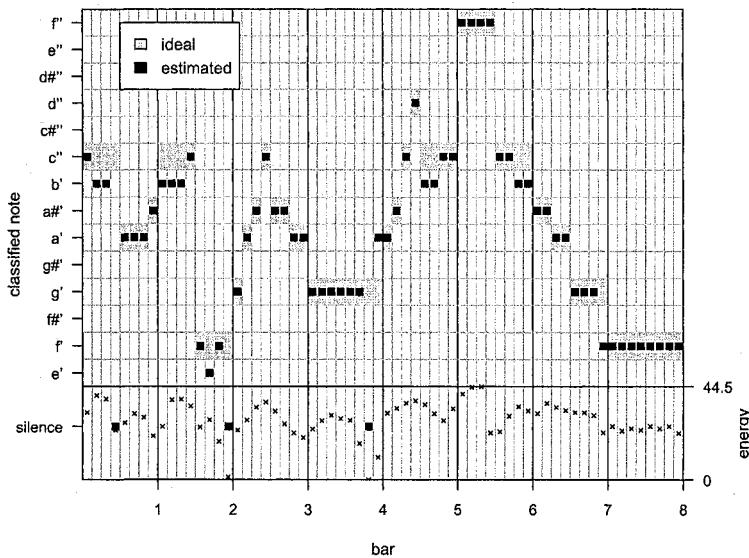


Fig. 6. Estimated vs. ideal notes of singer S05 after quantization



Fig. 7. Original sheet music of “Tochter Zion”



Fig. 8. Transcription of singer S05 using Song-Info (error rate: 0.20)



Fig. 9. Transcription of singer S05 using Voice-Info (error rate: 0.21)

- errors from vibrato (since our smoothing methods are not able to smooth off all of it)

6 Conclusion

With the presented methods and algorithms we were able to transcribe our examples in a sufficient way in order to avoid the need of extensive postprocessing. Unfortunately, some restrictive assumptions were included, particularly the knowledge about:

- the shortest note length (an eighth note), which is e.g. important for the smoother
- accompaniment pitch (at least correctly estimated)
- bar locations (beginning, end; we left off the automatic quantization step in our examples)
- notes to be sung in song

Knowledge about those notes which appear in a song is unrealistic. Nevertheless both our best algorithms use such information, not only *SNC* with “Song-Info” and “smoothing”, but also *RBFSVM* with “HMM” and “smoothing” since it is trained on the first A part and predicts the last A part of the same song. Using only the voice type as information, instead, reduces the classification performance.

Almost all note segments are identified correctly after smoothing, except for vibrato. Thus the question arises how to better smooth out vibrato, or how to include a model for vibrato in order to improve note classification. Moreover, in order to avoid counting errors from inaccurate timing, automatic dynamical quantization directly on the singers’ performances is planned to be considered.

References

- VON AMELN, F. (2001): *Blind source separation in der Praxis*. Diploma Thesis, Fachbereich Statistik, Universität Dortmund, Germany.
- CANO, P., LOSCOS, A., and BONADA, J. (1999): Score-Performance Matching using HMMs. In: *Proceedings of the International Computer Music Conference*. Beijing, China.
- CEMGIL, T., DESAIN, P., and KAPPEN, B. (2000): Rhythm Quantization for Transcription. *Computer Music Journal* 24 (2), 60–76.
- CHAPELLE, O., VAPNIK, V., BOUSQUET, O., and MUKHERJEE, S. (2001): Choosing Multiple Parameters for Support Vector Machines. *Machine Learning*, 46, 131.
- DIXON, S. (1996): Multiphonic Note Identification. *Australian Computer Science Communications*, 17 (1), 318–323.
- GARCZAREK, U., WEIHS, C., and LIGGES, U. (2003): Prediction of Notes from Vocal Time Series. *Technical Report 1/2003*, SFB475, Department of Statistics, University of Dortmund. <http://www.sfb475.uni-dortmund.de>.
- HYVÄRINEN, A., KARHUNEN, J., and OJA, E. (2001): *Independent Component Analysis*. Wiley, New York.
- IHAKA, R. and GENTLEMAN, R. (1996): R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5 (3), 299–314.
- LIGGES, U., WEIHS, C., and HASSE-BECKER, P. (2002): Detection of Locally Stationary Segments in Time Series. In: W. Härdle and B. Rönz (Eds.): *CompStat2002 – Proceedings in Computational Statistics – 15th Symposium held in Berlin, Germany*. Physika Verlag, Heidelberg, 285–290.
- NIENHUIJS, H.-W., NIEUWENHUIZEN, J., et al. (2002): *GNU LilyPond – The Music Typesetter*. Free Software Foundation, <http://www.lilypond.org/>, Version 1.6.5.
- PREUSSER, A., LIGGES, U., and WEIHS, C. (2002): Ein R Exportfilter für das Notations- und Midi-Programm LilyPond. *Arbeitsbericht* 35, Fachbereich Statistik, Universität Dortmund, Germany. <http://www.statistik.uni-dortmund.de>.
- ROSSIGNOL, S., DEPALLE, P., SOUMAGNE, J., RODET, X., and COLLETTE, J.-L. (1999): Vibrato: Detection, Estimation, Extraction, Modification. In: *Proceedings 99 Digital Audio Effects Workshop*.
- SCHÖLKOPF, B. (1998): Support-Vektor-Lernen. In: G. Hotz et al. (Eds.): *Ausgezeichnete Informatikdissertationen*. Teubner, Stuttgart, 135–150.
- SCHWAIGHOFER, A. (2002): *SVM toolbox for Matlab*. <http://www.igi.tugraz.at/aschwaig/software.html>.
- SEIDNER, W. and WENDLER, J. (1997): *Die Sängerstimme*. Henschel, Berlin.
- VAPNIK, V. (1995): *The Nature of Statistical Learning Theory*. Springer, New York.
- 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, Bergisch-Gladbach, Köln, 395–410.
- WEIHS, C. and LIGGES, U. (2003): Automatic Transcription of Singing Performances. *Research Report 03/1*, Department of Statistics, University of Dortmund. <http://www.statistik.uni-dortmund.de>.

Parsimonious Segmentation of Time Series by Potts Models

Gerhard Winkler¹, Angela Kempe², Volkmar Liebscher¹, and Olaf Wittich¹

¹ Institute of Biomathematics and Biometry,
GSF - National Research Center for Environment and Health,
D-85758 Neuherberg/München, Germany

² Graduate Programme Applied Algorithmic Mathematics,
Center for Mathematical Sciences, Munich University of Technology,
D-85747 Garching

Abstract. Typical problems in the analysis of data sets like time-series or images crucially rely on the extraction of primitive features based on segmentation. Variational approaches are a popular and convenient framework in which such problems can be studied. We focus on Potts models as simple nontrivial instances. The discussion proceeds along two data sets from brain mapping and functional genomics.

1 Introduction

The purpose of the present note is twofold: We want to give an elementary introduction to variational approaches to the analysis of one- and multi-dimensional data, and further to illustrate by way of simple data sets and statistical models what we mean by parsimonious statistics.

We will briefly discuss a particularly simple parsimonious approach to the statistical analysis of real-world data sets from life-sciences. Frequently, there is little or no ground truth, and the stochastic mechanism generating (noisy) data is essentially unknown. The only way to associate data to some hidden real event is to verify or falsify rough and basic criteria which characterize the event in question. Such criteria frequently are based on primitive signal features. In images these may be boundaries between regions of different intensity or texture, in time series they may be morphological features like modes or ‘ups and downs’, domains of monotony, or plateaus where the signal is constant. We start the discussion with two one-dimensional data sets, one from brain mapping and one from functional genomics. We expect that in these examples the observation period can be partitioned into intervals where the underlying signal can reasonably be represented by a constant. This is a primitive morphological feature, and the resulting step functions allow sound biological interpretations.

To extract piecewise constant ‘regressions’ from data, we adopt the simplest variational approach based on the Potts model. It is well known to physicists as the straightforward generalization of the Ising model for binary spins to multiple states. For a detailed discussion see Winkler (2003).

2 Two data sets from life sciences

In order to introduce and illustrate the concept, we present two sets of data. The first one consists of time series from functional magnetic resonance imaging (fMRI) of the human brain, and the second one of melting or fractionation curves for spots on a cDNA microchip.

Example 1 (fMRI Brain Data: Identification of Response Regions). The final aim is to identify regions of increased activity in the human brain in response to outer stimuli. Typically such stimuli are boxcar shaped as indicated in Fig. 1. They may represent ‘light or sound on and off’, i.e. visual or acoustic stimuli, or tactile ones like finger tipping on a desk. Functional magnetic resonance imaging (fMRI) exploits the BOLD effect which basically is a change of paramagnetic properties caused by an increase of blood flow in response to the demand of activated neurons for more oxygen. The degradation mecha-

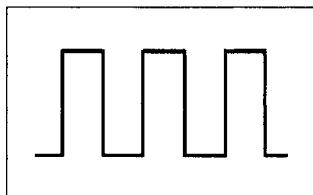


Fig. 1. A box car shaped signal representing ‘on-off’ stimuli in fMRI brain mapping.

nism along the path ‘(complex) eye - (highly complex) brain - (complicated) measuring device’ is only partially known. Moreover, measurement is indirect, since the recorded BOLD effect is a physiological quantity related to increase of blood flow and not a direct function of cortical activation. Hence a parsimonious approach based on significant plateaus should be appropriate.

Example 2 (Fractionation Curves from Gene Expression). The aim of this experiment is to explore the structure of unknown genes. To this end, single stranded sections of *known* cDNA are put on spots of microchips, which typically consist of about 20.000 spots. Each section is a finite sequence of four nucleic acids, which are coded by the letters A(denin), C(ytosin), G(uanin), and T(hymin). If further nucleic acids are added then they tend to bind to the known nucleic acids where T binds to A, and G binds to C. Hence sections of single stranded *unknown* cDNA tend to pair with DNA of similar sequence. The binding energy is maximal for perfect matches like

A	C	T	A	C	A	G	T	A	C	C	A
T	G	A	T	G	T	C	A	T	G	G	T

and such a perfect match means high stability. With perfect match the unknown sequence could be identified perfectly. A main problem is *cross-hybridisation*, which means that DNA sections pair with DNA of similar - but not precisely equal - sequence, for example

A C T A C | A G T A C | C C A
T G A T T | T C A T G | A G T

Perfect match and mismatch are illustrated in Figure 2. A new and innovative experiment provides data which hopefully will allow to identify mismatch dissociation at low stringency. It is called 'Specificity Assessment From Frac-

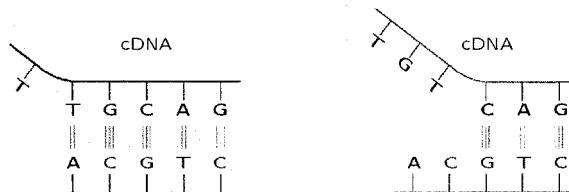


Fig. 2. Specific and unspecific hybridization.

tionation Experiments' or in short-hand notation 'SAFE', see Drobyshev et al.(2003). It is plausible that 'the melting temperature' of double stranded DNA depends on length and contents of specific sequences. It is also plausible that increasing temperature has similar effects as increasing washing stringencies with *formamide* solutions. Both decrease the binding energies and thus cross-hybridisation. This is the basis for the measurement of specific and cross-hybridisation. In the experiment, the chips are washed repeatedly (29 times) with formamide solutions of increasing concentration, and fractionation curves like in Fig. 3 are recorded. The aim of the statistical analysis is

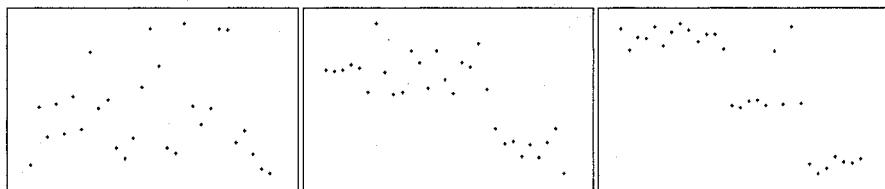


Fig. 3. Typical fractionation curves of single spots: lousy, fairly good, intermediate. (Data from Drobyshev et al. (2003))

to identify locations and heights of abrupt decreases, since they indicate that a certain type of cross-hybridizing cDNA was washed away.

In view of such data, one may doubt about too 'specific' methods or too detailed models for their analysis, and in fact we do so. A way out of this misery is to try a parsimonious approach as indicated above, see Davies (1995). There are attempts like in Davies and Kovac (2001), who adopt the taut string algorithm and its relatives for certain types of data. We tried a parsimonious variational approach called the *Potts Model* for our data.

3 The Potts Model: Rigorous results

The relevant features in Example 1 are successions of high and low plateaus, and in Example 2 the positions of rapid decreases and their height. Therefore we try to fit piecewise constant functions to our data.

The *Potts functional* is defined by

$$x = (x_1, \dots, x_n) \longmapsto P_{\gamma,y}(x) = \gamma|J(x)| + \sum_{k=1}^n (x_k - y_k)^2, \quad (1)$$

where $y = (y_1, \dots, y_n)$ denotes real (fixed) data, and $J(x)$ is the set of time points k where $x_k \neq x_{k+1}$, $k = 1, \dots, n-1$. $|J(x)|$ denotes the cardinality of $J(x)$. The second term rates fidelity of the signal x to data y , and the first one penalizes undesired properties of x . Formally, the functional is a penalized (negative log-) likelihood function.

A minute of contemplation reveals that there are three elementary concepts combined in this model:

- (i) A notion of a ‘jump’ or ‘break’: In the Potts model such a jump is present, where the values of the signal x in two subsequent time points differ from each other.
- (ii) A notion of smoothness: this concerns the behaviour of the signal between two subsequent jumps. It is a consequence of (i) that in the Potts model a signal is constant there.
- (iii) A notion of fidelity to data, i.e. some measure of distance between data y and the signal x .

Note that (ii) is a rather strict notion of smoothness: the signal on a discrete interval is ‘smooth’ only if it is constant. The first term penalizes the number of jumps irrespective of their size and the parameter $\gamma > 0$ controls the degree of smoothness.

Given data y , a ‘filter output’ $T_\gamma(y)$ is defined as a signal which minimizes $P_{\gamma,y}$. In general, it is not unique, but fortunately the following result from the forthcoming thesis Kempe (2003) guarantees uniqueness almost surely:

Theorem 1. *Suppose that the law of data y admits a Lebesgue density. Then for almost all y the functional $P_{\gamma,y}$ in (1) has one and only one minimizer.*

If the hypotheses hold a *filter* is defined uniquely for almost all y by the signal

$$T_\gamma(y) = \operatorname{argmin}_x P_{\gamma,y}(x).$$

We are going now to report essential properties of the filter. It is crucial that the range of hyperparameters γ can be partitioned into intervals, on which the estimate does not change as shown in Kempe (2003). Dependence on hyperparameters is illustrated in Fig. 4.

Theorem 2. For almost all data y the following is true: There are an integer k and hyperparameters $\infty = \gamma_0 > \gamma_1 > \dots > \gamma_k > \gamma_{k+1} = 0$ such that $T_\gamma(y)$ is unique for all $\gamma \in (\gamma_{j+1}, \gamma_j)$. Moreover, it is the same time-series for all $\gamma \in (\gamma_{j+1}, \gamma_j)$. $T_\gamma(y)$ is a constant signal for each $\gamma > \gamma_1$, and $T_\gamma(y) = y$ for $\gamma < \gamma_k$. The number of jumps of $T_\gamma(y)$ on the intervals (γ_{j+1}, γ_j) increases in j . For each $0 < i \leq k$, the functional $P_{\gamma_i, y}$ has precisely the two minimizers belonging to the γ -intervals adjacent to γ_i .



Fig. 4. Intervals on which estimates do not change.

The number of jumps in adjacent intervals may differ by more than one.

Significant simulations can only be carried out with an exact algorithm for the computation of the minimizer. This rules out stochastic algorithms like simulated annealing. For one dimension, an algorithm based on ideas from dynamic programming was presented in Winkler and Liebscher (2002).

Theorem 3. There is an algorithm to compute a minimizer of $P_{\gamma, y}$ in time complexity $O(n^3)$ for all $\gamma \in \mathbb{R}$ simultaneously.

The filter has some more pleasant properties. In particular, the iteration of the filter stops after one step. More precisely, a repeated application returns the same signal as a single one, or in other words the filter is *idempotent* in the sense that $T_\gamma \circ T_\gamma = T_\gamma$. This implies that T_γ is a morphological filter in the sense of Serra (1982, 1988), see Winkler and Liebscher (2002):

Theorem 4. The Potts filter is idempotent.

The filter has continuity or consistency properties like the following one:

Theorem 5 (A. Kempe, V. Liebscher, O. Wittich, unpublished). Let $\gamma \in \mathbb{R}$ and $y^\infty \in \mathbb{R}^n$. Suppose that $T_\gamma(y^\infty)$ is unique. If y^∞ is degraded by random noise ε_n according to $Y^n = y^\infty + \varepsilon_n$, and noise ε_n tends to zero in probability, then $T_\gamma(Y^n)$ tends to $T_\gamma(y^\infty)$ in probability.

We are not interested in a ‘restoration’ but in feature extraction. This is reflected by the theorem since we recover $T_\gamma(y^\infty)$ - and not y^∞ - in the limit.

4 Back to data

Scanning the filter outputs $T_\gamma(y)$ along the hyperparameter γ in view of Theorem 2, is illustrated in the Figures 5 and 6 for the brain and gene data. Visual inspection of the plots reveals clearly what the desired outputs of the method are. On the other hand, we did not yet find an overall satisfying unsupervised method for the identification of an appropriate hyperparameter γ .

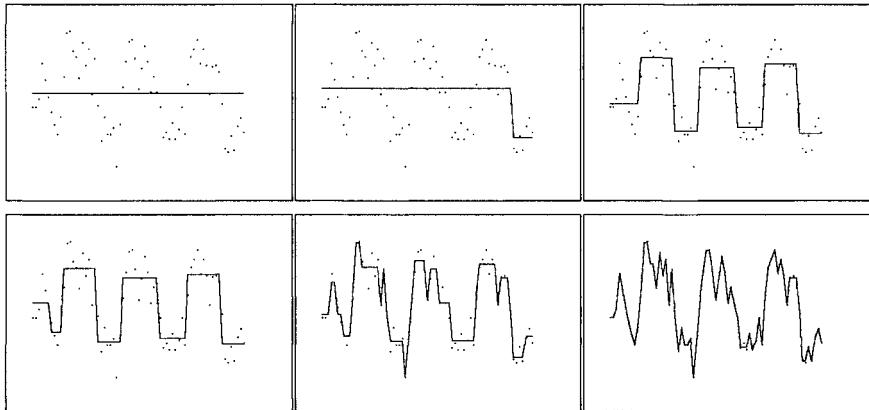


Fig. 5. Some steps of a scan through $T_\gamma(y)$ along decreasing hyperparameters γ for fMRI brain data. Dots indicate data y . Upper right is desired. (Data from D. Auer)

This problem is crucial for such and many similar approaches. For example, there is a host of model selection criteria like the classical ones from Akaike (1974) and Schwarz (1978). Cavanaugh (1997) develops exact criteria which can be adapted to our case, see Kempe (2003). Unfortunately, estimators based on these methods basically return data, as shown in Fig. 7. Therefore we watched out for a criterion which is stable under changes of the hyperparameter γ and of data y . Our first naive idea was to choose $T_{\gamma^*}(y)$ with γ^* from the longest interval of γ -values according to Theorem 2. For the brain data, this simple method outruled the classical criteria. Its results for brain data are contrasted to the classical criteria in Fig. 7.

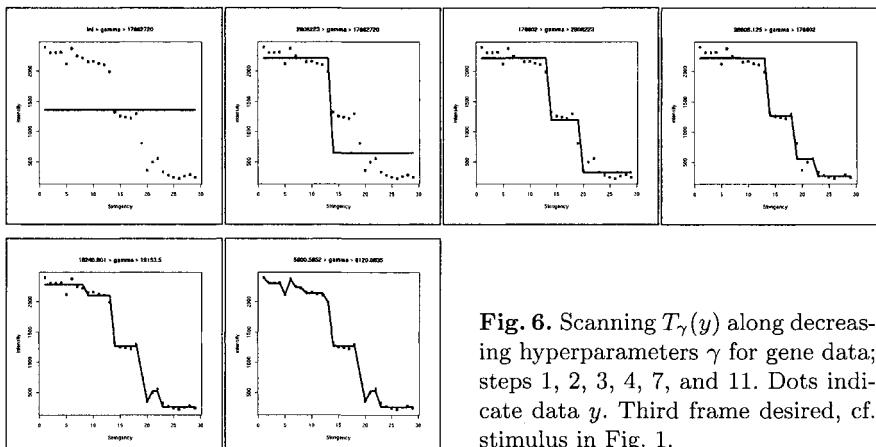


Fig. 6. Scanning $T_\gamma(y)$ along decreasing hyperparameters γ for gene data; steps 1, 2, 3, 4, 7, and 11. Dots indicate data y . Third frame desired, cf. stimulus in Fig. 1.

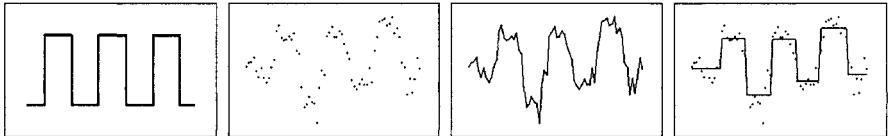


Fig. 7. Stimulus, data, $T_\gamma(y)$ for hyperparameter from Akaike's and Schwarz' information criterion and longest interval criterion: brain data. The latter gives a decent estimate whereas the former basically return data.

For the gene data, we have the additional restriction that the ‘true’ signal should be decreasing. Therefore, we modified our strategy to choose γ from the leftmost γ -interval on which $T_\gamma(y)$ decreases. Fig. 8 displays some of these estimates.

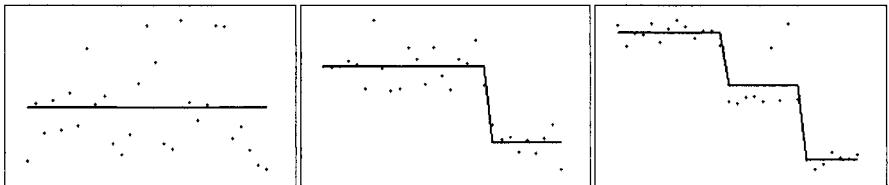


Fig. 8. Estimate for leftmost γ -interval with decreasing $T_\gamma(y)$: gene data.

We obtained partial results about such ‘estimators’ of intervals, but a satisfying rigorous justification is still missing. This is work in progress.

5 Summary and outlook

The Potts functional discussed above is a simple instance of a family of similar functionals. There are modified and more complicated penalties, or other data terms, for example with the sum of squares replaced by absolute deviations. The functionals may live on signals with discrete or continuous time.

For example, the Blake-Zisserman functional

$$x \mapsto BZ(x) = \sum_{k=1}^{n-1} \min\{(x_{k+1} - x_k)^2 / \mu^2, \nu\} + \sum_{k=1}^n (x_k - y_k)^2$$

considers a deviation $\delta = |x_{i+1} - x_i|$ as jump if $\delta > \nu^{1/2}\mu$. Between subsequent jumps it returns a signal which is smooth in the L^2 -sense. A comprehensive treatment is Blake and Zisserman (1987). This functional was constructed as a discrete approximation of the Mumford-Shah functional

$$E_{\mu, \nu, g}(f_S) := \nu|S| + \sum_{k=1}^{n+1} \int_{J_k(S)} (f_k(x) - g(x))^2 + \frac{1}{\mu^2} |f'_k(x)|^2 dx, \quad (2)$$

where time varies over a continuous time interval and the functions f are combined of pieces from functions in Sobolev spaces; it was introduced in Mumford and Shah (1989). A discussion of such functionals in the spirit of the above considerations is work in progress.

Acknowledgement: All simulations were performed by means of the software package ANTS IN FIELDS developed by Friedrich (2003a) in cooperation with the University of Heidelberg. The CD-ROM is attached to Winkler (2003). For a free download see Friedrich (2003b). We are also indebted to J. Beckers, A. Drobyshev, and D. Auer for providing data and introducing us to their subjects.

References

- AKAIKE, H. (1974): A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723.
- BLAKE, A. and ZISSERMAN, A. (1987): *Visual Reconstruction*. The MIT Press Series in Artificial Intelligence, MIT Press, Massachusetts, USA.
- CAVANAUGH, J.E. (1997): Unifying the derivations for the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, 33, 201–208.
- DAVIES, P.L. (1995): Data features. *J. of the Netherlands Society for Statistics and Operations Research*, 49 (2), 185–245.
- DAVIES, P.L. and KOVAC, A. (2001): Local extremes, runs, strings and multiresolution. *Ann. Stat.*, 29, 1–65.
- DROBYSHEV, A.L., MACHKA, CHR., HORSCH, M., SELTMANN, M., LIEBSCHER, V., HRABÉ DE ANGELIS, V., and BECKERS, J. (2003): Specificity assessment from fractionation experiments, (SAFE): a novel method to evaluate microarray probe specificity based on hybridization stringencies. *Nucleic Acids Res.*, 31 (2), 1–10.
- FRIEDRICH, F. (2003a): *Stochastic Simulation and Bayesian Inference for Gibbs fields*. CD-ROM, Springer Verlag, Heidelberg, New York.
- FRIEDRICH, F. (2003b): ANTSINFIELDS: *Stochastic simulation and Bayesian inference for Gibbs fields*, URL: <http://www.AntsInFields.de>.
- KEMPE, A. (2003): *Statistical analysis of the Potts model and applications in biomedical imaging*. Thesis, Institute of Biomathematics and Biometry, National Research Center for Environment and Health Munich, Germany.
- MUMFORD, D. and SHAH, J. (1989): Optimal approximation by piecewise smooth functions and associated variational problems. *Comm. Pure Appl. Math.*, 42, 577–685.
- SCHWARZ, G. (1978): Estimating the dimension of a model. *The Annals of Statistics*, 6 (2), 461–464.
- SERRA, J. (1982, 1988): *Image analysis and mathematical morphology*. Vol. I, II. Acad. Press, London.
- WINKLER, G. (2003): *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods. A Mathematical Introduction*. volume 27 of *Applications of Mathematics*, Springer Verlag, Berlin, Heidelberg, New York, second edition.
- WINKLER, G. and LIEBSCHER, V. (2002): Smoothers for Discontinuous Signals. *J. Nonpar. Statist.*, 14 (1-2), 203–222.

Part V

Marketing, Retailing, and Marketing Research

Application of Discrete Choice Methods in Consumer Preference Analysis

Andrzej Bąk and Aneta Rybicka

Wroclaw University of Economics,
ul. Komandorska 118/120, 53-345 Wrocław, Poland

Abstract. Stated consumer preferences refer to hypothetical market behaviour of consumers. In this case the analytical methods are based on data collected *a priori* by means of surveys to register intentions stated by consumers at the moment of survey taking. The methods used for stated preference analysis include, for instance, discrete choice methods. The general concept of discrete choice methods results from random utility theory. The consumer preference analysis by means of discrete choice methods is based on probability regression models. In this paper a conditional logit model is used to analyse consumer preferences measured against a nominal scale. An example discussed is the result of analysing the preferences of light beer consumers on the basis of a sample of 235 respondents.

1 Introduction

Preference analysis involves the concept of utility, which refers to individual customer's satisfaction from achieving a definite structure of consumption. Since a direct measurement of satisfaction level is not possible, theory of economics applies the concept of consumer preference instead, to quantify utility to some extent.

Consumer preferences denote consumer's ability to evaluate, prioritize and choose goods offered on the market on specific terms. In case the evaluation refers to goods or services of the same class, it may be possible to quantify relations between those products. In theory of economics those relations are referred to as preferences since they provide information on customer's attitudes towards specific products which allows building product hierarchy from the least to the most preferred (ordinal scale measurement) or quantitative evaluation of each product (interval scale measurement). Preference analysis makes it possible to describe and account for consumer behaviour with reference to marketed goods or services.

Consumer preference analysis involves construction of models to reflect actual consumer behaviour as well as methods for measuring (quantifying) consumer preferences (see Zwerina (1997), p. 2). Consumer preference modelling is to clarify the process of consumer behaviour, resulting in evaluation of products offered and finally choosing one of them. The objective of consumer preference measurement is preference quantification, i.e. constructing a scale of measurement that would allow quantitative description of relations between evaluations of individual products. Consumer preference measurement

utilizes track record and projections that describe intentions (declarations) of consumers. Therefore there is a distinction between methods of analysing revealed and stated preferences (see Green and Srinivasan (1990); Zwerina (1997), pp. 2-3;).

Stated consumer preferences refer to hypothetical (declared) market behaviour of consumers. In this case the analytical methods are based on data collected *a priori* by means of surveys to register intentions stated by consumers at the moment of survey taking. The methods used for stated preference analysis include, for instance, discrete choice methods. The general concept of discrete choice methods results from random utility theory. The process of selecting the profiles is of probabilistic nature, as the behaviour of goods or services consumers is not always consistent and predictable. This means that – under identical conditions and from an identical set of options – consumer choices may differ with time.

The consumer preference analysis by means of discrete choice methods is based on probability regression models. In this paper a conditional logit model is used to analyse consumer preferences measured against a nominal scale. An example discussed is the result of analysing the preferences of light beer consumers on the basis of a sample of 235 respondents.

2 Discrete choice method

2.1 Data collection

Within the method based on discrete choice analysis, the respondent can make the choice of profile that seems most attractive to him/her. The respondent can also refrain from making any choice if none of the suggested profiles meets his/her expectations. During the data preparation process, one-step or two-step procedures of generating factorial designs that may be symmetrical (identical number of levels for each attribute) or asymmetrical (different number of levels for each attribute), while the generated option subsets may include fixed or variable number of profiles.

One-step procedures involve fractional-factorial designs of L^{MP} type, where L denotes the number of levels assigned to each attribute, M is the total number of attributes, and P is the number of profiles within each subset. This is a typical symmetrical design for $M \times P$ factors, each of which has L levels. It is a one-step procedure since the generated design is composed of rows (subsets), each of which includes P options and each option is defined with M attributes. The number of subsets results from the adopted plan of total experiment reduction (full factorial design) of L^{MP} size.

With two-step procedures, the fractional-factorial designs are applied twice to arrive at option subsets. First, the set of profiles for respondents' choice is generated. Then the profiles are divided into subsets. The respondent chooses one profile from each subset (or makes no choice). Selected two-step procedures are presented, for instance, in Zwerina (1997), p. 55.

2.2 Model estimation

Factorial design makes it possible to prepare profiles used in respondent preference measurement. The collected data are used for parameter estimation with models reflecting relations between profile evaluation and the values of attributes that characterize them. These relations are of regressive nature, but the specific models applied may vary. The basic determiner of the regression model type and of the range of applicable methods of parameter estimation is the type of scale used for preference measurement; this in turn results from the character of research problem and from the method of preference data collection. With empirical studies on stated consumer preferences, the most frequently used models are multiple regression models (in traditional conjoint analysis) and multinomial (conditional) logit models (in choice analysis).

In choice analysis, part-worths of attribute levels are estimated on the basis of conditional logit models (see McFadden (1974)) that are applicable when attributes characterize profiles of products to be chosen by the respondents. The respondents are presented with an offer in the form of profile subsets. A respondent can select his preferred profile from each subset or decide not to make a choice. This decision process is described by the theory of random utility which says that profile choice is determined by profile utility. However, the utility is not fully determined as it is influenced not only by systematic but also by random terms.

The model of probability of i -th respondent choosing the j -th profile represents an optimization problem, since the buyer tries to maximize his profits from choosing a specific product or service by applying his/her own decision-making principles. The probability of the i -th respondent choosing the j -th profile from a subset of n elements takes the form of a conditional logit model (see Louviere and Woodworth (1983), pp. 352-355; Haaijer and Wedel (2000), p. 335; Kuhfeld (2001), p. 72):

$$p_{ij} = \frac{\exp(\mathbf{x}_{ij}^T \boldsymbol{\beta})}{\sum_{k=1}^n \exp(\mathbf{x}_{ik}^T \boldsymbol{\beta})} = \frac{1}{\sum_{k=1}^n \exp[(\mathbf{x}_{ik}^T - \mathbf{x}_{ij}^T) \boldsymbol{\beta}]}, \quad (1)$$

where: \mathbf{x}_i – vector of attributes characterizing the j -th profile as perceived by the i -th respondent; $\boldsymbol{\beta}$ – vector of part-worths of attribute levels.

The parameters of a conditional logit model are estimated by means of the maximum likelihood method or alternative least squares method. A conditional logit model is frequently used in procedures of discrete choice modelling for estimation of probability of choosing the j -th profile from a subset of n elements. Therefore this model type is widely used for measuring stated consumer preferences.

2.3 Available software

The SAS/STAT package includes applications to support preference analysis performed both as a conjoint analysis and as a discrete choice analysis. The

instructions included in the package support data collection and part-worth utility estimation. At the data collection step, the applications generating full and fractional factorial designs – PROC PLAN, PROC FACTEX and PROC OPTEX – are especially useful with preference modelling. At the part-worth estimation step, the most frequently used applications are PROC TRANSREG and PROC PHREG.

With stated consumer preference analysis, the SAS/STAT applications can support the following tasks (see Kuhfeld (2001)): generating full factorial design, generating fractional orthogonal factorial design for traditional conjoint analysis, generating efficient fractional factorial design for discrete choice analysis, attribute coding with dummy variables, parameter estimation for metric models in conjoint analysis (strong scale of preference measurement), parameter estimation for non-metric models in conjoint analysis (weak scale of preference measurement), simulation of market share, parameter estimation for discrete choice models (nominal scale of preference measurement).

3 An example of discrete choice method application

3.1 Data collection and model estimation

An example of discrete choice method application is the study on preferences of light beer consumers carried out in July, 2002 on a sample of populations of Jelenia Góra and Lwówek Śląski. Data were collected especially for this work using questionnaires (235 correctly filled in forms).

Since the early 1990s production and consumption of beer in Poland have been continuously growing (see Table 1). Therefore the beer market in Poland has been dynamically developing. As a result, the structure of beverage consumption has changed specifically in favour of beer. The supply is predominated by beers produced in Poland, i.e. Polish brands or foreign brands made in Polish breweries. The share of imported brands is negligible. It is result of consumer preference structure. Discrete choice methods are useful to discover it. These methods are relatively new and not popular in marketing practice in Poland.

Indicator	1994	1995	1996	1997	1998	1999	2000
Production	1.390	1.600	1.670	1.900	2.102	2.336	2.489
Exports	0.007	0.017	0.017	0.012	0.011	0.012	0.010
Imports	0.004	0.005	0.008	0.013	0.017	0.030	0.030
Supply	1.387	1.588	1.661	1.901	2.108	2.354	2.500

Table 1. Beer market in Poland (in billions of litres) (Source: Kupczyk, A., Current state and perspectives of the beverage market in Poland. URL: <http://www.agrotech.mediator.pl/bmp/>)

The objective of study was to identify determinants of consumers' choice of specific types and brands of beer. The analysis covered attributes and levels presented in Table 2. The outcome was a factorial design of L^{MP} type (where L denotes the number of levels assigned to each attribute, M is the total number of attributes, and P is the number of profiles within each subset) and $3^{4 \times 5}$ size. The full factorial design was reduced by means of iterative Fedorov algorithm that makes it possible to construct an optimal non-orthogonal factorial design on the basis of a predetermined candidate set (see details in Table 3).

Attribute	Levels
Country of origin	Poland, Germany, Czech Republic, Holland, Denmark
Price range	Up to 2.00 PLZ, 2.00-4.00 PLZ, above 4.00 PLZ
Alcohol content	Up to 1.0%, 1.8-5.0%, above 5.0%
Packaging type	Bottle, can, mug
Packaging volume	0.33 l, 0.5 l, above 0.5 l

Table 2. Attributes and levels characterizing light beer

Characteristic	Description
Factorial experiment	<ul style="list-style-type: none"> - full factorial design: $3^{4 \times 5}$ subsets (20 variables, 3 levels each) - candidate data set: 19,683 subsets - minimum number of subsets : $20 * (3 - 1) + 1 = 41$ - adopted for study: 45 subsets - efficiency of factorial design: \$D=85.96\$
Study structure	<ul style="list-style-type: none"> - number of blocks: 3 - number of subsets per block: 15 - number of profiles per subset: 6 (5 plus 'no choice' option)
Number of questionnaires	<ul style="list-style-type: none"> - distributed: 300 - filled in: 235

Table 3. Detailed characteristics of the study

Each respondent was to evaluate 15 subsets comprising 6 profiles each (5 profiles representing different levels of beer attributes, plus the 'no choice' option). As a result, 235 properly filled-in questionnaires were returned, with the following block spread: Block 1 – 75, Block 2 – 78, Block 3 – 82. The total number of data entries thus collected was 21,150 ($15 \times 6 \times 235$).

Through estimation of the conditional logit model, parameter estimates were obtained as presented in Table 4.

Variable Label	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Country of origin						
Poland	1	-0.19009	0.08048	5.5779	0.0182	0.827
Germany	1	-0.84030	0.08626	94.8857	< .0001	0.432
Czech Republic	1	-0.66955	0.08359	64.1529	< .0001	0.512
Denmark	1	-0.96916	0.09008	115.7588	< .0001	0.379
Holland	1	-1.37733	0.09538	208.5386	< .0001	0.252
none	0	0
Price range						
none	0	0
2-4 PLN	1	0.35644	0.05740	38.5609	< .0001	1.428
up to 2 PLN	1	0.50510	0.05521	83.7052	< .0001	1.657
above 4 PLN	0	0
Alcohol content						
none	0	0
1,8-4,5%	1	-0.02562	0.05012	0.2614	0.6092	0.975
up to 1,0%	1	-0.85350	0.05995	202.7177	< .0001	0.426
above 5,0%	0	0
Packaging type						
none	0	0
bottle	1	0.18360	0.05470	11.2653	0.0008	1.202
mug	1	0.25960	0.05625	21.3002	< .0001	1.296
can	0	0
Packaging volume						
none	0	0
0,33 l	1	-0.31304	0.05465	32.8142	< .0001	0.731
0,5 l	1	0.00462	0.05182	0.0080	0.9289	1.005
above 0,5 l	0	0

Table 4. Model parameters

The estimated part-worth utilities of individual attribute levels were then used to calculate total worth of each profile utility. Next, the likelihood of profile choice was calculated for each subset and for the whole set. The results, in the form of five profiles of highest choice probabilities and another five of lowest choice probabilities, selected out of the total 270 (15 subsets \times 6 profiles \times 3 blocks), are presented in Table 5.

3.2 Hazard ratio values

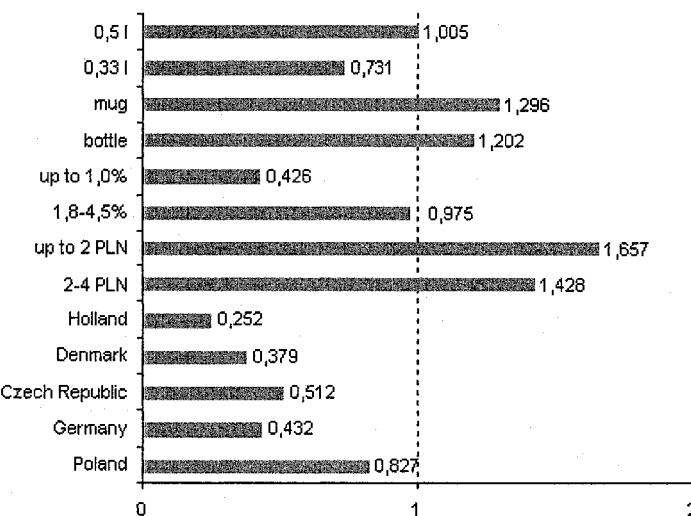
Interpretation of the estimated parameters of a conditional logit model is made easier by the hazard ratio that indicates a positive or negative impact of the estimate on a profile choice likelihood. Hazard ratio values are computed by exponentiating the parameter estimates ($\exp(\beta_l)$). If $\exp(\beta_l) > 1$ then attribute level impact for an event (choice) is positive. If $\exp(\beta_l) < 1$ then

Country of origin	Price range	Alcohol content	Packaging type	Packaging volume	Probability
Highest probabilities of choice					
Poland	up to 2 PLN	above 5.0%	mug	above 0.5 l	0.0104
Poland	up to 2 PLN	1.8-4.5%	mug	above 0.5 l	0.0102
Poland	up to 2 PLN	above 5.0%	bottle	0.5 l	0.0097
Poland	up to 2 PLN	1.8-4.5%	bottle	above 0.5 l	0.0094
Poland	2-4 PLN	above 5.0%	mug	above 0.5 l	0.0090
Lowest probabilities of choice					
Holland	above 4 PLN	up to 1.0%	bottle	above 0.5 l	0.0008
Denmark	above 4 PLN	up to 1.0%	can	0.33 l	0.0007
Holland	2-4 PLN	up to 1.0%	can	0.33 l	0.0007
Holland	above 4 PLN	up to 1.0%	can	above 0.5 l	0.0006
Holland	above 4 PLN	up to 1.0%	bottle	0.33 l	0.0006

Table 5. Selected values of choice probabilities

attribute level impact for an event is negative. And if $\exp(\beta_l) = 1$ then attribute level impact for an event is neutral.

Figure 1 presents hazard ratio values computed for parameters whose values are different from zero ($\exp(\beta_l) \neq 1$). A hazard ratio value above 1 indicates a positive parameter impact, while hazard ratio value under 1 indicates a negative parameter impact.

**Fig. 1.** Hazard ratio values for parameter values different from zero

4 Conclusions

Discrete choice methods can be applied for the analysis of results of stated preference measurement. These methods allow simulation of actual market choices. At the data collection stage of preference measurement, optimal factorial designs can be used. At the analysis stage conditional logit models can be used for estimation part-worths of attributes levels.

Computer software supports the data preparation stage (construction of factorial design, preparing profile subsets, and generation of simulation data), and the discrete choice model estimation stage (estimation of part-worths and total utility, estimation of choice probabilities).

Consumption of nationally-brewed beer dominates on the Polish beer market. Choice of beer profile is determined (in the order of importance) by: price range, packaging type, packaging volume, alcohol content, and country of origin. Choice probabilities is positively stimulated by: price range up to 2 PLN, price range of 2-4 PLN, packaging type – mug, packaging type – bottle, packaging volume – 0.5 l.

Presented analysis can be improved by segmentation of the respondents. It can be done using latent class discrete choice models.

Acknowledgements: The research presented in the paper was partly supported by the project KBN 5 H02B 030 21.

References

- GREEN, P.E. and SRINIVASAN, V. (1990): Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice. *Journal of Marketing, October*, 54, 3-19.
- HAAIJER, R. and WEDEL, M. (2000): Conjoint Choice Experiments: General Characteristics and Alternative Model Specifications. In: A. Gustafsson, A. Herrmann and F. Huber (Eds.): *Conjoint Measurement: Methods and Applications*. Springer, Berlin, 319-360.
- KUHFELD, W.F. (2001): *Multinomial Logit, Discrete Choice Modeling*. URL: <http://ftp.sas.com/techsup/download/technote/ts643.pdf>, SAS Institute.
- LOUVIERE, J.J. and WOODWORTH, G. (1983): Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data. *Journal of Marketing Research, November*, 20, 350-367.
- MCFADDEN, D. (1974): Conditional Logit Analysis of Qualitative Choice Behavior. In: P. Zarembka (Ed.): *Frontiers in Econometrics*. Academic Press, New York-San Francisco-London, 105-142.
- ZWERINA, K. (1997): *Discrete Choice Experiments in Marketing*. Heidelberg-New York, Physica-Verlag.

Competition Analysis in Marketing Using Rank Ordered Data

Reinhold Decker and Antonia Hermelbracht

Department of Economics and Business Administration,
University of Bielefeld, D-33615 Bielefeld, Germany

Abstract. A new approach to evaluating competition between different products of one category using individual rankings is introduced. The approach is based on the principle of ordinal regression and focusses on the determination of the sources of existing market shares. To account for this the success of a particular product is put down to the consumers' perception of relevant attributes. The functionality of our approach is demonstrated by applying it to an empirical data set from the car industry.

1 Introduction and motivation

Sufficient familiarity with the main strengths and weaknesses of competitors is an elementary prerequisite for any promising marketing strategy. Marketing databases which are generated, for example, from commercial retail or household panels usually contain sales and marketing mix data. But the suitability of such data for explaining individual purchase decisions is limited. Furthermore, data of this kind are primarily available for consumer goods for everyday use. If the marketing management is interested in more than just the "influence of price and promotion" on product sales having its own survey with a special focus on individual consumer preferences becomes indispensable. In this case the subjective perception and assessment of the products under consideration have to be investigated. On the other hand, both perception and assessment depend on relevant product attributes. In the following we are going to use the term "attribute" in a more general way, which goes beyond the usual term "marketing variable".

Up to now, numerous models and methods have been suggested in the literature to measure the effect of individual attributes on the overall utility of the respective products. Conjoint analysis in particular provides a comprehensive set of methods that can be used to ascertain the way and the extent to which relevant attributes contribute to individual preferences. But, if these preferences are measured with the widespread full profile method, it might happen that some respondents feel unable to give a holistic judgement because this implies that multiple attributes are taken into account simultaneously. This, for example, means that a respondent who prefers product A to product B with respect to attribute I and B to A with respect to attribute

II will hardly be able to submit a valid overall judgement. If this appears with a certain frequency the resulting partworth utilities may be crucially biased. Besides, the handling of "large" numbers of attributes provokes problems with some of the conventional methods. Particularly traditional conjoint analysis works efficiently only when the number of attributes doesn't exceed six (cf. Albrecht (2000, p. 33), Green and Srinivasan (1978, p. 108)).

If the ranking of products is done for each attribute separately the risk of overburdening the respondents is less probable. Attribute-wise product rankings facilitate the consideration of a comparatively large number of attributes. Therefore, the ranking we are starting from in the following looks like this: each respondent is asked to order all the products of his evoked set in such a way that the most preferable one with respect to attribute 1 gets rank 1, the second best gets rank 2 and so on. Then the whole ranking is repeated with respect to attribute 2. The ranking terminates when all attributes have been considered.

The rest of this article is organized as follows: in section 2 we sketch the basic methodical elements of our approach. This is followed by the description of an empirical application to a data set from the car industry in section 3. We conclude with a few critical remarks and some suggestions for future research.

2 Methodical foundation

The following methodical considerations start from four assumptions which, among other things, simplify the mathematical formalism:

1. The independent explanatory variables (the rankings) included in the analysis are ordinally scaled whereas the dependent response variable (the individual choice probability) is metric.
2. The maximum number of ranks is equal to the number of products to be characterized.
3. The product rankings of each attribute have a monotone descending ordering where the highest rank (rank 1) has the strongest effect on the response variable and vice versa.
4. There is no information about inter-rank distances; in particular, equidistance is not automatically assumed.

The last assumption represents an important advantage of this approach over other models which have similar objectives but require equidistance.

Starting from the above assumptions and referring to corresponding considerations of Hilbert (1996), a transformation of the available ranks into binary values is possible. With reference to Toutenburg (1992, p. 253) and

Gifi (1990, p. 67) the following dummy coding is selected (alternatively, the effect coding, the Helmert coding, or the split coding (cf. Tutz (2000, p. 18)) might be applied):

$$x_{ilmk} = \begin{cases} 1 & \text{if product } i \text{ is assigned rank } m \text{ regarding} \\ & \text{attribute } l \text{ by respondent } k \\ 0 & \text{otherwise} \end{cases} \quad \forall i, l, m, k$$

with $i = 1, \dots, I$ = number of products to be evaluated, $k = 1, \dots, K$ = number of observations or respondents, $l = 1, \dots, L$ = number of attributes to be taken into account, and $m = 1, \dots, M$ = number of rank categories.

In contrast to Toutenburg (1992) we do not only take $M - 1$ but also M dummy variables into consideration, since the last category cannot be captured implicitly by the others. This is due to the fact that each respondent is just ranking those products that are included in his/her individual evoked set. Therefore, the last category cannot be neglected as it includes the information whether a product is in the evoked set or not.

The availability of individual rank ordered data put us in a position to compute relative weights b_{il} for each product and attribute which will be used for model calibration later on. First, parameters $a_{(il)m}$ ensuring the descending ordering have to be specified by the market researcher. The brackets indicate that both a specification restricted exclusively to ranks and one regarding products and attributes as well as ranks is imaginable. In the following empirical study we exemplarily define $a_1 = 2^{M-1}$, $a_2 = 2^{M-2}$, ..., $a_M = 2^0$ to explicitly abstract from equidistance. Naturally, other definitions with the same focus are possible as well (e.g., $a_1 = M$, $a_2 = M - 1$, ..., $a_M = 1$ in the case of equidistance). Further on we define:

$$b_{il} = \frac{1}{K} \cdot \sum_{k=1}^K \frac{\sum_{m=1}^M a_{(il)m} \cdot x_{ilmk}}{\sum_{m=1}^M a_{(il)m}} \quad \forall i, l.$$

The real individual choice probabilities can be approximated in different ways. On the one hand the respondents may be asked to allocate a constant sum of points to the available products to express their individual preferences ("constant sum method"). In this case the relative shares of points serve as an approximation p_{ik}^{appC} of the real choice probabilities. On the other hand the unknown choice probabilities can also be approximated by the following transformation, which explicitly implies that individual preferences are directly related to the attribute-wise rankings of products:

$$p_{ik}^{appT} = \frac{\sum_{l=1}^L \sum_{m=1}^M a_{(il)m} \cdot x_{ilmk}}{\sum_{j=1}^I \sum_{l=1}^L \sum_{m=1}^M a_{(jl)m} \cdot x_{jlmk}} \quad \forall i, k.$$

The transformation formula can be applied if constant sum values are not available and if one is interested in a kind of benchmark for the fit the model can attain at best. Both methods provide an approximation of the real market shares if not available. This problem appears especially when the study concerned is directed at durable goods where valid data on the item level is frequently not accessible. Instead of this, statistical descriptions of such markets are mostly restricted to the aggregated brand level. For the car market, which we are focussing on in the empirical part of the paper, this unfortunately applies as well. The corresponding formulas look like this (presuming an adequate sampling scheme):

$$MA_i^{appC} = \frac{1}{K} \cdot \sum_{k=1}^K p_{ik}^{appC} \quad \text{or} \quad MA_i^{appT} = \frac{1}{K} \cdot \sum_{k=1}^K p_{ik}^{appT} \quad \forall i.$$

Strictly speaking the above measures are not market shares but aggregated relative preferences which are assumed to be appropriate approximations of the unknown data.

The strength of the interrelation between binary variable x_{ilmk} and individual choice probability p_{ik} is expressed by means of parameters $\tilde{\eta}_{ilm} \geq 0$ $\forall i, l, m$ which have to be estimated from the empirical data. In this way we explicitly take into account that each product which was evaluated (and therefore belongs to the individual evoked set of the respondent concerned) has a positive probability of being chosen. Consequently, all those products which are not included in the individual evoked set have choice probabilities equal to zero. Because of the monotonicity assumption it must apply: $\tilde{\eta}_{il,m-1} \geq \tilde{\eta}_{ilm} \forall i, l, m > 1$. To ensure this and referring to assumption 4 as well we define:

$$\begin{aligned} \tilde{\eta}_{ilM} &= \exp(\eta_{ilM}), \\ \tilde{\eta}_{il,M-1} &= \exp(\eta_{ilM}) + \exp(\eta_{il,M-1}) = \tilde{\eta}_{ilM} + \exp(\eta_{il,M-1}), \\ &\vdots \\ \tilde{\eta}_{il1} &= \exp(\eta_{ilM}) + \cdots + \exp(\eta_{il1}) = \tilde{\eta}_{il2} + \exp(\eta_{il1}) \quad \forall i, l. \end{aligned}$$

Thus the prerequisites for a simple model linking attribute-wise rankings and individual choice probabilities are given:

$$p_{ik} = \frac{\prod_{l=1}^L \sum_{m=1}^M \tilde{\eta}_{ilm} \cdot x_{ilmk}}{\sum_{j=1}^I \prod_{l=1}^L \sum_{m=1}^M \tilde{\eta}_{jlm} \cdot x_{jlmk}} \quad \forall i, k.$$

To estimate the unknown parameters $\eta_{111}, \dots, \eta_{ilm}, \dots, \eta_{ILM}$ the following objective function proves to be suited (MA_i and θ_{il} will be defined in a moment):

$$\min(f_1 + f_2 + f_3),$$

where

$$\begin{aligned} f_1 &= \frac{1}{K \cdot I} \cdot \sum_{k=1}^K \sum_{i=1}^I \frac{(p_{ik} - p_{ik}^{app(C/T)})^2}{\sqrt{p_{ik} \cdot p_{ik}^{app(C/T)}}} \\ f_2 &= \frac{1}{I} \cdot \sum_{i=1}^I \frac{(MA_i - MA_i^{app(C/T)})^2}{\sqrt{MA_i \cdot MA_i^{app(C/T)}}} \\ f_3 &= \frac{1}{I \cdot L} \cdot \sum_{i=1}^I \sum_{l=1}^L \frac{(\theta_{il} - b_{il})^2}{\sqrt{\theta_{il} \cdot b_{il}}}. \end{aligned}$$

The special structure of this multi-objective function (for multi-objective optimization see, e.g., Deb (2001)) allows for simultaneously taking into account the disaggregated as well as the aggregated data level and the interdependency between both. The resulting estimates can not only be used to explain the observed preferences but also – and this is still much more interesting in competition analysis – to predict consumer choices and to perform response simulations. The latter one is an important tool for assessing the sales potentials of new products and requires sufficient scope for generalization.

With $\theta_{il} = \sum_{m=1}^M \tilde{\eta}_{ilm} \forall i, l$ the originally ordinal scale of the data can be raised to the metric level, analogously to the additive approach of conjoint analysis. Parameter θ_{il} is interpretable as the contribution of attribute l to the total assessment of product i : the larger this value, the higher this contribution. The attribute contributions, for their part, may be used to estimate the market shares of the relevant products:

$$MA_i = \frac{\sum_{l=1}^L \theta_{il}}{\sum_{j=1}^I \sum_{l=1}^L \theta_{jl}} \quad \forall i.$$

We are now able to compare the products under consideration regarding the influence of individual attributes on the respective market shares. Thus the inequality relation $\theta_{2l} > \theta_{3l} > \theta_{1l}$, for instance, indicates a superiority of product 2 to product 3 and 1 with respect to attribute l . Further, due to the metric level of these parameters, it can be stated, for instance, that product 2 is twice as "good" as product 1 regarding attribute l . Comparisons of this kind prove to be useful for a deeper understanding of competitive advantages in established as well as developing consumer markets.

3 Empirical example

To verify the theoretical considerations above we applied our approach to rank orderings obtained from a recent survey at the University of Bielefeld. The underlying questionnaire was filled in mainly by students in the Department of Economics and Business Administration. Besides other aspects,

each respondent was asked to rank four automobiles by means of three different attributes. The four competitive products (items) were: "Audi A4", "BMW 3-er", "Mercedes C-Klasse", and "Volkswagen Passat". Using Auto Strassenverkehr (2002) as our guideline, we selected the attributes "traveling comfort/handling", "safety/chassis", and "price level/costs". Due to the simplicity and brevity of our survey it was possible to use the data of 246 correctly filled in questionnaires. The implementation of our approach in SAS was done with the NLP procedure (using general nonlinear optimization with the Newton-Raphson method). Due to the unavailability of the automobiles' real market shares (calculated from observed sales volumes on the item-level) we had to be satisfied with appropriate approximations.

In Table 1 both the model-based estimations of market shares and the approximated initial values are depicted. On the whole, and as expected, all products have similar market shares (or rather relative preference values), whereby the Mercedes leads slightly. The estimations reveal themselves to be very good if the approximation of the initial values is done with the transformation formula whereas the reproduction of market shares does not succeed to this extent if the approximation is done by means of the constant sum method.

	MA_i^{appT}	\widehat{MA}_i^{appT}	MA_i^{appC}	\widehat{MA}_i^{appC}
"Audi A4"	24.61 %	24.67 %	26.64 %	24.85 %
"BMW 3-er"	25.85 %	25.84 %	29.74 %	26.10 %
"Mercedes C-Klasse"	27.49 %	27.51 %	27.24 %	27.51 %
"VW Passat"	22.05 %	21.97 %	16.38 %	21.54 %

Table 1. Approximated and estimated market shares

The same (regarding the goodness of fit) applies to the estimation of the individual choice probabilities as well. In Figure 1 the corresponding results for the BMW, using the data which we acquired by means of the transformation formula, are displayed exemplarily. The visibly good fit finds its expression in a mean squared error equal to 0,00186 and a weighted mean squared error equal to 0,00725. The goodness of fit concerning the remaining products is very similar. It was possible to ensure statistical significance for 40 out of 48 model parameters.

To test the predictive power of our approach the available sample was repeatedly divided into two sub-samples. The first one was used to estimate the relevant parameters whereas the second one serves for validation. The whole procedure was repeated for different divisions of the data set. Each time the real choices of the testing group could be predicted adequately. The prediction of the choices of 10 arbitrarily selected respondents concerning the

BMW is depicted in Figure 2.

The main reason for different model adjustments (transformation formula versus constant sum method) lies in the depth of our questionnaire. Due to the experimental nature of the data collection only three attributes (frequently emphasized in car magazines like Auto Strassenverkehr (2002)) have been taken into account. However, some of the respondents seemingly included further aspects (attributes) in their assessment by means of the constant sum method.

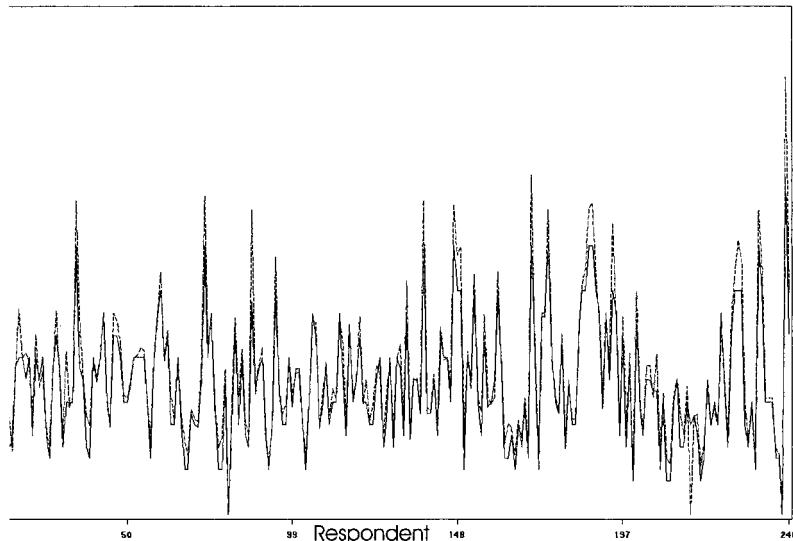


Fig. 1. Approximated (—) and estimated (- -) individual choice probabilities for the BMW (transformation formula)

In product management practice the decision makers are especially interested in a reliable explanation of the causes of existent market shares or rather relative preferences. With this in mind, the estimated attribute contributions to the individual market shares are depicted in Table 2 (based on the transformation formula approach).

The Mercedes obviously owes its market share primarily to its outstanding travelling comfort and safety facilities, whereas the price level contributes little. As regards the VW Passat we have exactly the reverse situation, while Audi and BMW hold a position somewhere between these two extremes. In the case of BMW a certain accent on travelling comfort and safety can be stated at most. All in all the estimated attribute contributions on hand are not only easy to interpret, but also intuitively plausible.

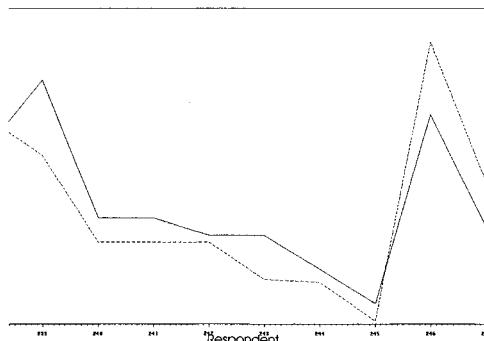


Fig. 2. Predicted (-) and stated (- -) individual choice probabilities for the BMW (constant sum method)

	Trav. comfort/ handling	Safety/ chassis	Price level/ costs	\widehat{MA}_i^{appT}
"Audi A4"	0.26266	0.23185	0.26077	24.67 %
"BMW 3-er"	0.30662	0.28846	0.19601	25.84 %
"Mercedes C-Klasse"	0.32398	0.37354	0.14471	27.51 %
"VW Passat"	0.13321	0.13158	0.40785	21.97 %

Table 2. Estimated attribute contributions

4 Discussion

The suggested approach enables the explanation of market shares of durable goods in a competitive environment and the prediction of future preferences regarding these products. To be able to successfully calibrate the model a data set of adequate size is required, i.e. a sufficient number of observations must be available for each attribute. In this context special attention should also be paid to the appropriate selection of relevant attributes. The absence of attributes which strongly determine individual preferences may bias the resulting estimates. On the other hand the number of observations required increases above average with the number of attributes concerned. For this reason it is of crucial importance to determine the correct attribute set exactly. But a bad approximation can also be an indicator of inconsistencies within the response behavior of certain respondents.

Future research might be concentrated particularly on two points. On the one hand further experience with different product categories seems to be necessary in order to gain deeper knowledge concerning the behavior of the model in practice. On the other hand an appropriate consideration of consumer heterogeneity is a methodically highly interesting and demanding challenge. A promising approach in this direction might be the use of suitable

heterogeneity distributions, which have proven to be an elegant solution for this problem in stochastic modelling of buying behavior.

References

- ALBRECHT, J. (2000): *Präferenzstrukturmessung*. Lang, Frankfurt.
- AUTO STRASSENVERKEHR (2002): Die schönsten Autos 2002. *Auto Strassenverkehr, Heft 8, 32–38*.
- DEB, K. (2001): *Multi-Objective Optimization using Evolutionary Algorithms*. Wiley, Chichester.
- GIFI, A. (1990): *Nonlinear Multivariate Analysis*. Wiley, Chichester.
- GREEN, P. and SRINIVASAN, V. (1978): Conjoint Analysis in Consumer Research: Issues and Outlook. *Journal of Consumer Research, 5, 103–123*.
- HILBERT, A. (1996): *Regressionsansätze zur Analyse ordinaler Daten*. Arbeitspapier 143/1996, Institut für Statistik und Mathematische Wirtschaftstheorie, Universität Augsburg.
- TOUTENBURG, H. (1992): *Lineare Modelle*. Physika, Heidelberg.
- TUTZ, G. (2000): *Die Analyse kategorialer Daten*. Oldenbourg, München.

Handling Missing Values in Marketing Research Using SOM

Mariusz Grabowski

Department of Computer Science,
Cracow University of Economics, ul. Rakowicka 27, 31-510 Kraków, Poland

Abstract. Many fields of research suffer from incomplete data. In marketing the problem of incomplete information is particularly important as data losses occur quite frequently. In practice, researchers using various methods deal with this problem in many less or more satisfactory ways. This paper validates the use of SOM (Self-Organizing Map) in estimating missing data in a marketing field and refers to another non-trivial method of handling missing values called expectation maximization (EM).

1 Problem of incomplete data

Incomplete data constitute a serious difficulty in many fields of research. In certain disciplines, e.g. technical, replacements for missing data can be easily obtained by repetition of the experiment, in other, e.g. socio-economic, this is not so obvious and in the majority of cases simply impossible (Kordos (1988)).

The problem of incomplete information is particularly important in marketing (Pociecha (1996)), where gaps occur in gathered statistical material quite frequently. In many cases data losses across the variables may exceed 30%. Such situation, especially when data losses are distributed across the variables, may even make further analysis of the data file impossible.

In general missing data fall into two categories (Pawełek (1996)): *cross-sectional* where as incomplete we usually consider empty components of data vectors, and *time series* where we observe empty places in corresponding time-sequences of the given variables. Marketing data usually do not concern time series and for this reason in this paper we will concentrate on structural data.

Little and Rubin (1987) define three types of assumptions that help to classify missing data as regards the possibility of their estimation:

- In case of the data *missing completely at random (MCAR)*, the mechanism causing data missingness is not related to the rest of the variables in a data set. In this case it is possible to impute missing values by the mean of observed values of the given variable.
- When data are *missing at random (MAR)*, variables containing the missing data are in a functional relationship with the rest of observed variables within a data set. However, the mechanism of data missingness does not depend on the missing data themselves. The missing values are estimable on the basis of the rest of the available data.

- The last case, *nonignorable*, occurs when data are not missing at random. The missingness mechanism depends on the values of the variable that contains missing data. This implies that some variables explaining missingness are not included in the data set.

The trivial missing data handling techniques, including variable deletion or listwise and pairwise data deletion, discard the information contained in rejected portions of data. Although the loss of information connected with even a partial lack of data is inevitable, it seems important to develop methods estimating the missing data that would aim at preserving as much of the information from the data space as possible. When estimating the missing data, it is desirable to maintain the distribution of the data file as close as possible. For this reason there is a number of methods that perform this task in a less or more satisfactory way. Most of them impute a multivariate normal distribution and MAR assumption.

More advanced missing data handling techniques include among others: regression methods, clustering or taxonomy methods, expectation maximization, raw maximum likelihood and multiple imputation. An extensive discussion may be found in Little and Rubin (1987) and Woethke (2000) where it is indicated that raw maximum likelihood and multiple imputation methods are superior to casewise or pairwise deletion and mean substitution.

Below a short description of some data estimation methods related to the subject of this paper, SOM and EM, is submitted.

In *regression methods* the idea for estimating missing values is based on the premise that variables containing the missing information are in certain functional dependency from the rest of the variables. Based on this premise, the regression model is built and the type of the relationship is imputed. Then the missing values are estimated.

Clustering or taxonomy methods use the idea of similarities between cases. The most successful method in this group is *hot-deck imputation*. The idea is to find the most similar case to the one including missing data and substitute them with the present ones from this particular case. This method has been successfully utilized by The United States Census Bureau for a long time. The main weakness of this group of methods is the arbitrariness of the choice of the similarity measure.

Expectation maximization is an iterative procedure that runs in two steps. In the first one, expectation step, the expected value of the complete "present" part of the data file likelihood is computed and then, in the second, maximization step, the expected values for the missing data obtained in the expectation step are substituted. Then the likelihood function is maximized to obtain new parameters. Those two steps are taken iteratively until the convergence is obtained. The method is widely utilized and implemented in Missing Values Analysis module of SPSS, and is considered as a benchmark method in the estimation of missing values.

2 Application of SOM to missing data estimation

SOM (Self-Organizing Map) is probably the most biologically inspired artificial neural network. Usefulness of this unsupervised neural net is proved by its uniqueness, e.g. lack of direct similarities to other methods of data analysis (Sarle (1994)), robustness and possibility of analyzing huge data sets containing even corrupted data (Kohonen (2001)). The main area of SOM applications is data analysis – discipline in which no additional information about data set (e.g. number of groups, group membership, etc.) is provided. From this point of view, in many papers, SOM is presented as visualization (Ultsch (1993)) and a clustering method (Murtagh (1995)). SOM, as a clustering algorithm, is used especially in case of huge data sets.

However, the SOM algorithm is conceptually very simple, it is surprisingly very resistant to any formal analysis Kohonen (2001). The great number of simulation experiments led to the conclusion that SOM may be perceived as:

- *Nonlinear transformation* of any multidimensional continuous space into one- or two-dimensional discrete space, which preserves order of objects from input space. In this sense SOM may be seen as similar to a principal component analysis Sarle (1994), multidimensional scaling (Jajuga (1990, 1993)) and clustering methods;
- *Nonparametric, nonlinear regression* (Cherkassky and Lari-Najafi (1991); Kohonen (2001)), which attempts to fit number of ordered codebook vectors to the distribution of objects in the input space.

As it was said before, the best data estimation method should try to maintain distribution of the data space or approximate it as close as possible. The mentioned above characteristics of SOM seem to be very useful from this point of view. We may assume that its non-parametric, distribution-mapping characteristics may be very useful in the missing data estimation because of the robustness to distribution assumptions.

Taking into account the typology of the data estimation algorithms submitted in the previous section, SOM might be qualified to the group of taxonomic methods with its high similarity to hot deck imputation. However, its regression capabilities at the same time recall some similarities to the group of regression methods. The advantage of SOM over classical regression results from the fact that it is not necessary to impute the type of relationship between dependent and independent variables, and in fact it is not necessary to build a regression model at all. It may be said that SOM, due to its non-parametric capabilities, constitutes a universal regression model, which may be utilized in the task of missing data estimation.

The first attempt of using SOM as a method estimating incomplete information was presented in Grabowski (1998). The data file used in the research was a commonly known IRIS data set and the compared data imputation method was mean substitution. Data losses were randomly generated across all 4 variables and 150 cases. The experiments were conducted for various

missing data ratios: 10%, 20%, 30%, 40%. The results showed that a relative estimation error was on average 4-5 times smaller in favour of SOM.

The main objective of this paper is to check if SOM is still so promising when comparing with a more advanced, benchmark data estimation method, i.e. expectation maximization.

3 Data file and software used in experiment

The data file, on which the analysis was based, came from the survey *Public utilities in opinions and budgets of Krakow citizens* conducted in 1998. Originally, the results of the survey form the matrix of 1014 cases and over 200 variables. The variables were measured on different scales including quantitative, ordinal and logical scales. The file was highly corrupted by a great number of missing data in almost every variable. The degree of data losses across the variables varies and reaches from several percent to even 95%. Obviously such highly corrupted variables are completely useless for further analysis and for this reason should be deleted from the file.

For the needs of the experiment, only some, namely eleven, variables of the file were selected. Two of them are quantitative and nine are ordinal. Quantitative variables are the following: P1D – *the cost of water* and P1F – *the cost of garbage disposal*. Qualitative variables: P2A – *opinion concerning water supply*, P2B – *opinion about public transportation*, P2C – *opinion about taking out sewage*, P2D – *opinion about garbage disposal*, P7A – *opinion about pressure of the water*, P7B – *opinion about continuity of water supply*, P7C – *opinion about the smell of the water*, P7D – *opinion about the color of water* and P7E – *opinion about water taste* were measured on eleven-point (0 – 10) Likert Scale. Some information concerning the characteristics of the variables is submitted in Table 1.

	P1D	P1F	P2A	P2B	P2C	P2D	P7A	P7B	P7C	P7D	P7E
min	1	1	1	1	1	1	1	1	1	1	1
max	218	73	10	10	10	10	10	10	10	10	10
% of incomplete values	3.2	2.0	1.1	12.6	4.8	1.0	2.0	2.1	3.4	3.4	5.6

Table 1. Some characteristics of the variables

A brief histogram analysis of the variables showed that in general the distribution of the variables was far from normal with the exception of P7C and P7E. The degree of the correlation differs across the variables. While computing the correlation, two correlation measures were used - classical Pearson correlation coefficient for quantitative variables and Spearman coefficient for ordinal variables. On average the correlation between the variables ranged from 0.2 - 0.5 except for P7C with P7D - 0.73 and P7D with P7E - 0.69.

Summarizing the above discussion it may be concluded that although the variables are not highly correlated, they are correlated in such a degree that the MAR assumption cannot be excluded.

The Little's MCAR test for all cases of experimental data yields p-value of 0, which results in rejecting the null hypothesis about MCAR assumption. This however does not exclude that data are MAR (unfortunately test for MAR assumption does not exist) – such an assumption was accepted for the experiments.

In order to measure the accuracy of a given method, additional data losses were randomly generated. The data losses were generated in the same proportion as the losses already existing in the variables. Consequently additional 3.2% of data losses were generated in P1D variable, 2.0% in P1F, 1.1% in P2A etc. Additional data losses generated as described above, would not weaken MAR assumption.

SOM belongs to the family of clustering methods and therefore it is sensitive to the chosen distance metric. In order to avoid gaining significance by P1D and P1F, which ranges are much higher than the rest of the variables, all the variables were standardized to the range of $<0, 1>$ using the following formula:

$$\hat{x} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

The experiment was carried out for two cases: not standardized and standardized data file. Missing Values Analysis module of SPSS 11.5 for Windows was used as the implementation of EM algorithm. SOM-PAK (Kohonen et al. (1995)), original SOM algorithm developed by Prof. Kohonen and his team in Helsinki University of Technology was utilized for SOM implementation. The parameters of SOM were chosen according to the general guidelines recommended in Kohonen et al. (1995) and were the following: topology: hexagonal, rectangular dimension of the map – relative to the sample size, i.e. 30x50, neighborhood function – Gaussian, alpha training coefficient – starting from 1 and linearly decreasing to 0, distance metric – Euclidean.

4 Results of the experiment

When comparing the results, the following two error measures were used:
Absolute error:

$$E = |x_w - x|$$

and relative error:

$$E_{rel} = \frac{E}{x_w} 100\%$$

where: x_w – true value, x – estimated value.

The relative error seems to be a preferred measure of the accuracy because it relates the value of error to the value of estimated data. However, in ordinal variables computing such an error is questionable and therefore it was decided to use the absolute error in this case. In case of numeric variables there was no obstacle to use the relative error. In case of every estimated variable, the following characteristics were computed: maximum value of the error (max), minimum value of the error (min) as well as mode (mod), median (med) and average (avg), with the exception that the mode was not computed for quantitative. Since the Likert scale is considered as interval one by many researchers, it was decided to compute an average for qualitative variables. The results of the experiment are submitted in Tables 2-5.

	P1D	P1F	P2A	P2B	P2C	P2D	P7A	P7B	P7C	P7D	P7E
max	2087.5%	126.0%	3	7	9	5	3	2	6	4	6
min	0.1%	0.8%	0	0	0	0	0	0	0	0	0
mod	n/a	n/a	0	1	1	2	1	0	3	1	1
med	39.6%	43.9%	1	2	1	2	1	0	2	1	1
avg	147.7%	42.6%	1.0	1.9	2.0	2.0	1.0	0.7	2.1	1.7	1.8

Table 2. Results of experiment; method – SOM

	P1D	P1F	P2A	P2B	P2C	P2D	P7A	P7B	P7C	P7D	P7E
max	3660.0%	430.0%	3	6	9	4	3	2	8	5	5
min	1.1%	0.8%	0	0	0	0	0	0	0	0	0
mod	n/a	n/a	1	1	1	1	1	1	2	1	1
med	39.6%	51.2%	1	2	1	2	1	1	2	2	2
avg	196.8%	77.2%	1.1	1.8	2.0	2.0	1.0	0.8	2.0	1.9	1.9

Table 3. Results of experiment; method – EM

	P1D	P1F	P2A	P2B	P2C	P2D	P7A	P7B	P7C	P7D	P7E
max	2920.9%	583.9%	5	8	10	6	3	4	6	7	8
min	3.8%	2.7%	0	0	0	0	0	0	0	0	0
mod	n/a	n/a	1	1	0	1	1	0	1	0	1
med	38.1%	50.7%	1	2	1	1	1	1	1	1	1
avg	175.0%	115.1%	1.8	2.2	1.9	2.2	0.8	0.8	1.4	1.7	1.2

Table 4. Results of experiment; method – SOM; standardized data file

	P1D	P1F	P2A	P2B	P2C	P2D	P7A	P7B	P7C	P7D	P7E
max	3662.3%	432.2%	3	6	9	4	3	2	8	5	5
min	1.0%	0.4%	0	0	0	0	0	0	0	0	0
mod	n/a	n/a	1	1	1	1	1	1	2	1	1
med	39.8%	51.3%	1	2	1	2	1	1	2	2	2
avg	196.9%	77.4%	1.1	1.8	2.0	2.0	1.0	0.8	2.0	1.9	1.9

Table 5. Results of experiment; method – EM; standardized data file

5 Conclusions

In general both methods performed the estimation of missing data with a comparable accuracy. However, the results require the following supplementary comments:

1. In case of quantitative variables, SOM outperformed EM in both experiment runs. SOM produced a much lower maximum error. Also the average error was smaller. This may be due to the fact that SOM, as a non-parametric method, maps the distribution of data samples in a greater degree, whereas EM imposes multinormality.
2. Standardization of the data file did not impose any differences in case of EM algorithm.
3. Standardization of data file in case of SOM caused higher error in case of numeric variables while a smaller error was generated in case of the ordinal ones. The reason was the range of numerical variables compared to categorical ones (10-20 times higher). After normalization the significance of numerical variables in the distance measure was lowered. It is worth stressing that the standardized case of SOM estimated all ordinal variables with a greater accuracy. The above is demonstrated in the lowest values of modes and medians of relative errors.

The experiment showed that SOM might be utilized as a non-trivial method for estimation of missing data. In this comparative study on a given data set, SOM algorithm proved to be at least as good if not better than EM algorithm. It seems that SOM as a non-parametric method fits better to the distribution of data samples. The features of SOM in missing data estimation demand further studies, which should include comparison with other data estimation methods across different data files with different distributions.

References

- CHERKASSKY, V. and LARI-NAJAFI, H. (1991): Constrained Topological Mapping for Nonparametric Regression Analysis. *Neural Networks*, 4, 27–40.
- GRABOWSKI, M. (1998): Application of Self-Organizing Maps to Outlier Identification and Estimation of Missing Data. In: A. Rizzi, M. Vichi, and H.H. Bock (Eds.): *Advances in Data Science and Classification*. Springer, Berlin, 279–286.
- JAJUGA, K. (1990): *Statystyczna teoria rozpoznawania obrazu*. PWN, Warszawa.
- JAJUGA, K. (1993): *Statystyczna analiza wielowymiarowa*. PWN, Warszawa.
- KOHONEN, T., HYNNINEN, J., KANGAS, J., and LAAKSONEN, J. (1995): *SOM_PAK The Self-Organizing Map Program Package Version 3.1*. Helsinki University of Technology, http://www.cis.hut.fi/research/som_pak/
- KOHONEN, T. (2001): *Self-Organizing Maps*. Springer, Berlin.
- KORDOS, J. (1988): *Jakość danych statystycznych*. PWE, Warszawa.
- LITTLE, R.J.A. and RUBIN, D.A. (1987): *Statistical analysis with missing data*. John Wiley and Sons, New York.
- MURTAGH, F. (1995): Unsupervised catalog classification. In: D. Shaw, J. Payne, and J. Hayes (Eds.): *Astronomical Data Analysis Software and Systems IV*, ASP, 264–267.
- PAWEŁEK, B. (1996): *Metody szacowania brakujacych informacji w szeregach przekrojowo-czasowych*. Niepublikowana praca doktorska. Akademia Ekonomiczna w Krakowie.
- POCIECHA, J. (1996): *Metody statystyczne w badaniach marketingowych*. PWN, Warszawa.
- SARLE, W.S. (1994): Neural Networks and Statistical Models. In: SAS Institute Inc.: *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, NC, 1538–1550.
- ULTSCH, A. (1993): Self organized feature maps monitoring and knowledge acquisition of a chemical process. In: S. Gielen and B. Kappen (Eds.): *Proceedings of the International Conference on Artificial Neural Networks (ICANN93)*. Springer-Verlag, London 864–867.
- WOTHKE, W. (2000): Longitudinal and multi-group modeling with missing data. (Adobe pdf format) In T.D. Little, K.U. Schnabel, and J. Baumert (Eds.): *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples*. Mahwah, NJ: Lawrence Erlbaum Associates.

Applicability of Customer Churn Forecasts in a Non-Contractual Setting

Jörg Hopmann and Anke Thede

Institut für Informationswirtschaft und -Management,
Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

Abstract. As selling a product to an existing customer is much more cost effective than acquiring new customers companies increasingly focus on retaining profitable customers rather than concentrating all marketing actions on the acquisition of new customers. For retaining customers it is very important to be able to predict whether a customer is still active. Effectless marketing expenses directed towards already inactive customers can be avoided and more intensive marketing actions can be taken in order to support active customers' purchase intentions. Several methods exist that can be used to predict customer activity. In this paper we apply a stochastic and a data mining method to real-life B2B purchase histories and compare the usability and the quality of churn prediction of each of the methods in a non-contractual B2B environment.

1 Introduction

In recent years, marketing strategies changed from a product-centric perspective to a customer-focused approach. Zeithaml and Berry (1996) state that reducing customer migration has a more positive impact on profits than market share, scale economies, and other commonly used strategic variables. Krafft and Bromberger (2001) find a negatively correlated relation between the company's value and the aggregated customer defection rate in their study. Losing customers not only leads to opportunity costs because of reduced sales but also to an increased need for attracting new customers which is five to six times as expensive as selling to existing customers (Bhattacharya (1998), Dyché (2002)). Being able to predict which customers are likely to defect and at what time allows for special marketing actions with the objective of retaining profitable customers. In this paper we give an overview over several possible methods for forecasting customer churn (**change & turn**) and apply the most promising ones to real-life B2B purchase histories to evaluate the quality of churn prediction of each of the methods with respect to their practical usability.

The churn behaviour of customers in contractual scenarios where the date of churn can easily be identified has recently been investigated (Rosset et al. (2002), Mani et al. (1999)). However, in non-contractual settings the starting date of a customer relation is known, but not the end. Estimating the date of defection in those contexts requires complex models. The research related

to this area concentrates on modelling repeat-buying behaviour based on the Negative Binomial Distribution Model (NBD) by Ehrenberg (1972). Schmittlein et al. (1987) proposed an NBD-based model extending the traditional model by a stochastic deflection process.

As an alternative to a complicated stochastic model of customer behaviour also standard classification techniques can be applied for predicting customer churn. In this paper we apply the NBD-based model and simple, alternative methods to real-life purchase histories to evaluate the prediction quality and the practical usability of the models.

The data is from a German B2B catalogue retailer for electronics and computer accessories. The purchase histories cover a period from 1997/01/01 to 2002/03/21. We use the last 27 months as a control period for our predictions and take the years 1997–1999 for our analysis. In the company, inactivity is usually assumed after 12 months without purchase, 27 months is thus sufficient for testing the actual activity. The data comprises a total of 72,026 customers and 502,965 orders with 1,924,077 single transactions. Customers who made their first purchase before January 1st, 1997 are excluded from our analysis in order to avoid left-censored data. Because customer transaction patterns vary from one lifecycle stage to the other we use only cohort data and restrict our study to the largest cohort with 1133 individuals.

2 A stochastic model for customer churn prediction

Many churn prediction applications used by practitioners are based on data mining algorithms. Designed for contractual relationships and depending on the knowledge of the exact time of churn, they are not fully applicable in the context of catalog retailers or other non-membership business models. Schmittlein et al. (1987) have proposed a stochastic model (Pareto/NBD) for determining the individual customer's probability of being alive based on past transaction data. It is based on the repeat-buying theory (Ehrenberg (1972)) and models transactional behavior of customers according to an NBD distribution. Extending the traditional repeat-buying model a Pareto distribution is incorporated to define customer migration processes.

In the Pareto/NBD model each customer is described by the number of purchases X and the time t of the last transaction since the initial purchase incident. Together with the time of observation T a customer's individual probability of being alive can be computed as $P(\text{alive} | X, t, T)$. The exact formula is given in Schmittlein et al. (1987). The Pareto/NBD model is based on the following behavioural assumptions:

1. **Purchase processes** of active customers follow a Poisson distribution with parameter λ . The interpurchase times are exponentially distributed.
2. The **heterogeneity** between individual purchase rates is captured by a gamma distribution of λ with parameters r and α ($E[\lambda | r, \alpha] = r/\alpha$ and $Var[\lambda | r, \alpha] = r/\alpha^2$).

Parameter z	r	α	s	β
Estimated z	2.29	2.58	0.421	0.7895
$\sigma(z)$	1.1298	1.3771	0.0815	0.1678
$CV(z)$	49.29%	53.34%	19.36%	43.08%

Table 1. Estimated Pareto/NBD parameters for 549 customers with $X > 1$.

3. **Customer lifetimes**, i.e. the time from trial until defection, are exponentially distributed with deathrate μ .
4. **Heterogeneity** in terms of different deathrates is modeled by a gamma distribution of μ with parameters s and β ($E[\mu | s, \beta] = s/\beta$ and $Var[\mu | s, \beta] = s/\beta^2$).
5. Purchase and defection processes are **independent**.

The coefficient of variation is defined as $CV(x) = \frac{\sigma(x)}{E(x)}$. $CV(\lambda) = r^{-\frac{1}{2}}$ serves as an index for homogeneity in the purchase rates. For the following evaluation, we use a varying threshold θ (0.1 - 0.9) to transform the probability of being alive into a dichotomic variable ($\theta \leq P(alive)$)).

The parameters to be estimated in order to apply the Pareto/NBD model are r , α , s , and β . For the studied cohort with 1133 individuals the parameter estimation was done by the bootstrap method of moments' estimate which is described in detail by Krafft (2002). Among this cohort 584 customers (51.5%) bought only once during the observed period. Due to very high parameter variation and poor classification results we excluded the one-time buyers (OTB) from the sample. OTB behaviour is not captured very well by the standard Pareto/NBD model as it is based on repeat-buying theory.

Limited by the underlying database, the minimum sample size of 1600 subjects as recommended by Schmittlein and Peterson (1994) could not be attained. Tab. 1 shows our estimates for the four model parameters. It is interesting to note that our estimates for the heterogeneity parameters r and s are very close to respective estimates in similar studies by Schmittlein et al. (1987), Reinartz and Kumar (2001), and Krafft (2002). While the gamma distribution shape parameter r indicates a relatively high homogeneity in transaction rates across customers, the corresponding parameter of the unobserved death rate s is very low, representing a high level of heterogeneity across customers. The variation coefficients shown in the bottom line of Tab. 1 are much higher than in other studies in the B2B sector (Reinartz and Kumar (2001)) which seems to be the result of a poor data fit. As a simple measure for excessive purchase regularity we computed the coefficient $R = \left(\frac{E(IPT)}{\sigma(IPT)} \right)^2$

Threshold θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
type I error	10.2%	11.3%	12.2%	12.8%	14%	14.7%	15.1%	15.8%	17.9%
type II error	12%	10.4%	7.8%	7.6%	6.9%	6.6%	5.7%	4.2%	2.7%
total error	22.2%	21.7%	20%	20.4%	20.9%	21.3%	20.8%	20%	20.6%

Table 2. Pareto/NBD classification error rates for customers with $X > 1$.

	type I error	type II error	total error
OTB heuristic	6.2%	0%	6.2%
Combined with NBD ($\theta = 0.3$)	9.1%	3.8%	12.9%

Table 3. Results for simple OTB heuristic and for Pareto/NBD combination.

with the interpurchase time spans IPT described by Morrison (1981). If interpurchase times are regular, the standard deviation σ is close to 0 and R is high whereas exponentially distributed interpurchase times with an average close to σ lead to an R value around 1. Crie (2001) points out that any value of R below 6 should be sufficiently close to the ideal one for practical modelling purposes. In the given sample 34 customers have an R value of $R \geq 6$ and 59 of $3 \leq R < 6$. Exclusion of customers with excessive regularity measures did not improve the variation coefficients and a further segmentation yields too small segments for reliable parameter estimation (Schmittlein and Peterson (1994)). Our classification accuracy is nevertheless comparable with that of a study by Krafft (2002). No major change in purchase behavior was observed by the company during the period 1997-1999.

To assess the classification results we take as hypothesis $H_0 = \text{customer is active}$ that has to be inspected for every customer from the selected cohort by the end of the classification period, namely Dec 1999. The actual activity can be assessed by checking whether the customer made at least one more purchase in the following period from January 2000 until March 2002. A type I error for H_0 implies that the model classifies an actually active customer as churned, a type II error that a churned customer is classified as active.

Tab. 2 shows the results achieved with the Pareto/NBD model for varying cutoff thresholds θ . The lowest total error rate is achieved at two different thresholds $c = 0.3$ and $c = 0.8$. Varying the threshold mainly influences the proportion of type I to type II error. High threshold values lead to a high misclassification of active customers while low values of θ yield an almost equal distribution of the error between the two error types. For the practical use the threshold should thus be chosen depending on the concrete loss function.

A simple heuristic regarding the OTB who are not included in the analysis so far is to classify all OTB as inactive. This can serve as an extension to the Pareto/NBD model which models only repeat buyers. The error rates are shown in the upper line of Tab. 3. 6.2% of the customers that bought only once during the analysis period repurchased during the control period. The lower line shows the results for a combination with the Pareto/NBD model using a threshold of $\theta = 0.3$. The combination works by applying the Pareto/NBD model on the repeat buyers and using the threshold to get an active/inactive classification and by classifying all OTB as inactive.

3 Defection forecasts using RFM models

In this section we examine the usability and the quality of a standard classification technique for predicting customer defection. RFM (Recency, Frequency,

probit				glm					
all customers		without OTB		all customers		without OTB			
	coeff.	sign.	coeff.	sign.	coeff.	sign.	coeff.	sign.	
R	0.06	0.001	0.058	0.001	R	0.02	0.001	0.021	0.001
M	-0.0001	0.001	-0.0001	0.001	F	-0.003	0.001	-0.0026	0.01
const	-0.356	0.05	-0.295	0.1	const	0.3	0.001	0.290	0.001
AIC	753.46		480.83		AIC	618.57		506.5	

Table 4. Results for probit regression and glm (coeff. = coefficients, sign. = significance level, const = constant).

Monetary) models analyze customer behaviour based on purchase history data, namely on the time of the last transaction, the number of transactions and the total amount of money spent during the observed period (Colombo and Weina (1999)). Given a segmentation for a set of customers optimal

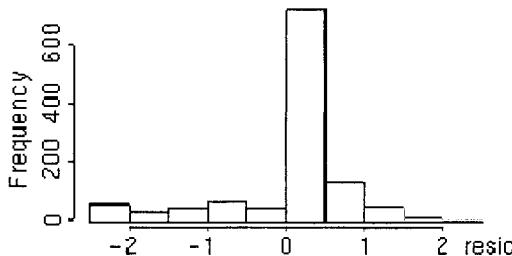


Fig. 1. Frequencies of the residual errors of the probit RFM model.

weights for recency, frequency and monetary value can be calculated using regression models. For predicting customer activity the segmentation classifies the customers into active and churned. For such a binary, dependent variable logistic regression is appropriate. The model then produces for each customer the probability between 0 and 1 that this customer belongs to the 1 segment. The probability can be turned into a classification by using a threshold probability ζ that assigns to each probability the classification into one of the two classes. By definition of the model $\zeta = 0.5$ is most obvious, but other thresholds can be useful as we demonstrate later.

The error distribution of residuals shown in Fig. 1 strongly supports the choice of a probit model. In fact, runs using the logit model yielded a worse fit and worse results compared to the probit model. We thus concentrate our detailed analysis on the probit model. To be able to compare the results with the Pareto/NBD performance the calculation is based on the same cohort of 1133 customers whose actual activity is drawn from the information of the control period. The software R, version 1.6.1, is used to perform the regression. A run of a complete RFM model was first performed. Tab. 4 shows the results of a model restricted to the variables that were significant at a level of at least 0.5 in the RFM run. The Akaike Information Criterion (AIC) is given to indicate the model fits. The restriction slightly improved the model fit. Tab. 5 shows the errors of the model using a threshold $\zeta = 0.5$.

method	error	all customers	without OTB	combination
probit	type I error	9%	12%	9%
	type II error	3.6%	6.7%	3.2%
	total error	12.7%	18.7%	12.2%
glm	type I error	8.7%	11.5%	8.7%
	type II error	4.3%	8.9%	4.3%
	total error	13.0%	20.4%	13.0%

Table 5. Errors for probit and glm regressions, with all customers, only OTB, and RFM-OTB heuristic combination, $\zeta = 0.5$.

To test the performance of simpler models we ran the same calculation using a simple general linear model (glm) with gaussian distribution, restricted to the significant parameters R and F. The results are shown in Tab. 4. This model does not restrict the resulting values to lie between 0 and 1 but application of a threshold ζ again transforms the results into a binary classification. The errors for $\zeta = 0.5$ are shown again in Tab. 5. The model fit as indicated by the AIC of the glm is better than the fit of the probit model. The errors are comparable to those of the probit analysis with the difference that the glm yields a lower type I error while allowing a higher type II error and a higher total error.

To directly compare the model performance with that of the Pareto/NBD model we ran a second analysis without the OTB and we calculated the error rates of the combination of the RFM methods with the OTB heuristic. The results for the analyses without OTB and the errors for this analysis and the combination are included in Tab. 4 and Tab. 5, respectively. The model fit of the analyses without OTB is much better compared to the models run on the complete customer base. Nevertheless this does not have a significant positive effect on the error rates.

For discussing the practical applicability of these models we have to consider the loss functions a marketer would associate with the different error types. For a marketer, a type I error corresponds to the cost of losing an active customer as a customer who is active and in principle willing to purchase might refrain from doing so if he no longer receives the usual marketing material. It can thus be calculated as the *cost of a new customer acquisition + direct marketing expenses*. The loss of the average profit margin is compensated by the newly acquired customer. The cost of a type II error is the direct marketing cost per customer caused by the fruitless marketing actions towards a defected customer. For the company studied, the cost of acquiring a new customer is approximately EUR 180 and the direct marketing cost per customer is approximately EUR 25. This means that a type I error costs 6.2 times more than a type II error. The weighted total error of a method corresponding to the total cost can be calculated as $(6.2 * mis_I + mis_{II}) / (6.2 * \text{total active} + \text{total churned})$ with mis_i the number of misclassifications of error type i . The left hand side of Fig. 2 shows the weighted total error for different thresholds at steps of 0.01 for the glm with

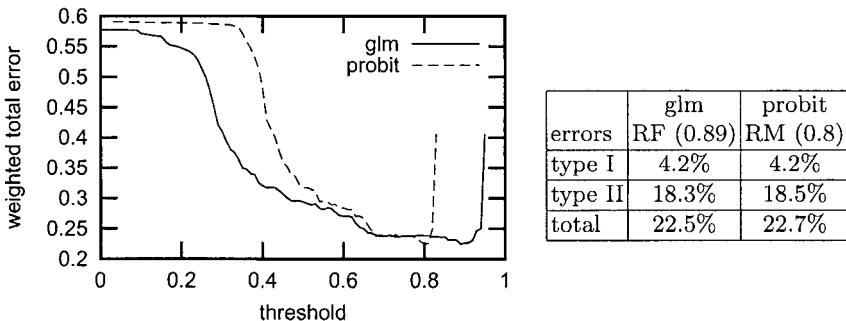


Fig. 2. Distribution of the weighted total error for different thresholds (left) and errors for minimal thresholds (right, threshold in brackets).

independent variables R and F and for the probit with independent variables R and M, including all customers. The minimum weighted total error can be reached applying a threshold of 0.89 for the glm and 0.8 for the probit model which yield the error values depicted in the table in Fig. 2.

4 Discussion

The results of the probit and the glm model are in all three variations comparable. There are slight differences concerning the proportion of the two error types and the lower total error of the probit model. Combining the models reduced to repeat-buyers with the OTB heuristic reaches almost the same results as the models run on the whole customer base. The increase in model fit quality is probably due to the reduced model input sample size. Tuning the threshold ζ in order to minimize the weighted total error as defined in Sec. 3 attains comparable results for both models. Fig. 3 indicates that the glm model ($\zeta \in [0.66, 0.93]$) is more robust towards the specification of ζ than the probit model ($\zeta \in [0.67, 0.81]$).

The results achieved by the Pareto/NBD model classifying all repeat-buyers (Tab. 2, $\theta = 0.3$) are worse than the outcomes of the probit model (middle column of Tab. 5) which is basically due to a worse type II error. The total error rate is similar to that of the glm but the type I error is higher. The Pareto/NBD model which is based on repeat-buying behavior obtains acceptable results for customers with a purchase frequency greater than one in the analysis period. One-time buyers have to be modeled separately. Nevertheless, the results are worse than those of the RFM-based models. This may be due to the poor model fit that may be caused by the fact that in the B2B sector purchases do not always appear at random but are influenced by external factors like e. g. budgeting policies. Extensions to the Pareto/NBD model that address the particular conditions of business customers are subject to future research. Another important drawback of the Pareto/NBD model is the restriction to a customer base with similar purchase patterns that limits the analysis to cohorts. In this case, even the largest cohort from data

covering three years is normally still not large enough for a sound analysis. This makes the model inappropriate for smaller and even medium sized B2B companies.

The probit and glm model both seem to be appropriate for practical modelling purposes. The RFM model does not make any assumptions about the input data. OTB are modeled well by RFM methods and do not require a separate treatment. We expect that extending the analysis to the whole customer base including other cohorts will lead to a further increase of the model performance.

References

- BHATTACHARYA, C.B. (1998): When Customers are Members: Customer Retention in paid Membership Contexts. *Journal of the Academy of Marketing Science*, 26(1), 31–44.
- COLOMBO, R. and WEINA, J. (1999): A Stochastic RFM Model. *Journal of Interactive Marketing*, 13(3), 2–12.
- CRIE, D. (2001): Active versus Inactive Customer or from Client to ex-Client: Concepts, Definitions and Measures. *Cahier de la recherche de l'IAE*.
- DYCHE, J. (2002): *The CRM Handbook: A Business Guide to Customer Relationship Management*. Addison-Wesley, Boston.
- EHRENBERG, A.S.C. (1972): *Repeat Buying: Theory and Applications*. North-Holland Pub.Co., American Else, Amsterdam.
- KRAFFT, M. and BROMBERGER, J. (2001): Kundenwert und Kundenbindung. In: S. Albers, M. Clement, K. Peters, and B. Skiera (Eds.): *Marketing mit Interaktiven Medien: Strategien zum Markterfolg*. F.A.Z.-Institut Frankfurt, 160–174.
- KRAFFT, M. (2002): *Kundenbindung und Kundenwert*. Physica-Verlag, Heidelberg.
- MANI, D.R., DREW, J., BETZ, A., and DATTA, P. (1999): Statistics and Data Mining Techniques for Lifetime Value Modeling. *Proceedings of KDD-99*, 94–103.
- MORRISON, D.G. (1981): Modeling Consumer Purchase Events: a Reply to Lawrence. *Journal of Marketing Research XVIII*, 465–469.
- REINARTZ, W.J. and KUMAR, V. (2001): The Impact of Customer Relationship Characteristics on Profitable Lifetime Duration. *Journal of Marketing*, 67(1), 77–99.
- ROSSET, S., NEUMANN, E., EICK, U., VATNIK, N., and IDAN, Y. (2002): Customer Lifetime Value Modeling and its Use for Customer Retention Planning. *Proceedings of the Eighth ACM SIGKDD, KDD-02*.
- SCHMITTLEIN, D.C., MORRISON, D.G., and COLOMBO, R. (1987): Counting your Customers: Who are they and what will they do next? *Management Science*, 33, 1, 1–24.
- SCHMITTLEIN, D.C. and PETERSON, R.A. (1994): Customer Base Analysis: An Industrial Purchase Process Application. *Marketing Science*, 12(1), 41–67.
- ZEITHAML, V.A. and BERRY, L.L. (1996): The Behavioral Consequences of Service Quality. *Journal of Marketing*, 60, 2, 31–46.

A Gravity-Based Multidimensional Unfolding Model for Preference Data

Tadashi Imaizumi

School of Management & Information Sciences,
Tama University, 4-1-1 Tama-shi, Tokyo, Japan

Abstract. A new model for analyzing two-way, two-mode preference data is proposed. MultiDimensional Unfolding models (MDU) have been used widely. In these model, the observed preference value is related to the distance between the ideal point and object point only. The market share of each brand is ignored or assumed to be the same for all objects. The attraction of each object, such as the market share of that object, must be incorporated in the analysis of marketing data. A gravity-based multidimensional unfolding model will be proposed. One specific characteristic of preference data of N subjects is that observed preference values of individuals are often not compatible between individuals. The de-generated configuration problem on applying the non-metric MDU method to a real data set will be caused by the weak condition on the data matrix. A linearly constrained non-metric approach is also proposed to try to rescue from obtaining the de-generated configuration.

1 Introduction

Product positioning has been argued in marketing research before designing new products. After analyzing the market structure, and trying to predict the future market share of an object, the new product will be designed. The typical data we analyze for product positioning are brand switching data or preference choice data. Analyzing brand switching data is useful to predict the market structure in the future. MultiDimensional Scaling (MDS) models and method have been used to explore the hidden market structure in these data. The similarity between two objects or the preference to an object of each individual are related to the distance between 2 points, which represent two objects, or object and individual, respectively. The MDS model is very a useful model for applying these data to obtain the spatial representation of objects. The multidimensional unfolding models (MDU), such as the ideal point model, are models for analyzing preference choice data. They assume that individuals share common dimensions of choice. Each individual is represented as 'ideal' point in the t -dimensional space as the distances from the ideal point to the objects points are related to the preference value to the objects of that individual. These unfolding models have also the same characteristics of ordinary MDS models. It is focused on the analysis of marketing data, e.g. how to estimate the degree of the brand loyalty and how to predict the market share. However, the degree of the loyalty of each individual to a

brand is not represented in MDS model since the only distances between two points are related to the data.

Multidimensional unfolding models are difficult to apply in the case that each brand is differently weighted in preference choice. The major model which tries to represent the brand loyalty or the attraction-mass is the so-called 'Gravity model'. This gravity model has been applied in many research fields such as the regional sciences, economics, social interaction and consumer research (Bottum (1989)) as cited in DeSarbo et al. (2002). DeSarbo et al. (2002) discussed the gravity model in the marketing areas and proposed a gravity based Multidimensional unfolding model,

$$F_{ij} = \beta S_i^\alpha M_j^\lambda / d_{ij}^2 \quad (1)$$

where F_{ij} is preference value for object o_j of individual I_i , d_{ij} is a Euclidean distance between ideal point y_i and object point x_j in t -dimensional space, S_i^α is consumer mass of individual I_i , and M_j is brand mass of object o_j , α is mass parameter for individual and λ is mass parameter for object. In this model, the preference value F_{ij} is rapidly increased as the object point x_j approaches to the ideal point y_i . The F_{ij} is rapidly decreasing with d_{ij}^2 in this model. The process of preference choice evaluation will be complex and not be well described as compared with the physics system. DeSarbo et al. (2002) also discussed the stochastic gravity model by introducing an multiplicative error term into their proposed model. However, some analysis of real data sets such as that of Rushton (1969)'s study indicates the preference value tends to vary milder with distance d_{ij} . One rational modification will be that M_j and d_{ij} vary. As F_{ij} is linear function of M_j with fixing other parameters, the randomness related to M_j will be accounted by the error term. But the randomness related to d_{ij} will not be accounted. Let D_{ij} be a random variable with mean d_{ij} and variance σ_{ij}^2 , then we have, not

$$E(D_{ij}^2) = d_{ij}^2, \quad (2)$$

but

$$E(D_{ij}^2) = d_{ij}^2 + \sigma_{ij}^2. \quad (3)$$

This suggests that the above gravity model is modified as

$$F_{ij} = \beta S_i^\alpha M_j^\lambda / (d_{ij}^2 + \sigma_{ij}^2). \quad (4)$$

Several sub-models on σ_{ij}^2 will be introduced such as

$$\sigma_{ij}^2 = \sigma_i^2 + \sigma_j^2, \quad \sigma_{ij}^2 = \sigma_i^2, \quad \sigma_{ij}^2 = \sigma_j^2, \quad \sigma_{ij}^2 = \sigma^2. \quad (5)$$

Since some individual will be mild in the preference choice evaluation, and the other will not. So, it is assumed that the parameter σ_{ij} is dependent on individuals only in this paper, $\sigma_{ij} = \sigma_i$.

We can apply this model to a data set directly if the data is metric. If an observed preference value is rank-ordered, or is not linearly related to F_{ij} , the other approach to analyze data is needed. The non-metric method by Kruskal (1964) has been widely used in MDS methods and MDU methods. This paper presents a modified gravity model with iterative non-metric approach. In applying non-metric MDU method, we must pay attention to the data conditionality. Sometimes, the data matrix is row-conditional which means that the preference values are not compatible between data of individuals but compatible within data of individual. In the non-metric MDS methods and MDU methods, the configuration is estimated by using ordinal information among data. It is pointed out that the degeneracies can occur in non-metric MDU methods. And the non-metric method for the unfolding model has not been applied successfully. How to avoid the degeneracies is very important for applying MDU methods. One procedure is to obtain a good initial configuration, another one is to constrain the regression function on $\{F_{ij}\}$ for the row-conditional data. In this paper, each of the monotone regression functions among N individuals is constrained to be linearly related.

2 Model

Let $\{o_j; j = 1, 2, \dots, n\}$ denote n objects, $\{I_i; i = 1, 2, \dots, N\}$ denote N individuals, f_{ij} denote the observed F_{ij} . And the distance d_{ij} is defined as an Euclidean distance between y_i and x_j ,

$$d_{ij} = \sqrt{\sum_p^t (y_{ip} - x_{jp})^2}. \quad (6)$$

As we analyze this data by the non-metric method,

$$f_{ij} = Mono_i[\beta S_i^\alpha M_j^\lambda / (d_{ij}^2 + s_i^2)], \quad (7)$$

where $Mono_i(\cdot)$ is a monotonous non-decreasing function for individual I_i . The parameter β , S_i , and λ are not identified in the non-metric method. And the consumer-mass S_i will be ignored in this model. Finally, we propose the modified gravity model in this paper as

$$f_{ij} = Mono_i[M_j / (d_{ij}^2 + s_i^2)]. \quad (8)$$

3 Method

The iterative non-metric procedure consists of three main steps.

- (1) Obtain an initial joint configuration, initial $\{M_j\}$, and an $\{s_i\}$.

- (2) Normalize the current joint configuration, compute the value of the loss function, and check whether this iterative process is converged or not.
- (3) If it is not converged, update the joint configuration, $\{M_j\}$ and $\{s_i\}$. repeat step 2 and 3. Each of $\{M_j\}$ and $\{s_i\}$ are restricted to be positive by adjusting an step-size parameter of an updating procedure.
- (4) The several solutions with the different dimensionality are compared on the value of the loss function and the degeneracies. And the optimal one is chosen.

3.1 Initial configuration

When the obtained solution of the higher dimensionality $s, s > t$, is available, the principal axes solution of dimensionality t is used as the initial joint configuration. And $\{M_j\}$ and $\{s_i\}$ at the dimensionality s are used as initial value of $\{M_j\}$ and $\{s_i\}$, respectively. When they are not available, then the initial M_j is set to 1, and the initial s_i is also set to 0. The initial joint configuration is derived as follows. With the initial setting for $\{M_j\}$ and $\{s_i\}$, d_{ij} will be expressed as

$$d_{ij}^2 = 1/f_{ij}. \quad (9)$$

Let $\mathbf{X}^{(0)}$ denote an initial object configuration, and $\mathbf{Y}^{(0)}$ denote an initial subject configuration, at iteration 0, then and the initial joint configuration will be derived by

- (1) double-center the matrix $\mathbf{C} = [1/\sqrt{f_{ij}}]$, and
- (2) apply the singular value decomposition of the matrix $\mathbf{C} = \mathbf{U}\Lambda\mathbf{V}'$,
- (3) set the initial joint configuration as $\mathbf{X}^{(0)} = \mathbf{V}\Lambda^{1/2}, \mathbf{Y}^{(0)} = \mathbf{U}\Lambda^{1/2}$.

3.2 Normalization

The current joint configuration will be normalized to

$$\sum_i^N y_{ip} + \sum_j^n x_{jp} = 0, \quad p = 1, 2, \dots, t, \quad (10)$$

$$\sum_i^N \sum_p^t y_{ip}^2 + \sum_j^n \sum_p^t x_{jp}^2 = N + n. \quad (11)$$

at each iteration.

3.3 Linearly constrained monotone function

The monotone transformed value of $\{f_{ij}\}$ is computed with using $\{F_{ij}\}$ at each iteration. One definition of the monotone regression function for the unconditional data is

$$\text{if } f_{ij} < f_{ik}, \text{ then } \text{Mono}(F_{ij}) \leq \text{Mono}(F_{ik}),$$

and that for the row-conditional data is

$$\text{if } f_{ij} < f_{ik}, \text{ then } \text{Mono}_i(F_{ij}) \leq \text{Mono}_i(F_{ik}), \text{ for } i = 1, 2, \dots, N$$

We define the linearly constrained monotonous function as follows,

$$\text{if } f_{ij} < f_{ik}, \text{ then } a_i \text{Mono}(F_{ij}) \leq a_i \text{Mono}(F_{ik}), \text{ for } i = 1, 2, \dots, N,$$

To obtain the monotone function $\text{Mono}(\cdot)$ which is common to all individuals, we adopt the following algorithm:

- (1) To set each vector length of F_{ij} to same one, compute $F_{ij}^+ = b_i F_{ij}$, where b_i is the normalizing factor to $\sum_i^N (F_{ij}^+)^2 = c$,
- (2) obtain $\{\hat{F}_{ij}^+\}$ such that,

$$\text{if } f_{ij} < f_{ik}, \text{ then } \hat{F}_{ij}^+ \leq \hat{F}_{ik}^+,$$

- (3) compute $\hat{F}_{ij} = \text{Mono}_i(F_{ij}) = a_i \text{Mono}(F_{ij}) = \hat{F}_{ij}^+ / b_i$, and $a_i = 1/b_i$.

As the above transformation on F_{ij} for each individual is a linear transformation, this will also minimizes

$$\sum_j^n (F_{ij} - \hat{F}_{ij})^2, i = 1, 2, \dots, N. \quad (12)$$

3.4 The loss function

For a given dimensionality t , we try to find the optimal solution \mathbf{y}_i , \mathbf{x}_j , M_j and s_i which minimize S

$$S = \sqrt{1/N \sum_i^N [\sum_j^n (F_{ij} - \hat{F}_{ij})^2 / \sum_j^n (F_{ij} - \bar{F}_i)^2]}, \quad (13)$$

where \hat{F}_{ij} is a disparity such that

$$\text{if } f_{ij} < f_{ik}, \text{ then } \hat{F}_{ij} \leq \hat{F}_{ik},$$

these disparities $\{\hat{F}_{ij}; j = 1, 2, \dots, n\}$ are derived which minimize

$$\sum_j^n (F_{ij} - \hat{F}_{ij})^2, i = 1, 2, \dots, N, \quad (14)$$

with the above linearly constraint.

3.5 Update joint configuration

All model parameters $\{y_i\}$, $\{x_j\}$, $\{M_j\}$, and $\{s_i\}$ are updated with using gradient method,

$$y_{ip}^{(q+1)} = y_{ip}^{(q)} + \gamma(c)^{(q)} \partial S / \partial y_{ip}, \quad (15)$$

$$x_{jp}^{(q+1)} = x_{jp}^{(q)} + \gamma(c)^{(q)} \partial S / \partial x_{jp}, \quad (16)$$

$$M_j^{(q+1)} = M_j^{(q)} + \gamma(M)^{(q)} \partial S / \partial M_j, \quad (17)$$

$$s_i^{(q+1)} = s_i^{(q)} + \gamma(s)^{(q)} \partial S / \partial s_i, \quad (18)$$

where q is the iteration number, $\gamma(c)^{(q)}$, $\gamma(M)^{(q)}$, and, $\gamma(s)^{(q)}$ is the step-size found by linear search at iteration q , respectively.

4 Application

We applied this proposed model to the food items data set that Green and Rao (1972) gathered. This data set was rank-ordered data for 15 food items from 42 individuals. These food items were ranked from the most preferred to the last preferred by each individual. So, this data is row-conditional data. The joint configuration and the other parameters were estimated for each dimension t , $t = 5, 4, 3, 2$ and 1. The S value obtained were 0.354, 0.434, 0.462, 0.523, and 0.621, respectively. This suggests that the solution of the dimensionality $t = 5$ or $t = 2$ is appropriate. As the visual representation is important to utilize the results, we chose the 2-dimensional solution for this data. Figure 1 shows the joint configuration. The 15 foods items are shown in Table 1.

Code	Food Item	Code	Food Item
TP	Toast pop-up	BTJ	Buttered toast and jelly
BT	Buttered toast	TMn	Toast and margarine
EMM	English muffin and margarine	CB	Cinnamon bun
JD	Jelly donut	DP	Danish pastry
CT	Cinnamon toast	GD	Glazed donut
BMM	Blueberry muffin and margarine	CC	Coffee cake
HRB	Hard rolls and buffer	CMB	Corn muffin and butter
TMd	Toast and marmalade		

Table 1. Food Items and it's code

Figure 1 is similar to the obtained configuration from M-D-SCAL V Row Split Analysis of Green and Rao (1972) or that of Borg and Groenen (1997). But, the object points were separately located for both of that of Green and Rao and that of Borg and Groenen. The ideal points were nearly located at

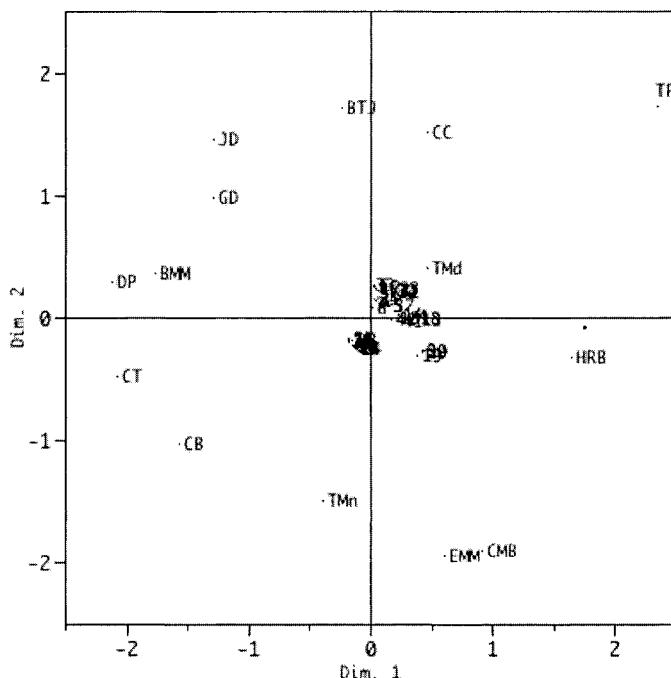


Fig. 1. Joint Configuration

the center of the final joint configuration. This result is similar to that of Borg and Groenen. Figure 2 shows the relative values of M_j as $\max_j M_j = 1$. The food items DP and CC had larger Brand-mass. The food item TMd had the smallest Brand-mass, and the smallest distance from each individual. s_i

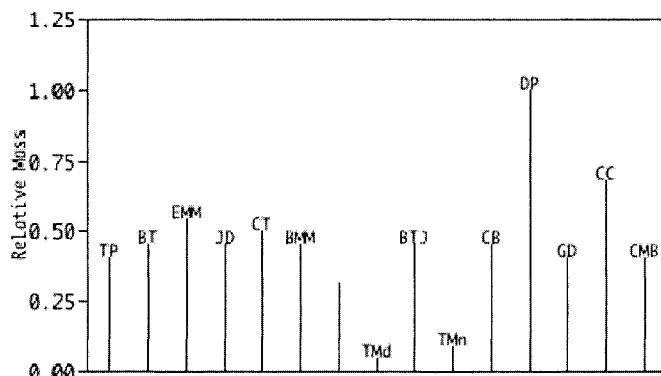


Fig. 2. M_j : Brand-mass of each food item

varied from 0.00 to 0.23, and the mean and standard deviation of them was 0.050 and 0.040, respectively. From normalization on a joint configuration and all ideal points being located at the center, the mean of square of distance is about 1. The preferences to the food items would rapidly decline with distances for many individuals, but they would decline milder for several individuals.

5 Conclusion

The gravity between object o_j and individual I_i is the exponential decreasing function of d_{ij} . And the configuration will become more locally clustered than that by other MDU methods. These tendencies are relaxed by introducing the expectation of D_{ij}^2 and the linearly constrained monotone function. The application indicated the proposed model and method recovered the hidden structure in the 15 food items.

We assumed that the variation of D_{ij} was expressed by individual variation only. These are several sub-models on the decomposition of σ_{ij} as mentioned. We must develop a measure that indicates which model is more appropriate.

Acknowledgements: The author acknowledges the valuable comments of two anonymous referees and the editors of this book.

References

- BOTTUM, M.S. (1989): Retail Gravity Model. *The Appraisal Journal*, 557, 166–172.
- BORG, I., and GROENEN, P. (1997): *Modern Multidimensional Scaling*. Springer, New York.
- DESARBO, W.S., KIM, J., CHOI, S.C., and SPAULDINGS, M. (2002): A Gravity-Based Multidimensional Scaling Model for Deriving Spatial Structures Underlying Consumer Preference Judgements. *Journal of Consumer Research*, 29, 91–100.
- GREEN, P.E. and RAO, V.R. (1972): *Applied Multidimensional Scaling*. Dryden Press, Hinsdale, IL.
- KRUSKAL, J.B. (1964): Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika*, 29, 115–129.
- RUSHTON, G. (1969): The Scaling of Locational Preferences. In: K.R. Cox and R.G. Gooedge (Eds): *Behavioral Problems in Geography: A Symposium*. Studies in Geography, 17, 192–227, Department of Geography, NorthWestern University.

Customer Relationship Management in the Telecommunications and Utilities Markets

Robert Katona¹ and Daniel Baier²

¹ Strategy and Business Architecture (Accenture Service Line),
Accenture, Campus Kronberg 1, D-61476 Kronberg, Germany

² Institut für Wirtschaftswissenschaften,
Brandenburgische Technische Universität Cottbus, D-03013 Cottbus, Germany

Abstract. Customer relationship management (CRM) gained an increasing popularity in theory and practice. It is a business philosophy, a holistic approach, which integrates a set of marketing, sales, and service capabilities. In Germany, CRM has entered the telecommunications and the utilities markets most recently. We will present in this paper, to which extent telecommunications companies are ahead with their CRM capabilities compared to companies in the utilities markets.

1 Conceptual background

1.1 From transaction to relationship marketing

The past two decades have witnessed a rapid emergence of relationship marketing in response to significant changes in the economic, social and technological environment of organizations (Palmer (2002)). Companies are being challenged by an increase in competitive intensity, changing consumer patterns and considerable advancements in information technology. These factors are the basis for a major shift from a transaction to a relationship oriented marketing in theory and practice (Grönroos (1994)). With the advent of relationship marketing, not merely the traditional acquisition of customers is crucial for the success of organizations, but in addition customer retention and loyalty is becoming exceedingly important (Diller (1996)). A series of theoretical and empirical studies show that customer loyalty is positively influencing long-term profitability (Peters (1999)). However, some of the major differences between the transactional and relational approach are shown in Table 1.

Although the interest in relationship marketing is continuously growing, a commonly accepted definition does not exist, neither for researchers nor for managers. The term "relationship marketing" was first introduced by Berry in 1983. He defines relationship marketing as "attracting, maintaining and enhancing customer relationships" (Berry (1983)). Many authors have contributed to define relationship marketing (e.g. Grönroos (1990), Ballantyne (1994), Morgan and Hunt (1994), Gummesson (2002)) emphasizing on the following aspects:

- Create, maintain and develop successful relational exchanges
- Exchanges of value for related parties (win-win situations)
- Strong and lasting relationships
- Interaction between the relationship actors.

Transactional focus	Relationship focus
Orientation to single sales	Orientation to customer retention
Discontinuous customer contact	Continuous customer contact
Focus on product features	Focus on customer value
Short time scale	Long time scale
Little emphasis on cust. service	High customer service emphasis
Limited commitment to meeting customer expectations	High commitment to meeting customer expectations
Quality is the concern of production staff	Quality is the concern of all staff

Table 1. The shift to relationship marketing (Christopher and McDonald (1995))

1.2 Operating relationship marketing with CRM

The relationship marketing philosophy is put into practice by means of customer relationship management (CRM) which in turn is based on database marketing. Link and Hildebrand developed in 1993 a first closed-loop approach for database marketing focusing on market analysis, marketing planning, and market reaction monitoring based on a dynamic and continuously updated customer database (Link and Hildebrand (1993)). The database contains individual customer as well as prospective customer level data which are used to build targeted, commercial relationships, to improve the cost-effectiveness of marketing programs and to stimulate sales and repeat purchases (For definitions of database marketing see e.g. Stone and Shaw (1987), Shani and Chalasani (1992), Link and Hildebrand (1993)).

However, database marketing is the underlying approach for CRM, which most recently gained an impressive popularity. Galbreath defines CRM as

- “Activities a business performs to **identify, qualify, acquire, develop and retain increasingly loyal and profitable customers.**
- CRM **integrates sales, marketing, service functions** through business process automation, technology solutions, and information resources to maximize each customer contact.
- CRM **facilitates relationships among enterprises, their customers, business suppliers, and employees.**“ (Galbreath (1998))

In this context, the company Accenture published the results of a comprehensive CRM study, which was conducted to quantify any correlation between CRM performance and financial performance. A major result of this study was the identification of 54 independent CRM capabilities grouped into five categories (Accenture (2000)).

Generating and applying customer insight: capturing the relevant information across all customer touch points, and using it to build a unique, fact based understanding of the customers needs. Customer insight allows a company to define most valuable customers and to determine how to work with them to maximize mutual value.

Developing customer offers: drawing on customer insight to configure products and services into differentiated solutions tailored to meet customers needs and intentions better than competitive alternatives.

Interacting with customers: enabling customers to interact seamlessly across all touch points and provide insightful and integrated actions (e.g. integrating marketing, sales and service; demonstrating insight into customer needs and preferences; providing consistent messages across each type of contact; reflecting the customers value to the company).

Integration of the organization: creating and nurturing the kind of environment that attracts, develops and retains the best customer skills and experience across the enterprise. The rational behind is, that human performance delivers a customer experience so unique that it cannot be replicated.

Integration of the enterprise: aligning customer-facing functions with all other functions - both internal and external to the organization - involved in the satisfaction of customer demands.

For each of these CRM categories, substantial statistical techniques (e.g. multiple regression and factor analysis) were used in order to derive the corresponding and independent CRM capabilities. To our knowledge, there exists no statistical validation of capabilities within other CRM oriented frameworks, like for example the 30R (Gummesson (1994)) or the 11C (Gordon (1998)). The statistical relevance of the Accenture CRM model made us choose this capability framework as a ground for our empirical study.

2 Empirical study

2.1 Research issue

In Germany, CRM has entered the telecommunications and the utilities markets most recently. Fixed calls in the telecommunications market as well as the electricity market were liberalized in 1998. Nevertheless, as rates of churn readiness indicate (Prof. Homburg and Partner (2002); dimap (2001)) today the companies in the telecommunications market (fixed and mobile) are facing a far more intense competition than the companies in the utilities markets. This observation brought us to examine, to which extend the companies in

the telecommunications markets are ahead with their CRM capabilities compared to companies in the utilities markets. For this purpose, we performed a survey in both markets based on a comprehensive questionnaire, which refers to the mentioned CRM capabilities. The study was conducted between July and October 2001.

2.2 Methodology

Procedures and sampling

A number of different associations for telecommunications and utilities companies in Germany provided us with comprehensive contact lists. We called over 400 companies in order to get in contact with marketing, sales or service executives and to ask for participation in our survey. Electronic and/or paper-based questionnaire surveys were mailed to a total of 237 companies. We received 78 completed questionnaires - representing a response rate of 33%. The proportions of groups responded are 49% from fixed telecommunications provider, 13% from mobile telecommunications provider, and 38% from utility companies. As the distribution of key characteristics are appropriately reflected, we assume that the three markets are quite properly represented by these three samples. The test persons were evaluating their own company in one of the three specific industries.

Questionnaire items

The questionnaire is grouped into the five CRM categories, which were described before. Each category covers a number of independent CRM capabilities resulting from the research by Accenture (Accenture (2000)). All survey questions are referring to these CRM capabilities in order to assess the level of fulfilling the corresponding CRM capability. Most of the 45 questions were based on multi-point evaluation scales between the values 0 or 1 up to 5. The stated opinion increases towards the maximum value of the scales with explanatory information at the two ends of each scale.

Analysis approach

With regard to the objective of our research the hypothesis of our empirical study was, that the level of meeting CRM capabilities is different, between companies in the fixed telecommunications market, companies in the mobile telecommunications market, and companies in the utilities market.

Data analysis proceeded in a three step approach for each CRM capability. We tested the homogeneity of variances with the Levene test to verify the one-way ANOVA assumption that the variances of the groups are all equal. Then, with one-way ANOVA we determined, whether significant differences exist among the group means. Pairwise multiple comparisons on the basis of the Scheffe test provided us the difference between each pair of means and indicated significantly different group means. We decided to choose the Scheffe test, which is relatively robust to violation of assumptions.

2.3 Hypothesis testing

In this paper we will present an excerpt of our empirical study focusing on one single CRM category namely customer insight (see section 1.2) and the corresponding CRM capabilities. In Table 2, these seven CRM capabilities are listed as well as the group means of fixed/ mobile telecommunications companies and utility companies.

The significance values resulting from the Levene test exceed .05, suggesting that with regard to all CRM capabilities the variances for the three groups are equal. For most of the CRM capabilities, the significance level of the F-test performing ANOVA is less than .05; that means, that at least one of the groups differs from the others. Though, for *customer data integration* and *customer segmentation* the significance level of the F-test is higher than .05. That the level of *customer data integration* and *customer segmentation* differs significantly between the groups is therefore not statistically verified.

CRM capability	Group Means			Levene test	One-Way ANOVA
	Fixed	Mobile	Utility	Sig.	Sig.
Internal data gathering	2,50	3,10	2,33	,712	,035
External data gathering	1,68	2,30	,93	,067	,003
Customer data integration	1,72	1,70	1,32	,303	,301
Customer data availability	1,95	2,80	1,34	,965	,000
Customer valuation	1,32	2,50	1,14	,982	,012
Customer segmentation	2,24	2,40	2,33	,341	,905
Data mining	,81	2,89	1,00	,665	,000

Table 2. Group means, Levene test, and one-way ANOVA for customer insight capabilities

The Table 3 lists the pairwise comparisons of the group means for the Scheffe test. Mean difference lists the differences between the sample means and Sig. the probability that the population mean difference is zero. This test was calculated with a significance level of .05.

The first CRM capability identified within the CRM category of generating and applying customer insight is the collection of customer information from internal sources. The level of *internal data gathering* within the mobile group is, according to the results in Table 3, significantly higher than in the utility group.

The acquisition of data from external sources (e.g. third party vendors) to enhance the customer database (*external data gathering*) is significantly more extensive with mobile respectively fixed communications companies than with utility companies.

Customer data integration is the expression for the integration of customer

identifiers and customer information across transaction (operational), interaction, and analytical data warehouses. As to the results of Table 3, it is not statistically verified that the level of *customer data integration* differs significantly between the groups.

Information sharing across channels and business units (*customer data availability*) is significantly on the highest level within mobile communications companies followed by fixed communications companies and then by utility companies.

Customer valuation is a data analysis that calculates a customers value with certain measures of value that can be used to profile customers and set customer strategies. Within the group of fixed communications enterprises *customer valuation* is significantly to the highest extent compared to the other two groups.

The CRM approach includes also the capability of *customer segmentation*. Our research reveals that a difference regarding the level of *customer segmentation* is not statistically verified.

Data mining represents a key CRM capability obtaining customer insights. As churn is a critical issue in these three industries, data mining plays a crucial role. It can be employed to analyze why customers churn and which customers are most likely to churn in the future. Customer data can be utilized to monitor and highlight customers who may, by the signature in their usage pattern, be thinking of migrating. This information is helpful for marketing departments in order to better target retention campaigns. This progressive data analysis is significantly most sophisticated within the group of mobile communications companies.

CRM capability	Group I	Group J	Mean Diff. I-J	Sig.
Internal data gathering	Fixed	Utility	,17	,691
	Mobile	Utility	,77	,035
External data gathering	Fixed	Utility	,75	,034
	Mobile	Utility	1,37	,007
Customer data integration	Fixed	Utility	,40	,329
	Mobile	Utility	,38	,626
Customer data availability	Fixed	Utility	,60	,037
	Mobile	Utility	1,46	,000
Customer valuation	Fixed	Utility	,19	,830
	Mobile	Utility	1,36	,013
Customer segmentation	Fixed	Utility	-,10	,946
	Mobile	Utility	,07	,988
Data mining	Fixed	Utility	-,19	,898
	Mobile	Utility	1,89	,004

Table 3. Scheffe test for customer insight capabilities

2.4 Discussion

Internal data gathering is an essential part of the data generation process. The collection of customer information from internal sources reflects direct transactions or interactions between a company and its customers. In a number of telecommunications companies, data are effectively captured from multiple sources and for most customer interactions, whereas in most utility companies data collection is transactional system focused (e.g. billing system).

The acquisition of data from external sources (*external data gathering*) becomes more valuable with the increase of data quality and the accessibility of necessary information. At telecommunications companies, external customer-level data is used on an ad hoc basis, but is not periodically refreshed. Third party data supplement internal data for specific purposes such as adding or validating core customer data.

A real challenge communications and utility companies face includes *customer data integration*. Today, there is little consistency of customer data, and the processes for linking multiple data are mostly manual. In both industries, a typical company shares only some data among interaction, transaction, and analytical data systems.

Customer data availability requires data architectures to support customer data access and analysis for different departments. Telecommunications companies generally have quite common interaction databases for most of the contact channels. Limited integration between interaction systems or rather stand-alone interaction systems make it difficult for utility companies to apply any customer insight.

Understanding the value of customers as well as prospective customers (*customer valuation*), helps companies to improve the profitability of interactions. Telecommunication companies are performing current and future valuation using dynamic data, whereas the majority of utility companies evaluate customers at segment level.

The experiences in *customer segmentation* are on a high level with advanced analytics, which lead to customer-specific strategies and value propositions.

Data mining is used in mobile telecommunications companies on the basis of a variety of techniques with application to customer strategy and future trend prediction. Fixed line communications companies focus mainly on the acquisition, but especially on the retention/defection of customers. The primary focus of data mining within utility companies is on tactical objectives with limited or no retention-focused models.

Our entire study reveals that with regard to most of the CRM capabilities the telecommunications industry is in the lead compared to utility companies. This is underlined by table 4, which summarizes the key results of three chosen and remarkable capabilities associated with the CRM category *customer interaction*. *Channel capacity management* is the ability to examine capacities across channels, and to direct users to selected channels so that

CRM capability	Fixed	Group	Levene	One-Way
		Means	test	ANOVA
Channel capacity management	2,02	3,00	1,77	,289 ,020
Channel integration	2,34	2,70	1,90	,063 ,014
Cust. infor. availability	1,92	2,60	1,77	,517 ,048

Table 4. Group means, Levene test, and one-way ANOVA for a choice of customer interaction capabilities

optimal operating levels can be achieved. The main task of *Channel integration* is featured by the synchronization of channels to ensure consistent customer treatment. And, *Customer information availability* stands for the ability to provide relevant customer information to a customer contact point.

However, utility companies can improve their CRM capabilities to reach the level met in the telecommunications market. The success of extending CRM capabilities depends among other things on a deeper understanding and knowledge of the holistic CRM approach. This framework is composed of integrated CRM capabilities, and therefore it is important to analyze furthermore interdependencies of the CRM capabilities - a task for further researches.

References

- ACCENTURE (2000): *How Much Are Customer Relationship Management Capabilities Really Worth? What Every CEO Should Know.* <http://www.accenture.com/>.
- BALLANTYNE, D. (1994): Marketing at the Crossroads. *Asia-Australia Marketing Journal*, 2(1), 1-7.
- BERRY, L.L. (1983): Relationship Marketing. In: L.L. Berry, G. Shostack, and G.D. Upah (Eds.): *Emerging Perspectives on Service Marketing*. Chicago, IL: American Marketing Association, 25-38.
- CHRISTOPER, M. and MCDONALD, M. (1995): *Marketing*. MacMillan Press Ltd., Houndsills et al..
- DILLER, H. (1996): Kundenbindung als Marketingziel. *Marketing ZFP*, 2, 81-93.
- DIMAP (2001): Der Markt für Telekommunikation in Deutschland: Kenntnis-Erfahrungen-Bewertungen. Ergebnisse einer Repräsentativerhebung. www.vatm.de/images/dokumente/umfrage.pdf.
- GALBREATH, J. (1998): Relationship management environments. *Credit World*, 87 (2), 14-22.
- GRÖNROOS, CH. (1990): *Service Management and Marketing*. Lexington Books, Lexington, Mass.
- GRÖNROOS, CH. (1994): From Marketing Mix Marketing to Relationship Marketing: Towards a Paradigm Shift in Marketing. *Management Decision*, 32(2), 4-20.

- GUMMESON, E. (2002): Relationship Marketing in the New Economy. *Journal of Relationship Marketing*, 1(1), 79–94.
- PROF. HOMBURG and PARTNER (2002): Marketing im Energiemarkt. Ein State of Practice Bericht für den Privatkundenbereich www.homburg-und-partner.de/knowhow/evupk180202.pdf.
- LINK, J. and HILDEBRAND, V. (1993): *Database Marketing und Computer Aided Selling*. Verlag Vahlen, München.
- MORGAN, R.M. and HUNT, S.D. (1994): The commitment-trust theory of relationship marketing. *Journal of Marketing*, 58(7), 20–38.
- PALMER, A. (2002): The Evolution of an Idea: An Environmental Explanation of Relationship Marketing. *Journal of Relationship Marketing*, 1(1), 79–94.
- PETERS, S.I. (1999): *Kundenbindung als Marketingziel. Identifikation und Analyse zentraler Determinanten*. Gabler Verlag, Wiesbaden.
- SHANI, D. and CHALASANI, S. (1992): Exploiting Niches Using Relationship Marketing. *Journal of Consumer Marketing*, 9, 33–42.
- STONE, M. and SHAW, R. (1987): Database Marketing for Competitive Advantage. *Long Range Planning*, 20, 12–20.

Strengths and Weaknesses of Support Vector Machines Within Marketing Data Analysis

Katharina Monien and Reinhold Decker

Department of Economics and Business Administration,
University of Bielefeld, D-33615 Bielefeld, Germany

Abstract. Support vector machines are not only promising for solving pattern recognition tasks but have also produced several successful applications in medical diagnostics and object detection to date. So it is just natural to check whether this methodology might also be a helpful tool for classification in marketing and especially in sales force management. To answer this question both strengths and weaknesses of support vector machines in marketing data analysis are investigated exemplarily with special attention to the problem of selecting appropriate kernel functions and determining the belonging parameters. Difficulties arising in this context are illustrated by means of real data from health care.

1 Introduction

In many industries sales force is an important but also very cost-intensive element of marketing. Focussing on those customers presumed to be worth an intensive treatment logically becomes indispensable to most firms in competitive market environments. Therefore, methods are required to separate "valuable" customers from "non-valuable" ones in appropriate data bases in order to be able to classify new potential customers correctly. If one succeeds in this respect to a sufficient extent costs in customer relationship management can be reduced significantly. Against this background the paper on hand investigates the suitability of support vector machines within the decision making process concerned.

Support vector machines are a class of methods which stem from machine learning theory and have been introduced by Boser et al. in 1992. Meanwhile, they have turned out to be a successful technique for pattern recognition, e.g. in medical diagnostics and object detection. In contrast to this applications in marketing are rather seldom up to now. Viaene et al. (2001) have used support vector machines for feature extraction in marketing by iteratively applying them to the underlying database. The property of reducing the number of input vectors was used in Yang (2002). In this paper the selection of important customers served as a starting point for the subsequent case-based reasoning focusing on the development of individual marketing strategies. As shown in Cheung et al. (2003) even content-based recommender systems can be constructed with the help of support vector machines. Finally, Decker and

Monien (2003) used this methodology to classify new customers by starting from the feature vectors of regular customers. In this paper, as an extension of these basic considerations, we concentrate on the selection of adequate kernels and parameters from an application-oriented point of view.

The remainder of the paper is structured as follows: In section 2 we are going to sketch the underlying methodology before applying it to pharmaceutical data in section 3. After the discussion of the technical problems mentioned before in section 4 we complete the paper with some final remarks and an outlook to future work.

2 Theoretical foundation

The aim of support vector machines is to compute a hyperplane or rather a decision function $f : \mathbb{R}^n \rightarrow \{1, -1\}$, based on given input vectors (observations), such that f classifies new observations as correct as possible. The basic elements to achieve this aim are margins, duality, and kernels. Applying alternative kernels allows to cover several classification architectures like for example neural networks (cf. Bennett (2000)).

Given l input vectors $\mathbf{x}_i \in \mathbb{R}^n$ with corresponding class labels $y_i \in \{1, -1\}$ the separating hyperplane $\{\mathbf{x} | \mathbf{w}\mathbf{x} + b = 0\}$ has to be determined in such a way that the margin between class 1 and class -1 is maximized as shown in Figure 1, where $\mathbf{w}\mathbf{x}$ denotes the simple dot product of vectors \mathbf{w} and \mathbf{x} . Considering that the distance of a point \mathbf{x}_i from the hyperplane, defined by

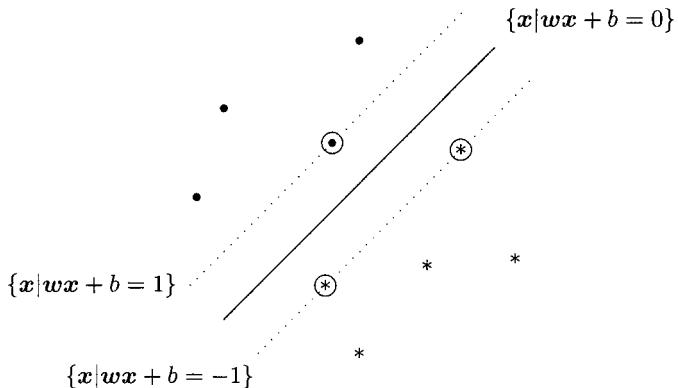


Fig. 1. Separating hyperplane (cf. Schölkopf and Smola (2002))

\mathbf{w} and b , is computed as $d(\mathbf{x}_i; \mathbf{w}, b) = \frac{|\mathbf{w}\mathbf{x}_i + b|}{\|\mathbf{w}\|}$, the margin is equal to

$$\min_{i:y_i=1} (d(\mathbf{x}_i; \mathbf{w}, b)) + \min_{i:y_i=-1} (d(\mathbf{x}_i; \mathbf{w}, b)).$$

After some transformations the maximization of this margin leads to the following optimization problem:

$$\min \frac{1}{2} \|\mathbf{w}\|^2$$

with

$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1 \quad \forall i \in \{1, \dots, l\},$$

where \mathbf{w} denotes a normal vector of the separating hyperplane and b indicates the distance from hyperplane to origin.

To solve this task it suffices to deal with the dual optimization problem (for more details see Vapnik (1998))

$$\max W(\boldsymbol{\alpha}) = \max \left\{ \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j \right\}$$

with

$$\sum_{i=1}^l \alpha_i y_i = 0 \quad \text{and} \quad \alpha_i \geq 0 \quad \forall i.$$

Those input vectors \mathbf{x}_i with $\alpha_i > 0$ are called support vectors and are the only ones to influence the position of the hyperplane.

An important property of support vector machines is the extendibility to non-linear separation problems. The trick is to map the original input vectors \mathbf{x}_i to a higher dimensional feature space where they can be separated linearly. In order to avoid complex computations Boser et al. (1992) introduced kernels which substitute the dot product of the mapped input vectors by means of Mercer's theorem. For a detailed description of this method see , e.g., Schölkopf and Smola (2002) chapter 2 and 7. The resulting optimization problem looks like this:

$$\max W(\boldsymbol{\alpha}) = \max \left\{ \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \right\}$$

with

$$\sum_{i=1}^l \alpha_i y_i = 0 \quad \text{and} \quad 0 \leq \alpha_i \leq C \quad \forall i,$$

where C is a constant which can be used to emphasize either the maximization of the margin or the minimization of the errors (cf. Schölkopf and Smola (2002)) and $k(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function that has to be chosen by the analyst. The most common specifications of $k(\cdot, \cdot)$ (cf. Schölkopf et al. (1999)) are the polynomial kernel of degree $d \in \mathbb{N}$ $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \mathbf{x}_j + 1)^d$, the radial

basis kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma|\mathbf{x}_i - \mathbf{x}_j|^2)$ with kernel width $\gamma \in \mathbb{R}$, and the neural network kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\kappa \mathbf{x}_i^\top \mathbf{x}_j + \theta)$ with parameters κ and $\theta \in \mathbb{R}$.

When the optimization of a support vector machine is completed, a new data point \mathbf{x} can be assigned to one of the two classes by means of classification function

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^l \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b\right).$$

The computed values of α_i give a clue to the influence of input vectors \mathbf{x}_i on the position of the resulting hyperplane. By choosing kernel parameters d , γ , κ , and θ as well as trade-off parameter C the user is able to control – to a certain degree – the number of support vectors and the goodness of separation, i.e. the number of misclassified input vectors. Increasing, for example, the parameters C , γ , and d leads to a decreasing number of misclassified input vectors, but might also cause overfitting.

3 Application to pharmaceutical data

In this section support vector machines are applied to a real data set from pharmaceutical industry containing 2922 observations with several variables. The data is capturing the extent to which general practitioners prescribe a certain medical drug and was divided into a training sample with 2192 observations and a test sample with 730 observations. Each observation is characterized by demographical variables concerning the catchment area of the respective general practitioner. In particular, these are the share of foreign households, the share of households with a low (as well as high) social status, the share of one family houses, the share of seven and more family houses, and the total population of the community or city concerned.

The empirical study focusses on the examination of the methodology on hand with respect to its ability of reducing the target group of pharmaceutical industry (in our case the general practitioners) to those who are worth to be visited by sales force for instance. Therefore, two classes are distinguished. The first one contains those physicians who comparatively often prescribe the relevant drug (class code -1) and who therefore are potential customers and thus the relevant target group. In contrast to this the members of class two (encoded with 1) rather rarely prescribe the drug under consideration. A special treatment of this group within marketing appears to be less promising. Finally, it has to be mentioned that our data are highly overlapped which makes classification a comparatively difficult task.

Using this data different support vector machines including the kernels mentioned in section 2 have been trained. Particular emphasis was laid on

the radial basis kernel. The corresponding results for both the training as well as the test data applying different parameters are depicted in Table 1.

For a given C we have chosen parameter γ of the radial basis kernel such that we get an acceptable result for both the training and the test data. At the same time the number of support vectors ought to be as small as possible. This number symbolizes the data that is necessary to represent the input data with given parameters. It turned out that choosing C relatively high (equal to 50000 or 100000) produces the best results with respect to these objectives. Particularly the numbers of support vectors (985 or 984 for $\gamma = 1$ or $\gamma = 0.5$) seem to be justifiable against the fact of highly overlapping data which must be the cause of a hit rate in the test data not higher than approximately 67 % as well.

Kernel	C	Parameter ($\gamma, d, \kappa, \theta$)	Number of sup. vectors	Hit rate	
				training data	test data
radial basis	1	500	2190	98.8%	66.4%
	1	0.0001	1755	71.8%	63.3%
	100	60	2074	99.1%	66.2%
	100	0.0005	1643	82.0%	58.2%
	50000	1000	2191	99.1%	66.4%
	50000	1	985	96.7%	65.8%
	100000	50	2021	99.1%	66.3%
	100000	0.5	984	96.2%	65.8%
polynomial	100	2	1469	66.7%	66.5%
neural network	10000	$10^{-14} - 10^{-6}$	1492	66.5%	66.6%
dot product	1000	—	2192	63.2%	64.6%

Table 1. Selected results of applying alternative kernels

As a benchmark, we also applied discriminant analysis which is traditionally used for classification in marketing. In the best case (using 3-nearest-neighbor method) we attained a maximum hit rate of 76.6 % for the training and 56.7 % for the test data. Furthermore, we compared the results of the support vector machines with a decision tree approach to additionally refer to a nonlinear benchmark. In this case the results were similar to those of the support vector machines, at least when non radial basis kernels are used, namely 66.2 % for the test and 66.4 % for the training data. However, the latter hit rate is significantly lower than those of the support vector machines with a radial basis kernel.

Starting from the results above we carried out an extensive simulation in order to examine the misclassification rate depending on different settings of γ in the radial basis kernel. In Figure 2 this dependency is visualized for both the training and the test sample. In both cases it becomes apparent

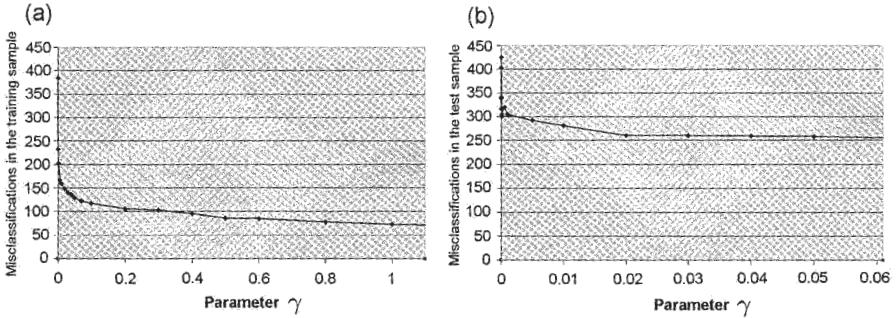


Fig. 2. Misclassifications in the training sample (a) depending on $\gamma \in [0; 1]$ as well as in the test sample (b) depending on $\gamma \in [0; 0.05]$ (each with $C = 50000$)

that the higher γ is, the lower the number of misclassifications becomes. In both samples the main variation of misclassifications in dependence of γ concentrates on a comparatively small range. Increasing γ reduces the number of misclassified data points in the training sample. If γ exceeds 35.0 the misclassification rate equals 19.

In another study based on a synthetic sales data set (for more details see Decker and Monien (2003)) the corresponding plot for the training sample looks quite similar, whereas the number of misclassifications increases for the test sample when γ increases. Further on, for both the pharmaceutical and the synthetic data the number of resulting support vectors increases when γ increases. Therefore, the user has to weigh up carefully the pros and cons when adjusting parameter γ .

4 Further remarks on parameter selection

Adapting the parameters of a support vector machine such that both the final number of support vectors and the number of misclassifications take a value near minimum can occupy a lot of time and requests some experience. Focussing on the trade-off parameter C allows to minimize misclassifications during the training step. Recently, Schölkopf and Smola (2002) propagated a default value of $C = l \cdot 10$ which could be confirmed by our own analysis as well, at least for sales data of the present type.

Unfortunately, no such recommendation can be given with respect to the selection of the kernel parameters. For both the polynomial kernel and the radial basis kernel it turned out that the adaptation of the hyperplane to the training data significantly depends on degree d and parameter γ respectively. For the pharmaceutical data the value of γ , which produces the best hit rate in our investigation, is equal to 1. In contrast to this the synthetic data men-

tioned above implies $\gamma = 10^{-6}$. Obviously, the parameter selection strongly depends on the underlying data set. Large values of γ must not always be the best choice. With regard to the neural net parameters κ and θ the user has to take into account that, depending on the available data set, only certain combinations of κ and θ are valid for this kernel due to the condition of Mercer, which has to be satisfied for the introduction of kernels (cf. Vapnik (1998)). All kernel parameters have to be chosen carefully with respect to the hit rate in the training as well as in the test data set to insure an optimal customer classification so that a better customer support becomes possible.

Finding optimal kernel parameters for a given data set is also a question of computational time. Using, for example, an implementation of support vector machines provided by Royal Holloway in cooperation with AT&T (cf. Saunders et al. (1998)) the computation of the optimization problem with a polynomial kernel took a few days for large values of C . In contrast to this the radial basis kernel just required a few seconds for most combinations of C and γ under consideration. This might be traced back to the fact, that support vector machines with a radial basis kernel need a smaller number of iterations due to the curve progressions. Hence, at least for highly overlapping data as to be expected regularly in personnel selling or direct marketing the radial basis kernel appears to be more practicable for customer classification tasks.

5 Conclusions and outlook

The purpose of this paper was to empirically motivate and discuss the use of support vector machines within marketing data analysis and particularly within sales force management. Special emphasis was put on the kernel selection and the determination of corresponding parameters.

Systematically classifying customers (e.g. general practitioners) with the help of support vector machines can significantly improve sales force efficiency and, as a consequence, reduce expenditures by exclusively focussing on the "valuable" customers. Systematically arranging the sales force depending on the results on an upstream customer classification can defuse the "bottleneck" of customer management. In particular the promising improvement of the hit rate of support vector machines compared to traditional discriminant analysis of approximately 10 %, together with its flexibility in separating non-linear data, show the necessity of further investigations on this topic. With the possibility of substantial interpretations of the output, support vector machines might develop to a powerful supplement or even substitute of the traditional classification toolbox in marketing.

But especially in direct marketing and sales force management a non-negligible problem arises from the fact that several important characteristics

of customers, such as gender and level of education, can not be measured directly on a metric scale. Against this background and taking into account the principally high potential of support vector machines for marketing and marketing research it is a great challenge to think about their adaption to binary predictor variables. Another important challenge to research on this methodology is the adequate processing of real data. In this context, especially the absence of general recommendations with respect to kernel and parameter selection might hamper a broader diffusion of support vector machines within the marketing community in the near future.

References

- BENNETT, K. and CAMPBELL, C. (2000): Support Vector Machines: Hype or Halleluja?, *SIGKDD Explorations*, 2, 2, 1–13.
- BOSER, B.E., GUYON, I.M., and VAPNIK, V.N. (1992): A Training Algorithm for Optimal Margin Classifiers. In: D. Haussler (Ed.): *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. ACM Press, Pittsburgh, 144–152.
- CHEUNG, K.-W., KWOK, J.T., LAW, M.H., and TSUI, K.-C. (2003): Mining Customer Product Ratings for Personalized Marketing. *Decision Support Systems – Special Issue on Web Data Mining*, 35, 2, 231–243.
- DECKER, R. and MONIEN, K. (2003): Support-Vektor-Maschinen als Analyseinstrument im Marketing am Beispiel der Neukundenklassifikation. *Der Markt*, 42, 1, 3–13.
- SAUNDERS, C., STITSON, M.O., WESTON, J., BOTTOU, L., SCHÖLKOPF, B., and SMOLA, A. (1998): *Support Vector Machine - Reference Manual*, Royal Holloway Technical Report CSD-TR-98-03, Royal Holloway.
- SCHÖLKOPF, B., BURGES, C., and SMOLA, A. (1999): Introduction to Support Vector Learning. In: B. Schölkopf, C. Burges, and A. Smola (Eds.): *Advances in Kernel Methods: Support Vector Learning*, Cambridge, MIT Press, 1–15.
- SCHÖLKOPF, B. and SMOLA, A. (2002): *Learning with Kernels*, MIT Press, Cambridge.
- VAPNIK, V.N. (1998): *Statistical Learning Theory*, Wiley, New York.
- VIAENE, S., BAESENS, B., VAN GESTEL, T., SUYKENS, J., VAN DEN POEL, D., DEDENE, D., DE MOOR, B., and VANTHIELEN, J. (2001): Knowledge Discovery in a Direct Marketing Case Using Least Squares Support Vector Machines. *International Journal of Intelligent Systems*, 16, 9, 1023–1036.
- YANG, Q. (2002): Towards Statistical Planning for Marketing Strategies. In: M. Ghallab, J. Hertzberg, and P. Traverso (Eds.): *Proceedings of the Artificial Intelligence Planning Conference*. Menlo Park, AAAI Press.

Classification of Career-Lifestyle Patterns of Women

Miki Nakai

Department of Social Sciences, College of Social Sciences,
Ritsumeikan University, 56-1 Toji-in Kitamachi, Kita-ku, Kyoto 603-8577 Japan

Abstract. This paper attempts to classify women's lifestyles according to similar patterns in their commitment to four aspects of life. Using cluster analysis, we analyze the occupational history and family data, as well as the materialistic affluence and cultural consumption of 644 women, aged between 30 and 49, taken from a national sample in Japan. Career-lifestyle typology of 8 clusters are presented. By multiple correspondence analysis of both these clusters and the social stratification positions, we reveal that the differences in lifestyle commitment are closely related to the women's social background and position. Our findings suggest that focusing not only on paid work but also on devotion to other aspects of life helps identify women with similar career-paths but different life-priorities.

1 Introduction

In most sociological research that focuses on women's work-family issues from a life course perspective, career typology focuses exclusively on paid work. Only such events as entry into or exit from the workforce are utilized. In such cases, categories of career-paths such as 'continuous working', 'intermittent working' or 'retiring after marriage' are often used (Moen (1996)). However, each of the categories of this sort of typology is comprised of various socioeconomic groups of women (Nakai and Akachi (2000), Crompton (1999)). In addition, classification of various women into these typical categories is sometimes arbitrary, especially for the women with atypical career paths.

The interface between work and family is, of course, an important subject to study. In Japan today, most women tend to work after finishing secondary or tertiary education, a trend seen in western industrial countries for both men and women. However, the various aspects of life should be taken into account when considering the complexities of women's lifestyles as well as their career-paths. This diversity is caused partly by women's greater freedom and availability of choice, and partly by the existence of obstacles in their lives (Esping-Andersen (1999), Hakim (1997), Oppenheimer (1982)).

From a comparative perspective, the difference is observed between Japan and the US and European countries in the relationship between education and labour force participation. In the Western societies, higher levels of female educational attainment have been associated with greater labour force participation during the last several decades, whereas in Japan, there is no

linear relationship, or rather, the association tends to be reversed (Brinton (1993)). In this paper, we consider how women's educational level affects their lifestyle as well as their occupational career.

Furthermore, classification based not only on occupational career but also on other aspects of life enables us to examine whether some '*a priori*' assertion is acceptable. It is repeatedly asserted that a rise in female labour force participation has caused a decline in birth rates despite the fact that OECD countries with relatively low female labour force participation have also found the downturn in the fertility rate for the past decade. Likewise, it is also alleged that an increasing number of women are likely to be childless because of their pursuit of fulfillment of post-materialistic and/or materialistic values. However, most of these arguments are not empirically appraised.

Moreover, until now, the arguments about the relationship between women's lifestyle types and their social backgrounds have tended to be superficial or inconsistent.

The aim of this paper is to classify women's lifestyles in terms of the differences in commitment to four important aspects of life. These are: the commitment to the workforce; the commitment to the family domain; the commitment to materialistic affluence and the commitment to post-materialistic values. These four aspects of life determine women's lifestyles as per the cluster analysis revealed later in this paper. The relationships between the lifestyle clusters and social stratification are also examined.

2 Data

The data comes from a representative survey conducted in 1995 of the social stratification and social mobility of Japanese society. Out of approximately 4000 men and women sampled aged 20-69, 2653 were successfully interviewed. Using cluster analysis, we analyzed the occupational history and family data, as well as two other aspects of life, of a subsample of 644 women aged between 30 and 49.

There are three reasons why our focus is primarily on this age range. First of all, we are interested in identifying recent trends amongst women. The respondents here were born and raised in the more unconventional society of the postwar period. Therefore, they are appropriate subjects of this study. Secondly, this age bracket is critical because in Japan the participation of the female labour force is strikingly different from that of other countries, giving rise to the so-called "M-shaped" working pattern. There is a drop in labour force participation for Japanese women in the late twenties and a resurgence by the late thirties, producing the M-shaped curve. This means that women's lifestyles diverge during the period between 30 and 50 years-of-age and the distinction among women becomes more noticeable at that period of life. Finally, utilizing the data of women of this age range allows us to analyze the successive occupational history of relatively young cohorts.

We use four concepts as primary lifestyle choice considerations for women:

- (1) The commitment to the workforce, or devotion of time and energy in the workforce. This was measured by the percentage of time spent in the workforce since leaving school.
- (2) The commitment to the family domain, or devotion of their resources to the family. This was measured by the number of children that they had.
- (3) The commitment to materialistic affluence. This was measured by the number of household items that they possessed out of a specific list of five items.
- (4) The commitment to post-materialistic values. This was measured by the composite score of a sum of seven questions which were already known to reflect the concept of post-materialistic cultural activities.

As the commitment to the workforce is a percentage variable, an inverse sine (arcsine) transformation is applied. The material affluence of a household is measured by the number of items possessed out of the following: car, piano, living-room furniture, works of art and house ownership. The development of cultural activities that embody post-materialistic values is measured by the sum of the responses to the following question regarding to seven cultural activities. The interviewee is asked how often she participates in the following activities: (a) classical music performances and concerts, (b) museums and art exhibitions, (c) traditional Japanese kabuki, noh play or Japanese puppet show (bunraku) performances, (d) flower arrangement, tea ceremony, calligraphy, (e) Japanese poetry (tanka or haiku), (f) community or voluntary work, and (g) reading novels, or books about history. Before adding up, responses on the five-point scale to seven variables were standardized to have unit variance.

3 Results

3.1 Women's career-lifestyle clusters

For purposes of differentiation, cluster analysis is applied. We used a hierarchical clustering method and chose Ward's method in this application. Indices of these four variables were standardized to have zero mean and unit variance over all the respondents, as they were not of the same scale values.

Eight groups of women were identified after the cluster analysis was applied. The eight-cluster solution was adopted based on careful consideration of hierarchical tree and the information of the increase of the proportion of the total variance (R-square) and the change of pseudo t-square statistics, as well as on careful comparison of the interpretation of the characteristics of each cluster at several selected levels. The proportion of the total variance explained by these eight groupings was 0.609. Using listwise deletion, the number of valid cases was reduced to 578. Table 1 shows the actual number of cases in each of these eight groups, as well as their percentage of the

Cluster	Description	Cases	(%)
A	Continuous working without children	83	(14.4)
B	Continuous working and child rearing	76	(13.1)
C	Housewife without household possessions	93	(16.1)
D	Continuous working and child rearing enjoying both cultural activities and household possessions	93	(16.1)
E	Child-rearing housewife with neither household possessions nor participation in cultural activities	46	(8.0)
F	Housewife with neither children nor participation in cultural activities	87	(15.1)
G	Child-rearing housewife with household possessions	74	(12.8)
H	Housewife enjoying both cultural activities and household possessions	26	(4.5)

Table 1. Career-lifestyle patterns.

whole. Women are divided into distinct groups, referred to as career-lifestyle clusters.

Three clusters (A, B and D) are characterized as the women who devote a large part of their life to paid work in the labour market. These three clusters combined make up 44% of the 578 women interviewed. They mostly consist of women who, according to conventional career typology, have been classified as 'continuous working'-type women. Cluster A represents women who choose childlessness and/or singleness, and makes up 14% of the women interviewed. Cluster B is characterized by the women's greater commitment to the workforce throughout their lives, possession of fewer household items, and less-developed cultural activities. Cluster D is characterized by the women's commitment to work, a higher-than-average number of children, more household possessions, and preference for cultural activities.

The other five clusters (C, E, F, G and H) indicate a lesser commitment to work, and often include women who work part-time. The five clusters combined make up 56% of the 578 women. These five groups include the majority of women who either work intermittently or who have left the work-force (for example, after marriage or the birth of their first child). It may reflect women's balancing of family life and work, or family responsibilities. However, this also suggests that most of the women of these groups are disadvantaged in the labour market by low-paid or non-career part-time jobs. Cluster E is characterized by the women's extensive commitment to parenting, but fewer household possessions and fewer cultural activities. The patterns of both occupational career and the possession of household of cluster C are similar to cluster E, but the difference between these two clusters is the women's devotion to child rearing. Cluster H is relatively small but it is remarkable for the women's significant commitment to cultural activities.

3.2 Women's career-lifestyle patterns and social background

The above career-lifestyle patterns seem to reveal the differences in women's social backgrounds, as well as their preferences and life-priorities. The respondents' social origins and their educational levels are associated with the possibilities that are open to them as individuals, which may well be the determining factors of their lifestyles. Furthermore, it is expected that these clusterings are also related to the respondents' husband's occupational position. To clarify this issue, multiple correspondence analysis (MCA) is applied. The variables included are:

- father's occupational status: professional (**F-prof**), managerial (**F-mang**), clerical (**F-cler**), sales (**F-sale**), skilled manual (**F-skil**), semi-skilled manual (**F-sskl**), non-skilled manual (**F-nskl**), agricultural (**F-agri**)
- father's education: primary, secondary, tertiary
- respondent's education: primary, secondary, tertiary
- husband's occupational status: professional (**H-prof**), managerial (**H-mang**), clerical (**H-cler**), sales (**H-sale**), skilled manual (**H-skil**), semi-skilled manual (**H-sskl**), non-skilled manual (**H-nskl**), agricultural (**H-agri**), not-married (**Single**)

The words in bold in parentheses are used to represent each occupational category, and abbreviation F and H stand for the father's and husband's indicator in the figure respectively.

Figure 1 (overleaf) represents the position of the different groups in two-dimensional space. The points depicting the fathers' occupational status form a semi-circle when viewed collectively. The main dimension, Dimension 1 (horizontal), primarily correlates with the fathers' occupational status and educational achievements.

Cluster A (representing work-centered and childless women) and cluster H (mostly housewives with active cultural participation) are associated with the father's high occupational and educational levels, as well as their own high educational levels. On the other hand, clusters B and E (women having a higher-than-average number of children but fewer household possessions and little participation in post-materialistic activities) clearly correspond to the father's low education, as well as to the father's relatively low social status which would be the case, for example, with agricultural and non-skilled manual work. Therefore, the above results show that amongst those who are classified into one of the work-centered lifestyle clusters, there are women who are from both disadvantaged social backgrounds (B) and privileged social backgrounds (A). The results also show that amongst those who are classified into one of the non-work-centered lifestyle clusters, there are women from privileged social backgrounds who are actively involved in cultural activities (H), as well as women of deprived social origins who have little opportunity for participation in cultural activities (E). Some argue that full-time housewives

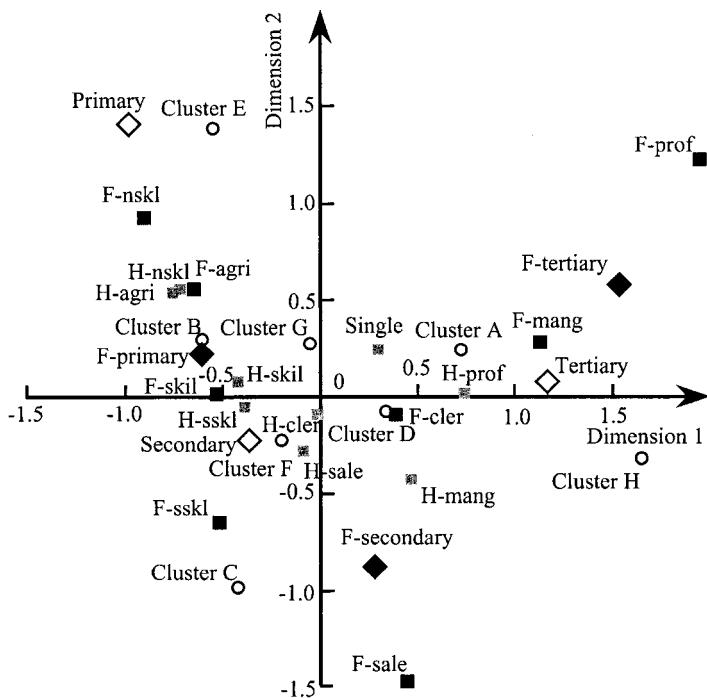


Fig. 1. Career-lifestyle clusters and social background, education and husband's status: plot of multiple correspondence analysis.

are the privileged who, in recent years, can choose not to work (Yamada (1999)). However, the evidence of our study does not necessarily support that position.

With regards to the husband's variable, cluster A, professionals, and singleness are associated. Similarly, the association among cluster H and having husbands of managerial position is shown, as these points are located in the same direction from the origin. These women with advantaged origins tends to be married to men with higher occupational status using their cultural capital, and that, in turn, could facilitate their post-materialistic activities more. Cluster B seems to depict the married women assumed by the 'Douglas-Arisawa effect', which hypothesizes that the decision as to whether or not a woman becomes a paid worker is influenced by her husband's income. The lower the husband's income, the more the wife tends to choose paid work (Douglas (1934), Arisawa (1956)). The results of the MCA suggest an association between the women of cluster B and the husband's low occupational status, as well as between cluster B and their own relatively low social origins. Nowadays, it is sometimes claimed that the argument of negative correlation between the husband's income and the wife's participation in paid work does

not hold true any more. Some argue, instead, that work-centred women are the ones who can provide most adequately for their children in a variety of ways, or that they are the elite who are on a level with men (Mifune and Shigekawa (1999)). However, in reality, a certain proportion of women whose husbands' occupations are of a low level continue to work. This shows that the relationship mentioned in the Douglas-Arisawa effect is still valid. However, some women whose fathers have relatively high occupational and educational levels, and who themselves have completed tertiary education, decide to continue to work even after marriage and childbirth (D).

On top of that, the women pursuing their own career (A) also tend to enjoy more post-materialistic attributes than materialistic ones. Likewise, the women who choose not to work with a high social status (H) also enjoy more post-materialistic attributes than materialistic ones. On the other hand, the women of the two remaining clusters (F and G), located between both high and low social positions, may be seeking more materialistic assets than post-materialistic ones. It has often been argued that economic development brings about a change in value priorities, that is, a move away from materialism towards post-materialism (Inglehart (1977)). The results of this paper can be understood in the light of that argument. The different patterns of commitment to materialistic and post-materialistic enrichment express people's personal positions. The upper to upper-middle class women give priority to post-materialism, or opportunities for self-improvement and enjoyment, and results in a tendency to move away from acquiring material possessions. In contrast, the lower-middle class makes materialistic affluence a priority. It is suggested that those who aspire to post-materialism are committed to participating in cultural activities, irrespective of whether they engage in paid work or not. Besides, the evidence could be found that the women of poor background (B, C, E and F) cannot be generous consumers. Even continuous working does not compensate them (B) for their disadvantage, because they might investigate primarily in their children, not consume household properties, within budget constraint.

4 Discussion

Based on the respondents' occupational history, family data and levels of commitment to materialistic and post-materialistic values, our findings identify women with similar career-paths but different life-priorities. The career-lifestyle typology, rather than the type of career-path, is significant because of its implications. The reason why female labour force participation is relatively low in Japan, though the educational level of Japanese women is one of the highest in the world, seems to be bipolarization of the higher educated. Above all, closer relation between the women who choose career and childlessness and tertiary education suggests that women's rising educational levels over recent decades may bring about an increasing number of women who

choose this kind of career-lifestyle in the future unless the work situations improve.

It is the career-lifestyle typology that influences women's life events such as, for example, the timing of marriage, childbirth, and leaving (as well as possibly re-entering) the workforce (Mayer and Tuma (1990), Blossfeld (1995)). In all probability, this typology will also have an impact on future generations of women (Elder (1974), Bourdieu (1979)). The type of occupation and the sequence thereof, as well as the timing and duration of any absence from the workforce are also important issues. Therefore, the outcome of classification using cluster analysis can be used in further research. We might be able to compare and assess the change in the career-lifestyle patterns of subsequent generations by longitudinal analysis utilizing the data of a follow-up study.

References

- ARISAWA, H. (1956): Structure of Wages and Structure of Economy. In: I. Nakayama (Ed.): *Basic survey of Wages*. Toyokeizaishinposha. (in Japanese)
- BLOSSFELD, H.P. (1995): *The New Role of Women: Family Formation in Modern Societies*. Westview Press, Colorado.
- BOURDIEU, P. (1979): *La Distinction: Critique Sociale du Judgement*. Minuit, Paris.
- BRINTON, M.C. (1993): *Women and the Economic Miracle: Gender and Work in Postwar Japan*. Univ. of California Press, Berkeley.
- CROMPTON, R. (1999): *Restructuring Gender Relations and Employment*. Oxford Univ. Press, New York.
- DOUSLAS, P.H. (1934): *The Theory of Wages*. Macmillan, New York.
- ELDER, G.H. Jr. (1974): *Children of the Great Depression: Socials Change in Life Experience*. Univ. of Chicago Press, Chicago.
- ESPING-ANDERSEN, G. (1999): *Social Foundations of Postindustrial Economies*. Oxford University Press, Oxford.
- HAKIM, C. (1997): *Key Issues in Women's Work*. Athlone, London.
- INGLEHART, R. (1977): *The Silent Revolution: Changing Values and Political Styles among Western Politics*. Princeton University press, Princeton.
- MAYER, K.U. and TUMA, N.B. (1990): *Event History Analysis in Life Course Research*. Univ. of Wisconsin Press, Madison.
- MIFUNE, M. and SHIGEKAWA, J. (1999): Wife's Career Patterns and Family Budget. In: Y. Higuchi and M. Iwata (Eds.): *Contemporary Japanese Women*. Toyokeizaishinposha, Tokyo, 127-145. (in Japanese)
- MOEN, P. (1996): Gender, Age, and the Life Course. In: R. Binctock and L. George (Eds.): *Handbook of Aging and the Social Sciences*. Academic Press, New York, 171-187.
- NAKAI, M. and AKACHI, M. (2000): Labour Market and Social Participation. In: K. Seiyama (Ed.): *Gender, Market, and Family*. University of Tokyo Press, Tokyo, 111-131. (in Japanese)
- OPPENHEIMER, V.K. (1982): *Work and the Family: A Study in Social Demography*. Academic Press, New York.
- YAMADA, M. (1999): *Restructuring of Family*. Shinyosha, Tokyo. (in Japanese)

Joint Space Model for Multidimensional Scaling of Two-Mode Three-Way Asymmetric Proximities

Akinori Okada¹ and Tadashi Imaizumi²

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. A joint space model and an associated nonmetric algorithm to analyze two-mode three-way asymmetric proximities (object \times object \times source) are presented. Each object is represented as a point and a circle (sphere, hyper sphere) in the common joint configuration which is common to all sources. Each source is represented as a point in the common joint configuration. For each source, the radius of an object is stretched or shrunk according to the distance between the dominance point representing the source and the point representing the object. An application to intergenerational occupational mobility data is shown.

1 Introduction

Several procedures of multidimensional scaling (MDS) for analyzing two-mode three-way asymmetric proximities, or two-mode three-way asymmetric MDS, have been introduced (DeSarbo et al.(1992), Okada and Imaizumi (1997), Zielman (1991), Zielman and Heiser (1993)). Two-mode three-way asymmetric proximities usually consist of a set of proximity matrices. Each matrix consists of proximities among a set of objects from a source, and is not necessarily symmetric.

These two-mode three-way asymmetric MDS represent differences among sources or individual differences when a source corresponds to an individual. The differences among sources come from symmetric relationships among objects and from asymmetric relationships as well. While each of the two-mode three-way asymmetric MDS uses its own way of representing the differences among sources, all of them utilize weights showing the salience of symmetric or asymmetric relationships among objects for each source to represent the differences among sources. These weights can represent differences among sources. But they are not related to the characteristics of objects, or they are not related to the locations of objects represented in the configuration of objects when the model is based on the spatial representation. The purpose of the present study is to develop a joint space model and an associated algorithm of two-mode three-way asymmetric MDS, where objects and sources

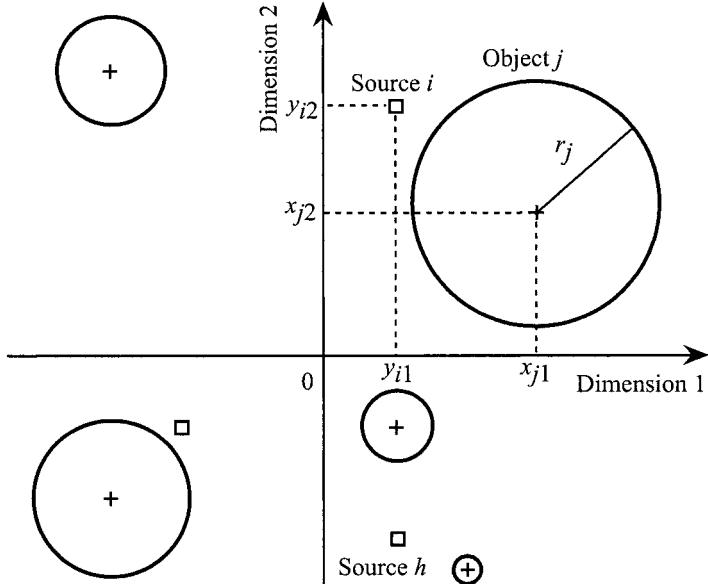


Fig. 1. The common joint configuration.

are represented in the same multidimensional space so that differences among sources are related directly to the locations of objects.

2 The model

The model consists of the common joint configuration, symmetry weights, and asymmetry weights. In the common joint configuration, an object is represented as a point and a circle (sphere or hypersphere) centered at that point, which was inherited from Okada and Imaizumi (1987), and a source is presented as a point called the dominance point. The common joint configuration represents the relationships among objects, among sources, and between objects and sources which are common to all sources. Figure 1 shows a two-dimensional common joint configuration. Object j is represented as a point (x_{j1}, x_{j2}) and a circle of radius $r_j (\geq 0)$, where x_{js} is the coordinate of object j along dimension s of the common joint configuration. Source i is represented as a point (y_{i1}, y_{i2}) , where y_{is} is the coordinate of source i along dimension s of the common joint configuration.

Each source has its own configuration of objects where each object is represented as a point and a circle. The configuration of points representing objects is derived from the configuration of points in the common joint configuration by uniformly stretching or shrinking dimensions of the common joint configuration. In the configuration of objects for source i , object j is represented as a point $(w_i x_{j1}, w_i x_{j2}, \dots, w_i x_{jp})$ and a circle, where $w_i (\geq 0)$

is the symmetry weight for source i which shows the salience of symmetric proximity relationships among objects for source i , and p is the dimensionality of the common joint configuration. The radius of a circle of an object for a source is determined by adjusting the radius r_j in the common joint configuration using (a) the distance between the dominance point representing the source and the point representing the object, and (b) the asymmetry weight for the source. The radius of object j for source i is given by

$$(1 - \beta_i \exp(-d_{ij}^2 c))r_j,$$

where $\beta_i (\geq 0)$ is the asymmetry weight for source i which shows the salience of asymmetric proximity relationships among objects caused by the dominance point for source i , d_{ij} is the Euclidean distance between the dominance point representing source i and the point representing object j in the common joint configuration;

$$d_{ij} = \sqrt{\sum_{s=1}^p (y_{is} - x_{js})^2}, \quad (1)$$

and c is the constant to adjust the effect of the distance between the dominance point and the point representing the object. The radius of an object for a source becomes smaller as the distance between the dominance point and the point representing the object increases.

Let s_{jki} be the observed proximity from objects j to k for source i . It is assumed that s_{jki} is monotonically decreasingly (when s_{jki} depicts similarity) or increasingly (when s_{jki} depicts dissimilarity) related to m_{jki} defined as

$$m_{jki} = w_i d_{jk} - (1 - \beta_i \exp(-d_{ij}^2 c))r_j + (1 - \beta_i \exp(-d_{ik}^2 c))r_k, \quad (2)$$

where $w_i d_{jk}$ is the distance between objects j and k in the configuration of objects for source i . The definition of m_{jki} is exactly the same as that of the predecessor (Okada and Imaizumi (1997)) except for the adjustment of the radius. Figure 2 shows m_{jki} and m_{kji} in the two-dimensional configuration of objects for source i . In Figure 2 objects j and k are shown. Each object is represented as a point and a circle. The left panel shows m_{jki} , and the right panel shows m_{kji} . m_{jki} is larger than m_{kji} , because the radius representing object j is smaller than the radius representing object k for source i .

3 The algorithm

An associated algorithm to derive the common joint configuration, w_i , β_i , and c from observed two-mode three-way asymmetric proximities was developed. Let n be the number of objects, and N be the number of sources. A nonmetric iterative algorithm to derive the common joint configuration ($x_{js}; j = 1, \dots, n; s = 1, \dots, p : y_{is}; i = 1, \dots, N; s = 1, \dots, p : r_j; j = 1, \dots, n$), the

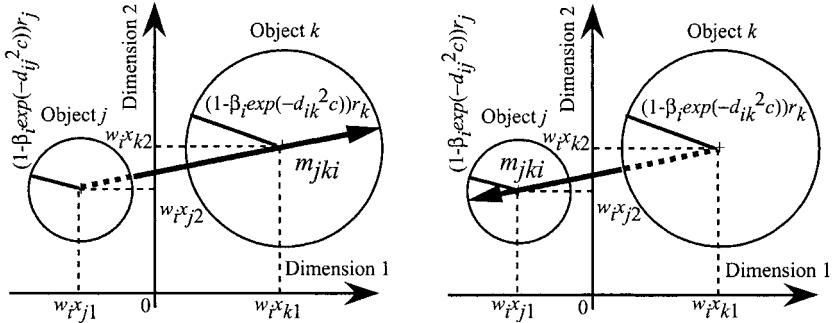


Fig. 2. m_{jki} (left panel) and m_{kji} (right panel) in the configuration of objects for source i .

symmetry weight ($w_i; i = 1, \dots, N$), the asymmetry weight ($\beta_i; i = 1, \dots, N$), and c from observed proximities s_{jki} ($j, k[j \neq k] = 1, \dots, n; i = 1, \dots, N$) was extended from the one for the predecessor which had been developed based on Kruskal's nonmetric algorithm (Kruskal (1964)). The badness-of-fit measure called the stress

$$S = \sqrt{\frac{1}{N} \sum_{i=1}^N \left[\sum_{\substack{j=1 \\ j \neq k}}^n \sum_{k=1}^n (m_{jki} - \hat{m}_{jki})^2 / \sum_{\substack{j=1 \\ j \neq k}}^n \sum_{k=1}^n (m_{jki} - \bar{m}_i)^2 \right]}, \quad (3)$$

is defined, where \hat{m}_{jki} is the monotone transformed s_{jki} , and \bar{m}_i is the mean of m_{jki} for source i . The common joint configuration, the symmetry weight, the asymmetry weight, and c which minimize the stress are sought for a given dimensionality.

The joint configuration was normalized so that the origin is at the centroid of the object and dominance points, and that the sum of squared coordinates of objects and dominance points along p dimensions is equal to $n + N$. For source i m_{jki} is normalized so that the sum of squared differences from the mean \bar{m}_i (the denominator of Equation (3)) is equal to n .

4 An application

The present asymmetric MDS was applied to analyze intergenerational occupational mobility data. The data consist of four tables of occupational mobility among eight occupational categories from fathers to sons, where each table corresponds to the data collected in 1955, 1965, 1975, and 1985. The eight occupational categories are; (a) Professional, (b) Non-manual employed by large enterprises, (c) Non-manual employed by small enterprises, (d) Non-manual self-employed, (e) Manual employed by large enterprises,

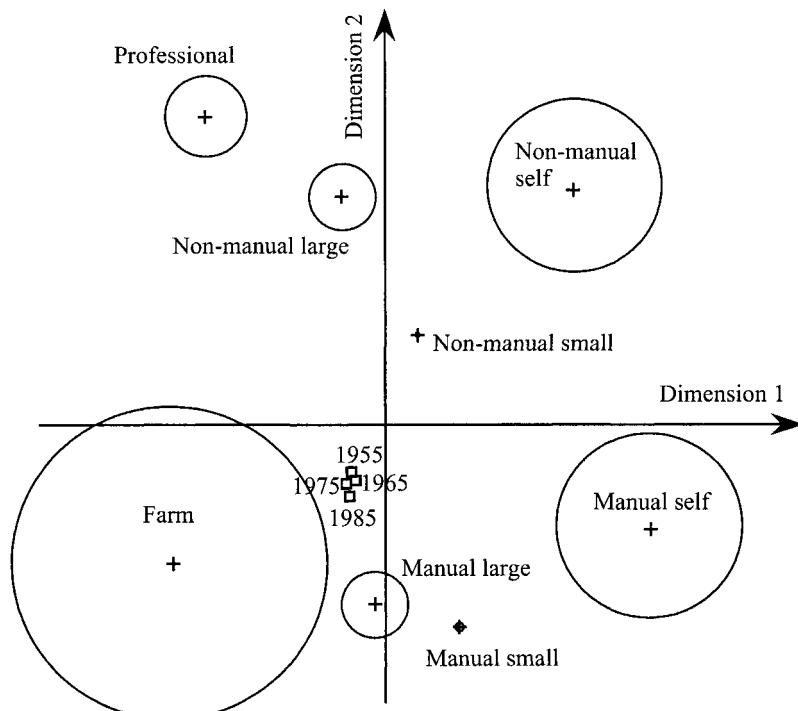


Fig. 3. The two-dimensional common joint configuration of eight occupational categories and four years.

(f) **Manual** employed by **small** enterprises, (g) **Manual self-employed**, and (h) **Farm**. Emboldened word(s) will be used to represent each occupational category hereafter. The data were rescaled to remove the differences in the share of manpower among the eight occupational categories for each year, and the rescaled data are shown in Okada and Imaizumi (1997).

The analysis was done in seven- through unidimensional spaces. The analysis suggests to adopt the two-dimensional result as the solution, which has the stress of 0.328, as the solution. To validate the two-dimensional result, the analysis using 50 different random initial common joint configurations with initial $y_{is} = 0$, $r_j = 0$, $\beta_i = 1$, and $c = 1$ were done in the two-dimensional space, yielding the smallest stress of 0.331. Thus the two-dimensional result of stress 0.328 was adopted as the solution. Figure 3 shows the two-dimensional common joint configuration of the solution after the orthogonal rotation of the originally obtained configuration.

In Figure 3, each occupational category is represented as a point and a circle, and each year is represented as a point. The configuration shown in Figure 3 was derived by orthogonally rotating the originally obtained configuration so that two dimensions can easily be interpreted. The vertical dimension

seems to represent the difference between non-manual and manual occupational categories, and the horizontal dimension seems to represent differences among (a) self-employed, (b) employees of small enterprises, (c) employees of large enterprises, and (d) professional or farm occupational categories (Okada and Imaizumi (1997, pp. 218-219)). The configuration of eight occupational category is very close to the common object configuration in earlier studies analyzing the same data (Okada and Imaizumi (1997), (2002)). Correlation coefficients between the coordinates of occupational categories of the present study and those of the earlier studies are 0.921 and 0.985 for dimensions 1 and 2 in the case of the former study, and 0.932 and 0.988 for dimensions 1 and 2 in the case of the latter study. In this application, the larger radius means that the sons of fathers in the occupational category have the larger tendency of moving out from the corresponding occupational category. The radius of the farm occupational category is the largest, and the radius of the non-manual small occupational category is the smallest, that coincides with the earlier study. Correlation coefficients between the present radii and those of the earlier studies are 0.993 and 0.994 for the former and the latter.

Four years are located near the origin, and four dominance points moved downward along dimension 2 or moved toward the points representing the manual occupational categories and away from the points representing the non-manual occupational categories. Thus the radius of the manual occupational categories becomes smaller, and the radius of the non-manual occupational categories becomes larger, as the year passes from 1955 through 1985, suggesting the asymmetric occupational mobility from manual to non-manual occupational categories decreased. This is consistent with Seiyama et al. (1990, p. 30).

Year	Symmetry weight	Asymmetry weight
1955	0.401	0.349
1965	0.300	0.052
1975	0.244	0.155
1985	0.235	0.100

Table 1. Symmetry weight and asymmetry weight.

Table 1 shows the symmetry weight w_i and the asymmetry weight β_i . And c is 0.940. The symmetry weight decreased monotonically from 1955 to 1985, showing the symmetric occupational mobility among occupational categories decreased. This is consistent with the earlier study.

In the present model, the asymmetric relationships among objects for a source are represented by (a) the radius of the source, (b) the location of the dominance point representing the source, and (c) the asymmetry weight for the source. The first is common to all sources, while the second and

the third are specific to the source. The radii shown in Figure 3 show that eight occupational categories have substantive differences in the asymmetry on the occupational mobility among them. Four dominance points in Figure 3 are located closely, suggesting the distance between the dominance point representing the year and the point representing the occupational category are similar for four years. This shows that the distance causes the similar effect on the radius of the object for four years, which makes it not too difficult to interpret the resulting radius for each year. In the case where dominance points spread over a configuration, the interpretation may be more difficult than the interpretation of the present result. The asymmetry weight in Table 1 is larger for 1955 than for 1965, 1975, and 1985. Thus the salience of asymmetric relationships among eight occupational categories caused by the dominance point or d_{ij} is larger for 1955 than for the other three years.

5 Discussion

The present model was extended from the predecessor, and has inherited several aspects from it. Each object is represented as a point and a circle, and the radius of the circle has the same meaning (the larger radius suggests the larger outward tendency from the corresponding object, and the smaller inward tendency into the corresponding object). The symmetry weight has the same meaning as well. But the present model differs from the predecessor in some respects, which comes from employing the joint space model, representing both objects and sources in a same multidimensional space.

The orientation of dimensions of the common joint configuration of the present model is determined uniquely only up to the orthogonal rotation, because the orthogonal rotation does not affect the badness-of-fit measure. The obtained common joint configuration can be orthogonally rotated if necessary. On the other hand, the introduction of the asymmetry weight of the predecessor is benefited in getting uniquely oriented dimensions of the common joint configuration up to the reflections and permutations.

One characteristic which the present model inherited from the predecessor is that the sum of squared deviations of m_{jki} is additively decomposed into two terms; the one term represents the magnitude of symmetric component of the configuration of objects for source i , and the other term represents the magnitude of asymmetric component of the configuration of objects for source i (Okada and Imaizumi (1997, p. 209)).

The common joint configuration shows the distances from each source to objects. The differences among sources are related to the configuration of objects. The relationships among sources represented in the common joint configuration are more easily be interpreted than in the common object configuration of the predecessor. But the radius for each source of the present model seems more difficult to interpret especially when dominance points spread over a common joint configuration than those of the predecessor. A

dominance point representing a source shows a location or a hypothetical object whose radius is minimized for the corresponding source, suggesting the object has the largest tendency of inward from the other objects and the smallest tendency of outward to the other objects for the source (cf. Krumhansl (1978)). The radius can be negative when the asymmetry weight β_i is larger than 1 (cf. Okada and Imaizumi (2003)). In the case of the brand switching, a dominance point represents a hypothetical brand which has the relatively strongest power in the brand switching relationships among brands for the corresponding source.

References

- DESARBO, W.S., JOHNSON, M.D., MANRAI, A.K., MANRAI, L.A., and EDWARD, E.A. (1992): TSCALE: A New Multidimensional Scaling Procedure Based on Tversky's Contrast Model. *Psychometrika*, 57, 43–69.
- KRUMHANSL, C.L. (1978): Concerning the Applicability of Geometric Model to Similarity Data: The Interrelationship between Similarity and Spatial Density. *Psychological Review*, 85, 445–463.
- KRUSKAL, J.B. (1964): Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika*, 29, 115–129.
- OKADA, A. and IMAIZUMI, T. (1987): Nonmetric Multidimensional Scaling of Asymmetric Proximities. *Behaviormetrika*, No. 21, 81–96.
- 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. (2002): A Generalization of Two-Mode Three-Way Asymmetric Multidimensional Scaling. In: W. Gaul and G. Ritter (Eds.): *Classification, Automation, and New Media*. Springer, Berlin, 113–122.
- OKADA, A. and IMAIZUMI, T. (2003): Two-Mode Three-Way Nonmetric Multidimensional Scaling with Different Directions of Asymmetry for Different Sources. In: H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J.J. Meulman (Eds.): *New Developments in Psychometrics*. Springer, Tokyo, 495–502.
- SEIYAMA, K., NAOI, A., SATO, Y., TSUZUKI, K., and KOJIMA, H. (1990): Stratification Structure of Contemporary Japan and its Trend, In A. Naoi and K. Seiyama (Eds.): *Social Stratification in Contemporary Japan Vol. 1. Structure and Process of Social Stratification*. Tokyo University Press, Tokyo, 15–50. (in Japanese)
- ZIELMAN, B. (1991): Three-Way Scaling of Asymmetric Proximities. Research Report RR91-01, Department of Data Theory, University of Leiden.
- ZIELMAN, B. and HEISER, W.J. (1993): Analysis of Asymmetry by a Slide-Vector. *Psychometrika*, 58, 101–114.

Structural Model of Product Meaning Using Means-End Approach

Adam Sagan

Chair of Market Analysis and Marketing Research,
Faculty of Management, Cracow University of Economics,
ul. Rakowicka 27, 31-510 Kraków, Poland

Abstract. The aim of the paper is to model motivational and cognitive structures of product meaning based on means-end chain framework. Applying SEM to means-end provides new analysis to help validate "hard laddering" measurement scales, model relationships among multiple latent predictors (bundles of product attributes) and criterions (consequences and values), as well as error of measurement and test a priori substantive assumptions against the data. This methodology introduces Guttman scaling to means-end framework and a confirmatory instead of classical exploratory approach to MEC analysis.

1 Review of laddering research

The basic assumption of means-end theory is that consumers buy product attributes which in their consequences of usage help them to solve important problems or satisfy important ends. These consequences vary in the levels of abstraction and form hierarchical linkages or networks called means-end chains. Means-end chains form the hierarchical structures of consumer's knowledge, motivation and decision-making, which are based on product domain related ladders that involve: concrete and abstract attributes, functional and psychological consequences as well as instrumental and terminal values (Reynolds and Gutman (1988)). In the process of data gathering, a special technique of interview called laddering has been used for this purpose, because it forces consumers to "move up the ladder of abstraction" along the attributes-consequences-values continuum. The application of means-end approach is very broad across different product ranges and topics of marketing (Reynolds and Gutman (1988), Pieters et al. (1995), Lin (2002)).

The research process can be divided into three steps: 1. identification of salient attributes, 2. elicitation of individual ladders and 3. analysis and mapping of means-end structures.

1.1 Identification of salient attributes

The first step in laddering interview is to determine product salient attributes. Several methods of attribute elicitation can be used during this stage: freelist elicitation, free, sequential and multiple pilesort, triadic sorting (rep grid),

paired comparisons and attribute selection task. These methods of attributes' identification are rooted in emic perspective of cultural domain analysis, where a particular product domain is revealed from the respondents' point of view (Spredaly (1988)). These techniques enable respondents to define which attributes they use to distinguish among products. When the most important salient attributes have been established, laddering interview can be adopted to answer the most crucial structural question: why particular attributes are important for the consumer in a given situation.

1.2 Elicitation of individual ladders

Many techniques of data gathering are used to obtain consumers' cognitive and motivational structures in various areas of marketing. So called "soft laddering" that is a kind of qualitative laddering interview, is predominantly applied in order to answer the question above. Despite numerous advantages of this qualitative approach, it seems to be time and cost consuming, has limited external validity and arises many problems with coding and individual ladders identification. To overcome this problems more structured techniques of "hard" laddering have been developed. Both techniques have similar convergent and predictive validity (Botschen and Thelen (1998)), nevertheless hard laddering seems to be easier to administer and less costly. Hard laddering is also more suitable for larger samples because of more generalizability and higher external validity. Among various approaches to laddering data gathering the following are most representative for this approach:

A. Semistructured interview - respondents are presented with a series of boxes connected by arrows that facilitate data collection. This paper and pencil method allows to identify up to four attributes and explore up to three consequences for each attribute (Botschen and Hemetsberger (1998)).

B. Card sorting approach - respondents are shown three kinds of cards: attributes, consequences and values cards. From the pile of attribute cards they have to select the most important attribute for the product in question. Then from the pile of consequence cards - the most important consequence, and from the pile of values - the most important value following from the consequence. The procedure is repeated with the second-most and third-most important attribute. Additional attributes, consequences and values can be added on separate blank cards (Roehrich and Vallette-Florence (1991)).

C. Verbal rating scales - respondents' rate statements containing whole predefined chain consisting of attribute - consequence and value link. Particular component links can be rated on 5-7 point scale in a similar manner (Vanden Abeele (1990)).

D. Product-profile rating scales - respondents are supposed to rank product profiles (based on an orthogonal plan as in conjoint analysis) with regard to specific consequences and values. The covariance matrix of attributes, consequences and values is analyzed and HVM is estimated by structural equation modelling (Grunert (1997)).

E. Association Pattern Technique - respondents are presented empty attributes-consequences and consequences-values as rows and columns of separate tables and they have to mark those cells that the given relationships exist (they mark an intersection where particular attribute leads to particular consequence in A-C matrix and similarly where a given consequence leads to a given value in C-V matrix) (ter Hofstede et al. (1998)).

These techniques of data gathering are more suitable for quantitative research, they are less time and effort consuming however they are exploratory and have lower internal validity in comparison to traditional soft laddering qualitative interviews.

1.3 Analysis of means-end structures

Hard laddering data enables to utilize broad range of data analysis and classification methods. Traditional as well as relatively new methods of laddering data analysis have been developed in this area.

A. Hierarchical value map (HVM) or consumer decision map (CDM) - is one of the most often used methods of data analysis and data visualization. It is a tree-type graph containing information concerning the frequency of selection and the frequency of hierarchical links among particular ACV items.

B. Correspondence analysis and multidimensional scaling - is used to obtain spatial representation of the relationships between attributes consequences and values elicited by the respondents in reduced Euclidean space (Valette-Florence and Rapacchi (1991)).

C. Social network analysis - gives opportunity to analyze ACV chains as social networks that are characterized by several node-, path- and net-based properties. These are described by various indices like centrality, abstractness reachability and transitivity (Pieters et al. (1995), Sagan (2001)).

D. Factor analysis and multiple regression - are used to reduce attribute variables and classify them in a smaller number of independent common factors. Then multiple regression is applied to analyze consequence-attribute and value-consequence linear relations (Lin (2002)).

E. Loglinear analysis - is adopted to describe the probabilities of ACV linkages in the ladders and test conditional independence of AC and CV matrices in APT data (ter Hofstede et al. (1998)).

F. Cognitive differentiation analysis - is used to describe the relationship of the pairwise judgements between brands. Respondents rate pairs of brands, preference intensity and attribute, consequence and value specific to each brand (Reynolds and Perkins (1987)).

G. Constrained cluster analysis - solves problems of identification of dominant means-end chains on the basis of frequency of individual ladders that represents the chain, and representativeness of ladders that is accuracy with which means-end chain represents a set of ladders (Aurifeille and Valette-Florence (1995)).

2 Development of means-end scale of product meaning

2.1 Dimensions of product meaning

The area of many studies in marketing (Solomon (1983)), psychology (Friedmann and Zimmer (1988)), sociology and cultural anthropology (McCracken (1990)) is the symbolic significance of a product. These studies show that products are important for the consumers not only because of what they do, but also because of what they mean. The meaning of a product is determined in context of social group, where the product serves as social stimulus that indicates consumers' role performance and role knowledge (Solomon (1983), (Leigh and Gabel (1992)).

The conceptualization of product meaning is based on J. N. Sheth, B. I. Newman, and B. L. Gross theory of consumption values (Sheth et al. (1991)) and W. Nöth's concept of product semiotizations processes (Nöth 1988). Sheth, Newman and Gross theory identifies functional, social, emotional, epistemic, and conditional values that influence consumer behavior.

According to Nöth products can be viewed as indexical signs. He distinguishes three types of product meaning that the process of product semiotization can lead to: utilitarian, commercial and socio-cultural.

Means-end product meaning scale was developed on the basis of the dimensions of product meaning. We assumed that the hierarchy of product meaning is reflected by reflexive and cumulative indicators (items) that form a Guttman scale.

2.2 Conceptual model of a product meaning means-end scale

In exploratory studies several Guttman scales have been developed that cover functional and psychological consequences and terminal value with respect of utilitarian, hedonic and social meaning of products. On the basis of the freelist data obtained in a qualitative stage of the analysis, three attributes, dimensions of product meaning and consequences in means-end chain were selected. Attributes and consequences were selected on the basis of scree test of freelist data, where top of mind core attributes were chosen.

Three attributes of mobile phones were selected: price, brand name and innovation of mobile phone (MMS). Then a Guttman cumulative scale for measurement of three dimension of meaning was developed. The first dimension was called "utilitarian", the second "hedonic" and the third - "social recognition". Each dimension was measured by an unidimensional, hierarchical and cumulative Guttman means-end scale. The conceptual model of the measurement is shown in Figure 1.

2.3 Measurement model

An example of a nine-item scale that reflects utilitarian, hedonic and social meaning of price is given below:

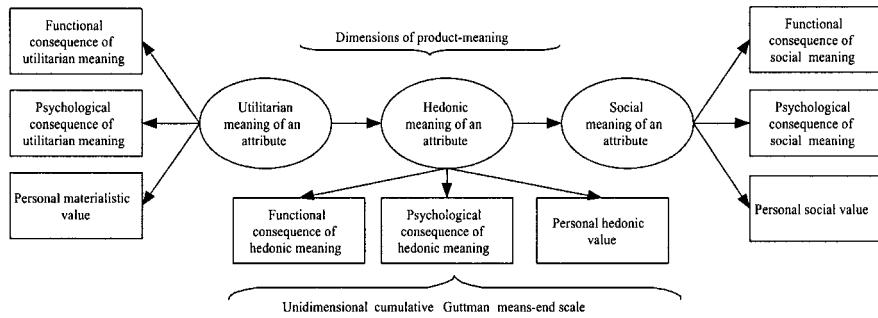


Fig. 1. Conceptual model of means-end scale

1. Higher price of mobile phone N is important for me because it gives me higher durability of mobile phone if Yes go to 2 if No go to 4
2. Higher durability of mobile phone N is important for me because I feel that I save time and money in the long run if Yes go to 3 if No go to 4
3. Saving time and money is important for me because providence is the main quality of my life if Yes go to 4 if No go to 4
4. Higher price of mobile phone N is important for me because I feel that I can afford this phone if Yes go to 5 if No go to 7
5. Feeling that I can buy what I want gives me a lot of pleasure in spending money if Yes go to 6 if No go to 7
6. Pleasure in spending money is important for me because I like an exciting and joyful life if Yes go to 7 if No go to 7
7. Higher price of mobile phone N is important for me because it is not so popular and the others probably don't have this phone if Yes go to 8 if No go to 10
8. When I have something that others don't, I feel unique and I am myself if Yes go to 9 if No go to 10
9. Being different from others is important because I want to be recognized in my society Yes No

This scale is a kind of "forced" Guttman scale, because respondents are asked to follow instructions below the statements. This constrain, however, tends to give a "perfect" cumulative Guttman scale as a result.

Item response theory (IRT) is very often used in Guttman cumulative scale development, validating and scoring. All IRT measurement models, like parametric Rasch and Birnbaum models or nonparametric Mokken model assume unidimensionality of latent construct.

Structural equation modelling (SEM) belongs to CTT tradition of scale development and validation. However the SEM approach enables to identify multidimensional causal and correlational relationships between latent constructs of product meaning. Because the response scale consists of yes-no

answers on particular statements, tetrachoric correlations between items were calculated under the assumption that these correlations between dichotomous manifest indicators, underlie latent continuous dimension of product meaning.

The analysis of dimensionality and reliability of Guttman scale is based on testing the hypothesis of a simplex structure in the measurement model. In contrary to Likert-type parallel items, where a common factor model is used in unidimensionality and reliability assessment, Guttman cumulative scale analysis is based on a simplex model. In that case the scatterplot of factor loadings on 2-dimensional factors is U-shaped what is called "horseshoe" or "arch" effect. When this effect is present it suggests an unidimensional Guttman scale. A second factor in this situation is a simple mathematical artefact. This hypothesis was tested with the help of confirmatory factor analysis, where sub-diagonal coefficients were set as equal. An example of unidimensionality analysis, where below-diagonal tetrachoric correlations are set as equal is given below:

$$\begin{pmatrix} 1 & & & & & & \\ .81 & 1 & & & & & \\ .70 & .81 & 1 & & & & \\ .54 & .70 & .81 & 1 & & & \\ .34 & .54 & .70 & .81 & 1 & & \\ .11 & .34 & .54 & .70 & .81 & 1 & \\ -.11^* & .11 & .34 & .54 & .70 & .81 & 1 \\ -.32 & -.11^* & .11 & .34 & .54 & .70 & .81 & 1 \\ -.52 & -.32 & -.11^* & .11 & .34 & .54 & .70 & .81 & 1 \end{pmatrix}$$

* $p > 0.05$, Chi-Square = 36.27, d.f. = 36, $p = 0.45$
 $CFI = 0.99$, $TLI = 0.99$, $RMSEA = 0.02$

The CFA shows that the hypothesis of unidimensionality of the extended utilitarian construct should be accepted (items form the simplex structure, where correlations get smaller and are equal below diagonal).

3 Structural model of means-end

After the unidimensionality of measurement models based on Guttman scales has been checked, a structural model was estimated. A two-step approach of model estimation suggested by Anderson and Gerbing (1988) was implemented, because the measurement part of the model is based on a different theoretical assumption (means-end theory). Next the structural part of the model (product meaning categorizations) was estimated. To estimate the model a tetrachoric correlation matrix was calculated and WLMSV estimator (weighted least square using a diagonal weight matrix with robust standard

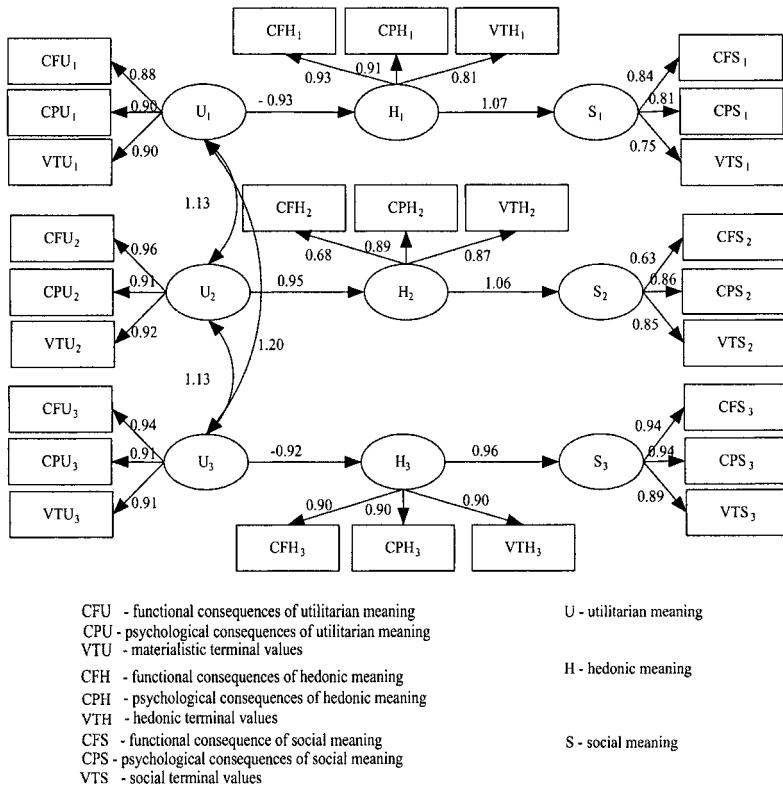


Fig. 2. Structural model of product meaning

errors and mean- and variance adjusted chi-square test statistic) were used. Muthén's Mplus 3.12 was selected for model computation.

Parameters of the model of product meaning show (Figure 2) that utilitarian dimensions have a negative impact on the hedonic dimension, except the brand name where a positive relation between utilitarian, hedonic and social dimensions is observed.

All model parameters were significant ($p=.05$). Covariances between utilitarian consequences of price, brand and innovation were also significant. Fit indices showed an unacceptable level of goodness of fit (significant Chi-Square and RMSEA greater than .06) although a comparative fit index (CFI greater than .95) suggests that the theoretical model explains the data pretty well when compared with the basement model.

4 Conclusions

The result of the research shows the applicability of Guttman scaling of utilitarian, hedonic (fun and enjoyment) and social (impress for others and social

recognition) dimensions of product meaning. Each dimension is measured by reflexive cumulative and hierarchically ordered items along the attributes - consequences - values continuum. Confirmatory factor analysis of simplex structure tests the unidimensionality of Guttman scale. The structural equation model is based on tetrachoric correlations that underline latent continuous product meaning dimensions. This approach introduces a confirmatory and quantitative view on product meaning and means-end analysis. Of course, further research and replications improving measurement models and comparative multigroup analysis is needed in order to scale validation and model refinement.

References

- ANDERSON, J.C. and GERBING, D.W. (1988): Structural Equation Modeling in Practice; a Review and Recommended Two-Step Approach. *Psychological Bulletin, 103*, 411-423.
- AURIFEILLE, J-M. and VALETTE-FLORENCE, P. (1995): Determination of the Dominant Means-End Chains: A Constrained Clustering Approach. *International Journal of Research in Marketing, 12*, 267-278.
- BOTSCHEN, G. and HEMETSBERGER, A. (1998): Diagnosing Means - End Structures to Determine the Degree of Potential Marketing Program Standardization. *Journal of Marketing Research, 42*, 151-159.
- BOTSCHEN, G. and THELEN, E. (1998): Hard versus Soft Laddering: Implication for the Appropriate Use. In: I. Balderjahn, C. Mennicken, and E. Vernette (Eds.): *New Developments and Approaches in Consumer Behavior Research*. Schäffer - Poeschel, Stuttgart, 321-339.
- FRIEDMAN, R. and ZIMMER, M. (1988): The Role of Psychological Meaning in Advertising. *Journal of Advertising, 1*, 31-40.
- GRUNERT, K.G. (1997): What's in a Steak? A Cross - Cultural Study on the Quality Perception of Beef. *Food Quality and Preference, 8*, 157-174.
- LEIGH, J.H. and GABEL, T.G. (1992): Symbolic Interactionism: Its Effects on Consumer Behavior and Implications for Marketing Strategy. *Journal of Consumer Marketing, 9*, 27-38.
- LIN, F. (2002): Attribute - Consequence - Value Linkages: A New Technique for Understanding Consumer's Product Knowledge. *Journal of Targeting, Measurement and Analysis for Marketing, 10*, 339-352.
- MCCRACKEN, G. (1990): *Culture and Consumption: A Theoretical Account of the Structure and Movement of The Cultural Meaning of Consumer Goods*. Indiana University Press, New York.
- NÖTH, W. (1988): The Language of Commodities. Grundwork for a Semiotics of Consumer Goods. *International Journal of Research in Marketing, 4*, 173-186.
- PIETERS, R., BAUMGARTNER, H., and ALLEN, D. (1995): A Means - End Chain Approach to Consumer Goal Structures. *International Journal of Research in Marketing, 12*, 227-244.
- REYNOLDS, T.J. and GUTMAN, J. (1988): Laddering Theory. Method, Analysis and Interpretation. *Journal of Advertising Research, 2*, 11-31.

- REYNOLDS, T.J. and PERKINS, W.S. (1987): Cognitive Differentiation Analysis: New Methodology for Assessing the Validity of Means-End Hierarchies. *Advances in Consumer Research*, 4, 109–113.
- ROEHRICH, G. and VALLETTE-FLORENCE, P. (1991): A Weighted Cluster-Based Analysis of Direct and Indirect Connections in Means-End Chains: An Application to Lingerie Retail. In: K.G. Grunert and P. Vallette-Florence (Eds.): *Workshop on Values and Lifestyle Research in Marketing*. EIASM, Brussels.
- SAGAN, A. (2001): Metody sieciowe w analizie środków-celów z wykorzystaniem programu UCINET. [Social Network Analysis of Means-End Chains - Application of UCINET Statistical Package] *Zeszyty Naukowe*. AE, Krakw 558.
- SHETH, J., NEWMAN, B., and GROSS, B. (1991): Why We Buy What We Buy: A Theory of Consumption Values. *Journal of Business Research*, 22, 159–170.
- SOLOMON, M.R. (1983): The Role of Products as Social Stimuli: A Symbolic Interactionism Perspective. *Journal of Consumer Research*, 10, 319–329.
- SPREADLY, J. (1988): *The Ethnographic Interview*. Harcourt Brace Jovanovich, 1988.
- TER HOFSTEDE, F., ANDENAERT, A., STEENKAMP, J.B., and WEDEL, M. (1998): An Investigation into the Association Pattern Technique as a Quantitative Approach to Measuring Means-End Chains. *International Journal of Research in Marketing*, 15, 37–50.
- VALETTE-FLORENCE, P. and RAPACCHI, B. (1991): Improvements in Means-End Chain Analysis. Using Graph Theory and Correspondence Analysis. *Journal of Advertising Research*, 2, 30–45.
- VANDEN ABELE, P. (1990): A Means-End Study of Diary Consumption Motivation. No. EC Regulation 1000/90-43ST. EC.

The Concept of Chains as a Tool for MSA Contributing to the International Market Segmentation

Elżbieta Sobczak

Wroclaw University of Economics,
ul. Komandorska 118/120, 53-345 Wrocław, Poland

Abstract. New tendencies in international market segmentation methodology lead towards joining similar market segments, functioning in different countries, into one international segment, defined as an inter-market one. The paper presents suggestions of applying multivariate statistical analysis methods for identifying international segments. The basis of this segmentation is made up of the concept of means-end chains, which assumes that products' attributes become for the consumer the means, which enable obtaining set objectives: i.e. consequences and personal values. Configurations of bonds between the attributes, consequences and values called the hierarchical cognitive structure of the product, diversify consumers and due to this, can become the criteria of their segmentation. The article is of methodological character.

1 Introduction

Market segmentation is one of the basic marketing strategies of a company functioning at an international market. Its goal is to utilize the advantage of standardization, and as a result of it increase the company's competitiveness at an international market (Kotler and Armstrong (1988), Rutkowski and Wrzosek (1985)). The concept of a two-stage international market segmentation includes two integrated levels.

STAGE I: Macroeconomic segmentation involving the choice of countries in which the company is planning to start up its activities.

STAGE II: Microeconomic segmentation, within which the identification of homogenous groups of consumers takes place (Komor (2000)).

The traditional approach towards the international market segmentation involves macroeconomic segmentation. The result of its application is the division on the world market into groups of countries. Limiting foreign markets segmentation only to the macroeconomic segmentation is incorrect, because it is based on a false idea of non-divisibility of particular countries and does not take into consideration the advancements of the process of globalization, which refers also to attitudes, preferences and behaviours of consumers.

Microeconomic segmentation is referred to as an integral segmentation. Its purpose is the direct selection of homogenous groups of purchasers, at the selected international markets, disregarding country boarders.

2 The idea of consumers' concepts of means-end chains

The concept of means-end chains developed in 1972 by A. Newell and H. Simon may become the conceptual basis for inter-market segmentation. It turned out very useful in analysing consumers' behaviours, since it finally leads towards understanding their preferences and choices.

Means-end chains, also called the cognitive structure of a product, are hierarchical links between the levels of the knowledge about the product, identified by a consumer.

The levels of the knowledge about the product are represented by three types of notions, selected with regard to their level of abstractness: product's attributes, consequences resulting from using the product and consumer's personal values. The general hierarchy of a cognitive structure of a consumer's product is presented on figure 1.

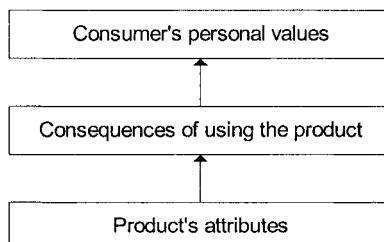


Fig. 1. The hierarchy of a cognitive structure of a consumer's product (Source: own research)

The attributes become the characteristic features of a product, while consequences are the desired advantages resulting from its use (products present attributes and the consumer receives their consequences). The top level of abstraction is occupied by personal values achieved by the consumer as the result of using the product. Attributes and consequences make up the means indispensable for accomplishing the required objectives by a consumer, i.e. personal values. Hence the phrase "means-end chains" (MEC) (Abbott (1955), Claeys et al. (1995), Gutman (1982)).

Configurations of linkages between attributes, consequences and values may diversify consumers. "Means-end chains" become therefore the basis for the identification of market segments in an international intersection, they enable capturing differences in behaviours of consumers living in different countries. This concept may also be useful in the quantification of an intra-national diversification of consumers.

The objective of this paper is to present the procedure, which enables the transition from a theoretical concept of means-end chains towards its practical utilization in the inter-market segmentation process, with the application of multivariate statistical analysis methods.

3 The proposal of inter-market segmentation based on the means-end chains concept

STAGE I: Defining the scope of the notion: "product cognitive structure"

This stage refers to the identification of basic notions necessary for the construction of the cognitive structure of the product (means-end chains) and may be carried out by:

1. Establishing *a priori* by the researcher these notions, which define the attributes of the product, consequences and personal values resulting from its use.
2. Application of the "laddering" procedure on the basis of the limited in size, statistical sample (Sagan (1998)).

STAGE II: The construction of detailed matrices of associations: AC (attributes – consequences) and VC (values – consequences)

Quantitative method of means-end chains measurements is used here. It is called the *Association Pattern Technique* and was suggested by F. ter Hofstede et al.

a. The construction of AC matrix (attributes – consequences)

$$AC^r = [x_{ij}^r]_{(n \times m)} = \begin{bmatrix} x_{11}^r & x_{12}^r & \dots & x_{1m}^r \\ x_{21}^r & x_{22}^r & \dots & x_{2m}^r \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}^r & x_{n2}^r & \dots & x_{nm}^r \end{bmatrix}_{(n \times m)}, \quad (1)$$

where: AC^r – numerical illustration of the product's attributes association structure with consequences identified by an r -th consumer;

x_{ij}^r – value number of associations of i -th attribute with j -th consequence identified by the r -th consumer; $r, s = 1, \dots, R$ – consumer's number; $i = 1, \dots, n$ – product's attribute number; $j = 1, \dots, m$ – number of consequence resulting from using the product;

$x_{ij}^r \in \{0, 1\}$; $x_{ij}^r = 1$ if an r -th consumer identifies the association relation of an i -th attribute with the j -th consequence; $x_{ij}^r = 0$ if an r -th consumer does not identify the association relation of an i -th attribute with the j -th consequence.

b. The construction of CV matrix (consequences – values)

$$CV^r = [y_{jk}^r]_{(m \times p)} = \begin{bmatrix} y_{11}^r & y_{12}^r & \dots & y_{1p}^r \\ y_{21}^r & y_{22}^r & \dots & y_{2p}^r \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1}^r & y_{m2}^r & \dots & y_{mp}^r \end{bmatrix}_{(m \times p)}, \quad (2)$$

where: CV^r – numerical illustration of the consequences association structure with personal values, identified by the r -th consumer; y_{jk}^r – value number of associations of j -th consequence with k -th personal value identified by the r -th consumer; $r, s = 1, \dots, R$ – consumer's number; $j = 1, \dots, m$ – number of consequence resulting from using the product; $k = 1, \dots, p$ – number of consumer's personal value; $y_{jk}^r \in \{0, 1\}$; $y_{jk}^r = 1$ if an r -th consumer identifies the association relation of the j -th consequence with the k -th personal value; $y_{jk}^r = 0$ if an r -th consumer does not identify the association relation of the j -th consequence with the k -th personal value.

STAGE III: The construction of detailed block matrices of ACV associations

The following matrices have to be constructed:

$$AVC^r = \begin{bmatrix} AC^r \\ \cdots \\ CV^{rT} \end{bmatrix}_{[(n+p) \times m]} = \begin{bmatrix} x_{11}^r & \cdots & x_{1m}^r \\ \vdots & \ddots & \vdots \\ x_{n1}^r & \cdots & x_{nm}^r \\ \hline y_{11}^r & \cdots & y_{1m}^r \\ \vdots & \ddots & \vdots \\ y_{p1}^r & \cdots & y_{pm}^r \end{bmatrix}, \quad (3)$$

where: AVC^r – numerical illustration of the product's cognitive structure, identified by the r -th consumer, T – transposition.

STAGE IV: Diversified quantification of consumers with regard to product's cognitive structure

The level of consumers' diversification with regard to the binary coded cognitive structure of a product may be quantified on the basis of four-grid contingency table, presented as figure 2.

		K_s		K_r
		1	0	
1	a	b	a+b	
	0	c	d	c+d
		a+c	b+d	(n+p) x m

Fig. 2. 2×2 contingency table

where: K_r, K_s – respectively r -th, s -th consumer, $r, s = 1, \dots, R$ consumer's number, $a(d)$ – number of association relations consistently identified (not identified) by K_r and K_s consumers, respectively (1; 1) and (0; 0); $b(c)$ –

number of association relations identified exclusively by the $K_r(K_s)$; respectively (1; 0) and (0; 1); $(n+p) \times m$ – number of potential association relations (resulting from the *AVC* matrix size)

On the basis of the contingency table constructed in such a way, one could define the value of the selected association coefficient applied in the multivariate statistical analysis. It is suggested to use one of the basic measurement units of this type, the Sokal and Michener's coefficient (Newell, Simon (1972)):

$$d_{rs} = \frac{b + c}{a + b + c + d} = \frac{b + c}{(n + p) \times m}, \quad (4)$$

where: d_{rs} – the measure of consumers' K_r and K_s differentiating, with regard to the cognitive structure of the product (means-end chains).

The Sokal and Michener's coefficient measures the share of association relations identified inconsistently by the consumers K_r and K_s in the general number of potential relations.

$d_{rs} \in [0; 1]$, $d_{rs} = 0$, in case of identical cognitive structures of the product identified by the consumers K_r and K_s , $d_{rs} = 1$, when none of the potential association relations was consistently identified by the consumers K_r and K_s .

The determination of distance matrices, on the basis of coefficient (4) association values, for all analysed pairs of the product cognitive structures, brings to an end the discussed stage of the research procedure.

STAGE V: Specification of market segments with regard to the cognitive structure of the product

The division of consumers into homogenous groups, with regard to the identified cognitive structure of the product, does not constitute a separate classification problem, that is why the generally recognized methods of the multivariate statistical analysis (Grabiński et al. (1989), Hellwig (1981), Jajuga (1993)) may be applied here. If consumers coming from different countries undergo the analysis, the segmentation may be defined as the inter-market one.

STAGE VI: The construction of aggregate association matrices

For each market segment obtained as the result of completing the previous stage of the research procedure, one should establish the presented below, so called, aggregate association matrix:

$$S^z = \begin{bmatrix} f_{11}^z & \dots & f_{1m}^z \\ \vdots & \ddots & \vdots \\ f_{n1}^z & \dots & f_{nm}^z \\ \hline g_{11}^z & \dots & g_{1m}^z \\ \vdots & \ddots & \vdots \\ g_{p1}^z & \dots & g_{pm}^z \end{bmatrix}_{(n+p) \times m}, \quad (5)$$

where: $f_{ij}^z = \frac{\sum_{r \in G_z} x_{ij}^r}{U_z}$; $g_{kj}^z = \frac{\sum_{r \in G_z} y_{kj}^r}{U_z}$; $f_{ij}^z (g_{kj}^z)$ – frequency of identifying the association procedure of an i -th attribute with j -th consequence (k -th personal value with j -th consequence) in the z -th group of consumers; $z = 1, \dots, Z$ – consumers' group number; G_z – z -th group of consumers; U_z – size of the z -th group of consumers.

Each element of S^z matrix is the frequency, with which the attribute as the element of the row leads towards the consequence as the element of the column (f_{kj}^z), or the frequency with which consequence, as the element of a column leads towards the value as the element of the row (g_{kj}^z).

At this stage one should:

1. Define α – the level of elimination of unusual association relations, using the formulas below:

$$\alpha = \max_j \min_l a_{lj}, \quad (6)$$

or

$$\alpha = \min_j \max_l a_{lj}, \quad (7)$$

where: a_{lj} – matrix S^z element, $l = 1, 2, \dots, (n + p)$.

2. Check the following relations for each f_{kj}^z and g_{kj}^z element

$$f_{ij}^z \leq \alpha \quad (8)$$

$$g_{kj}^z \leq \alpha. \quad (9)$$

In each S^z matrix, the frequency of identifying association relations (f_{kj}^z , g_{kj}^z), which meet the above relations, one should replace with zeros, and leave the remaining frequencies unchanged. In this way modified, aggregate matrices of association S_1^z will be constructed.

STAGE VII: The construction of the hierarchical value map for each market segment

The hierarchical value map is constructed on the basis of the modified, aggregate association matrix.

This map makes up a graphic representation of the means-end chains set, and may be regarded as an aggregate map of the product's cognitive structure, identified by a defined market segment. It is composed of nodes and connections joining these nodes. Nodes represent notions classified as the attributes, consequences and personal values (Valette-Florence and Rapacchi (1991)). Lines joining these notions represent relations between them and are marked when the frequency of the association relations' occurrence between them is bigger than the parameter. Exemplary hierarchical value map for market segment is presented on figure 3.

Comparative analysis of the product's cognitive maps received for the particular segments may supply the following information:

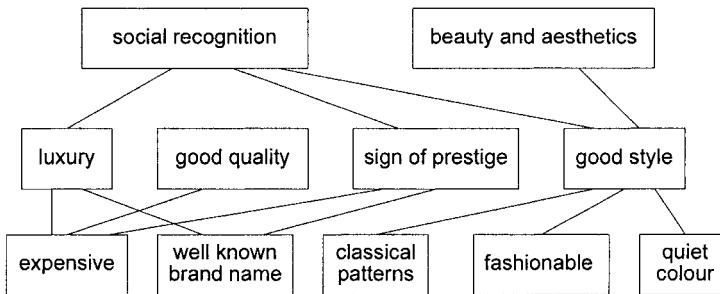


Fig. 3. Two-stage international market segmentation concept (Source: own research)

1. number and type of the most developed features (in the presented segment they refer to such values as social respect, beauty and aesthetics);
2. number of the selected categories and the degree of interrelations' density (fig. 3 presents 11 categories and 13 interrelations, there is 0,85 interrelation per one category);
3. specific interrelations occurring in the segment;
4. common interrelations for all selected segments.

Practical implication of the received research results may become the fact that different advertising strategies or variants of products were used in case of transnational segments, selected due to a different motivation picture presented on the product's cognitive map. In case when a given enterprise is not willing to diversify the assortment structure of the product or its advertising campaign for the selected segments, it should design and plan them taking advantage of the common elements resulting from the analysis of cognitive maps.

The analysis of hierarchical maps of values reflecting specialized, international market segments may become helpful for enterprises in taking up decisions regarding purposefulness and the scope of globalization in conducting marketing activities, and consequently increase chances for elaborating more effective marketing programmes. Therefore it seems that the method of inter-market segmentation, conducted on the basis of means-end chains, may prove its usefulness in global marketing activities of enterprises.

4 Summary

The identification of market segments based on „means-end chains” increases the company's capacity for taking up activities aiming at the development of the product and its adjustment to the expectations of potential buyers. Effective marketing strategy requires the knowledge of links between the product's attributes, consequences and values. The presented proposal of utilizing

means-end chains concept for the purpose of international market segmentation proves the usefulness of multivariate statistical analysis methods for this type of applications.

References

- ABBOTT, L. (1955): *Quality and Competition*. Greenwood Press, Westwood.
- CLAEYS, C., SWINNEN, A., and ABEELE, P.V. (1995): Consumer's Means-End Chains for „Think” and „Feel” Products. *International Journal of Research in Marketing*, 12, 193–208.
- GRABIŃSKI, T., WYDYMUS, S., and ZELIAŚ, A. (1989): *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*. PWN, Warszawa.
- GUTMAN, J. (1982): A Means-End Chain Model Based on Consumer Categorization Processes. *Journal of Marketing*, 46 (Spring), 60–72.
- HELLWIG, Z. (1981): Wielowymiarowa analiza porównawcza i jej zastosowanie w badaniach wielocechowych obiektów gospodarczych. In: Welfe (Ed.): *Metody i modele ekonomiczno-matematyczne w doskonaleniu zarządzania gospodarką socjalistyczną*. PWE, Warszawa.
- JAJUGA, K. (1993): *Statystyczna analiza wielowymiarowa [Multivariate Statistical Analysis]*, PWN, Warszawa.
- KOMOR, M. (2000): *Euromarketing. Strategie marketingowe przedsiębiorstw na eurorynku*. PWN, Warszawa.
- KOTLER, P. and ARMSTRONG, G. (1988): *Principles of Marketing*. Prentice Hall, Englewood Cliffs.
- NEWELL, A. and SIMON, H.A. (1972): *Human Problem Solving*. Prentice Hall, Englewood Cliffs.
- RUTKOWSKI, I. and WRZOSEK, W. (1985): *Strategia marketingowa*. PWE, Warszawa.
- SAGAN, A. (1998): *Badania marketingowe. Podstawowe kierunki*. Wydawnictwo AE, Kraków.
- TER HOFSTEDE, F., AUDENAERT, A., STEENKAMP, J.-B.E.M., and WEDEL, M. (1998): An Investigation into the Association Pattern Technique as a Quantitative Approach to Measuring Means-End Chains. *International Journal of Research in Marketing*, 115, 37–50.
- VALETTE-FLORENCE, P. and RAPACCHI, B. (1991): Improvements in means-end chain analysis. Using graph theory and correspondence analysis. *Journal of Advertising Research*, 1, 30–45.

Statistical Analysis of Innovative Activity

Marek Szajt

Faculty of Management, Chair of Econometrics and Statistics,
Technical University of Czestochowa,
Al. Armii Krajowej 19 b, 42-200 Czestochowa, Poland

Abstract. Nowadays, a significant relationship between innovation and economic growth is emphasised more and more often. European countries are characterised by similar innovative activity and innovation-creating policy, which is confirmed by research, carried out by national government agencies and international institutions. In spite of differences arising out of geographical, historical and social factors, general tendencies in the development of high technology industries and promotion of research and development (R&D) activities are similar. In this paper, an empirical analysis of the endogenous innovative activity is presented. In the research the linear-discrimination function and logit function were used to model processes of innovative activity. Received information can be useful in planning of state innovative politics. In conclusion we can say that the countries characterized by a large increase and a high level of expenditure on R&D activity can reach a higher level of innovative activity and, therefore, generate technological development both on their territories and outside, through export of knowledge.

1 Characteristics of innovative activity determinants

Nowadays, a significant influence of innovation on economic growth is emphasized more and more often. Along with the forming of general economic theories, a lot of different definitions connected with innovations have been created. In Schumpeter's theory of economic growth innovation denotes the achievement of developing an already discovered element for practical (or commercial) use (Schumpeter (1960)). This definition, although very simple, is clear and equivalent to many terms that have emerged in recent years.

In connection with new problems that arise, we also encounter difficulties connected with their precise quantification. First of all, as main source of innovation, patents are adopted. Persons or institutions from home or abroad may submit the patents on the territory of a given country. The number of patents submitted by the residents reflects the activity of a given country in the sphere of research and development (R&D). In order to obtain a better comparability of data concerning the number of patents, the data is quantified per area units or the number of inhabitants.

Innovative activity measured as a number of patents per one thousand of inhabitants is influenced by various factors. Today, as main determinants of innovative activity, the following factors are enumerated: first of all, gross expenditure on the R&D (GERD) sector, employment in this sector and the level of economic growth.

In this paper, an empirical analysis of the above mentioned problems is presented on the basis of information published by the Central Statistical Office and Organization of Economic Cooperation and Development (OECD).

2 Regression line as an indicator of the R&D activity

It would be difficult not to notice the ever-increasing expenditure on research and development activity, both in Poland and in other countries. Despite the fact that Poland spends more money on R&D than results from the tendency observed in the OECD countries (Zolkiewski (1999)), these expenses, due to a low level of national wealth do not have considerable effects. The straight

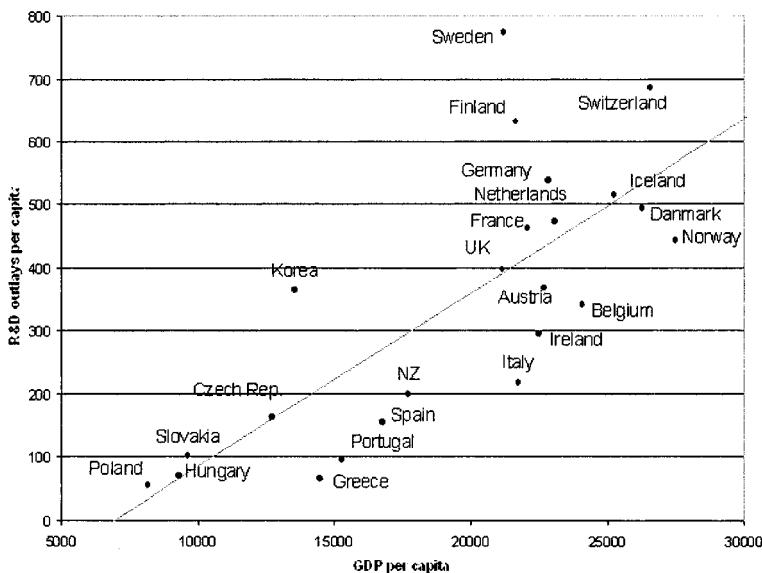


Fig. 1. Expenditure on the R&D activity and GDP in some OECD countries (PPP in USD at 1995 prices)

line depicted in Figure 1 is the regression function in the form:

$$y = -189,8365 + 0,0275 * x$$

$$(104,3242)(0,0052)$$

estimated for the data of 1999, concerning expenditure on R&D per capita in relation to GDP per capita in the OECD countries. Although the fit of this equation measured by the coefficient of determination $R^2 = 0,575$ (adjusted $R^2 = 0,555$) is not convincing, the value of the parameter standing by the

independent variable is significantly different from 0. The countries above this line are characterized by the tendency of investing in this sphere in an higher degree than the OECD average. The countries below this line incur expenses that are lower than expected. However, the presented regression line is not the only way to show lower or higher commitment of individual countries in their investing in R&D. In the case of the analyzed data, the exponential function seems to be a more appropriate discriminant (Morrison (1990)). If for the 90-s we determined both linear and exponential functions, their course would be as follows: Figure 2 suggests that the expenditure on

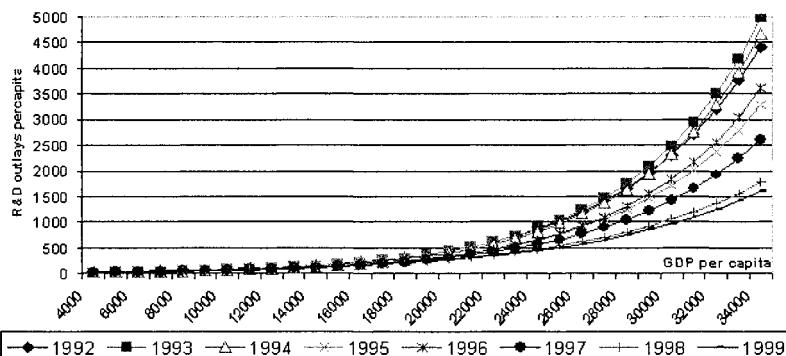


Fig. 2. The trends of outlays on R&D per capita and GDP per capita in the years 1992 - 1999

the R&D activity increases much more slowly in less wealthy countries than in the richer ones. Exceeding a certain wealth point results in an automatic increase in expenditure on the investigated sector. At the beginning of the 90-s this increase was much more dynamic than now as it even exceeded the rate of growth of wealth. The current slow-down can be explained by the present recession all over the world. The recession also affected the R&D sector. From the above charts it follows that these functions look similar in different periods of time, especially in the case of the linear functions. After an evaluation of the fit the analytical form of the proposed functions based on the coefficient of determination, the opinion presented above is confirmed.

The values of the coefficients of determination for exponential functions are in all cases higher than in the case of linear functions. It also should be added that in case of the linear functions, their analytical form in the years 1995 - 96 and 1998 - 99 is improper (Greene (2000)). The conclusions from the above calculations would suggest that the expenditure in some countries was lower than the average determined by the regression line. However, this presentation does not take into account the capabilities of individual countries connected with the wealth of their societies and their R&D policies. For the

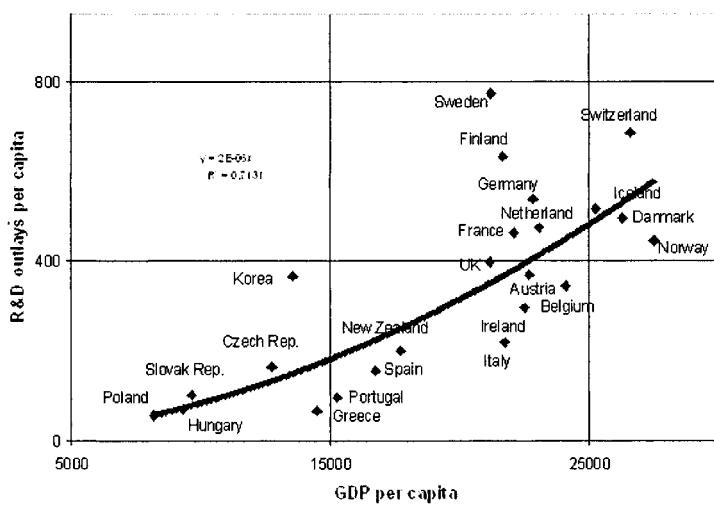


Fig. 3. Outlays on R&D per capita and GDP per capita in 1999 in some OECD countries (PPP in USD at 1995 prices)

Year	R ²	R ² adj.	Run number	R ²	R ² adj.	Run number	S ₁	S ₂
1992	0,706	0,692	12	0,752	0,740	12	7	17
1993	0,738	0,723	12	0,782	0,770	9	5	15
1994	0,748	0,733	9	0,843	0,834	11	6	15
1995	0,712	0,698	5	0,820	0,811	10	7	17
1996	0,732	0,719	7	0,876	0,870	13	7	17
1997	0,698	0,683	8	0,814	0,805	15	7	17
1998	0,640	0,624	7	0,793	0,784	13	7	17
1999	0,685	0,671	7	0,823	0,815	11	8	18

Table 1. Fit of the similar functions

purpose of further analysis, we will use the linear discriminative function (Aczel (2000)). This function was determined on the basis of dichotomic variables both on the explained and explaining side and took up the following values:

$$Y = -1,0367X_1 + 1,4667X_2$$

where variable X_1 - concerns the GDP value per capita,
 X_2 - expenditure on the R&D activity per capita,

The percent of grouped cases correctly classified is on the level 65,22% (hit ratio) and it is higher than the level of arbitrator classification - 52,17%, and the level of proportional cancels criterion - 50,10%.

The values calculated on the basis of this function allow us to determine whether individual countries with their wealth and proper R&D policies obtain results similar to others. As a point of reference, we take the position of

points determined for the data of individual countries in relation to the regression line. The information obtained in this way indicates that Denmark, Italy, Ireland, Belgium, Norway and Austria reach the value of the function that could indicate their higher commitment to financing the R&D activity. An opposite situation concerns the Czech Republic and the Slovakian Republic that, at a relatively low wealth level, are characterized by the B+R activity financing level higher than that indicated by the regression function.

A certain flaw of the Polish system is the unquestionable domination of budget means in financing the R&D activity. In the future, the share of the company sector in financing innovative activity is expected to be bigger. The Polish structures of financing innovations through the company sector are not adequately developed, as a result of which the state budget is excessively burdened. The innovations depend to a higher and higher degree on effective interactions between the scientific base and business sector.

In order to examine the innovative activity, we can use the data concerning either an investigated country or a group of countries. Unfortunately, there are many difficulties connected both with the availability of full data and their comparability. In such a case, we can only use the spatial and time analysis or use analyses based on discrete programming. In this paper, we will present possibilities using discriminative analysis as a method for obtaining division of countries according to the innovative activity criterion. Figure 4 shows

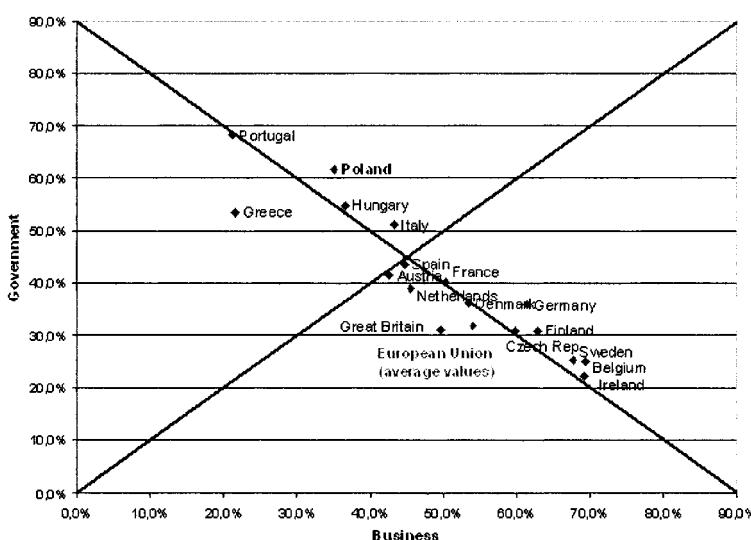


Fig. 4. Share of government and of industry in financing the R&D expenditure

how much financing structure of the R&D sector in Poland differs from that in most EU countries (EUROSTAT (2000)). The upper triangle contains

countries, in which the bigger role in the R&D financing is played by the government. Poland can also be found in this triangle, whereas most EU countries are characterized by higher expenditure incurred by the industry, or these expenditures are balanced. It is also worth mentioning that countries in which the government participation in the R&D financing is the highest are outsiders from the point of view of the expenses presented earlier. Similarly, countries that are in the lead are characterized by a high participation of the private company (industry) sector in the R&D financing.

The R&D expenditure per one inhabitant in the European Union reached in 1998 an average level of \$383.5 and in comparison to 1992, it was an increase by nearly 17% (OECD (2000a, b)). A similar increase (16.1%) characterized OECD countries that reached the level of \$470.1. However, if we take into account the annual level of GERD increase for Poland, it amounted to 11% in 1998, whereas for the European Union, it did not exceed 3.5%.

The difference in the level of development of the R&D sector in the countries of Central and Eastern Europe in comparison with the European Union is clearly visible if we look at them as a whole, taking into account not only their share in the GDP or employment, but also the value of their expenditure. In this breakdown, Poland is head and shoulders above other countries

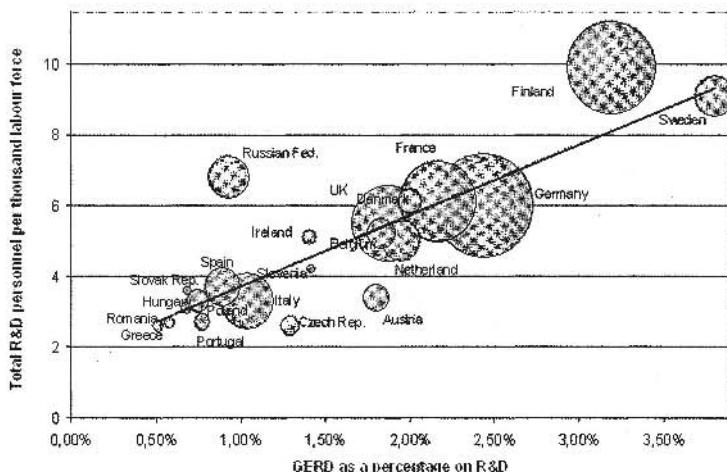


Fig. 5. GERD as a GDP percentage and in millions of USD. Persons employed in the R&D sector per 1000 employed persons in 1999

in this region except for Russia. However, it is easy to notice two essential groups of countries. The first one includes the Mediterranean countries and the ex-Soviet bloc countries that are characterized by a relatively low share of expenditures in the GDP, not exceeding 1.5%. The other group is a group of countries with an employment in the R&D sector exceeding 5 persons per

1000 employed persons and with a budget exceeding 1.5% of the GDP. The weakest member from this point of view is Ireland, in which the expenditures amount only to 1.41%. Slovenia is on its way to join this group. Austria and Russia are atypical countries. In Austria, the expenditure is disproportionate to scientific workers; in Russia, the situation is opposite. What is left is a group of leaders, comprised of Sweden and Finland, where the share of expenditure exceeds 3% and with an employment exceeding 9 persons per 1000 employed persons.

3 A logit model of patent activity

Taking into consideration both the differences and similarities in financing innovations described above, an attempt to measure the innovative activity of Poland, measured by the number of patents submitted in a given country by its residents, was proposed. Due to the nature of the data that we have, this attempt was extended by the information concerning other countries.

Treating the number of patents submitted by the residents per one thousand inhabitants as an empirical probability, we may use models of the logit or probit-type. At the first stage, in order to confirm the correctness of our assumptions, we check the normality of the distribution of the data that we possess. For this purpose, we will use the Shapiro-Wilk test, which is an appropriate test in the case of investing low-size samples. The values obtained

	B	CS	Dk	Fr	H	I	Nl	No	P1	P	E	S
Number of observations	10	4	16	14	8	4	10	10	6	8	12	8
Value of test W	0,88	0,73	0,90	0,85	0,77	0,69	0,83	0,89	0,75	0,84	0,82	0,84
Value $W_{n;0,01}$	0,78	0,69	0,84	0,83	0,75	0,69	0,78	0,78	0,71	0,75	0,81	0,75

Table 2. Values of the Shapiro-Wilk statistics for selected countries

confirm at the level of 99% the normality of the distribution of the investigated variables. Therefore, we move on to further studies.

Taking into consideration that the graphical layout of the expenditure on the R&D activity resembles a logistic function, we may venture to describe it approximately by means of the logit-type function. The essential feature of this function is the nature of the explained variable. It is a logit transformation of the probabilities of occurrence of a given feature. In our case, we will adopt as this feature the share of patents per one thousand inhabitants of a given country in the overall number of patents in a group of countries (PAT). Thus, all our observations, that are, in fact, structure indices, take up values from the 0:1 range, i.e. the range that corresponds to the probability distribution. As explaining variables were proposed:

- G - gross expenditure on the research and development activity in USD, according to the purchasing power parity (\$ PPP) in fixed prices of 1995, per capita.
- R - the number of researchers employed in the research and development sector per one thousand inhabitants.

Data concerning 15 European countries, characterized by similar dependencies between the course of two examined variables in time, were used in the test. Being the OECD members, these countries are in fact the countries that passed the normality of the distribution test. What is favorable in the case of the logit models is the fact, that for the observations we have, we may use two time segments of different length. In our case, lengths of time series amount to 4 - 17 and are annual data. For these time series, harmonic means were calculated and used in the model.

The model was estimated with the help of the Generalized Least Squares Method (Nowak (1998)), calculating the components of the \mathbf{W} matrix diagonal according to the formula (Wisniewski (1986)):

$$w_j = \frac{1}{m_j p_j (1 - p_j)}, \quad j = 1, \dots, m$$

where p_j are empirical values of the probabilities (in our case, the number of patents per 100 people), m_j - the size of the sample for a given country. Calculating the theoretical values of the model, we use the following dependence:

$$\log \frac{\hat{p}}{1 - \hat{p}} = \hat{L} \quad \text{and} \quad \hat{p} = \frac{10^{\hat{L}}}{1 + 10^{\hat{L}}}$$

The form of the logit function in this case is as follows:

Country:	Number of observations	Average number of patents	Average GERD	Number of researchers	Logit	Theoretical values of Logit	Probabilities
Spain	12	0,050	0,075	0,077	-1,279	-1,258	0,052
Hungary	8	0,097	0,062	0,113	-1,971	-1,150	0,066
Poland	6	0,064	0,046	0,137	-1,167	-1,080	0,077
Belgium	10	0,088	0,252	0,172	-1,016	-0,813	0,133
Denmark	16	0,220	0,253	0,216	-0,551	-0,667	0,177
Netherlands	10	0,151	0,336	0,204	-0,749	-0,646	0,184
France	14	0,222	0,394	0,226	-0,544	-0,533	0,227

Table 3. Empirical and theoretical values estimated in the model

$$L = -1,56888 + 0,729641G + 3,316663R \quad R^2 = 0,91542 \quad \bar{R}^2 = 0,896625$$

$$(0,012759) \quad (0,048267) \quad (0,098553)$$

Theoretical GERD \$/ person	Theoretical number of researchers	Theoretical values of Logit	Probabilities	Growth of probabilities
0,001	0,1	-1,23648	0,054831	-
0,1	0,1	-1,16425	0,064117	0,009286
0,2	0,1	-1,09128	0,074967	0,010850
0,3	0,1	-1,01832	0,0877482	0,012515
0,4	0,1	-0,94536	0,101857	0,014374
0,5	0,1	-0,87239	0,118286	0,016430
0,5	0,2	-0,54073	0,223555	0,105269
0,5	0,3	-0,20906	0,381927	0,158372
0,5	0,4	0,122606	0,570113	0,188186
0,5	0,5	0,454273	0,740006	0,169893

Table 4. Simulative probability increases for the model

As we may conclude from the above simulation, for the investigated group of countries, an increase in the expenditure on the R&D activity per capita by 0.1 thousand \$ PPP will result in an increase in the probability (of the share of a given country in the overall number of patents of the selected group of states) on average by 0.168. It is worth mentioning that this effect is getting stronger and stronger. Similarly, when the number of researchers increases by 0.1 person per one thousand, the probability (of the share of a given country in the overall number of patents of the selected group of countries) will decrease on average by 0.004.

4 Conclusions

The above data confirm the importance of the expenditure on research and development activity in creating patent activity and, what follows, innovative activity. As can be easily noticed, these dependencies do not only concern the situation in Poland, but also in other countries. Taking into consideration a relatively stable status of scientific workers in comparison with the investigated expenditure, it is difficult not to give prominence to this factor. Theoretically, the logit model used presents the existence of a very strong dependence between the investigated features. It is shown not only by the value of the model fit, but also by its estimate. In particular, the assessment of the parameter explaining the influence of expenditure on the patent activity shows that its role is significant; it is confirmed by the simulations. To sum it up, we can say that the countries characterized by a large increase and a high level of expenditure on the R&D activity can reach a higher level of innovative activity and, therefore, generate technological development both on their territories and outside, through export of knowledge.

One may say that not every country can afford the great expenses connected with research. The answer is that no one will wait for this country

on the way to economic development. If a given country's budget has a low potential, then, despite its best intentions, this country is usually not able to equal other, financially more powerful countries.

References

- ACZEL, A.D. (2000): *Statystyka w zarządzaniu (Complete Business Statistics)*. PWN, Warszawa.
- EUROSTAT (2002): *R&D and innovation statistics in candidate countries and the Russian Federation*. Office for Official Publications of the European Communities, Luxembourg.
- GREENE, W.H. (2000): *Econometric Analysis*. Prentice Hall, New Jersey.
- MORRISON, D.F. (1990): *Wielowymiarowa analiza statystyczna (Multivariate Statistical Methods)*. PWN, Warszawa.
- NOWAK, E. (1998): *Prognozowanie gospodarcze, metody, modele, zastosowania, przykłady (Economic Forecast, Methods, Models, Applications, Examples)*. Placet.
- OECD (2000a): *Science and Technology Industry Outlook 2000*. OECD, Paris.
- OECD (2000b): *Main Science and Technology Indicators, No2*. OECD, Paris.
- SCHUMPETER, J.A. (1960): *Teoria rozwoju gospodarczego (Theory of Economic Development)*. PWN, Warszawa.
- WISNIEWSKI, J. (1986): *Ekonometryczne badania zjawisk jakociowych. Studium metodologiczne (Econometric Research of Qualitative Phenomenon. Methodological Study)*. Wydawnictwo Uniwersytetu im. M. Kopernika, Toruń.
- ZOLKIEWSKI, Z. (ed.) (1999): *Rachunek Satelitarny Nauki 1996-1997 (A Satellite Account of Science 1996-1997)*. Studia i Prace z Prac Zakładu Badan Statystyczno-Ekonomicznych, GUS, Warszawa.

The Prospects of Electronic Commerce: The Case of the Food Industry

Ludwig Theuvsen

Institute of Agricultural Economics,
Georg-August-University Göttingen,
Platz der Göttinger Sieben, D-37073 Göttingen, Germany
theuvsen@uni-goettingen.de

Abstract. Is it possible for food manufacturers to sell their products to consumers on the Internet? This paper analyzes this question by identifying the prerequisites for establishing successful business-to-consumer online sales activities. Applying ideas taken from the resource-based and the market-based view in strategic management and the marketing literature, it is argued that successful electronic commerce in the food sector depends on rareness, non-substitutability, high consumer involvement and differentiation.

1 Electronic commerce and the food industry

The food industry is one of Europe's largest industries. Unlike many other industries, it is still dominated by small and medium-sized companies in many countries. In Germany, e.g., more than 75 % of all food manufacturers have less than 100 employees and only 0.4 % of the food companies have more than 1,000 employees (BMVEL (2002)). Retailing on the other hand is dominated by large companies in many countries. In Germany, e.g., the 10 largest retailers have a market share of about 84 %, and the leading 30 retailers share about 98 % of the market. For these large chains it is easy to play small and medium-sized food manufacturers off against each other since most of these manufacturers lack strong and therefore irreplaceable branded goods. This results in heavy pressure on prices and a low financial performance of many small and medium-sized food manufacturers. Many sectors of the food industry, thus, face deep structural changes which threaten the financial success and the survival of minor and even some major national food manufacturers (Menrad (2001)).

In view of this situation an alternative distribution channel like the Internet is an attractive prospect for small and medium-sized manufacturers; nevertheless, the recent significance of electronic commerce (eCommerce) in the food sector is still very low. In Germany, only 0.1 % of foods were sold on the Internet in 2001 predominantly by companies which also operate traditional retailing activities (Morath and Doluschitz (2002)). So far only very few food manufacturers have entered eCommerce.

At first glance this is surprising since the Internet promised to make things easier, cheaper, and faster. Authors argue that the Internet enhances

sell channels as well as buy channels (Deise et al. (2000)). And the Internet promised to enable direct sales to consumers, i.e. to establish successful business-to-consumer (B2C) activities. This results in disintermediation, i.e. a shortening of value chain by skipping some of its parts, e.g. retailers (Shapiro and Varian (1999)). But sceptics question the validity of these effects in the food industry. In fact, food is different from other products for several reasons:

- Food is a non-digital product which has to be delivered by postal services. Unfortunately, the ratio between product value and logistic costs is unfavorable for most food products.
- Food is a little bit unwieldy and cannot be put in a letter-box in many cases. Consumers, thus, have to organize delivery when they are not at home during the arrival of the products.
- Fresh products are perishable and characterized by high quality uncertainty. Consumers prefer personal inspection before buying these products.
- The market share of branded products is low in some important product groups (e.g. meat). Consumers, thus, cannot reduce quality uncertainty by relying on branded products.
- Refrigerated and frozen products are difficult to handle and to deliver due to the necessity of uninterrupted logistic chains.
- Most consumers view food as a convenience product which they want to buy easily and cheaply. Consumers, thus, are not willing to search for food on the Internet, crawl through sometimes difficult to handle websites, pay high postage and packing costs or waste their time with organizing deliveries.

For these reasons many authors see very bad opportunities for establishing successful B2C eCommerce activities in the food industry (Wilke (2002); Schmidt (2003)). In fact, fresh, refrigerated and frozen products are certainly rather unsuitable for B2C eCommerce activities. In other cases it is really hard to see why consumers should be willing to bear the costs and strains of eCommerce since many products can be bought simpler and cheaper just around the corner in the next supermarket. Authors like de Figueiredo (2000) obviously were a little bit too optimistic about the prospects of eCommerce in the food sector.

Does this mean that the Internet will not offer any chances for food manufacturers at all? We do not think so. Nevertheless we admit that eCommerce is not a viable option for the vast majority of food manufacturers. But there are some interesting exceptions to the rule that food cannot be sold successfully on the Internet. Several small and medium-sized German food manufacturers were able to establish successful online shops which now contribute a considerable share of their sales and profits. What is different about these manufacturers? Why are consumers willing to buy the products of these manufacturers on the Internet?

This paper has a closer look at these exceptions to the rule. Based on several case studies in German food manufacturers with online activities, the paper argues that some companies were able to establish successful eCommerce activities since their products meet the conditions of rareness and non-substitutability, high consumer involvement and differentiation. These prerequisites for successful B2C eCommerce in the food industry are carved out in more detail in the next chapters and stated as propositions. A model of successful B2C eCommerce activities in the food industry is derived as a starting point for further empirical research. Since fresh, refrigerated and frozen products are unsuitable for B2C eCommerce activities for technical reasons, the analysis is limited to packed products with a minimum durability of several weeks up to several years. Due to space limitations, the paper is restricted to the analysis of direct sales by manufacturers through online shops on their websites. It does not analyze the possibility of offering food in online auctions (e.g. Ebay) or in Internet shopping malls.

2 The precondition of rareness and non-substitutability

The resource-based view in strategic management argues that rare and non-substitutable resources can be a source of sustained competitive advantage (Barney (2002)). These resources enable a company to develop core competence and to offer products or services that are valuable to customers (Prahalad and Hamel (1990)). "Pedestrian" resources (Montgomery (1995)) on the other hand do not create a competitive advantage. Coined on the eCommerce problem we can argue that food which can be bought almost everywhere or can easily be substituted cannot be sold successfully on the Internet since consumers will not be willing to bear the costs and strains of eCommerce. Only rare products give manufacturers a competitive advantage over retailers and motivate consumers to buy on the Internet. In retailing the rareness of products is determined by the size of the sales territory and the depth of distribution (Kotler (2002)). A limited sales territory is typical for many small and medium-sized manufacturers with a regional focus. Such manufacturers have the opportunity of successfully selling their products on the Internet. Online activities give consumers outside the companies' traditional sales territory the opportunity to buy these products. This leads to the first proposition:

Proposition 1: The rarer a food product is outside the Internet, i.e. the more restricted its sales territory is, the more likely it is to be successfully sold on the Internet.

Companies which do not restrict their sales territory can create rareness by limiting the depth of distribution which is defined by the number of marketing intermediaries which distribute a product. Intensive distribution strategies are characterized by broad distribution through as many retailers and stores as possible. Exclusive and selective distribution strategies on the

other hand offer opportunities for online sales to consumers, if the number of retailers and stores is strictly limited. The most pronounced form of selective distribution is the creation of exclusive online offers which are not available offline and which might be able to attract consumers to the manufacturer's website.

Very often food is distributed intensively. These products cannot be sold successfully on the Internet due to a lack of rareness. But there are exceptions to the rule: Many producers of luxury goods, e.g., have exclusive distribution strategies and restrict their sales to specialty shops. For other companies, e.g. wineries and distillers, it is quite easy to create special versions of their products by using different bottles, labels or brands which are sold exclusively on the Internet. Thus,

Proposition 2: The rarer a food product is outside the Internet, i.e. the more limited the depth of distribution is, the more likely it is to be successfully sold on the Internet.

The resource-based view in strategic management emphasizes the importance of non-substitutability for gaining and sustaining competitive advantage (Barney (2002)). Non-substitutability is an important aspect in eCommerce, too. If consumers can easily substitute a product by buying another brand, private-label products, or different products which approximately meet the same needs, they will be reluctant to buy the original product on the Internet. Whether two products are viewed as substitutable depends on product features, e.g. ingredients, taste and quality, and personal characteristics, e.g. brand loyalty.

Some small and medium-sized food manufacturers are specialized in regional specialties or non-standard production technologies (high ecological standards, traditional production methods and so on). These products are sometimes hard to substitute due to a lack of comparable alternatives and offer opportunities for online sales.

Proposition 3: The harder to substitute a food product due to product characteristics or consumers' brand loyalty, the more likely it is to be successfully sold on the Internet.

3 The precondition of high consumer involvement

Consumer involvement is defined as "the intensity of interest with which consumers approach their dealings with the marketplace" (Loudon and Della Bitta (1993), 341). High involvement describes situations in which consumers are emotionally or cognitively involved with the product or the buying decision. Consumers are motivated to collect information about products and to weigh pros and cons against each other, i.e. to decide carefully before spending money. Under low-involvement conditions consumers are not interested in the product at all, show little attention, and do not actively search for information (Loudon and Della Bitta (1993)).

Ecommerce depends on the motivated consumer who is willing to overcome the obstacles of buying food on the Internet, i.e. to search for online offers, to crawl through websites, to accept transportation costs, to organize delivery and so on. High consumer involvement, thus, is an important precondition for successful online activities. Low involvement products will not be able to motivate the consumer to search for food on the Internet.

The intensity of involvement is the result of a complex mix of product, personal and situational characteristics as well as moderating factors. Important personal characteristics are needs, values, experiences and interests. On the product level the perceived technical and social risk, the relatedness with personal values, interests, experiences, etc. and the promotional strategy (media choice, nature of the message) determine the consumers' interest. The situation embraces the use which will be made of the product, the occasion for which a purchase is being made, or new information about the health risks, e.g., associated with product use. Moderating factors are the opportunity to process information (reduced, e.g., by a lack of time or distractions) and the ability to process information (depends on, e.g., the consumers' knowledge and information processing capacities) (Loudon and Della Bitta (1993)).

Due do the large number of influencing and moderating factors it is impossible to generally attribute high or low involvement to specific products. Nevertheless, food is often viewed as a low-involvement or convenience product. Buying decisions are made superficially; consumers have an evoked set of alternatives from which they choose. They are just trying to minimize the costs and strains of buying food (Nieschlag et al. (1997), 154). In principle, the low-involvement character of food restricts the chances for successfully selling food on the Internet. But there are some situations in which food becomes a high involvement product due to personal, product or situational characteristics:

The personal characteristics favor online sales, if food relates to the personal needs, values, experiences and interests of consumers. Food, e.g. wine, beer or whiskey, may have become a hobby for consumers. Or food reminds consumers of outstanding situations in their life like holidays (e.g. regional specialties). Food might also indicate the belonging to an attractive social group (e.g. Mediterranean products). Thus,

Proposition 4: The higher the consumer involvement in a food product is, i.e. the closer food relates to the personal needs, values, experiences and interests of consumers, the more likely it is to be successfully sold on the Internet.

The product characteristics favor online sales, if consumers want to avoid technical or social risks they attribute to food. Consumers may, e.g., reject food containing genetically modified organisms or breaching religious principles. In these cases consumers may be highly motivated to find food which avoids these risks on the Internet.

Proposition 5: The higher the consumer involvement in a food product is, i.e. the more consumers attribute technical or social risks to food consumption, the more likely it is to be successfully sold on the Internet.

Situational characteristics favor online sales, if consumers have a special use or occasion in mind. Chocolate, e.g., is a typical convenience product, but it becomes a high-involvement product as soon as it is considered a gift.

Proposition 6: The higher the consumer involvement in a food product is, i.e. the more consumers have a special use or occasion in mind when buying food, the more likely it is to be successfully sold on the Internet.

In these cases food becomes a high-involvement product which consumers might be motivated to buy on the Internet. Manufacturers producing articles which relate to personal needs, values, experiences and interests, avoid technical or social risks or are considered for special uses or occasions, e.g. wineries, distillers, and producers of regional specialties or luxury goods, might be able to establish a successful web presence.

4 The precondition of differentiation

Due to a rapidly growing number of websites the Internet is characterized by the 'economics of attention' (Shapiro and Varian (1999)). Companies establishing an Internet presence, thus, compete for consumers' attention and hits on their websites as a precondition for selling their products. The battle for attention is fought by giving a website or the products offered something special which attracts and ties consumers. Otherwise the companies risk that nobody will take notice of their web presence and that their online shops will fail.

According to Michael Porter (1980) companies have two strategic options for overcoming a lack of attention: cost leadership and differentiation (or a mix of both). This is still true for the Internet economy (Porter (2001)) so a successful website can be created by offering products at the lowest prices or by differentiating the website or the products offered. In many countries, e.g. in Germany, hard and soft discounters have the cost leadership position in food retailing. Since sending foods by parcel services causes comparatively high postage and packing costs, companies offering their products on the Internet cannot successfully implement a cost leadership strategy. Selling food on the Internet, thus, requires a differentiation strategy.

There are numerous ways of differentiating a product; "the potential in any product or service for differentiation is limited only by the boundaries of human imagination" (Grant (1998), 219). In B2C eCommerce the most important differentiation variables are (Theuvsen (2002)): the product itself, the placement of banners on heavily-frequented websites, state-of-the-art product placements that use the multimedia features of the Internet, above-average customer support (e.g. complete information about food ingredients), the introduction of one-to-one marketing, e.g. personalized buying recommenda-

tions, the use of brand names, customer loyalty programs (e.g. long-service bonuses), and the establishment of virtual communities on a company's website.

For many small and medium-sized manufacturers it is difficult to differentiate their products or websites. They often lack strong brands, have limited financial resources necessary for successful branding or effective customer loyalty programs and so on. But there are also some companies with a comparatively strong financial base, strong brands or a well-known company name attracting online visitors. These companies can utilize these resources for establishing an online shop. Thus,

Proposition 7: The better differentiated a web presence or the products are, the more likely it is to successfully sell food on the Internet.

The analysis so far revealed several preconditions for a successful web presence in the food sector. These results were set out in the form of propositions. Figure 1 integrates these propositions into a model that provides a conceptual basis for assessing the chances of food manufacturers in B2C eCommerce.

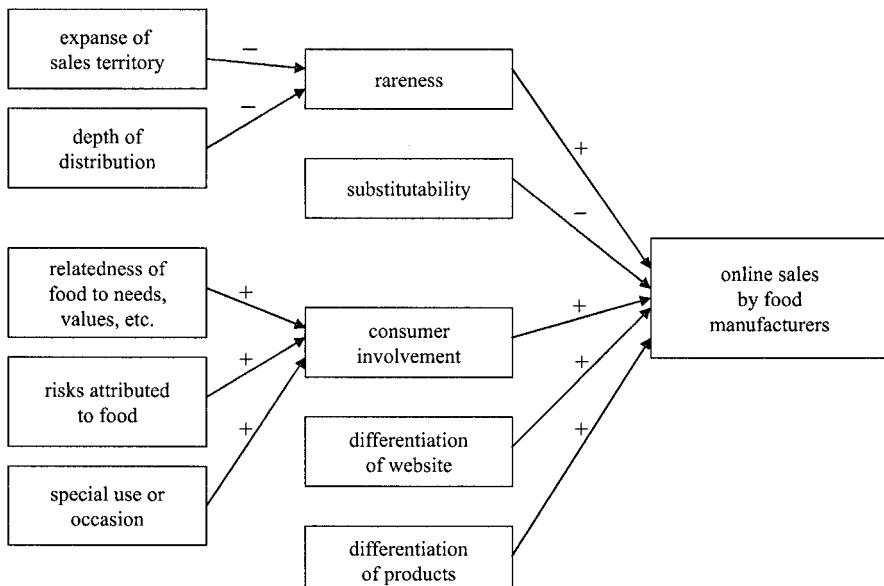


Fig. 1. The prospects of eCommerce in the food industry

5 Food sales on the internet: Some empirical results

According to the grounded theory approach, management researchers must generate formal theories in order to advance understanding of the economic world. However, to be valid, Glaser and Strauss (1967) insist that formal theory must be developed from a substantive grounding in concrete social or economic situations. Substantive theory is often derived from qualitative research, for example case studies. In early phases of theorizing, case studies offer the opportunity of testing prior theory by replicating previous cases and extending theory by choosing cases that provide the opportunity of filling in theoretical formulations or are the polar opposite of previous cases (Yin (1994), Locke (2001)). Table 1, thus, shortly presents five case studies which in general are in line with the propositions presented above. The cases, thus, set the stage for the development of the model presented above as well as more advanced empirical work.

Company/URL	Rareness	Substitutability	Involvement
Klosterbrauerei Neuzelle, Neuzelle www.klosterbrauerei.com	limited offline sales territory	unusual new and traditional recipes	beer reminds consumers of the great day spent at the famous monastery of Neuzelle
St. Martin Brauerei, Lahnstein www.st-martin-brauerei.de	limited offline sales territory	standard offers (e.g. Pilsener beer) as well as specialties	low
Welt-des-Bieres GmbH, Wiesbaden www.welt-des-bieres.de	low for standard offers high for special offers (e.g. beer bottles as a M. & R. Schumacher special edition) (often out-of-stock in offline sales channels)	high very low due to unique product features	low high for "beer collectors" and Formula One fans
Caviar House GmbH Deutschland www.caviar-house.de	limited depth of distribution	low due to unique product features	moderate (emotional luxury product)
Balik Räucherei im Toggenburg AG, Schweiz www.balik.com	limited depth of distribution	comparatively low due to above-average product quality	moderate (emotional luxury product)

Table 1. Electronic commerce in the food industry: Some empirical results

Company/URL	Differentiation	Online success
Klostereibrauerei Neuzelle	neighboring monastery publishes name; consumer-specific labels; specialties	very high (5 % of total sales compared to an average of 0.1 % in the food industry)
St. Martin Brauerei	regional specialist; good corporate citizenship	low (about one online order per month)
Welt des Bieres	low	very moderate
	high due to unique product features	very high
Caviar House	trusted and well-known market leader, but: Internet does not offer personalized buying experience	moderate (highly frequented website; most orders, though, still by telephone calls)
Balik Räucherei	secret receipt can be traced back to the court of the Russian Tsar; hand-made products; but: no personalization	moderate (highly frequented website; most orders, though, still by telephone calls)

Table 2. Electronic commerce in the food industry: Some empirical results (continued)

6 Summary and conclusions

In this paper we developed a model of successful eCommerce activities in the food industry. This model emphasizes the prerequisites of rareness, non-substitutability, high consumer involvement and differentiation. Without doubt, these characteristics will only apply to the products of a minority of food manufacturers. The well-known large international food manufacturers and national major branders, e.g., cannot sell their products on the Internet due to a lack of rareness. Other companies produce highly replaceable products for which there are numerous substitutes. But a handful of companies which fulfill the above-mentioned preconditions can be successful in eCommerce. Some examples have already been mentioned in earlier chapters: small vine-growers and distillers with high-quality products, producers of regional specialties with a limited sales territory, producers of food which respects certain religious or other principles which are important to consumers who want to avoid social or technical risks attributed to food consumption. In-depth empirical research is necessary to verify the results of the case studies and to test the model presented above.

References

- BARNEY, J.B. (2002): *Gaining and Sustaining Competitive Advantage*. 2nd edn., Prentice Hall, Upper Saddle River, NJ.
- BMVEL (Ed.) (2002): *Statistisches Jahrbuch über Ernährung, Landwirtschaft und Forsten der Bundesrepublik Deutschland*, 46, Münster.
- DE FIGUEIREDO, J.M. (2000): Finding Sustainable Profitability in Electronic Commerce. *Sloan Management Review* 41, 4, 41–52.
- DEISE, M.V., NOWIKOW, C., KING, P., and WRIGHT, A. (2000): *Executive's Guide to E-Business*. Wiley, New York.
- GLASER, B.G. and STRAUSS, A.L. (1967): *The Discovery of Grounded Theory*. Aldine, Chicago.
- GRANT, R.M. (1998): *Contemporary Strategy Analysis*. 3rd edn., Blackwell, Malden, MA.
- KOTLER, P. (2002): *Marketing Management*. 11th edn., Prentice Hall, Upper Saddle River, NJ.
- LOCKE, K. (2001): *Grounded Theory in Management Research*. Sage, London.
- LOUDON, D.L. and DELLA BITTA, A.J. (1993): *Consumer Behavior*. 4th edn., McGraw Hill, New York.
- MENRAD, K. (2001): Entwicklungstendenzen im Ernährungsgewerbe und im Lebensmittelhandel in Deutschland. *Berichte über Landwirtschaft*, 79, 597–627.
- MONTGOMERY, C.A. (1995): Of Diamonds and Rust: A New Look at Resources. In: C.A. Montgomery (Ed.): *Resource-Based and Evolutionary Theories of the Firm: Towards a Synthesis*. Kluwer, Boston, 251–268.
- MORATH, C. and DOLUSCHITZ, R. (2002): Lebensmittelhandel im Internet: Konzepte, Erfahrungen, Potentiale. *Zeitschrift für Agrarinformatik*, 10, 60–65.
- NIESCHLAG, R., DICHTL, E., and HÖRSCHGEN, H. (1997): *Marketing*. 18th edn., Duncker & Humblot, Berlin.
- PORTER, M.E. (1980): *Competitive Strategy*. Free Press, New York.
- PORTER, M.E. (2001): Strategy and the Internet. *Harvard Business Review*, 79, March, 63–78.
- PRAHALAD, C.K and HAMEL, G. (1990): The Core Competence of the Corporation. *Harvard Business Review*, 68, 3 79–91.
- SCHMIDT, H. (2003): Internet 3.0. *Frankfurter Allgemeine Zeitung* March 11, 13.
- SHAPIRO, C. and VARIAN, H.R. (1999): *Information Rules*. Harvard Business School Press, Boston, MA.
- THEUVSEN, L. (2002): Lebensmittelvertrieb über das Internet: Chancen und Strategien kleiner und mittelständischer Hersteller. *Zeitschrift für Agrarinformatik*, 10, 41–50.
- WILKE, K. (2002): Schwieriges Feld. *Handelsjournal*, 2, 34–35.
- YIN, R.K. (1994): *Case Study Research*, 2nd ed., Sage, Thousand Oaks, CA.

Part VI

Finance, Capital Markets, and Risk Management

Macroeconomic Factors and Stock Returns in Germany

Wolfgang Bessler and Heiko Opfer

Center for Finance and Banking,
Justus-Liebig-University Giessen, D-35394 Giessen, Germany

Abstract. The objective of this study is to investigate the importance of various macroeconomic factors in explaining the return structure for a bank index and five German industrial indices for the period from 1974 to 2000. The empirical analysis focuses on the decomposition of variances and the estimation of beta coefficients for various macroeconomic factors. A rolling regression technique is applied in order to identify a possible time variation of the regression coefficients. Overall we find empirical evidence of the time variation of the explanatory power and the regression coefficients. Moreover, banks are especially exposed to interest rate risk.

1 Introduction

There has been a long history in the area of empirical capital market research to analyze and explain the factors that determine stock returns. In addition to the traditional equilibrium and arbitrage based models, such as the CAPM and the APT, a number of multi-factor asset pricing models have been developed. These models are based on the assumption that the stock returns are generated by a limited number of economic variables or factors. Most of the empirical studies concentrate on analyzing the stock returns of industrial firms for the capital market in the United States. In most studies, however, it is assumed that the parameters are time invariant. The objective of our own empirical study is to investigate the time-variability of the explanatory power and the beta coefficients in multi-factor models. Thus, we employ an estimation procedure that allows for the time variability of the model coefficients. In the empirical analysis we include various industry indices for the German capital market. In order to investigate whether banks are special and require a special asset-pricing model we first focus on the index of financial institutions and then compare these results to five industrial indices.

2 Literature review

Multi-factor asset pricing models are based on the assumption that stock returns are influenced directly or indirectly by a number of different economic factors. This view is empirically supported, for example, by the study of Chen et al. (1986). In their model the authors identify the following variables as

the most important factors that help to explain stock returns: the growth rate of industrial production, the expected and unexpected inflation rates, the default risk premium, and the maturity risk premium.

When analyzing and modelling the return generating process for financial institutions it becomes apparent that changes in interest rates have a major impact on bank stock returns. This special sensitivity of banks to changes in interest rates can be explained by the fact that banks traditionally perform maturity transformation. This view is supported, for example, by the empirical studies of Flannery et al. (1984), Bessler et al. (1994), and Bessler et al. (2003a). Their empirical results confirm the greater sensitivity of bank stocks to interest rates changes compared to industrial firms. Moreover, the sensitivity depends on the risk exposure, i.e. assets-liability structure, of the bank. There is little empirical research in explaining stock returns for the German capital market with respect to the impact of macroeconomic factors. The research so far includes studies by Nowak (1994) and Sauer (1994). These studies, however, focus either on the analysis of individual stocks, randomized stock-portfolios or specific market segments such as the DAX or the OTC-market. None of this research so far differentiates between various industry groups or analyzes the return generating process for banks.

3 Model

The basic assumption in factor models is that there exists a return generating process that can be explained by a number of economic factors. In order to analyze the impact of various economic factors and to test their significance the following model structure is usually assumed

$$r_i - r_f = \alpha_i + \beta_{1,i}f_1 + \dots + \beta_{k,i}f_k + \epsilon_i. \quad (1)$$

The excess return (return minus the risk free rate r_f) of a risky security i is a function of k common factors that are each weighted by their relevant beta coefficient. The security specific constant α_i represents the security return that is not explained by the k common factors. Moreover, there exists an error term ϵ_i that explains the unsystematic risk of the security.

One important aspect of this study is to investigate the decomposition of variance. The variance of the excess return of a security depends on the variance of the k factors that are weighted with the squared beta coefficients as well as a security specific variance.

$$V(r_i - r_f) = \beta_{1,i}^2 V(f_1) + \beta_{2,i}^2 V(f_2) + \dots + \beta_{k,i}^2 V(f_k) + V(\epsilon_i) \quad (2)$$

The total variance of the returns is the sum of the variance that can be explained by the k factors and the unexplained variance, which is called specific variance. The last component includes all security specific factors that cannot be explained by the common factors. The coefficient of determination of the

regression is identical with the part that is explained by the common factors. The variance decomposition relies on the assumption that the factors are uncorrelated. Therefore, we need to orthogonalize the factors. In addition, we employ a rolling regression technique in this study in order to investigate the time variability of the regression coefficients. The estimation period for the rolling regression coefficients is 60 months. In our empirical results we report the regression coefficient for the date of the last data point in the estimation period. The parameters are estimated by employing an OLS regression technique that has autocorrelation and heteroscedastic consistent standard errors as suggested by Newey et al. (1987).

4 Data

The empirical analysis focuses on six industry-research indices (DAFOX) that were constructed by the University of Karlsruhe. These are all market weighted performance indices. They include all stocks of a specific industry that are traded on the Frankfurt Stock Exchange in the official market segment. One focus of this paper is on banks and insurance companies that are represented in the index of financial institutions. To simplify we will refer to this index as the bank index. For comparison reasons we also include the following five industrial indices in our analysis: chemicals, utilities, vehicles, construction companies as well as consumer goods. We use monthly excess returns for the period from March 1974 to December 2000. As the risk free rate we employ the three months money market rate. The following macroeconomic factors (abbreviations in parenthesis) are used as independent variables in the analysis:

- The maturity risk premium which is the difference between the returns of zero coupon bonds with maturities of ten and one year (TERM).
- The return of a zero coupon bond with ten years to maturity (LTIR).
- The DM/US\$ foreign exchange rate (USD).
- The ifo business climate index (IFO)
- and finally the return of the DAFOX is included as the market index which represents the impact of the stock market in general (DAFOX).

The selection of these factors is based on previous research on multi-factor models for the German market (see Sauer (1994) and Nowak (1994)). The factors are the monthly changes of the variables. In order to correct for the multi-collinearity between the various variables we orthogonalize the time series in specified sequence. Thus, every factor is corrected for the impact of the factors with a higher order.

5 Empirical results

The objective of this study is to analyze the importance of various macroeconomic factors in explaining the return structure for six German industry

indices for the period from 1974 to 2000. The empirical analysis concentrates first on decomposing the variances and then on estimating the beta coefficients for various macroeconomic factors.

5.1 Average factor specific explanations

In Table 1 we report the average factor specific explanations of the variance for the five factors that were employed in our empirical analysis. It is evident

Industry	TERM	LTIR	USD	IFO	DAFOX	Specific
Banks	3.1%	16.8%	1.8%	2.8%	60.6%	15.0%
Chemicals	3.6%	7.2%	6.9%	3.9%	56.0%	22.3%
Utilities	2.1%	8.2%	2.6%	1.9%	48.1%	37.1%
Vehicles	1.8%	4.6%	5.4%	5.8%	66.9%	15.6%
Constructions	1.7%	6.7%	4.0%	1.9%	40.8%	44.9%
Consumer Goods	2.7%	4.9%	4.8%	2.6%	52.6%	32.5%

Table 1. Average factor specific explanations (1974-2000)

that the market index is the predominant factor in explaining stock returns in the German capital market. On average this factor explains between 40.8% and 66.9% of the variance. The four macroeconomic factors are able to explain between 11.8% and 24.4% of the variance. Given the predetermined specification and ranking of the selected economic variables it is evident that the model is best suited for explaining the stock returns for banks. Given the nature of the banking business the predominant macroeconomic factor for banks are changes in long-term interest rates (16.8%). In contrast, for the other five indices only 4.6% to 8.2% of the variation can be explained by long-term interest rates. The highest explanatory power (8.2%) is observed for utilities which is consistent with findings for other countries. One explanation for this result is the relative importance of long-term liabilities on the balance sheet of utilities.

A comparison of the contribution of the exchange rate (DM/US\$) to the variance of the indices reveals that there is a clear separation between the industries that are export and import oriented and those that usually concentrate on domestic business. The indices for which the exchange rate has the lowest explanatory power are banks (1.8%) and utilities (2.7%). For the export oriented industries the exchange rate has a much higher explanatory power (5.4% for vehicles and 6.9% for chemicals). We also find a different impact of the ifo business climate index. This factor should explain the impact of the expected business climate on the index returns. As expected we observe the highest explanatory power for the business cycle sensitive vehicles index (5.8%) whereas the relatively stable utilities show the lowest

sensitivity (1.9%) to this factor. For the maturity risk premium we observe a relative small difference in explanatory power between the various indices. The specific part, i.e. the variation that cannot be explained by the five factors, includes additional variables that may be due to industry specific or firm specific factors. The specific part has the highest values for the utilities (37.1%) and for the construction index (44.9%). However, these results are much stronger compared to other studies that employ various stock indices as independent variables. A comparison of the explanatory power of the tested model structure for the six indices provides clear evidence that the bank index is best explained by this model.

5.2 Decomposing the variance over time

After analyzing the average explanatory power of the factors over the entire period we now concentrate on the decomposition of variance over time. The importance of the various factors becomes even more evident when the focus is on the time-variability instead of the averages. In Figure 1 we present the decomposition of variance for the bank index over time. It is quite evident that the DAFOX and the long-term interest rate are the dominant factors.

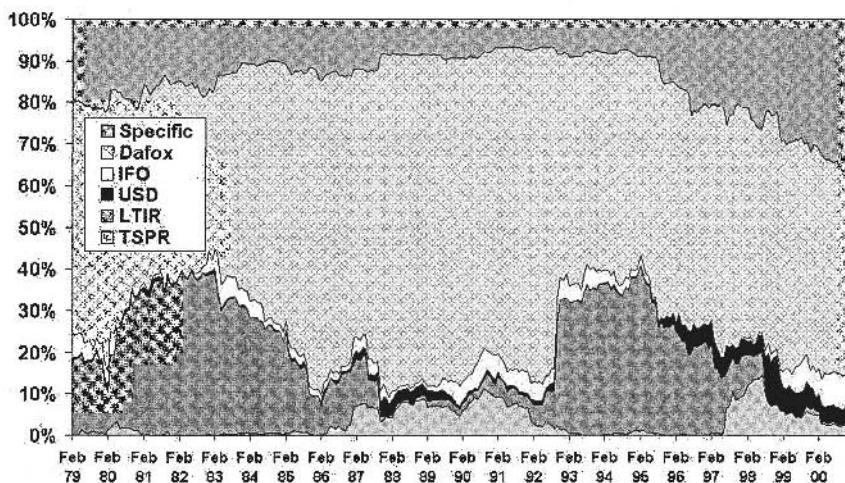


Fig. 1. Decomposition of variance for the bank index over time

Moreover, the importance of the various macroeconomic factors is time variant. For the time period from 1979 to 1987 and from 1993 to 1997 the long-term interest rate represents the most important economic factor in that it explains up to 40% of the variability of stock returns. The period in between, however, is characterized by a lower impact of the macroeconomic factors. Interestingly, the returns for this period as well as for the period after 1997

are best explained by the maturity risk premium. Thus, the importance of the long term interest rate increases with the level of interest rates. This result is consistent with the observation that banks are especially exposed to interest rate risk in a high interest rate environment (Bessler (2001)). This is due to the usually positive maturity transformation of banks. The importance of the foreign exchange rate (DM/US\$) increases substantially since 1996. Before 1996 this factor has contributed only marginally to the explanation of stock returns in this index. In nearly every period the DAFOX shows the highest explanatory power. However, we observe a decreasing contribution since 1995 that results in an increase of the specific part. The other five indices show a similar behavior (for details see Bessler et al. (2003b)). Consequently, we find an increase in the contribution of the specific component during the second half of the 1990s. This result can be explained with the pronounced diversion in the development between various industries during the 1990s. However, the explanatory power of the model exceeds the one from comparable studies that employ world, country, and industry indices as independent variables (see Financial Markets and Portfolio Management (1990-2003)).

The long-term interest rate does not have the same importance for the five industrial indices as for the banks. The explanatory power of the factor shows a similar pattern, however, at a different level. This result confirms the special interest rate sensitivity of banks due to the very nature of the banking business. The development of the explanatory power of the DM/US\$ foreign exchange rate is quite interesting. For all five industrial indices the importance of this factor increases since the middle of the 1980s and exhibits another shift at the end of the 1990s. This result can be explained on the one hand with the globalization of trade relationships and on the other hand with the emerging of the single European currency. Thus, the euro-dollar relationship has become more important (Entorf (2000)).

5.3 Analysis of beta coefficients

The time series of beta coefficients for the long-term interest rate factor for the bank, utilities, and chemicals indices are presented in Figure 2. The beta coefficients for all indices show a similar trend. The bank index has almost always the highest sensitivity to the interest rate factor. This is an indication for the significance of long-term interest rates for banks. During the first half of the 1990s we observe an increase in the sensitivity for all three indices. Interestingly, this pattern changes completely in that the beta coefficients are insignificant after 1998. This result may be due to the fact that banks have changed their business model during the second half of the 1990s in that they first reduced the maturity transformation and second used derivatives to hedge their remaining risk exposure. One reason for this change in risk management is the higher bank equity standards resulting from Basle II. We also test for significant differences in the betas over time. A standard t-test for non-overlapping sub-periods for 60 months intervals reveals that in many

periods the beta coefficients exhibit a significant time variability for all three industries (banks 61%, utilities 28%, chemicals 62%).

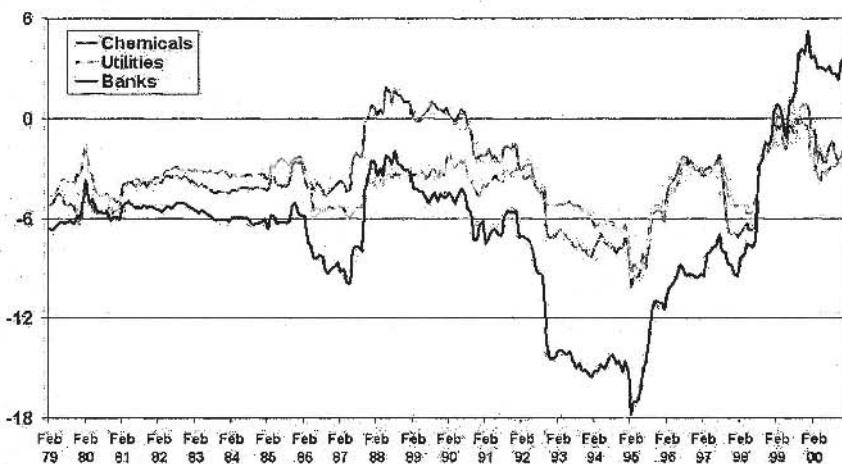


Fig. 2. Beta coefficients for long-term interest rates for the period 1979-2000

Moreover, we observe a significant increase of the sensitivity over time with respect to the exchange rate factor (for details see Bessler et al. (2003b)). Most of the beta coefficients are positive, i.e. an increase in the value of the US\$ results in positive stock returns. The increase in the sensitivity during the second half of the 1990s is quite interesting. One explanation for this observation is the increase in the likelihood that the European Monetary Union would be established. An increase in the probability of a successful introduction of the Euro resulted in a decrease of the foreign exchange rate risk for the other currencies in the Euro. This may have led to a stronger orientation towards the US\$. This result is supported in a subsequent study of the FF and Lira. For both currencies we find a significant impact only for the period of the EMU-crisis in 1992-93. In nearly all periods the other currencies lag the explanatory power of the US\$ (results are available on request).

6 Conclusion

The objective of this study was to analyze the importance of various macroeconomic factors in explaining the return structure for six German stock indices for the period from 1974 to 2000. Of special interest was to investigate whether the bank index shows a different behavior relative to industrial indices. A comparison of the results for banks and for five industrial indices reveals the greater sensitivity of the banks to changes in long-term interest

rates. Moreover, we find strong evidence that the relationship between interest rates and returns for banks has not been stable over time but time variant. This is especially evident in periods of high interest rates, i.e. usually an inverted yield curve. This relationship may be due to the positive maturity transformation of banks. One interesting result of our study is that the interest rate sensitivity of bank stock returns has significantly decreased towards the end of the 1990s. This decrease of the net exposure to interest rate risk can be explained with a reduction in maturity transformation and the use of derivatives by banks. Especially interesting is the relationship for the DM/US\$ since the middle of the 1990s. Not surprisingly, banks have the lowest exposure to exchange rate changes due to their well known hedging activities.

In sum, the empirical results confirm the time variability of the explanatory power and the beta coefficients for all indices. Thus, the impact of various macroeconomic factors in explaining stock returns is not stable but time variant. The order and the combination of macroeconomic factors that are employed in this study seem to be well suited to explain the stock returns especially for banks.

References

- BESSLER, W. (2001): Maximalbelastungstheorie und Zinsrisikomanagement. In: H. Schmidt, H. Ketzel, and St. Prigge (Eds.): *Moderne Konzepte fuer Finanzmaerkte, Beschaeftigung und Wirtschaftsverfassung*. Mohr Siebeck, Tuebingen, 15–48.
- BESSLER, W. and BOOTH, G.G. (1994): Interest Rate Sensitivity of Bank Stock Returns in a Universal Banking System. *Journal of International Financial Markets, Institutions and Money*, 3, 117–136.
- BESSLER, W. and MURTAGH, J.P. (2003a): *An International Study of the Risk Characteristics of Banks and Non-Banks*. WP, University of Giessen.
- BESSLER, W. and OPFER, H. (2003b): *Eine empirische Untersuchung zur Bedeutung makroökonomischer Einflussfaktoren auf Aktienrenditen am deutschen Kapitalmarkt*. WP, University of Giessen.
- CHEN, N.-F., ROLL, R., and ROSS, S.A. (1986): Economic Forces and the Stock Market. *Journal of Business*, 59, 383–403.
- ENTORF, H. (2000) Der deutsche Aktienmarkt, der Dollar und der Aussenhandel. *Zeitschrift fuer Betriebswirtschaft*, 70, 515–539.
- FLANNERY, M.J. and JAMES, C.M. (1984): The Effect of Interest Rate Changes on the Common Stock Returns of Financial Institutions. *Journal of Finance*, 39, 1141–1154.
- NEWHEY, W.K. and WEST, K.D. (1987): A Simple, Positive Semi-Definite, Heteroskedasticity and Autokorrelation Consistent Covariance Matrix. *Econometrica*, 55, 703–708.
- NOWAK, T. (1994): *Faktormodelle in der Kapitalmarktheorie*. Botermann & Botermann Verlag, Koeln.
- SAUER, A. (1994): *Faktormodelle und Bewertung am deutschen Aktienmarkt*. Fritz Knapp Verlag, Frankfurt am Main.

Application of Classification Methods to the Evaluation of Polish Insurance Companies

Marta Borda and Patrycja Kowalczyk-Lizak

Department of Financial Investments and Insurance,
Wroclaw University of Economics,
ul. Komandorska 118/120, 53-345 Wrocław, Poland

Abstract. The evaluation of the financial standing is in the interest of both the insurance companies and other groups of businesses operating in the insurance market. The paper presents the selection of ratios characterizing the financial standing of the insurance companies in the Polish market. The authors have applied the k -means method and Ward's method in order to classify the insurance companies according to their financial situation. The obtained results show the variation in the financial standing of the analyzed insurers and changes in this field over the last few years.

1 Introduction

In view of the specific nature of insurance activity, insurance companies are more exposed to insolvency risk than other enterprises. Insurance business is based on risk transfer from the customer to the insurer, and calculation of insurance premiums takes place before knowing the real cost of insurance cover (especially the amount of insurance claims or benefits). We can also point to a number of external factors, which increase the risk of insolvency for insurance companies. These are as follows (see Babbel and Santomero (1997), Black and Skipper (2000)):

- globalisation and liberalisation of insurance business,
- processes of concentration and consolidation in the insurance markets,
- intensive competition not only within the insurance sector, but also from other financial intermediaries (e.g. banks, investment funds, etc.),
- growth of risk in the financial markets,
- development of alternative risk transfer methods.

Thus, there exists a clear demand for the comprehensive evaluation of the financial standing of the insurance companies. From the insurance company's point of view, the evaluation of the financial condition enables the identification of risk appearing in various areas of the business, which consequently gives reasons for taking appropriate decisions concerning the management of risk. The evaluation of the financial standing of the insurance companies is also in the interest of various groups of businesses functioning in the insurance market or directly cooperating with the insurance sector, e.g. customers,

present and potential shareholders, the regulators, creditors, insurance agents and brokers.

In the Polish insurance market, the homogenous and comprehensive evaluation system of the financial standing of the insurance companies is still lacking. First of all, it is caused by the immaturity of the Polish insurance market and problems with fixing critical values of financial ratios. There are also difficulties in access to credible and up-to-date data. Only two insurance companies functioning in Poland are rated. They are TUiR Warta S.A. (BBBpi S&P - based on freely available data) and AIG Polska (AAA S&P which is the reflection of the financial power of the entire holding) (Jaworski (2002)). Along with the development of the Polish insurance market and further globalisation of the insurance business, the introduction of the evaluation system of the insurance companies will be inevitable. Research concerning the application of classification methods in the Polish insurance companies rating was previously performed by Ronka-Chmielowiec and Kuziak (1999), Jajuga et al. (2001). World research in the insurance companies' classification was carried out by e.g. Ambrose and Seward (1988), Klein (1992), Singh and Power (1992).

2 Selection of ratios characterizing the financial standing of the insurance companies in the Polish market

In the first stage of our study, the content-related selection of diagnostic variables was conducted. We used some ratios applied by the Polish Institution of Insurance Supervision in the financial analysis of insurance companies in the Polish market (see KNUiFE (2001)). The set of diagnostic variables consists of the most important measures from all areas of the insurer's financial operations. The stress is put on measures characterizing the level of solvency, profitability and activity costs of the insurance companies. Taking into account the essential differences in existing life insurance companies and property and casualty insurance companies, we separately conducted both the selection of ratios and the classification for each insurance section. In the case of life insurance companies the set of diagnostic variables is formed by the following ratios:

- 1) Debt ratio (X_1)

$$\frac{\text{Technical reserves (net of reinsurance)}}{\text{Equity}}$$

- 2) Return on sales (X_2)

$$\frac{\text{Net account}}{\text{Gross written premiums}}$$

3) Return on investment (X_3)

$$\frac{\text{Investment income} - \text{Investment expenses}}{\text{Total investments}}$$

4) Return on sales (X_4)

$$\frac{\text{Total investments}}{\text{Technical reserves (net of reinsurance)}}$$

5) Return on sales (X_5)

$$\frac{\text{Claims paid} +/- \text{Change in reserve for claims (net of reinsurance)}}{\text{Earned premiums (net of reinsurance)}}$$

6) Return on sales (X_6)

$$\frac{\text{Technical reserves (net of reinsurance)}}{\text{Written premiums (net of reinsurance)}}$$

The catalogue of ratios selected for the evaluation of the non-life insurance companies consists of five measures (marked by symbols Z_1, Z_2, \dots, Z_5). It is assumed, that the ratios Z_1, Z_2 and Z_3 are identical to the ratios X_1, X_4 and X_5 respectively, considered in the case of life insurance companies, and the other two ratios (Z_4 and Z_5) are as follows:

- Rate of claims reserves (Z_4)

$$\frac{\text{Gross claims reserves}}{\text{Gross earned premiums}}$$

- Premium retention ratio (Z_5)

$$\frac{\text{Written premiums (net of reinsurance)}}{\text{Gross written premiums}}$$

All the above-mentioned ratios are expressed in percentages, and their interpretation is presented in KNUiFE (2001). The dynamic development and consequential growing number of insurance companies is a characteristic feature of the Polish insurance market. For example, in 1999 the insurance business was run by 27 insurers in section I (life insurance) and 30 insurance companies in section II (property and casualty insurance). By 2001 35 life insurance companies and 35 non-life insurers functioned in the market (www.knuife.gov.pl). For the study, we chose insurance companies, which have operated in the Polish market since at least 1997 (according to the date of selling of the first insurance policy), and some of them have been there since the market's beginning. Moreover, the chosen insurance companies differ from each other in market share, structure of equity and their insurance portfolios. We calculated the selected financial ratios for 19 life insurance

companies and 14 property and casualty insurance companies based on the financial data between 1999 and 2001 taken from their financial statements. "Younger" insurance companies, owing to their deviating values of analysed ratios and the recurrent incompleteness of financial data or lack of access to them, were not included in the study.

3 Classification of the insurance companies according to the financial standing

In the next stage of the study, making use of the previously selected financial ratios, we clustered the insurance companies according to their financial situation. We applied the k -means method and Ward's method, both considered by Hartigan (1975, 1979), Ward (1963), Jajuga (1990). We assumed that the number of classes would be four, taking into account four basic groups into which the insurance ratings were divided. We chose exactly these methods, because they belong to different groups of classification methods. The k -means method is one of the iterative optimization methods and Ward's method represents the large group of the hierarchical clustering methods. In our research we also used other hierarchical methods but obtained very similar results. The classification results of life and non-life insurance companies between 1999 and 2001 are presented in the tables 1 and 2 respectively. Comparing the results of clustering of life insurance companies in 1999 we notice that both applied methods resulted in a similar classification of insurers. The only difference in the composition of each class is the movement of four objects from class O (by application of the k -means method) to class P (by application of Ward's method). In 2000 the higher variation of class content, depended on the used classification method, is noticeable. The insurance companies belonging to class M (by application of the k -means method) were classified into two classes (by application of Ward's method), while the insurance companies belonging to class N and P (k -means method) formed together class P (Ward's method). In 2001 the movement of four insurers (Allianz Życie, Ergo Hestia Życie, Gerling Życie and Winterthur Życie) from class O (by application of the k -means method) to class P (by application of Ward's method) took place.

Generally, class M contains the average-sized insurance companies characterized by the lowest (negative) return on sales and a very low level of technical reserves (net of reinsurance), both in relation to the equity and to the amount of assets allocated to cover them. Between 2000-2001 the improvement of ratios connected with technical reserves of these companies was observed, however the companies achieved a high level of claims ratio. Class N (with the exception of 2000 by application of Ward's method) is formed by specific insurance companies. Rejent Life is the only mutual insurance company, running a business in the field of life insurance in Poland, while PZU Życie still has a dominant, though no longer monopolistic position in the

market. Both insurance companies are characterized by non-negative return on sales, unfavourably indebtedness of equity and a high rate of technical reserves. In 2001 PZU Życie reduced their indebtedness of equity and thanks to that the company moved to class P, which includes all of the other large life insurance companies. Class O comprises the average-sized life insurance companies, which in most cases have existed for at least four years. Unfortunately, all insurance companies in this class obtained a negative return on sales in the analyzed period, however the values of the other ratios achieved acceptable levels. As it has already been mentioned, large life insurance companies (characterized by positive return on sales, high indebtedness of equity and high level of technical reserves) and also smaller insurers with a relatively stable financial situation belong to class P.

Comparing the classification results of property and casualty insurance companies, we see that both applied methods between 1999-2001 classified the insurers in a similar way. It's worth noticing that in 1999, both Commercial Union and AIG belonged to their own separate classes, depending on which method was applied. The probable reason was, that Commercial Union had a very high reserves cover ratio, which differed considerably from the value of this ratio calculated for other insurance companies. AIG in 1999 had a very low level of equity capital. This resulted in a low debt ratio 2, deviated both from the recommended upper level of 330% and the value of this ratio for other insurers. In 2000 two insurance companies - Compensa and Filar belonged to separate classes (by application of both methods). The first one, Compensa, had the lowest and deviated from the others value of debt ratio 2. The second one, Filar, had the lowest value of claims ratio among all the analyzed insurers. In 2001 Compensa, which had a very high and considerably different from others debt ratio 2 and the highest claims ratio, formed a separate class.

In order to verify the results of conducted classification the distance matrix and variances within each class and between classes were analyzed. In the most cases the similarity (measured by the Euclidean distance) of the insurance companies belonging to one class was bigger than the similarity of insurers belonging to different classes. We noticed that the variances within the classes were usually smaller than the variances between classes.

4 Conclusion

At this stage of the development of the Polish insurance market it is not easy to evaluate the insurance companies according to their financial standing. The variety of financial ratios and the lack of available and complete financial statements make it impossible to conduct a comprehensive study comprising all the insurance companies in the market. However, the classification results, presented in the paper, allow us to formulate some conclusions. In the case of life insurance companies the basic factors, which affected the clustering

	1999		2000		2001	
	k-means method	Ward's method	k-means method	Ward's method	k-means method	Ward's method
M	CompŻ FilarŻ	CompŻ FilarŻ	CompŻ FilarŻ InterŻ PolisaŻ PolonŻ	CompŻ FilarŻ	FiatŻ FilarŻ HerosL InterŻ PolonŻ	FiatŻ FilarŻ Heros InterŻ PolonŻ
N	PZU Ż RejentL	PZU Ż RejentL	PZU Ż RejentL	InterŻ PolisaŻ PolonŻ	RejentL	RejentL
O	ErgoHŻ FiatŻ Finlife InterŻ RoyalŻ WinterŻ	FiatŻ InterŻ	AlliaŻ ErgoHŻ FiatŻ Finlife GerlŻ HerosL RoyalŻ WartaV WinterŻ	AlliaŻ ErgoHŻ FiatŻ Finlife GerlŻ HerosL RoyalŻ WartaV WinterŻ	AlliaŻ CompŻ ErgoHŻ Finlife GerlŻ HerosL RoyalŻ RoyalŻ WinterŻ	CompŻ Finlife PolisaŻ RoyalŻ
P	AlliaŻ AmpliL ComUŻ GerlŻ HerosL ING-NN PolisaŻ PolonŻ WartaV	AlliaŻ AmpliL ComUŻ ErgoHŻ Finlife GerlŻ HerosL ING-NN PolisaŻ PolonŻ RoyalŻ WartaV WinterŻ	AmpliL ComUŻ ING-NN	AmpliL ComUŻ ING-NN PZUŻ RejentL	AmpliL ComUŻ ING-NN PZUŻ WartaV	AlliaŻ AmpliL ComUŻ ErgoHŻ GerlŻ ING-NN PZUŻ WartaV WinterŻ

where:

AlliaŻ - Allianz Życie,
 ComUŻ - Commercial Union Życie,
 ErgoHŻ - Ergo Hestia Życie,
 FilarŻ - Filar Życie,
 HerosL - Heros Life,
 ING-NN - ING Nationale-Nederlanden,
 PolonŻ - Polonia Życie,
 RejentL - Rejent Life,
 WartaV - Warta Vita,

AmpliL - Amplico Life,
 CompŻ - Compensa Życie,
 FiatŻ - Fiat Życie,
 GerlŻ - Gerling Życie,
 InterŻ - Inter Życie,
 PolisaŻ - Polisa Życie,
 PZUŻ - PZU Życie,
 RoyalŻ - Royal PBK Życie,
 WinterŻ - Winterthur Życie.

Table 1. The classification results of life insurance companies according to the financial standing between 1999 and 2001

	1999		2000		2001	
	<i>k</i> -means method	Ward's method	<i>k</i> -means method	Ward's method	<i>k</i> -means method	Ward's method
M	CommU	CommU	Comp	Comp	Comp	Comp
N	Allianz PZU Warta	AIG	Filar	Filar	Europa Warta	PZU Warta
O	AIG Comp ErgoH Gerling Heros Polonia	Allianz CIGNA Comp ErgoH Gerling Heros Polonia PZU Warta Winter	AIG Allianz CIGNA Europa Gerling Heros Polonia PZU Warta	AIG Allianz CIGNA Europa Gerling Heros Polonia PZU Warta	Allianz CIGNA ErgoH Filar Gerling Heros Polonia PZU	Allianz CIGNA ErgoH Filar Gerling Heros Polonia
P	CIGNA Europa Filar Winter	Europa Filar	CommU ErgoH Winter	CommU ErgoH Winter	AIG CommU Winter	AIG CommU Europa Winter

where:

CIGNA - CIGNA STU,
Comp - Compensa,
Winter - Winterthur,

CommU - Commercial Union,
ErgoH - Ergo Hestia,

Table 2. The classification results of property and casualty insurance companies according to the financial standing between 1999 and 2001

results, are the profitability (measured by the return on sales) and the financial security (measured by the level of technical reserves in relation to other financial values). In the case of property and casualty insurance companies the level of indebtedness of equity and the claims ratio largely decided on the results of the classification. There are only a few life- and non-life insurance companies with good and relatively stable financial condition in the Polish market. Most of analyzed insurers have improved their financial situation but the pace of changes is still too slow. We can anticipate that this trend will continue in the future and after integration into the European Union Polish insurance companies will be forced to rise their competitiveness and financial security level.

References

- AMBROSE, J. M. and SEWARD, A. (1988): Best's Ratings, Financial Ratios and Prior Probabilities in Insolvency Prediction. *Journal of Risk and Insurance*, 55, 2, 229–244.
- BABEL, D.F. and SANTOMERO, A.M. (1997): Financial Risk Management by Insurers: An Analysis of the Process. *Journal of Risk and Insurance*, 64, 231–270.
- BLACK, K. Jr. and SKIPPER, H.D. Jr. (2000): *Life & Health Insurance*, 13th ed., Prentice-Hall, Upper Saddle River, NJ.
- HARTIGAN, J.A. (1975): *Clustering Algorithms*. Wiley, New York.
- HARTIGAN, J.A. and WONG, M.A. (1979): A K -means Clustering Algorithm, Algorithm AS136. *Applied Statistics*, 28, 1, 100–108.
- JAJUGA, K. (1990): *Statistical pattern recognition*, (in Polish language). Warsaw: PWN.
- JAJUGA, K., KUZIAK, K., and WALESIAK, M. (2001): An attempt to the application of classification methods in insurance rating (in Polish language). *Taxonomy*, 8, Wrocław.
- JAWORSKI, W. (2002): *Insurance Rating* (in Polish language). Poznań: AE.
- KLEIN, R. (1992): *Insurance company rating agencies: A description of their methods and procedures*. NAIC, Kansas City.
- KNUiFE, Department of Analysis of the Polish Insurance System (2001): *Methodology of financial analysis of the insurance companies* (in Polish language). Warsaw.
- RONKA-CHMIELOWIEC, W. and KUZIAK, K. (1999): Clustering of Insurance Companies under Financial Condition, (in Polish language). *Taxonomy*, 6, Wrocław.
- SINGH, A.K. and POWER, M.L. (1992): The Effects of Best's Rating Changes on Insurance Companies Stock Prices. *Journal of Risk and Insurance*, 59, 310–317.
- WARD, J.H. (1963): Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58, 236–244.

Analytic Hierarchy Process – Applications in Banking

Czesław Domański¹ and Jarosław Kondrasiuk²

¹ Chair of Statistical Methods, University of Lódź,
ul. Rewolucji 1905, 41, 90-214 Lódź, Poland
czedoman@uni.lodz.pl

² LG Petro Bank S.A., Dept. of Economic Planning and Analyses,
ul. Rzgowska 34/36, 93-172 Lódź, Poland
jaroslaw.kondrasiuk@lgpetro.pl

Abstract. In our article we want to present Analytic Hierarchy Process (AHP) as support methodology for optimizing decision making processes. We will focus on making strategy decisions in a bank with applying both basic and adjusted AHP application models. Due to our research we will present four groups of banking application models. We also describe guidelines of Thomas L. Saaty's AHP methodology. The AHP provides an objective way for reaching an optimal decision for both individual and group decision makers with a limited level of inconsistency. It makes it possible to select the best alternative (under several criteria) from a number of alternatives through carrying out pairwise comparison judgements. Overall priorities for ranking the alternatives are being calculated on the basis of pairwise comparisons.

1 The Analytic Hierarchy Process

The AHP is a multicriteria decision support method created by Thomas L. Saaty (Saaty (1986)). It provides an objective way for reaching an optimal decision for both individual and group decision makers. The AHP is designed to select the best from a number of alternatives evaluated with respect to several criteria by carrying out pairwise comparison judgements with using some measures for limiting inconsistency in judgements (unavoidable in practice). In this method the judgements are used to develop overall priorities for ranking the alternatives.

There are four main stages in the AHP process:

1. Hierarchy model building - the basic AHP model consists of three levels: goal, criteria level and alternatives. It is possible to add as many as necessary levels of subcriteria depending on the complexity of the problem.
2. Identification of decision makers preferences - in the AHP method it is being performed by collecting information about pairwise judgements due to a goal (for criteria), a specified criterion (for alternatives or subcriteria) or a subcriterion (for alternatives).
3. Synthesis - obtained by a process of weighting and adding down the hierarchy leading to multilinear form in two possible modes:

- the distributive mode in which the principal eigenvector is normalized to yield a unique estimate of ratio scale underlying the judgements;
 - the ideal mode in which the normalized values of alternatives for each criterion are divided by the value of the highest rate alternative.
4. Sensitivity analysis giving an answer to a question whether the alternative chosen as the best would be changed in case of modifying criteria/subcriteria preferences.

The key to the problem is to identify decision maker/makers preferences by collecting information about pairwise judgements. It is being performed separately due to a goal (for criteria), each criterion (for alternatives or sub-criteria) and all subcriteria (for alternatives). We present one of the possible scales of converting the intensity of judgements into numeric form in Table 2. Having a set of information we have built a matrix of ratio comparison for a given goal/criterion. A matrix \mathbf{A} (matrix of ratio comparison) might be converted into the vector of priorities \mathbf{w} in various ways. However, the need of consistency make us to choose the eigenvalue formulation $\mathbf{Aw} = n\mathbf{w}$. Under assumption that the priorities $\mathbf{w} = (w_1, \dots, w_n)^T$ with respect to a single criterion are known - we can examine what we have to do to recover them. Having the matrix \mathbf{A} :

$$\mathbf{A} = \begin{bmatrix} \frac{w_1}{w_1} & \frac{w_1}{w_2} & \dots & \frac{w_1}{w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{w_n}{w_1} & \frac{w_n}{w_2} & \dots & \frac{w_n}{w_n} \end{bmatrix}$$

we multiply it on the right by \mathbf{w}

$$\mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

to obtain $n\mathbf{w}$.

Each element a_{ij} of the matrix of ratio comparison describes the importance of alternative i over alternative j . In order to guarantee the consistency of the judgements, relevant groups of the matrix elements have to follow the equation: $a_{ij} \cdot a_{jk} = a_{ik}$. When we do not have a scale at all, or do not have it conveniently as in the case of some measuring devices - we can only estimate of w_i/w_j . It leads to the problem:

$$\mathbf{A}'\mathbf{w}' = \lambda_{max}\mathbf{w}'$$

where λ_{max} is the principal eigenvalue of $\mathbf{A}' = (a'_{ij})$ the perturbed value $\mathbf{A} = (a_{ij})$ with the reciprocal $a'_{ji} = 1/a'_{ij}$ forced.

The solution is obtained by raising the matrix to sufficient large power - then summing over the rows and normalizing to obtain the priority vector $\mathbf{w}' = (w'_1, \dots, w'_n)^T$. This process is stopped when the difference between

components of the priority vector obtained at k -th power and at the $(k+1)$ -st power is less than some predetermined small value. The vector of priorities is the derived scale associated with the matrix of comparisons. The value zero in this scale is assigned to an element that is not comparable to the elements considered.

With the eigenvector for $n \leq 3$ normalizing the geometric means of the rows leads to an approximation of the priorities. In all the cases it is possible to get an approximation by normalizing the elements of each column of the judgement matrix and then averaging over each row. However, it is important to remember that such steps can lead to rank reversal (in spite of closeness of the eigenvector solution).

Having the exact value of \mathbf{w}' in normalized form we can obtain the exact value (or an estimate) of λ_{max} by adding the columns of \mathbf{A}' and multiplying the resulting vector by the priority vector \mathbf{w} .

After obtaining the principal eigenvector estimate \mathbf{w} we have to consider the question of consistency. The fact that the original matrix \mathbf{A} need not to be transitive (for example A_1 may be preferred to A_2 and A_2 to A_3 but A_3 may be preferred to A_1) arises the problem. The solution is the consistency index (*C.I.*) of a matrix of comparison defined as:

$$C.I. = \frac{\lambda_{max} - n}{n - 1}.$$

The consistency ratio (*C.R.*) is obtained by comparing the *C.I.* with the appropriate one of the following set of numbers (Table 1) each of which is an average random consistency index derived from a sample of randomly generated reciprocal matrices. The study of the problem and revision of the judgements should be completed if $\frac{C.I.}{R.I.} \geq 0.10$.

N	1	2	3	4	5	6	7	8	9	10
Random Consistency Index	0.00	0.00	0.52	0.89	1.11	1.25	1.35	1.40	1.45	1.49

Table 1. Average Random Consistency Index *R.I.* published in Saaty (1986)

The above described solution to the problem is consider to be classical Saaty solution (Saaty (1994) and Domański and Kondrasik (2000 and 2002)) and is used for reaching both local and global vectors of priorities - necessary for synthesis (either distributive or ideal mode).

Sensitivity analysis is the final step that gives an answer to a question whether the alternative chosen as the best would be changed in case of modifying criteria/subcriteria preferences.

Intensity of Importance	Definition	Explanation
1	Equal importance	Two activities contribute equally to the object
2	Weak	
3	Moderate importance	Experience and judgement slightly favour one activity over another
4	Moderate plus	
5	Strong importance	Experience and judgement strongly favour one activity over another
6	Strong plus	
7	Very strong or demonstrated importance	An activity is favoured very strongly over another; its dominance demonstrate in practice
8	Very, very strong	
9	Extreme importance	The evidence favouring one activity over another is of the highest possible order of affirmation

Table 2. The Fundamental Scale published in Saaty and Vargas (1994)

2 Applications of the AHP

2.1 Establishing banking rates

After occurring of a market destabilization factor we should build two AHP models. The first model will lead us to an optimal solution for changing the average deposit rate (chapter 2.1.1) and the second one will be an answer to the question how to change the base loan rate (chapter 2.1.2). The models may be used individually or together. The general idea of using them follows the steps:

1. Occurring of the market destabilization factor (e.g. a change of base interest rate/rates).
2. Actualization of preference matrices for both models (including checking the inconsistency of preferences and limiting them according to the AHP method).
3. Solving models.
4. Sensitivity analyses.
5. A final decision based on optimal solutions reached in previous step concerning deposit and loan products of the bank (Domański and Kondrasiuk (1999)).

2.1.1 Establishing the price of the bank deposits

Following AHP methodology we have structured the AHP model with four criteria:

- COMPETITION - marketing point of view on pricing deposits according to deposit rates of competitive banks;
- MARKET - treasury point of view, including possible acquiring bank deposits (and alternative costs);
- PLAN - financial planning and forecasting future benefits and costs of the bank;
- PORTFOLIO - presents assets portfolio of the bank as the measure of efficiency of the already acquired deposits.

Due to simplification of the model, we have decided to limit possible alternatives to changes of the average deposit rate from increasing to decreasing the rate by 1.00 % with 0.25 % step (Domański and Kondrasiuk (2000)).

2.1.2 Establishing the base loan rate of the bank

The base structure uses the following criteria:

- COMPETITION - loan rates of competitive banks;
- DEMAND - present and possible market share;
- DEPOSITS - the source of the money converting into loans;
- INTERBANK MONEY MARKET - as an alternative source of the money converting into loans.

The possible alternatives are also limited from increasing to decreasing the base loan rate by 1.00 % with 0.25 % step (Domański and Kondrasiuk (2000)).

2.2 Marketing strategy decisions

In this application, our problem is a necessity of choosing a new marketing strategy.

Due to the AHP methodology we have structured the AHP hierarchy presented in Figure 1. Our modification of the original AHP methodology is the level of criteria. An organisation part of a bank is a criterion in this case. The pairwise preferences due to a specified criterion in this approach are visualizations of generalized preferences between possible alternatives of a department (represented by its director) or a committee (achieved by internal negotiations or even voting). The consequence of this methodological adjustment is a different meaning of pairwise comparison due to a goal - representing our main decision maker (for example: a president of the Management Board) pairwise preferences between suggestions made by departments and committees. It is

crucial for the success of the whole project to assure the full confidentiality of these preferences or even allow the decision maker to finish the model (with proper software support and internal training). Lack of a proper protection of the final decision maker's preferences might discourage him from using this adjusted AHP methodology for making strategic decisions. The final version of the model used for choosing new marketing strategy consists of four AHP criteria:

- Management Board - in case when the main decision maker is a president this criterion describes aggregated pairwise preferences of the other members of Management Board (instead of this criterion it is possible to get separate criteria for each Management Board member);
- ALCO - Assets and Liabilities Management Committee provides an opinion on bank's needs from a liquidity point of view (in general: an answer to the question whether the bank needs to increase a pace of growth of loans or deposits);
- Marketing Department - based on in-depth analysis of bank's marketing weaknesses and opportunities pairwise summarization of Departments preferences;
- Planning Department - a combined efficiency, available cost budget and development targets based presentation of preferences.

In the presented model possible alternatives have been simplified to:

- General Image - focusing in media on brand name promotion;
- Deposits - marketing of liability products;
- Loans - marketing of asset products;
- Services - promotion of other products like banking cards, bank assurance products etc.

Our model is adjusted to a bank with a centralized system of decision making. In such an organisation the most important thing is pairwise comparison of criteria level - personally made by the bank's president. After reaching an optimal solution according to the AHP method - sensitivity analysis provides the final decision maker with a possibility to reconsider his pairwise preferences.

We shall not forget that the presented model is very general - giving only an idea of marketing policy goals. In practice, alternatives will be more specific due to specific product categories, target group of clients, media selection, timing, geographic area selection etc. (Domański and Kondrasiuk (2002)). In fact, the presented approach might support the decision making process of any crucial decision for the bank - with a unique AHP model adjusted to the problem complexity and a corporate culture.

2.3 Merger related decision

In this chapter we will describe a few hierarchy models which can be used for supporting merger related decisions - a case of merging two banks. We

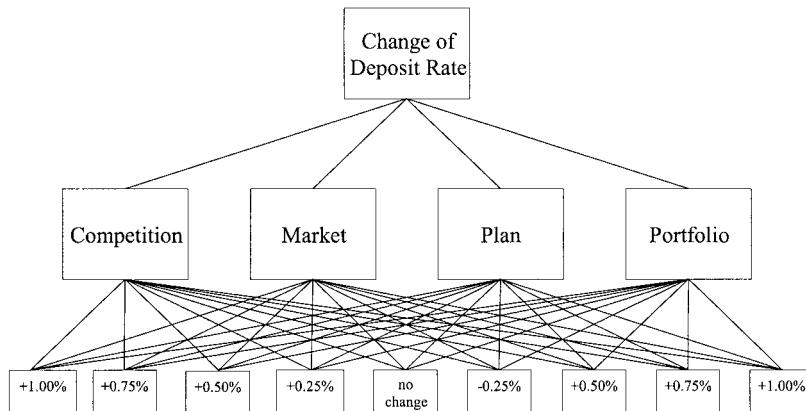


Fig. 1. The three level hierarchy used for selecting marketing strategy of the bank (Domański and Kondrasiuk (2002)).

do not consider the problem of a merger decision - whether to merger or not, we only focus on the supporting decision how to do it in optimal way. Furthermore, our models are adjusted to the merger of two banks being parts (daughter-companies) of other (much bigger) financial institutions.

2.3.1 The Head Office localization

The problem is where the Head Office of the new merged bank should be placed. In this case the model uses four criteria:

- EFFICIENCY - which alternative would be better considering business potential of a new bank (rapid growth of customers number, branches network to be served, etc.);
- LIQUIDATION COSTS - which part of the merged bank would be cheaper to write down;
- ORGANISATIONAL STRUCTURE - how close each of the alternatives is to the preferred (by mother company) business model;
- CUSTOMER RELATIONS IMPACT - danger of discouraging present customers (e.g. municipalities) closely related to "their own" - local Head Office;

and three alternatives:

- Head Office of Bank A;
- Head Office of Bank B;
- Mixed - some responsibilities placed in the previous Head Office of Bank A, the rest of responsibilities placed in the previous Head Office of Bank B.

We have to remember that this problem appears only when two separate cities are involved. In other circumstances, the decision making process is a pure economic calculation of rental costs or costs of already owned buildings with choosing the cheapest solution.

2.3.2 The Head Office departments evaluation

A model for choosing the best parts of an organisation is used for each pair of departments from both banks. Depending on the elasticity of the business model implied, it is possible to consider two solutions instead - when a specific responsibility field in one bank is spread among two departments in one bank and focused in one department of the second bank. The base structure uses five criteria:

- PRESENT COMPETENCY EVALUATION - value added to the company, fulfilling bank/corporate culture needs;
- TARGET COMPETENCE AREA - possible future fulfilling of the new bank needs;
- COST EFFICIENCY - in case of profit centres it might be profitability ratio;
- EMPLOYEE POTENTIAL - human resources evaluation: knowledge, experience, etc.;
- FLEXIBILITY - to unexpected needs of merger in view of past achievements.

In this model we have five alternatives:

- Department from Bank A [A1];
- Department from Bank B [A2];
- Merged department mostly based on Bank A department [A3];
- Merged department mostly based on Bank B department [A4];
- Completely new department [A5].

In fact, the alternative [A1] is almost equal to the alternative [A3] and the alternative [A2] is almost equal to [A4] making it possible to limit the problem to the three alternatives: [A3], [A4] and [A5].

2.3.3 IT system (implementation) choosing

The problem of IT system (implementation) choosing should be a technological - fact based decision. Unfortunately, experience shows other than objective reasons to be taken into consideration. The below presented selection of criteria is being limited to objective ones:

- EFFICIENCY - present efficiency of system implemented (for new one: theoretical);

- THEORETICAL AFTER MERGER EFFICIENCY - how the system would behave working for the merged bank (technological tests based));
- CURRENT COSTS - maintenance, development, licence and other costs;
- IMPLEMENTATION/ADJUSTMENT COSTS - costs of making necessary changes in a system to make possible fulfilling the merged bank needs (in case of a new system: full implementation costs);
- DEVELOPMENT POSSIBILITIES - flexibility to future system adjustments due to new products, growing business volumes, new law requirements (also in view of costs).

In this model we have four alternatives:

- System from Bank A;
- System from Bank B;
- Keeping both systems - one system for some operation (like cost administration or for General Ledger functions), second for the rest of operations;
- New system.

2.4 Human recourses decisions

The below presented models were prepared on the basis of banking cases but human resources problems are common to all organisations, therefore, there is no problem to use them in any company.

2.4.1 The Management Board member substitution

The problem is a sudden disappearance (due to resignation or other reasons) of one of the Management Board members. In a normal (not crisis) situation the AHP model would be very similar to a model from chapter 2.4.2, but in such circumstances a time limitation makes possible only to use internal resources. Of course we have to keep in mind legal procedures like final acceptance of the bank president's choice by shareholders through the Supervisory Board (including in some cases external bodies confirmation). In this case the model might use the following criteria of the bank president (a final decision maker):

- EXPERIENCE;
- FLEXIBILITY;
- PRESENT PLACEMENT IN ORGANISATIONAL STRUCTURE - appointing somebody from the present managers should not disturb the present field of responsibility of such a person (solving one problem by creating a bigger one);
- SHAREHOLDERS RELATIONS - in general: theoretical possibility of acceptance by the Supervisory Board.

2.4.2 Key employee selection

In this case, the problem is choosing the best candidate for a key post (top managers, a unique specialist, etc.). Depending on the post and organisational structure, a final decision maker will be either the bank president or a direct manager.

We can use as criteria:

- PAST ACHIEVEMENTS [C1];
- PAST EXPERIENCE & KNOWLEDGE VERSUS POST REQUIREMENTS [C2];
- HUMAN RESOURCES - IQ, EQ and other necessary evaluations [C3];
- DIRECT IMPRESSION - by an interviewer (depending on the number of people present during the interview and interview levels we can have several such criteria) [C4...n].

Depending on the approach (open: both external and internal or limited: internal) the additional steps might be applied:

1. Pre-selection of candidates - due to criteria [C1] and [C2].
2. Preliminary selection - due to a criterion [C3]
3. Secondary selection - due to criteria [C4...k] where k means all interviewers involved till final level (in case of one level interviews: $k = n$).

The described steps should lead to a very limited number of candidates (due to the AHP methodology they will be the alternatives), e.g. three or five. Having alternatives we solve the model in normal way.

3 Summary

The described solutions to the decision making problems in banking are only examples of the AHP methodology elasticity. Applying the AHP method might as well lead to limiting internal conflicts concerning problems in which different groups (even alliances) are involved as clearing (restructuring) a decision path under not fully consistent preferences. Other advantages of the AHP applications in banking are:

- speeding up decision making process for crucial decisions,
- enlarging consistence of decision makers preferences,
- replacing intuitive decision systems with objective (measurable) ones.

The AHP applications should not be used in decision making problems with full financial information available, like choosing the cheapest tender for paper supply (or other administrative cost items). This limitation to the AHP is weakened by one factor: most of the decisions involve some preferences - making new applications possible.

Thanks to its high flexibility, this method is adjustable to any corporate culture and internal limitations of the final decision maker/makers (e.g. a management board or board of directors) - by building multi-level models or a system of several models.

References

- DOMAŃSKI, C. and KONDRAŚIUK, J. (1999): Analytic Hierarchy Process in Banking. *Bulletin of the International Statistical Institute*, T. LVIII, B. 1, Helsinki.
- DOMAŃSKI, C. and KONDRAŚIUK, J. (2002): AHP as Support for Strategy Decision Making in Banking. In: K. Jajuga, A. Sokołowski, and H.-H. Bock (Ed.): *Classification, Clustering, and Data Analysis*, Springer, Heidelberg-Berlin.
- DOMAŃSKI, C. and KONDRAŚIUK, J. (2003): Establishing the Bank Rates with Using of Statistical Methods, *Acta Universitatis Lodzienensis - Folia Oeconomica*, 164, WUL, Lodz.
- DOMAŃSKI, C. and KONDRAŚIUK, J. (2000): Implementing of Analytic Hierarchy Process in Banking. *Acta Universitatis Lodzienensis - Folia Oeconomica*, 152, WUL, Lodz.
- DOMAŃSKI, C., KONDRAŚIUK, J., and MORAWSKA, I. (1997): Zastosowanie analitycznego procesu hierarchicznego w przedsiębiorstwie. In: T. Trzaskalik (Ed.): *Zastosowania Badan Operacyjnych*, Absolwent, Lodz.
- SAATY, T.L. (1986): Axiomatic Foundation of the Analytic Hierarchy Process. *Management Science*, 32, 7, 841–855.
- SAATY, T.L. (1994): *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*, Vol. VI, RWS Publications, Pittsburgh.
- SAATY, T.L. and VARGAS, L.G. (1994) : *Decision Making in Economic, Political, Social and Technological Environments with the Analytic Hierarchy Process*, Vol. VII, RWS Publications, Pittsburgh.

Tail Dependence in Multivariate Data – Review of Some Problems

Krzysztof Jajuga

Department of Financial Investments and Insurance,
Wroclaw University of Economics,
ul. Komandorska 118/120, 53-345 Wrocław, Poland

Abstract. Tail dependence is understood as the dependence between the variables assuming that these variables take the values from the tails of univariate distributions. In the paper two approaches of tail dependence determination are discussed: conditional correlation coefficient and tail dependence coefficients. It can be argued that both approaches are the generalizations of the well-known univariate approach, based on conditional excess distribution. In the paper the proposal is also given to extend tail dependence coefficients to the general multivariate case and to represent these coefficients through copula function.

1 Introduction

Classical statistical analysis is usually concentrated on typical observations, coming from the “center” (“core”) of the distribution of a random variable. However, in practice there is growing interest in the outlying observations, coming from the tail of the distribution. Such observations reflect the existence of “rare events”, that is the events for which the probability of the occurrence is very small. The analysis of outlying observations is very often called tail analysis. It is particularly important for the analysis of risk, including financial risk. Here the observations coming from the tail correspond to large losses.

Different methods can be used to perform the analysis of outlying observations in the univariate case. One of the most common methods is based on the so-called Extreme Value Theory. Here two different approaches are:

- the analysis of the distribution of the maximum (or minimum); here as the limiting distribution one uses Generalized Extreme Value Distribution (Fréchet, Gumbel or Weibull distribution);
- the analysis of the conditional excess distribution, being distribution in the tail, defined as the distribution of the variable given that this variable exceeds some value; here as a good approximation one uses Generalized Pareto Distribution (Pareto, Pareto type II or exponential distribution).

An extended description of Extreme Value Theory is given for example in Embrechts et al. (1997).

In this paper we discuss the tail analysis for the multivariate case. We concentrate mostly on bivariate data, however at the end of the paper the general multivariate case data is being approached as well.

First of all, it should be mentioned that in the multivariate case there are two important issues in tail analysis, namely:

- analysis of values in the tail of univariate marginal distributions;
- analysis of tail dependence between two (or more) univariate distributions.

The classical approaches used in multivariate statistical analysis to cope with these problems are, for example: multivariate classification and clustering, analysis of mixtures of multivariate distributions, analysis of multivariate distributions with heavy tails (e.g. multivariate stable, multivariate t distribution). Of course, the first problem mentioned above (not discussed here) can be solved by performing a separate tail analysis for each considered variable.

In this paper we concentrate on the analysis of tail dependence - this is the dependence in the tails of the distributions. Apart from conducting the separate analysis of the tails of univariate distributions this can give an useful insight in the structure of multivariate data.

We discuss two different approaches in the analysis of tail dependence:

- conditional correlation coefficient;
- tail dependence coefficients.

As we will see, these approaches can be regarded as the generalizations of the univariate tail analysis.

2 Conditional correlation coefficient

We now turn to bivariate case. The presented approach is the modification of the classical Pearson correlation coefficient. We analyze correlation under condition that one or both variables take values in the tail (tails). This approach is based on the ideas presented by Malevergne and Sornette (2002). There are at least two possible models to define conditional correlation coefficient.

The first model is given as:

$$\rho_u = \frac{COV(X, Y|X > u)}{\sqrt{V(X|X > u)V(Y|X > u)}} \quad (1)$$

So here the underlying condition states that just one of two considered variables (here denoted by X) takes value from the tail.

Of course the value of the conditional correlation coefficient given by (1) depends on the underlying bivariate distribution of variables X and Y . If we consider bivariate normal distribution, standardized in such a way that

variance of variable X is equal to 1, only the asymptotic results (for large sample) are known. We have then:

$$\rho_u \rightarrow \frac{\rho}{\sqrt{1 - \rho^2}} \frac{1}{u} \quad (2)$$

So we see that in this case the conditional correlation coefficient depends (asymptotically) on unconditional correlation coefficient and on the threshold u . If the threshold approaches infinity then the conditional correlation coefficient goes to zero, so we have asymptotic independence.

If, on the other hand, we consider bivariate t distribution, we have the following asymptotic result:

$$\rho_u \rightarrow \frac{\rho}{\sqrt{\rho^2 + (\nu - 1)\sqrt{\frac{\nu-2}{\nu}(1 - \rho^2)}}} \quad (3)$$

So we see that in this case the conditional correlation coefficient depends (in limit) on unconditional correlation coefficient, the number of degrees of freedom and on the threshold u . If the threshold approaches infinity, then (except for the case of zero unconditional correlation coefficient) the conditional correlation coefficient goes to non-zero constant, so we have asymptotic dependence.

The second model to determine conditional correlation coefficient is given as:

$$\rho_u = \frac{COV(X, Y|X > u, Y > u)}{\sqrt{V(X|X > u, Y > u)V(Y|X > u, Y > u)}} \quad (4)$$

So here the underlying condition states that both the considered variables take values from the tails.

The value of the conditional correlation coefficient given by (4) depends on the underlying distribution. If we consider bivariate normal distribution, standardized in such a way that the variances of both variables are equal to 1, only the asymptotic results (given the sample increases) are known. We have then:

$$\rho_u \rightarrow \rho \frac{1 + \rho}{1 - \rho} \frac{1}{u^2} \quad (5)$$

So we see that in this case the conditional correlation coefficient depends (asymptotically) on unconditional correlation coefficient and on the threshold u . If the threshold approaches infinity then the conditional correlation coefficient goes to zero, so we have asymptotic independence. For bivariate normal distribution the second model does not bring a significant improvement over the first model.

The presented approach, based on the conditional correlation coefficient, can be understood as the generalization of the univariate approach, where one

analyzes conditional excess distribution (mentioned before). If one extends these results to more dimensions, one should speak about the conditional covariance matrix. However in this case there are no results available (yet).

3 Copula analysis - short introduction

Before we present the second approach, based on the so-called tail dependence coefficients, it is necessary to present the idea of copula analysis, which gives the framework for the calculation of these coefficients.

The main idea of copula analysis lies in the decomposition of the multivariate distribution on two components. The first component consists of the marginal distributions. The second component is the function linking these marginal distributions to give the multivariate distribution. This function reflects the structure of the dependence between the components of the multivariate random vector. Therefore the analysis of multivariate distribution function is conducted by “separating” univariate distributions from the dependence.

This idea is reflected in Sklar theorem (Sklar (1959)), given as:

$$F(x_1, \dots, x_n) = C(F_1(x_1), \dots, F_n(x_n)) \quad (6)$$

Where: F - the multivariate distribution function; F_i - the distribution function of the i -th marginal distribution; C - copula function.

Thus the multivariate distribution function is given as a function of the univariate (marginal) distribution functions. This function is called copula function and it reflects the dependence between the univariate components. Using (6) we have in the bivariate case:

$$P(X_1 \leq x_1, X_2 \leq x_2) = C(F_1(x_1), F_2(x_2)) \quad (7)$$

There is the other function, strictly related to copula function, called survival copula function, given as:

$$P(X_1 > x_1, X_2 > x_2) = \bar{C}(F_1(x_1), F_2(x_2)) \quad (8)$$

$$\bar{C}(u_1, u_2) = 1 - u_1 - u_2 + C(u_1, u_2) \quad (9)$$

The advantage of the approach based on copula function is that for given marginals by assuming different copula functions we obtain different data structures. Of course there are very many possible copula functions. For example one may use (in bivariate case):

- Normal (Gaussian) copula:

$$C(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right) dx dy \quad (10)$$

- t copula:

$$C(u_1, u_2) = \int_{-\infty}^{t^{-1}(u_1)} \int_{-\infty}^{t^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{x^2 - 2\rho xy + y^2}{p(1-\rho^2)}\right)^{-\frac{p+2}{2}} dx dy \quad (11)$$

- Gumbel copula:

$$C(u_1, u_2) = \exp\left(-\left((- \ln u_1)^\theta + (- \ln u_2)^\theta\right)^{\frac{1}{\theta}}\right) \quad (12)$$

A detailed description of copula functions is given in Nelsen (1999) and Joe (1997).

4 Tail dependence coefficients

The objective of tail dependence coefficients is to measure the dependence in tails of distribution. There are two such coefficients, defined as:

- Lower tail dependence coefficient:

$$\lambda_L = \lim_{u \rightarrow 0} P\left(Y \leq G^{-1}(u) \mid X \leq F^{-1}(u)\right) \quad (13)$$

- Upper tail dependence coefficient:

$$\lambda_U = \lim_{u \rightarrow 1} P\left(Y > G^{-1}(u) \mid X > F^{-1}(u)\right) \quad (14)$$

In the formulas (13) and (14) F and G denote the distribution functions.

The main idea behind tail dependence coefficient is based on the calculation of the probability that one variable takes value from the tail (lower or upper) given that the other variable takes value from the tail (lower or upper). The value in the tail is defined as a (lower or upper) quantile. It is worth to notice, however, that this probability as taken as limiting probability given one goes with the probability in the tail to 0 (lower tail dependence) or to 1 (upper tail dependence). It is worth to notice that the tail dependence coefficients are direct generalizations of the approach based on the conditional excess distribution.

Both tail dependence coefficients can take values from the interval [0;1]. If tail dependence coefficient is equal to 0, we call this asymptotic independence. If tail dependence coefficient is higher than 0, we call this asymptotic dependence.

The most important property of tail dependence coefficients is that they can be represented through copula functions. This is given in the following formulas:

$$\lambda_L = \lim_{u \rightarrow 0} \left[\frac{C(u, u)}{u} \right] \quad (15)$$

$$\lambda_U = \lim_{u \rightarrow 1} \left[\frac{1 - 2u + C(u, u)}{1 - u} \right] \quad (16)$$

It can be proved that:

- For normal copula we get the asymptotic tail independence (upper and lower) if the underlying correlation coefficient is different from 1;
- For t copula we get the asymptotic tail dependence (upper and lower) if the underlying correlation coefficient is different from -1;
- For Gumbel copula we have asymptotic lower tail independence and asymptotic upper tail dependence, since:

$$\begin{aligned} \lambda_L &= 0 \\ \lambda_u &= 2 - 2^{\frac{1}{\theta}}, \quad \theta > 1 \end{aligned}$$

There are some results known for other copula functions, they are given by Heffernan (2000).

Example.

As illustration of the tail dependence coefficients we present here a simple example based on real data coming from the financial market. We took into account the following financial time series:

- Two indices of Warsaw Stock Exchange: WIG (the index of most stocks traded on this exchange), WIG20 (the index of the 20 stocks of large capitalization);
- US market index: S&P500;
- UK market index: FTSE-100.

The financial time series of logarithmic rates of return come from the period January 2, 1995 – October 3, 2003. In addition, we studied the logarithmic rates of return for the following exchange rates: USD/PLN, EUR/PLN. Here period January 1, 1999 – October 3, 2003 was taken into account. Tail dependence coefficients were calculated for the following pairs: WIG20 and WIG, WIG20 and S&P500, WIG20 and FTSE-100, USD/PLN and EUR/PLN.

For each pair we use the formulas presented above and calculate the approximation of tail dependence by taking level of u (probability) close to 0 (for lower tail) or close to 1 (for upper tail). We present here the results of lower tail dependence coefficients in the case of Clayton copula. They are given in the Table 1.

Probability	WIG20 and WIG	WIG20 and S&P500	WIG20 and FTSE-100	USD/PLN and EUR/PLN
0.1	0.769	0.267	0.363	0.400
0.05	0.769	0.220	0.328	0.371
0.01	0.769	0.164	0.290	0.341
0.001	0.769	0.133	0.273	0.329

Table 1. Lower tail dependence coefficients (Clayton copula)

The results given in this table indicate that in all cases there is asymptotic lower tail dependence. This means that very low returns (that is very large losses) in one variable are associated with very low returns (very large losses) in the other variable. In the presented example this is particularly true for returns on two Warsaw Stock Exchange indices. The interesting conclusion is that the lower tail dependence between the indices of Warsaw Stock Exchange and London Stock Exchange is higher than between the indices of Warsaw Stock Exchange and New York Stock Exchange.

5 Extension to multivariate case

The presented tail dependence coefficients are defined for the bivariate case. It might be an interesting task, however, to consider more general case. Now we give simple proposals to extend tail dependence coefficients to multivariate case. First of all, it is worth to remind that tail dependence coefficient is understood as the conditional probability that one variable takes value from the tail given the other variable takes value from the tail. It is then natural approach to consider the probability that each of many, say, $n-1$, variables take values from the tails, given that the remaining, say, n -th variable takes value from the tail. Therefore we propose the following definition in trivariate case:

- Lower tail dependence coefficient:

$$\lambda_L = \lim_{u \rightarrow 0} P(Y \leq G^{-1}(u), Z \leq H^{-1}(u) \mid X \leq F^{-1}(u)) \quad (17)$$

- Upper tail dependence coefficient:

$$\lambda_U = \lim_{u \rightarrow 1} P(Y > G^{-1}(u), Z > H^{-1}(u) \mid X > F^{-1}(u)) \quad (18)$$

It can be proved that these tail dependence coefficients can be represented through copula functions and copula survival functions, using the following formulas:

$$\lambda_L = \lim_{u \rightarrow 0} \left[\frac{C(u, u, u)}{u} \right] \quad (19)$$

$$\lambda_U = \lim_{u \rightarrow 1} \left[\frac{\bar{C}(u, u, u)}{1 - u} \right] \quad (20)$$

Finally, we can write the tail dependence coefficients in general multivariate case, as:

- Lower tail dependence coefficient:

$$\lambda_L = \lim_{u \rightarrow 0} P(X_1 \leq F_1^{-1}(u), \dots, X_{n-1} \leq F_{n-1}^{-1}(u) \mid X_n \leq F_n^{-1}(u)) \quad (21)$$

- Upper tail dependence coefficient:

$$\lambda_u = \lim_{u \rightarrow 1} P(X_1 > F_1^{-1}(u), \dots, X_{n-1} > F_{n-1}^{-1}(u) \mid X_n > F_n^{-1}(u)) \quad (22)$$

Of course, similar representation through copula functions as in (19) and (20) can be derived. The properties of the tail dependence coefficients should be studied.

In the presented approach tail dependence coefficients were defined as conditional ones, given that one variable takes value from the tail. Of course, one can also consider the approach where conditioning is on more than one variable. For example, in trivariate case we can take limiting probability that one variable takes value from the tail given the other two variables take values from the tails. However, in this case, the representation of tail dependence coefficients through copula function is not so straightforward as in the proposed approach.

References

- EMBRECHTS, P., KLÜPPELBERG, C., and MIKOSCH, T. (1997): *Modelling Extremal Events for Insurance and Finance*. Springer–Verlag, Berlin.
- HEFFERNAN, J.E. (2000): A directory of coefficients of tail dependence. *Extremes*, 3, 279–290.
- JOE, H. (1997): *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.
- MALEVERGNE, Y. and SORNETTE, D. (2002): *Investigating extreme dependences: concepts and tools*, manuscript, www.gloriamundi.org.
- NELSEN, R.B. (1999): *An Introduction to Copulas*. Springer, New York.
- SKLAR, A. (1959): Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

The Stock Market Performance of German Family Firms

Jan Kuklinski¹, Felix Lowinski¹, Dirk Schiereck², and Peter Jaskiewicz²

¹ Institute for Mergers&Acquisitions, Universität Witten/Herdecke, D-58448 Witten, Germany

² Endowed Chair for Banking and Finance, European Business School Oestrich-Winkel, Schloss Reichartshausen, D-65375 Oestrich-Winkel, Germany

Abstract. This paper investigates the long-run performance of initial public offerings of 208 family firms floated in Germany between 1977 and 1998. Five years after the listing the market-adjusted return was on average -43.39% compared to a broad market index. However, in absolute terms, investors realized a positive return of 32.11% after 60 months.

1 Introduction

During recent decades an increasing number of companies chose an initial public offering (IPO) to raise new equity capital. As 395 German companies went public during the period of 1977 to 1998 the majority of these IPOs took place during the last 8 years, i.e. 1991 to 1998. Financial aspects seem to be important in almost every IPO (see Rydqvist/Hoegholm (1995), pp. 292). As predominantly companies associated with the typical German Mittelstand show decreasing equity ratios, this development could explain the high fraction of family-owned "middle market" companies going public.

So far, neither academics, professionals, nor jurisprudence were able to develop a precise definition of family firms. In addition, family firms can be organized in any possible legal form. Consequently, the percentage of equity ownership is often employed and seems to be a reasonable criterion for the purpose of an empirical study. But objections are raised that other factors are more important for the characterization as a family-owned company, e.g. family members who define the long term strategy, the founder's power and authority or the company's mentality and common values. Thus, the fractional equity ownership as a sole criterion appears insufficient and more individual examinations should be applied. Due to their specific governance structures, i.e. privately-held family firms and public corporations represent two extremes of the relation between ownership and control, this paper analyzes the capital market's reactions to family firm IPOs. Conclusions can be drawn, which influence changes in the governance structure of family firms have on firm value.

Chapter 2 presents a summary of previous empirical findings on post-IPO long-run performance. Chapter 3 describes the data sample and discusses the

methodology. Empirical findings are presented and discussed. The remaining Chapter summarizes the main findings and provides an outlook.

2 Previous empirical research

Numerous papers examine the short- and long-run performance of IPOs in different countries and for various periods of time. A majority of these studies documents a long-run underperformance of IPOs firms compared to some type of benchmark. Table 1 comprehensively summarizes the results of some landmark studies.

Country	Authors	Period	Data Sample	Months	Average Abnormal Return (%)
Germany	Ljungqvist (1997)	1970-1993	180	36	-12.11
Germany	Stehle/Ehrhardt (1999)	1960-1992	187	36	-5.0
Austria	Aussenegg (1997)	1984-1996	67	36	-56.9
UK	Espenlaub et al. (2000)	1985-1992	561	60	-73.95
USA	Ritter (1991)	1975-1984	1,526	36	-29.1
USA	Loughran (1993)	1967-1987	3,656	60	-33.3
USA	Loughran/Ritter (1995)	1970-1990	4,753	36	-26.9

Table 1. Empirical studies regarding the long-term performance of IPOs
Source: Adapted from Loughran et al. 1994.

The majority of former research documents a long-run underperformance of IPOs compared to some type of benchmark. Differences in the levels of underperformance can be interpreted as the result of different institutional settings in different countries (see Loughran et al. (1994)). The most recent strand of IPO research regards misspecifications of models to be responsible for the widely documented long-run underperformance. Emphasizing these model misspecification problems it is shown that IPOs do not perform statistically different from seasoned firms. The most common excuse formulated by the efficient market hypothesis is the failure to properly adjust for risk (see Shleifer (2000), p. 20). E.g. Brav (2000), Lyon et al. (1999), Ehrhardt (1999) and Zutter (2001) provide extensive discussions regarding the choice of methods and benchmarks.

Evidence on family-firm long-run performance is very scarce: Using a sample of 31 Austrian family-owned companies going public between 1984 and 1991, Aussenegg (1997) finds a statistically significant underperformance of -118.60% after 5 years. Analyzing a sample of 105 firms issuing shares between

1970 and 1991 Ehrhardt and Nowak (2001) compute a non-significant underperformance of -8.10% after 36 months. This result can be attributed primarily to the underperformance of issuers employing dual-class share structures, i.e. common stock and non-voting preferred stock. Whereas the performance of non-dual class share IPOs is neutral and not significantly different from zero, the underperformance of dual-class share issues is higher with -19.60% and statistically significant at the 5 percent level. Although similar findings are scarce, other types of empirical research, e.g. using accounting data, seem to confirm the long-run underperformance of family firms (e.g. Morck et al. (1998)).

3 Data sample and methodological approach

3.1 Data sample

The initial data sample includes 395 IPOs between 1977 and 1998. Companies were identified primarily by the DAI-Factbook 2001 issued by the Deutsches Aktieninstitut and by publications of Schürmann and Körfgen (1987, 1997). Additional information was obtained by various sources. Where possible, data was verified by multiple sources. Using the definition developed below, in a first step 208 IPOs were identified as flotations of family-owned firms. To achieve a homogenous data basis the initial sample had to meet the following criteria, partially suggested by Ritter (1991): (i) gross proceeds of 1 million Deutsche Mark or more, (ii) daily stock prices available from Karlsruher Kapitalmarktdatenbank (KKMDB), (iii) companies being taken public by an investment bank, (iv) no other outstanding shares of the firm being traded before the IPO and (v) the company being listed at the Geregelter Freiverkehr (until May 1988), Geregelter Markt (since May 1998), Amtlicher Handel or Neuer Markt (since 1997). Flotations at the Ungeregelter Freiverkehr (until May 1988), Freiverkehr (since 1988) and other over-the-counter (OTC) segments were excluded. The final number of companies in the sample totals 174. The final sample contains 3 companies whose return series are shorter than the 5-year event window. In order to prevent a survivorship bias, truncated return series are used, i.e. all firms in the final sample are included regardless of the length of their return series.

Since its introduction the regulated market (Geregelter Markt) was the preferred choice of most issuers. The mean size of the 174 IPOs is 16.3 million in nominal value and 144.3 million Deutsche Mark in market value, respectively. As these figures seem low, they are negatively biased by the smaller IPOs before 1983 and the flotations on the Neuer Markt in 1997 and 1998.

As discussed above, the main criterion to define a company as family-owned or as a family firm seems to be the percentage of family ownership. Parts of the literature mention a threshold of 50% plus one vote held by family members (see Klein (2000), p. 107). Mainly due to problems obtaining precise information about ownership structures before and immediately after

the IPO, the fractional equity ownership as a sole criterion appears insufficient. It is argued that the ownership variable should be accompanied by measures that account for the ability to exercise control and the influence on the strategic and operating development of the firm. Thus, the definition used in this paper is based on Klein (2000) and proposes ownership, leadership and control as the defining attributes. With a required family ownership fraction of at least 25% the formal definition of a family-owned company (FC) is as follows:

$$FC \cong \left(\frac{E_{fam(\min 25\%)}}{E_{total}} \right) + \left(\frac{Mgt_{fam}}{Mgt_{total}} \right) + \left(\frac{SB_{fam}}{SB_{total}} \right) \geq 1.$$

The first addend denotes the equity fraction held by family members (E_{fam}) compared to the company's total equity (E_{total}). The second addend describes the proportion of family members in the board of directors (Mgt_{fam}) to the overall number of members in the management board (Mgt_{total}). Finally, the third addend refers to the supervisory board and calculates the number of family members (SB_{fam}) in comparison to the total number of members (SB_{total}) in this committee.

This study does not employ a criterion with regard to the company's age. Although, in comparison to other capital markets, German companies are comparatively old with a mean (median) age of 55.89 (40) years when going public. The research design of long-run event studies addresses two major issues: the calculation of returns and the choice of benchmark. This paper computes buy-and-hold abnormal returns (BHARs) as this measure represents the most common market-adjusted measurement method (see e.g. Ritter (1991) or Brown/Warner (1980)). Due to the size and scope of this paper and its plausible character abnormal performance will be measured against the Deutscher Aktien Forschungsindex (DAFOX). A comparable proceeding can be found e.g. in Ljungqvist (1997).

3.2 Empirical results regarding the long-term performance

Table 2 summarizes the results. Compared to the market performance of 75.49%, the performance of family-owned firms is 32.11% at the end of the event window of 5 years. The difference represents a significant abnormal return of -43.39% (t-value 3.64). The results are in line with previous research conducted e.g. by Ehrhardt and Nowak (2001) or Ljungqvist (1997), documenting an underperformance of family-owned firms.

Among others, Ehrhardt and Nowak (2001) propose that the results could be biased by the inclusion of family firms listed at the Neuer Markt. In order to avoid this problem a sub-sample is formed which excludes IPOs in 1996 to 1998. The exclusion does not alter the results significantly.

In order to find possible explications for the observed patterns multivariate regressions are performed with 60-months BHARs as the depended

	1977-1998	1977-1995
Average BHR_{IPOs}	1.3211	1.3283
Standard deviation	1.3255	1.3458
Median	1.0601	1.0601
Average BHR Dafox	1.7549	1.7455
Wealth Relative	0.8425	0.8480
Average BHAR	-0.4339	-0.4172
Standard deviation	1.4405	1.4613
Median	0.3983	-0.3536
t-value	-3.6392	-3.3782
Min. BHAR	-3.4110	-3.4110
Max. BHAR	10.1516	10.1516
No. of pos. BHAR	43	42
No. of neg. BHAR	103	98
No. of IPOs (total)	146	140

Table 2. Long-term abnormal returns after 60 months.

variable. A listing on the Amtlicher Handel as the market segment with the highest listing requirements, consequently attracting candidates with more reputation than on other market segments, is considered as the first variable (AH). The issuance of preferred stock (VZ), the free float of common stock as well as the reputation of the investment bank are used as other variables. However, none of these variables offers a significant explanation. Both, the coefficient of determination R2 and the F-statistics indicate that the model has almost no explanatory power (see Tab. 3).

Regressor	Coefficient	t-statistics	p-value
Constant	-0.3957210	-0.8836040	0.3784
AH	-0.2089800	-0.82206920	0.4132
VZ	0.0813100	0.1864620	0.8524
Free float	-0.2737330	-0.2420040	0.8091
Underwriter	0.1273050	0.5045060	0.6147
<i>R</i> ²	0.98%		
Adj. <i>R</i> ²	0.00%		
F stat.	0.35		0.8469
Sample Size	143		

Table 3. Multivariate regression¹

¹ The formula for the calculation of the OLS-regression is: $BHAR60_i = \alpha_0 + \alpha_1 * AH_i + \alpha_2 * VZ_i + \alpha_3 * FREEFLOAT_i + \alpha_4 * UNDERWRITER_i + \varepsilon_i$. The abnormal return over the 60 months period represents the dependent variable and refers to family businesses that went public between 1977 and 1998. AH:

The quality of an underwriter cannot be viewed as an indicator for post-IPO performance, contrary to the hypothesis formulated in other studies. Ehrhardt and Nowak (2001), also reject the underwriter hypothesis as the majority of IPOs in their sample was taken public by the Deutsche Bank.

4 Summary and outlook

As part of a wave of IPOs on German stock exchanges, a growing number of small and medium-sized family-owned businesses have taken the challenge of going public. This paper investigates the long-run performance of IPOs of 174 family firms floated in Germany between 1977 and 1998. The fundamental change in ownership structure induced by the flotation represents a change in the governance of the firm as for the first time dispersed outsiders hold equity capital.

While earlier studies already pointed out that family-owned firms tend to underperform, the results indicate that this situation has worsened during the 1990s. Compared to the market, shareholders of family-owned businesses realized a negative abnormal return of 43.39%.

As for the expectations regarding future IPOs the results presented in this paper indicate that many firms do not achieve the expected lower cost of capital. This can be considered as a sign that German family businesses are only partially suited for the capital market. In any case, a in-depth analysis of a respective IPO candidate should be performed in order to make a more specific judgement.

References

- AUSSENEGGER, W. (1997): Die Performance österreichischer Initial Public Offerings. *Finanzmarkt und Portfolio Management*, 11(4), 413–431.
- BRAV, A. (2000): Inference in Long-Horizon Event Studies: A Bayesian Approach with Application to Initial Public Offerings. *Journal of Finance*, 55, 5, 1979–2016.
- BROWN, S. and WARNER, J. (1980): Measuring Security Price Performance. *Journal of Financial Economics*, 8, 205–258.
- EHRHARDT, O. (1997): *Börseneinführungen von Aktien am deutschen Kapitalmarkt*. Wiesbaden.

Dummy variable with the value of 1 if the shares are registered on the official market, otherwise the value is 0; VZ: Dummy variable, with the value of 1 if only preferred shares are traded at the stock exchange, otherwise it takes the value of 0; FREEFLOAT: percentage of dispersed common stock ownership in relation to the company's equity; UNDERWRITER: Dummy variable with the value of 1 if the Deutsche Bank, Dresdner Bank or Commerzbank are consortium banks or leaders at the IPO. In all others cases this variable is 0.

- EHRHARDT, O. and NOWAK, E. (2001): Private Benefits and Minority Shareholder Expropriation - Empirical Evidence from IPOs of German Family-Owned Firms. *Center of Financial Studies, Johann Wolfgang Goethe University, CFS-Working Paper No. 2001/10*. Frankfurt.
- ESPENLAUB, S., GREGORY, A., and TONKS, I. (2000): Re-assessing the Long-Term Underperformance of UK Initial Public Offerings. *Journal of Finance*, 32, 737–748.
- KLEIN, S. (2000): *Familienunternehmen - Theoretische und empirische Grundlagen*. Wiesbaden.
- LJUNGQVIST, A. (1997): Pricing initial public offerings: Further evidence from Germany. *European Economic Review*, 41, 1309–1320.
- LOUGHREN, T. (1993): NYSE vs. NASDAQ Returns: Market Microstructure or the Poor Performance of Initial Public Offerings? *Journal of Financial Economics*, 33, 2, 241–260.
- LOUGHREN, T. and RITTER, J. (1995): The New Issues Puzzle. *Journal of Finance*, 50, 23–51.
- LOUGHREN, T., RITTER, J., and RYDQVIST, K. (1994): Initial Public Offerings: International Insights. *Pacific Basin Finance Journal*, 2, 165–199.
- LYON, J., BARBER, B., and TSAI, C.-L. (1999): Improved Methods for Tests of Long-Run Abnormal Stock Returns. *The Journal of Finance*, 54, 165–201.
- MORCK, R., STRANGELAND, D., and YEUNG, B. (1998): Inherited Wealth, Corporate Control and Economic Growth: The Canadian Disease? *NBER Working Paper No. 6814*. Cambridge.
- RITTER, J. (1991): The long-run performance of initial public offerings. *Journal of Finance*, 1, 46, 3–27.
- RYDQVIST, K. and HOEGHOLM, K. (1995): Going Public in the 1980s: Evidence from Sweden. *European Financial Management*, 1, 287–315.
- SCHÜRMANN, W. and KÖRFGEN, K. (1987): *Familienunternehmen auf dem Weg zur Börse*. 2nd ed., München.
- SCHÜRMANN, W. and KÖRFGEN, K. (1997): *Familienunternehmen auf dem Weg zur Börse*. 3rd ed., München.
- STEHLE, R. and EHRHARDT, O. (1999): Renditen bei Börseneinführungen am deutschen Kapitalmarkt. *Zeitschrift für Betriebswirtschaft*, 69, 1395–1409.
- SHLEIFER, A. (2000): *Inefficient Markets: An Introduction to Behavioral Finance*. Oxford.
- ZINGALES, L. (1995): Insider Ownership and the Decision to Going Public. *Review of Economic Studies*, 62, 425–448.
- ZUTTER, C. (2001): The Long-run Consequences of Dual-Class IPO's: A Comparison of Dual- and Single-Class Long-run Performance, *Katz Graduate School of Business, Working Paper*. University of Pittsburgh.

Testing of Warrants Market Efficiency on the Warsaw Stock Exchange – Classical Approach

Agnieszka Majewska and Sebastian Majewski

Department of Insurance and Capital Markets,
University of Szczecin, Ul. Mickiewicza 64, 71-101 Szczecin, Poland

Abstract. The efficiency of different markets was a subject of research by plenty of analysts. Most market research on the derivatives market was concentrated on valuation, but only a little part was concentrated on market efficiency. The goal of this article is to provide an empirical test of efficiency of the warrants quoted on the Warsaw Stock Exchange. One of the approaches of the derivatives' market efficiency testing is researching a relationship between implied and historical volatility. The efficient market hypothesis assumes that volatility prediction, which is build on the sign from market, its named implied volatility, could be estimator of empirical volatility in the future, named historical volatility. Using standard procedures for estimating regression line by the OLS and for verification of econometric models, researchers could conclude about rejection or the lack of bases' disallowable the hypotheses' about market efficiency. The research includes testing weak and strong efficiency. There are two ways of testing the warrants market efficiency in the paper. The first approach includes empirical tests based on the following stages: 1. The calculation of underlying assets' historical volatility. 2. The estimation of implied volatility from warrants prices. 3. Verification of hypothesis. The second approach consists in comparing the actual warrants prices and the estimated prices generated from the Black-Scholes pricing model.

Introduction

This paper examines the efficiency of the Polish warrants market using transaction data of warrants quoted on the Warsaw Stock Exchange in the year 2002. We test efficiency by means of:

- comparison of theoretical prices (from Black and Scholes Option Pricing Model) and actual real prices of warrant;
- testing relationships between implied and historical volatility on the example of warrants written on shares of TPSA.

We use different methods of volatility's estimation for BSOPM pricing model in the first approach. Thus there are classical methods using standard deviation, exponential moving averages procedure and econometric models $ARCH(q)$ of the type. Existence of significant differences in warrant prices (theoretical and empirical) decides of lack of efficiency. We examine the relationship between implied and historical volatility on only one example because warrants on shares of TPSA constitute 20% of all quoted warrants and

27% of equity warrants in the analysed period. The second cause was the fact, that we had longer time series for TPSA warrants.

1 Polish warrants market

Warrants belong to instruments, that are very similar in their nature to options. J. Hull (1997) describes them as "options, which are constructed in a different way" because their writer could be a company or a financial institution. A warrant gives his owner rights, but its obligation is to buy or sell a particular underlying asset (share, bond) at a specified price at or before a specified date. Initially writers of warrants were only companies emitting shares or bonds, so this kind of warrants was named – equity warrants. Currently most warrants on the stock exchange are covered warrants, which can be emitted by famous financial institutions (banks). They are emitted in series and give a holder the right to get the difference between the strike price and the spot price of an underlying asset at a specified date. There are two kinds of warrants: call and put, as in the case of options. They could also be part of the three types: European type, American or Bermudian by the criteria of the expiration date. They are quoted at the stock exchanges and over the counter markets as well as they concern different kinds of assets. Except for shares and bonds, underlying assets could also be: share indexes, interest rates, currency and noble metals. Companies – sellers can turn the purses without necessity of payments of dividends and percentage in case of equity warrants.

Derivatives with asymmetrical risk have been traded on the Warsaw Stock Exchange since October 20th 1999¹. Empirical data in this paper were warrants, which were quoted on the Warsaw Stock Exchange from December 2001 to October 2002. They were the following securities:

- warrants on shares: Agora, Big, Budimex, Elektrim, KGHM, PKNORLEN, Prokom, Softbank, TPSA;
- warrants on futures on index WIG 20;
- warrants on indices: NIF, TECHWIG, WIG 20.

2 The warrants market efficiency

Financial Market Efficiency is very important for allocating capital and risk management. Efficiency could be understood as operational efficiency, allocating efficiency or pricing (informational) efficiency.

E. Fama (1970) defined a market efficiency hypothesis as situation, when all information, which has influence on securities prices, is reflected in the

¹ The first quotation of asymmetrical risk derivatives was on the Poznan Commodity Exchange in 1995 and they were call options.

actual price. It means this information is analysed by investors and has influence on prices of securities changes. Polish market efficiency was described many times by several authors, but research kept major attention on the share market.

One of the assumptions using Black and Scholes Options Pricing Model (BSOPM) is that there are no arbitrage conditions on the market, so it is a simple definition of market efficiency. Other assumptions are for example, the returns of underlying asset follow the log normal distribution; the value of returns is known and directly proportional to time; all information on the market is reflected in prices. Capelle-Blancard and Chaudhury (2001) examined French option market efficiency and showed that on large and established markets while efficiency occurs on small markets we could often meet lack of efficiency. The interesting fact is that arbitrageurs are not interested in doing business on small exchanges in spite of good conditions for arbitrage.

Testing of warrants market efficiency could be done by three approaches:

1. The first one is based on research on the comparison of two kinds of warrants' prices: theoretical and empirical (market price). This approach is one of two goals of research in this paper.
2. The second approach is research on the relationship between implied volatility and historical volatility. This hypothesis assumes that actual implied volatility is an unbiased estimator for future historical volatility (for example for a next trading day). This is the second goal of this paper. In such tests we are checking possibilities of having profit from prediction of volatility.
3. The third approach is a search for violations of non-arbitrage relationships among the prices (see Capelle-Blancard and Chaundhury (2001)). This kind of research could not be done for the Polish warrants market, because of the lack of short sell.

2.1 Comparison among real and theoretical warrants' prices

This part of efficiency tests is based on Black-Scholes pricing formula with six kinds of volatility used in valuation process. Classical BSOPM formula could be presented by equations: For a call option at t -th moment:

$$c = S \cdot N(d_1) - X \cdot e^{-r(T-t)} N(d_2)$$

where:

$$d_1 = \frac{\ln\left(\frac{S}{X}\right) + \left(r + \frac{\sigma^2}{2}\right) \cdot (T - t)}{\sigma\sqrt{T - t}}$$

$$d_2 = \frac{\ln\left(\frac{S}{X}\right) + \left(r - \frac{\sigma^2}{2}\right) \cdot (T - t)}{\sigma\sqrt{T - t}} = d_1 - \sigma\sqrt{T - t}.$$

Where:

S — the value of the underlying asset at moment t ,

X — strike price of option,

r — domestic risk-free interest rate,

T — time to option expiry,

s — volatility of the underlying asset,

$N(d)$ — normal cumulative distribution function of d_1 or d_2 , respectively.

With using put-call parity (Stoll (1969)) we get an equation for the European put option value:

$$p = X \cdot e^{-r(T-t)} N(-d_2) - S \cdot N(-d_1)$$

The six approaches used in the estimation of volatility as follows:

1. standard deviation;
2. exponential weight moving averages (EWMA) with $\lambda = 0.5$ (smoothing parameter);
3. EWMA with $\lambda = 0.7$;
4. EWMA with $\lambda = 0.9$;
5. EWMA with $\lambda = 0.95$;
6. the volatility from econometric model ($ARCH(q)$ type).

The first kind of volatility is a classical parameter taken from original BSOPM. Volatility from point 2 and 3 are taken form Alexander (1996), but 4 and 5 are taken from J.P. Morgan methodology RiskMetrics (1996). The last one type of volatility is taken from popular capital market methodology based on autoregression relationships among returns. Such spectrum of methods allows looking at different opportunities of warrants' prices valuations. We chose one autoregression model using maximum likelihood method with criteria of Akaike (Weron and Weron (1998)).

2.2 Relationships between implied volatility and historical volatility

This approach consists of two hypotheses: weak market efficiency and strong market efficiency. The weak market efficiency could be written as follows:

Let I mean implied volatility and H — historical volatility, so we could write hypothesis by equation:

$$H_t = \alpha + \beta \cdot I_{t-i} + \varepsilon_t$$

where:

H_t — historical volatility,

I_t — implied volatility,

α, β — parameters of linear regression function,

ε_t is random component with null hypothesis:

$H_0 : \alpha = 0$ and $\beta = 1$,

ε_t was from normal distribution and its values are independent of each other.

Strong market efficiency is given by the following hypothesis:

$$H_t = \alpha + \beta \cdot I_{t-i} + \gamma \cdot I_{t-i-1} + \varepsilon_t$$

with null hypothesis:

$H_0 : \alpha = 0, \beta = 1$ and $\gamma = 0$.

Such tests were done for each trading day in the analysed period (for Mondays, Tuesdays, Wednesdays, Thursdays and Fridays), because of auto-correlation the sixth regression line was estimated for non-overlapping data (see Dunis and Keller (1993)). Such a regression line comes from building new input data by averages for each working week. In this approach we could find time series, that are free from any seasonal effects.

3 Empirical tests

Tables 1–2 contain percentage errors of warrants' pricing using BSOPM formula. The actual prices of warrants are original data from quotations on the Warsaw Stock Exchange (where S — standard deviation; e.m. — econometric model: $ARCH(1)$, $ARCH(2)$ or $GARCH(1,1)$; None — means that any parameter wasn't significant; Ac — autocorrelation; Gamma — name of significant parameter; 1, 2, 3, 4, 5 and 6 — six approaches of estimating volatility).

Intuitively we assume that the warrants market on the Warsaw Stock Exchange is non-effective, because of the fact that Polish market has a short history of derivatives trading. First we examine prices: quoted on WSE and estimated (with BSOPM), assuming that if the average percentage difference among prices will be under 5% it will mean market efficiency. Second – the smallest differences should be near 0%, and third – the biggest differences could not be higher than 10%.

Only in a few cases (one contract on TPSA) percentage average errors were bellow 5%. Generally the average was on levels from 20% to 40%, although in case of warrants on TechWig index it reached 3 432%. This means that market prices were not independent of speculative activities on the market.

We could underline, that only two warrants are near the limit of 5% and they are: TPSL012BRE (average error 12.75%) and W20U130BRE (average error 10.5%) and only one of thirty five indicates market efficiency (average error 1.51%). Fifteen analysed valuations of warrants were very unstable. Their average errors were higher than 100%.

In the sense of warrants market efficiency we understand such situation rather as the lack of efficiency.

Unfortunately a third test for comparison of prices was a total failure. Only in two cases of two hundred ten, percentage errors of valuation were

SYMBOL OF A CONTRACT	S	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$	$\lambda = 0.95$	e.m.
AGOI054BDM	30.13	35.86	30.48	29.33	28.89	39.07
AGOK060BRE	35.05	39.66	54.73	60.99	69.21	16.11
AGOU054BDM	86.41	325.85	156.17	81.31	81.94	62.65
BDXU032BDM	21.80	54.53	47.20	24.95	23.65	8.97
BIGI004BDM	-	-	-	-	-	-
ELEI004BDM	153.88	-	24 468.18	2 589.16	1 029.64	569.91
ELEL004BRE	361.05	1 940.57	1 188.99	1 499.82	1 484.54	1 192.03
ELEX003BRE	13.18	23.58	24.96	21.49	17.72	17.36
KGHB013BRE	20.58	180.60	73.23	31.72	20.67	12.12
KGHK014BRE	58.86	431.42	100.21	50.08	43.35	25.96
KGHU015BDM	13.41	9.37	12.26	18.09	20.06	5.97
PKML130BRE	159.73	73.23	43.02	34.63	43.65	53.23
PKMU155BDM	8.90	22.09	16.99	14.58	16.49	17.62
PKNI019BRE	215.90	1 960.48	546.39	213.67	210.78	347.00
PKNU018BRE	20.56	37.49	27.10	25.52	30.31	29.44
PKNX017BRE	26.62	82.73	59.48	22.86	12.32	48.29
SFTB020BRE	84.08	171.54	64.53	52.80	59.48	265.03
SFTL018BRE	10.49	83.71	31.38	52.88	54.77	67.93
SFTU024BDM	15.33	46.23	40.85	40.16	34.83	15.11
TPSB013BRE	27.36	85.82	49.12	30.42	26.48	29.96
TPSI014BRE	146.74	-	1 629.755	116.502	116.35	147.77
TPSK015BRE	37.79	-	2 152.78	71.54	41.04	28.04
TPSL012BRE	7.66	26.85	15.75	9.29	8.07	8.64
TPSU013BRE	18.63	41.95	29.00	20.19	18.56	18.79
TPSU014BDM	1.20	2.23	1.81	1.32	1.17	1.3
TPSX011BRE	34.22	1 515.56	308.63	66.94	39.62	36.77
FW2L111BDM	12.94	128.71	75.34	3.34	10.00	11.93
NIFI055BRE	57.82	57.33	59.68	58.17	57.88	42.70
NIFL060BRE	204.12	279.41	258.06	231.92	220.34	80.78
TECL050BRE	249.27	3 432.87	865.47	374.46	285.99	97.89
W20L150BRE	262.45		786.92	84.23	124.27	64.86
W20U130BRE	9.65	11.05	10.71	12.16	13.55	6.13
W20X100BRE	207.31	523.96	20.10	167.26	137.47	116.00
W20X135BRE	21.72	21.38	20.10	21.95	24.66	12.73

Table 1. The average percentage differences among actual and theoretical prices resulting from BSOPM model with all types of volatility estimations (Source: own research)

near 10% and in other cases maximal error formed on the level of about 50%. We think that such results are not satisfying. The comparison of market prices and estimated prices showed that differences among prices are deep and create an opportunity for arbitrage.

SYMBOL OF A CONTRACT	S	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$	$\lambda = 0.95$	e.m.
AGOI054BDM	0.66	14.02	6.46	2.71	2.27	5.40
AGOK060BRE	34.37	32.94	22.39	24.54	58.19	5.34
AGOU054BDM	3.44	0.64	6.35	3.95	2.77	2.41
BDXU032BDM	3.31	6.00	4.52	3.48	3.37	5.79
BIGI004BDM	5.54	5 698.50		304.78	220.79	7.89
ELEI004BDM	20.95	24.41	24.68	25.87	34.44	29.59
ELEL004BRE	-	-	-	-	-	-
ELEX003BRE	8.08	6.84	10.40	8.64	8.71	9.17
KGHB013BRE	1.13	3.95	9.96	8.04	7.46	3.18
KGHK014BRE	2.49	9.84	9.20	0.07	1.63	1.11
KGHU015BDM	2.13	4.02	5.24	7.35	6.82	1.86
PKML130BRE	72.19	50.14	40.62	12.50	15.00	35.51
PKMU155BDM	1.45	1.19	2.54	1.82	1.40	8.26
PKNI019BRE	24.84	0.37	3.02	2.52	3.42	41.81
PKNU018BRE	3.17	0.86	3.34	2.40	0.78	3.97
PKNX017BRE	22.12	8.83	9.05	8.34	3.26	37.57
SFTB020BRE	9.94	55.33	11.09	19.00	43.88	34.13
SFTL018BRE	4.95	51.91	8.55	43.20	45.60	0.06
SFTU024BDM	2.48	2.48	2.48	2.45	2.40	2.48
TPSB013BRE	-	-	-	-	-	-
TPSI014BRE	0.18	15.09	7.53	2.33	3.73	1.05
TPSK015BRE	9.19	14.46	2.01	0.39	0.20	0.28
TPSL012BRE	0.92	5.35	1.95	0.10	0.94	0.36
TPSU013BRE	0.14	0.36	0.18	0.17	0.05	0.09
TPSU014BDM	-	-	-	-	-	-
TPSX011BRE	1.55	65.98	21.79	1.61	2.45	1.13
FW2L111BDM	7.93	107.20	73.55	0.29	5.59	5.15
NIFI055BRE	44.61	33.58	42.06	48.00	46.79	31.01
NIFL060BRE	110.54	96.49	102.81	110.23	109.59	57.52
TECL050BRE	114.01	974.75	755.34	160.11	121.35	49.96
W20L150BRE	13.47	40.75	44.22	13.36	63.41	12.19
W20U130BRE	0.27	0.24	0.26	0.05	0.09	0.00
W20X100BRE	-	-	-	-	-	-
W20X135BRE	6.84	0.12	0.29	2.84	4.57	0.26

Table 2. The minimal percentage differences among actual and theoretical prices resulting from BSOPM model with all types of volatility estimations (Source: own research)

In case of the warrant on TechWig index we could see very strong activities of speculating investors in quotations of new technology shares. This sector of the market, we are sure, is non efficient.

The best results we got for TPSA's warrants valuation, so we decided to use that data to examine market efficiency using the second approach. First

of all we had to decide, which relationships we would analyse. We built three econometric models of relationships between implied and historical volatility with different time-lags: one workday, two workdays and three workdays. For daily interval of data the most logical should be first time lag and the verification of model hypothesis confirmed that assumption. The econometric model with one workday time lag gave better parameters. It means that we assume: implied volatility today is an unbiased estimator of historical volatility tomorrow.

Only for one workday we got weak efficiency (Table 3), but strong test showed its weakness (Table 4). For averages we got the best results, independently of the estimation volatility method used. After weak efficiency tests we examined strong efficiency and just like a theory says the next time-lag of implied volatility did not have a significant effect for historical volatility. The worst results of estimation we obtained for volatility estimated from EWMA with $\lambda = 0.95$ (in almost all cases we got autocorrelation).

This seems to imply that volatility in our research was not a good estimator for future historical volatility. So we could say, probably the Polish warrants market is inefficient.

4 Conclusion

In this paper we examined the efficiency of warrants market quoted on Warsaw Stock Exchange from December 2001 to October 2002.

If BSOPM is a right model for the Polish warrants market we conclude as follows:

- We used two approaches to prove that small, developing markets could not be efficient. Almost all analysed warrants did not pass the test of weak efficiency. It could be understood as a fact that simple market strategies could give investor profits higher than the market average.
- In the first phase of estimating of econometric models, we used daily data, but we could not obtain effects because of autocorrelation. It means that among the workdays there are some relationships and it could be eliminated by means of several methods. One of the most effective method was average in a week.
- In most cases investors could use opportunities to arbitrage to get a profit. Such information could be ideal for speculating investors, but some conditions must be satisfied:
 - fluency of the market (about 72.4% of market trading value was concentrated in 5% of domestic companies²);
 - short sale possibility;
 - costs of transactions.

The lack of efficiency would be reflected in the valuation of derivatives using theoretical pricing models based on the log-normal distribution.

² FIBV

Mondays				
V	Model	R ²	DW statistic	Notice
1	$H_t = 0.028 - 0.016I_{t-1}$	0.247	1.17	None
2	$H_t = 0.034 - 0.303I_{t-1}$	0.278	2.36	None
3	$H_t = 0.035 - 0.272I_{t-1}$	0.339	2.089	None
4	$H_t = 0.032 - 0.114I_{t-1}$	0.220	1.142	None
5	$H_t = 0.029 - 0.046I_{t-1}$	0.102	0.756	Ac
6	$H_t = 0.028 - 0.018I_{t-1}$	0.145	2.182	None
Tuesdays				
1	$H_t = 0.028 - 0.004I_{t-1}$	0.012	0.985	None
2	$H_t = 0.024 + 0.039I_{t-1}$	0.003	1.263	None
3	$H_t = 0.028 - 0.051I_{t-1}$	0.009	1.399	None
4	$H_t = 0.030 - 0.080I_{t-1}$	0.090	1.149	None
5	$H_t = 0.029 - 0.060I_{t-1}$	0.117	0.867	Ac
6	$H_t = 0.028 - 0.001I_{t-1}$	0.000	1.402	None
Wednesdays				
1	$H_t = 0.028 + 0.001I_{t-1}$	0.002	2.330	None
2	$H_t = 0.033 - 0.261I_{t-1}$	0.019	0.994	Weak efficiency
3	$H_t = 0.031 - 0.162I_{t-1}$	0.017	0.993	None
4	$H_t = 0.030 - 0.080I_{t-1}$	0.085	1.149	None
5	$H_t = 0.029 - 0.001I_{t-1}$	0.000	1.249	None
6	$H_t = 0.029 - 0.023I_{t-1}$	0.009	1.095	None
Thursdays				
1	$H_t = 0.028 + 0.001I_{t-1}$	0.001	1.361	None
2	$H_t = 0.047 - 0.658I_{t-1}$	0.170	2.009	None
3	$H_t = 0.039 - 0.404I_{t-1}$	0.111	1.672	None
4	$H_t = 0.030 - 0.087I_{t-1}$	0.025	1.156	None
5	$H_t = 0.029 - 0.043I_{t-1}$	0.018	0.825	Ac
6	$H_t = 0.031 - 0.076I_{t-1}$	0.106	2.221	None
Fridays				
1	$H_t = 0.028 - 0.020I_{t-1}$	0.540	1.218	None
2	$H_t = 0.023 + 0.207I_{t-1}$	0.061	1.105	None
3	$H_t = 0.029 - 0.016I_{t-1}$	0.001	1.064	None
4	$H_t = 0.031 - 0.073I_{t-1}$	0.056	1.147	None
5	$H_t = 0.030 - 0.023I_{t-1}$	0.020	1.257	None
6	$H_t = 0.026 + 0.125I_{t-1}$	0.360	1.041	None
Averages				
1	$H_t = 0.054 - 1.030I_{t-1}$	0.013	1.302	Weak efficiency
2	$H_t = 0.056 - 0.122I_{t-1}$	0.654	2.320	None
3	$H_t = 0.054 - 0.107I_{t-1}$	0.668	2.098	None
4	$H_t = 0.034 - 0.257I_{t-1}$	0.311	0.742	Ac
5	$H_t = 0.041 - 0.522I_{t-1}$	0.472	1.152	None
6	$H_t = 0.031 - 0.113I_{t-1}$	0.529	2.515	None

Table 3. Testing weak efficiency for each workday (Source: own research)

V	Model	R^2	DW statistic	Notice
Wednesdays				
2	$H_t = 0.028 - 0.050I_{t-1} - 0.024I_{t-2}$	0.246	1.060	Gamma
Averages				
1	$H_t = 0.050 + 0.264I_{t-1} - 1.453I_{t-2}$	0.467	0.862	None

Table 4. Testing strong efficiency for Wednesdays and averages (Source: own research)

And at last if BSOPM is a weak formula of valuation of warrants on the WSE we conclude that we could not test market efficiency this way, because of the possibility of strongly misleading results. Thus we have to test the efficiency using other methods, for example models with stochastic volatility.

This paper could also be a first part of research of warrants market efficiency – the classical part. Thus next research using stochastic volatility models, could give an answer to the question if the Polish warrants market is efficient. Concluding, an aim of this paper could be reversed – if there are differences between prices of warrants, and if there is no significant relationship among implied and historical volatility, could we examine the Polish warrants efficiency?

References

- ALEXANDER, C. (1996): *Risk Management and Analysis*. John Wiley & Sons, London.
- BLACK, F. and SCHOLES, M. (1973): The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81, 637–659.
- CAPELLE-BLANCARD, G. and CHAUDHURY, M. (2001): *Efficiency test of the French Index (CAC 40) Option Market*, September 2001.
- CHRISTIANSEN, B.J. and PRABHALA, N.R. (1998): The relation between implied and realized volatility. *Journal of Financial Economics* 50, 125–150.
- DUNIS, Ch.L. (2001): *Forecasting financial markets. Exchange rates, interest rates and asset management*. ABC, Cracow, 252–276.
- JAJUGA, K. and JAJZGA, T. (1996): *Investment*. PWN, Warsaw.
- J.P. MORGAN/REUTERS (1996) *Risk MetricsTM - Technical Document*, New York.
- STOLL, H.R. (1969): The relationship between put and call option prices. *Journal of Finance*, 31, 319–32.
- TARCZYŃSKI, W. (1997): *The efficiency of activities of the Warsaw Stock Exchange*, Ekonomista 7.
- WERON, A. and WERON, R. (1998): *Financial Engineering*, WNT, Warsaw.

Group Opinion Structure: The Ideal Structures, their Relevance, and Effective Use

Jan W. Owsiński

Systems Research Institute,
Polish Academy of Sciences,
Newelska 6, 01-447 Warszawa, Poland
E-mail: owsinski@ibspan.waw.pl

Abstract. One is often interested in the “behind-the-scenes” of a group decision. This interest may refer to knowing whether the “vote” distribution’s mode coincides with the outcome, determining the structure of the set of opinions (any “blocks of votes”?), or finding the biggest subgroup of (relatively) consistent opinions. The potential uncovered structures may take the form of “ideal” or “perfect” structures, and their derivatives, which may be of a far broader significance than just for the group decision making. They may also shed light on the definitions of such basic notions as “consensus”. The paper presents several conditions to be fulfilled by such structures, in decreasing order of strength, and their properties, with a perspective on potential determination and applications. In addition, the conditions presented are “positive cluster definitions” of non-probabilistic character.

1 Introduction

We are often interested in knowing the “behind-the-scenes” of a group decision outcome, be it a simple voting exercise or a more complex procedure. This interest may itself be of quite differentiated nature: to know whether the “vote” distribution’s mode coincides with the outcome, to determine the structure of the set of opinions (are there any distinct “blocks of votes”?), or to find the biggest subgroup of (relatively) consistent opinions.

This interest is largely stirred by the known – and evident – cases, in which the opinion structure is well defined and the opinion groups are well separated. Such cases incline to look for the general features of similar “ideal” or “prefect” structures, with an important bearing for a much wider class of situations, including standard data analysis circumstances.

Thus, the paper deals with (i) formulation of conditions of “ideal” or “perfect” structures; (ii) characteristics of these structures, and relations between them; (iii) existing algorithmic procedures, which can be used to determine the “ideal” and “perfect” structures, perhaps with certain modifications; (iv) significance thereof for the broader data analysis domain.

It should be emphasized that the structures defined should correspond to the notion of “common opinion (decision)”, possibly referring to individual

groups (subsets of the decision makers), with emphasis on the idea of “consensus”, which may therefore be identified through a natural data structure rather than through an “external” criterion of whatever character.

The structures, and the conditions they satisfy, constitute, in fact, the “positive definitions” of clusters or of partitions. Although such positive definitions, if at all appearing, tend to refer to the probabilistic framework, the formulations here forwarded may eventually serve to bridge the gap between such definitions and the classical progressive merger procedures on the one hand, and the k-means-type approaches on the other.

2 Notations

We consider the group decision situation, with decision-makers (DMs) indexed i , $i \in I = \{1, \dots, n\}$, and options (“candidates”), being the object of decisions, indexed k , $k \in K = \{1, \dots, m\}$. The options are specified by the aspects (“criteria”) indexed g , $g \in G = \{1, \dots, f\}$. The subspaces \mathbf{G}_g of values of aspects are the subspaces of option descriptions, $x_k = \{x_{k1}, \dots, x_{kg}, \dots, x_{kf}\} \in \mathbf{G} = \mathbf{G}_1 \times \dots \times \mathbf{G}_f$. We denote the set of descriptions of the options considered by X_I , $X_I \subseteq \mathbf{G}$.

The DMs express their opinions through a function of the set K , $C_I(K)$. This function may take various shapes – a single element of the set K (“the best option”), an ordering (not necessarily complete) of the elements of K , a set (not necessarily complete nor consistent) of pair-wise comparisons of options $k, l \in K$, etc. For particular DMs the function $C_I(K)$ takes the values $C_i(K)$, $i \in I$. We will denote by $C(K)$ the aggregate opinion, and by $c_A(K, T)$ the procedure, leading to a “decision”. Note that although we start with interpretation from the domain of group decision making, and this will be the primary subject of our interest, the considerations herein have a definitely broader application, encompassing, in fact, the entire area of cluster-oriented data analysis.

3 The group opinion structure and its regularity

Our interest in group opinion structure results primarily from the conviction that the distribution of $C_i(K)$ standing behind the aggregate output (e.g. who was elected) is both far from decently unimodal and at the same time legibly regular. Thus, in particular, the apparently tendentious example of Fig. 1 actually occurs in thousands of court cases (“did the accused kill in justified self [somebody else’s] defence?”). We may cite here the recent case of the double killing at the streetcar stop in Warsaw, the clear opinion split concerning not just the judges, but the eyewitnesses themselves, some being convinced that the accused committed a cool-blooded double murder, others maintaining that he acted first in the defence of threatened streetcar passengers and then in his own defence. It could be inferred, on a superficial,

but intuitively obvious, level, that we deal here with a kind of group-wise consensus (“guilty” vs. “innocent”), with virtually no middle ground (if at all possible).

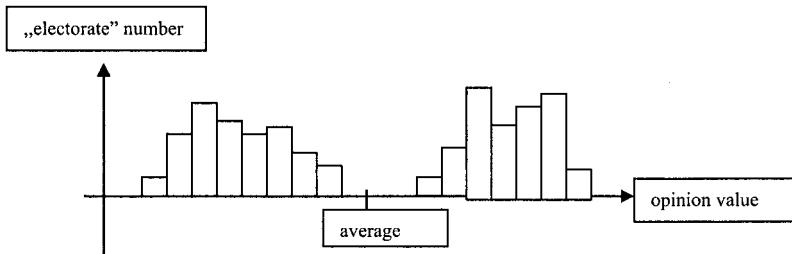


Fig. 1. Opinion structure divided about an abstract average: just a tendentious example?

A similar example is provided by the case analysed by the present author (see Owsíński and Zadrożny (2000)) concerning 96 votings in the Polish Diet, at the beginning of the 1993–1997 term. More than 400 MPs were accounted for, each characterized by 96 variables, these variables taking five values, namely: (i) YES, (ii) NO, (iii) ABSTENTION, (iv) NOT VOTING, (v) ABSENT.

The method of the author, based on the global objective function (Owsíński (1990)) was applied, allowing for (a) simultaneous determination of the clusters and their number, and (b) natural determination of the index of hierarchy, $r \in [0, 1]$, which accompanies the course of the procedure, decreasing from 1 for all objects (opinions) constituting individual clusters, along with successive sub-optimal aggregations to 0.5, for which the sub-optimal solution is found, and below, until all objects are clustered together.

The results are shown in Table 1. Let us emphasize that the MPs were not *a priori* characterized by their party membership. It is only *ex post* that the party labels were attached to them. In spite of this, clustering recovers with extremely high precision the political structure, divided into (A) the then ruling formally leftist coalition (SLD+PSL), (B) the centrist-rightist opposition (BBWR+KPN+UW+MN), and (C) the definitely leftist UP (currently, in 2003, in the ruling coalition with SLD), except for one or two MPs from the UP, who would either join the opposition or the ruling coalition.

Owing to generation of the index of hierarchy, r , the method allows for characterization of clusters forming the sub-optimal solution. And thus, the SLD+PSL coalition cluster (A) was formed so that the MPs of SLD grouped together already at r equal roughly 0.8, while those from PSL at r below 0.7. On the other hand, the opposition cluster, (B), took shape only at the r level lower than 0.6. Hence, we obtain a very distinct structure of separate clusters, whose internal compactness, though, is highly differentiated.

Clustering run/group	n^q	%PSL	%SLD	%UP	%BBWR	%KPN	%UW	%MN
I/1	298	100	100	—	—	—	—	—
I/2	115	—	—	3	100	100	100	100
I/3	40	—	—	97	—	—	—	—
II/1	298	100	100	—	—	—	—	—
II/2	113	—	—	3	100	100	100	100
II/3	42	—	—	97	—	—	—	—
III/1	301	100	100	3	—	—	—	—
III/2	113	—	—	3	100	100	100	100
III/3	39	—	—	94	—	—	—	—
IV/1	297	100	100	—	—	—	—	—
IV/2	114	—	—	3	100	100	100	100
IV/3	42	—	—	97	—	—	—	—

PSL – Peasant Party; SLD – Alliance of Democratic Left; UP - Union of Labour; BBWR – Block for Reforms; KPN – Confederation of Sovereign Poland; UW – Union of Freedom; MN – German Minority; n^q – number of MPs in groups obtained

Table 1. Groups of MPs in Polish Diet of 1993–97 for the first 96 votings

These two examples show that the appearance of very well discriminate sub-groups of opinions is not an incidental occurrence, but in definite situations, like those of group decision over a hot issue, may rather be a rule, and that an insight into their structure may be highly informative.

4 The consensus and distance

Now, what is the relation of these well-defined structures to anything like group-wise consensus? The notion of consensus comes from the complete agreement of all DMs $i \in I$, that is

$$C_i(K) = C(K), \quad \forall i \in I. \quad (1)$$

Attainment of such a complete agreement may be very difficult. Yet, it is not unrealistic, as demonstrated by over 200 years of effective practice of the Polish Diet in the 15th–17th century (and of the EU up to 15 as well). It requires (i) ample time (for discussion and in terms of horizon), (ii) skill in reaching (satisfactory) compromise, and (iii) feeling of responsibility of the DMs.

Stepping back from the requirement of complete agreement is being justified in a variety of ways: e.g. existence of the issues, over which compromise is impossible: there is no intermediate ground between “guilty” and “not guilty”, if even we admit different degrees of guilt. Another group of justifications is related to broadly conceived “uncertainty”.

There is, though, still another important argument for a weaker definition of “consensus”: the very natural fact that consensus may exist when (1) is not exactly fulfilled: some DMs “do not have an opinion”, either at all (“complete abstention”), or with respect to an aspect of $C_i(K)$. Thus, (1) can be replaced by a definition, according to which “consensus” is reached when

$$C_i(K) \text{ are not inconsistent with } C(K), \forall i \in I. \quad (2)$$

In (2) the symmetric “lack of inconsistency” means that $C_i(K)$ might not be identical with $C(K)$ and that $C(K)$ cannot imply a sequence of options that would be in conflict with individual $C_i(K)$. If some DMs do not decide on the preference as to the pair of options $k, l \in K$, while all the remaining DMs have all the same opinion as to this preference, then preservation of this preference in $C(K)$ is not inconsistent with all of the $C_i(K)$. This “extension” of the notion of consensus is fully natural and uncontroversial.

Now, let us further refer to the notion of distance between the opinions. According to (1) [full] consensus means that all distances between $C_i(K)$, $d(i, i') = 0$. The extension from (2) means that the [proper] consensus may take place also when $d(i, i') = 0$ does not hold. A further “natural” extension of the notion is as follows: if two formally different opinions $C_i(K)$ and $C_{i'}(K)$ have an identical effect in terms of the shape of $\mathbf{c}_A(K, I)$, which can be the result of the adopted aggregation procedure (like “only two first options matter”), then these opinions – for this case – can be considered to be in agreement, that is – a definite form of distance between them, denoted $d^*(i, i')$, is equal 0. Hence, we can assume that (an “operational definition”)

$$\begin{aligned} &\text{the set } \{C_i(K)\}_i \text{ is in the state of consensus, if the existing differences} \\ &\text{between the particular } C_i(K) \text{ do not impact upon } \mathbf{c}_A(K, I) \end{aligned} \quad (3)$$

corresponding to $d^*(i, i') = 0 \forall i, i'$. We apparently cannot go further in this direction without losing “naturality”.

5 The ideal structures and some of their properties

5.1 Introduction and notation

We may, however, start from another end of the spectrum: the one suggested by the two cases cited. Consensus might namely be understood not as an “absolute” agreement, in terms of (1)–(3), but as a relative agreement, relative – with respect to other opinions and/or a definite issue. Thus, in case of Fig. 1 we could say that there is consensus within each of two groups of eye-witnesses consisting in negating the other group’s opinion as to the guilt of the accused (or: consensus as to the guilt). They may not be of an identical opinion, but when asked a definite question, they will respond all in a manner distinctly different than the (an) other group. Note that we do mean here any

discriminant or decision-tree type of analysis (i.e. search for the “optimum” discerning question, which may turn out impractical), but identification of data structures, which by themselves offer “sufficient” separability.

In order to go further we will introduce now some new notions. Assume the partition P of the set I into subsets (groups of opinions, clusters) denoted A_q , $q = 1, 2, \dots, p$ (in cases we distinguish partitions the cardinality of the set of indices q will be denoted $p(P)$). The partitions $P = \{A_q\}_q$ are disjoint and exhaustive, i.e. $\bigcup_q A_q = I$, and $A_q \cap A_{q'} \neq \emptyset \forall q \neq q'$. We will also define subsets associated with a partition in a different manner, namely: $A(i) = \{A_q | i \in A_q\}$, and the corresponding q will be denoted $q(i)$. In addition, let d^q denote the cluster diameter, and $d^{qq'}$ – the distance between clusters q and q' . Although the definitions of d^q and $d^{qq'}$ may vary, we will refer to the instance of perhaps most intuitive definitions, namely

$$d^q = \max_{i,j \in I} d_{ij} \quad \text{and} \quad d^{qq'} = \min_{i \in A_q, j \in A_{q'}} d_{ij}. \quad (4)$$

We will now present some conditions on data (opinion) structures that (apparently) correspond to “natural” separation of subgroups, and the related problems. There are few references in the literature, devoted to this issue, in which conditions are either yardsticks on the “correctness” of the clustering procedures (Rubin (1967)), or the starting points for some clustering procedures (Luccio and Sami (1969), Nieminen (1980), Nowicki and Stańczak (1980)).

5.2 The globally ideal structure – and the problems

With the notations introduced we can formulate the strongest of the group-defining conditions:

$$\max_q d^q < \min_{q \neq q'} d^{qq'}. \quad (5)$$

This, indeed, is a very strong condition, to which we will refer as the “globally ideal (strong) structure” (GISs). Verbally, and intuitively, it is equivalent to: “*all the distances within clusters are shorter than any of the distances between the clusters*”. This condition is only slightly weakened with the replacement of “ $<$ ” by “ \leq ”, leading to the weak GIS (GISw). If we admit the definitions (4), the condition (5) turns into

$$\max_q \max_{i,j \in A_q} d_{ij} < \min_{q' \neq q} \min_{i \in A_q, j \in A_{q'}} d_{ij}. \quad (6)$$

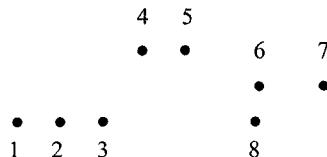
This condition, even though very strong, does not imply a uniquely defined “globally ideal strong structure”. If, namely, we assume, quite naturally, that $d_{ii} = 0 \forall i \in I$, then even the initial set I , conceived as a partition ($P \equiv I$), satisfies condition (5). Moreover, the partition formed by merging two objects (opinions), i' and i'' , such that $d_{i'i''} = \min_{i,j} d_{ij}$, and leaving all the other

objects apart, is also a GIS. Thereby, we arrive at the necessity of devising a procedure for identifying the *proper* GIS, that is – the maximum structure (in terms of cardinality of A_q or, otherwise, the level of aggregation procedure) satisfying the condition (5).

In view of the used definition of d^q , involving maximum, it appears that the classical agglomerative *complete linkage* algorithm may be of use. Thus, we hope to obtain the *proper* GIS by consecutively merging the closest clusters, starting with the initial set of objects (opinions) I . This is the general principle of the agglomerative algorithms. They differ by the manner, in which the matrix of distances, $\{d^{qq'}\}$, is updated after each merger. In the complete linkage algorithm, if clusters indexed q, q' are merged to form a new cluster, say q^* , then the new distances from the remaining clusters q'' are obtained as $d(q^*, q'') = \max\{d(q, q''), d(q', q'')\}$. Thereby, if a $d(q^*, q'')$ turns out afterwards the smallest one, leading to the merger of q^* and q'' , we can be sure we refer to d^q 's in the sense of the definition here adopted.

Complete linkage, though, does not control for the $d^{qq'}$ according to our definition. So, at each step the condition of GIS will have to be checked. Even if we do not have the proof that thereby we will actually obtain the proper GIS, we may at least hope for a good approximation.

Yet, there is one more bad surprise on this, already cumbersome road. Let us consider the following eight-object example:



in which we use the Euclidean distances d_{ij} . In this case the first four partitions, obtained, for instance, through complete linkage, are

$$P^0 = I, \quad P^1 = P^0 - \{1\} - \{2\} \cup \{1, 2\}, \quad P^2 = P^1 - \{4\} - \{5\} \cup \{4, 5\},$$

$$P^3 = P^2 - \{6\} - \{8\} \cup \{6, 8\}$$

and all of them are GISs. Suffice now, though, that $d_{67} < d_{13} < d_{78}$, as in the figure, that the next partition, obtained through any of the classical progressive merger schemes, will not be a GIS. In fact, if in the step leading to P^4 we merge $\{1, 2\}$ with $\{3\}$, then $d^{\{1,2,3\}} > d^{\{6,8\}\{7\}}$, violating (5), and similarly, if we merge $\{6, 8\}$ with $\{7\}$, then we have $d^{\{6,7,8\}} > d^{\{1,2\}\{3\}}$, also violating (5). Yet, by proceeding further, for instance with complete linkage, we would obtain $P^5 = \{\{1, 2, 3\}, \{4, 5\}, \{6, 7, 8\}\}$, a GIS! Hence, no progressive merger procedure would have avoided the trap of falling “on the way” into a non-GIS, itself being a direct consequence of (5), even though

the proper GIS could be obtained with an additionally controlled progressive merger procedure.

Until now, we have been considering the problem of determining the GIS for a given data set. We have established that a GIS exists for any set I and that it is not unique. Thus, the problem is transformed into the one of finding the “proper” (“maximum”) GIS. Yet, there is also, less demanding, the problem of verifying whether a given partition P of I is a GIS. Here, the algorithm is very simple: to check all the inter-cluster distances (their minimum) against the intra-cluster ones (their maximum).

In both these problems (determination of the proper GIS and verification whether a P is a GIS) there is a side-problem, of very high importance for the analysis of data: namely that of the “outliers”. In the framework of the analysis here presented an outlier is an object, whose removal makes out of a non-GIS a GIS. Again, the respective procedure is relatively straightforward for the case of verification, while it is not so at all in the case of construction. Roughly, the procedure would consist in the establishment and analysis of the subsets of distances, which do not comply with (5).

5.3 The locally ideal structures

In order to step down from the very stringent requirement of (5) let us take another intuitively obvious, but local, condition, which, at the object level looks as follows:

$$\forall i : j' \in A(i) \wedge j'' \notin A(i) \Rightarrow d_{ij'} < d_{ij''} \quad (7)$$

that is: “for any object, all its distances to objects from the same cluster are smaller than distances to objects from different clusters”. It is easy to see that (7) implies

$$\forall q : i, j \in A_q, j' \in A_q : d_{ij} < d_{ij'} \quad (8)$$

and further

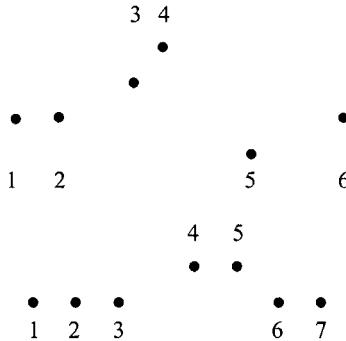
$$\forall q : d_{ij} < d_{ij'} \quad (9)$$

which, conform to the definitions adopted, is equivalent to

$$\forall q : d^q < \min_{q' \neq q} d^{qq'} \quad (10)$$

This condition will be referred to as LISs (local ideal structure, strong). It appears to be similar to (5), but is definitely not, see the example below:

In this example $d_{56} > d_{23}$. Hence, partition $P = \{\{1, 2\}, \{3, 4\}, \{5, 6\}\}$ is LISs, but not GISs. On the other hand, though, would we really consider $\{5, 6\}$ a cluster (even though (7) appears to be quite strong)? Note also, though, that the situation is quite different for clusters with high cardinalities, where this kind of locality may be quite natural. A less “strange” example



of a partition, which is LISs, but not GISs, is $P = \{\{1, 2, 3\}, \{4, 5\}, \{6, 7\}\}$ of the data set as below:

We can also formulate a “dual” to LIS at the cluster level, namely

$$\forall q \neq q' : \max\{d^q, d^{q'}\} < d^{qq'} \quad (11)$$

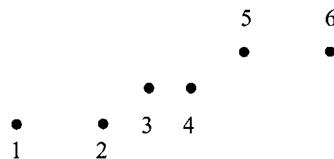
to which we will refer as LISds, and which is, in fact, equivalent to (10) (can be, similarly as (10), derived directly from (7)).

Here, again, we may ask for the constructive and verifying procedures. The answer is similar as for the GIS: an approximate technique of a progressive merger type with checking for the fulfilment of the respective condition in case of the constructive procedure, and a simple verification through direct application of (7).

Now, having the formulation (11) we can apply a reverse condition, also at the cluster level, namely:

$$\forall q, q' : \min\{d^q, d^{q'}\} < d^{qq'} \quad (12)$$

which can be referred to as LBSs (local biased structure). Despite the weakening, we can still perceive a certain sense in this condition: the diameter of the “smaller” cluster is always smaller than the distance between two clusters. Indeed, “small” clusters may happen to lie “close” to the “big” ones. Would we accept, though, as proper partition a LBSs structure like the one below?



Again, as before, this kind of structure may look much more “reasonably” with clusters larger in terms of cardinality. Let us note, though, that in this

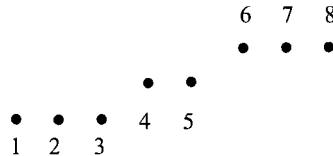
case we can no longer count on recovering the LBS for a given I through a simple progressive merger procedure, since the nearest-neighbourhood condition of the type of (7) is violated. We are limited, at least in terms of (relatively) simple procedures, to verify whether a given P is LBS. This, however, is no longer as simple as before. The procedure would start from the $\min_q d^q$, checking condition (8) against all the other clusters, then move to the next “smallest” cluster, etc.

It is highly probable that structures like (8) may result rather from the k -means-like procedures, especially those vested with variable scale devices, than from the hierarchical merger algorithms.

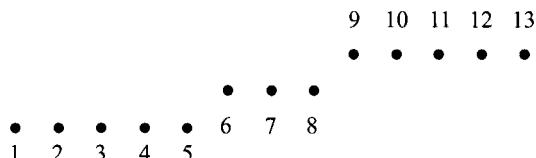
We will now introduce the last of the conditions considered, a condition appearing to really go very far in the direction of weakening requirements on the structure, but not necessarily so. Thus,

$$\forall q : \delta_{\max}^q < d^{qq'} \forall q' \neq q \quad (13)$$

where $\delta_{\max}^q = \max_{i \in A_q} \min_{i' \in A_q - i} d_{ij}$, meaning that it is the biggest nearest neighbour distance in cluster q , which was used here to replace the diameter from LISs. Now, a simile of the preceding example, satisfying (13), would look as follows:



and another example, as follows



In this case, we may be tempted to use single-linkage-like procedures, in which individual nearest-neighbour relations are most important, and the structures obtained often resemble those from the last example. Yet, these procedures have, as in case of (5) and the complete linkage, to be accompanied by checking of the other side of the condition. Here, again, verification of the existing P 's through direct application of (13) is relatively easy.

6 Conclusions

In numerous situations we would like to know, on the one hand, the structure of the set of opinions composing a definite aggregate (“election results” or a “decision”), and on the other hand – the consistency (“strength” or “stability”) of this structure, especially with respect to the notion of “consensus”, either for the entire population in question, or for the individual subsets, constituting the components of structure. It is essential to recognize the existence of such a need, the possibility of carrying out the analysis, and the interpretation of the results.

The conditions of an “ideal” or “perfect” structure are in fact positive cluster definitions, even if they are not unambiguous. These definitions, and especially in the context of several of them, form a kind of bridge between the classical progressive merger procedures, which refer to the object-level properties, and the k -means-like algorithms, which rely on the local cluster properties. At the same time, they provide a toolbox for verification of potential ideal or perfect character of concrete partitions obtained, in particular, through decision making procedures. In addition, we may be obtaining a new tool for the identification of outliers.

Thus, the work should be continued in order to (i) establish the relations between particular conditions and the associated sets of structures satisfying them, (ii) test the algorithms of determination of (proper) structures and of their verification, (iii) define for the general cases the interdependence between the output from cluster analysis (e.g. in terms of the objective function and/or the index of hierarchy) and the classes (definitions) of degree of agreement (“consensus”).

References

- LUCCIO, F. and SAMI, M. (1969): On the decomposition of networks in minimally interconnected subnetworks. *IEEE Trans. Circuit Theory*, CT-16, 2, 184–188.
- NIEMINEN, J. (1980): On minimally interconnected subnetworks of a network. *Control & Cybernetics*, 9, 1-2, 47–52.
- NOWICKI, T. and STAŃCZAK, W. (1980): Partitioning a set of elements into subsets due to their similarity. In: E. Diday, L. Lebart, J.P. Pagès, and R. Tomassone (Eds.): *Data Analysis and Informatics*. North Holland, Amsterdam, 583–591.
- OWSIŃSKI, J.W. (1990): On a new naturally indexed quick clustering method with a global objective function. *Applied Stochastic Models and Data Analysis*, 6, 1, 157–171.
- OWSIŃSKI, J.W. and ZADROŻNY, S. (2000): Structuring the set of MPs in Polish Parliament: a simple clustering exercise. *Annals of Operations Research*, 97, 15–29.
- RUBIN, J. (1967): Optimal Classification into Groups: An Approach for Solving the Taxonomy Problem. *J. Theoret. Biol.*, 15, 103–144.

Volatility Forecasts and Value at Risk Evaluation for the MSCI North America Index

Momtchil Pojarliev¹ and Wolfgang Polasek²

¹ INVESCO Asset Management,
Bleichstrasse 60-62, D-60313 Frankfurt, Germany

² Institute of Advanced Studies, Vienna, Stumpergasse 56, 1060 Wien
Email: polasek@ihs.ac.at

Abstract. This paper compares different models for volatility forecasts with respect to the value at risk performance (VaR). The VaR measures the potential loss of a portfolio for the next period at a given significance level. We focus on the question if the choice of the appropriate volatility forecasting model is important for the VaR estimation. We compare the forecasting performance of several volatility models for the returns of the MSCI North America index. The resulting VaR estimators are evaluated by comparing the empirical failure rate with the forecasting performance.

1 Introduction

In April 2000 the NASDAQ 100 index dropped by about 20%. Similar developments in other stock markets have led to more research in risk assessments: Investors are interested in reliable risk measures of their portfolios, like the VaR measure, the lower quantiles of the predicted returns distribution. There are numerous approaches to calculate the VaR. Many analysts use the popular variance-covariance approach by J.P. Morgan, better known as RiskMetrics (1996). Recently, new methods have been developed on the basis on extreme value theory (Danielson and Vries (1997))and a comparative approach for Germany can be found in Zucchini and Neumann (2001). The aim of the paper is to investigate the following question: Can time series forecasts improve the VaR estimate and what evaluation criteria for volatility forecasts are good diagnostics for the VaR performance? Therefore we will compare the performance of the following models: The naive model, where the variance estimator is just the historical variance, the RiskMetrics model, the GARCH, an asymmetric GARCH (AGARCH) and a bivariate BEKK model.

The paper is organized as follows: In the next section we estimate the volatility of the MSCI North America index with the above models. The forecasting performance of the models is compared in section 3 using a rolling sample of 800 observation and an evaluation period of approximately two years. As forecasting evaluation criteria we use an auxiliary linear regression

model (Pagan and Schwert (1990)) and the Christoffersen tests (Christoffersen (1998)) for VaR. In section 4 we estimate the VaR of a hypothetical portfolio of 1 Mio \\$ and compare the results. In the last section we conclude.

2 Volatility models

We will investigate the volatility of the daily returns of the MSCI North America index from May 1st 1995 until April 3rd 2000. The first 800 observations (from 1st May 1995 until 22 May 1998) are used as training sample and the remaining observation are used for out-of-sample comparison.

2.1 The naive model

The naive model uses the variance of a moving sample of 800 observation (approximately 3 years) as forecast for the next period (1000 observations were used in Dockner and Scheicher (1999)).

$$\hat{\sigma}_{t+1}^2 = \frac{1}{799} \sum_{i=1}^{800} (r_{t+1-i} - \hat{\mu})^2 \quad (1)$$

where r_t are the returns at time t and $\hat{\mu}$ is the estimated average return of the sample.

2.2 The RiskMetrics model

The model proposed by J. P. Morgan (1995) is an exponentially weighted moving average model. The volatility of the next period can be calculated as a weighted average of the current volatility and the squared returns:

$$\sigma_{t+1}^2 = \lambda \sigma_t^2 + (1 - \lambda) r_t^2 \quad (2)$$

where λ is the weight factor. As proposed by RiskMetrics we set λ equal to 0.94. As an initial value for σ^2 , we use the squared returns.

2.3 The GARCH model

The simplest form of observation model we can use for a returns vector is:

$$r_t = \mu + \epsilon_t, \quad (3)$$

where the conditional distribution of ϵ_t is assumed to be Gaussian with mean zero and variance σ_t^2 . A GARCH(p,q) model uses the following parameterisation for the variance:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p \alpha_i \epsilon_{t-i}^2 + \sum_{i=1}^q \beta_i \sigma_{t-i}^2. \quad (4)$$

The information criteria AIC and BIC in Table 1 suggest to select a GARCH(1,1) model for the period from 1st May 1995 until 22 May 1998 for the daily returns of the MSCI North America index.

The estimated GARCH(1,1) model of the returns of the MSCI North America index is (t-values in parentheses):

$$\hat{\sigma}_t^2 = 10^{-6} 1.23 + 0.083 \epsilon_{t-1}^2 + 0.904 \hat{\sigma}_{t-1}^2. \quad (5)$$

$(t - val.) \quad (2.48) \quad (7.80) \quad (57.47)$

2.4 The asymmetric GARCH model

An asymmetric GARCH(p,q) model (AGARCH) has the form

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^p (\alpha_i + \gamma_i S_{t-i}) \epsilon_{t-i}^2 + \sum_{i=1}^q \beta_i \sigma_{t-i}^2 \quad (6)$$

where the dummy variable for the negative residuals is defined as

$$S_{t-i} = \begin{cases} 1 & \text{if } \epsilon_{t-i} < 0; \\ 0 & \text{if } \epsilon_{t-i} \geq 0. \end{cases} \quad (7)$$

The idea is that asymmetric behaviour of the negative deviations are sources for additional risk. We estimated the AGARCH(1,1) model as follows:

$$\hat{\sigma}_t^2 = 10^{-6} 3.3 + 0.003 \epsilon_{t-1}^2 + 0.18 S_{t-1} \epsilon_{t-1}^2 + 0.86 \hat{\sigma}_{t-1}^2. \quad (8)$$

$(t - val.) \quad (4.23) \quad (0.19) \quad (7.27) \quad (44.21)$

The asymmetry parameter γ has a significant t value which is a sign of a rather strong asymmetric persistence effect. The sum of the parameters is less than 1 for positive shocks but is larger than 1 for negative shocks.

2.5 The BEKK model

We want to investigate to what extend multivariate models can improve the variance forecasts. Let $r_t = (r_{1t}, r_{2t})'$ be a vector of returns, then a Gaussian multivariate GARCH model is given by

	AIC	BIC
GARCH(1,0)	-5402	-5374
GARCH(1,1)	-5477*	-5440*
GARCH(2,1)	-5475	-5428
GARCH(2,2)	-5455	-5398

Table 1. AIC and BIC for different GARCH models, MSCI North America index. The star (*) denotes the smallest value.

$$r_t | I_{t-i} \sim N[\mu_t, H_t] \quad , t = 1, \dots, T \quad (9)$$

where we assume $\mu_t = \mu$ for the conditional mean and H_t for the conditional covariance matrix given the information set I_t up to time t .

The BEKK(p,q) model of Engle and Kroner (1995) assumes the following parameterisation of the conditional covariance matrix:

$$H_t = A_0 A_0' + \sum_{i=1}^p A_i (\epsilon_{t-i} \epsilon_{t-i}') A_i' + \sum_{i=1}^q B_i H_{t-i} B_i'. \quad (10)$$

where the transposed matrix pairs for each of the coefficient matrices A_i and B_i guarantee symmetry and non-negative-definiteness of the conditional covariance matrix H_t . Using the AIC criteria, we select a BEKK(2,1) model for the returns of the MSCI Europe and MSCI North America indices. For the VaR estimation we use the second diagonal element of the H_t which is the variance of the MSCI North America index. An interesting result is that a bivariate BEKK(1,1) model, where the MSCI World index is used as leading indicator, performs worse with respect to the forecasting performance for the volatility of the MSCI North America index.

3 Forecasting performance

Using the estimation results of the previous two sections and a rolling sample of 800 trading days we forecast the volatility of the returns of the MSCI North America index for the out-of-the-sample period from 25 May 1998 until 3rd April 2000 (we estimate each model 486 times with 800 observations).

We use the auxiliary linear regression approach by Pagan and Schwert (1990) to evaluate the forecasting performance of the volatility models. We simply regress the "realized volatility" (i.e. the squared returns) on a constant and the forecasted volatility:

$$r_t^2 = \alpha + \beta \hat{\sigma}_t^2 + \epsilon_t, \quad t = 1, \dots, T. \quad (11)$$

The constant α should be equal to 0 and the slope β equal to 1. The t-statistic of the coefficients is measure for the bias from the ideal prediction and the R^2 is an overall measure of the forecasting performance. Table 2 summarizes the results of the auxiliary regression (11).

There is a clear sign that multivariate modeling improves the forecasting performance. The BEKK model leads to the smallest "bias" for the intercept and to the largest R^2 .

4 VaR comparison

We assume a hypothetical portfolio of 1 Mio US \$ which follows the MSCI North America index and the VaR is estimated for the next trading day (in

	$\alpha(t\text{-st.})$	$\beta(t\text{-st.})$	R^2
Naive	0(3.34)	-1.134(-1.47)	0.004
RiskMetrics	0(2.46)	0.575(3.52)	0.025
GARCH(1,1)	0(2.83)	0.571(3.98)	0.032
AGARCH(1,1)	0(3.45)	0.593(5.45)	0.058
BEKK(2,1)	0(0.13)	1.034(5.83)	0.062

Table 2. Model comparison by auxiliary regression of the MSCI North America index (1998/05/25 - 2000/04/03).

the period from 25 May 1998 to 3rd April 2000). Assuming that the returns are normally distributed, the 95%-VaR is computed as 95% quantile of the returns distribution i.e.

$$\widehat{VaR}_{t+1} = -10^6 1.645 \hat{\sigma}_{t+1} \quad (12)$$

where $\hat{\sigma}_{t+1}$ is the forecast of the standard deviation given all information up to time t . A 95%-VaR of 1000 US \$ means that with 5% probability the maximum loss in the next trading day will be more than 1000 US \$.

Figure 1 plots the actual portfolio changes ($10^6 r_t$) by tracking the MSCI North America index, the VaR estimates based on the AGARCH(1,1) model and the VaR estimates based on the naive model. The naive model doesn't capture the risk for the first 100 observations and overestimates it in the rest of the period. It is clear that using the VaR estimates of the naive model overestimates the risk if the volatility is small and underestimate it in bear markets.

4.1 Evaluating the VaR by interval forecasts

To evaluate the performance of the different $(1 - \alpha)\%$ -VaR estimators we are using the following four criteria:

1. The failure rate (F)
2. Likelihood ratio test of unconditional coverage (LR_{uc})
3. Likelihood ratio test of independence (LR_{ind})
4. Joint test of coverage and independence (LR_{cc})

We define the failure rate (F) as the number of times for which the actual loss is larger than the estimated VaR. For $t = 1, \dots, T$ we define the test statistic

$$F = \sum_{i=1}^T D_t \quad (13)$$

with the dummy variable

95%-VaR based on the AGARCH(1,1) model and on the naive model

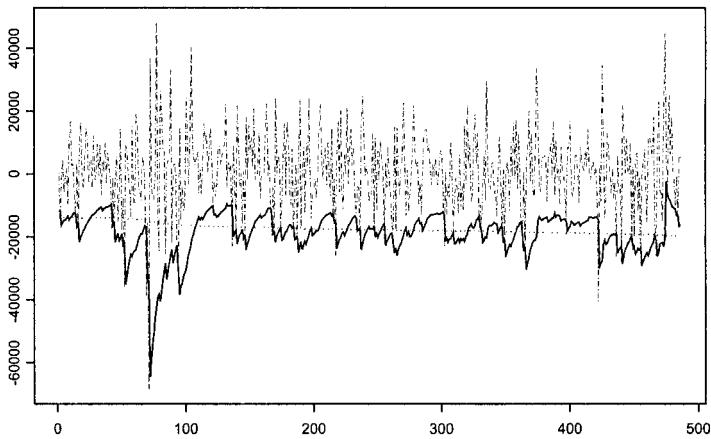


Fig. 1. Actual portfolio changes of the MSCI North America index and the 95%-VaR estimates for the AGARCH(1,1) model and the naive model from 25 May 1998 until 3rd April 2000

$$D_t = \begin{cases} 1 & \text{if } VaR_t - P_t^a > 0; \\ 0 & \text{if } -VaR_t - P_t^a \leq 0. \end{cases} \quad (14)$$

The following LR_{uc} , LR_{ind} and LR_{cc} tests are proposed by Christoffersen (1998) to evaluate forecasts over a certain horizon. The likelihood ratio (LR) test of unconditional coverage LR_{uc} tests if $E(D_t) = (1 - \alpha)T$ against $E(D_t) \neq (1 - \alpha)T$ where α is the probability level for the VaR and T is the number of trading days ($T = 486$) in the evaluation period.

The LR test of independence LR_{ind} tests the hypothesis of independence against a first order Markov chain. Independence would mean that the days for which the actual losses are larger than the estimated value-at-risk ($D_t = 1$) are independent from each other.

The above tests are combined into LR_{cc} test, where the null of the unconditional coverage test is tested against the alternative of the independence tests. The three tests are numerically related by the following equation (see Christoffersen (1998)):

$$LR_{cc} = LR_{uc} + LR_{ind}. \quad (15)$$

The results of the Christoffersen tests for the 90% to 99% VaR based on the BEKK(2,1) model are presented in Figure 2. The three panels plot the LR

statistics for the $(1-\alpha)$ -VaR in steps of 1%. The horizontal line corresponds to the 5 per cent critical value of the relevant chi-squared distribution. We see that the null hypothesis is not rejected for α -levels between 3% and 10%. Since the LR_{uc} and LR_{cc} tests are rejected for the 99%-VaR we conclude that the BEKK model might not be an optimal model for an $\alpha = 1\%$ level.

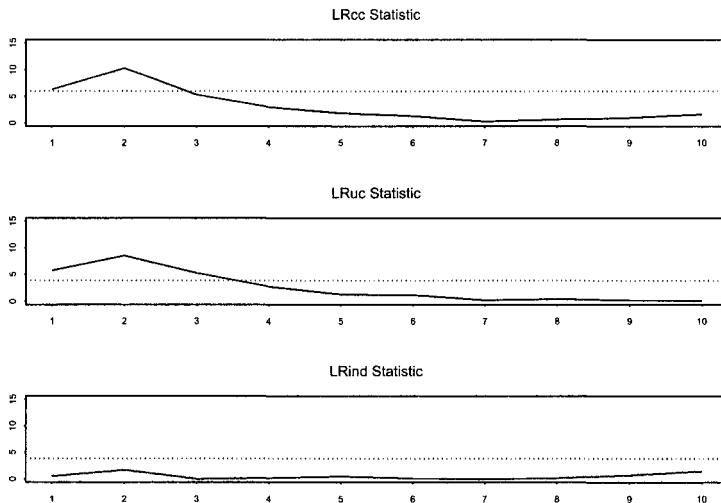


Fig. 2. Christoffersen tests of the $(1 - \alpha)\%$ -VaR (for α between 1% and 10%) for the BEKK(2,1) model of the MSCI North America index

Table 3 summarizes the result for the out-of-the-sample performance of the investigated models for the different indices. The + and - for the Christoffersen tests mean that the null hypothesis is accepted or rejected at a level of 5% respectively. Note that the BEKK(2,1) model leads to the lowest failure rate.

	LR_{uc}	LR_{ind}	LR_{cc}	failures rate (in %)
Naive	-	+	-	39 (8.02)
RiskMetrics	+	+	+	31 (6.38)
GARCH	+	+	+	34 (6.99)
AGARCH	-	+	+	35 (7.20)
BEKK	+	+	+	30 (6.17)

Table 3. 95%-VaR performance for 486 days, MSCI North America index.

5 Conclusions

In this paper we have compared VaR estimates based on volatility forecasts. Therefore we have evaluated the volatility forecasting performance of 3 univariate models together with the historical variance and a bivariate BEKK model. The BEKK model yields the largest R^2 , followed by the AGARCH model in the auxiliary regression. Comparing the resulting VaR estimates we found that the multivariate volatility forecasts model yields the best results. The bivariate BEKK model with the MSCI Europe index as a leading indicator leads to the smallest failure rate for 95%-VaR. Surprisingly, the AGARCH model performs worse than the GARCH model with respect to the VaR. The R^2 of the auxiliary regression model can be used as indicator for the goodness of the resulting VaR estimates even if the R^2 values are very small. The Christoffersen tests are useful as diagnostic tests for a good VaR model when the α -level is very small. In a further research paper (see Pojariiev and Polasek (2000)) we extend this approach to evaluate VaR models for stock returns for Europe and the Pacific area.

References

- CHRISTOFFERSEN, P. (1990): Evaluating Interval Forecast. *International Economic Review*, 39, 841–862.
- DANIELSON, J. and VRIES, C. (1997): Value-at-Risk and Extreme Returns. In: No 98-017/2 in Tinbergen Institute Discussion Papers from Tinbergen Institute <http://www.tinbergen.nl/discussionpapers/98017.pdf>.
- DOCKNER, E. and SCHEICHER, M. (1999): Evaluating Volatility Forecasts and Empirical Distributions in Value at Risk Models. *Finanzmarkt und Portfolio Management*, 1, 39–55.
- ENGLE, R. and KRONER, K. (1995): Multivariate Simultaneous GARCH. *Econometric Theory*, 11, 122–150.
- MATHSOFT (1996): S-Plus, S+Garch User's Manual: *Data Analysis Products Division, MathSoft, Seattle*.
- PAGAN, A. and SCHWERT, W. (1990): Alternative Models for Conditional Stock Volatility. *Journal of Econometrics*, 50, 267–290.
- POJARLIEV, M. and POLASEK, W. (2000): Value at Risk estimation for stock indices using the Basle Committee proposal from 1995: *University of Basel*, <http://www.ihs.ac.at/polasek>.
- RiskMetrics (1996): Risk Metrics Technical Document, fourth edition <http://www.riskmetrics.com/research/>.
- ZUCCHINI, W. and NEUMANN, K. (2001): A Comparison of Several Time Series Models for Assessing the Value at Risk of Shares. *Applied Stochastic Models for Business and Industry* 2000, 135–148.

Selected Methods of Credibility Theory and its Application to Calculating Insurance Premium in Heterogeneous Insurance Portfolios

Wanda Ronka-Chmielowiec and Ewa Poprawska

Department of Financial Investments and Insurance,
Wroclaw University of Economics,
ul. Komandorska 118/120, 53-345 Wrocław, Poland

Abstract. In the first part of the paper the Bühlmann and Bühlmann-Straub models will be reviewed. Both are useful to model portfolios of insurance policies in which some individual contracts or groups of contracts are characterized by non-standard claim experience. The insurance premium for such heterogeneous portfolios (the credibility premium) is calculated by maximum likelihood estimation. In the second - empirical part of the paper several examples of heterogeneous policy portfolios will be presented and appropriate credibility premium calculated.

Introduction

For an insurance company calculating the risk premium, which is in fact the price that an insured pays for insurance, is the essential stage of the risk management process. One of the basic rules that should be followed while calculating a pure risk premium is the equivalence of premiums and expected amount of claims. This means that pure risk premium (P) for an individual risk X should be equal to the mean value of X .

In homogeneous portfolios the pure risk premium is estimated by an overall mean of historical observations of claim amounts. But the situation varies when not all the risks are homogeneous or when the policyholders have some previous risk experience, which should be taken into consideration in determining the risk premium. Moreover it is relevant to use a credibility premium for the policyholders that systematically reduces their individual risk and the amount of losses. This helps to avoid negative selection. If the amount of data is too small to allow an insurer to estimate the expected loss based only on the individual experience of the contract, one should combine an experience of a whole portfolio with an individual one. The credibility theory provides with several models and methods that are helpful to obtain the appropriate risk premium.

The main problem is to answer the question how credible should the individual and the collective experience be. There are two extreme solutions possible:

- one is to charge a premium estimated by the overall mean of the data;
- the other one is to calculate a premium based on the average claim for the individual contract or group of contracts. This approach is justified for individually negotiated contracts, especially when catastrophic risks are involved in the contract, but can only be applied if the claim experience is large enough.

In practice one can find a solution, which compromises these two extremes. The calculation of the credibility premium is reduced to a simple form of a weighted average of individual and collective mean:

$$z_j E(X_j) + (1 - z_j) E(X) \quad (1)$$

when X_j is a random variable which represents risk (claim distribution) for the j -th contract or group of contracts of the same terms, X is the claim distribution for a single claim in a portfolio, z_j is called the credibility factor and $0 \leq z_j \leq 1$.

In order to calculate the credible premium one should find an appropriate estimator for the credibility factor z_j , which is in fact, a weight attached to individual claim experience and expresses how 'credible' the individual experience of contract j is. One of the methods that allows to find the credibility factor is based on 'maximum credibility theory' which was presented in details by Bailey (1950) and then developed by Bühlmann (1967).

So as to clarify the idea of credibility, we note that we have n historical data of claim amounts for each cell (contract or group of contracts) j , which may be written in form of a vector of n elements $X_j = (X_{1j}, X_{2j}, \dots, X_{nj})$. Moreover we consider all X_{jt} to be random variables with mean m . The sample mean for a contract is then equal to:

$$\bar{X}_j = \frac{X_{1j} + X_{2j} + \dots + X_{nj}}{n} \quad (2)$$

and may differ from the overall mean m . We assume that the differences are random. We assume that the risk exposure is not the same for each contract. It may be characterized by a risk parameter. We assume this parameter to be a random variable. The distribution of this random variable describes the risk structure in the portfolio of insurance policies.

1 The balanced Bühlmann model

All the symbols are based on Kaas et al. (2001). In this model we assume that a random variable X_{jt} represents a claim statistic for cell j in time t . For the sake of simplicity it is assumed that each cell contains a single contract or group of them (there is the same number of contracts in all groups). The claim amount for cell j in time t may be decomposed into a sum of three separate components. The first one - m is the average claim for a whole

portfolio of policies. The second one is the deviation from the overall mean characteristic for cell j and the third one characterizes the deviation specific for time period t (it describes good and bad year for a policyholder). We assume deviations to be independent random variables. We can decompose a claim statistic for cell j in time t as follows:

$$X_{jt} = m + \Xi_j + \Xi_{jt} \quad (3)$$

with $j = 1, \dots, J$, $t = 1, \dots, T$, and Ξ_j , Ξ_{jt} are independent random variables and $E(\Xi_j) = E(\Xi_{jt}) = 0$, $\text{Var}(\Xi_j) = \alpha$, $\text{Var}(\Xi_{jt}) = s^2$. In this model we assume all variance of X_{jt} to be equal.

We can interpret the components of X_{jt} as follows:

- m is the overall mean of the whole portfolio of policies - the expected claim amount for an average policyholder;
- Ξ_j is a random deviation from this mean for contract j . The conditional mean, given $\Xi_j = \xi$ of the random variables X_{jt} equals to $m + \xi$ which is a long-term average claim for contract j . The component Ξ_j describes the quality of risk specific to the j -th contract. The $\text{Var}(\Xi_j) = \alpha$ characterizes the differences between contracts, the distribution of Ξ_j describes the risk structure of the portfolio.
- the component Ξ_{jt} characterizes the deviation from the long-term average for time t .

Then one may show that the predictor for the claim amount to be paid in the next period $T+1$ (which may be described by an unknown random variable $X_{j,T+1}$), which is a linear combination of all observable data X_{11}, \dots, X_{JT} with the same mean as $X_{j,T+1}$ is of the form of

$$z_j \bar{X}_j + (1 - z_j) \bar{X}. \quad (4)$$

It may be also proven that for such predictor the mean squared error is minimal. Moreover the obtained predictor is the appropriate premium that is the base of the price for this contract. All the statements written below may be formally summarized in the

Theorem:

Assume that claim amount for contract j in time t may be written as a sum of stochastically independent components:

$$X_{jt} = m + \Xi_j + \Xi_{jt}, \quad j = 1, \dots, J, t = 1, \dots, T+1, \quad (5)$$

where all Ξ_j are identically distributed with $E(\Xi_j) = 0$ and $\text{Var}(\Xi_j) = \alpha$, similarly all Ξ_{jt} are identically distributed with $E(\Xi_{jt}) = 0$ and $\text{Var}(\Xi_{jt}) = s^2$ for all j and t .

Then the linear combination

$$g_{11}X_{11} + \dots + g_{JT}X_{JT}, \quad (6)$$

which is the best predictor of $X_{j,T+1}$ in sense of minimal mean squared error

$$E[(X_{j,T+1} - g_{11}X_{11} - \dots - g_{JT}X_{JT})^2] \quad (7)$$

is equal to the credibility premium

$$z\bar{X}_j + (1-z)\bar{X}, \text{ where } z = \frac{\alpha T}{\alpha T + s^2} \quad (8)$$

is the credibility factor, which is in this case equal for all j . The collective estimator of m (an overall mean) is of the form of:

$$\bar{X} = \frac{1}{JT} \sum_{j=1}^J \sum_{t=1}^T X_{jt} \quad (9)$$

and the individual estimator of mean for an individual contract j is of the form:

$$\bar{X}_j = \frac{1}{T} \sum_{t=1}^T X_{jt}. \quad (10)$$

The optimal credibility factor has several asymptotic properties:

- If $T \rightarrow \infty$, then $z \rightarrow 1$. This means that the more claim experience one has, the more credible the individual experience is.
- If $\alpha \rightarrow 0$, then $z \rightarrow 0$. This means that when the differences between individual contracts are low (the portfolio is in fact not heterogeneous) the collective mean m is the optimal estimator of the risk premium.
- If $\alpha \rightarrow \infty$, then $z \rightarrow 1$. This is intuitively clear, because if the differences between contracts in a portfolio are significant, the information about the other contracts does not provide information about risk j .
- If $s^2 \rightarrow \infty$, then $z \rightarrow 0$. The interpretation is as follows: if for a fixed risk parameter the claim observations are extremely variable, individual contract experience is useless for charging the risk premium.

2 Example 1 - estimating the credibility premium in the Bühlmann model

The purpose of the example is to illustrate some practical aspects of using the Bühlmann model in charging an appropriate risk premium in a heterogeneous portfolio of insurance policies. The calculations are done on pseudo-random observations, which are considered to simulate claims having occurred in the portfolio of a house insurance. The contracts are divided into 12 groups for the sake of two categories:

1. the age of the building - under this category policies are divided into 3 subgroups:

- new buildings (of the age of less than 10 years);
 - old buildings (of the age of more than 30 years);
 - of the age between 10 and 30 years;
2. location of the building - with its reference to flood and extreme weather risks (buildings are divided into 4 subgroups).

Moreover we assume that observations of 10 periods for each group of contracts are available and that there is exactly the same number of policies in each group. The claim amounts are counted in thousands of Euro.

The way of calculating the credibility factor in this case is based on the method described in details in Kaas et al. (2001).

Every single data had been generated from a gamma distribution. The mean values for each group are different and they are also random (normally distributed) what expresses the risk structure of the portfolio. The overall mean of the whole portfolio is equal to 100 and the variance 60.

To estimate the credibility factor we calculate the following statistics:
The mean-square-between:

$$MSB = \frac{\sum_{j=1}^J T (\bar{X}_j - \bar{X})^2}{J - 1} \quad (11)$$

and the mean-square-within:

$$MSW = \frac{\sum_{j=1}^J \sum_{t=1}^T (X_{jt} - \bar{X}_j)^2}{J(T - 1)}. \quad (12)$$

It can be shown (see Kaas et al. (2001)) that the statistic defined as $F = MSB/MSW$ has the mean $\alpha T + s^2$ while MSW has the mean s^2 , hence we can use $(1 - 1/F)$ to estimate z .

In the analysed example the statistics defined above are equal to:

$$MSB = 161,43 \quad MSW = 68,55 \quad F = 2,3549$$

and the resulting credibility factor is $z = 0,575353$.

It is worth to remind that the credibility factor is equal for each group of policies. The resulting credible premiums for each group of policies are displayed in Table 1.

3 The Bühlmann-Straub model

There are many modifications of the basic model possible. One of them is the model described below. Similarly to the previous one, in this model the observations can be decomposed as follows:

$$X_{jt} = m + \Xi_j + \varepsilon_{jt}, \quad j = 1, \dots, J, \quad t = 1, \dots, T + 1, \quad (13)$$

Group	\bar{X}_j	credibility premium
1	107,19	104,28
2	94,13	96,76
3	98,67	99,38
4	98,07	99,03
5	98,01	98,99
6	103,13	101,94
7	98,09	99,04
8	103,09	101,92
9	100,15	100,23
10	97,19	98,53
11	107,19	104,28
12	99,18	99,67

Table 1. Credibility premium for groups of policies

where the unobservable risk components Ξ_j are independent and identically distributed with mean 0 and variance α ; the components Ξ_{jt} are also independent with mean 0. Furthermore the components Ξ_j and Ξ_{jt} are assumed to be independent too. The main difference between this model and the previous one is that in the Bühlmann-Straub model the variance of the components equals to:

$$\text{Var}(\Xi_{jt}) = \frac{s^2}{w_{jt}}, \quad (14)$$

where w_{jt} is the weight of the observation X_{jt} . These weights are adequate to the relative precision of the observations. The weights have the most natural interpretation if we assume that the cell j is not a single contract but a group of them and the number of claims in different groups are not equal. In this case one may consider X_{jt} to be an average of replications. Hence

$$X_{jt} = \frac{1}{w_{jt}} \sum_k X_{jtk}, \quad (15)$$

where $X_{jtk} = m + \Xi_j + \Xi_{jtk}$ and Ξ_{jtk} are independent and identically distributed random variables with mean 0 and variance s^2 . They describe the deviation from $m + \Xi_j$ for an individual contract k in group j and period t . It is possible to derive the best predictor of the risk premium $m + \Xi_j$ in the sense of minimal mean squared error in the form of:

$$z_j X_{jw} + (1 - z_j) X_{zw}, \quad (16)$$

where:

$$z_j = \frac{\alpha w_{j\Sigma}}{s^2 + \alpha w_{j\Sigma}}, \quad (17)$$

$$z_{\Sigma} = \sum_{j=1}^J z_j, \quad w_{j\Sigma} = \sum_{t=1}^T w_{jt}, \quad w_{\Sigma\Sigma} = \sum_{j=1}^J w_{j\Sigma}, \quad (18)$$

$$X_{jw} = \sum_{t=1}^T \frac{w_{jt}}{w_{j\Sigma}} X_{jt}, \quad X_{zw} = \sum_{j=1}^J \frac{z_j}{z_{\Sigma}} X_{jw}, \quad X_{ww} = \sum_{j=1}^J \frac{w_{j\Sigma}}{w_{\Sigma\Sigma}} X_{jw}. \quad (19)$$

Hence for the contract j the credible premium is of the form of (16), where X_{jw} is the individual estimator for the risk premium, X_{zw} is the credibility weighted collective estimator, and z_j is the credibility factor for policy j .

4 Example 2 - estimating the credibility premium in the Bühlmann-Straub model

Similarly to the first example, we analyze a portfolio of a house insurance. As in the previous example the observations are divided into 12 groups, but the assumption of equality of numbers of policies in all groups and in all periods is rejected. The portfolio is dynamic and the number of contracts is changing during the observable 10 years. That makes this situation more adequate to reality. Under these assumptions it is necessary to use a Bühlmann-Straub model in which weights are introduced. It should be stressed that in this example natural weights are used.

The credibility estimators in this model depend on the unknown structure parameters: m , α and s^2 . The estimators of s^2 and α are based on the following statistics:

weighted sum-of-squares-within:

$$SSW = \sum_{j,t} w_{jt} (X_{jt} - X_{jw})^2 \quad (20)$$

and the weighted sum-of-squares-between:

$$SSB = \sum_j w_{j\Sigma} (X_{jw} - X_{ww})^2. \quad (21)$$

Hence the unbiased estimators of the structure parameters are as follows (the proof is presented in detail in Kaas et al. (2001)):

$$\tilde{m} = X_{ww}, \quad \tilde{s}^2 = \frac{SSW}{J(T-1)}, \quad \tilde{\alpha} = \frac{SSB - (J-1)\tilde{s}^2}{w_{\Sigma\Sigma} - \sum_j (w_{j\Sigma}^2 / w_{\Sigma\Sigma})}. \quad (22)$$

In the analyzed example we obtain the following values of estimators described above:

Group	X_{zw}	credibility factor z_j	credibility premium
1	341,88	0,8977	335,63
2	280,08	0,9096	280,15
3	220,35	0,9429	223,80
4	291,51	0,9652	291,14
5	242,41	0,9559	244,11
6	234,39	0,9236	237,94
7	288,19	0,9600	287,90
8	249,12	0,9503	250,70
9	299,07	0,9533	298,22
10	312,93	0,9441	311,14
11	257,72	0,9630	275,92
12	339,51	0,9048	333,93

Table 2. Credibility premium for groups of policies

$X_{zw}=280,88$; $SSB=2\ 496\ 464$; $SSW=1\ 322\ 651$; $s^2=12\ 246,77$; $\alpha=1\ 130,65$. The resulting appropriate credibility factor and the pure credibility premium for contracts in each 12 groups are displayed in Table 2.

In spite of the fact that the models described in the paper seem to be rather simple, surprisingly they may be very useful in practice.

Summary

The main advantage of the models described in the paper is that they are one of the simplest. There is no need to assume any form of the distribution of individual claim amount and no need for parameter estimation. The calculations of the credibility factor are based on very simple statistics (mean value and variance). Moreover they may be applied even when the amount of available data is too small for using more sophisticated methods like for example GLM based models. Another advantage is their flexibility. The weights included in Bühlmann-Straub models may be either chosen as natural weights or arbitrary chosen ones. And finally both models are easy to extent by dividing observations in more components.

References

- BAILEY, A. (1950): Credibility procedures. *Proceedings of the Casualty Actuarial Society*, XXXVII, 7-23, 94-115.
- BÜHLMANN, H. (1967): Experience rating and credibility. *ASTIN Bulletin* 4, 199-207.
- KAAS, R., GOOVAERTS, M., DHAENE, J., and DENUIT, M. (2001): *Modern Actuarial Risk Theory*. Kluwer Academic Publishers, Boston.

Support Vector Machines for Credit Scoring: Extension to Non Standard Cases

Klaus B. Schebesch and Ralf Stecking

Institut für Konjunktur- und Strukturforschung,
Universität Bremen, D-28359 Bremen, Germany

Abstract. Credit scoring is being used in order to assign credit applicants to good and bad risk classes. This paper investigates the credit scoring performance of support vector machines (SVM) with weighted classes and moderated outputs. First, we consider the adjustment of support vector machines for credit scoring to a set of non standard situations important to practitioners. Such more sophisticated credit scoring systems will adapt to vastly different proportions of credit worthiness between sample and population. Different costs for different types of misclassification will also be handled. Second, sigmoid output mapping is used to derive default probabilities, important for constructing rating systems and a step towards more “personalized” credit contracts.

1 Introduction

A more accurate prediction of the expected defaulting behavior of new credit applicants is an important contribution to risk management in finance. New regulations in the credit business, and the option for more “personalized” credit contracts call for a better understanding of the available data on the evolution of past credit contracts and to extract more statistical information in order to enhance credit risk models. Credit scoring is a basic binary classification task in finance. But even this task is not a “standard” problem of cost-neutral predictions on representatively sampled data. Recorded credit data are extremely biased in favor of the non-defaulting cases. Adapted modeling would make a huge difference in terms of profit or losses avoided.

From an applications point of view, **weighting classes** is most useful (Lin (1999), Lin et al. (2002)), as suggested by the following class related weights for credit scoring: (i) asymmetric misclassification costs (of false positive and false negatives), and (ii) unequal proportions of class representatives in samples and the populations. A method adapted to this non standard situations would reduce the e.g. notoriously high number of credit applicants turned down by a bank.

In fact, knowledge about the geometry of labeled credit data sets is very poor. What efficient and general method, potentially superior to linear approaches, can be used? A good candidate for classifying high dimensional nonlinear and relatively sparse data are Support Vector Machines (Schölkopf and Smola (2002)). Stecking and Schebesch (2003) show for real life credit

scoring data, that “standard” SVM are superior to traditional linear and logistic regression. But does this carry over to the non standard situation of asymmetric misclassification costs and unequal sample / population class representatives? For the same data, a logistic regression with optimized cut off proved superior to the standard SVM. But with minor modifications to a standard SVM the results improve dramatically, outperforming logistic regression with cut off significantly. In order to further improve real world applicability of the SVM, sigmoid output mapping is used to derive default probabilities. They refine the credit scoring predictions and they can be used for credit rating and thus for establishing more “personalized” terms and conditions for credit contracts.

2 From standard to non standard classification

Classification for credit scoring uses the data of $N > 0$ randomly supplied past credit applicants as training cases $\{x_i, y_i\}$, with $i = 1, \dots, N$. Input vector $x_i \in \mathcal{R}^d$ describes credit applicant i and output or class label $y_i \in [-1, 1]$ records whether credit applicant i was “bad”(defaulting, $y_i = +1$) or “good”(non-defaulting, $y_i = -1$). SVM build a decision rule directly from such data, i.e. without using information from the joint probability $p(x, y)$ prior to learning.

In order to understand the action of the standard nonlinear SVM, first consider a map $\phi : \mathcal{R}^d \rightarrow \mathcal{F}$, which “lifts” data points from input space into a higher dimensional (potentially infinite dimensional) feature space \mathcal{F} . With map ϕ one trades complicated patterns in input space for simpler but higher dimensional patterns in feature space. Following Schölkopf and Smola (2002) and Stecking and Schebesch (2003), the “hard margin” SVM can **linearly separate** any labeled data set via ϕ from the primal optimization problem $\min_{w, b} \left\{ \frac{1}{2} w' w \mid y_i(w' \phi(x_i) + b) - 1 \geq 0, \text{ for } i = 1, \dots, N \right\}$. Over-fitting is controlled by allowing for misclassification, i.e. by using $C > 0$ and ζ in the “soft margin” primal $\min_{w, b, \zeta} \left\{ C e' \zeta + \frac{1}{2} w' w \mid y_i(w' \phi(x_i) + b) \geq 1 - \zeta_i, \text{ for } i = 1, \dots, N \right\}$, with unit vector $e = (1, 1, \dots, 1)$ and slacks $\zeta_i \geq 0$. Note that unknown w and ϕ have the same dimension as \mathcal{F} ! Using $w = \sum_{i=1}^N y_i \alpha_i \phi(x_i)$, which follows from the first order optimality conditions, the associated dual program reads $\max_{\alpha} \left\{ e' \alpha - \alpha' Y K Y \alpha \mid e' Y \alpha = 0, 0 \leq \alpha_i \leq C, \text{ for } i = 1, \dots, N \right\}$, with duals $\alpha_i \geq 0$, output matrix Y , with $Y_{ii} = y_i$ and $Y_{ij} = 0, i \neq j$, and kernel matrix K , with $K_{ij} = k(x_i, x_j) \geq 0$ functions of input pairs, which contain ϕ implicitly. In this paper we use RBF kernels $k(u, v) = \exp(\frac{-||u-v||^2}{\sigma^2})$, with $\sigma > 0$. RBF are not biased by location, i.e. $k(u, u) = 1$, for every u , and their superposition can model locally non-linear boundaries between classes, which might exist in credit scoring data (Schebesch and Stecking (2003)). A **support vector** is a data point x_i with associated optimal $\alpha_i > 0$. After

selecting C and σ , the resulting support vectors, with $C > \alpha_i > 0$, compactly describe the decision margins between the classes. Standard SVM produce a decision function (left expression) associated with an idealized Bayesian probabilistic point of view (Lin (1999), rhs. expression)

$$\text{SIGN}\left(\sum_{i=1}^N y_i \alpha_i k(x, x_i) + b\right) \text{ or, for } N \rightarrow \infty, \quad \text{SIGN}\left(p(y=1|x) - \frac{1}{2}\right),$$

which implicitly assumes equal costs of predicting each case (and each class), and which also assumes approximately the same number of class representatives for “in sample” and for “out of sample” data. Variables x, y of the rhs. expression now stand for “possible” as opposed to “given” data.

Non Standard SVM is slightly modified in order to incorporate **prior class weights** $L(y_i) > 0$ on the training data. Lin, Lee and Wahba (2002) show that an application dependent cut-off in the idealized probabilistic view of the SVM decision rule, i.e.

$$\text{SIGN}\left(p(y=1|x) - \frac{L(-1)}{L(-1) + L(1)}\right),$$

merely implies $0 \leq \alpha_i \leq CL(y_i)$ as a change to the N restrictions of the above dual program. Computation of the optimal b is also slightly adjusted. In SVM implementations like *SVM-Light* from the GMD, a control parameter $L = L(1)/L(-1)$ accounts for the misclassification costs of the two classes.

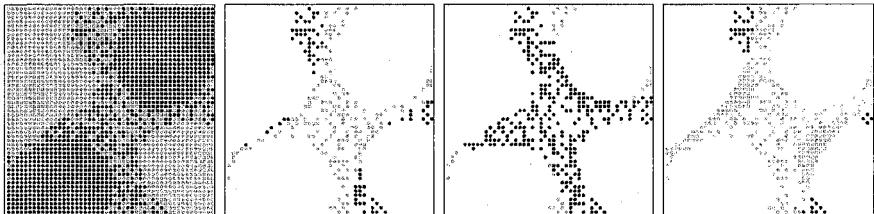


Fig. 1. Example of nonlinear binary classification problem with two dimensional inputs (left). Positives are marked with dark and negatives with light circles. Coordinate axes are the input dimensions. Standard SVM (middle left): false negatives (dark circles) and false positives (light circles). Using the same nomenclature, the results of non standard SVM assigning a higher cost to false positives (middle right) and, respectively, to false negatives (right) is shown.

Figure 1 depicts the effect of different class weights on the SVM. The detailed effect of class weight variation is a non trivial change in the correctly classified cases. Hence, class weights are an “aggregate” parameter to influence emerging class boundaries, different from that of C (overall misclassification) and that of σ (local nonlinearity). The overall effect of class weights is to increase cost related performance, mainly at the expense of cost neutral prediction accuracy, i.e. the result of the standard SVM (middle left).

3 SVM for credit scoring

3.1 Non standard situation

We use data of 658 credit applicants sampled from 17158 credit applicants of a population from a German building and loan association. While the sampled data are available in full detail, information about population data is only summary (distributional). Depending on historical credit performance, persons are grouped into “defaulting” (class label $y_i = 1$) and “non defaulting” ($y_i = -1$). Sixteen input variables describing the credit applicant from the sample are used for model building. In Stecking and Schebesch (2003) it was already shown, that in the case of “neutral cost” credit classification standard SVM outperforms traditional methods like linear discriminant analysis and logistic regression. However, the data distributions of sample and population are extremely different. While in the true (population) data the share of defaulting applicants amounts to only 6.70%, the in-sample share of defaulting and non defaulting applicants is of nearly equal size. Furthermore, there are different costs for different types of misclassification. A “bad accepted” credit applicant causes much higher misclassification costs than a “good rejected”, because up to the whole amount of credit can be lost when accepting a “bad”, while only the interest payments are lost when rejecting a “good”. According to our cooperation partner, a cost relation between “bad accepted” and “good rejected” of 5 : 1 would be adequate.

Non standard SVM dealing with such situations require just one additional parameter (see previous section): a cost factor L describing the relation between the two types of misclassification cost and the proportions of class membership in sample and population. How can we initialize this cost parameter? The answer is based on the modified Bayesian rule for the idealized probabilistic view of the SVM decision rule from the previous section. Define $L = L(1)/L(-1) = (c^- \pi^+ \pi_s^-)/(c^+ \pi^- \pi_s^+)$ with c^+ the cost of false positives (“good rejected”) and c^- the cost of false negatives (“bad accepted”) respectively (Lin et al. 2002). The parameters π^+ (π^-) and π_s^+ (π_s^-) denote the proportion of defaulting (non defaulting) in population and sample. By using $c^+ = 1$, $c^- = 5$, $\pi^+ = 0.0670$, $\pi^- = 0.9330$, $\pi_s^+ = 0.4909$ and $\pi_s^- = 0.5091$ one gets an initial L of 0.3723 (L needs tuning, see impact of L in fig. 1).

Our SVM classification model uses radial basis function kernels with fixed kernel width parameter $\sigma = 0.05$ (see previous section and Stecking and Schebesch (2003)). This parameter controls the width of the area of strong activation of a radial basis function around its center vector in input space. To optimize the SVM with regard to expected classification performance, grid search over varying penalty terms C and over varying cost factors L is employed. The best results are observed at $C = 4$ and $L = 0.30$. A cost factor of $L < 1$ typically allows for bigger misclassification error on defaulting applicants. For a first inspection, a ROC (receiver operating characteristics) curve is shown, comparing the results of standard to non standard SVM.

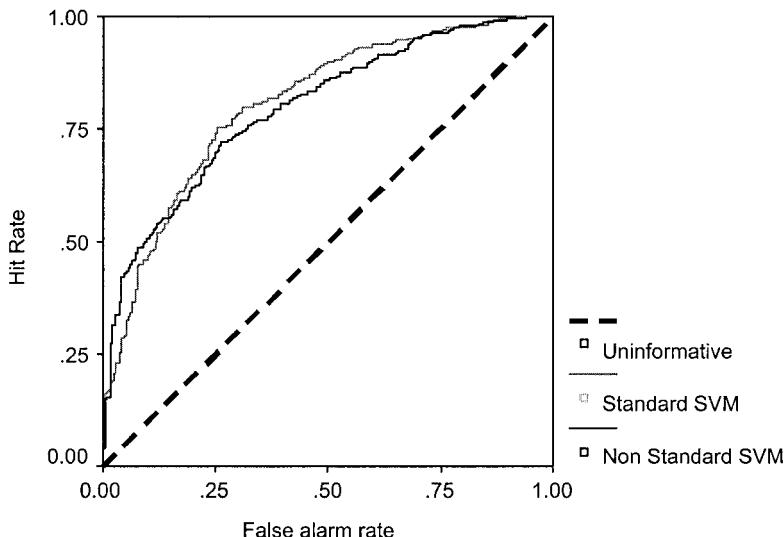


Fig. 2. ROC curves of standard vs. non standard SVM.

The first step in computing a ROC curve is arranging the SVM results in descending order. The ordinate of the ROC curve denotes the hit rate (= the proportion of rejected bad compared to all bad applicants). The abscissa denotes the false alarm rate (= the proportion of rejected good compared to all good applicants). In this way, the overall ability of the models to separate "good" and "bad" cases can be measured regardless of the cut off values, namely as the deviation of the model ROC curve from a "pure chance" performance line (the "uninformative" diagonal). In figure 2 no substantial difference in the overall separation ability of the models can be observed. However, for a small accepted false alarm rate (less than 20%, say) the hit rate is consistently higher for the non standard SVM.

With ROC curves crossing, neither model dominates the other in all cases. In order to evaluate real world model performance, the results of a tenfold cross validation were randomly re-sampled exposing the trained SVM to the true number of defaulting and non defaulting applicants in the population. Subsequently, by resampling one thousand times, the variability of the model results is estimated. Resampling techniques have been proven to give an accurate approximation of the true population variability if the original sample was drawn at random. Otherwise, the results of the resampling process cannot readily be extended to the full population, but, resampling can at least be used to estimate the variability of model performance (Stein (2002)).

Table 1 shows the tenfold cross validation results after resampling. The costs of misclassification are calculated as five times "bad accepted" plus

Tenfold Cross Validation (Resampling)						
Model	Statistic	Good rejected	Bad accepted	Bad rejected	Good accepted	Cost of misclassification
SVM						
Non Stand.						
	Median	668	665	484	15341	3996
	Lower 5%	625	637	457	15298	3847
	Upper 5%	711	692	512	15384	4138
SVM						
Standard						
	Median	3824	323	826	12185	5442
	Lower 5%	3732	298	799	12091	5282
	Upper 5%	3918	350	851	12277	5599
Logistic Regression						
	Median	1815	506	643	14194	4344
	Lower 5%	1695	447	593	14075	4196
	Upper 5%	1934	556	702	14314	4502

Table 1. Credit scoring model performance and cost of misclassification.

“good rejected”. This is our final measure to compare model performance. In terms of misclassification costs the non standard SVM clearly outperforms the standard SVM. To add a more traditional benchmark, logistic regression with cut off optimization (Stecking (2003)) was also used. Logistic regression was shown to give comparable (or even slightly better) credit scoring results than linear discriminant analysis (Stecking and Schebesch (2003)). Logistic regression with optimized cut off is superior to standard SVM but is significantly inferior to non standard SVM. The lower 5% of the misclassification costs of logistic regression are even higher than the upper 5% of the non standard SVM. For both standard and nonstandard SVM, roughly one out of six input patterns are essential support vectors (i.e. $C > \alpha_i > 0$, see previous section). This fairly high number of support vectors does not signal “poor expected generalization”! Owing to nonlinearity in the credit scoring data, these support vectors are needed to construct the decision rule in feature space.

3.2 Default probabilities

Default probabilities are very important to refine credit scoring, especially for constructing a system of rating classes. The output of the SVM is an uncalibrated real value which is not a probability, but there are several ways to derive posterior class probabilities from the SVM (Kwok (1999)), (Platt (1999)):

By using a sigmoid logit function, with function values bounded between 0 and 1, each SVM output $s(x)$ (i.e. the value which enters the sign function of the decision rule) can be mapped into a probability. Two approaches of output mapping are compared (1) the ad-hoc probability estimation by Kwok (1999): $p(y = 1|x) = \frac{\exp(-1 + s(x))}{\exp(-1 + s(x)) + \exp(-1 - s(x))}$, and (2) the fitted sigmoid proposed by Platt (1999): $p(y = 1|x) = \frac{1}{1 + \exp(as(x) + b)}$, with a and b to be optimized. For example, an SVM output of zero would lead to an ad-hoc probability of 0.5. Output values lower (higher) than zero lead to probabilities beneath (above) 0.5. The ad-hoc probability estimation is a special case of the fitted sigmoid, with $a = -2$ and $b = 0$. Fitted sigmoid can be used to construct more accurate probability distributions, especially useful in non standard situations of classification. The parameters a and b can then be estimated by maximum likelihood. The output of the non standard SVM of the previous section is used to compare both approaches.

Observed vs. estimated default probabilities					
SVM-Output Percentiles	No. of applicants	Defaulting applicants	Def. prob. observed	Def. prob. SVM-Ad-hoc	Def. prob. SVM-Sigmoid*
0 – 10%	1680	7	0.42%	0.66%	0.58%
10 – 20%	1742	21	1.23%	1.41%	1.18%
20 – 30%	1712	39	2.29%	2.14%	1.75%
30 – 40%	1685	60	3.59%	2.83%	2.30%
40 – 50%	1726	53	3.09%	3.85%	3.11%
50 – 60%	1744	71	4.08%	5.00%	4.04%
60 – 70%	1692	68	3.99%	6.60%	5.40%
70 – 80%	1730	153	8.84%	9.12%	7.70%
80 – 90%	1705	128	7.51%	12.73%	11.45%
90 – 100%	1743	548	31.44%	22.35%	29.05%
Total	17158	1149	6.70%	6.70%	6.70%

* with $a = -1.84$ and $b = 1.47$

Table 2. Observed vs. estimated default probabilities.

Table 2 describes the derived probabilities by calculating the k -percentiles of the original SVM output (with k running from 10%, 20% to 90%) to give the borders for ten approximately equally sized intervals of the (weighted) input data. For each interval, the share of observed defaulting applicants is computed and compared to the two alternative mean probability estimations in each interval. Both probability estimates from table 2 are conforming quite well, with larger differences between observation and prediction in the two riskiest classes 80 – 90% and 90 – 100% only. Furthermore, global measures

with regard to the quality of probability calibration are (*i*) mean squared error (Brier score: $B = 1/N \sum_i (p_i - (y_i + 1)/2)^2$) and (*ii*) log likelihood measure ($LL = \sum_{y_i=1} \ln(p_i) + \sum_{y_i=-1} \ln(1 - p_i)$). For both calibration measures SVM-Sigmoid is slightly superior to SVM-Ad-hoc (Brier score 0.0531 against 0.0548 and log likelihood measure -3423.22 against -3497.62).

4 Conclusion

Using some minor modification of the standard SVM dramatically improves cost related performance of out of sample classification. This is especially useful for practitioners from the credit banking sector, which can use the non standard SVM for successfully handling asymmetries in misclassification costs, in sub-sample sizes or in class-wise risk assumptions. Clear superiority of the non standard SVM is established against the otherwise well performing logistic regression with optimized cut off. The robustness of the results is extensively checked by bootstrapping and cross validation. Moderation of the real valued SVM output in order to obtain default probabilities is also worthwhile. They are useful for refining credit scoring by constructing rating systems – a step towards more “personalized” credit contracts.

References

- KWOK, J.T. (1999): Moderating the Outputs of Support Vector Machine Classifiers. *IEEE Transactions on Neural Networks*, 10, 5, 1018–1031.
- LIN, Y. (1999): Support Vector Machines and the Bayes rule in classification. Technical Report 1014, Dept. of Statistics, University of Wisconsin.
- LIN, Y., LEE, Y., and WAHBA, G. (2002): Support Vector Machines for Classification in Nonstandard Situations. *Machine Learning*, 46, 1-3, 191–202.
- PLATT, J.C. (1999): Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: A.J. Smola, P. Bartlett, B. Schölkopf, and C. Schuurmans (Eds.): *Advances in Large Margin Classifiers*, MIT Press, Cambridge, MA.
- SCHEBESCH, K.B. and STECKING, R. (2003): Support Vector Machines for Credit Applicants: Detecting Typical and Critical Regions. In: *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, MA.
- STECKING, R. (2003): Credit Scoring im Baukreditwesen, in H. Schaefer (Ed.): *Kredit und Risiko: Basel II und die Konsequenzen für Banken und Mittelstand*, Metropolis-Verlag, Marburg, 45–56.
- 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.
- STEIN, R.M. (2002): Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation. Technical Report 020305, Moody's KMV, New York.

Discovery of Risk-Return Efficient Structures in Middle-Market Credit Portfolios

Frank Schlottmann^{1,2} and Detlef Seese¹

¹ Institut für Angewandte Informatik und Formale Beschreibungsverfahren,
Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

² GILLARDON AG financial software, Research, Alte Wilhelmstr. 4,
D-75015 Bretten, Germany

Abstract. We discuss a hybrid approach that combines Multi-Objective Evolutionary Algorithms and quantitative methods of portfolio credit risk management to support the discovery of downside risk-return efficient structures in middle-market credit portfolios. In an empirical study, we compare the performance of the solutions discovered by our hybrid method to the solutions found by a corresponding non-hybrid algorithm on two different real-world loan portfolios.

1 Introduction

The management of portfolio credit risk has recently attracted many research activities both in academic and financial institutions. This is caused by the fact that there has been a steadily increasing number of bankruptcies in many countries due to the economic downturn. Moreover, there has been an intensive development of new methods for portfolio credit risk management (cf. e. g. Crouhy et al. (2000) for an overview), and the banking supervision authorities have recently proposed new supervisory rules which lead to new constraints for investors that hold credit exposures in their portfolios, cf. e. g. Basel Committee for Banking Supervision (2001).

In this setting, the discovery of risk-return efficient credit portfolio structures with respect to constraints is essential for many financial institutions, particularly for investors holding middle-market portfolios¹. This is e. g. the case for many German universal banks. A typical real-world problem arising from this task is summarized as follows:

Given is a bank which holds a middle-market portfolio of $m > 1, m \in \mathbb{N}$ obligors in $t = 0$ and has a fixed time horizon $T \in \mathbb{R}_+$.

Each obligor $i \in \{1, \dots, m\}$ is subject to the risk of default and it is characterized by the following data which is considered to be constant within the time period $[0, T]$: net exposure $e_i \in \mathbb{R}_+$ (loss given default of obligor i), expected payoff rate $r_i \in \mathbb{R}$ based on e_i (net of funding cost), expected default probability $p_i \in \mathbb{R}_+$, capital budget requirement $w_i \in \mathbb{R}_+$ based on e_i .

¹ These portfolios typically consist of obligors which are small- and medium-size enterprises not having direct access to the capital market.

Moreover, we assume that the bank has a given capital budget $B > 0$ (e. g. supervisory capital limit) and the possible portfolio structures are described by a vector of binary decision variables $x := (x_1, x_2, \dots, x_m)^T$, $x_i \in \{0, e_i\}$. The binary decision about holding the net exposure in the portfolio or not reflects the decision about hedging the credit risk, for instance by buying default protection or by transferring the net exposure to the capital market.

Let $F_x(X)$ denote the cumulative probability distribution function of the random variable X representing the aggregated losses from the portfolio for a given portfolio structure x . For our empirical study, we assume that the investor uses the CreditRisk+ model described in CreditSuisse Financial Products (1997) to approximate $F_x(X)$ and measures the aggregated downside risk for a given portfolio structure x by evaluation of the Credit-Value-at-Risk (CVaR) at the arbitrary, but fixed confidence level $\alpha \in (0, 1)$:

$$\text{risk}(x) := F_x^{-1}(\alpha) - E[X] \quad (1)$$

where $F_x^{-1}(\alpha)$ is the α -percentile of F_x and $E[X]$ is the expectation of X . This is a standard measure of unexpected loss (i. e. downside risk) in many real-world applications, cf. e. g. Gordy (2000).

The expected payoff (i. e. expected return in monetary units) of a portfolio structure x is given by

$$\text{exppay}(x) := \sum_{i=1}^m r_i x_i - \sum_{i=1}^m p_i x_i = \sum_{i=1}^m (r_i - p_i) x_i. \quad (2)$$

Note that both the $\text{risk}()$ and the $\text{exppay}()$ function require more arguments which we omitted for notational convenience since the only parameters which are modified are the x_i decision variables.

We assume that the search space of possible portfolio structures is restricted by constraints such that we have to search for feasible solutions. In our empirical study, we assume that a portfolio structure x is feasible if it satisfies the following budget constraint:

$$\sum_{i=1}^m x_i w_i \leq B \quad (3)$$

Moreover, a portfolio structure x dominates another portfolio structure $y \neq x$ if and only if one of the following cases is true:

$$\text{exppay}(x) > \text{exppay}(y) \wedge \text{risk}(x) \leq \text{risk}(y) \quad (4)$$

$$\text{exppay}(x) \geq \text{exppay}(y) \wedge \text{risk}(x) < \text{risk}(y) \quad (5)$$

If x dominates y , we will denote this relationship by $x >_d y$.

Now we can formulate the bank's problem of the discovery of a diverse set of feasible downside risk-return efficient portfolio structures as follows:²

² Cf. the work of Markowitz (1952), but we use a different risk measure and discrete decision variables here.

Find the largest set of non-dominated feasible portfolio structures for the given data, denoted by

$$PE^* := \{x \mid x \text{ is feasible} \wedge \forall y : y \text{ is feasible} \Rightarrow \neg(y >_d x)\} \quad (6)$$

where x and y are again two different portfolio structures.

In Seese and Schlottmann (2002) we have shown that the computation of PE^* is NP-hard if the parameters (excluding the decision variables) are given as rational numbers, therefore we consider an approximation algorithm in the following section. Since particularly the *risk()* objective function has a non-linear and non-convex structure (see e. g. Pflug (2000) for the mathematical properties of such downside risk measures) and we have to deal with integrality conditions as well as further constraints we propose a heuristic approach.

2 A Hybrid Multi-Objective Evolutionary Approach

In the literature, Multi-Objective Evolutionary Algorithms (MOEAs) were successfully applied to many constrained multi-objective problems having a difficult structure, see e. g. Osyczka (2002) for a detailed discussion of design problems and the advantages of MOEAs concerning the discovery of feasible non-dominated solutions in non-convex problem settings. We have opted for a hybrid approach which combines elements from Multi-Objective Evolutionary Algorithms and a problem-specific local search algorithm to benefit both from the strengths of different MOEA concepts and the convergence speed of gradient-based local search. The basic scheme of this Hybrid Multi-Objective Evolutionary Algorithm (HMOEA) is shown below:

Input: Credit portfolio data as specified in section 1

$t := 0$

Generate initial population $P(t)$ by random initialisation

Initialise elite population $Q(t) := \emptyset$

Evaluate $P(t)$

Repeat

Select individuals from $P(t)$

Recombine selected individuals (variation operator 1)

Mutate recombined individuals (variation operator 2)

Apply local search to mutated individuals (variation operator 3)

Create offspring population $P'(t)$

Evaluate joint population $J(t) := P(t) \cup P'(t)$

Update elite population $Q(t)$ from $J(t)$

Generate $P(t+1)$ from $J(t)$

$t := t + 1$

Until $t > t_{max} \vee Q(t) = Q(\max \{0, t - t_{diff}\})$

Output: $Q(t)$

In addition to the usual population denoted by $P(t)$ we propose the use of an elite population $Q(t)$ in our algorithm that contains the feasible non-dominated solutions found so far at each population step t . Besides the almost sure convergence of such an elite population to PE^* for $t \rightarrow \infty$ shown by Rudolph and Agapie (2000), a further advantage is that the population $P(t)$ can be kept quite small since it does not need to store a lot of solutions during the search for new solutions.

The evaluation of $P(t)$ and $J(t)$ is based on the following relation proposed in Deb (2001), p. 288: Given are two distinct portfolio structures x and y . x constraint-dominates y if and only if one of the following cases is true:

$$\begin{aligned} x \text{ is feasible } &\wedge y \text{ is feasible } \wedge \text{exppay}(x) > \text{exppay}(y) \wedge \text{risk}(x) \leq \text{risk}(y) \quad (7) \\ x \text{ is feasible } &\wedge y \text{ is feasible } \wedge \text{exppay}(x) \geq \text{exppay}(y) \wedge \text{risk}(x) < \text{risk}(y) \quad (8) \end{aligned}$$

$$x \text{ is feasible } \wedge y \text{ is infeasible} \quad (9)$$

$$x \text{ is infeasible } \wedge y \text{ is infeasible } \wedge \sum_{i=1}^m x_i w_i < \sum_{i=1}^m y_i w_i \quad (10)$$

The non-dominated sorting procedure in our HMOEA uses this dominance criterion to classify the solutions in a given population, e. g. $P(t)$, into different levels of constraint-domination. The best solutions which are not constraint-dominated by any other solution in the population, obtain fitness value 1 (best rank). After that, only the remaining solutions are checked for constraint-domination, and the non-constraint-dominated solutions among these obtain fitness value 2 (second best rank). This process is repeated until each solution has obtained an associated fitness rank.

The selection operator is performed using a binary tournament: Two individuals x and y are randomly drawn from the current population $P(t)$, using uniform probability of $p_{sel} := \frac{1}{|P(t)|}$ for each individual. These individuals are checked for constraint-domination by comparison of their fitness values (ranks) and if, without loss of generality, x dominates y then x wins the tournament and is considered for reproduction. Tournament selection is considered to be favorable over other selection schemes, see e. g. Osyczka (2002) for a discussion of this topic.

The first variation operator applied to the selected individuals is the standard one-point crossover for discrete decision variables, i. e. the gene strings of two selected individuals are cut at a randomly chosen position and the resulting tail parts are exchanged with each other to produce two new offspring. This operation is performed with crossover probability p_{cross} on individuals selected for reproduction. In analogy to natural mutation, the second variation operator changes the genes of selected individuals randomly with probability p_{mut} (mutation rate) per gene to allow the invention of new, previously undiscovered solutions in the population and to prevent stalling into local optima.

Our third variation operator represents a problem-specific local search procedure that is applied with probability p_{local} to each selected solution x after crossover and mutation. Its working principle is shown below:

Input: $P(t)$, $Q(t)$, credit portfolio data as specified in section 1
For each $x \in P(t)$ apply the following instructions with probability p_{local}

If x is feasible Then $D := -1$
Else Choose D between $D := 1$ or $D := -1$ with uniform probability 0.5
Initialisation $\forall i : \hat{x}_i := x_i$
 $Step := 0$
Do

Copy $\forall i : x_i := \hat{x}_i$
 $expay_{old} := expay(x)$, $risk_{old} := risk(x)$
For each x_j calculate the partial derivatives $d_j := \frac{\partial}{\partial x_j} \left(\frac{expay(x)}{risk(x)} \right)$
If $D = -1$ **Then**
Choose the minimal gradient component $i := \arg \min_j \{d_j | x_j > 0\}$
Remove this obligor from the current portfolio structure: $\hat{x}_i := 0$
Else
Choose the maximal gradient component $i := \arg \max_j \{d_j | x_j = 0\}$
Add this obligor to the current portfolio structure: $\hat{x}_i := e_i$
End If
 $expay_{new} := expay(\hat{x})$, $risk_{new} := risk(\hat{x})$
 $Step := Step + 1$
While ($Step < Step_{max}$) $\wedge (\exists i : \hat{x}_i > 0) \wedge (\exists j : \hat{x}_j = 0) \wedge \hat{x} \notin P(t) \wedge$
 $\hat{x} \notin Q(t) \wedge ((D = -1 \wedge \hat{x} \text{ is infeasible}) \vee (D = 1 \wedge \hat{x} \text{ is feasible})) \wedge$
 $(expay_{new} > expay_{old} \vee risk_{new} < risk_{old}))$
Replace x in $P(t)$ by its optimised version
End For
Output: $P(t)$

This operator moves a given solution x into a direction which depends on the feasibility of the current solution at the start of the local search and on the elements of the gradient of the function relating the expected payoff of the current portfolio structure to its downside risk.³ A solution that is infeasible at the start of the local search is always moved into the direction of feasible solutions by removing obligors from the current solution. In the case of a start with a feasible solution, the choice about addition or removal of obligors from the current portfolio structure is made randomly to avoid the local search always stalling into the same local optima. The conditions for continuing the **Do...While** loop are violated if the maximum number of $Step_{max} \in \mathbb{N}$ local

³ From an economic point of view this rational function can be interpreted as a risk-adjusted performance measure, see e. g. Ong (1999), p. 215 ff. for a discussion of such measures.

search optimization steps has been reached or if the current solution can be modified further. Moreover, the loop is also terminated if a solution that was infeasible at the start of the local search variation operator execution has been made feasible, or if a feasible solution has not been improved during the current iteration.

We consider the hybridisation of the MOEA using this additional variation operator to be a significant improvement compared a standard, non-hybrid MOEA since the randomised search process of the MOEA can be guided a bit more towards the set of feasible non-dominated solutions. Therefore, such a local search operator improves the convergence speed of the overall algorithm towards the desired solutions, while the hybrid approach does not suffer from local optima problems due to the MOEA benefits (e. g. mutation). In addition to these arguments, e. g. the CreditRisk+ portfolio credit risk model provides additional local structure information for a current solution x beyond the objective function values that can be exploited in linear computing time $O(m)$ measured by the number of obligors.

By application of the variation operators to the selected individuals we obtain an offspring population $P'(t)$. The members of the joint population $J(t) := P(t) \cup P'(t)$ are evaluated using the non-dominated sorting procedure described above. In the next step, the elite population $Q(t)$ is updated by comparing the feasible non-dominated solutions from $J(t)$ to those in $Q(t)$.

Before finishing the population step t and setting $t := t + 1$ the members of the new parent population $P(t+1)$ have to be selected from $J(t)$. Because elitist EAs, which preserve the best solutions from both parents and offspring, usually show better convergence properties, we also use this mechanism in our algorithm. Besides elitism, we also need a diversity preserving concept to achieve a good distribution of the solutions in the whole objective space. We incorporate the concept of crowding-sort proposed in Deb (2001), p. 236 mainly because it does not need additional parameters.

The algorithm can be terminated if a maximum number of t_{max} population steps has been performed or if $Q(t)$ has not been improved for a certain number t_{diff} of population steps. Besides this stopping criterion for practical applications we use an alternative stopping criterion in our empirical study described in the next section that terminates the algorithm if given, fixed number of target function calls (denoted by t_{target}) is exceeded. This allows a direct comparison of the results between different algorithms.

3 Empirical results

We applied our hybrid algorithm to two middle-market loan portfolios provided by a German bank. The first portfolio (denoted by port25) consisted of 25 obligors and the second portfolio port386 contained 386 obligors. As mentioned in the introduction, we assume a constrained search space, i. e. a restricted capital budget B in the respective test case. We set $p_{cross} := 0.95$

and $p_{mut} := \frac{1}{m}$ in all tests which is a quite common parameter setting for elitist Evolutionary Algorithms. For the local search variation operator, we set the parameters to $p_{local} := 0.05$ and $Step_{max} := 4$ which yields a good trade-off between additional computational effort and resulting higher solution quality for many test portfolios (not only in the tests presented here).

Figure 1 shows a comparison between PE^* obtained by a complete enumeration of the search space which required more than 4 hours, and an approximation set obtained by a single run of the HMOEA within 40 seconds (stopping criterion: $t_{target} := 50000$ target function calls exceeded) on the reference PC. By visual inspection we observe that the approximation set

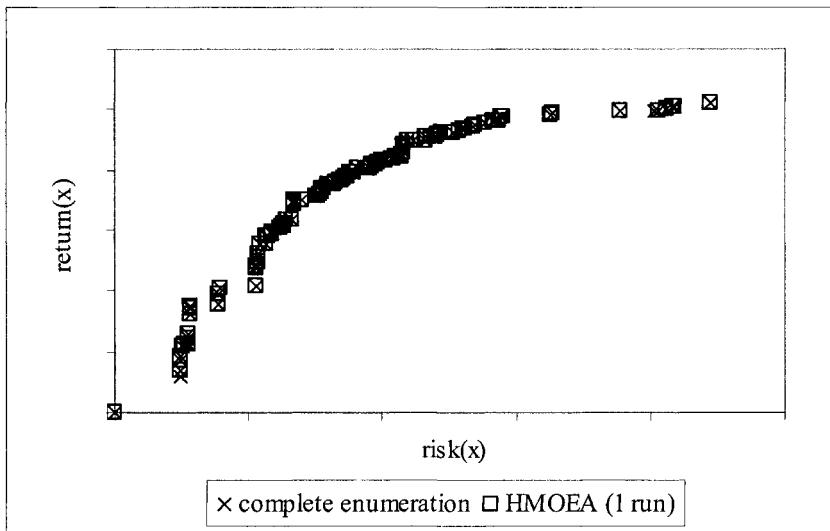


Fig. 1. Complete enumeration versus HMOEA output for port25

indicated by squares is very close (and in many points even identical) to the set of feasible non-dominated solutions (indicated by 'x' points) while being computed in a very small fraction of the upper computational bound determined by the enumeration.

Moreover, we tested the HMOEA against a benchmark algorithm which is its MOEA counterpart without the local search variation operator to check the benefits of hybridisation. We performed 50 independent runs of the respective algorithm on each test problem and stopped the respective run after exceeding $t_{target} := 50000$ target function calls for port25 and $t_{target} := 500000$ target function calls for port386 (due to the larger search space). In the following text, we compare the resulting approximation sets: The set PE_1 contains the solutions discovered by the non-hybrid MOEA in a single run, and PE_2 is the corresponding approximation set discovered by

the HMOEA. We check both the cardinality of the approximation sets (more solutions are desirable) and the solution quality concerning their domination (less domination is preferable) as well as concerning their maximum spread in the objective function space (larger spread is better).

criterion	mean			std. deviation		
	PE_1	PE_2	rel. chg.	PE_1	PE_2	rel. chg.
# solutions	114.82	114.62	-0.17%	2.49	2.31	-7.11%
# dominated solutions	3.08	1.82	-40.91%	2.28	1.49	-34.59%
# dominated solutions # solutions	2.67%	1.58%	-40.67%	1.96%	1.43%	-26.90%
normalized maximum spread	2.97	3.16	+6.55%	0.17	0.05	-71.22%
runtime [seconds]	39.38	38.66	-1.83%	0.70	0.63	-10.10%

Table 1. Results for port25

Table 1 indicates that on average, both approximation sets contain about the same number of solutions for port25, but PE_2 contains significantly less dominated solutions compared to PE_1 (we measured significant average absolute differences in favour of the HMOEA at the 99% confidence level). Hence, the relation between these two criteria is better for PE_2 . Moreover, PE_2 has a larger possible Euclidean distance between any pair of solutions in the objective space on average (indicated by normalized maximum spread) which is again statistically significant at 99% confidence concerning the mean absolute differences. Note that the runtime of both algorithms is quite similar. Moreover, the lower standard deviations indicate that the results of the HMOEA are less volatile compared to the results found by the non-hybrid MOEA which is a desired feature of a stochastic search algorithm.

criterion	mean			std. deviation		
	PE_1	PE_2	rel. chg.	PE_1	PE_2	rel. chg.
# solutions	1312.46	1486.50	+13.26%	58.84	49.66	-15.60%
# dominated solutions	1194.96	77.06	-93.55%	92.36	67.13	-27.32%
# dominated solutions # solutions	91.05%	5.18%	-94.31%	5.72%	4.44%	-22.37%
normalized maximum spread	1.85	1.91	+3.14%	0.07	0.07	+6.98%
runtime [seconds]	1258.08	1319.18	+4.86%	12.87	11.24	-12.71%

Table 2. Results for port386

The results for port386 shown in Table 2 reveal interesting differences between both algorithms in the larger search space: On average, PE_2 contains more solutions, less dominated solutions, and a larger spread of solutions (again the respective mean absolute difference is significant in favour of HMOEA at 99% confidence). Compared to the improvement of the solution

quantity and quality, the mean additional runtime required due to hybridisation is small. Except for the maximum spread, the performance measures again have a smaller standard deviation.

4 Conclusion

We discussed a hybrid approach supporting the discovery of risk-return efficient credit portfolio structures which combines MOEA-based search and problem-specific methods to obtain a faster discovery of promising solutions while not suffering from local optima. In our empirical study, the hybrid approach yielded a higher average solution quality and almost always less solution volatility compared to a non-hybrid MOEA. Moreover, the computation time required was reasonably small due to the integration of computationally efficient calculations within the CreditRisk+ portfolio credit risk model. We consider our results are a promising basis for further research concerning the integration of problem-specific knowledge into flexible problem solving concepts like MOEAs to obtain similar results for other complex problems in the area of credit risk and also in other financial problem contexts.

The authors would like to thank an anonymous German bank for providing the real-world portfolio data.

References

- BASEL COMMITTEE FOR BANKING SUPERVISION (2001): *The new Basel capital accord*. Bank for International Settlements, Basel.
- CROUHY, M., GALAI, D., and MARK, R. (2000): A comparative analysis of current credit risk models. *Journal of Banking and Finance* 24, 59–117.
- CREDITSUISSE FINANCIAL PRODUCTS (1997): CreditRisk+ (TM) - a credit risk management framework. www.csfp.co.uk/creditrisk/assets/creditrisk.pdf.
- DEB, K. (2001): *Multi-objective optimisation using evolutionary algorithms*. John Wiley & Sons, Chichester.
- GORDY, M. (2000): *Credit VaR models and risk-bucket capital rules: a reconciliation*. Proc. of 36th Annual Conference on Bank Structure and Competition, Federal Reserve Bank of Chicago.
- MARKOWITZ, H. (1952): Portfolio selection. *Journal of Finance* 7, 77–91.
- ONG, M. (1999): *Internal credit risk models*. Risk Books, London.
- OSYCZKA, A. (2002): *Evolutionary Algorithms for single and multicriteria design optimization*. Physica, Heidelberg.
- PFLUG, G. (2000): Some remarks on the value-at-risk and the conditional value-at-risk. In: S. Uryasev (Ed.): *Probabilistic constrained optimization*. Kluwer, Dordrecht, 272–281.
- RUDOLPH, G. and AGAPIE, A. (2000): Convergence properties of some multi-objective evolutionary algorithms. In: A. Zalzala (Ed.): *Proceedings of the 2000 Congress on Evolutionary Computation*. IEEE Press, Piscataway, 1010–1016.
- SEESE, D. and SCHLOTTMANN, F. (2002): The building blocks of complexity: a unified criterion and selected problems in economics and finance. Sydney Financial Mathematics Workshop 2002, www.qgroup.org.au/SFMW.

Approximation of Distributions of Treasury Bill Yields and Interbank Rates by Means of α -stable and Hyperbolic Distributions*

Witold Szczepaniak

Department of Financial Investments and Insurance,
Wroclaw University of Economics,
ul. Komandorska 118/120, 53-342 Wroclaw, Poland

Abstract. In this paper α -stable and hyperbolic distributions are presented and proposed as alternatives to the normal distribution in approximation of treasury bill yields and interbank rates distributions.

1 Introduction

The analysis of distributions of treasury bill yields and interbank rates shows that they have kurtosis greater than 3 and tails heavier tails than the normal distribution, see for instance Das (2002), Johannes (2001). The fatness of tails is the effect of jumps caused by The Central Bank decisions and by unanticipated macroeconomic events. From term structure of interest rate modeling point of view, jumps and heavy tails are believed to be essential because they imply that the Brownian motion used in most interest rate models should be replaced with Lévy processes: jump-diffusion and pure jump processes which have increments whose distribution is heavy-tailed, see Bas and Das (1996), Björk et al. (1997), Eberlein and Özkan (2002), Eberlein and Raible (1999), Glasserman and Kou (1999) or Glasserman and Merener (2001).

The aim of this paper is to present α -stable and hyperbolic distributions as alternatives to the normal distribution in approximation of yields of the Polish 52-week Treasury bills (T-bills 52w) and of interbank rates such as 3-month WIBOR¹ (WIBOR 3M).

2 α -stable distributions

Stable distributions are a rich class of heavy-tailed distributions which were first characterized by Paul Lévy in his study of sums of independent and

* A project supported by The Polish State Committee for Scientific Research, Grant No. 2H02B01724.

¹ *Warsaw InterBank Offered Rate*-daily rate fixing in Warsaw, analogous to LIBOR or EURIBOR.

identically distributed random variables in the 1920's. The adjective *stable* is used to account for the fact that the distribution of a sum of iid stable r.v.'s belongs to the same class of distributions as the variables' distribution itself. Thus the distributions in question are invariant under the operation of summation.

They have been proposed as a model for many types of physical and economic phenomena because there are strong reasons, both theoretical and practical, for using them. Firstly, the Generalized Central Limit Theorem (GCLT) states that the only possible non-trivial distributional limit of properly normalized sums of iid random variables is stable distribution. Secondly, many large data sets exhibit heavy-tailedness and skewness—this evidence combined with the GCLT is used by many to justify the use of stable models for instance in finance, see Mandelbrot (1963), Fama (1965), Embrechts, Klüppelberg and Mikosch (1997).

A random variable X is said to be stable, and more precisely α -stable², if and only if for all $n > 1$ there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$ such that

$$X_1 + \dots + X_n = c_n X + d_n, \quad (2.1)$$

where X_1, \dots, X_n are iid copies of X ; it is well-known that the only possible choice for c_n is $c_n = n^{1/\alpha}$ for some $\alpha \in (0, 2]$.

The lack of exact analytical formulas for densities and distribution functions in this class (with the exceptions of Gaussian distribution, Cauchy distribution and Lévy distribution) implies that the most convenient way to describe all the possible density functions of stable distributions is through the stochastic inverse Fourier transform

$$s_{\alpha, \beta, \gamma, \delta}(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp(itx) \varphi(t) dt, \quad (2.2)$$

where $\varphi(t)$ is a characteristic function defined as follows:

$$\varphi(t) = \begin{cases} \exp(i\delta t - \gamma^\alpha |t|^\alpha [1 - i\beta \text{sign}(t) \tan \frac{\alpha\pi}{2}]) & \alpha \neq 1, \\ \exp(i\delta t - \gamma |t|^\alpha [1 + i\beta \frac{2}{\pi} \text{sign}(t) \ln |t|]) & \alpha = 1, \end{cases} \quad (2.3)$$

where $\alpha \in (0, 2]$ is the *index of stability*, $\beta \in [-1, 1]$ —skewness parameter, $\gamma > 0$ —scale parameter and $\delta \in \mathbb{R}$ —location parameter. For $\alpha = 2$ we have the normal distribution and for $\alpha \in (0, 2)$ a distribution with tails heavier than these of the normal. Note that for $\alpha \in (0, 2)$ $E|X|^p$ is finite if $p \in (0, \alpha)$, and that $E|X|^p = +\infty$ if $p \geq \alpha$. The above fact implies that for $\alpha < 2$ $E|X|^2 = +\infty$ and stable distributions do not have finite moments beginning with the second one and further on.

² As opposed to \diamond -stable distributions; in the latter case stability has been extended into the operation of multiplication and the taking of maximum, see Mitnik and Rachev (2001).

3 Hyperbolic distributions

Generalized hyperbolic distributions were introduced and first applied to model grain size of wind blow sands in the late 1970's by Ole Barndorff-Nielsen (1977). The adjective *hyperbolic* reflects the fact that the graph of the logarithm of density function is a hyperbola which is in contrast to the normal distribution where we obtain a parabola. An important aspect is that generalized hyperbolic distributions embrace many special and limit cases such as hyperbolic distributions, normal inverse Gaussian distribution, Student's-t, variance-gamma and normal distribution. All of them have been used to model financial returns. Statistical properties of these univariate distributions are well-known, see Blæsild (1999). Recently, generalized hyperbolic distributions and their subclasses have been proposed as models for the increments of price processes, see Barndorff-Nielsen (1995, 1998), Eberlein and Keller (1995), Eberlein, Keller and Prause (1998), Prause (1999).

The one-dimensional generalized hyperbolic distribution is given by the density function

$$\begin{aligned} gh_{\lambda,\alpha,\beta,\delta,\mu}(x) &= \frac{\sqrt{(\alpha^2 - \beta^2)^\lambda}}{\sqrt{2\pi}\alpha^{\alpha-\frac{1}{2}}\delta^\alpha K_\lambda(\delta\sqrt{\alpha^2 - \beta^2})} \sqrt{\left[\delta^2 + (x - \mu)^2\right]^{(\alpha-\frac{1}{2})}} \\ &\quad \times K_{\lambda-\frac{1}{2}}\left(\alpha\sqrt{\delta^2 + (x - \mu)^2}\right) \exp[\beta(x - \mu)], \quad x \in \mathbb{R}, \end{aligned} \tag{3.1}$$

where K_λ is the modified Bessel function of third kind with index λ and the parameters describe: $\alpha > 0$ shape, $0 \leq |\beta| < \alpha$ skewness, $\delta > 0$ scale, $\mu \in \mathbb{R}$ location and $\lambda \in \mathbb{R}$ the class of distribution. Note that the normal distribution is obtained as the limit of the generalized hyperbolic distributions (3.1) when letting $\delta \rightarrow \infty$, and $\delta/\alpha \rightarrow \sigma^2$. For $\lambda = 1$, from (3.1), one can obtain the hyperbolic distribution

$$hyp_{\alpha,\beta,\delta,\mu}(x) = \frac{\sqrt{\alpha^2 - \beta^2}}{2\alpha\delta K_1(\delta\sqrt{\alpha^2 - \beta^2})} \exp\left[-\alpha\sqrt{\delta^2 + (x - \mu)^2} + \beta(x - \mu)\right], \tag{3.2}$$

where $\alpha > |\beta|$, $\delta \geq 0$ and $\mu \in \mathbb{R}$. Hyperbolic distributions have all moments unlike the stable ones.

4 Empirical studies

The subject of the studies was the behavior of the (most liquid on the Polish market) 52-week Treasury bills (T-bills 52w) quoted from 2000-01-03 to 2002-12-16 (146 observations) as well as 3-month interbank rates WIBOR (WIBOR

3M) covering period from 2002-01-02 to 2002-12-31 (227 observations). Recently, the most popular term structure models, that is the so-called *market models* or *LIBOR and swap market models*, have assumed that the interest rate increments (changes) $r_{t+\Delta} - r_t$ are log-normally distributed. Therefore, this study will be carried out upon transformed data $\ln r_{t+\Delta} - \ln r_t$ rather than the data itself. The normality assumption of the transformed data will be verified.

In the case of both T-bills 52w and WIBOR 3M one can observe a clear decreasing tendency in interest rate movements. The increments of the interest rates under consideration reveal jumps which are possible to occur only if the distribution of the increments $\ln r_{t+\Delta} - \ln r_t$ is heavy-tailed. These jumps reflect decisions of The National Bank of Poland as well as some unanticipated macroeconomic events.

4.1 Analysis of normality

Preliminary analysis of normality was conducted on the basis of qq-plot, see Figure 1. In both cases heavy tails are observed which exclude normal

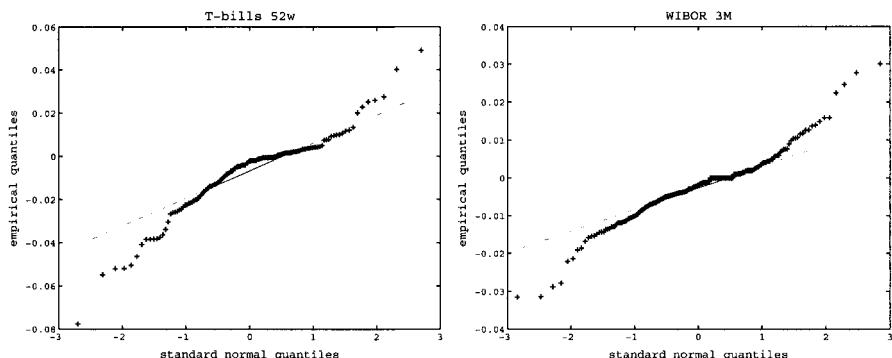


Fig. 1. QQ-plots.

distribution. A more formal analysis of normality was conducted on the basis of Jarque-Bera and Lilliefors test of goodness of fit (see Table 1 and 2).

statistics	value of statistics	critical value
Jarque-Bera	39.07861	5.991464
Lilliefors	0.135967	0.073578

Table 1. Test of normality of T-bills 52w at 5% level.

statistics	value of statistics	critical value
Jarque-Bera	61.33302	5.991464
Lilliefors	0.109481	0.058936

Table 2. Test of normality of WIBOR 3M at 5% level.

The hypothesis that T-bills 52w and WIBOR 3M interest rate log-increments have normal distribution can be rejected at the assumed level of significance.

4.2 Estimation of parameters

For all three distributions, that is normal, α -stable and hyperbolic, the estimation of parameters was conducted by means of the maximum likelihood estimation method (MLE). The results are shown in Table 3. In the case of

distribution	T-bills 52w		WIBOR 3M	
	parameters	errors	parameters	errors
normal	$\hat{\mu} = -0.0071$ $\hat{\sigma} = 0.0178$	$\Delta\hat{\mu} = 0.0029$ $\Delta\hat{\sigma} = 0.0021$	$\hat{\mu} = -0.0024$ $\hat{\sigma} = 0.008608$	$\Delta\hat{\mu} = 0.0011$ $\Delta\hat{\sigma} = 0.0008$
α -stable	$\hat{\alpha} = 1.1796$ $\hat{\beta} = -0.5724$ $\hat{\gamma} = 0.00703$ $\hat{\delta} = -0.0023$	$\Delta\hat{\alpha} = 0.2273$ $\Delta\hat{\beta} = 0.2743$ $\Delta\hat{\gamma} = 0.0014$ $\Delta\hat{\delta} = 0.0019$	$\hat{\alpha} = 1.5049$ $\hat{\beta} = -0.120$ $\hat{\gamma} = 0.0044$ $\hat{\delta} = -0.002$	$\Delta\hat{\alpha} = 0.2065$ $\Delta\hat{\beta} = 0.3907$ $\Delta\hat{\gamma} = 0.0006$ $\Delta\hat{\delta} = 0.0009$
hyperbolic	$\hat{\alpha} = 94.6300$ $\hat{\beta} = -27.72$ $\hat{\delta} = 4.86e - 13$ $\hat{\mu} = -0.000357$	$\Delta\hat{\alpha} = 8.57859$ $\Delta\hat{\beta} = 6.77999$ $\Delta\hat{\delta} = 2.28e - 11$ $\Delta\hat{\mu} = 0.19e - 3$	$\hat{\alpha} = 174.38$ $\hat{\beta} = -35.03$ $\hat{\delta} = 1.54e - 10$ $\hat{\mu} = -1.07e - 10$	$\Delta\hat{\alpha} = 12.078$ $\Delta\hat{\beta} = 9.0172$ $\Delta\hat{\delta} = 4.01e - 9$ $\Delta\hat{\mu} = 2.62e - 7$

Table 3. Estimated parameters and errors of estimation for 52w T-bills and WIBOR 3M.

α -stable distribution some difficulty is caused by the necessity of estimation through the stochastic inverse Fourier transform (2.2) as well as the choice of a suitable form of characteristic function (2.3), see Nolan (1999). Estimation of parameters of hyperbolic distribution is relatively simpler than that of α -stable distribution, which is due to the explicit analytical form of the density function, see Prause (1999). Below, in Figure 2, left tail fitted density functions are presented against the background of the left tail kernel density of empirical distribution on a log-scale.

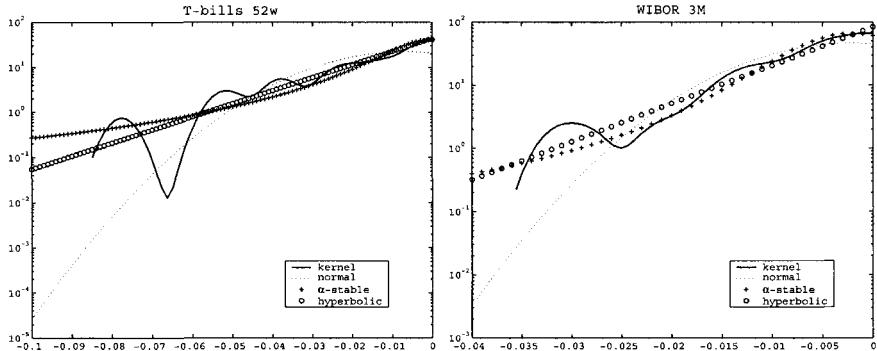


Fig. 2. Left tail fitted density functions and kernel density of empirical distribution on a log-scale.

4.3 Goodness of fit

Goodness of fit between the empirical distribution $F_e(x)$ and the theoretical $F_t(x)$ was tested by means of Kolmogorov \mathcal{K} statistic

$$\mathcal{K} = \max_x |F_e(x) - F_t(x)|, \quad (4.1)$$

and Anderson-Darling \mathcal{AD} statistic

$$\mathcal{AD} = \max_x \frac{|F_e(x) - F_t(x)|}{\sqrt{F_t(x)[1 - F_t(x)]}}. \quad (4.2)$$

The smaller values the statistic \mathcal{K} and \mathcal{AD} take, the better fit to the studied distribution is.

There is a problem when one uses the \mathcal{K} statistic in the testing of heavy-tailedness. Large values of the statistic yet are not evidence that the empirical data and the considered model differ in the tails since the \mathcal{K} statistic gives the best results when investigating the goodness of fit about the median. It happens so because the variance of the term $|F_e(x) - F_t(x)|$ is proportional to $F_t(x)[1 - F_t(x)]$ which is largest in the neighborhood of the median ($F_t(x) = \frac{1}{2}$) and smallest in the tails of distribution (where $F_t(x)$ is close to 0 or 1). If extremal observations are mainly under consideration, which is connected with maximal risk assessment, the \mathcal{K} statistic is therefore of little value to the researcher.

Anderson-Darling statistic (4.2) is a more suitable one to apply in order to compare goodness of fit in the tails. It is a version of weighed Kolmogorov statistic such that the term $\frac{|F_e(x) - F_t(x)|}{\sqrt{F_t(x)[1 - F_t(x)]}}$ takes on the largest values in the far ends of the empirical distribution.

For T-bills 52w as well as WIBOR 3M α -stable and hyperbolic distributions are by far a better fit than the normal distribution (see Table 4 and

5). Treasury yield of T-bills 52w is best reflected by the 1.1796-distribution about the median and by hyperbolic distribution (with appropriate parameters) in the tails (see Table 4). Interbank rates WIBOR 3M observations

statistics		
distribution	Kolmogorov \mathcal{K}	Anderson-Darling \mathcal{AD}
normal	0.12796	1.00919
α -stable	0.0506	0.18570
hyperbolic	0.06673	0.1725

Table 4. Kolmogorov and Anderson-Darling statistics for T-bills 52w.

concentrated about the median are best described by hyperbolic distribution and for observations coming from the tails the best fit is 1.5049-stable distribution (see Table 5).

statistics		
distribution	Kolmogorov \mathcal{K}	Anderson-Darling \mathcal{AD}
normal	0.10165	0.52731
α -stable	0.07889	0.1654
hyperbolic	0.0734	0.22223

Table 5. Kolmogorov and Anderson-Darling statistics for WIBOR 3M.

5 Conclusions

From the above analysis follows that for the distribution of log-increments of T-bills 52w as well as of WIBOR 3M the hypothesis of normality has to be definitely rejected. The 1.1796-stable distribution is the best fit for the yields of T-bills 52w distribution about the median and hyperbolic distribution is more appropriate in the tail areas. The situation is slightly different in the event of WIBOR 3M distribution. This time, hyperbolic distribution describes empirical data which are concentrated about the median better than that in the tails. However, the best fit for the tails is 1.5049-stable distribution.

The above results imply that on the Polish market the Brownian motion, which is currently used in most interest rate models, should be replaced with Lévy processes—jump-diffusion such as the Brownian-Poissonian processes or pure jump motions such as α -stable and hyperbolic ones—which have increments whose distribution is heavy-tailed.

References

- BARNDORFF-NIELSEN, O.E. (1977): Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society A 353*, 401–419.
- BARNDORFF-NIELSEN, O.E. (1995): Normal inverse Gaussian processes and the modeling of stock returns, Research Reports No. 300, University of Aarhus.
- BARNDORFF-NIELSEN, O.E. (1998): Processes of normal inverse Gaussian type. *Finance and Stochastics 2*, 41–68.
- BAS, B. and DAS, S. (1996): Analytical approximation of the term structure for jump-diffusion processes: a numerical analysis. *Journal of Fixed Income, June*, 78–86.
- BJÖRK, T., KABANOV, Y., and RUNGGALDIER, W. (1997): Bond market structure in the presence of marked point processes. *Mathematical Finance 7*, 211–239.
- BLÆSILD, P. (1999): Generalized hyperbolic and generalized inverse Gaussian distributions, Working Paper, University of Aarhus.
- DAS, S. (2002): The surprise element: jumps in interest rate models. *Journal of Econometrics, 106*, 27–65.
- EBERLEIN, E. and KELLER, U. (1995): Hyperbolic distributions in finance. *Bernoulli 1*, 281–299.
- EBERLEIN, E., KELLER, U., and PRAUSE, K. (1998): New insights into smile, mispricing and value at risk: the hyperbolic model. *Journal of Business, 71*, 371–405.
- EBERLEIN, E. and ÖZKAN, F. (2002): The Lévy LIBOR Model, Working Paper, University of Freiburg.
- EBERLEIN, E. and RAIBLE, S. (1999): Term structure models driven by general Lévy processes. *Mathematical Finance 9*, 31–53.
- EMBRECHTS, P., KLÜPPELBERG, C., and MIKOSCH, T. (1997): *Extremal events for insurance and finance*. Springer-Verlag, Berlin Heidelberg.
- FAMA, E. (1965): The behavior of stock market prices. *Journal of Business, 38*, 34–105.
- GLASSERMAN, P. and KOU, S. (1999): The term structure models of simple forward rates with jump risk, Working Paper, Columbia University.
- GLASSERMAN, P. and MERENER, N. (2001): Cap and swaption approximation in LIBOR Market Models with jumps, Working Paper, Columbia University.
- JOHANNES, M. (2001): The statistical and economic role of jump in continuous-time interest rate models, Working Paper, Graduate School of Business Columbia University.
- MANDELBROT, B. (1963): The Variation of Certain Speculative Prices. *Journal of Business, 36*, 394–419.
- MITNIK, S. and RACHEV, S. (2001): *Stable Paretian models in finance*, Wiley, New York.
- NOLAN, J. (1999): Maximum likelihood estimation and diagnostics for stable distributions, Working Paper, American University, Washington.
- PRAUSE, K. (1999): The generalized hyperbolic model: estimation, financial derivatives and risk measures, Ph.D. dissertation, University of Freiburg.

Stability of Selected Linear Ranking Methods - An Attempt of Evaluation for the Polish Stock Market

Waldemar Tarczyński and Małgorzata Łuniewska

Department of Insurance and Capital Markets,
University of Szczecin, ul. Mickiewicza 64, 71-101 Szczecin, Poland

Abstract. In the paper the original applications of linear methods are shown. The authors have considered three different ranking methods for constructing portfolios basing on a stable classification i.e: Generalised Distance Measure *GDM*, synthetic development measure *TMAI*, No-Pattern Method (standardized value sums methods). The results of linear ranking of the companies according to *GDM* have been evaluated in relation to the synthetic development measure *TMAI* and No-Pattern Method. The research was carried out for the companies listed on the Warsaw Stock Exchange in the period 2001-2002.

1 Introduction

There are three big groups of methods within the analysis methods applied to the capital market: technical analysis, fundamental analysis and portfolio analysis. Among those methods it is solely the portfolio analysis that allows a combined analysis of all the stocks. The problem that arises when performing such analyses concerns the base of companies for the construction of a portfolio of securities. An investor applying those methods expects, on the one hand, the diversification of the risk decreasing the portfolio's risk, resulting from the increased number of securities in the portfolio (Evans and Archer (1986), Dobbins et al. (1992), Tarczyński (1997)). On the other hand, on the basis of the rate of return and portfolio risk effect, the investor can at the beginning define the marginal parameters of the investment, that is the lowest expected rate of return and the highest portfolio risk that is possible (Tarczyński (1997, 2002)). The advantages of such a portfolio analysis are thus extensive.

Evaluating the classical concepts of portfolios of securities from the pragmatic point of view we can state that they are the techniques of long-term investment analyses and investing. It results mainly from a low flexibility of a portfolio of securities. The construction of a portfolio is pointless if it takes several weeks while its change will be necessary in, for instance, a month. Even if we decide, basing on the evaluation of the current market situation, that it is necessary to reconstruct the portfolio, the actual reconstruction is impossible in a short term due to the limited liquidity of the stock exchange (on the Warsaw Stock Exchange approximately 1% of the stock of every

company is traded every day). Therefore, it seems obvious that a portfolio of securities should be constructed for a long term.

The analysis of the stock market proves that the criteria of the rate of return and risk determined by the covariance of the rate of return are not the best measures due to their instability. It can be observed particularly in developing markets (Polish market being one of them), with low liquidity, where classical portfolios do not allow to yield above-average returns. It encourages the search for new solutions allowing the construction of a portfolio of securities, which naturally uses long-term bases for investment decisions. The combination of fundamental analysis methods and the idea of construction of a portfolio of securities seems purposeful. It is impossible in a direct way, however, since the fundamental analysis is too extensive and its formalisation for the purpose of construction of a portfolio requires serious simplifications. Fundamental analysis, as a typical technique of analysis for long-term investing, seems to be a good basis to construct a portfolio of securities. The problem which needs to be solved is the transformation of the multi-element results of the fundamental analysis to a form that ensures their applicability in the construction of a portfolio of securities¹. It is a proposal of constructing a fundamental portfolio of securities, which is going to be a long-term portfolio, taking into consideration all the significant advantages of fundamental analysis, i.e. respecting the real power of the companies at the cost of the resignation of financially and economically weak entities, called speculation companies. The portfolio constructed on such bases will be solid and secure. The advantages of such an approach seem obvious to long-term investors. The main criterion subject to optimisation is the sum of synthetic measures describing the fundamental power of the companies comprising the portfolio weighted with the share of their stocks in the portfolio. Such a construction should ensure the long-term security and stability of the portfolio.

In previous works we have proposed the application of the methods of multidimensional comparative analysis to limit the set of companies comprising the portfolio to those which are best in terms of fundamental analysis (economic and financial condition)². The paper analyses the stability of the classification of companies for the purpose of construction of a portfolio of securities and compares the ranking according to various most popular methods. Basing on a stable classification of, say, top 20 companies we may construct portfolios of securities consistent with their long-term and fundamental character. For this purpose we suggest the use of the linear rank-

¹ A proposal how to construct such a portfolio was presented among others by Tarczyński (1995, 1997, 2002).

² *Wybrane metody wielowymiarowej analizy porównawczej w procesie budowy portfela papierów wartościowych*, paper presented at the Conference Financial investments and insurance in Szklarska Poręba in 2002, co-author: M. Luniewska; Teoria dywersyfikacji ryzyka - podejście fundamentalne, paper presented at the Conference Preference Modelling versus Risk in Ustroń in 2003, co-author: M. Luniewska.

ing method: synthetic development measure *TMAI* (Taxonomic Measure of Investment's Attractiveness), generalised distance measure *GDM* and the No-Pattern Method - standardised value sums method. We have made an attempt to find out which of the analysed methods seems to be the best in Polish conditions and how often the base of companies should be verified.

2 Description of the analysed methods

The paper analyses the synthetic development measure *TMAI*, generalised distance measure *GDM* and the No-Pattern Method — standardised value sums method. To calculate the synthetic development measure the following formulae have been used (Tarczyński (2002)):

$$TMAI_i = 1 - \frac{d_i}{d_0}, \quad (i = 1, 2, \dots, n), \quad (1)$$

where:

$TMAI_i$ — synthetic development measure for the object i ,

d_i — distance between the object i and the pattern object calculated according to the formula:

$$d_i = \sqrt{\sum_{j=1}^m w_j \cdot (z_{ij} - z_{0j})^2}, \quad (i = 1, 2, \dots, n),$$

z_{0j} — maximum value of z_{ij} for the object i ,

z_{ij} — standardized value of the attribute j for the object i ,

d_0 — norm assuring that $TMAI_i$ reaches values ranging from 0 to 1:

$$d_0 = \bar{d} + a \cdot S_d.$$

\bar{d} — average value of d_i ,

S_d — standard deviation of d_i .

Using the relation (1) and the information that $0 \leq TMAI_i \leq 1$ and $d_i > 0$, we may determine the margin value for the constant a :

$$a \geq \frac{d_{imax} - \bar{d}}{S_d},$$

where d_{imax} is the maximum value of d_i .

The weights w_i in the distance formula have been calculated according to the following formula based on the coefficient of variation of the variable:

$$w_j = \frac{V_j}{\sum_{j=1}^m V_j}, \quad (j = 1, 2, \dots, m),$$

In the case of the general distance measure GDM the following formulae have been used³:

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^m w_j \cdot a_{ikj} \cdot b_{kij} + \sum_{j=1}^m \sum_{l=1, l \neq i, k}^n w_j \cdot a_{ilj} \cdot b_{klj}}{2 \cdot \left[\sum_{j=1}^m \sum_{l=1}^n w_j \cdot a_{ilj}^2 + \sum_{j=1}^m \sum_{l=1}^n w_j \cdot b_{klj}^2 \right]^{1/2}}, \quad (2)$$

where:

d_{ik} – distance measure,

w_j – weight of the variable j meeting the conditions:

$$w_j \in (0; m), \quad \left(\sum_{j=1}^m w_j = m \right),$$

For the variables measured with the ratio scale the following substitution takes place:

$$a_{ipj} = x_{ij} - x_{pj} \quad \text{for } p = k, l,$$

$$b_{krj} = x_{kj} - x_{rj} \quad \text{for } r = i, l,$$

where x_{ij} (x_{kj} , x_{lj}) – observation i (k, l) for the variable j .

This approach uses the idea of a generalised correlation coefficient, which includes the Pearson's linear correlation coefficient and Kendall's rank correlation coefficient, tau.

The idea of the sums of standardised values has been used to construct the relative development level ratio. The ratio has been calculated according to the formulae (Grabiński et al. (1989), Pociecha et al. (1988)):

$$W_i = \frac{\sum_{j=1}^k z_{ij}}{\sum_{j=1}^k \max_i \{z_{ij}\}},$$

$$z_{ij} = x_{ij}^* + \left| \min_i \{x_{ij}^*\} \right|, \quad (3)$$

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{S_j},$$

where:

W_i – relative development level ratio,

³ The method has been proposed by Walesiak in 2002.

x_{ij} – value of the attribute j for the object i (diagnostic variable),
 \bar{x}_j , S_j – arithmetic mean and standard deviation of the attribute j , respectively.

The ratio, like $TMAI$ is standardised and reaches values ranging from 0 to 1. The closer the value of the measure to 1, the better the object in terms of the general criterion.

For selected methods of analysis, two sets of diagnostic variables have been proposed: for banks and financial institutions and for other companies. In the case of banks the set of diagnostic attributes is comprised by: profitability ratios: return on equity (ROE), return on assets (ROA); liquidity ratio: current liquidity ratio; security ratios: capital adequacy, equity/total assets.

For the rest of the companies the following variables have been used: profitability ratios: return on equity (ROE), return on assets (ROA); liquidity ratio: current liquidity ratio; activity ratios: receivables turnover (in days), inventory (in days), liabilities turnover (in days), assets turn-over; debt ratio: debt margin.

Those variables are generally available and published, for instance by Notoria Service, by quarters of the year and years for all the companies, which should make it easy to carry out the analyses proposed in this paper in terms of data access.

Among the variables presented above turnover of receivables, inventory and liabilities and the debt margin have been considered to be destimulants. The current liquidity ratio for non-financial companies is a nominate, while all the other variables have been assumed as stimulants.

In all the analysed methods the same system of standardisation of the data was used to assure its comparability: the 0–1 standardisation. In all the methods related to the synthetic measure of development $TMAI$ and GDM , the weight system was based on the coefficient of variables variation and a variant was calculated without weighing the variables. In the no-pattern method the same preferences for all the variables have been used, i.e. weights equal to 1.

3 Example

The research covers all the companies listed on the Warsaw Stock Exchange since at least 31.01.2000 and quoted in continuous trading on 31.12.2002, excluding financial institutions (banks, investment funds), 157 companies in total. The data necessary to identify individual classifications come from Notoria Service No. 4(38)/2002. The calculations have been based on the formulas (1) — (3) for the year 2000, 2001, 2nd and 3rd quarters of the year 2002.

To evaluate the stability and similarity of various variants of classifications of the listed companies, we may analyse the relation of ranking in time and

space. Basing on the results of calculation, which are measured with the rating scale, the similarity of rankings of the set of companies with time has been evaluated by means of the Spearman rank correlation coefficient and Kendall's tau⁴. The following approximate relation exists between those coefficients, which is true for a large sample and the values of correlation coefficients significantly different from 1⁵:

$$\rho_s \approx \frac{3}{2} \cdot \rho_k, \quad (4)$$

$$-1 \leq 3 \cdot \rho_k - 2 \cdot \rho_s \leq 1,$$

where:

ρ_s – Spearman rank correlation coefficient,

ρ_k – Kendall's tau coefficient (see: Kendall (1975), Gibbons (1985), Hays (1981)).

The relation of the classifications with time has been analysed individually for every method in the groups with weights and without weights.

TMAI					
	Variant with weights		Variant without weights		
Coefficient	00/01	01/II02	II02/III02	00/01	01/II02
Spearman Rank	0.08859	0.9187	0.8646	0.7972	0.8850
Kendall's Tau	0.7418	0.7986	0.7389	0.6479	0.7583
GDM					
Spearman Rank	0.6100	0.6664	0.8272	0.7413	0.7256
Kendall's Tau	0.4579	0.5459	0.6805	0.5644	0.5953
No-Pattern Method					
Spearman Rank			0.8305	0.8630	0.9404
Kendall's Tau			0.6636	0.7104	0.7972

Table 1. Measures of relations of the rankings in time (Source: the authors' own calculations with *Statistica 6.0PL*)

The data presented in Table 1 proves that for all the analysed methods there is a significant relation between the analysed periods measured both

⁴ Kendall's tau coefficient should be considered here as better, see: Walesiak (2002, p. 60).

⁵ Steczkowski and Zeliaś (1981, p. 169), Siegel and Castellan (1988).

with the Spearman rank correlation coefficient and Kendall's tau. It means that the classifications are stable and there are no statistically significant differences in ranking of those companies. It is particularly important since it allows to state that the period when the database should be verified may be one year when the companies publish their annual reports. In the variant with weights, in the case of the Spearman rank correlation coefficient there is a stronger relation for *TMAI* than *GDM*. In the variant without weights for the periods 2000/2001 and II 2002/III 2002 the strongest correlation (i.e. the highest stability of the ranking) exists in the case of the no-pattern method. Analogous relations have been obtained for Kendall's tau coefficient in the variant with and without weights.

Basing on the classifications obtained, an attempt has been made to check if the introduction of the system of weights in *TMAI* and *GDM* methods significantly affects the stability of the classification. The research has been carried out for the periods 2000, 2001, II 2002 and III 2002. The correlation coefficients describing those relations have been presented in Table 2.

<i>TMAI</i>				
Coefficient	2000	2001	II 2002	III 2002
Spearman Rank	0.8145	0.9477	0.9302	0.9206
Kendall's Tau	0.6439	0.8274	0.7965	0.8092
<i>GDM</i>				
Spearman Rank	0.8228	0.9227	0.9531	0.9663
Kendall's Tau	0.6478	0.7857	0.8318	0.8468

Table 2. Measures of relations for variants with and without weights in the same period (Source: the authors' own calculations with *Statistica 6.0PL*)

The data presented in Table 2 proves unambiguously that the introduction of the system of weights based on the coefficient of variation does not have a significant impact on the rankings of the companies obtained with both *TMAI* and *GDM* method in the analysed period.

An interesting issue is whether between the rankings obtained with various methods significant relations exist in given period in variants with and without weights. The correlation coefficients describing those relations have been presented in Table 3.

The data presented in Table 3 proves there is a great relation between all the methods for all the variants and analysed periods. Despite the high conformity we cannot state unambiguously, however, which of those methods is better for the purpose of the portfolio analysis. It is necessary here to analyse the efficiency of the portfolios obtained on grounds of the base of companies constructed with each method. Such an analysis enables us to

Variants with weights				
TMAI/GDM	2000	2001	II 2002	III 2002
Spearman Rank	0.9752	0.6708	0.5541	0.5926
Kendall's Tau	0.9404	0.5231	0.4318	0.4728
Variants without weights				
TMAI/GDM	2000	2001	II 2002	III 2002
Spearman Rank	0.9527	0.7903	0.7045	0.6522
Kendall's Tau	0.8932	0.6252	0.5661	0.5185
Variants without weights				
TMAI/BWZ	2000	2001	II 2002	III 2002
Spearman Rank	0.7887	0.7695	0.8156	0.8276
Kendall's Tau	0.6234	0.6034	0.6118	0.6593
Variants without weights				
GDM/BWZ	2000	2001	II 2002	III 2002
Spearman Rank	0.7349	0.7058	0.6324	0.6580
Kendall's Tau	0.5839	0.5569	0.5192	0.5362

Table 3. Relations between the classifications in the same periods according to various methods (Source: the authors' own calculations with *Statistica 6.0PL*)

judge whether the differences between the classifications which are stable in time are significant.

For synthetic variables, on the basis of which the analysed classifications have been constructed, another approach to the measure of similarity (stability) of the set of companies in time can be used. The concept can be applied to synthetic variables measured by means of interval or ratio scale. The measure can define not only the rank of the deviation from the values of synthetic variables but also the rank of the deviations resulting from: the differences between average values of synthetic variables, differences in dispersion of values of synthetic variables and the inconformity of the direction of changes in the values of synthetic variables⁶. Formally, such a measure can be described by the following formula⁷:

$$W^2 = \frac{1}{n} \cdot \sum_{i=1}^n (p_{it} - p_{iq})^2, \quad (5)$$

where:

n – number of objects subject to classification (companies),

p_{iq} , p_{it} – values of synthetic measure in analysed periods t and q .

The measure W^2 takes the value 0 if there are no differences in the values of synthetic measures for consecutive periods. An average rank of deviations of the classifications compared is measured with the square root from (5).

⁶ See: Walesiak (2002, pp. 57-58).

⁷ *Ibidem*.

Partial measures mentioned above, whose sum equals W^2 , are given with the following formulae:

$$W_1^2 = (\bar{p}_t - \bar{p}_q)^2, \quad (6)$$

$$W_2^2 = (S_t - S_q)^2, \quad (7)$$

$$W_3^2 = 2 \cdot S_t \cdot S_q \cdot (1 - r), \quad (8)$$

where:

$\bar{p}_t, S_t, (\bar{p}_q, S_q)$ – arithmetic mean and standard deviation for the value of the synthetic variable $t(q)$, respectively,

r – Pearson linear correlation coefficient between the synthetic measures t and q .

The values of the measures calculated on the basis of formulae (5)÷(8) for given synthetic variables have been presented in Table 4.

TMAI						
	Variant with weights			Variant without weights		
Ratio	00/01	01/II02	II02/III02	00/01	01/II02	II02/III02
W	0.0696	0.1049	0.0625	0.0396	0.1272	0.0401
W_1	0.0576	0.1013	0.0379	0.0007	0.1240	0.0020
W_2	0.0065	0.0022	0.0173	0.0175	0.0062	0.0032
W_3	0.0386	0.0275	0.0468	0.0356	0.0278	0.0401
GDM						
W	0.0815	0.1589	0.1001	0.0681	0.0999	0.0592
W_1	0.0241	0.1343	0.0834	0.0353	0.0731	0.0437
W_2	0.0167	0.0460	0.0172	0.0253	0.0357	0.0155
W_3	0.0763	0.0718	0.0527	0.0527	0.0582	0.0368
No-Pattern Method						
W				0.0545	0.0204	0.0303
W_1				0.0487	0.0007	0.0151
W_2				0.0016	0.0066	0.0044
W_3				0.0246	0.0193	0.0350

Table 4. Measures of relations for synthetic variables obtained in time (Source: the authors' own calculations)

The data presented in Table 4 has proved the conclusions reached for the ranks. In the variant with weights, greater conformity in time for synthetic variables has been obtained for *TMAI*. Only for the years 2000/2001 the value W_1 for *GDM* was lower and for the period II 2002/III 2002 — W_2 . In the variants without weights for W , the overall measure of conformity in time, in 2000/2001 *TMAI* proved to be the best and for the other periods

— the no-pattern method. For the measure of the difference between the averages W_1 , *TMAI* rankings proved to be the best in periods 2000/2001 and II 2002/III while for the periods 2001/II 2002 — the no-pattern method. In the case of dispersion measured with W_2 in the years 2000/2001, the best method is the no-pattern method, while in the rest of the periods — *TMAI*. In the case of the inconformity of the direction of changes W_3 the no-pattern method proved to be the best.

The research carried out proves the linear ranking methods (*TMAI*, *GDM* and no-pattern method) enable us to construct a stable-in-time classification of companies listed on the Warsaw Stock Exchange, which opens up a possibility of construction of an appropriate base of companies for the portfolio analysis with respect to its long-term and fundamental character. It has also been proved that in the conditions of the Warsaw Stock Exchange the introduction of weights based on the system of the coefficient of chance variation does not lead to significant changes in the classification.

References

- DOBBINS, R., FRĄCKOWIAK, W., and WITT, F. (1992): *Praktyczne zarządzanie kapitałami firmy*. PAANPOL, Poznań.
- EVANS, J.L. and ARCHER, S.H. (1968): Diversification and the Reduction of Dispersion: An Empirical Analysis. *Journal of Finance*, 23, 761–767.
- GIBBONS, J.D. (1985): Nonparametric statistical inference. 2nd ed., Marcel Dekker, New York.
- GRABIŃSKI, T., WYDYNMUS, S., and ZELIAŚ, A. (1989): *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*. PWN, Warszawa.
- HAYS, W.L. (1981): *Statistics*. 3rd ed., CBS College Publishing, New York.
- KENDALL, M.G. (1975): *Rank correlation methods*. 4th ed., Griffin, London.
- POCIECHJA, J., PODOLEC, B., SOKOŁOWSKI, A., and ZAJĄC, K. (1988): *Metody taksonomiczne w badaniach społeczno-ekonomicznych*. PWN, Warszawa.
- SIEGEL, S. and CASTELLAN, N.J. (1988): *Nonparametric statistics for the behavioral sciences*. 2nd ed., Mc Grow-Hill, New York.
- STECZKOWSKI, J. and ZELIAŚ, A. (1981): *Statystyczne metody analizy cech jakościowych*, PWE, Warszawa.
- TARCZYŃSKI, W. (1995): O pewnym sposobie wyznaczania składu portfela papierów wartościowych, *Przegląd Statystyczny*, No. 1.
- TARCZYŃSKI, W. (1997): *Rynki kapitałowe — metody ilościowe*. Vol. 2, Placet, Warszawa.
- TARCZYŃSKI, W. (2002): *Fundamentalny portfel papierów wartościowych*. PWE, Warszawa.
- WALESIAK, M. (2002): *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*. Akademia Ekonomiczna we Wrocławiu, Wrocław 2002.

Part VII

Production, Logistics, and Controlling

A Two-Phase Grammar-Based Genetic Algorithm for a Workshop Scheduling Problem

Andreas Geyer-Schulz and Anke Thede

Schroff-Stiftungslehrstuhl Informationsdienste und elektronische Märkte,
Institut für Informationswirtschaft und -management,
Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

Abstract. In this contribution we present a two-phase grammar-based genetic algorithm that we use to solve the problem of workshop scheduling in an educational environment which respects partial preferences of participants. The solution respects constraints on workshop capacities and allows for different schedule types. We approach this problem by defining a grammar which defines a language for expressing the restrictions on workshops and participants. A word of this formal language represents a solution which by definition of the language is always feasible. For each feasible schedule the fitness is the result of optimizing the group's social welfare function which is defined as the sum of the individual utility functions as expressed by the partial preferences. This optimization is achieved with an order based genetic algorithm which assigns to each participant his personal schedule.

1 Introduction

Genetic algorithms are widely used for solving combinatorial optimization problems (Starkweather et al. (1991)). They work by imitating evolution and use the concept of survival of the fittest by creating new solution candidates on the basis of the best parent candidates. Selection, mutation and recombination are the basic methods for generating new candidate generations (Goldberg (2001)). However, for several problem types (e.g. project scheduling problems (Neumann et al. (2002), Merkle et al. (2002))) the application of genetic algorithms proved to be difficult because of the increasing complexity of generating random solution representatives. Gottlieb and Raidl (2000) have addressed this problem for knapsack problems with repair and decoder algorithms whereas Bruhn and Geyer-Schulz (2002) approached the problem with an extension of genetic programming (Koza (1998)).

We use a genetic algorithm based on constraint context-free languages to solve a workshop scheduling problem with the following properties:

- 50 workshops take place on one day during two time slots: in the morning or in the afternoon. Workshops take place in both or in only one time slot.
- Each workshop has a limited number of places that must not be exceeded and that may be different for the two time slots of the same workshop.

- 419 participants (girls) have to be assigned to the workshops. Each girl visits a morning and an afternoon workshop or only one of these.
- Each person must not visit the same workshop twice.
- Each girl can assign positive and negative priorities or preferences to the workshops. Preferences range from -4 (do not want to visit at all) over 0 (no priority = do not care) to +4 (would like to visit very much).
- Only a partial preference function (two of each kind of priorities for each girl) is allowed. This reduces the girls' cost for specification of their preferences and it preserves more degrees of freedom for the optimization algorithm. All other preferences are set by default to 0.

The optimization problem deals with assigning the girls to the workshops such that all constraints mentioned above are met and that the sum of the partial preference functions of each girl which corresponds to the social welfare function is maximized. Details are described in sec. 2.

The genetic algorithm we developed is based on a context-free language with linear constraints (Bruhn (2000), Bruhn and Geyer-Schulz (2002)). A *context-free grammar* is a set $G = (V, T, E, S)$ with V a set of non-terminal symbols, T a set of terminal symbols with $V \cap T = \emptyset$, $E \subset V \times (V \cup T)$ a set of rules and the start symbol $S \in V$. T^* denotes a finite sequence of symbols of T . A string is derived from the grammar G by starting with the start symbol, repeatedly applying rules from E to the symbols in the string deriving new strings and finally stopping when the resulting string consists only of terminal symbols. The *context-free language* $L \subset T^*$ of the grammar G is the set of all strings of terminal symbols that can be derived with the aid of the grammar.

Context-free languages can be used for genetic algorithms to define the structure of all possible individuals. As an example, the following, very simple grammar can be used to derive all 0-1 strings (we denote the non-terminal symbols with a capital letter and a rule $e = (v \in V, x \in V \cup T)$ as $v \rightarrow x$, alternative derivations for the same non-terminal symbol are separated by a |. The sets of symbols are given by the symbols present in the rules, the start symbol is always S .)

$$\begin{aligned} S &\rightarrow SP \mid P \\ P &\rightarrow 0 \mid 1 \end{aligned}$$

Grammars can be extended with constraints adding the possibility to represent linear constraints of optimization problems that cannot be represented by the grammar itself (Bruhn (2000), Bruhn and Geyer-Schulz (2002)). Constraints are represented by assigning “usage”- and “limit”-vectors to symbols of the grammar. A derivation for a string of the language which can be represented as a tree is extended by adding the usages and limits of each symbol in the tree. To illustrate this consider the above grammar and a problem with the constraint that the number of 1s in the resulting string may not exceed 5. As there is only one constraint the vectors are one-dimensional. We assign an initial usage of 1 to the terminal symbol 1 and a limit of 5 to the non-terminal

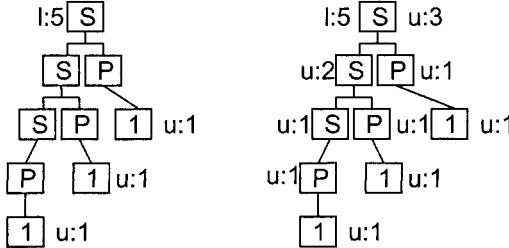


Fig. 1. Initial usage (u) and limit (l) values (left) and recursive propagation (right).

symbol S which represents the whole resulting string. The default initial usage for the other symbols is 0 and the default initial limit is ∞ . The usage of each non-terminal symbol is then recursively defined as its own usage plus the sum of the usages of its sons. The derivation tree in Fig. 1 shows the effect of usage. A derivation tree is now defined as *valid* if for each symbol x in the tree $usage(x) \leq limit(x)$ which corresponds to the fact that the linear constraints are met. This defines *constraint context-free grammars* and correspondingly *-languages* (Bruhn (2000), Bruhn and Geyer-Schulz (2002)). The purpose of constraint context-free grammars is to limit the generation of individuals to feasible ones with respect to the additional constraints. The advantage is that no measures have to be taken a posteriori to eliminate invalid individuals from the population like introducing a penalty function or repair algorithms. If a tree is found infeasible it either has to be destroyed and recreated or a backtracking algorithm can be deployed to find a valid tree. The extension of the operators for mutation and recombination to constraint languages can be found in Bruhn (2000), Bruhn and Geyer-Schulz (2002).

2 The workshop scheduling problem

Let $T_I = \{I, II, III\}$ denote the types of individuals I which participate only in morning, afternoon or in both sessions, respectively. $i \in I = I_I \cup I_{II} \cup I_{III}$ is the index for an individual, I the set of indices of all individuals, the index sets for the different types of individuals are $I_I = \{1, \dots, n_1\}$, $I_{II} = \{n_1 + 1, \dots, n_2\}$, and $I_{III} = \{n_2 + 1, \dots, n\}$. These three subsets are mutually disjoint.

Let $T_W = \{V, N, B1, B2\}$ denote the types of workshops W which are organized only in morning, afternoon or in both sessions where $B1$ denotes the morning and $B2$ the afternoon session of wholeday workshops. $j \in W = W_V \cup W_N \cup W_{B1} \cup W_{B2}$ is the index of a workshop, the index sets for the different types of workshops are $W_V = \{1, \dots, w_1\}$, $W_N = \{w_1 + 1, \dots, w_2\}$, $W_{B1} = \{w_2 + 1, \dots, w_3\}$, $W_{B2} = \{w_3 + 1, \dots, w\}$. These four subsets are mutually disjoint. In the sets W_{B1} and W_{B2} the indices are ordered such that indices at the same positions belong to the same workshop.

Each individual assigns a raw score value $r_{ij} \in \{-4, -3, -2, -1, 1, 2, 3, 4\}$ to a maximum of 16 workshops where each value may be selected at most twice because each girl may specify only a partial preference function as specified in the previous section. All other workshops are implicitly rated with $r_{ij} = 0$. The matrix of raw scores is $R = [r_{ij}]$ with r_{ij} meaning individual i assigns a raw score of r_{ij} to workshop j . Different sessions of the same workshop have the same rating. The scores are computed by the following monotonously increasing mapping $f : R \rightarrow S$:

$$s_{ij} = \begin{cases} r_{ij}^2 + b, & \text{if } r_{ij} > 0 \\ r_{ij}^2, & \text{otherwise} \end{cases} \quad (1)$$

with b a sufficiently large positive constant. $X = [x_{ij}]$ is the assignment matrix. $x_{ij} = 1$ means individual i is assigned to workshop j , else i is not assigned to j . $x_{ij} \in \{0, 1\}$. The partial preference function for individual i is $\sum_{j \in W} s_{ij} \cdot x_{ij}$.

The objective function of the workshop assignment problem is formulated as a social welfare function. We maximize:

$$\max \sum_{i \in I} \left(\sum_{j \in W} s_{ij} \cdot x_{ij} \right) \quad (2)$$

subject to the following constraints:

1. on workshops, upper bounds on workshop participants are respected:

$$\sum_{i \in I} x_{ij} \leq u_j \quad \forall j \in W \quad (3)$$

with u_j denoting the maximum number of participants for workshop j .

2. on individuals of

(a) type T_I (visit only morning workshops):

$$\sum_{j \in (W_V \cup W_{B1})} x_{ij} = 1 \quad \text{and} \quad \sum_{j \in (W_N \cup W_{B2})} x_{ij} = 0, \quad \forall i \in I_I \quad (4)$$

(b) type T_{II} (visit only afternoon workshops):

$$\sum_{j \in (W_N \cup W_{B2})} x_{ij} = 1 \quad \text{and} \quad \sum_{j \in (W_V \cup W_{B1})} x_{ij} = 0, \quad \forall i \in I_{II} \quad (5)$$

(c) type T_{III} : visits two workshops:

$$\sum_{j \in W} x_{ij} = 2, \quad \forall i \in I_{III} \quad (6)$$

and with different topics:

$$x_{ij} + x_{i,j+w_3-w_2} \leq 1, \quad \forall i \in I_{III}, \forall j \in W_{B1} \quad (7)$$

Each of these constraints has less than n constraints in $n \cdot w$ variables. The problem size is bounded from above by $n \cdot w$ variables, with at most $2w$ constraints on workshops and $4n$ constraints on individuals. For 500 individuals and 50 workshops of types W_{B1} or W_{B2} the problem has approximately 50,000 binary variables and about 1,200 constraints.

3 Grammar and linear constraints

In Fig. 2 and 3 we present the grammar and the linear constraints we used to code the workshop scheduling problem. This grammar codes the constraints in equ. (7) and assures that the number of the different workshop types match the number of corresponding girls (v_0, n_0 code the empty morning and afternoon workshop). The limits on participants of equ. (3) are represented by the matrix in Fig. 3 which shows the initial usage and limit values. The assignment of the girls to the workshop pairs (equ. 4, 5, and 6) is done with a penalty function and is described in detail in sec. 4.

$S \rightarrow P \dots PV \dots VN \dots N$	$(I_{III} P's, I_I V's, I_{II} N's)$
$P \rightarrow v_{w_2+1}N_{w_3+1} \mid \dots \mid v_{w_3}N_w \mid VvNa$	$(P: \text{wholeday girls})$
$V \rightarrow Va n_0$	$(V: \text{morning girls})$
$N \rightarrow v_0Na$	$(N: \text{afternoon girls})$
$Va \rightarrow v_1 \mid \dots \mid v_{w_1} \mid v_{w_2+1} \mid \dots \mid v_{w_3}$	$(Va: \text{all morning workshops})$
$Na \rightarrow n_{w_1+1} \mid \dots \mid n_{w_2} \mid n_{w_3+1} \mid \dots \mid n_w$	$(Na: \text{all afternoon workshops})$
$Vv \rightarrow v_1 \mid \dots \mid v_{w_1}$	$(Vv: \text{only-morning workshops})$
$N_{w_3+1} \rightarrow n_{w_3+2} \mid n_{w_3+3} \mid \dots \mid n_w$	
$N_{w_3+2} \rightarrow n_{w_3+1} \mid n_{w_3+3} \mid \dots \mid n_w$	
\vdots	
$N_w \rightarrow n_{w_3+1} \mid n_{w_3+2} \mid \dots \mid n_{w-1}$	

Fig. 2. Grammar for the workshop scheduling problem.

	symbol	v_1	\dots	v_{w_3}	n_{w_1+1}	\dots	n_w
limits:	S	u_1	\dots	u_{w_3}	u_{w_1+1}	\dots	u_w
usages:	v_1	1	\dots	0	0	\dots	0
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	v_{w_3}	0	\dots	1	0	\dots	0
	n_{w_1+1}	0	\dots	0	1	\dots	0
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	n_w	0	\dots	0	0	\dots	1

Fig. 3. Initial usage and limit vectors (one vector per line).

4 The two-phase algorithm

The grammar and the linear constraints presented in sec. 3 can generate feasible workshop schedules that respect the workshop limitations. Strings or schedules generated by the defined context-free language are of the form shown in the upper line of Fig. 4. For a complete solution of the problem the girls have to be assigned to the schedule in a way that each girl is assigned to a v - n pair (gene-pair of the schedule). The simplest, natural way to do this

$$\begin{array}{c|c|c|c|c|c|c} v_{i_1} n_{j_1} & v_{i_2} n_{j_2} & \dots & v_{i_{n-n_2+1}} n_0 & \dots & v_0 n_{j_{n-n_2-n_1+1}} & \dots & v_0 n_{j_n} \\ \hline girl_1 & girl_2 & \dots & girl_{n-n_2+1} & \dots & girl_{n-n_2-n_1+1} & \dots & girl_n \end{array}$$

Fig. 4. Workshop schedule s generated by grammar and assignment of girls.

$$s = \frac{v_{i_1} n_{j_1} | v_{i_2} n_{j_2} | \dots | v_{i_{n-n_2+1}} n_0 | \dots | v_0 n_{j_{n-n_2-n_1+1}} | \dots | v_0 n_{j_n}}{girl_{17} | girl_{202} | \dots | girl_{43} | \dots | girl_3 | \dots | girl_{398}}$$

Fig. 5. Example for assignment of girls after inner optimization.

would be to assign the first workshop pair to the first girl $girl_1$, the second pair to $girl_2$ etc, as shown in the lower line of Fig. 4. The fitness of a thus generated solution candidate is calculated as the sum of the partial preference functions of the girls for the assigned workshops.

$$fitness(s) = \sum_{k=1}^n (s_{ki_k} + s_{kj_k}),$$

Optimizing this solution can be done by applying the genetic operators to the individuals and thus varying the gene combinations on the one hand and the gene positions on the other hand. The disadvantage is that good workshop combinations do not have any positive effect on the fitness unless they are at the right position in the gene and that mutation and recombination operators are computationally quite expensive as they have to take care of not violating any of the linear constraints for the schedule to remain feasible. This can be avoided by decomposing the problem into two optimization problems:

1. finding a feasible workshop schedule
2. assigning the girls to a given schedule such that the fitness is maximized

This introduces a schedule-specific optimization problem and leaves the creation of feasible schedules with good workshop combinations to the first, “outer” optimization problem. The second, “inner” optimization problem for a given workshop schedule is an order-based problem, searching for the best permutation of girls, with the constraint of correctly associating girls to workshop pairs with respect to their type. This problem is currently solved by an order-based genetic algorithm using a penalty function for assuring the generation of feasible solutions. However, in the future a further problem decomposition per participant type will be studied. The result generated by the inner genetic algorithm may look as depicted in Fig. 5. The algorithm for the combined phases is shown in Fig. 6.

The individuals worked with in the inner algorithm are composed of a numeric vector of indices where each girl is represented by one gene. The fitness calculated by the inner algorithm is an approximation of the value of the workshop schedule of an individual of the outer algorithm.

The genetic algorithm for the workshop scheduling problem is implemented in the Objective Caml programming language. The genetic operators are implemented as follows:

```

create_initial_population;
while no_of_outer_evaluations < max_outer_evaluations do
    evaluate_population;
    mutate_population;
    recombine_population;
    select_population;
    foreach el in new_population do
        create_initial_population;
        while no_of_inner_evaluations < max_inner_evaluations do
            evaluate_population;
            mutate_population;
            recombine_population;
            select_population;
        } “inner” genetic algorithm
    done
done
done
return best_value;

```

Fig. 6. Two phase genetic algorithm.

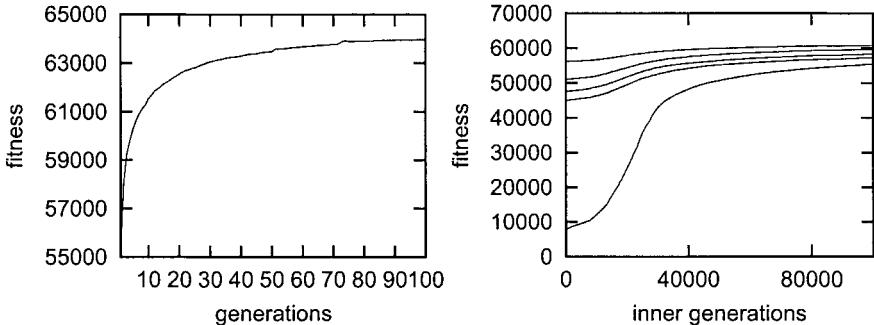
- Selection in the inner and outer algorithm are done by roulette wheel selection proportional to the individuals' fitnesses. The best individual is always transferred into the new population (elitist selection).
- Mutation of inner individuals is implemented using a two-change heuristic (inversion). The selection of the two positions is done by roulette wheel selection inversely proportional to the partial preferences of single genes.
- Evolution of inner individuals is guided using simulated annealing (Winkler (1995)): increasing temperature decreases the probability to accept an inversion that degrades the overall fitness of this individual.
- Recombination and outer mutation is implemented purely stochastically.

5 Results

Tab. 1 shows the results of the workshop assignments actually used for the workshop on November 29th, 2002 in line 1. These were computed by a first version of the genetic algorithm with a greedy inversion operator with 40 hours of computation time and an additional 40 hours of manual effort to reach an acceptable schedule. The results achieved now by the genetic algorithm described in sec. 4 are given for comparison in line 2 in Tab. 1. This version used about 18 hours of computation time, both running on a 1.2 MHz Athlon CPU with 1 GB of main memory and a Linux operating system. Interesting is the percentage of workshops assigned to a girl with a positive priority for it which we were able to increase significantly. No negative priority (“neg.”) is distributed and no girl is assigned to the wrong type of workshop (“false”).

To test the stability of our results we ran the algorithm 25 times. The average fitness development over the generations can be seen in the left diagram

	fitness	neg.	false	prio 0	prio 1	prio 2	prio 3	prio 4	% positive
workshop day	56,808	1	0	318	24	51	99	285	59 %
genetic algorithm	64,282	0	0	208	37	87	125	321	73.3%

Table 1. Results achieved by genetic algorithm and values used on workshop day.**Fig. 7.** Average fitness values over 25 runs (left) and development of inner fitness values over outer generations 1, 2, 10, 40 and 50 with 20 outer individuals (right).

in Fig. 7. These are the fitness values of the outer individuals after the inner optimization has been executed on each. The optimization was done with 20 individuals and 100 generations in the outer algorithm and 2 individuals during 75,000 generations in the inner algorithm. It shows that fitnesses of approx. 64,000 can be expected and the standard deviation of 401.5 shows that the algorithm is robust with regard to random restarts. The right diagram in Fig. 7 gives an impression about the average development of the inner fitnesses over all 20 individuals during the different outer generations. The lowest line shows the fitness during the 1st outer generation, the highest of the 50th and the ones in the middle correspondingly. Only selected generations are shown for clarity. The shape of especially the lowest curve shows the effect of the rising temperature: a slow start is followed by a range of rapid increase in fitness phasing out in slower increase as the optimal fitness value is approached. As the best inner results of each outer individual are passed on to the next generation the fitness for each following generation starts at a better value than the preceding one. In fact, very little optimization can be reached during late outer generations which is a promising source of further performance improvement.

6 Conclusion

In this paper we introduced a two-phase genetic algorithm based on constraint, context-free languages for solving a real-life workshop scheduling problem with positive and negative partial preferences for participants. The

aim of the algorithm is a distribution of participants to workshops that respects the constraints and maximizes the social welfare function expressed by the sum of the partial preferences. The context-free grammar is defined in a way that many of the problem's constraints are already incorporated and only feasible solutions with respect to these constraints are generated. Additionally, constraints are added for modelling the participants limits of the workshops. The resulting optimization procedure is then further improved by decomposing the problem into two phases. The first phase generates feasible workshop schedules and the second phase optimizes the assignment of participants to workshops. The second phase algorithm is realized with a very simple, order-based genetic algorithm. Further tuning of the performance of the algorithm remains to be done.

We evaluated the algorithm by running it for a number of times to test the stability of its results and compared the results generated to the distribution actually used on the last workshop day which we take as a lower bound for acceptable solutions. We managed to improve the results significantly.

References

- BRUHN, P. (2000): *Genetische Programmierung auf Basis beschränkter kontextfreier Sprachen zur Lösung kombinatorischer Optimierungsprobleme*. Phd Thesis, Wirtschaftsuniversität Wien.
- BRUHN, P. and GEYER-SCHULZ, A. (2002): Genetic Programming Over Context-Free Languages with Linear Constraints for the Knapsack Problem: First Results. *Journal of Evolutionary Computation*, 10 (1), 51–74.
- GOLDBERG, D.E. (2001): *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Pub Co, New York. Ed. 22.
- GOTTLIEB, J. and RAIDL, G.R. (2000): Characterizing locality in decoder-based eas for the multidimensional knapsack problem. In: C. Fonlupt, J.-K. Hao, E. Lutton, E. Ronald, and M. Schoenauer (Eds.): *Proceedings of the 4th Conference on Artificial Evolution*. Springer-Verlag, Berlin, 38–51.
- KOZA, J.R. (1998): *Genetic Programming : On the Programming of Computers by Means of Natural Selection*. MIT Press, Cambridge, MA. Ed. 6.
- MERKLE, D., MIDDENDORF, M., and SCHMECK, H. (2002): Ant Colony Optimization for Resource-Constrained Project Scheduling. *IEEE Transactions on Evolutionary Computation*, 6 (4), 333–346.
- NEUMANN, K., SCHWINDT, C., and ZIMMERMANN, J. (2002): *Project Scheduling with Time Windows and Scarce Resources*. Springer-Verlag, Berlin Heidelberg.
- STARKWEATHER, T., McDANIEL, S., MATHIAS, K., WHITLEY, D., and WHITLEY, C. (1991): A Comparison of Genetic Sequencing Operators. In R.K. Belew and L.B. Booker (Eds.): *Proceedings of the Fourth International Conference on Genetic Algorithms*. Morgan Kaufmann Publishers, San Meteo, California, 69–76.
- WINKLER, G. (1995): *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer-Verlag, Berlin Heidelberg.

Classification and Representation of Suppliers Using Principle Component Analysis

Rainer Lasch and Christian G. Janker

Lehrstuhl für Betriebswirtschaftslehre, insbesondere Logistik,
Technische Universität Dresden, D-01062 Dresden, Germany

Abstract. Supplier rating as a critical component for successful Supply Chain Management (SCM) has become more and more important. An empirical study among 193 companies, focused on supplier rating and selection, found out that the existing methods of supplier rating do not satisfy the needs in practice. A new model, based on principal component analysis, closes this gap and demonstrates the utility of classification and representation of suppliers for the supplier management process. A case study illustrates the model.

1 Introduction

With increasing significance of the purchasing function, purchasing decisions have become more important during the last decade. According to the SCM concept, a strategic, long-term partnership between buyer and supplier should be reached. As organisations become more dependent on suppliers, the consequences of poor decision making become more severe. The selection of the "right partners" is therefore more than ever the critical component of a successful supplier management process. Thus, a systematic, objective-oriented process of supplier management is required. Moreover, new demand for supplier rating arise, to which the existing procedures do not fit exactly.

After a short overview of the process of supplier management, the designed supplier-rating-system using principal component analysis is presented. A case study illustrates the procedure and shows managerial implications.

2 Suppliers management process

Within the framework of SCM supplier management is responsible for the arrangement of the supplier-buyer-connection. According to Arnolds et al. (1998), identification and preservation of suppliers, who distinguish themselves by continuance, capacity and readiness of delivery, should be established.

Besides, the decisive frame of supplier management is determined by the corporate strategy via sourcing strategy, object and source (cf. Figure 1). For example, if an enterprise pursues cost-leadership, a possible purchasing strategy can be, to realize high amounts via single-sourcing. For the procurement

object this means that criteria like offer price, discounts etc. are of high importance. Coming from a specific need, the first step is to identify suppliers who offer the required object; this step of the process is called supplier identification. To do so, the relevant buying market is marked out and searched for potential suppliers who produce the required object or are able to produce it (Glantschnig (1994)).

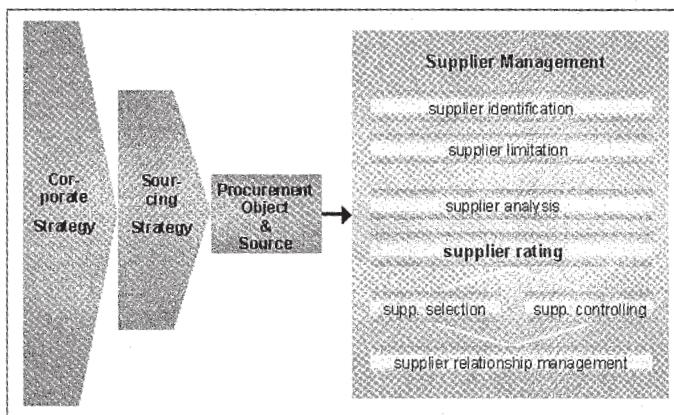


Fig. 1. Suppliers Management Process

Due to the high effort of the following steps, all potential suppliers can not be considered; thus, a supplier limitation is required. Moreover, specified main evaluation criteria of the suppliers are demanded from procurement research - on the bases on a supplier's form or via audits, for example. Both steps together - supplier identification and supplier limitation - are called supplier pre-qualification.

In the supplier analysis the results from self-information, procurement research and audits are gathered and processed for the following supplier rating. The subsequent rating should only consider those suppliers which fulfil the mentioned demand of the main evaluation criteria. Supplier rating includes the systematic and extensive evaluation of the supplier's capacity with the aim to choose new and control already set-up suppliers. Therefore, all the relevant evaluation criteria and the methods to be applied have to be fixed.

The supplier rating is followed either by supplier selection or by supplier controlling. Supplier controlling defines the regular examination of the capacity, which takes place for the duration of the supplier-buyer-connection. Thus, companies are able to uncover deficits of the supplier in time and introduce corresponding sanctions. Together with the integration and development of existing suppliers this step is called supplier relationship management.

Based on the process above, the central position of the supplier rating will be clear. Among other questions within the framework of a 2001 executed empirical study of the chair of business management, especially logistics, of the Dresden University of Technology, the demand for a method to evaluate suppliers were asked. Nearly 70% of the companies asked for a procedure applicable on all purchasing situations (cf. Figure 2).

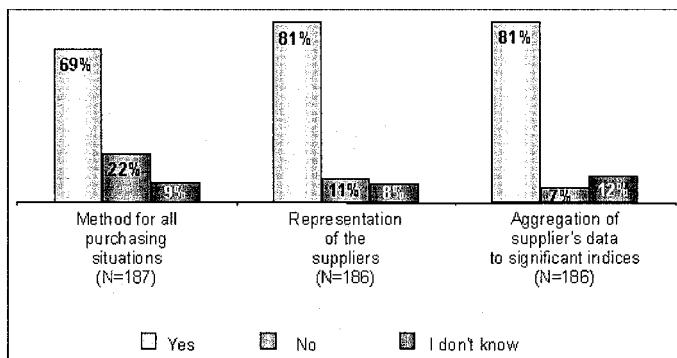


Fig. 2. Statements to Supplier Rating

A clear majority of over 80% wishes a representation of the relative positions of competing suppliers which can be used, for example, as an aid within negotiations. In the same size it was agreed, that an aggregation of supplier's data to few significant indices is helpful to describe the current situation of the suppliers. If these empirical demand for supplier rating are supplemented with the theoretical ones and placed opposite to the existing methods, it will appear first of all that especially representation and classification are not fulfilled so far.

3 Supplier-rating-system using principal component analysis

Within the construction of a new method all evaluated demand should be fulfilled. Not least the conditions of the representation and the aggregation of data suggested the application of a factor analysis, which therefore creates the basis for the supplier-rating-system presented in Figure 3. The system consists of a subjective component and an objective one, just like the factor analysis itself, which has also fixed steps and such ones with a necessary subjective intervention.

As shown in Figure 3 the subjective component considers the present purchasing situation as well as the corporate strategy and is reflected in the variable choice (i.e. the choice of the criteria) and in the construction of the ideal

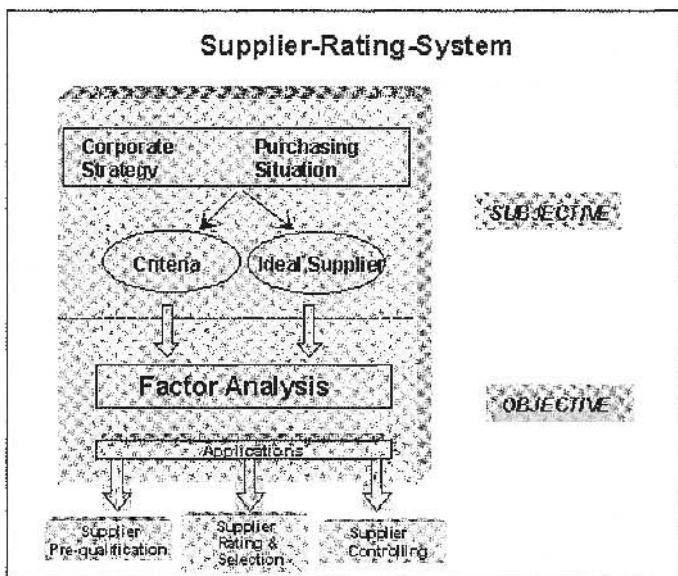


Fig. 3. Supplier-Rating-System

supplier. If an enterprise pursues, for example, the aim of cost-leadership, many sub-criteria of the price are to be included within the decision making.

Next, the factor analysis is executed as an objective procedure which analyses and represents the suppliers to rate and classify them. As factor extraction procedure the principal component method was used due to the lack of pre-set distribution models. Furthermore, this method represents a purely data-manipulating procedure and a priori-estimates of the communalities are not necessary (Fahrmeir et al. (1996)).

The supplier-rating-system is executed in four steps. In the first step the evaluation criteria and the ideal supplier has to be fixed. According to the decision situation, the average number of the main evaluation criteria varies in the supplier pre-qualification. For example, in case of a new product introduction the buyer is confronted with new suppliers, while he can fall back on already existing supplier-buyer-relationships by an assortment change only. Therefore, an enterprise uses more evaluation criteria for a new product introduction than in all other situations (Lasch et al. (2001)). The selection and the number of the criteria are determined furthermore by the partial step of the supplier management process. Supplier pre-qualification provides many suppliers, but only few main criteria; opposite, only few suppliers are rated in supplier selection, however on the basis of many specified criteria.

The ideal supplier is defined by the procuring enterprise which fixes the ideal scores of the criteria. The rating team should consist of several departments of the enterprise (procurement, production, controlling etc.).

Step 2 of the supplier-rating-system, the application of descriptive statistics, provides a quantitative survey of the suppliers and their features. It grants an impression of the main evaluation criteria of the potential suppliers and yields first indications on the basis of the correlation matrix, whether a meaningful factor analysis can be expected.

The factor analysis (step 3) takes place in four partial steps, which includes besides the above-mentioned correlation matrix the factor extraction (via principal component method), the factor rotation and the determination of the factor values (Hartung/Elpelt (1999)).

Using the resulting factor values, a plot is constructed in the last step (only possible in the case of two or three extracted factors, of course). Because factor analysis was designed originally for purely metric data material, the supplier's position in the plot must be confirmed (in case of including several ordinal and/or binary criteria) by performing a cluster analysis or a multi-dimensional scaling (MDS). In conclusion the plot is interpreted according to the aim of the appreciation: in the supplier pre-qualification all suppliers in the cluster of the ideal supplier are submitted to a specified appreciation; later on in the supplier selection the supplier(s) with the lowest distance to the ideal supplier will be chosen exactly. Besides, the cluster is stretched by an ellipse (its centre exactly in the position of the ideal supplier) whose height and width arise from the reciprocal explanation ratio of the factors. Thus, the same numerical distance is vertical less than horizontal, because in general the (horizontal) first factor comprises more information than the second one. In other words, a horizontal distance to the ideal supplier of the factor 2-explanation ratio equals a vertical distance of the factor 1-explanation ratio; so the constructed ellipse is always higher than wider.

The distances to the ideal supplier can also be calculated for selection or pre-qualification of the suppliers. Therefore, for every supplier the city-block-distance is determined between the factor values of this supplier and the ideal supplier and these distances must be weighted with the explanation ratios of the separate factors. A case study shall illustrate the supplier-rating-system's possibilities of application in practice.

4 Case study

The company ARGASSI (name changed) which produces high-quality rubber parts, is planning a new product introduction and has decided within the framework of its purchasing strategy for binary sourcing, i.e. the new product should be delivered by two yet unknown suppliers. In the supplier pre-qualification the data of 20 potential suppliers as well as the ideal supplier (fixed by a cross-functional team of ARGASSI) were recorded. By means of a standardized questionnaire, ARGASSI collected the following characteristic features: PRICE, QUANTITY, QUALITY, LOGISTICS and SERVICE. The price was converted into Euro due to the international character of the

suppliers. The other criteria were scaled 5-stage ordinal with the ratings insufficient (=1 point), sufficient (=2 points), mean (=3 points), well (=4 points) and very well (=5 points). The aim of this supplier pre-qualification was to identify approximately five suppliers who should be submitted to a specified appreciation in the next step.

Some high correlations found out in the descriptive analysis pointed a meaningful application of the factor analysis. After factor extraction and factor rotation, the first factor explains 43,1% and the second factor 36,1% of the total variance, what implies with altogether 79,2% a very good explanation ratio. The first factor is stretched by the criteria QUANTITY, QUALITY and PRICE and forms therefore a product-related component, while the second factor with the criteria LOGISTICS and SERVICE describes the additional performances (cf. Table 1).

	factor 1	factor 2	communality
price	0,6677	-0,5083	0,704166
quantity	0,9017	0,0183	0,813338
quality	0,9396	-0,0708	0,887847
logistics	-0,0182	0,9392	0,882348
service	-0,1053	0,8118	0,670030
explanation ratio	0,4310	0,3610	—

Table 1. Rotated Factor Loadings

Figure 4 shows the Varimax-rotated plot. The relative positions of the suppliers were confirmed by using a cluster analysis as well as a multidimensional scaling.

Finally, the ellipse with a height-to-width-ratio of 43,1:36,1 and centre in the ideal supplier identified six suppliers (No 1, 2, 6, 7, 9, 18) which entered into the following supplier selection. On the basis of 25 specified criteria the six suppliers were valued again to select two suppliers in accordance to the supplier selection's aim. An analogous execution of the supplier-rating-system provides (under inclusion of the a priori defined ideal supplier) the choice of the two suppliers No 1 and 2.

Regular controls of the whole supplier base are an important component to get an overview of the performance structure of all or selected suppliers. Analogies can be derived in the treatment of similar suppliers. Such a performance structure can be represented by building clusters of suppliers in the plot; thus, the factor analysis forms the basis of a supplier's structural analysis. It is important to mention that all suppliers do not deliver the same product (like in the case before), so that the quantitative criteria, for example PRICE, are no more directly comparable and therefore ordinal scales (e.g. (strongly) below average, average, (strongly) outstanding) have to be used, which requires the confirmation of the plot by a cluster analysis or a MDS.

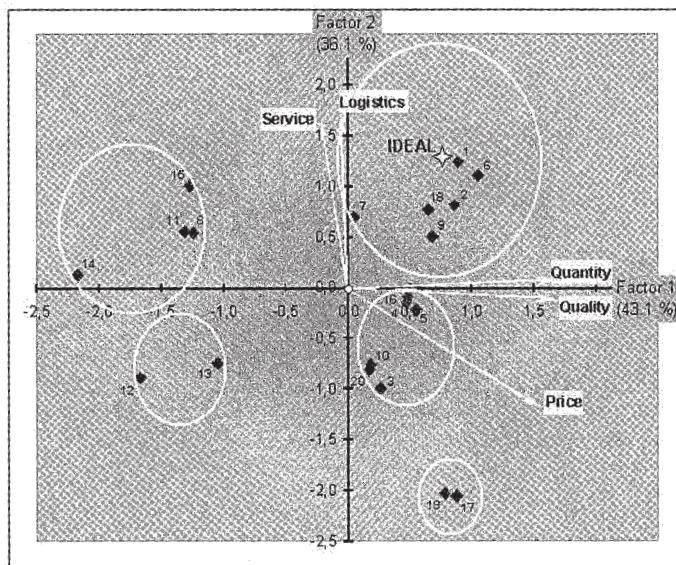


Fig. 4. Varimax-Rotated Factor Plot: Supplier Pre-qualification

Moreover, for further analysis of the plot attention is to be paid to scaling: the fulfilment degree has to increase with rising score. For example, the criterion PRICE has to reach its top score for being "very below average". The inclusion of an ideal supplier is not feasible, because every procurement case places different demand on the fulfilment degree of the criteria.

For 20 existing suppliers of ARGASSI a supplier's structural analysis with regard to five main criteria (PRICE, QUANTITY, QUALITY, LOGISTICS and SERVICE, all submitted 5-stage ordinal) has been carried out.

After execution of the supplier-rating-system, factor 1 (explanation ratio 46,7%) describes a price/additional performance component, while factor 2 includes direct product characteristics. Finally, clusters (with a height-to-width-ratio of 46,7:36,9) can be constructed in the plot (cf. Figure 5): the members of the clusters C and F represent top performance suppliers ("stars"), as they show outstanding performance while offering a price below the average price. The suppliers in the cluster B ("Dogs") must be submitted to reinforced controls and/or could be developed, because they only achieve below-average performances. Furthermore, suppliers of the clusters A, C, D or even F can be identified as relevant Benchmarking partners. The clusters A, D and E show partly below-average, partly above-average performances; depending on the strategic position and the delivered good of the supplier, these classes may be acceptable. If not, suppliers of the clusters C and F can be used for their development.

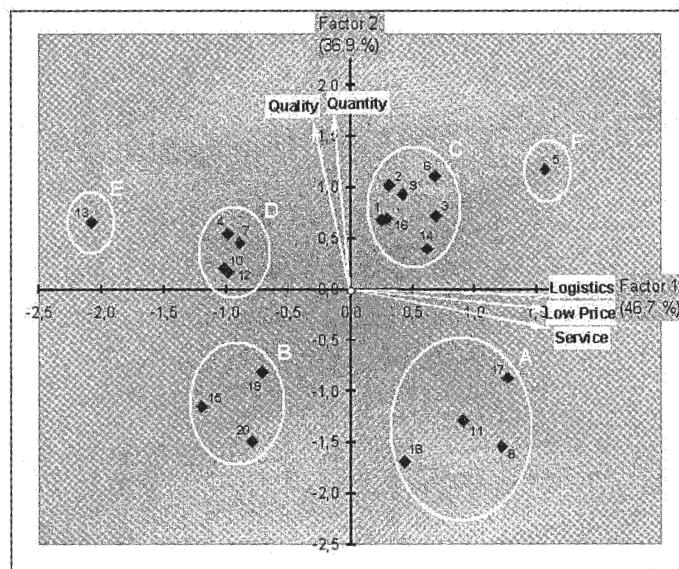


Fig. 5. Varimax-rotated Factor Plot: Supplier Structure

In summary, the designed supplier-rating-system using principal component analysis represents an appropriate rating system which has to run computer-assisted. The inclusion of the company's standard software supplies the data of many quantitative criteria. The resulting plot can be used for pre-qualification, selection and for controlling of suppliers. If the procedure is applied repeatedly, a dynamic observation of the suppliers is guaranteed. In this dynamic monitoring, changing circumstances, to the requirements of the market's adjusted criteria safeguard a performance measurement of the own supplier's base, always corresponding to real market conditions.

References

- ARNOLDS, H., HEEGE, F., and TUSSING, W. (1998): *Materialwirtschaft und Einkauf*. Gabler, Wiesbaden.
- FAHRMEIR, L., HAMERLE, A., and TUTZ, G. (1996): *Multivariate statistische Verfahren*. de Gruyter Verlag, Berlin/New York.
- GLANTSCHNIG, E. (1994): *Merkmalsgestützte Lieferantenbewertung*. Fördergesellschaft Produkt-Marketing, Köln.
- HARTUNG, J. and ELPELT, B. (1999): *Multivariate Statistik*. Oldenbourg, München/Wien.
- LASCH, R., JANKE, C.G., and FRIEDRICH, C. (2001): Identifikation, Bewertung und Auswahl von Lieferanten. *Dresdner Beiträge zur Betriebswirtschaftslehre*, 56, Dresden.

A Knowledge Based Approach for Holistic Decision Support in Manufacturing Systems

Uwe Meinberg^{1,2} and Jens Jakobza²

¹ Lehrstuhl Industrielle Informationstechnik, Brandenburgische Technische Universität Cottbus, Universitätsplatz 3-4, D-03044 Cottbus, Germany

² Fraunhofer-Anwendungszentrum für Logistiksystemplanung und Informationssysteme (ALI), Universitätsplatz 3-4, D-03044 Cottbus, Germany

Abstract. Within production planning and control (PPC), orders are usually steered according to the production goal system. The priority rules are derived from production. The efficiency of these rules for goal-oriented control is often questioned. Besides this procedure itself can affect conflicts with the goals of other divisions and can also negatively affect the total result of the company, usually with a time-delay. In this article a concept for a knowledge-based system is to be discussed, which helps in overcoming many problems in order to approach the goal of a holistic planning and controlling process of production more closely.

1 Introduction

Companies which produce order-oriented have to master two challenges. On the one hand production has an extensive influence on the external effect of the company by the way of controlling the production and therefore on the short and long-term entrepreneurial success. A condition for a sensible control strategy is that the same appropriate valuation criteria of the orders are considered within production planning and control. On the other hand, their production depends directly on the concrete incoming orders and/or customer demand. Forecast-based production to satisfy future demand is often not economical. The customer requirements are usually not foreseeable and hence products can not be specified before the intake of orders. This leads to fluctuations in the extent of utilisation of production capacity which directly depend on the incoming order stream. External factors influencing production mean that different strategies must be adopted to ensure efficiency.

Most PPC-systems used cannot react to changes on the shop floor and are based on simple functions and single goals (Smed et al. (2000), Yeo (1999)). Knowledge-based systems can contribute to making order ranking in the production planning and control of order-oriented manufacturing more effective in terms of business success.

2 Criteria for order ranking

The functions or rules for order ranking based on the production goal system are used in nearly all PPC-systems (Figure 1). Most popular are the

first-come-first-served rule and the shortest-operation-time rule. By using the first-come-first-served rule, the average lead time of the orders will be held constant. By knowing and realising this time, the goal of minimising time-delay is realisable (Wiendahl (1987)). By using the shortest-operation-time rule, the goal of minimising lead time will be fulfilled (Conway (1960)).

The exclusive validity of the production goal system is only true if the company's operating profit is influenced solely by the manufacturing division at the cost level (Kurbel (1995)). But if yield- or liquidity-oriented goals could be influenced on the receipt site by decisions made by the manufacturing division, this exclusive validity would not be given. The framework for supporting decisions in the PPC-process has to be widened in these cases.

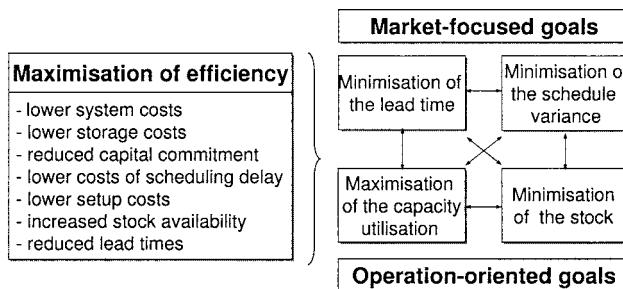


Fig. 1. Target system of Production (VDI 3633).

An approach for the classification of the order valuation criteria has been carried out by Doerner. He characterizes problems, which are to be solved in socio-technical systems by humans, on the basis of the urgency and the relevance (Doerner (1992)). This approach can be transferred to orders.

2.1 Urgency

An order possesses first a concrete date for delivery to the customer and/or from the view of the production a completion date. Moreover, one can have agreed to contractual agreements such as penalties or damage claims for late delivery. For example, an order with agreed payments of compensation for disregarding the date of delivery has different effects on the cost situation of the company than an order with the same schedule slippage situation without these additional agreements in the case of date delay. The first one will therefore be processed first because of the greater urgency.

The long-term negative effect of late delivery on customer relations is however undisputed. Short term bottleneck situations in the manufacturing process or during material procurement (Figure 2) require that such aspects flow fast and objectively into the prioritization of orders.



Fig. 2. Conflict areas of order-oriented production.

2.2 Relevance

Orders are released directly by customers, i.e., the concrete customer acquisition and support is mostly aligned with the public relations, marketing and selling activities of companies manufacturing order-oriented. This expenditure has to be considered within the order control of the manufacturing process accordingly.

Each customer does not have the same share in the turnover of a company. Customers with a high turnover rate have, therefore, mostly a higher relevance for a company than customers with small turnover rates.

However, if the margins of the individual orders are not equivalent to each other, the turnover rate is only of limited suitability for the exclusive characterising of the relevance of orders.

Sometimes the customer honors short term deliverability with price surcharges. This means a higher margin for the manufacturer. If heterogeneous products, for example products with different market positions, are manufactured on the same machines, the orders will be valued differently. Certain product groups could have special relevance for the company, irrespective of the concrete customer. They could be viewed as new on the market and could have great importance in the future. Also they could be so-called "stars" (BCG (1974)) because today they have a great market position in a fast growing market which has to be held or extended.

Only the schedule urgency is related to the production target system, especially to the goal of minimizing deviation. The other above mentioned examples for ranking criteria such as turnover rate, margin, product placement and market position are not included in the conventional ranking concepts. These criteria are used in company divisions other than the production division. These are often reasons for the conflicts between the different company divisions. It could be loosely compared to "building the tower of Baby-

lon" because different divisions value each order with different criteria. They speak about them with another language (Figure 2).

2.3 Compatibility

In the sense of an economic handling of lot sizes

- to keep low setup costs,
- to increase system workload by lower setup costs and
- to lower lead time by lower setup time,

the compatibility of two or more jobs in a queue can be measured (Hopf (1996)). The calculatory costs of an order can only be achieved if the real production conditions correspond to the planed calculation, which an economic lot size presupposes. An order is less compatible, if the necessary production capacity of the related workable orders in the queue is lower than the necessary production capacity of another order package for an operation on the same machine. Apart from the characterisation of an order, the valuation of the neighbouring relation of orders (related to a possible processing sequence) would be sensible.

Aside from the compatibility of the jobs themselves, the compatibility of jobs and the necessary working material (machines, tools) can also be very important. For example, tools can indicate an abrasion level, so that only handling of jobs with a lower level of tolerance are permitted, whereas jobs with high precision require a setup of the tools.

An order classification can be made by using the relevance, the urgency and the compatibility (Figure 3). The weighting of the criteria themselves depends on the present company's and competitor's situation.

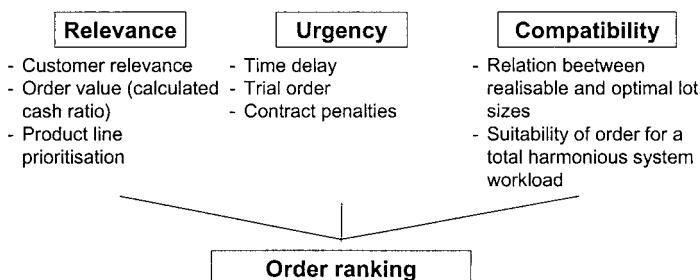


Fig. 3. Example of multi criteria order classification.

3 Strategy determining factors

The criteria for order ranking can be assigned to company goals. Their relevance in the company's goal system depends on the situation, in which the company finds itself. In the inverse conclusion, the height of the weighting of the criterion for order valuation depends on the situation of the company. Figures 4 and 5 show how the margin level and the system load of a company influence the weighting of company goals. The height of the situation-dependent weighting is assignable by simulation. Other connections are also possible. Such connections can depend for example on branches or regional aspects.

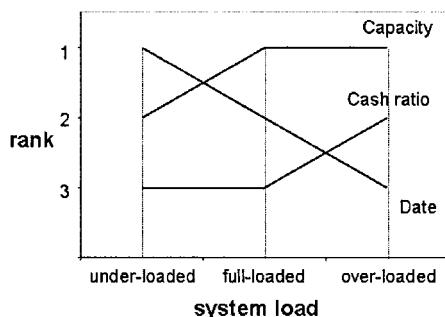


Fig. 4. The strategy determining aspect systems utilisation.

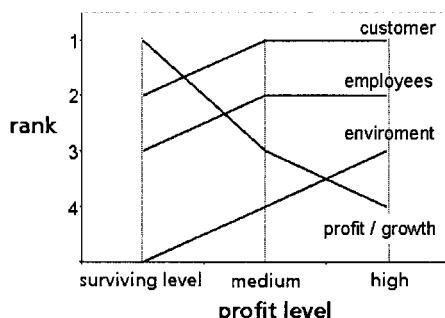


Fig. 5. The strategy determining aspect profit situation (Meffert (2000)).

The approach for a knowledge based decision support system is aimed at supporting decision makers by detecting deviations of general conditions of pre-defined models. Normally, the models for order ranking were used statically and nearly independent from the actual company situation. Only extreme situations would be used as trigger for redefining these models (fire department strategy).

But the models were developed based on the goals of the company. The ranking, importance and dependence of the goals are strongly related to the environment of the company. The relationship between goals and the company's situation and/or environment and the dependency between goals and ranking criteria build the knowledge base for the approach. Changes in the environment or the company situation require a verification of the goal system and the related system models. Today, the reaction time between a change in the environment and the implementation of a modified system model is often too long. Permanent supervision of the environment, combined with an automatic adaption of the related goals and models for required changes, provides a great advantage for decision-making in a dynamic environment.

4 Conclusion

The concept of a knowledge-based decision support system presented allows flexible control of production processes by defining strategies and their situation-dependent utilisation. Companies can react quickly to changing market and framework situations.

This concept was nearly completely implemented for the first time in the PPC-system of a medium-sized cold rolling steel mill with high dependency on their suppliers (steel mills) and a large variation of ordered products.

Different possible company situations were analysed. For each situation a strategy for the weight of each criteria for valuation the orders was determined and implemented into the PPC-system . Total automatic supervision of the company situation and adapting the control strategy could not be realised because the actual integration concept of the software systems in the company did not permit it and could not be changed. The supervision and adaption is carried out regularly by regularly task for the production division manager. He decides upon a control strategy daily. The part of this concept which was implemented showed a high acceptance by all planning managers due to the targeted transparency of the valuation of orders.

References

- THE BOSTON CONSULTING GROUP (BCG) (1974): *Perspectives on Experience*. Boston.
- CONWAY, R.W., JOHNSON, B.M., and MAXWELL, W.L. (1960): An Experimental Investigation of Priority Dispatching. *The Journal of Industrial Engineering*, 11, 3, 221-229.
- DOERNER, D. (1992): *Die Logik des Mißlingens - Strategisches Denken in komplexen Situationen*. Rohwolt, Hamburg.
- HOPF, W. (1996): *Fuzzy Logic zur Steuerung auftragsorientierter Werkstattfertigung*. Peter Lang, Frankfurt am Main.
- KURBEL, K. (1995): *Produktionsplanung und -steuerung: methodische Grundlagen von PPS-Systemen und Erweiterungen*. München, Wien.

- MEFFERT, H. (2000): *Marketing - Grundlagen marktorientierter Unternehmensführung*. Gabler, Wiesbaden.
- SMED, J., JOHTELA, T., JOHNSSON, M., PURANEN, M., and NEVALAINEN, O. (2000): An Interactive System for Scheduling Jobs in Electronic Assembly. *The International Journal on Advanced Manufacturing Technology*, 16, 6, 450–459.
- VDI 3633: *Simulation*. Blatt 1, Beuth, Berlin.
- WIENDAHL, H.-P. (1987): *Belastungsorientierte Fertigungssteuerung*. München, Wien.
- YEO, S.H. and NEW, A.K. (1999): A Method for Green Process Planning in Electric Discharge Machining. *The International Journal on Advanced Manufacturing Technology*, 15, 4, 287–291.

Intelligent Fashion Interfaces – Questions to New Challenges of Classifying

Astrid Ullsperger^{1,2}

¹ Klaus Steilmann Institut für Innovation und Umwelt GmbH,
Bochum/Cottbus, Germany

² Lehr- und Forschungsgebiet Tragbare Elektronik und Rechentechnik
/Juniorprofessur,
Brandenburgische Technische Universität Cottbus, Universitätsplatz 3-4,
D-03044 Cottbus, Germany

Abstract. This paper describes challenges of new intelligent fashion interfaces (ifi's) for context awareness systems, ambient intelligence, wearable electronics, other smart applications and classification consequences. Classifying possibilities in different dimensions will be discussed. One approach for structuring intelligent fashion interfaces could be the 7th skin model. Especially the collection and utilization of data from sensor networks in textiles and clothing will provide new supporting structures for context awareness systems.

1 Introduction

Could fashion products include special electronic functions that make interactivity to the surroundings possible and become part of ambient intelligence? The integration of many formerly discrete physical devices/packages into a compressible, homogeneous, intelligent textile system will provide a new set of revolutionary capabilities. Fibers or textiles, as key-intelligent fashion interfaces, will be able to sense, act, store or emit data from context situations and in future also to listen or to speak. Interfaces to communicate with e.g. machines, artifacts, devices with embedded technology or networks, like WPANs, WLANs are nearly unknown for textile, apparel and fashion industry. In this context intelligent interfaces are defined as items with integrated electronics or computing ability, communications, power resources, software and with stylish demands. For the discussion e-textiles/fibers are the most important ifi's of the future, like wearable computing technology with new forms of human computer interaction and textile based technology, e.g. the Sensate Liner (Lind et al. (1997)). The new field, defined as pervasive or ubiquitous computing, needs new methods, classifications and standards to create useful user interfaces and surfaces to make the new technology accessible for special markets and later for everybody. In this paper effort will be made to classify the fashion interfaces as basis for the convenience of consumers, retailers and developers and for the improvement of quality standards. The paper focuses on technical intelligent fashion interfaces like e-textiles, sensor

buttons, blue-tooth collars, infrared shoes as well as social intelligent fashion interfaces. Another approach is to classify the effects of these interfaces for consumers and consumer behavior, to reduce risks for privacy, health and other issues. Instead of slogging to create brand-new interfacing techniques for each new product a classification scheme could help to combine only the most efficient solutions.

2 State of the art and prospects

With the miniaturization of electronic devices the kind of utilizing computers will change rapidly too. New technologies, new designs, new processes and, most importantly, a new philosophy for designing and fabricating large-scale information systems are crucial for success. As one symbol of new generation a bulky precursor of an IFI, the Wearable Personal Computer (WPC), was invented by Steve Mann in 1980.

Even if not fashionable or textile still now, examples as MA V, Poma, CharmIt, show possibilities and new applications for personal information devices (Post et al. (1997)) with related input/output-devices to wear. The computer infrastructure provides processor power, additional sensing, personal, familiar, dependable and specified interfaces. As progress Lind believes that the next step will be the integration of "...textiles and computers by designing and producing a weavable computer that is also wearable like any other textiles" (Lind et al. (1997)). Main requirements are the wireless connectivity with different networks, like WLANs, thus allowing information to be accessed whenever and wherever the user is in the environment (Barfield and Caudell (2001)). To find new unobtrusive solutions military institutions as well as technical and medical research institutes worldwide were carrying out projects in the textile field and have been presenting innovative ways of integrating smart functions into clothes as demonstrators for some years. A new research field "e-textiles" has been arising (DARPA) with all problems of classifying the new industrial branch of textile electronics (US: Electronic Industry 450 B+, textile industry 480 B+). In the NAICS (North American Classification System) there is no keyword placed even in the library of congress classification. In the 5th framework program of the European Commission Disappearing Computer project, along with National Microelectronics Research Center (NMRC), Ireland, Swiss Material Testing and Research Lab (EMPA), Switzerland, Integrated Systems Design (ISD), Greece and KSI Klaus Steilmann Institut for Innovation and Environment, Germany, the integration and implementation of computational fibers into different objects, machines and even as part of the human will be explored (Ullsperger et al. (2003)). The vision of fiber computing (FICOM) is that whole computers might be made from man made fibers people are comfortable wearing or used to live with. Smart-Textiles are creating a multi-disciplinary infrastructure

that will establish design tools and manufacturing processes to support the integration of capabilities into our everyday use. The underlying conceptual advancement or technology is the development and insertion of intelligent mixed signal yarns and textiles into everyday fabrics and fabric-based infrastructures. Potential computational fibers include glass fibers, silicon fibers, polymeric fibers with electrical properties and metallic fibers. Computational capability has to be coupled with flexibility, washability and wearing comfort. One possibility is to fabricate self-contained modules in large quantities, distributed along the length of a given fiber, where each module would have a power source, sensor, small amount of processing power and an actuator. Simple tasks, as an example, could be changing visual patterns on a shirt based on certain local conditions, thermal regulation could be another one. The resulting physical implementation has implications on the human-computer interaction issues. With intelligent textile technology integrated into e.g. clothes, the user will expect personal interface access. As personalized system it should ensure appropriate responses to every day tasks and be connected to learning context awareness systems. Only with this kind of "invisible" computing integration the basic requirements for wearable computing like constancy and interaction, augmentation and mediation will be fulfilled. The wearable or fiber system can become a soft and user friendly permanent mediator for computers, electrical devices and other interfaces. With complex user models and corresponding software agents, such interfaces can be extended to recognize and predict the resources needed by the user (Starner et al. (1997)). Permanent context and situation information helps the system to work efficient, be affective and learn relationships between context, situation and affective response. For these complicated tasks the research has to find out new methods of data analyzing as well as intelligent solutions for storing and emitting data. Especially legal and standardizing problems will be arising.

One classification attempt could be reflected in relation to the following questions:

1. How can user interact with nearly "invisible" modules that are distributed over clothing? Here should data be generated about the 6 senses (hearing, vision, touch, smell, degustation, thinking/feeling), categories of i/o-devices and user interfaces like woven conductive icons, light signals, special speech recognition systems, fiber displays, pictorial symbols etc..
2. Which consequences will arise if the technology is strongly related to context or situation computing databases in the environment? The community will need new rules and codices in relation to individual authentication and for protected working with data. Who handles, collects and utilizes data? Who needs which kind of context information?
3. Should be explicit control of the modules implemented or built some knowledge into the system for autonomous decision making? Here it is

necessary to categorize groups of devices and interfaces, which summarize data and to build wearable intelligence with self learning systems.

4. How can the user stay in control at all times even if context information systems collect, storage, process and classify millions of information? This process will be strongly connected with processes of standardizing to make sure, that everybody is aware of the embedded systems and has possibilities to interrupt data analyzing processes.

And in relation to the developments of affective computing solutions arise more open questions for the interface design and standardizing processes e.g.:

5. What are the consequences, when a computer "can recognize the user's affective state" (Picard, (1997))? For measuring the "mood" of persons for example integrated sensors in wearable systems might listen to the talk, watch gestures, sense changes in the heart rate, blood pressure, and electrodermal response. Almost any bodily signal might be analyzed for clues to the wearer's affective state.

In this paper there will not be a discussion of the assessment of wearables from the moral/ethical point of view. Well-suited to wearable technology are signals that currently require physical contact to sense, such as electromyogram and skin conductivity. In combination with an intelligent agent the system might learn in dependence to the context or situation to interrupt, provide certain types of information or be quiet. For this, the system has to combine measured factors with a kind of emotional intelligence, which will be a big problem. If the tools enable recognition of it's wearer's affective patterns, an interface should have multiple strategies for responding different emotional expressions. But it should not let the user feel manipulated or worse. And there must be strong limitations for e.g. insurances, businessmen to get open access to all this information to prevent data misuse, because in future we will be surrounded by elementary sensors, actuators, logic, and power sources embedded in different artifacts and combined with reconfigurable network architectures with fault tolerance and operational longevity. The legal and standardizing requirements seem to be infinite.

3 Classification approach with 7th skin model

There are several dimensions by which wearable computers, context awareness and augmented reality systems can be created, interlinked, evaluated and specified. A meta model has to combine data from the following dimensions:

1. Interface dimension (7 skin model: see Figure 1)
2. Context concept dimension (e.g. 15 subcategories)
3. Fashion dimension (e.g. clothes, accessories, colors, styles, forms)
4. Application dimension (e.g. Wellness, Health, Care, Information, Communication, Business, Entertainment, Sports, Security, Protection, Interaction, Control).

The functionality of future fibers, textiles and clothes will correspond to a range of traditional functions (like protection, identification, decoration) and innovative amplifications like communication, information, interaction and simulation. These new functions are the trigger for a paradigm change in the textile industry, because they are coming together with new categories of networks. Even if an individual fashion artifact has limited functionality, it can gain advanced behavior when grouped with other objects through interfaces in special netrastructures. Scalability, mobility, interactivity, flexibility, security, heterogeneity are new challenges in the realization of e.g. context awareness systems. For the creation of completely new applications around smart devices, the theoretical 7 skin model, which subdivides our physical world into seven skins in relation to computational embedded systems, serves as maintenance and infrastructure.

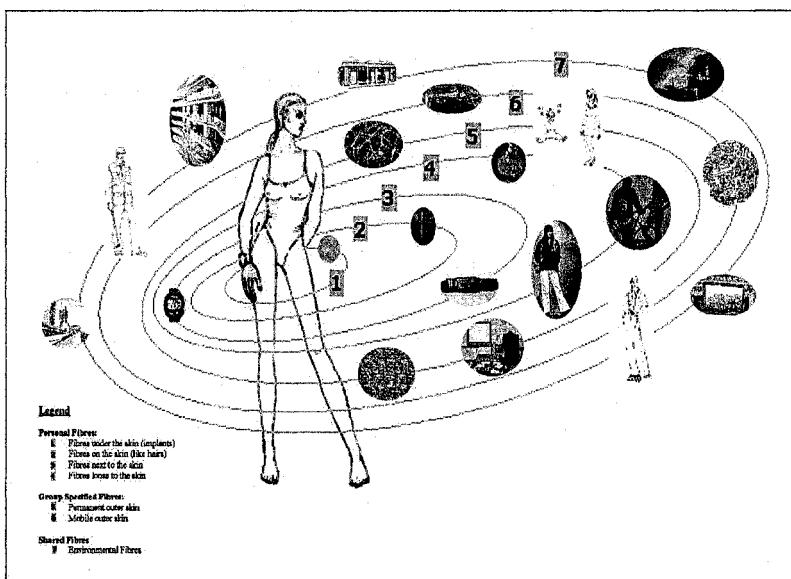


Fig. 1. FICOM: 7 skin model [adapted from Hartmann 5 skin model, 2000]

Disappearing computing applications expect a number of new integrating, standardizing and judging solutions. It should be seriously determined what applications of computing and which location (like under the skin, on the skin, in textiles, in the environment) are important and how to implement such technology. There are many benefits in comparison to integrated microcomputers within the nervous system and other biological systems. Applications could integrate cognitive and sensory prostheses and regulate physiological parameters. For characterising the functions of "smart artefacts" 3 main groups as classification basis for a hierarchical system have to be de-

fined: PRD - personal related devices; GSD - group specified devices and USD - unlimited shared devices. Characteristics of computational PRD's are e.g. collection and storage of individual information; processing of personal information; individual communication interfaces; flexible, comfortable, invisible, visible or touchable; personal input and output devices; human-compatible (healthy, soft, sensitive, wearable). The GSD's are characterized by collecting and storing of social or collective interesting information; processing of common information; open access for communication interfaces; flexible, stable, invisible or touchable; input and output devices for group specified persons and infrastructure compatible (Distinctive, Social, Ecological). Examples e.g. for fibers in the PRD group and in the different 7 skin levels are: 1st level - computational fibers (cf) under the skin (e.g. Implants, "get chipped"); 2nd level - cf on the skin (e.g. hairs); 3rd level - cf close to the skin (e.g. underwear); 4th level - cf loose to the skin (e.g. outerwear, accessories, bags, luggage, shoes); in the GSD group like in the 5th level - permanent outer skin (e.g. smart homes, furniture textiles, wallpapers, home textiles); 6th level - mobile outer skin (e.g. smart transportation - automobile). The 7th level refers to the global skin (e.g. Environment and Infrastructure) and belongs to the USF group. Apart from individual functions the mobility and flexibility are the most important distinctive marks.

For characterizing only the personal related devices, the following data categories have to be collected, processed, stored or transmitted and used by ifi's:

IFI categories	Data categories (samples)
1. personal body data	blood pressure, temperature, skin color, moisture level
2. user identity data	size, interocular distance, color of eyes, hair, age, profile
3. companions	humans (sister, friend), devices (PDA, memory stick, camera)
4. location	indoors (building, office)/outdoors (vehicle, walk, region, landscape)
5. time	e.g. morning, midday, night
6. physical environment	notebook, chair, lamp
7. computing resources	CPU, data rate, speed
8. context sensing	sounds, touch, smell
9. augmented reality	background information, history, price, circuit diagram
10. localized information	events, weather, government, business hours
11. context-based retrieval	former user, accident, highlight for experts, run
12. situated reminders	switch button, try the fish soup
13. remembrance agent	call mother, take gloves
14. appropriate interfaces	i/o-devices screen, keypad, color change
15. monitoring	engine, contest

4 New technical and social questions

Numerous databases with context information have to be build up and integrated in context awareness systems. For this, the interface dimension should be evaluated - that means available existing interfacing techniques from all over the world have to be examined and analyzed in relation to their relevance for new textile, clothing and fashion applications. International standardization is a must for progress and market acceptance by the customer and should be readily tackled for the benefit and risks of all users and developers. Small sensors are capable of detecting a multitude of different environmental parameters. Advances in short-distance wireless networks, like WLINK or the current Bluetooth standard improve possibilities of communication and information interchanges and wireless personal area networks. Such Blue-Tooth systems allow up to seven stand-alone devices to form a modular, reconfigurable system. Thus the user can take advantage of combined functionality while retaining the ability to exchange individual devices. However this type of integration does not allow for an efficient reuse of resource (e.g. user interface). Currently there exist two approaches to interface integration: embedded models (users interact directly with devices embedded in the environment) or portable (Trevor et al. (2002)) / wearable. Studies were worked out in relation to design personal interaction points (PIP's) for personalizing shared pervasive devices. The most common examples for portable interfaces are combinations of PDA's, mobile phones or digital watches with a camera, a GPS receiver or other I/O-devices. Drawbacks of such devices are their lack of modularity, which makes it impossible to exchange just a single appliance, the inflexibility and the lack of affectivity. In addition, such combined devices are often much bulkier than individual appliances. On the other side, by allowing several appliances to share interfaces, like textile sensory pressure applications the overall system size and weight could be significantly reduced. The whole system would become more user friendly.

Even the fashion dimension, which integrates all questions of typical parameters for the products of this branch, e.g. design, colour, style, cut, pattern, clothing type has to be analysed. Fashion designer have to become more industrial product designer with basic functional knowledge. The integration of "hardware" into jackets, bags etc. requires completely novel assembly and manufacturing technologies.

If the interface, context and fashion dimension have to be combined with the application dimension finally, to create a multidimensional space for IFIs, a quite large number of classifying and standardizing tasks have to be solved by international and interdisciplinary research and development groups. Market research activities carried out by KSI have shown that a great proportion of customers in various target groups are very interested in amplifications of textile functions, especially those related to health, care, wellness, protection, security and entertainment (Hartmann et al. (2002)).

5 Outlook

Interesting research questions for the field of classification and standardizing could be for instance the development of new categories for ambient intelligence. New production processes and perhaps even new machinery will have to be invented just for the integration of computational fibers with normal fibers. Various interfaces will have to be considered for ambient intelligent products. International standards and classifications should be established for more efficient utilization of research resources. To continue the discussion process for IFIs there will be space and links under ICEWES -www.icewes.net. Achieving standardization of interface techniques will contribute to the generation, distribution and use of knowledge for ambient intelligent products.

References

- BARFIELD, W. and CAUDELL, T. (2001): *Fundamentals of wearable computers and augmented reality*. Lawrence Erlbaum Associates, London.
- HARTMANN, W.D. (2000): *EVONETIK - 10 Erfolgstipps für Netzwerke*, transnovation 1.1., Essen.
- HARTMANN, W.D., STEILMANN, K., and ULLSPERGER, A. (2002): *High-Tech-Fashion*. Brainduct digital edition, Bochum.
- LIND, E.J., EISLER, R., BURGHART, G., JAYARAMAN, S., PARK, S., RAJAMANICKAM, R., and MCKEE, T. (1997): A Sensate Liner for Personnel Monitoring Applications. In: *First International Symposium on Wearable Computers, IEEE, October 13 - 14, 1997, Cambridge, Massachusetts*, 98-107.
- PICARD, R.W. (1997): *Affective Computing*. MIT Press, Cambridge.
- POST, E.R., ORTH, M., RUSSO, P.R., and GERSHENFIELD, N. (2002): Embroidery: Design and fabrication of textile-based computing, in: *IBM System Journal, 39, 3 and 4, 2002*.
- STARNER T., MANN, S., RHODES, B., LEVINE, J., HEALEY, J., KIRSCH, D., PICARD, R.W., and PENTLAND, A.P. (1997): Augmented reality through wearable computing. *Presence, Vol. 6, No. 4, August 1997*, 386-398.
- TREVOR, J., HILBERT, D.M., and SCHILIT, B.N (2002): Issues in Personalizing Shared Ubiquitous Devices. In: *Proceedings of UbiComp, 2002. 4th International Conference, Göteborg, Sweden, September/October 2002*, 56-72.
- ULLSPERGER, A. et al. (2003): Flexible Silicon Functional Fibres. In: *Seventh International Symposium on Wearable Computers, IEEE, October 21 - 23, 2003, NY, USA*.

Full Factorial Design, Taguchi Design or Genetic Algorithms – Teaching Different Approaches to Design of Experiments

Ralf Woll and Carina Burkhard

Lehrstuhl Qualitätsmanagement,
Brandenburgische Technische Universität Cottbus, D-03013 Cottbus, Germany

Abstract. Design of Experiments (DoE) is of high interest in engineering to get robust and reproducible results with a minimum of experiments. Based on a simple experiment different approaches for DoE and experiences from an educational program will be shown. The simple example demonstrates how DoE can be applied for student education in an effective way. It shows as well how to teach some of the stepping stones of DoE. A proposal for a complex training module in DoE will be given.

1 Introduction to Design of Experiments

There are many reasons to deal with Design of Experiments (DoE) in the education of engineering students. With this methods quality can be systematically improved, processes can be improved for cost-efficiency and experiments can lead to results in a short period of time. DoE is an approach to do experiments systematically, saving time and money due to the reduction of the number of experiments and to maximize significance. DoE is one element of the Six Sigma Strategy which is a zero-defect approach. Even though DoE leads to scientific correct experiments and to stable processes many engineers and scientists do not know about DoE. This is why DoE is an essential part of the educational program "Quality Engineering" at BTU Cottbus. It is part of this program in applied statistics where the main focus is on the Analysis of Variance (ANOVA), the DoE-methods of Shainin and Taguchi as well as Genetic Algorithms.

Design of Experiments has been invented to improve significantly product- and process quality. A detailed analysis should give necessary information with minimum efforts (Franzowski (1994)). In the beginning of systematic experimentation in the 16th century the one factor approach was applied. Only one factor of all factors was to be changed to measure its influence on the result of an experiment. With that an experiment with six factors only six combinations were possible. Another milestone in DoE was set by changing all factors against each other. Each factor could therefore be set to two or more positions. With his approach an experiment with six factors on two positions would lead to $2^6 = 64$ combinations - all possible combinations

on two positions. With this set of experiment all results of all combinations could be determined. However, the amount of experiments was very high with increasing amount of parameters.

In the 20th century in DoE research the target was to reduce the number of combinations while keeping the message of the experiment's outcome clear. Approaches, such as the Variance of Analysis, based on research of Sir Ronald Aylmer Fisher, were introduced (Kühlmeyer (2001)). His method is an approach to find out which factors influence the results most. Taguchi, another pioneer in DoE, used methods to find out how important the factors were concerning the results. And Shainin had put together methods that restricted many factors to get the most important factors. The most recent research in DoE are the genetic algorithms. This approach uses methods, which emulate the natural selection process (Krottmaier (1994)). In student education the recent streams of DoE are applied on a catapult.

2 The catapult experiment

The catapult experiment is a simple model for any experiment and very useful for education. With a small catapult balls are thrown with the arm to throw them as far as possible. The catapult has been first described of Schubert et al. (1992). Their intention was to use the catapult for engineering education to get practise in the application of statistical tools. They assume this to be necessary to learn methods which enable engineers to shorten development times and to increase quality. Their experiment showed the engineering approach as well as a blending of the engineering with a statistical approach. The engineering approach mainly consisted of analysis of mechanical effects and a detailed calculation on how far the ball would be thrown. For the study 200 man-hours were necessary. The statistical approach on the other side lead to an even more accurate result in only 6 man-hours. The catapult experiment is also part of many six sigma trainings, e.g. Magnussen et al. (2001).

For the experiments in the education of engineering students a wooden catapult is here used (see figure 1), manufactured by Quality Science, Produktionstechnisches Zentrum Berlin. This catapult is able to throw items the size of a ping pong ball from around 1 to 4 m. A ball is positioned in the ball holder, the arm will be drawn backwards, will let go off and the ball is thrown forward. To be able to always keep the same deflection angle an element is positioned behind the catapult. To make the point visible where the ball hits the ground, the ball is moistened with water before the throw.

For the experiment the catapult is fixed to a row of tables so that the ball will hit the ground the same height as the foot of the catapult. This is necessary to make the experiment's results comparable and independent from the table height. A measuring tape will then be placed on the table to measure the throwing distance of the ball. The goal of the experiment is to find the adjustment for the catapult which allows the catapult to throw

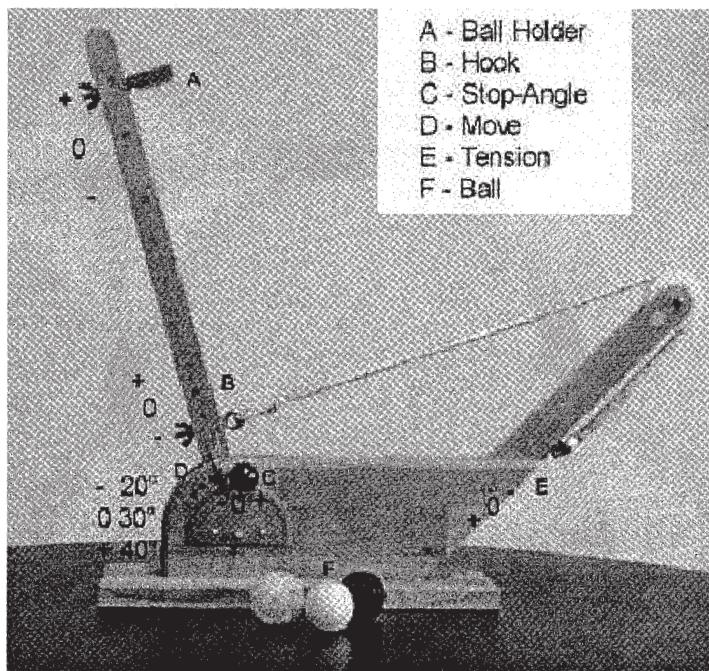


Fig. 1. A wooden catapult with six influence variables for education in DoE

the ball as far as possible with low scattering. The general approach to DoE is to carry out a system analysis, to choose a DoE strategy and to realize and analyse the experiment. The approach to the system analysis is in the case of the catapult quite clear: The catapult has six actuating variables which can all be changed into three positions. Possible influences and process parameters should be investigated. An aid can be the cause-effect-diagram to find the influence of the 6 M - man, machine, method, material, measurement, milieu. The main goal at the beginning of the experiments is to teach system analysis with regard to Shainin's famous saying: "Don't let the engineers do the guessing; let the parts do the talking" (Bhote (1991)).

3 Accomplishment of the experiments with different DoE strategies

For the student laboratory three DoE strategies are part of the educational program: Shainin, Taguchi and Genetic Algorithm. Groups of three to five students are working with one strategy on a catapult. One of the groups determines the reliability of the experiment by carrying out a test on normal distribution. This is however not referred to in this article.

The DoE strategy of Shainin identifies the most important influence factors by combining several methods, such as paired comparison, multi-vari charts, component search and others. In the student education the method of variables search is applied. Target of the variables search is to find the parameters with the main influence on throwing distance.

The variables search with six influence variables is carried out with sixteen experiments. Before the experiments low and high settings are specified. Low settings are those which are regarded to influence the result of the experiment negatively, high settings are those which are regarded to have a positive influence. The results of the variables search are shown in figure 2. The upper line shows the experiments where the determined variables are low while all other variables are on a high position. The lower line belongs to all experiments where the determined variable is high while all other variables are low. With the settings of the variables search, the balls are thrown from 77 to 398 cm. With the diagram above it is possible to identify interactions of influence factors. Further investigations have shown an interaction between hook (B) and move (D) setting.

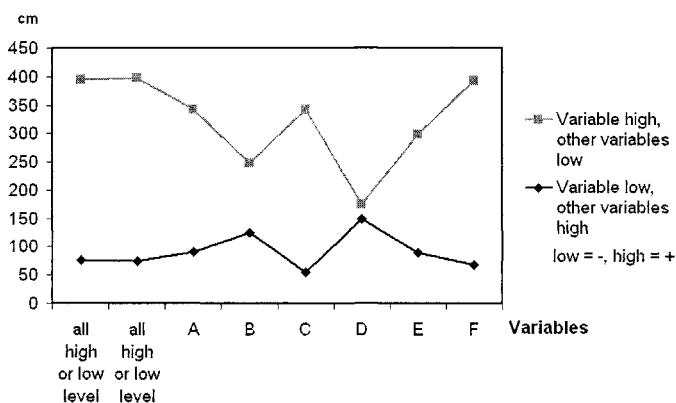


Fig. 2. Result of the variables search experiments

Parameter Design, the DoE strategy of Taguchi, is the second strategy used in the laboratory. Taguchi's philosophy implies a robust process must not only adjust to the target value but also be insensitive against scattering (Pfeifer (2002)). Nine experiments have been carried out following the Taguchi $L_9(3^4)$ orthogonal array (see figure 3). The abbreviation means nine experiments with four factors on three positions (Hicks and Turner (1999)). As the catapult has six influence variables two of them are set to the highest level, which is tension (E+) and the squash-ball (F+). The results from those

settings vary in average from 138 to 428 cm. With these results the influence of each factor is calculated, which is shown with the effect diagram in figure 3.

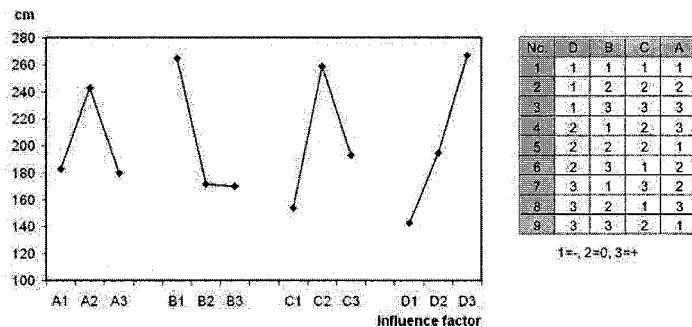


Fig. 3. Effect diagram of a Taguchi $L_9(3^4)$ and table with the settings of the catapult for the experiment. As Taguchi uses other designations than classical DoE methods, the pattern is explained below the table.

The third DoE strategy applied is the strategy of Genetic Algorithms. This strategy follows the principle of evolution. With the help of a random number strings are developed to the optimum. String is the term for the different settings of the catapult. With this strategy there is no specified number of tests. The number of tests is linked to random. The experiment is carried out on two positions with each variable factor. The best setting has been found after 44 different positions, with 398 cm. However, already with ten different settings the width of the ball has been 393 cm.

4 Discussions

Three different approaches of DoE have been applied to one type of catapult. The results are summarized in table 1. The results of a full factorial design are added to verify the results. This DoE has not been part of the student laboratory, but has been part of an other study project.

The results listed in table 1 are then discussed within the student laboratory. From the different designs, the Taguchi Design needed least amount of experiments for the best result. However, it is likely, that this result occurred randomly. As Taguchi Design like the approached neglects interactions, knowledge on interactions is necessary and further tests have to be carried out.

Every experiment with a certain design is part of full factorial design. For Shainin Design this is only possible if the experiments are repeated at each setting. It is further an important statement, that the Taguchi approach is not in contradiction with factorial experiments. A Taguchi $L_8(2^7)$, e.g., the

Design of Experiment	No. of experiments	Throw Distance [cm]	Combination
Full Factorial Design	729	428	A0, B-, C+, D+, E+, F0 A0, B-, C0, D+, E+, F+
Variables Search	16	398	A+, B+, C+, D+, E+, F+
Taguchi Design	9	428	A0, B-, C0, D+, E+, F+
Genetic Algorithms	44	428	A0, B-, C0, D+, E+, F+

Table 1. Results of different types of Designs of Experiment

most incomplete DoE for 7 parameters is equivalent to the factorial design 2(7-4) for three parameters (see figure 4).

The goal of the educational program is to empower students to do scientific and efficient experiments. Therefore a profound knowledge of statistical methods is necessary. First approach to gather this profound knowledge is to teach the basic knowledge on statistical methods. This approach is rather behaviouristic: state-of-the-art of DoE is taught and is considered to be objective and common sense on DoE in the class room. Each student should study, learn and reflect the same things to make the world objectively the same for everyone and to have a common starting point (Blumstengel (1998)).

However, the achievements are rather soon forgotten, if they are not practised. Therefore, the constructivist approach is here applied with the application of an experiment. The learning paradigm of the experiment is to construct, to profit from the cooperation of a team, to interact, to cope with different situations and to learn with a teacher who rather coaches than teaches (Holzinger (2000)). A feeling of insecurity, about what is right or wrong can occur with this approach (Arnold and Siebert (1995)). Students learn subjectively but search for guidelines on what they experience is among the possible results. This is why the two approaches, behaviouristic constructivist, can be found together in the educational approach here described.

The experiment has been shown to be useful for the students to realize that a systematic DoE approach together with engineering knowledge leads to results faster than a simple engineering approach, see also Schubert et al. (1992). It has been shown that engineering students with their technical background achieve rapidly ideas and solutions through brainstorming. Often they are not aware that their first ideas are very close to the solution they are seeking for. Here, the teacher-coach is needed to show students to be systematically when, in their own eyes, they are only wondering but what they really do is working with an engineer's tool.

Education should confront students with problems in DoE as well. One example for this is the ball holder. Position A+ should bring the best results, concerning to the students system analysis. However, A0 is a better adjustment than A+. This is because the ball is rolling back during the throw and

(2²)

A	B
-	-
+	-
-	+
+	+

(2³)

A	B	C
-	-	-
+	-	-
-	+	-
+	+	-
-	-	+
+	-	+
-	+	+
+	+	+

(2⁴)

A	B	C	D
-	-	-	-
+	-	-	-
-	+	-	-
+	+	-	-
-	-	+	-
+	-	+	-
-	+	+	-
+	+	+	-
-	-	-	+
+	-	-	+
-	+	-	+
+	+	-	+
-	-	+	+
+	-	+	+
-	+	+	+
+	+	+	+

(2⁴⁻¹)

A	B	C	D	AxBxC
-	-	-	-	-
+	-	-	-	+
-	+	-	-	+
+	+	-	-	+
-	-	+	-	-
+	-	+	-	-
-	+	+	-	-
+	+	+	-	-
-	-	-	+	-
+	-	-	+	-
-	+	-	+	-
+	+	-	+	-
-	-	+	+	-
+	-	+	+	-
-	+	+	+	-
+	+	+	+	-

(2⁷⁻⁴)

Nr.	A	B	C	D	AxB	(m)	AxC	F	BxC	G	AxBxC
1	-	-	-	+	+	+	-	-	-	-	-
2	+	-	-	-	-	+	+	+	-	-	-
3	+	+	-	-	+	-	+	-	-	-	-
4	+	+	+	+	-	-	-	-	-	-	-
5	-	+	+	+	-	-	-	-	-	-	-
6	+	+	+	-	+	-	-	-	-	-	-
7	+	+	+	-	-	+	+	-	-	-	-
8	+	+	+	+	+	+	+	+	-	-	-

Taguchi L₈ (2⁷)

No.	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2
a	b	a	c	a	b	a	
b			c	c	b	c	

Fig. 4. Different design of experiments can be transferred into other experimental designs

does not get the full power of the jib. This irregularity has not been assumed with the system analysis. Students did not have this information before the experiments. This is how they realized that even with a system analysis it can not always be assured that all problems can be found in a first experiment. The result of confronting students with this problem lead to vivid discussions on DoE, its advantages and disadvantages.

The most precise experimental results are not valuable without an accurate and detailed documentation. Engineers need to be excellent communicating their test results. This includes the precise system analysis and documentation of the main influence factors as well as to write down discussions beside the experiments. It has been observed, that most of the student groups had been quite close to the final result only two minutes after the beginning of the experiments. At that time the group members still played with the catapult instead of doing serious experiments. In this phase unconventional thoughts appeared which were close to the final solution. If these phrases would have been documented and analysed with the final test results than the utilisation of the DoE strategies could have been even more effective. This is one reason to focus on documentation.

5 Results and conclusions

Design of Experiments is a necessary qualification for engineering students as preparation for a career in production industry and science. It is a tool for cost-efficient and systematic experiments. To show differences, advantages and disadvantages of the different DoEs, student groups should do the same experiments with different DoE strategies. In this article we propose the use of a wooden catapult and the application of DoE of Shainin, Taguchi and Genetic Algorithms. A discussion should then intensify the gathered experiences. An approved starting point for a discussion is a property of the investigated object, here the catapult, which does not come up with a system analysis. The experiments in the student laboratory should show clearly the importance of documentation. All details should be documented which are necessary to understand the experiments any time later. Especially the documentation of the first thoughts before starting the experiments should be written down. Often these thoughts are the key to the solution. Besides all the DoE strategies an important experience should be transferred with the laboratory. Engineering students should be aware that they are engineers which means to first of all think about the problem before using any method.

References

- ARNOLD, R. and SIEBERT, H. (1995): *Konstruktivistische Erwachsenenbildung*. Schneider-Verlag, Hohengehren.
- BHOTE, K.R. (1991): *World Class Quality*. Amacom, New York.
- BLUMSTENGEL, A. (1998): *Entwicklung hypermedialer Lernsysteme*. Wissenschaftlicher Verlag Berlin, Berlin.
- FRANZOWSKI, R. (1994): Annahmestichprobenprüfung. In: W. Masing (Ed.): *Handbuch Qualitätsmanagement*. Hanser, München.
- HICKS, C.R. and TURNER, K.V. (1999): *Fundamental Concepts in the Design of Experiments*. Oxford University Press, New York.
- HOLZINGER, A. (2000): *Basiswissen Multimedia, Band 2: Lernen*. Vogel 2000, Würzburg.
- KROTTMAIER, J. (1994): *Versuchsplanung*. Verlag TÜV Rheinland, Köln.
- KÜHLMAYER, M. (2001): *Statistische Auswertungsmethoden für Ingenieure*. Hanser, Berlin.
- MAGNUSSON, K., KROSLID, D., and BERGMAN, B. (2001): *Six Sigma umsetzen*. Hanser, München.
- PFEIFER, T. (2002): *Quality Management*. Carl Hanser Verlag, München.
- SCHUBERT, K., KERBER, M.W., SCHMIDT, S.R., and JONES, S.E. (1992): The Catapult Problem: Enhanced Engineering Modeling using Experimental Design. *Quality Engineering*, 4(4), 463–473.

Part VIII

Medicine and Health Services

Requirement-Driven Assessment of Restructuring Measures in Hospitals

Werner Esswein and Torsten Sommer

Lehrstuhl für Wirtschaftsinformatik, insb. Systementwicklung,
Technische Universität Dresden, D-01062 Dresden

Abstract. From 2004 on German hospitals will face a changed environment and must meet new requirements as the payment system will then be based on DRGs (Diagnosis Related Groups). Hospitals will have to reorganise their internal structures and processes to make sure that the case-based lump sums will cover the costs of the respective treatments. To evaluate the reorganisation projects, a criteria catalogue based on hospitals' requirements is necessary. This article aims at proposing such a catalogue along with a process for the assessment of restructuring measures. This qualitative view is supplemented by a quantitative assessment of restructuring measures.

1 Introduction

The German government puts forward the introduction of a DRG-based prospective payment system which will replace the payment of treatment-day equivalents. This drastic change in hospital financing will result in a radical change in incentives. It has so far been most lucrative to keep patients in hospital for as long as possible and to thus redeem many treatment-day equivalents. In contrast, case-based lump sums with an amount almost independent from the length of stay are profitable when the length of stay is short and fewest possible treatments are performed. Considering this shift in incentives, the legislator has obligated the hospitals to quality management (QM). As a result, the hospitals face requirements of the DRG-introduction and of altered general conditions and have to implement restructuring measures in the short term.

2 Attribute scheme

The evaluation of these restructuring measures can be done in different ways. This article proposes a requirement-driven approach, that is to evaluate the restructuring measures based on previously set requirements. Requirements and measures being described by the same attributes is a crucial prerequisite to a requirement-driven assessment of restructuring measures.

Hence, an attribute scheme (or criteria catalogue) must be initially selected, which preferably embraces all aspects of hospitals' requirements. This way, adaptions of the attribute scheme later on can be avoided.

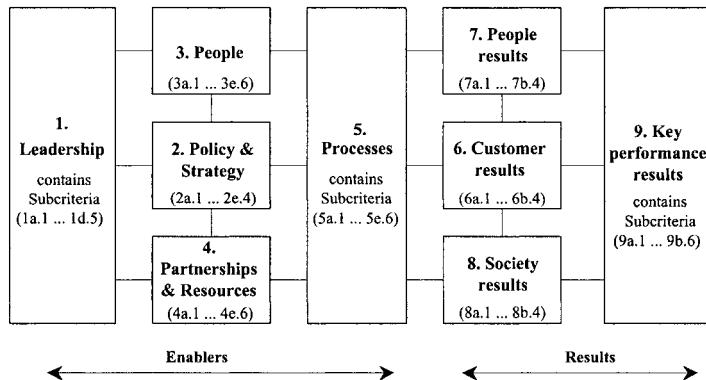


Fig. 1. Structure of the EFQM-Model, according to (Schubert (2001), p. 113)

In the authors' opinion, the EFQM-Model fulfills these criteria and can serve as an attribute scheme. In the context of this article the focus is not on the EFQM-Model's characteristic as a QM-model, but on its mere structure where subcriteria are used as attributes.

3 EFQM-Model

Along with other models of QM, the model of the European Foundation for Quality Management (EFQM) has achieved widespread acceptance (Brandt (2001), p. 7). It was initially developed to help European industry overcome their quality problems and supported by the European Union as a counterpart to the Baldrige Award in the U.S. (Zink (1998), p. 89). Noteworthy is the EFQM-Model's holistic perspective using nine blocks of criteria to assess an enterprise's QM-system (see figure 1). The processes as well as employees and key performance figures are taken into account. Therefore, the structure of the EFQM-Model shall be used to assign attributes to both requirements and restructuring measures.

This seems to be meaningful for a number of reasons: EFQM is a model that has become more and more relevant in the service sector and has been applied in hospitals over the past years. With its holistic concept EFQM covers an enterprise's range of aspects and doesn't only concentrate on particular areas such as structural quality or the like. Therefore, all requirements irrespective of what type shall be able to be assigned into this structure.

4 Process of the assessment of restructuring measures

Requirements are usually gathered in a goal-oriented manner, e.g. to gain a specification for future changes, to assign priorities or to be able to tell

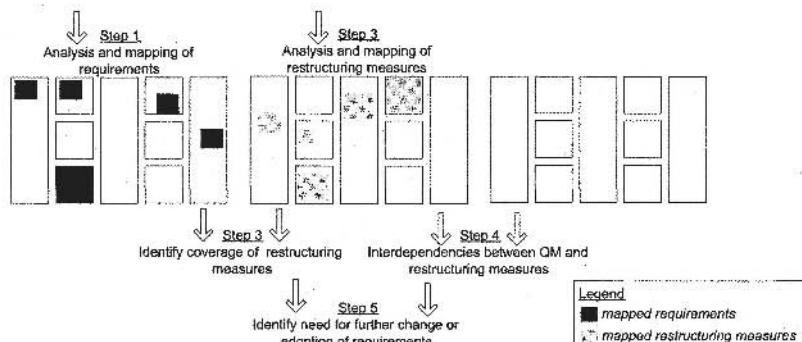


Fig. 2. Process of assessment of restructuring measures

whether restructuring measures (RM) have been successfully implemented. However, gathering the requirements and their specification is a necessary preliminary step.

In the first step of the proposed process (see figure 2) the refined requirements are assigned to the subcriteria of the EFQM-Model and given the respective subcriteria as attributes.

Afterwards the RM undergo the same procedure in step two. Their impact on the subcriteria is examined and they will be assigned attributes accordingly.

Step three is basically the same as step two. Here the RMs' impact on the subcriteria is examined and they will be assigned attributes accordingly. In step three, the measures' coverage regarding the requirements can now be evaluated. (This is a qualitative evaluation in first place, a quantitative approach is presented in section six.) This way the measures' capability of fulfilling the requirements become apparent and the need for further change can be identified.

As a result of the RMs' assignment into the structure of the EFQM-Model carried out in step three, it is now easy to reveal interdependencies between QM and the measures as step four.

In the fifth step it can be decided whether the requirements need further refinement or adaption. Also, the necessary style and extent of the RM can be determined. If certain requirements are not covered, it has to be decided whether additional measures have to be taken or whether the measures originally planned will have to be substituted by measures better covering the requirements. It can also be the case that, as a result of step four, changes in certain fields of QM are necessary or that the RM have to be adapted to meet certain quality standards.

5 Exemplary application of the process

According to the process described above, the restructuring measure 'introduction of CaseMaps' shall now be compared against the requirements 'efficient process execution' and 'consideration of medical guidelines'.

5.1 Preparation (Gathering and specification of requirements)

The change of the hospital payment system to a DRG-based system implies far-reaching consequences for the hospitals. If a treatment causes higher costs than covered by the case-based lump sum the hospital will have to shoulder the loss. A cost-effective treatment is therefore indispensable (Neubauer and Zelle (1996), p. 25) and processes have to be executed efficiently. This can only be achieved when processes are well documented, regularly reviewed and when process alternatives are assessed based on processes' performance figures. Monitoring the fulfillment of process goals requires the assignment of resources to the steps of processes. All persons involved need to know the ideal processes and have to be informed in case of changes to the processes.

The consideration of evidence-based guidelines (e. g. by means of clinical pathways) becomes an additional requirement and is intended by the legislator (section 137e (3) (1) SGB V). That implies that processes follow a pre-defined way - there must be an ideal model of the respective processes whose execution ensures that the guidelines are met. As deviations from guidelines are allowed in well-founded cases, a patient-related documentation has to be established where deviations from the original process can easily be seen and justifications can be documented.

5.2 Step 1 (Analysis and mapping of requirements)

After requirements have been gathered and refined they are now assigned into the structure of the EFQM-Model. That is, the requirements are provided with the EFQM-subcriteria as attributes.

Leadership: The implementation of regular process reviews is a question of leadership and requires uniform assessment factors and a systematic procedure. Thus the requirement is assigned the subcriteria 'Evaluation of key results' (1b.4) and 'Improvement of processes' (1b.5).

People: The people involved have to know the ideal processes and have to regularly be informed of changes. Hence, training is required, which alludes to the criterion 'Training- and development plans' (3b.2).

Partnerships and Resources: Regular information of people involved in the care process demands that information is gathered and analysed systematically and is available to these people. These requirements can be subsumed under the criteria 'Gathering of information and knowledge' (4e.1) and 'Enable access to users' (4e.2).

Processes: The description of pathways and their regular review is relevant to patient care and thus to the criterion 'Key process design and quality policy' (5a.1). As performance figures need to be defined and gathered, the criterion 'Performance figures and service-goals' (5a.4) is affected.

Results: The assessment of results is essentially based on the goals defined in the area of the Enablers-criteria (Brandt (2001), p. 299). Thus, in this context criteria of the Result-area seem to be irrelevant.

5.3 Step 2 (Analysis and assignment of restructuring measures)

As an example of restructuring measures to be introduced, the introduction of CaseMaps is chosen. CaseMaps essentially relate the days of a treatment plan to goals of the treatment in a matrix-like way. The intersections of days and goals represent tasks necessary to fulfill the given goals. Every instance of a CaseMap documents the treatment of a single patient. (Bollmann and Beck (2002), p. 242). The technique of CaseMaps shall now be assigned into the structure of the EFQM-Model.

Processes: CaseMaps depict the ideal sequence of tasks roughly and are no process descriptions in a narrower sense; this would require the description of dependencies between the tasks. Moreover, the assignment of resources to tasks is not explicitly provided. The rough level of aggregation hardly enables cost assignment and activity-based costing. The fulfillment of the overall goals of the treatment can be read off. Documentation of some performance indicators (e.g. infection severity or pain level) is possible. Summarizing, CaseMaps support the criteria 'Key process design and quality policy' (5a.1) and 'Performance figures and service-goals' (5a.4) to a moderate extent.

Partnerships and Resources: The information contained in CaseMaps can be used as a basis for the measurement and depiction of results and is available for the staff in principle. Therefore the subcriterion 'Gathering of information and knowledge' (4e.1) and 'Enable access to users' (4e.2) are supported.

5.4 Step 3 (Identify coverage of restructuring measures)

Leadership: CaseMaps provide no process of their own for the management of patient care processes. Hence, this subcriterion is not supported.

People: CaseMaps do not provide an own training concept.

Partnerships and Resources: The information gathered with and documented in CaseMaps supports hospitals in creating and maintaining a knowledge base. But who is actually given access to the information is not controlled by the CaseMaps but by the hospital.

Processes: CaseMaps combine a rough process description with a patient-related documentation. Performance indicators and resources can only be assigned to a certain extent. More complex evaluations regarding advanced performance figures can not be carried out on the basis of the information provided by CaseMaps.

5.5 Step 4 (Interdependencies between QM and restructuring measures)

Having used the EFQM-Model so far only as a skeleton to assign requirements and restructuring measures into, its original purpose as QM-model is now considered. In the example QM influences the requirements in so far, as systematic process design (criterion 5a) is based on the principle of continuous improvement (Brandt (2001), p. 209). Regular reviews and adaption of processes thus become requirements, too. CaseMaps do not cover this aspect.

Vice versa, the application of CaseMaps also has implications for QM: Monitoring the process of CaseMap design and assuring certain quality standards that CaseMaps must meet. That relates to the criteria 'Key process design and quality policy' (5a.1) and 'Performance figures and service-goals' (5a.4) (but compared with figures related to medical treatment on a meta-level) and 'Identification of Process improvement possibilities' (5b.1).

5.6 Step 5 (Identify need for further change or adaption of requirements)

It has become evident that the introduction of CaseMaps can not cover all of the requirements discussed. Further requirements arise as it needs to be assured that after evaluation of the CaseMaps, treatment plans are adjusted accordingly and that the persons affected are informed of these changes. It is also of great importance that in case of changes to the technique of CaseMaps itself (e.g. possibility to assign resources) a kind of 'method training' is performed for the people involved, and that a continuous adjustment of the technique according to the needs of its users is guaranteed. That is, a process management regarding the maintenance of CaseMaps has to be introduced.

There is also a need for change regarding the integration of further performance figures into the process of documentation. It has to be decided whether CaseMaps are to be supported by other techniques, or whether the deficits are perceived so crucial that CaseMaps will be completely substituted by other measures. If CaseMaps are used, there is a requirement to introduce other measures to collect performance indicators and to integrate them with the CaseMaps into a unified process to avoid inconsistency.

6 Quantification of the assessment

After requirements and alternative restructuring measures have been assigned to criteria of the EFQM-Model, it might be of interest which (bundle of) measures fulfill the requirements best - a quantitative assessment is needed.

Each requirement r_i out of a set of n requirements (with i running from 1.. n) can be viewed as a vector where index $k(1a.1 \dots 5e.6)$ represents the respective attributes (EFQM-criteria). The same applies to the restructuring measures where v_j represents a single measure.

$$r_i = \begin{pmatrix} r_{i, 1a.1} \\ \vdots \\ r_{i, 5e.6} \end{pmatrix} \quad v_j = \begin{pmatrix} v_{j, 1a.1} \\ \vdots \\ v_{j, 5e.6} \end{pmatrix}$$

It could be argued that the attributes' types are dichotom: a requirement may be or may not be assigned an EFQM-criterion. A closer look shows that this would be too broad and inappropriate for a differentiated assessment. It is therefore proposed to use parameter values of (0..1).

Now distances have to be calculated between the respective parameter values of requirements and of restructuring measures to determine, how well certain measures cover the requirements. However, it would not be meaningful to compare single requirements against restructuring measures. Instead, all of the requirements need to be aggregated and be compared against the bundle of restructuring measures to be introduced. That means that an aggregate requirements object \bar{r} has to be computed from the matrix R of all requirements.

$$R = \begin{pmatrix} r_{1, 1a.1} & \cdots & r_{n, 1a.1} \\ \vdots & & \vdots \\ r_{1, 5e.6} & \cdots & r_{n, 5e.6} \end{pmatrix} \quad \Rightarrow \quad \bar{r} = \begin{pmatrix} \bar{r}_{1a.1} \\ \vdots \\ \bar{r}_{5e.6} \end{pmatrix}$$

The aggregation shall be done differently for requirements and restructuring measures. If several requirements are assigned the same criteria (e.g. $r_{1, 5a.4} = 0.3$ and $r_{7, 5a.4} = 0.6$), it is not clear whether both parameter values overlap or not. Hence, the parameter value for $\bar{r}_{5a.4}$ could lie in between 0.6 and 0.9. To make sure that all requirements are taken into account the single parameter values are summarized.

With regard to the restructuring measures, it needs to be avoided that two measures which only partially fulfill a certain criterion can together appear to completely fulfill it. Thus, overlapping is not assumed and only the parameter value of the restructuring measure with maximum coverage is relevant.

$$\bar{r}_k = \min \left(\sum_{i=1}^n r_{ik}, 1 \right) \quad \bar{v}_k = \max(v_{1k} \dots v_{mk})$$

Focussing on a requirement-driven assessment, only distances of attributes are relevant, where requirements' parameter values are greater than those of the restructuring measures. Otherwise, a measure fulfilling a criterion whose fulfillment is not required could compensate for a deficit regarding a criterion necessarily to be fulfilled. Determining the distance between aggregated requirements and measures is then done as follows:

$$d_k(\bar{r}_k, \bar{v}_k) = \begin{cases} 0 & \text{for } \bar{r}_k < \bar{v}_k \\ (\bar{r}_k - \bar{v}_k)^2 & \text{otherwise} \end{cases}$$

Now an evaluation spanning all criteria can be carried out, delivering a figure that enables the comparison between different bundles of measures. It can be meaningful to assign weights to the criteria (w_k) according to their importance. Here the weights inherent in the EFQM-Model could be applied.

$$d(\bar{r}, \bar{v}) = \sqrt{\sum_{k=1a.1}^{5e.6} d_k(\bar{r}_k, \bar{v}_k) \times w_k}$$

7 Conclusion

This paper illustrates how restructuring measures can be effectively compared against the requirements that hospitals face. The EFQM-Model is used as an attribute scheme, such that the EFQM-criteria serve as attributes on upon which the comparison is based.

Apart from the comparison between requirements and restructuring measures, the very nature of the EFQM-Model additionally allows for a straightforward determination of the interdependencies between quality management and the restructuring measures planned.

Through the all-embracing characteristic of the EFQM-criteria and the universality of the approach suggested it is possible to assess restructuring measures in a wide range of contexts.

Furthermore, the back coupling to the requirements in step 5 and the possibility to carry out the assessment before and after the implementation of restructuring measures supports the principle of Plan-Do-Check-Act (PDCA) demanded by most of the quality management models (KTQ (2002)).

References

- BOLLMANN, M. and BECK, M. (2002): Geplante Behandlungsabläufe und CaseMaps - Wirkung, Nutzen und Anwendungsfelder im Krankenhaus der Zukunft. In: W. Hellmann (Ed.): *Klinische Pfade: Konzepte, Umsetzung, Erfahrungen*. ecomed, Landsberg/Lech, 239–248.
- BRANDT, E. (Ed.) (2001): *Qualitätsmanagement & Gesundheitsförderung im Krankenhaus: Handbuch zur EFQM-Einführung*. Luchterhand, Neuwied, Kriftel
- KTQ (2002): *KTQ-Manual inkl. KTQ-Katalog Version 4.0*. Deutsche Krankenhaus Verlagsgesellschaft, Dsseldorf.
- NEUBAUER, G. and ZELLE, B. (1996): Fallpauschalen: Ein Ansatz zu einer leistungsbezogenen Krankenhausvergütung. In: D. Adam (Ed.): *Krankenhausmanagement: auf dem Weg zum modernen Dienstleistungsunternehmen*. Gabler, Wiesbaden, 19–32.
- SCHUBERT, H.-J. (2001): Von Leistungs- und Prfvereinbarungen zur Umsetzung umfassender Qualitätsmanagementkonzepte. In: H.-J. Schubert and K.J. Zink (Eds.): *Qualitätsmanagement im Gesundheits- und Sozialwesen*. 2. Edn., Luchterhand, Neuwied, Kriftel/Ts., 106–119.
- ZINK, K.J. (1998): *Total Quality Management as a Holistic Management Concept*. Springer, Berlin et al.

Analyzing Protein Data with the Generative Topographic Mapping Approach

Isabelle M. Grimmenstein and Wolfgang Urfer

Fachbereich Statistik,
Universität Dortmund, D-44221 Dortmund, Germany

Abstract. The Generative Topographic Mapping (GTM) approach by Bishop et al. (1998) is used for the classification of sequences from a protein family and the graphical display of the group relationships on a two-dimensional map. The results are compared with the closely related Self-Organizing Map (SOM) approach of Kohonen (1982). A modification of the classical GTM approach is presented, better suited for the analysis of sequence data.

1 Introduction

To get a better understanding of the functional role of the members of a protein family in biochemical processes, it is important to know the internal organization of the family and to detect key regions where interactions with other molecules take place or which are essential for a special three-dimensional structure.

One possibility for the determination of evolutionary relationships between members of a given protein family is to use phylogenetic methods which estimate a phylogenetic tree for the considered sequences. However, phylogenetic methods have also its disadvantages, summarized in Grimmenstein et al. (2002). If possible, *maximum likelihood* (ML) methods should be preferred to other commonly used phylogenetic methods like parsimony, UPGMA (*unweighted pair group method using arithmetic averages*) and distance methods, which are all of heuristic nature. The ML approach has however the great disadvantage that it is especially for protein sequences computationally very demanding and might lead to overfitting in the case of many short sequences. Moreover, it is not necessary to estimate a full phylogenetic tree when the classification of a protein family on only special resolution levels is of interest. The assumption of a tree structure is in addition not always appropriate like in the case of evolutionary events with horizontal transfer of genetic information.

Our aim is to find a biologically meaningful classification of a given protein family into main groups with a visualization of the group relationships. In addition, we want to make inferences about key regions and residues in the proteins determinant for the derived groups and also for the whole family. We therefore choose a different approach for our purposes.

In a recent work (Grimmenstein et al. (2002)) we tried to analyze sequence data from a special protein family named septins by an improved *Self-Organizing Map* (SOM) approach of Andrade et al. (1997), which gives a classification with a visualization of the group relations and allows to detect key sites. However, we found some drawbacks of the SOM approach due to its heuristic nature. These are essentially:

- Different runs of the SOM algorithm with different initializations yield different results.
- There is no global optimization function for the assessment of results.
- The approach is not based on a statistical model.
- The selection of parameters like the learning rate and the neighborhood size has no theoretical basis.
- The convergence of the weight vectors, which represent the classes, is not assured.
- The neighborhood preservation is not guaranteed.
- The Euclidean distance is not a proper measure for protein sequences.

To overcome the limitations of the SOM approach, we try as an alternative for our purposes the *Generative Topographic Mapping* (GTM) approach by Bishop et al. (1998) for the analysis of sequence data from protein families. First, we outline briefly the original GTM approach as given in more detail in Bishop et al. (1998), apply it to a data set of proteins from the septin family and compare the results with former ones derived by the SOM approach. Second, we describe a modification of the classical GTM approach which accounts better for the structure of sequence data.

2 The GTM approach

The GTM approach is in spirit similar to the SOM approach. It models the high-dimensional data space in terms of some latent variables in a low-dimensional space leading also to a classification of the data points with a visual representation of the neighborhood relationships on a lattice, i.e. a map, when we assume a two-dimensional latent space. In contrast to the SOM however, the GTM approach has the advantage that it is founded on a probabilistic framework and overcomes the essential deficiencies of the SOM. The probability distribution of the data is modelled explicitly in terms of the latent variables and a likelihood function serves as global optimization criterion. The convergence of the likelihood – at least to a local maximum – is guaranteed as well as a topographic ordering of the groups on the map.

The mapping is modelled by the GTM approach in reversed direction compared to the SOM which projects the data points on a low-dimensional grid. Points $\mathbf{x} \in \mathbb{R}^2$ from the latent space are mapped with a non-linear mapping function $\mathbf{y}(\mathbf{x}; \mathbf{W})$ into a two-dimensional non-Euclidean manifold in the higher dimensional data space \mathbb{R}^D . The mapping is determined by a

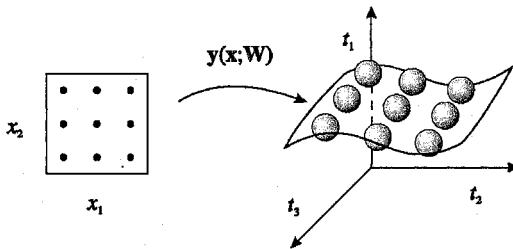


Fig. 1. Illustration of the mapping from the two-dimensional latent space to the data space ($D = 3$) by the non-linear mapping function $y(\mathbf{x}; \mathbf{W})$ (taken from Bishop et al., 1998). By choosing the prior distribution over the latent space as a sum of delta functions, a regular grid is formed in latent space from which each node \mathbf{x}_k , $k = 1, \dots, K$, constitutes the center of a corresponding Gaussian in data space.

parameter matrix \mathbf{W} . To account for noise in real data, a spherical Gaussian distribution is assumed for data vector $\mathbf{t} \in \mathbb{R}^D$ centered on $y(\mathbf{x}; \mathbf{W})$ with inverse variance (precision) β , given latent point $\mathbf{x} \in \mathbb{R}^2$ and parameter matrix \mathbf{W} , i.e. the density function is given by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi} \right)^{D/2} \exp \left\{ -\frac{\beta}{2} \|y(\mathbf{x}; \mathbf{W}) - \mathbf{t}\|^2 \right\}. \quad (1)$$

To determine the distribution of \mathbf{t} independent from latent points \mathbf{x} one has to calculate the integral

$$p(\mathbf{t}|\mathbf{W}, \beta) = \int p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta)p(\mathbf{x})d\mathbf{x} \quad (2)$$

over the latent space, whereas $p(\mathbf{x})$ is the prior distribution over the latent space. If we assume in analogy to the SOM that the distribution of \mathbf{x} in latent space is only centered on the nodes of a regular grid, i.e. given as a sum of delta functions

$$p(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^K \delta(\mathbf{x} - \mathbf{x}_k), \quad (3)$$

we obtain for the distribution of \mathbf{t} , given \mathbf{W} and β , a mixture of constrained Gaussians

$$p(\mathbf{t}|\mathbf{W}, \beta) = \frac{1}{K} \sum_{k=1}^K p(\mathbf{t}|\mathbf{x}_k, \mathbf{W}, \beta). \quad (4)$$

An illustration of the mapping with the distributions in latent and data space is given in Figure 1.

To find an optimal mapping $y(\mathbf{x}; \mathbf{W})$, \mathbf{W} and β (altogether $DM + 1$ parameters) are determined with maximum likelihood. The log likelihood

$$\mathcal{L}(\mathbf{W}, \beta) = \sum_{n=1}^N \ln \left\{ \frac{1}{K} \sum_{k=1}^K p(\mathbf{t}_n|\mathbf{x}_k, \mathbf{W}, \beta) \right\} \quad (5)$$

is maximized for the given data set $\{\mathbf{t}_1, \dots, \mathbf{t}_N\}$, assuming independence for the data points \mathbf{t}_n , with respect to \mathbf{W} and β by using the EM (expectation-maximization) algorithm (Dempster et al. (1977)).

To determine closed form solutions for \mathbf{W} and β , given in Bishop et al. (1998), the mapping function $\mathbf{y}(\mathbf{x}_k; \mathbf{W})$ is selected as generalized linear regression model

$$\mathbf{y}(\mathbf{x}_k; \mathbf{W}) = \mathbf{W}\phi(\mathbf{x}), \quad (6)$$

which has a universal approximation capability. The components of the M -dimensional vector $\phi(\mathbf{x})$ consist of fixed basis functions $\phi_j(\mathbf{x})$, $j = 1, \dots, M$, and \mathbf{W} is a $D \times M$ matrix mapping the output of the basis functions to the data space. The basis functions are chosen by Bishop et al. (1998) to be uniformly gridded Gaussians over the latent space, but also other functional forms are possible.

To assign the data points to groups and for data visualization the mapping from latent to data space has to be reversed. This is done by applying Bayes' theorem. For a given data point \mathbf{t}_n , $n = 1, \dots, N$, the posterior probability has to be calculated for each latent point \mathbf{x}_k , $k = 1, \dots, K$, and then the mode

$$\arg \max_k p(\mathbf{x}_k | \mathbf{t}_n, \mathbf{W}, \beta) \quad (7)$$

is determined, i.e. the data point \mathbf{t}_n is assigned to that lattice point \mathbf{x}_k with the highest posterior probability.

3 Data structure

For the analysis of protein data with the GTM approach we use a multiple alignment of length L from the given amino acid sequences to have a uniform dimensionality in data space. As in the data analysis with SOMs we transform the aligned sequences for numerical treatability in a binary form. Each alignment position in a sequence is described by 20 components according to the 20 possible amino acids. A "1" is assigned to that component, which corresponds to the amino acid present in the considered position. The other 19 components receive a "0". In case of a gap, all the 20 components are assigned a "0". We receive sequence vectors \mathbf{t}_n , $n = 1, \dots, N$, of dimension $D = 20L$.

4 Results

We applied the GTM methodology to a data set containing proteins from a conserved family named septins, which we analyzed before with the improved SOM approach of Andrade et al. (1997). Our data set consisted of the central part of a multiple alignment of 71 septins with altogether 382 positions from different species like humans, rats, mice, fruit flies and yeast. To avoid too much noise in the data, the two termini were cut from the original alignment with 1019 positions because of their large diversity and the many gaps.

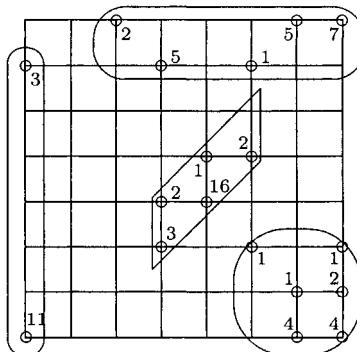


Fig. 2. Classification of 71 septins obtained by the GTM approach on a 8×8 grid with class sizes. The 4 main groups obtained with SOMs are marked.

The alignment was derived by the SP-trEMBL data base (<http://igbmc.strasbg.fr:8080/DbClustal/dbclustal.html>) with a fragment of the innocent bystander protein (SEP1_DROME) from *Drosophila melanogaster* (fruit fly) as query sequence. With the binary transformation of the data we received sequence vectors of dimensionality 7640.

We applied the GTM approach to our data set with different initial conditions. We varied the size of the squared grid in the latent space from 3×3 to 8×8 and also the number of basis function (M) between 4 and 16. With a 8×8 grid, i.e. 64 latent points, and 16 basis functions we obtained clearly the highest likelihood value. The corresponding mapping is given in Figure 2, showing that altogether 18 latent points of the grid are assigned sequences with varying numbers. This derived classification is apart from small differences very similar to the one we obtained previously by the SOM approach on a 5×5 grid with 16 clusters. Altogether 7 clusters of the two classifications are identical and 4 clusters of different sizes from the SOM classification are each subdivided by the GTM, whereas also three complete clusters from the SOM classification and additionally two partial clusters are joined together by the GTM to only one cluster. One of these two split SOM clusters contains a protein (Q9U334), which is wrongly grouped together with another one by the SOM, being classified correctly in a separate group by the GTM. It had however also ambiguous classifications with different SOMs. Namely, this protein turned out to be not a member of the septin family, but from the peptidase family by a check in the SWISS-PROT data base (Bairoch and Apweiler (2000)). With the SOM we could find a classification of the analyzed sequences into four main groups. These four groups could be retrieved with the GTM by joining together quite closely related groups on the map (cf. Figure 2).

By comparing the class representatives, which are given in the GTM solution by the projected nodes $\mathbf{y}(\mathbf{x}_k; \mathbf{W})$, $k = 1, \dots, K$, in the data space, it

is possible to detect key regions and residues in the sequences determinant for the groups or the whole protein family.

One problem we faced during the application of the GTM approach on our data is that the algorithm is computationally demanding, especially in terms of memory usage. We used for our calculations the GTM toolbox by Markus Svensén, written in MATLAB and freely available from the web site <http://www.ncrg.aston.ac.uk/GTM/>. One reason is that the GTM algorithm operates only in batch mode and our data vectors were of very high dimension. Another reason is that MATLAB itself is quite demanding in terms of memory and processing time. The SOM approach in contrast was applied in sequential mode on our data and didn't show large need in computer resources.

5 Modified GTM approach for sequence data

For our sequence data we implied with the GTM approach a Gaussian distribution in the data space, given a node \mathbf{x}_k from latent space and parameters \mathbf{W} and β . This assumption is not justified for the binary coded sequence data. We should choose instead a discrete distribution being more appropriate for sequence data.

Svensén (1998) recommends for the modelling of a discrete variable, indicating the membership of mutually exclusive classes, the use of the multinomial distribution with the soft-max function as mapping function. This can be transferred to the modelling of the distribution of an alignment position in an amino acid sequence. If \mathbf{t} is a binary vector $\in \mathbb{R}^{20}$ describing the alignment position with a "1" in the corresponding component of the considered amino acid, the probability for \mathbf{t} is

$$p(\mathbf{t}|\mathbf{x}, \mathbf{W}) = \prod_{a=1}^{20} [y_a(\mathbf{x}, \mathbf{w}_a)]^{t_a}, \quad (8)$$

given latent point \mathbf{x} and parameter matrix \mathbf{W} . The expectation values $y_a(\mathbf{x}, \mathbf{w}_a)$ for amino acid a , $a = 1, \dots, 20$, are determined with the soft-max function

$$y_a(\mathbf{x}, \mathbf{w}_a) = \frac{\exp [\phi^T(\mathbf{x})\mathbf{w}_a]}{\sum_{a'=1}^{20} \exp [\phi^T(\mathbf{x})\mathbf{w}_{a'}]} \quad (9)$$

as mapping function in dependence of latent point \mathbf{x} , where $\phi(\mathbf{x})$ is, as in the classical GTM model, an M -dimensional vector of basis functions and the \mathbf{w}_a are M -dimensional parameter vectors forming together the parameter matrix \mathbf{W} . The soft-max function ensures that the values for $y_a(\mathbf{x}, \mathbf{w}_a)$ lie between 0 and 1.

This approach can be extended to model the distribution of whole sequences of length L , if we assume independent alignment positions. If \mathbf{t} des-

ignates a whole sequence vector in binary form of dimension $20L$, its distribution is obtained by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{W}) = \prod_{l=1}^L \prod_{a=1}^{20} [y_{al}(\mathbf{x}, \mathbf{w}_{al})]^{t_{al}}, \quad (10)$$

given latent point \mathbf{x} and parameter matrix \mathbf{W} . The expectation values $y_{al}(\mathbf{x}, \mathbf{w}_{al})$ for amino acid a , $a = 1, \dots, 20$, in alignment position l , $l = 1, \dots, L$, are determined analogously with the soft-max function

$$y_{al}(\mathbf{x}, \mathbf{w}_{al}) = \frac{\exp [\phi^T(\mathbf{x})\mathbf{w}_{al}]}{\sum_{a'=1}^{20} \exp [\phi^T(\mathbf{x})\mathbf{w}_{a'l}]}, \quad (11)$$

\mathbf{w}_{al} being M -dimensional parameter vectors. In analogy to the classical GTM approach, we assume a sum of delta functions (cf. formula (3)) for the distribution in the latent space to obtain a regular grid with nodes \mathbf{x}_k , $k = 1, \dots, K$. This leads to different probabilities $y_{al}(\mathbf{x}_k, \mathbf{w}_{al})$ for the occurrence of an amino acid a in a given position l for each node \mathbf{x}_k . These probabilities can be used for the determination of a sequence profile for each node.

The components of the parameter vectors \mathbf{w}_{al} have to be estimated for an optimal mapping – altogether $20LM$ parameters. This is the same number of parameters to be estimated as in the classical case, apart from the inverse variance parameter β . To determine the parameters, the log likelihood

$$\mathcal{L}(\mathbf{W}) = \sum_{n=1}^N \ln \left\{ \frac{1}{K} \sum_{k=1}^K \prod_{l=1}^L \prod_{a=1}^{20} [y_{al}(\mathbf{x}_k, \mathbf{w}_{al})]^{t_{al}^{(n)}} \right\} \quad (12)$$

of the given data has to be maximized with respect to the parameter vectors \mathbf{w}_{al} summarized in matrix \mathbf{W} . However, there is no analytic solution with the EM algorithm in this case. One has to use for the maximization of $\mathcal{L}(\mathbf{W})$ numerical optimization technics like quasi-Newton methods or gradient ascent with a generalized EM approach.

So far, we considered in the multinomial approach only the case, when the alignment doesn't contain gaps. For modelling gaps, an extra letter could be introduced in the alphabet of amino acids, leading to 21 components per position in the binary coding of a sequence.

Instead of modelling each amino acid separately one could summarize related amino acids with similar biochemical and biophysical properties into groups. This would give a reduction in the number of components needed for the coding of each alignment position and hence in the number of parameters to be estimated. Also the phylogenetic difference between mutations to closely related amino acids and to more distant ones would be considered by this coding scheme.

6 Summary and outlook

We presented the application of the GTM approach of Bishop et al. (1998) for the analysis of protein sequence data as alternative to the SOM approach of Kohonen (1982), we tried in an earlier work (Grimmenstein et al. (2002)). The aim was to find a reasonable classification of protein families with class representatives for the detection of key regions and residues in the proteins. We applied the GTM approach to proteins of the septin family and compared our derived classifications with former results derived by SOMs. The results showed much similarities, but differed also in some aspects.

Since the implied Gaussian distribution with the GTM is not justified for sequence data, we proposed as alternative to use a distribution based on the multinomial. The implementation of the multinomial approach as well as the implementation of a sequential form of the GTM algorithm for a reduction of required computer resources in terms of memory and processing time will be subject of future research.

Acknowledgements: We would like to thank Dr. Dirk Husmeier from the Biomathematics and Statistics Scotland (BioSS) in Edinburgh for helpful discussions and comments as well as Dr. Ingrid Vetter and Professor Alfred Wittinghofer from the Max-Planck-Institute for Molecular Physiology in Dortmund for their advice in biological questions.

References

- ANDRADE, M.A., CASARI, G., SANDER, C., and VALENCIA, A. (1997): Classification of Protein Families and Detection of the Determinant Residues with an Improved Self-Organizing Map. *Biological Cybernetics*, *76*, 441–450.
- BAIROCH, A. and APWEILER, R. (2000): The SWISS-PROT Protein Sequence Database and its Supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*, 45–48.
- 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.
- 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*.
- KOHONEN, T. (1982): Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.
- SVENSÉN, M. (1998): *GTM: The Generative Topographic Mapping*. PhD Thesis, Aston University.

How Can Data from German Cancer Registries Be Used for Research Purposes?

Alexander Katalinic

Institut für Krebsepidemiologie e.V.,
Universität zu Lübeck, Beckergrube 43-47, D-23552 Lübeck,
Email: alexander.katalinic@krebsregister-sh.de

Abstract. In the past data from German population based cancer registries have been rarely used for research purposes. Based on examples of ongoing research of the cancer registry Schleswig-Holstein different access forms to the registry data are shown. Briefly three access ways could be identified: 1. use of anonymous epidemiological data, 2. use of person identifying data and 3. matching of independently obtained study data with cancer registry data. In summary it could be shown that research on the basis of cancer registry data is possible and can be effective. Therefore cancer registry data should be used more intensively than in the past.

1 Background

The main tasks of German cancer registries are to observe the incidence of all kinds of cancer and their trends, to analyse the collected data under statistical-epidemiological view and to provide data for health care planning and research purposes, including the evaluation of preventive, therapeutic and after-care procedures. These tasks are fixed in regional laws of the sixteen German federal counties (Bundesländer). In the past cancer registry data were rarely used for research purposes compared to other epidemiological cancer research. This is mainly based on the circumstance that there is only a geographically partial registration of cancer in Germany until now and that many of the existing cancer registries could not yet provide complete data for their region. Due to a state law all federal counties founded cancer registries in Germany until the year 2000, although not all regions are covered in total area yet (Schüz et al. (2002)). But for the future a complete and full registration of all cancer cases in Germany can be assumed. Then nationwide data for incidence and mortality provided by cancer registries will be available. Until this time only a couple of registries can offer data for cancer research. But how can the data, collected by German cancer registries, be used for research purposes? The research activities of the cancer registry of Schleswig-Holstein will serve as example for different accesses to cancer registry data.

2 Methods

The Cancer Registry covers the entire area of Schleswig-Holstein (SH, the northern Bundesland of the Federal Republic of Germany, on the border to

Denmark) with a population of almost 2.8 million inhabitants (Statistic Federal Office of SH (1998-2000)). Since 1998 all persons with a diagnosis of malignant cancer disease are statistically recorded area-wide in a central register. The legal instrument hereto is represented by law of the Bundesland, which implies the compulsory registration/obligation for all doctors to register without the possibility to dissent for the persons affected. On the whole, 3,300 recording institutes (general practitioners, hospitals, pathologists, public health departments) participate in the cancer registry. As a special feature, patients are asked together with the notification of their cancer case to the registry, whether they were willing to take part in future research projects. Until now (March 2003) there are about 80,000 persons with 85,000 cancer cases stored in the registry. Each year about 14,000 to 15,000 new cancer cases are expected for the Region of Schleswig-Holstein. For the year of diagnosis 2000 about 14.500 new cancer cases were collected by the registry (Katalinic et al. (2002)). The estimated completeness of registration was about 90 percent for all cancer case sites, whereas some localisations can be assumed to be completely registered. About 75 percent of the patients are willing to take part in research projects, if they had been asked by their doctors.

3 Results

Three possible ways to access cancer registry data for research projects were identified:

1. use of anonymous epidemiological data
2. use of person identifying data
3. matching of independently obtained data with cancer registry data

3.1 Use of anonymous data

The cancer registry can provide anonymous epidemiological data for selected questions both on the basis of individual data and on the base of aggregated data. Data can include variables according to figure 1. The only limitation is that single data sets (or aggregates thereof) will not deanonymize a person. In Schleswig-Holstein this kind of data is used for the routine cancer reporting, e.g. yearly reports concerning incidence (total numbers, crude and age-adjusted rates for both sexes), tumour stage (TNM-Staging), regional distribution(on the basis of districts) of incidence and tumour stage and the descriptive comparison of national and international data. The annual reports (Cancer in Schleswig-Holstein, Vol. 1 and 2 (Katalinic et al. (2001, 2002))) are available under www.krebsregister-sh.de. Further the anonymous data sets are used for explorative, hypothesis-generating analysis. An example for this is the regional analysis of tumour stages of breast cancer, which was done by the Cancer registry Schleswig-Holstein itself. Local differences in tumour size (T of TNM), which already were observed for single years in

Epidemiological data set of the cancer registry Schleswig-Holstein	
Sex, sibling	Side indication in case of paired organs
Date of birth, place of birth	TNM-stage
Postal code and municipal index number	Former tumour affections.
Tumour diagnosis acc. to ICD-10	Certainty of diagnosis
Histology and localization acc. to ICD-O-2	Therapies carried out
Month and year of tumour diagnosis	Month and year of death
	Cause of death
	Autopsy carried out

Fig. 1. typical data set of a cancer registry

the annual reports, could be shown on the basis of about 5,000 breast cancer cases for a longer time period. These findings led to the question whether regional differences in health care are responsible for earlier or later detection of breast cancer. To investigate this question more clearly, it was planned to identify patients with breast cancer from the cancer registry database and to carry out a survey. This kind of data access leads to the next possibility of usage of cancer registry data.

3.2 Use of person identifying data

In Germany the usage of person identifying cancer registry data is regulated by the different cancer registry laws. In Schleswig-Holstein the access to these data is possible on proposal, including study protocol, vote of an ethical committee and data protection statement. Patients, who gave their consent for future research at notification and fulfil defined inclusion criteria (e. g. kind and year of diagnosis), can be identified from the cancer registry database. This population can be contacted directly by the study group for further investigation. Starting from the hypothesis of regional differences in breast cancer care, a study design, based on cancer registry data, was developed to analyze the process of cancer care from diagnosis over therapy to aftercare. Patients with diagnosis of breast, prostate cancer or malignant melanoma will be identified from the cancer registry data base, be included in the study and asked by means of a questionnaire (figure 2). Within two years about 7,500 patients (3,800 with breast cancer, 2,400 with prostate cancer and 1,300 with malignant melanoma) would at most be eligible from the pool of cancer patients, but only about 50 percent of these are expected to be accessible for the questionnaire. This portion is smaller than expected (75 percent), because not all patients are invariably informed by their doctors about the possibility of participating in research projects. The study is supported by Deutsche Krebshilfe e.V. and was started under the acronym 'OVIS-Studie' in October 2002. The access to cancer patients by means of the cancer registry has several advantages. At first, patients come from a population-based pool, which is well described through the epidemiological variables according to table 1. Even when the database is split into responders and non-responders (at the time of notification or of the study itself, see figure 1) a control of response

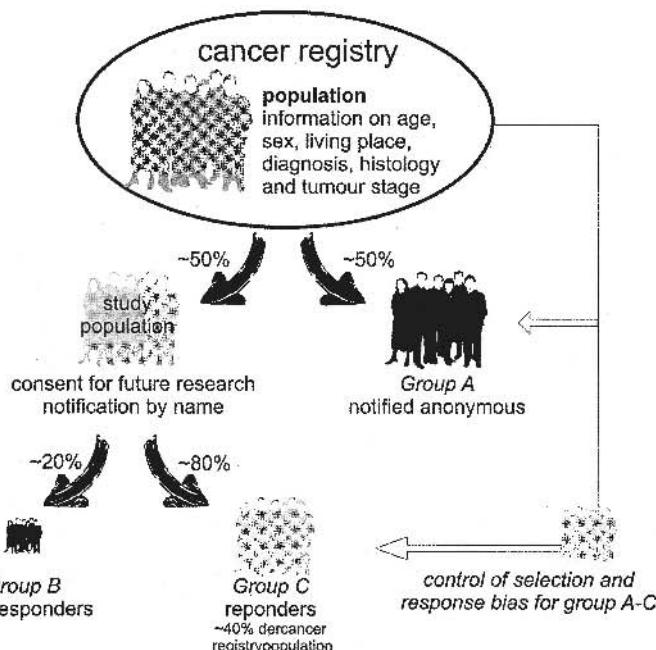


Fig. 2. Identification of patients from the cancer registry database and control of selection and response through epidemiological variables

and selection is possible. In most typical epidemiological studies the analysis of non-response hardly can be done, because no or only minimal data of the non-responders is available. A disadvantage of the cancer registry access to cancer patients is the rather long time delay until patients can be reached. The minimal time for patient access, induced through a delayed transmission of the notification to the registry, lies in the range of six to twelve months. On the other hand, in a clinical study comparing different treatments, patients are usually enrolled at the time of diagnosis or prior to surgery.

3.3 Matching of independently obtained study data with cancer registry data

A third possibility of usage of cancer registry data is a matching of own study data with the registry data. A common endpoint for cohort studies is the development of cancer or death of cancer. In this situation one has to take into account that it sometimes takes many years until the cohort members experience cancer events. The difficulty of long term follow up, with frequent loss of cohort members is well known. In this situation the matching of the separate cohort with an external independent data source could be helpful. Identification of cohort members as cancer patients, stored in the registry, could help to reduce loss to follow-up. Another application of the matching process with

cancer registry data is the evaluation of cancer mass screening programs and quality assurance projects. The example of evaluation of diagnostic procedures (e.g. mammography screening or quality assured mammography) can show the importance of the matching with a cancer registry. On one side it is important to compare the new tumours found within the project with the cancer registry data. Are tumours smaller under the intervention than tumours found in an earlier period? Are they smaller than tumours found in regions that do not have such a program? Will cancer survival be prolonged for patients diagnosed under the more intensive diagnostic process? These and other questions can be answered relatively easily with a direct analysis of registry data in time and space. On the other hand the tumours not found by means of the intervention are even more important. How many breast cancer cases were misdiagnosed or overlooked? How many cancer cases became clinically obvious between two planned diagnostic procedures? To answer these questions an individual matching of the study cohort and patients stored in the cancer registry has to be carried out. Technically this can be done with adequate data safety on a pseudonymous basis with stochastical record linkage methods. Through the matching process the cancer data supplied from the registry can be assigned to the 'hits' of the matching process. Continued matching rounds, e.g. yearly or two-yearly, can easily be performed and improve the validity of the study results. The process of matching is not standardized in the German Bundesländer. In some of them matched data can be given back to the researcher; in others the analysis can only be made within the registry; in others the legal conditions are unclear yet.

4 Discussion and conclusion

The examples given above show that research on the basis of cancer registry data is possible and can be effective. A broad range of different access ways, depending on the goal of the research question, is available. Many questions can be answered by using the anonymized cancer registry data. Others require direct access directly to the person identifying data with following contact to the patient. A persisting problem arises if cancer research with registry data should be performed across the borders of the German Bundesländer, involving more than one of them. In fact the possibility for research on the basis of registry data exists in almost all of the Bundesländer, but the different laws of individual data protection will make it hard to find a common easy and comparable way for studies across Bundesländer borders. The logistic expenditure and so the costs will be high. These circumstances should motivate a harmonizing debate of the current cancer registry landscape, not only in research, but also in registration itself. Meanwhile cancer registry data should be used more intensively than in the past. Only if the need of cancer registry data is visible, better, easier and expanded access to cancer registry data can be obtained.

References

- ARBEITSGEMEINSCHAFT BEVÖLKERUNGSBEZOGENER KREBSREGISTER IN DEUTSCHLAND (2002): *Krebs in Deutschland*. 3. erweiterte, aktualisierte Ausgabe, Saarbrücken.
- KATALINIC, A., HOLZMANN, M., BARTEL, C., GREULICH, K., PRITZKULIET, R., and RASPE, H. (2002): *Krebs in Schleswig-Holstein – Inzidenz und Mortalität im Jahr 2000* Band 2, Schmidt-Römhildt Verlag, Lübeck.
- KATALINIC, A., HOLZMANN, M., BARTEL, C., and RASPE, H. (2001): *Krebs in Schleswig-Holstein - Inzidenz und Mortalität im Jahr 1999* - Band 1, Schmidt-Römhildt Verlag, Lübeck.
- SCHÜZ, J., SCHÖN, D., BATZLER, W., BAUMGARDT-ELMS, C., EISINGER, B., LEHNERT, M., and STEGMAIER, C. (2000): Cancer registration in Germany: current status, perspectives and trends in cancer incidence 1973-93. *J.Epidemiol.Biostat.*, 5, 99–107.

Probabilistic Record Linkage of Anonymous Cancer Registry Records

Martin Meyer, Martin Radespiel-Tröger, and Christine Vogel

Population Based Cancer Registry Bavaria,
Friedrich-Alexander-Universität Erlangen-Nürnberg, D-91052 Erlangen, Germany

Abstract. In a cancer registry usually different cancer notifications for the same patient will arrive in the course of time. It is important to link all notifications to the same person even if some differences of identifying features are present, otherwise it would not be possible to get correct numbers of tumors and to calculate valid incidence measures. The registration office of the Bavarian cancer registry implemented a semiautomatic record linkage module. This program asks the user for an interactive user decision only when a borderline case has been detected. In most cases it is possible to perform an automatic link of records or at least to create a suitable preselection.

1 Different sources of cancer records

For the purpose of population based registration a cancer registry has to collect data from different medical institutions. An important source are the pathologists who may provide very detailed morphological information about a tumor itself, but only few items about the patient. Hospitals and medical centers could supply data about therapy, on the other hand the family doctor perhaps knows best the first date of diagnosis. Sometimes a second manifestation of a tumor will be recorded. If a person has died then the public health authorities inform the registry about date and cause of death. These very different types of cancer notifications usually do not arrive in a logical order, even records about a second manifestation of a tumor could reach the cancer registry before the primary tumor record. Often notifications for the same patient can be linked in an early state of data collection (e.g. a clinical tumor centre links pathological data to the corresponding clinical records), but in general the cancer registry must be able to link any type of incoming cancer record to the current content of the registry's database.

2 Anonymous data records

In Bavaria the cancer notifications are not sent directly to the registration office to be stored in the database. To provide a high level of data protection a confidential office collects the notifications and encodes all identity variables. Only these anonymous variables - together with the epidemiological information and tumor data - are forwarded to the registration office. The original

identity data will be erased after some months. The anonymous variables allow to separate different persons, but it is not possible to decipher any original names. As a consequence the Bavarian registry cannot link incoming records via patient name and address, because only anonymous identity variables are reaching the registry's database to follow the rules of data protection written in the Bavarian cancer registry law. It is an interesting fact, that anonymization does not influence the process of probabilistic record linkage, but there are some special conditions for the collection of reference probabilities for anonymous data.

3 Tasks of record linkage

The most important task of record linkage is to link the different notifications arriving at different times to the same person, otherwise it would not be possible to get correct numbers of tumors and to calculate valid incidence measures. This process should be automated as much as possible, because a high number of notifications must be handled every year.

As a consequence of input errors or typing mistakes sometimes the identifying variables may differ between different records for the same person. Distinct identity data could also be a consequence of removal or change of family name after marriage. Nevertheless the linkage process should bring together such records.

Two types of linkage errors must be considered:

- Homonym errors: Records are linked, although they belong to different persons. A reason for this error could be a random correspondence of identity variables, e.g. two persons with a widespread family name living in the same big city.
- Synonym errors: Records are not linked, although they belong to the same person. This error could appear if too much typing mistakes or other changes happened for two notifications of a certain person.

4 Linkage algorithm

Source of the linkage algorithm is the method described by Fellegi and Sunter (1969). This algorithm is used by many commercial software products and individual implementations of record linkage solutions.

In general there are two sets of notifications (A, B) to be linked. Then it is the task of the linkage algorithm to select two disjoint subsets (M, U) out of the set of all possible pairs ($A \times B$). Subset M ("matched") should contain the pairs of records belonging to identical persons. Subset U ("unmatched") should contain all other pairs of data records.

$$A \times B = \{(a, b); a \in A, b \in B\} \quad (1)$$

$$M = \{(a, b); s(a) = s(b), a \in A, b \in B\} \quad (2)$$

$$U = \{(a, b); s(a) \neq s(b), a \in A, b \in B\} \quad (3)$$

where a and b are representing cancer notifications and $s(x)$ is the function linking a notification x to a specific person.

For the implementation into the database application of the Bavarian cancer registry the definitions of the Subsets M and U were simplified to:

$$\begin{aligned} A &= \{ \text{single new data record} \} \\ B &= \{ \text{all existing data records} \} \end{aligned}$$

The linkage process then is repeated for each new data record.

The data records are handled as feature vectors, containing all important variables for identification of a person, e.g. last name, first name, date of birth, place of residence.

$$a = (a_1, \dots, a_n)$$

$$b = (b_1, \dots, b_n)$$

For each element of the feature vector two conditional probabilities have to be determined. These probabilities consider the value distribution of the match variables und their probability of changes in time:

- The probability "m" explains the probability that two records a and b match exactly in the i^{th} element of the feature vector, if both records belong to the same person:

$$m_{ik} = P(a_i = b_i \wedge a_i = x_{ik} | (a, b) \in M) \quad (4)$$

The probabilities m_{ik} can be determined from a validated set of person records. Theoretically these probabilities should be provided for all x_{ik} , i.e. for each possible expression of the i^{th} vector element. In practice it is sufficient to assume $m_{ik} = m_i$ for all values of k and to use an estimation like $m_i = 1 - P(\text{input errors or typing errors or other changes})$.

- The probability "u" explains the probability that two records a and b match randomly in the i^{th} element of the feature vector, although the records belong to different persons:

$$u_{ik} = P(a_i = b_i \wedge a_i = x_{ik} | (a, b) \in U) \quad (5)$$

Since the size of the set M is very small compared with the size of U, it is allowed to make again some simplification for the calculation of u_{ik} . Then the probabilities u_{ik} can be determined by examination of the population's frequency table for the i^{th} vector element. Anonymization must be recognized before the creation of these frequency tables.

The linkage weight for a pair of records is calculated by adding the logarithmic ratio of m- and u-probabilities for all items of the feature vector:

$$w = \sum_i w_i \quad (6)$$

with

$$w_i = \log \left(\frac{m_i}{u_{ik}} \right), \text{ if } a_i = b_i \wedge a_i = x_{ik}$$

$$w_i = \log \left(\frac{1-m_i}{1-u_{ik}} \right), \text{ if } a_i \neq b_i \wedge a_i = x_{ik}$$

If an item matches in both records, then the weight summand w_i will have a positive sign. A negative sign of w_i appears when the values for a feature vector item are different in both records.

To our knowledge the incorporation of information on the association structure between different feature vector items into the process of record linkage has not been applied so far. This would be an interesting topic for future research, but the collection of empirical data to create the reference tables for the probabilities m and u would be much more difficult.

5 Application of record linkage

As noted in the paper of Fellegi and Sunter (1969) it is neither practicable nor necessary to compare each data record with each other record, because the number of comparisons would be too large for a cancer data pool. It is sufficient to examine certain blocks, containing records with a similarity of some typical items, e.g. phonetic names and/or date of birth. Only members of the same blocks are handled as linkage candidates. The linkage weights are calculated for these candidate records. An example of the linkage weight distribution after linking a data pool of about 60,000 records is displayed in Figure 1. The histogram gives the impression of two overlapping distributions which can be easily interpreted as the distributions for the sets M and U (see formula 2 and 3). A minimization of linkage errors is achieved by setting an upper and a lower threshold value (Figure 2). Linkage weights below the lower threshold are classified as positive non links, weights above the upper threshold as positive links. Weights in the range between both limits need an interactive decision by the user, who compares not only the anonymous identification variables but also the tumor data to judge the relation between the linkage candidates. If necessary, additional information about critical cases will be queried from the original senders of tumor notifications. When a new notification arrives and the new information indicates, that a former linkage result was made in error, then it is possible to reassign older notifications to other persons.

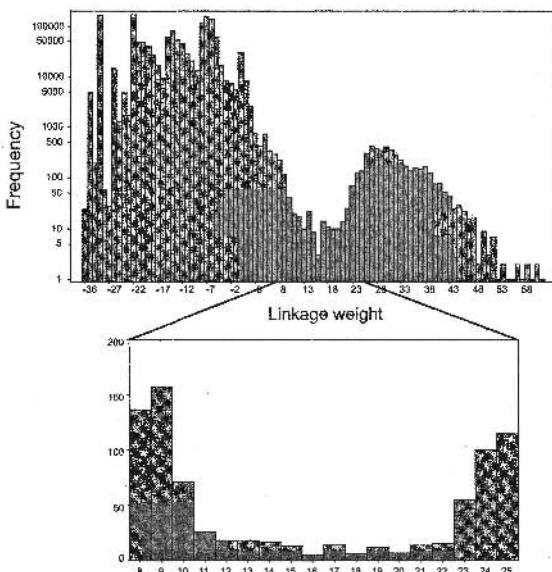


Fig. 1. Distribution of linkage weights for all candidate pairs of records

6 Validation

Anonymization limits the available validation methods. It is not possible to compare directly the results of record linkage with the original data stored at the senders of cancer notification. Nevertheless the success of record linkage can be proved: Many homonym errors will be detected and erased when the tumor data of the records intended to link do not fit, when considering a plausible patient's tumor history. Synonym errors were examined with a sample of records which were sent to the registry twice: a first version in an early state of documentation, a second version after an intensive data revision in the sender's database. The second version for example included naming correction after review of official death certificates and other data proof processing. The rate of records from the first package, which could not be linked to a record of the second version, was only 0.3 %.

7 Integration of record linkage into an automated data processing system

The registration office of the Bavarian cancer registry implemented a semiautomatic record linkage module. This program asks the user for an interactive user decision only when a borderline case has been detected. In most cases it is possible to perform an automatic linkage of records or at least to create a suitable preselection. The stochastic process assesses the level of connection

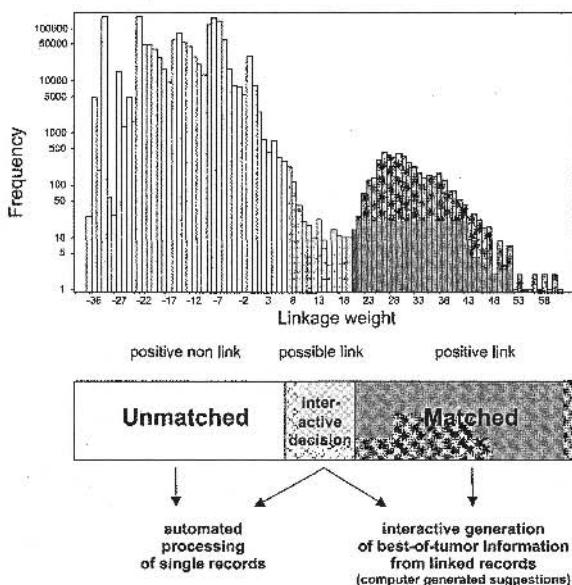


Fig. 2. Semiautomatic linkage process

between all relevant pairs of data records with a linkage weight. If only low linkage weights are found for a new data record, then this record is handled as a single record and can be stored and processed automatically (Figure 2).

If high linkage weights are calculated for the new record and some existing records, then these records are linked together and the user has to extract the best tumor information out of the selected data records. Only the borderline cases of the mid area of linkage weights force a decision of the user about linking. The record linkage module is embedded into a database application, where many steps were automated to achieve a high throughput of data records and to guarantee valid and reproducible results of all processing steps. Fully automated steps are the import of incoming cancer notifications, formal validation, simple validation and cross-field validation. Computer generated error reports can be distributed to the senders of records. The record linkage process as well as the extraction of the best-of-tumor information from linked records are semiautomated. When storing the result of the best-of-tumor task, the fully automated validation procedures are executed once again.

References

- FELLEGI, I.P. and SUNTER, A.B. (1969): A theory for record linkage. *American Statistical Association Journal*, 40, 1183-1220.
 SCHMIDTMANN, I. and MICHAELIS, J. (1994): Untersuchungen zum Record Linkage für das Krebsregister Rheinland-Pfalz. Technical report, Tumorzentrum Rheinland-Pfalz, Mainz.

An Empirical Study Evaluating the Organization and Costs of Hospital Management

Karin Wolf-Ostermann¹, Markus Lüngen², Helmut Mieth², and Karl W. Lauterbach²

¹ Department of Child and Adolescent Psychiatry, Philipps-University Marburg, D-35033 Marburg, Germany

² Institute for Health Economics and Clinical Epidemiology of the University of Cologne, D-50935 Köln, Germany

Abstract. In Germany so far, there exists no evaluation of the relationship between the organization and costs of hospital administrations and hospital characteristics. In a survey of hospital administration costs, structure, and salary level 126 hospitals participated for the years 1998 and 1999. Hospitals of medium size and non-profit ownership show the smallest expenditures for personnel in the administration per treated case. However, salary level was not uniformly linked to hospital size. Hospital ownership appeared to be a strong indicator for the level of personnel salaries. For the planned introduction of prospective payment via Diagnosis Related Groups (DRG) starting 2003 in Germany our study has substantial implications. Publicly owned hospitals, in particular, are likely to have their administrations most severely affected by the change.

1 Introduction

The intended adoption of a global reimbursement system for inpatient care starting from 2003 in Germany envisions identical payments for identical treatments at different hospitals. This prospective payment system using DRG (Diagnosis Related Groups) will represent a homogeneous classification of inpatient cases according to medical and economical principles. It will not represent for example nursing aspects or differences in administrative structures of hospitals. The changes of the financing mode will affect all hospitalization sectors. A substantial increase in efficiency is foreseeable. The change will also have substantial effects on the administration of hospitals: For the first time the product of the hospital will be accessible to an economical analysis, since Diagnosis Related Groups represent explicitly an appropriate classification (Fetter et al. (1980)). This definition of the products of a hospital makes it possible to establish a more precisely cost accounting. Thereby it becomes recognizable in which divisions profits or losses are gained. It is uncertain whether all hospital administrations in Germany are methodically and mentally sufficiently prepared in order to meet these requirements. As

the experiences of the foreign countries show, the introduction of a prospective payment system in hospitals led not rarely to a change of the (upper) management (Breßlein (2001)).

So far no detailed data exist concerning administration costs in German hospitals. The Institute of Health Economics and Clinical Epidemiology at the University of Cologne conducted a survey of hospital administration costs, structure and salary level which for the first time gives a comprehensive review of the actual condition of hospital administrations of the German Federal Republic.

1.	<i>Basic data</i> number of wards, beds, cases; days of inpatient stay; actual hospital costs; number of personnel; ownership
2.	<i>Management structure</i> organization of hospital management (size, type and capacity of boards, standing orders)
3.	<i>Organization of administration</i> structure of administrative departments; training of management; gross salary level of management
4.	<i>Type and costs of staff</i> number of positions and costs per full vigor in 1998 and 1999 by administrative department
5.	<i>External support, consultants</i> enlisted external support by administrative department; costs of external support and annual audit
6.	<i>Internal meetings</i> frequency and participants of internal meetings
7.	<i>Position of managing director</i> annual gross salary (including extended benefits); fraction of variable amount; appropriateness of salary; responsibility of managing director; length of employment in current position; temporary employment

Table 1. Structure of the questionnaire "Organization and costs of hospital management"

2 Methods

We outlined a special survey within the scope of the "f&w-Krankenhaus-Kompass" - a recurrent voluntary survey of hospital services and costs¹. An overview about the survey is given in Table 1. The results of the "f&w-Krankenhaus-Kompass" were solely spread to the participants of the survey. Here we give a summary with commentaries especially concerning the costs for personell in the administration.

¹ f&w Krankenhaus-Kompass Spezial 2000.

Actual costs were raised for the years 1998 and 1999. For each part of the questionnaire a detailed evaluation by means of descriptive, graphic and inductive statistic procedures was carried out. This comprised thorough one-dimensional as well as two- and multi-dimensional analyses. Methods of the variance and regression analysis as well as correlation and association calculations are used in order to identify and also to quantify relationships between individual characteristic. All analyses have been done by the use of the statistical analysis system SPSS 11.0.1.

3 Results

Altogether the data of 126 hospitals in Germany are included in our study. From the participating houses 55,5% are in free-non-profit, 40,2% in public and 4,3% in private ownership (Germany: 40,5%, 38,8%, 20,6%). The classification of hospitals according to the number of beds (1-200, 201-300, 301-500, >500 beds) was geared to both classifications used by the Federal Statistical Office and getting an almost evenly divided sample. Hospitals in private ownership have less beds and cases than the average of all hospitals and they are more often specialized on medical fields. In our sample privately owned hospitals are underrepresented so that we avoid statements concerning for smaller special hospitals which would require further data collection. For the interpretation of the results it is important that a significant association of bed number and ownership shows up (Cramér V=0,25, p=0,025). Hospitals in public ownership have a tendency to an above average number of beds (Figure 1).

Organization of administration

About three quarters of the participating hospitals (73,8%) set up a separate supervision and/or decision maker between the top management in the hospital and the highest owner committee. Here no significant differences between public and free-non-profit hospitals arise. The most frequent designation of the committee reads supervisory board (19,8%) or board of directors (15,1%). By far most cases this committee has an advisory/controlling function (38,9%), prior to a co-operating/decision making function (24,6%), while pure representation arises relatively rare (11,1%). The highest administrative level consists in the predominant number of the hospitals of the classical forefront with administrative manager, medical director and director of nursing. Altogether no significant characteristics stand out regarding ownership or hospital size. Basically hospitals seem to have identical structures in the highest leadership level.

Salaries of department managers

The department manager "personnel" obtains the highest gross salary with a median of 100.000 DM, followed from the department managers "accounting"

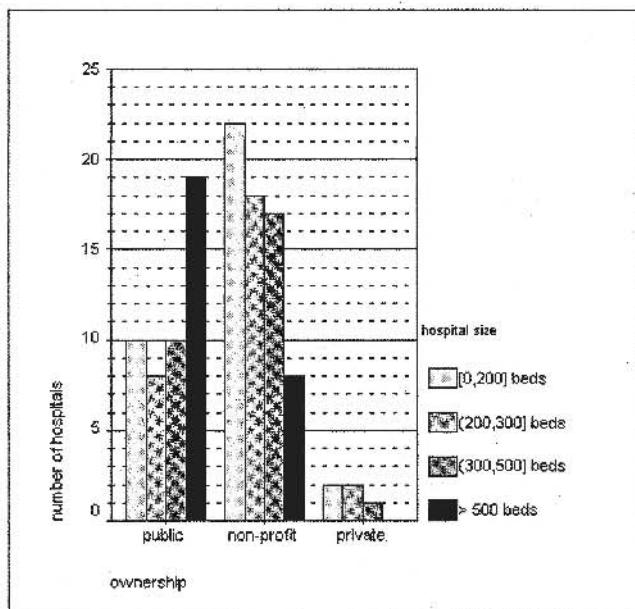


Fig. 1. Ownership of hospitals by number of beds

with 98.700 DM (in each case for the year 1999). The head of the "patient administration" and "stock management" have a median gross salary of 80.000 DM; followed by "organization/EDP" and "controlling" with yearly gross salary of the department managers about 90.000 DM (Figure 2). The yearly gross salary generally rises as a function of hospital size and the skills of the department managers.

Univariate analyses of variance were used to model the influence of hospital size, ownership, number of beds and training of department managers (and their two-way interactions) on the yearly gross salary of individual department managements. In particular for personnel management, accounting, stock management/purchase as well as organization/EDP hospital size and ownership are (significant) factors of influence (c.f. Table 2). Divisions with high staff responsibility, like the personnel management, can obtain the highest salaries.

Salary level and contracts of management

Details of the working contracts show an relation between ownership and yearly gross salary ($Cramér V=0,26$, $p=0,090$). In public hospitals in the year 1999 23,8% of the administrative managers obtained a salary greater than 200.000 DM per year. In non-profit hospitals this share raised to 42,1%. Independently of the ownership the portion of the administrative managers with a yearly salary higher than DM 250.000 DM is about 15%. Likewise an association exists regarding to the hospital size ($Cramér V=0,32$, $p=0,001$).

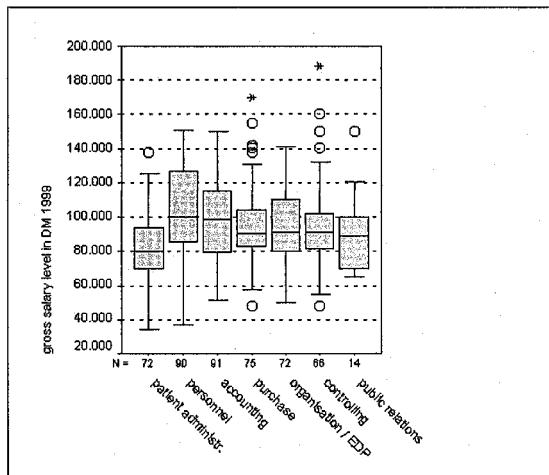


Fig. 2. Boxplot of yearly gross salary for department managers (1999)

Only 8,8% of hospitals with less than 200 beds pay a salary of 200.000 DM and over. This portion rises with hospitals size: 201-300 beds 30,4%; 301-500 beds 59,1%, >500 beds 55,5%. The higher portion of the public hospitals may be causal for the slight decline in the last category. A correlation analysis with consideration of ownership and number of beds (grouped) reveals ownership as an important factor of influence (Cramér V=0,44, p<0,001).

In 1999 30,8% of the administrative managers indicate limited work contracts (free-non-profit hospitals 23,8%, public hospitals - despite the smaller yearly gross salary - 34,9%). 42,7% of the work contracts of the administrative leaders contain a variable salary component. The higher the yearly gross salary of the administrative manager, the higher is the variable portion of the salary. Hospitals with more than 500 beds show an exception due to ownership. For hospitals in public ownership the variable portion of the yearly gross salary is clearly smaller than for other forms of ownership. Based on the entire gross salary of all hospitals the average portion of the variable salary portions is about 6,7%.

More than half (53,3%) the administrative leader indicated that they classify their yearly gross salary as appropriate. Ownership and satisfaction with the yearly salary are associated (Cramér V=0,28, p=0,018). Administrative managers of public hospitals are more dissatisfied with their objectively lower salary than those of non-profit hospitals. In addition the portion of time-limited work contracts is substantially higher for public hospitals. Altogether non-profit hospitals seem to pay the highest salary, and also the highest portion of variable salary. However their work contracts are less frequently limited. The administrative leaders of non-profit hospitals recompense this altogether with a higher satisfaction.

	p-value model *	R ²	Significant influence factors*
patient administration	0,122	0,213	• hospital size (p=0,041)
personnel management	0,006	0,437	• hospital size (p=0,009) • ownership (p=0,037)
accounting	0,086	0,360	
stock management/ purchase	<0,001	0,699	• hospital size (p<0,001) • training (p=0,004)
organization/EDP	0,056	0,657	• hospital size (p=0,009)
controlling	0,702	0,212	
public relations	0,989	0,470	

* p-values are used as explorative measures because of multiple testing

Table 2. Analyses of variance

Costs of hospital administration

The costs of the administrative staff per case is 183,90 DM (median). Personnel costs per case differ with hospital size (1-200 beds 182,10 DM, 201-300 beds 177,40 DM, 301-500 beds 197,70 DM, >500 beds 186,80 DM). Hospitals of medium size show the lowest costs per case for administration personnel. The computation of the Eta coefficient for a description of the combined influence of hospital size and ownership on personnel occupation in full vigours as well as on personnel costs per case leads to the results shown in Table 3.

	Eta-coefficient* for combined influence of hospital size and ownership	
	personnel in full vigours	personnel costs per case
patient administration	0,721	0,458
personnel management	0,639	0,416
accounting	0,618	0,446
stock management/ purchase	0,642	0,484
organization/EDP	0,503	0,216
controlling	0,633	0,486
public relations	0,561	0,566

*(c.f. Ostermann and Wolf-Ostermann, (1999, p. 131f))

Table 3. Eta-coefficient by department

In all departments the total variability in personnel occupation can be explained by more than 50% by the combined influence of hospital size and ownership. The impact of size and ownership varies between the departments (patient administration (72,1%), stock management/purchase (64,2%), personnel management (63,9%), controlling (63,3%) accounting (61,8%)). The influence on the personnel costs per case is altogether smaller, lies however - with exception of the department of organization/EDP - between 40% and 60%. Altogether it can be stated that the hospital size has a high impact

on personnel costs. This is little surprising. However, it is important that ownership leads in all cases to clearly increased values of the Eta coefficient. As result it can be held that ownership has the largest influence on the personnel costs of the administration. The trend is overlaid by the hospital size. Very large and very small hospitals exhibit unfavorable personnel costs in the management.

4 Discussion

The study shows that substantial differences exist between hospitals in reference to the personnel costs of the administration. As main criteria hospital size (in number of beds) and ownership were examined. The results show that hospitals of medium size have the smallest personnel expenditure (of the administration) per case. That approximately U-shaped distribution corresponds to results of the German Federal Office of Statistics and international results regarding the total costs per case. Sowden et al. (1997) summarized the literature concerning the connection between total costs and the size of a hospital and stated that hospitals with 300-600 beds exhibit the smallest costs per case. As a reason it was pointed out that smaller hospitals have higher fix costs despite very large hospitals suffer from inefficiencies, which could not be quantified in any study (Rivers and Bae (2000)).

The privately owned hospitals are underrepresented in our survey. The reason might be that these houses are mostly very small in Germany and often part of a holding or larger group. These holdings often have internal benchmarking projects so that the external benchmarking is not appropriate. We would not expect a radical change in our findings if further privately owned hospitals would have been included.

Our study has several implications for health policy and the management of hospitals. First, we showed that ownership and hospital size have a crucial influence on the personnel costs of the administration. Both characteristics are however hardly to be affected by the hospital. It is not discussed by the German government to give additional payments to large hospitals or publicly owned hospitals. Unless these hospitals could not improve their efficiency, they will see losses. In the long run large hospitals, and as well publicly owned hospitals, are necessary for a high quality and diverse provision of health care services. The further situation of large and publicly owned hospitals in the competition must be supervised.

Second, although the costs analyzed here reach only 6% of the total costs of a hospital in Germany, the overall effect should be estimated much larger. There are several hints that the wage agreements in publicly owned hospitals are much less oriented at efficiency than in privately owned hospitals. Personnel cost account for about 70% of the overall hospital costs. Publicly owned hospitals were often a part of a local system of social services, run by clerks with an deviant scope of goals. Further the wage agreements are based

on the Bundesangestelltentarif (BAT) which is less flexible than many other tarif systems. The reworking of the BAT is part of an intense discussion. We expect that in the next years especially publicly owned hospitals will face a major relaunch of tarif systems.

Third, in the short run the management could reduce the number of employees in the administration or try to restructure the mix of skilled and unskilled employees to reach lower costs. Both alternatives are not very helpful in the current situation because the challenges concerning the new prospective payment system, the strengthening of the competition and the new possibilities of treating outpatient patients starting 2004 demand for more and more skilled administration, not for less. One result of our study seems to be, that less people in the administration with high skills are more efficient than many employees with rather low skills. We think that this will affect the personell management of the future.

Acknowledgement: The authors thank Mrs. Meurer for her permission to use the data of the hospital-benchmark-project (f&w Krankenhaus-Kompaß), Melsungen, Germany.

References

- BREßLEIN, S. (2001): Die Krankenhäuser überleben, aber die Geschäftsführungen nicht. *f&w - führen und wirtschaften im Krankenhaus*, 3, 218–220.
- FETTER, R.B., SHIN, Y., FREEMAN, J.L., AVERILL, R.F., and THOMPSON J.D. (1980): Case mix definition by diagnosis-related groups. *Med Care*, 18(2 Suppl): iii, 1–53.
- F&W – FÜHREN UND WIRTSCHAFTEN IM KRANKENHAUS – KRANKENHAUS-KOMPASS SPEZIAL 2000 (2001): *Organisation und Kosten der Krankenhausverwaltung. Auswertung des Betriebsvergleichs f&w-Krankenhaus-Kompass Spezial 2000*. Hrsg. f&w - führen und wirtschaften im Krankenhaus. Bibliomed, Melsungen.
- LAUTERBACH, K.W. and LÜNGEN, M. (2001): *DRG-Fallpauschalen: eine Einführung*. 2. Ed. Schattauer-Verlag, Stuttgart.
- LÜNGEN, M., WOLF-OSTERMANN, K., and LAUTERBACH, K.W. (2001): *Krankenhausvergleich*. Schattauer-Verlag, Stuttgart.
- OSTERMANN, R. and WOLF-OSTERMANN, K. (1999): *Statistik*. 2. Ed., Oldenbourg Verlag, München.
- RIVERS, P.A. and BAE, S. (2000): The relationship between hospital characteristics and the costs of hospital care. *Health Services Management Research*, 13, 256–63.
- SOWDEN, A., ALETRAS, V., PLACE, M., RICE, N., EASTWOOD, A., GRILLI, R., FERGUSON, B., POSNETT, J., and SHELDON, T. (1997): Volume of clinical activity in hospitals and healthcare outcomes, costs, and patient access. *Quality in Health Care*, 6, 109–114.
- WOLF-OSTERMANN, K., LÜNGEN, M., MIETH, H., and LAUTERBACH, K.W. (2002): Eine empirische Studie zu Organisation und Kosten der Verwaltung im Krankenhaus. *Zeitschrift für Betriebswirtschaft*, 72(10), 1065–1084.

Index

- 2-3 AHC, 3
2-3 Hierarchy, 3
 α -Stable Distribution, 515
Adaptive Distance, 173
Additive Model, 146
Aggregation Method, 207
Analytic Hierarchy Process, 435
Asymmetric Distribution, 181
Asymmetry, 371
- Bąk, A.*, 305
Baier, D., 346
Balanced Incomplete Block Design, 111
Banking, 435, 498, 506, 515
Bartel, H.-G., 46
Bartoszewicz, B., 103
Bayesian Analysis, 19
Bernoulli Trial with Dependence, 154
Bertrand, P., 3
Bessler, W., 419
Bootstrap, 267
Borda, M., 427
Breakdown Point, 128
Brüggemann, R., 69
Burkhard, C., 567
Buy-and-Hold Abnormal Return, 454
- Cancer Record, 599
Cancer Registration, 593, 599
Capital Asset Pricing Model, 461
Capital Market, 419, 446, 454, 461, 482, 523
Car Industry, 313
Career-Lifestyle Cluster, 363
Ceranka, B., 111
Change Point, 154
Chechik, G., 224
Chelcea, S., 3
Chemical Balance Weighing Design, 111
Choice Mechanism, 239, 305, 435
Christoffersen Test, 482
Classification, 30, 283, 355, 427, 523, 544, 559
- Classification Tree, 207
Classifying Factor, 103
Clustering 3, 19, 38, 46, 54, 62, 69, 77, 81, 91, 128, 173, 471
Cluster-wise Consensus, 471
Competition Analysis, 313
Complexity, 3
Computational Learning, 191
Conditional Logit Model, 305
Consistency, 38
Consistency Test, 267
Constraint Context-Free Grammar, 535
Consumer Involvement, 406
Convex Hull, 11
Copula Analysis, 446
Corporate Governance, 454
Correlation Analysis, 199
Credit Risk, 498, 506
Customer Churn Prediction, 330
Customer Relationship Management, 346
- Dahms, S.*, 120
Data Coding, 249
Data Mining, 91, 207, 217, 249
Daub, C.O., 81
De Carvalho, F.A.T., 11, 173
Decision Making, 435
Decision Support, 552
Decision Tree, 199
Decker, R., 313, 355
Decomposition of Variance, 419
Density Estimation, 91
DeSarbo, W.S., 19
Design of Experiments, 111, 305, 567
Diagnosis Related Group, 605
Didelez, V., 259
Dimension Reduction, 224, 259
Discrete Choice Method, 305
Discrimination, 11, 54, 77
Dissimilarity Matching Function, 11
DNA Sequence, 154
Document Classification, 249
D’Oliveira, S.T., 11

- Domański, C.*, 435
 Double Sampling, 30
Duncan, K.H.F., 19
 Dynamic Clustering, 173
 Dynamic Regression, 259
 Education, 567
 Electronic Commerce, 406
 Empty Cells, 138
 Environmental Time Series, 275
 Epidemiology, 593
Esswein, W., 191, 577
 Exchange Algorithm, 62
 Expectation Maximization, 322
 Extreme Value Theory, 446
 Face Recognition, 224
 Family Firm, 454
 Family Relationship, 585
 Financial Ratio, 427
 Finite Mixture, 19
Flöter, A., 199
 Flow Cytometry Data, 69
 fMRI Brain Mapping, 295
 Food Industry, 120, 406
Fried, R., 259
Gamrot, W., 30
Garczarek, U., 283
Garlipp, T., 38
Gatnar, E., 207, 217
Gehlert, A., 191
 Gene Expression, 295
 Generalized Linear Model, 103, 165
 Generative Topographic Mapping, 585
 Genetic Algorithm, 146, 535, 567
Geyer-Schulz, A., 535
Globerson, A., 224
 Goodness of Fit, 62, 138
Grabowski, M., 322
Graczyk, M., 111
 Graphical Model, 259
 Gravity Model, 338
Grimmenstein, I.M., 585
 Group Choice, 239
 Group Opinion, 471
 Group Rationality, 239
 Guttman Scale, 379
 Hazard Ratio, 305
 Health Care Industry, 577, 605
Hennig, Ch., 128
Hermelbracht, A., 313
 Hierarchical Classification, 3, 54
 Hierarchical Value Map, 388
 Ho-Kashyap Algorithm, 30
Hopmann, J., 330
 Hospital Management, 577, 605
 Hyperbolic Distribution, 515
Imaizumi, T., 338, 371
 Individual Difference, 371
 Industry Index, 419
 Information Theory, 81
 Initial Public Offering, 454
 Insurance Industry, 103, 427, 490
 Insurance Rating, 427
 Intelligent Textile, 559
 International Market Segmentation, 388
 Interval-Type Data, 173
Jajuga, K., 446
Jakobza, J., 552
Janker, Ch.G., 544
Jaskiewicz, P., 454
 Joint Space, 338, 371
Katalinic, A., 593
Katona, R., 346
Kempe, A., 295
 Kernel Density Estimation, 38
 Knowledge Discovery, 91
 Knowledge-Based Approach, 552
Konczak, G., 138
Kondrasiuk, J., 435
Kosiorowski, D., 239
Kowalczyk-Lizak, P., 427
Krause, R., 146
Krauth, J., 154
Kuhnt, S., 165
Kuklinski, J., 454
Kurths, J., 81
 Laddering, 379
Lanius, V., 259
Lasch, R., 544
Lauterbach, K.W., 605
Leśkow, J., 267

- Liebscher, V.*, 295
Liechty, J., 19
Ligges, U., 283
 Linear Discriminant Function, 396
 Linear Ranking Method, 523
 Local Search, 506
 Logistic Regression, 120, 330
 Logit Function, 396
 Long-Run Performance, 454
Lowinski, F., 454
Lüngen, M., 605
Luniewska, M., 523
- Macroeconomic Factor, 419
Majewska, A., 461
Majewski, S., 461
 Marketing, 346, 355, 405
 Marketing Research, 305, 313, 322, 330,
 338, 371, 388
 Markov Chain, 239
 Masked Outlier, 165
 Means-End Chains Concept, 379, 388
Meinberg, U., 552
Mendola, D., 275
 M-Estimation, 38
 Meta Modelling, 191
 Metabolic Concentration Data, 199
 Metabolic Pathway, 199
 Method Base, 191
 Method Engineering, 191
Meyer, M., 599
 Microarray, 81
 Middle-Market Portfolio, 506
Mieth, H., 605
 Missing Data, 275, 322
 Model-Based Clustering, 46, 69
Monien, K., 355
 mRNA Expression, 81
Mucha, H.-J., 46, 69
Müller, Ch.H., 38
 Multidimensional Scaling, 338, 371
 Multi-Factor Asset Pricing Model, 419
 Multi-Objective Evolutionary Algo-
 rithm, 506
 Multiple Correspondence Analysis, 363
 Multiple Imputation, 275
 Multivariate Statistical Analysis, 388
 Multivariate Time Series Analysis, 259
 Mutual Information, 81
- Nakai, M.*, 363
 Negative Binomial Regression, 103
 Negative Moment, 181
 Nested Cluster Structure, 471
 Network Inference, 199
 Neural Network, 217
 N-Gram Coding, 249
 Noise Component, 128
 Nonparametric Regression, 295
 Nonparametric Test, 138
 Nonresponse, 30
 Number of Mixture Components, 128
- Okada, A.*, 371
Opfer, H., 419
 Ordinal Regression Approach, 313
 Outlier, 128, 165
 Outlier Identification Rules, 165
Owsiński, J.W., 471
- Parameter Selection, 355
 Pareto/NBD Model, 330
 Pareto's Law, 91
 Partial Moment, 181
 Pattern Recognition, 199, 224, 249
Pawlitschko, J., 165
Pirçon, J.-Y., 54
 Poisson Process, 54
 Poisson Regression, 103
Pojarliev, M., 482
Polasek, W., 482
Poprawska, E., 490
 Population Based Research, 593
 Post-Materialistic Value, 363
 Potts Model, 295
 Preference Data, 19, 305, 338
 Principal Component Analysis, 544
 Production, 535, 544, 552
 Protein Data, 585
 Proximity, 371
 P-spline, 146
- Quality Management, 120, 567
 Quantized Time Series, 267
- Radespiel-Tröger, M.*, 599
 Random Forest, 207
 Rank Ordered Data, 313

- Rasson, J.-P., 54
 Reference Modelling, 191
 Regression Clustering, 38
 Regression Tree, 120
 Resampling, 46
 Retailing, 406
 RFM Model, 330
 Risk Management, 427, 490, 498, 506
Rix, R., 62
 Robustness, 54, 165
Roland, F., 54
Ronka-Chmielowiec, W., 490
 Rough Set, 217
Rozmus, D., 217
Rybicka, A., 305
Sagan, A., 379
 Sales Data Analysis, 355
 Sampling, 30, 46, 77
Schaub, T., 199
Schebesch, K.B., 498
Schiereck, D., 454
Schlottmann, F., 506
Schwaiger, M., 62
Seese, D., 506
Selbig, J., 81, 199
 Self-Organizing Map, 322, 585
Simon, U., 69
 Simulated SAR Image, 11
 Singing Voice, 283
Skibicki, M., 77
 Smoothing Parameter, 146
Sobczak, E., 388
 Social Background, 363
 Social Welfare Function, 535
Sommer, T., 577
Souza, R.M.C.R., 173
 Spectral Density Estimation, 267
 Stability, 523
 Statistical Software, 46, 217
Stecking, R., 498
Steuer, R., 81
 Stochastic Dominance, 181
 Stock Exchange, 419, 446, 454, 461, 482, 523
 Stratification, 54, 77
 Stratified Sample, 77
 Structural Model, 379
 Supplier Management, 544
 Support Vector Machine, 355, 498
 Symbolic Classifier, 11
 Symbolic Data Analysis, 11, 173
Szajt, M., 396
Szczeponiak, W., 515
 Tail Dependence, 446
Tarczyński, W., 523
 Telecommunication Industry, 346
 Ternary Balanced Block Design, 111
 Textile Industry, 559
Thede, A., 330, 535
Theuvsen, L., 406
 Time Series Analysis, 259, 267, 275, 283, 295
 Time Variability, 419
Tishby, N., 224
 Transcription, 283
 Tree Model, 199, 207
Trousse, B., 3
Trzpiot, G., 181
Tutz, G., 146
 Two-Mode Clustering, 19, 62
 Two-Phase Algorithm, 535
 Two-Phase Sampling, 30
Ullsperger, A., 559
Ultsch, A., 91
 Unfolding, 338, 371
Urfer, W., 585
 User Profiling, 3
 Utility Industry, 346
 Validation, 46
 Value at Risk, 482, 506
 Variable Selection, 146, 259
 Visualization, 585
Vogel, Ch., 599
 Volatility Model, 482
Wagner, R., 249
 Web Content Mining, 249
Weihs, C., 283
Winkler, G., 295
Wittich, O., 295
Wolf-Ostermann, K., 605
Woll, R., 567
 Workshop Scheduling Problem, 535
Wronka, C., 267