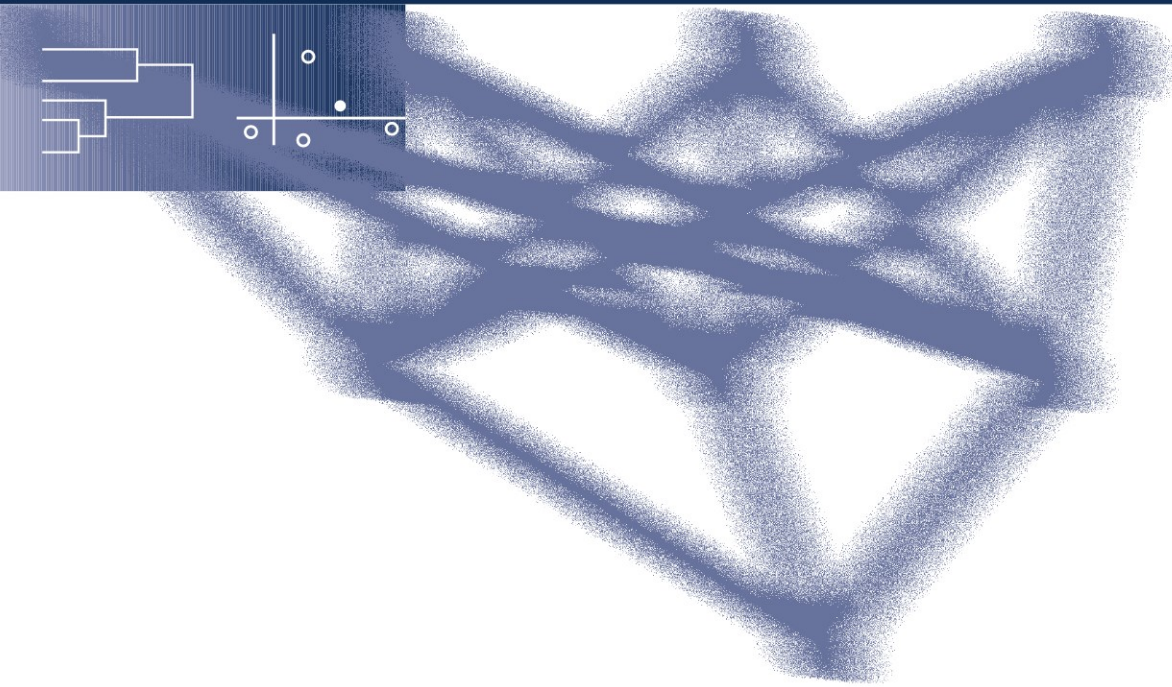
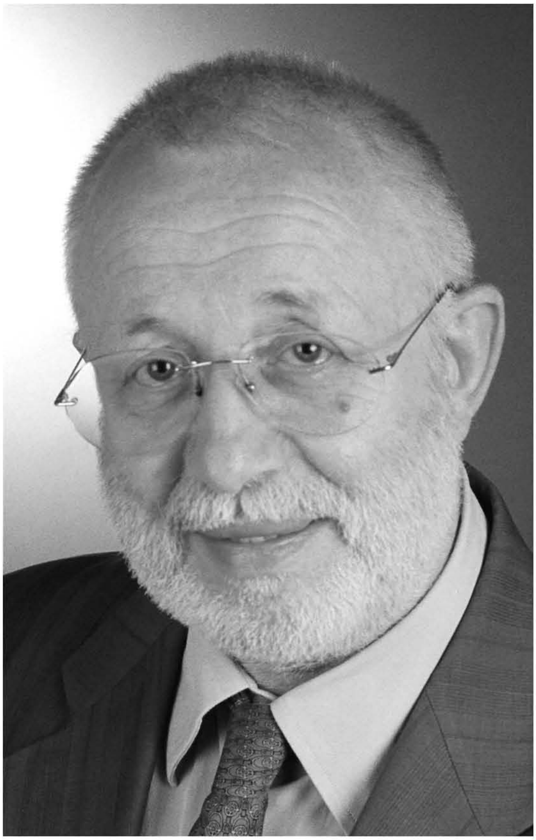
STUDIES IN CLASSIFICATION,
DATA ANALYSIS,
AND KNOWLEDGE ORGANIZATION

D. Baier
R. Decker
L. Schmidt-Thieme
Editors

Data Analysis and Decision Support



Studies in Classification, Data Analysis, and Knowledge Organization



Wolfgang Gaul

Daniel Baier · Reinhold Decker
Lars Schmidt-Thieme
Editors

Data Analysis and Decision Support

Foreword by Shizuhiko Nishisato

Springer

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. Reinhold Decker
Chair of Marketing
Department of Business Administration and Economics
Bielefeld University
Universitätsstr. 25
33615 Bielefeld
Germany
rdecker@wiwi.uni-bielefeld.de

Prof. Dr. Dr. Lars Schmidt-Thieme
Computer Based New Media Group (CGNM)
Institute for Computer Science
University of Freiburg
Georges-Köhler-Allee 51
79110 Freiburg
Germany
lst@informatik.uni-freiburg.de

ISSN 1431-8814
ISBN 3-540-26007-2 Springer-Verlag Berlin Heidelberg New York

Library of Congress Control Number: 2005926825

Springer · Part of Springer Science+Business Media

springeronline.com

Softcover-Design: Erich Kirchner, Heidelberg

SPIN 11427827 43/3153 – 5 4 3 2 1 0 – Printed on acid-free paper

Foreword

It is a great privilege and pleasure to write a foreword for a book honoring Wolfgang Gaul on the occasion of his sixtieth birthday. Wolfgang Gaul is currently Professor of Business Administration and Management Science and the Head of the Institute of Decision Theory and Management Science, Faculty of Economics, University of Karlsruhe (TH), Germany. He is, by any measure, one of the most distinguished and eminent scholars in the world today.

Wolfgang Gaul has been instrumental in numerous leading research initiatives and has achieved an unprecedented level of success in facilitating communication among researchers in diverse disciplines from around the world. A particularly remarkable and unique aspect of his work is that he has been a leading scholar in such diverse areas of research as graph theory and network models, reliability theory, stochastic optimization, operations research, probability theory, sampling theory, cluster analysis, scaling and multivariate data analysis. His activities have been directed not only at these and other theoretical topics, but also at applications of statistical and mathematical tools to a multitude of important problems in computer science (e.g., webmining), business research (e.g., market segmentation), management science (e.g., decision support systems) and behavioral sciences (e.g., preference measurement and data mining). All of his endeavors have been accomplished at the highest level of professional excellence.

Wolfgang Gaul's distinguished contributions are reflected through more than 150 journal papers and three well-known books, as well as 17 edited books. This considerable number of edited books reflects his special ability to organize national and international conferences, and his skill and dedication in successfully providing research outputs with efficient vehicles of dissemination. His talents in this regard are second to none. His singular commitment is also reflected by his contributions as President of the German Classification Society, and as a member of boards of directors and trustees of numerous organizations and editorial boards. For these contributions, the scientific community owes him a profound debt of gratitude.

Wolfgang Gaul's impact on research has been felt in the lives of many researchers in many fields in many countries. The editors of this book, Daniel Baier, Reinhold Decker and Lars Schmidt-Thieme, are distinguished former students of Wolfgang Gaul, whom I had the pleasure of knowing when they were hard-working students under his caring supervision and guidance. This book is a fitting tribute to Wolfgang Gaul's outstanding research career, for it is a collection of contributions by those who have been fortunate enough to know him personally and who admire him wholeheartedly as a person, teacher, mentor, and friend.

A glimpse of the content of the book shows two groups of papers, data analysis and decision support. The first section starts with symbolic data analysis, and then moves to such topics as cluster analysis, asymmetric multidimensional scaling, unfolding analysis, multidimensional data analysis, aggregation of ordinal judgments, neural nets, pattern analysis, Markov process, confidence intervals and ANOVA models with generalized inverses. The second section covers a wide range of papers related to decision support systems, including a long-term strategy for an urban transport system, loyalty programs, heuristic bundling, E-commerce, QFD and conjoint analysis, equity analysis, OR methods for risk management and German business cycles. This book showcases the tip of the iceberg of Wolfgang Gaul's influence and impact on a wide range of research. The editors' dedicated work in publishing this book is now amply rewarded.

Finally, a personal note. No matter what conferences one attends, Wolfgang Gaul always seems to be there, carrying a very heavy load of papers, transparencies, and a computer. He is always involved, always available, and always ready to share his knowledge and expertise. Fortunately, he is also highly organized - an important ingredient of his remarkable success and productivity. I am honoured indeed to be his colleague and friend.

Good teachers are those who can teach something important in life, and Wolfgang Gaul is certainly one of them. I hope that this book gives him some satisfaction, knowing that we all have learned a great deal from our association with him.

Toronto, Canada, April 2005 *Shizuhiko Nishisato*

Preface

This year, in July, Wolfgang Gaul will celebrate his 60th birthday. He is Professor of Business Administration and Management Science and one of the Heads of the Institute of Decision Theory and Management Science at the Faculty of Economics, University of Karlsruhe (TH), Germany. He received his Ph.D. and Habilitation in mathematics from the University of Bonn in 1974 and 1980 respectively.

For more than 35 years, he has been an active researcher at the interface between

- mathematics, operations research, and statistics,
- computer science, as well as
- management science and marketing

with an emphasis on data analysis and decision support related topics.

His publications and research interests include work in areas such as

- graph theory and network models, reliability theory, optimization, stochastic optimization, operations research, probability theory, statistics, sampling theory, and data analysis *(from a more theoretical point of view)* as well as
- applications of computer science, operations research, and management science, e.g., in marketing, market research and consumer behavior, product management, international marketing and management, innovation and entrepreneurship, pre-test and test market modelling, computer-assisted marketing and decision support, knowledge-based approaches for marketing, data and web mining, e-business, and recommender systems *(from a more application-oriented point of view).*

His work has been published in numerous journals like Annals of Operations Research, Applied Stochastic Models and Data Analysis, Behaviormetrika, Decision Support Systems, International Journal of Research in Marketing, Journal of Business Research, Journal of Classification, Journal of Econometrics, Journal of Information and Optimization Sciences, Journal of Marketing Research, Marketing ZFP, Methods of Operations Research, Zeitschrift für Betriebswirtschaft, Zeitschrift für betriebswirtschaftliche Forschung as well as in numerous refereed proceedings volumes.

His books on computer-assisted marketing and decision support – e.g. the well-known and wide-spread book "Computergestütztes Marketing" (published 1990 together with Martin Both) – imply early visions of the nowadays ubiquitous availability and usage of information-, model-, and knowledge-oriented decision aids for marketing managers. Equipped with a profound mathematical background and a high degree of commitment to his research topics, Wolfgang Gaul has strongly contributed in transforming marketing and marketing research into a data-, model-, and decision-oriented quantitative discipline.

Wolfgang Gaul was one of the presidents of the German Classification Society GfKl (Gesellschaft für Klassifikation) and chaired the program committee of numerous international conferences. He is one of the managing editors of "Studies in Classification, Data Analysis, and Knowledge Organization", a series which aims at bringing together interdisciplinary research from different scientific areas in which the need for handling data problems and for providing decision support has been recognized. Furthermore, he was a scientific principal of comprehensive DFG projects on marketing and data analysis.

Last but not least Wolfgang Gaul has positively influenced the research interests and careers of many students. Three of them have decided to honor his merits with respect to data analysis and decision support by inviting colleagues and friends of him to provide a paper for this "Festschrift" and were delighted – but not surprised – about the positive reactions and the high number and quality of articles received.

The present volume is organized into two parts which try to reflect the research topics of Wolfgang Gaul: a more theoretical part on "Data Analysis" and a more application-oriented part on "Decision Support". Within these parts contributions are listed in alphabetical order with respect to the authors' names.

All authors send their congratulations

"Happy birthday, Wolfgang Gaul"

and hope that he will be as active in his and our research fields of interest in the future as he had been in the past.

Finally, the editors would like to cordially thank Dr. Alexandra Rese for her excellent work in preparing this volume, all authors for their cooperation during the editing process, as well as Dr. Martina Bihn and Christiane Beisel from Springer-Verlag for their help concerning all aspects of publication.

Cottbus, Bielefeld, Freiburg
April 2005

Daniel Baier
Reinhold Decker
Lars Schmidt-Thieme

Contents

Part I. Data Analysis

Part II. Decision Support

Part I

Data Analysis

Optimization in Symbolic Data Analysis: Dissimilarities, Class Centers, and Clustering

Hans-Hermann Bock

Institut für Statistik und Wirtschaftsmathematik,
RWTH Aachen, Wüllnerstr. 3, D-52056 Aachen, Germany

Abstract. 'Symbolic Data Analysis' (SDA) provides tools for analyzing 'symbolic' data, i.e., data matrices $X = (x_{kj})$ where the entries x_{kj} are intervals, sets of categories, or frequency distributions instead of 'single values' (a real number, a category) as in the classical case. There exists a large number of empirical algorithms that generalize classical data analysis methods (PCA, clustering, factor analysis, etc.) to the 'symbolic' case. In this context, various optimization problems are formulated (optimum class centers, optimum clustering, optimum scaling,...). This paper presents some cases related to dissimilarities and class centers where explicit solutions are possible. We can integrate these results in the context of an appropriate k-means clustering algorithm. Moreover, and as a first step to probabilistically based results in SDA, we consider the definition and determination of set-valued class 'centers' in SDA and relate them to theorems on the 'approximation of distributions by sets'.

1 Symbolic data analysis

Classical data analysis considers single-valued variables such that, for n objects and p variables, each entry x_{kj} of the data matrix $X = (x_{kj})_{n \times p}$ is a real number (quantitative case) or a category (qualitative case). The term *symbolic data* relates to more general scenarios where x_{kj} may be an interval $x_{kj} = [a_{kj}, b_{kj}] \in I\!R$ (e.g., the interquartile interval of fuel prices in a city), a set $x_{kj} = \{\alpha, \beta, ...\}$ of categories (e.g., $\{green, red, black\}$ the favourite car colours in 2003), or even a frequency distribution (the histogram of monthly salaries in Karlsruhe in 2000). Various statistical methods and a software system SODAS have been developed for the analysis of symbolic data (see Bock and Diday (2000)). In the context of these methods, there arise various mathematical optimization problems, e.g., when defining the dissimilarity between objects (intervals in $I\!R^p$), when characterizing a 'most typical' cluster representative (class center), and when defining optimum clusterings.

This paper describes some of these optimization problems where a more or less explicit solution can be given. We concentrate on the case of *interval-type data* where each object $k = 1, ..., n$ is characterized by a data vector $x_k = ([a_{k1}, b_{k1}], ..., [a_{kp}, b_{kp}])$ with component-specific intervals $[a_{kj}, b_{kj}] \in I\!R$. Such data can be viewed as n p-dimensional intervals (rectangles, hypercubes) $Q_1, ..., Q_n \subset I\!R^p$ with $Q_k := [a_{k1}, b_{k1}] \times \cdots \times [a_{kp}, b_{kp}]$ for $k = 1, ..., n$.

2 Hausdorff distance between rectangles

Our first problem relates to the definition of a dissimilarity measure $\Delta(Q,R)$ between two rectangles $Q=[a,b]=[a_1,b_1]\times\cdots\times[a_p,b_p]$ and $R=[u,v]=[u_1,v_1]\times\cdots\times[u_p,v_p]$ from $I\!R^p$. Given an arbitrary metric d on $I\!R^p$, the dissimilarity between Q and R can be measured by the *Hausdorff distance* (with respect to d)

$$\Delta_H(Q,R) \;:=\; \max\{\delta(Q;R),\, \delta(R;Q)\} \tag{1}$$

where $\delta(Q;R) \;:=\; \max_{\beta\in R}\, \min_{\alpha\in Q} d(\alpha,\beta)$. The calculation of $\Delta_H(Q,R)$ requires the solution of two optimization (minimax) problems of the type

$$\min_{\alpha\in Q} d(\alpha,\beta) \;\longrightarrow\; \max_{\beta\in R} \;=\; \delta(Q;R). \tag{2}$$

A simple example is provided by the one-dimensional case $p=1$ with the standard (Euclidean) distance $d(x,y):=|x-y|$ in $I\!R^1$. Then the Hausdorff distance between two one-dimensional intervals $Q=[a,b], R=[u,v]\subset I\!R^1$, is given by the explicit formula:

$$\Delta_H(Q,R)=\Delta_1([a,b],[u,v]):=\max\{|a-u|,|b-v|\}. \tag{3}$$

For higher dimensions, the calculation of Δ_H is more involved.

2.1 Calculation of the Hausdorff distance with respect to the Euclidean metric

In this section we present an algorithm for determining the Hausdorff distance $\Delta_H(Q,R)$ for the case where d is the Euclidean metric on $I\!R^p$. By the definition of Δ_H in (1) this amounts to solving the minimax problem (2).

Given the rectangles Q and R, we define, for each dimension $j=1,...,p$, the three p-dimensional cylindric 'layers' in $I\!R^p$:

$$\begin{array}{lllcl}
A_{-1}^{(j)} & := & \{x=(x_1,\ldots,x_p)'\in I\!R^p & | \quad x_j<a_j\} & \text{"lower layer"} \\
A_{0}^{(j)} & := & \{x=(x_1,\ldots,x_p)'\in I\!R^p & | \; a_j\le x_j\le b_j\} & \text{"central layer"} \\
A_{+1}^{(j)} & := & \{x=x=(x_1,\ldots,x_p)'\in I\!R^p & | \quad b_j<x_j\} & \text{"upper layer"}
\end{array}$$

such that $I\!R^p$ is dissected into 3^p disjoint (eventually infinite) hypercubes

$$Q(\epsilon) \;:=\; Q(\epsilon_1,\ldots,\epsilon_p) \;:=\; A_{\epsilon_1}^{(1)}\times A_{\epsilon_2}^{(2)}\times\cdots\times A_{\epsilon_p}^{(p)}$$

with $\epsilon=(\epsilon_1,\ldots,\epsilon_p)\in\{-1,0,+1\}^p$. Note that $Q=A_0^{(1)}\times A_0^{(2)}\times\cdots\times A_0^{(p)}$ is the intersection of all p central layers. Similarly, the second rectangle R is dissected into 3^p disjoint (half-open) hypercubes $\tilde{R}(\epsilon) \;:=\; R\cap Q(\epsilon)=R\cap Q(\epsilon_1,\ldots,\epsilon_p)$ for $\epsilon\in\{-1,0,+1\}^p$. Consider the closure $R(\epsilon):=[u(\epsilon),v(\epsilon)]$ of $\tilde{R}(\epsilon)$ with lower and upper vertices $u(\epsilon),v(\epsilon)$ (and coordinates always among the boundary values a_j,b_j,u_j,v_j). Typically, several or even many of these hypercubes are empty. Let $\mathcal{E}$ denote the set of ϵ's with $R(\epsilon)\neq\emptyset$.

We look for a pair of points $\alpha^* \in Q$, $\beta^* \in R$ that achieves the solution of (2). Since R is the union of the hypercubes $R(\epsilon)$ we have

$$\begin{aligned} \delta(Q;R) &:= \max_{\beta\in R}\min_{\alpha\in Q} \parallel \alpha-\beta \parallel \\ &= \max_{\epsilon\in\mathcal{E}}\{\max_{\beta\in R(\epsilon)}\min_{\alpha\in Q} \parallel \alpha-\beta \parallel\} = \max_{\epsilon\in\mathcal{E}}\{m(\epsilon)\} \end{aligned} \tag{4}$$

with $m(\epsilon) := \|\alpha^*(\epsilon)-\beta^*(\epsilon)\|$ where $\alpha^*(\epsilon)\in Q$ and $\beta^*(\epsilon)\in R(\epsilon)$ are the solution of the subproblem

$$\min_{\alpha\in Q} \parallel \alpha-\beta \parallel \quad \rightarrow \quad \max_{\beta\in R(\epsilon)} \quad = \quad \parallel \alpha^*(\epsilon)-\beta^*(\epsilon) \parallel \; = \; m(\epsilon). \tag{5}$$

From geometrical considerations it is seen that the solution of (5), for a given $\epsilon\in\mathcal{E}$, is given by the coordinates

$$\begin{aligned} &\alpha_j^*(\epsilon)=a_j,\ \beta_j^*(\epsilon)=u_j && \text{for} \quad \epsilon_j=-1 \\ &\alpha_j^*(\epsilon)=\beta_j^*(\epsilon)=\gamma_j && \text{for} \quad \epsilon_j=0 \\ &\alpha_j^*(\epsilon)=b_j,\ \beta_j^*(\epsilon)=v_j && \text{for} \quad \epsilon_j=+1 \end{aligned} \tag{6}$$

(here γ_j may be any value in the interval $[a_j,b_j]$) with minimax distance

$$m(\epsilon) \;=\; (\sum_{j:\epsilon_j=-1}(a_j-u_j)^2+\sum_{j:\epsilon_j=0}0^2+\sum_{j:\epsilon_j=+1}(v_j-b_j)\;)^{1/2}.$$

Inserting into (4) yields the solution and the minimizing vertices α^* and $\beta^*of(2)$, and then from (1) the Hausdorff distance $\Delta_H(Q,R)$.

2.2 The Hausdorff distance with respect to the metric d_∞

Chavent (2004) has considered the Hausdorff distance (1) between two rectangles Q,R from $I\!R^p$ with respect to the sup metric $d=d_\infty$ on $I\!R^p$ that is defined by

$$d_\infty(\alpha,\beta) := \max_{j=1,\dots,p}|\alpha_j-\beta_j| \tag{7}$$

for $\alpha=(\alpha_1,...,\alpha_p),\beta=(\beta_1,...,\beta_p)\in I\!R^p$. The corresponding Hausdorff distance $\Delta_\infty(Q,R)$ results from (2) with δ replaced by δ_∞:

$$\begin{aligned} \delta_\infty(Q;R) \;&:=\; \max_{\beta\in R}\min_{\alpha\in Q} d_\infty(\alpha,\beta) \overset{*}{=} \max_{j=1,\dots,p}\{\ \max\{|a_j-u_j|,|b_j-v_j|\}\ \} \\ &= \max_{j=1,\dots,p}\{\ \Delta_1([a_j,b_j],[u_j,v_j])\} \end{aligned}$$

where $\overset{*}{=}$ has been proved by Chavent (2004). By the symmetry of the right hand side we have $\delta_\infty(Q;R)=\delta_\infty(R;Q)$ and therefore by (1):

$$\Delta_\infty(Q,R) = \max_{j=1,\dots,p}\{\Delta_1([a_j,b_j],[u_j,v_j])\} = \max_{j=1,\dots,p}\{\max\{\ |a_j-u_j|,|b_j-v_j|\}\ \}.$$

2.3 Modified Hausdorff-type distance measures for rectangles

Some authors have defined a *Hausdorff-type L_q distance* between Q and R by combining the Hausdorff distances $\Delta_1([a_j,b_j],[u_j,v_j])$ of the p one-dimensional component intervals in a way similar to the classical Minkowski distances:

$$\Delta^{(q)}(Q,R):=(\sum_{j=1}^{p}\Delta_1([a_j,b_j],[u_j,v_j])^q)^{1/q}=(\sum_{j=1}^{p}\max\{|a_j-u_j|^q,|b_j-v_j|^q\})^{1/q}$$

where $q \geq 1$ is a given real number. Below we will use the distance $\Delta^{(1)}(Q,R)$ with $q=1$ (see also Bock (2002)).

3 Typical class representatives for various dissimilarity measures

When interpreting a cluster $C=\{1,...,n\}$ of objects (e.g., resulting from a clustering algorithm) it is quite common to consider a *cluster prototype* (class center, class representative) that should reflect the typical or average properties of the objects (data vectors) in C. When, in SDA, the n class members are described by n data rectangles $Q_1,...,Q_n$ in $I\!R^p$, a formal approach defines the class prototype $G=G(C)$ of C as a p-dimensional rectangle $G \subset I\!R^p$ that solves the optimization problem

$$g(C,G):=\sum_{k\in C}\Delta(Q_k,G) \qquad \rightarrow \qquad \min_G \tag{8}$$

where $\Delta(Q_k,G)$ is a dissimilarity between the rectangles Q_k and G. Insofar $G(C)$ has minimum average distance to all class members. For the case of the Hausdorff distance (1) with a general metric d, there exists no explicit solution formula for $G(C)$. However, explicit formulas have been derived for the special cases $\Delta=\Delta_\infty$ and $\Delta=\Delta^{(1)}$, and also in the case of a 'vertex-type' distance.

3.1 Median prototype for the Hausdorff-type L_1 distance $\Delta^{(1)}$

When using in (8) the Hausdorff-type L_1 distance (2.3), Chavent and Lechevallier (2002) have shown that the optimum rectangle $G=G(C)$ is given by the *median prototype* (9). Its definition uses a notation where any rectangle is described by its mid-point and the half-lengths of its sides. More specifically, we denote by $m_{kj}:=(a_{kj}+b_{kj})/2$ the mid-point and by $\ell_{kj}:=(b_{kj}-a_{kj})/2$ the half-length of the component interval $[a_{kj},b_{kj}]=[m_{kj}-l_{kj},m_{kj}+l_{kj}]$ of a data rectangle Q_k (for $j=1,...,p$; $k=1,...,n$). For a given component j, let $\tilde{\mu}_j:=\text{median}\{m_{1j},\ldots,m_{nj}\}$ be the median of the n midpoints m_{kj} and $\tilde{\lambda}_j:=\text{median}\{\ell_{1j},\ldots,\ell_{nj}\}$ the median of the n half-lengths ℓ_{kj}. Then the optimum prototype for C is given by the *median prototype*

$$G(C)=([\tilde{\mu}_1-\tilde{\lambda}_1,\tilde{\mu}_1+\tilde{\lambda}_1],...,[\tilde{\mu}_p-\tilde{\lambda}_p,\tilde{\mu}_p+\tilde{\lambda}_p]). \tag{9}$$

3.2 Class prototype for the Hausdorff distance Δ_∞

When using the Hausdorff-type distance Δ_∞ induced by the sup norm in $\mathbb{R}^p$, Chavent (2004) has proved that a solution of (8) is provided by the rectangle:

$$G(C) = ([\hat{\alpha}_1, \hat{\beta}_1], ..., [\hat{\alpha}_p, \hat{\beta}_p])$$

with

$$\hat{\alpha}_j := (\max_{k\in C} a_{kj} + \min_{k\in C} a_{kj})/2 \qquad j = 1, ..., p$$
$$\hat{\beta}_j := (\max_{k\in C} b_{kj} + \min_{k\in C} b_{kj})/2 \qquad j = 1, ..., p.$$

In this case, however, the prototype is typically not unique.

3.3 Average-vertex prototype with the vertex-type distance

Bock (2002, 2005) has measured the dissimilarity between two rectangles $Q = [a, b]$, and $R = [u, v]$ by the *vertex-type distance* defined by $\Delta_v(Q, R) := ||u - a||^2 + ||v - b||^2$. Then the optimum class representative is given by

$$G(C) := ([\bar{a}_{C1}, \bar{b}_{C1}], ..., [\bar{a}_{Cp}, \bar{b}_{Cp}]) \tag{10}$$

where $a_{Cj} := \frac{1}{n}\sum_{k\in C} a_{kj}$ and $b_{Cj} := \frac{1}{n}\sum_{k\in C} b_{kj}$ are the averages of the lower and upper boundaries of the componentwise intervals $[a_{kj}, b_{kj}]$ in the class C.

4 Optimizing a clustering criterion in the case of symbolic interval data

Optimization problems are met in clustering when looking for an 'optimum' partition $\mathcal{C} = (C_1, ..., C_m)$ of n objects. In the context of SDA with n , with data rectangles $Q_1, ..., Q_n$ in $\mathbb{R}^p$ we may characterize each cluster C_i by a class-specific prototype rectangle G_i, yielding a *prototype system* $\mathcal{G} = (G_1, ..., G_m)$. Then clustering amounts to minimizing a clustering criterion such as

$$g(\mathcal{C}, \mathcal{G}) := \sum_{i=1}^{m} \sum_{k\in C_i} \Delta(Q_k, G_i) \quad \rightarrow \quad \min_{\mathcal{C}, \mathcal{G}}. \tag{11}$$

It is well-known that a sub-optimum configuration $\mathcal{C}^*, \mathcal{G}^*$ for (11) can be obtained by a k-means algorithm that iterates two partial minimization steps:
(1) minimizing $g(\mathcal{C}, \mathcal{G})$ with respect to the prototype system $\mathcal{G}$ only, and
(2) minimizing $g(\mathcal{C}, \mathcal{G})$ with respect to the partition $\mathcal{C}$ only.
The solution of (2) is given by a *minimum-distance partition* of the objects ('assign each object k to the prototype G_i with minimum dissimilarity $\Delta(Q_k, G_i)$') and is easily obtained (even for the case of the classical Hausdorff distance Δ_H by using the algorithm from section 1). In (1), however, the determination of an optimum prototype system $\mathcal{G}$ for a given $\mathcal{C}$ is difficult for most dissimilarity measures Δ. The importance of the results

cited in section 3 resides in the fact that for a special choice of dissimilarity measures, i.e. $\Delta = \Delta^{(1)}$, Δ_∞, or Δ_v, the optimum prototype system $\mathcal{G} = (G(C_1), ..., G(C_m))$ can be obtained by explicit formulas. Therefore, in these cases, the k-means algorithm can be easily applied.

5 Probabilistic approaches for defining interval-type class prototypes

Most papers published in SDA proceed in a more or less empirical way by proposing some algorithms and apply them to a set of symbolic data. Thus far, there exists no basic theoretical or probability-based approach. As a first step in this direction, we point here to some investigations in probability theory that relate to set-valued or interval-type class prototypes.

In these approaches, and in contrast to the former situation, we do not start with a given set of data vectors in $\mathbb{R}^p$ (classical or interval-type), but consider a random (single-valued or set-valued) element Y in $\mathbb{R}^p$ with a (known) probability distribution P. Then we look for a suitable definition of a **set-valued** 'average element' or 'expectation' for Y. We investigate two cases:

5.1 The expectation of a random set

In the first case, we consider a **random set Y** in $\mathbb{R}^p$, as a model for a 'random data hypercube' in SDA (for an exact definition of a random (closed) set see, e.g., Mathéron (1975)). We look for a subset G of $\mathbb{R}^p$ that can be considered as the 'expectation' $E[Y]$ of Y. In classical integral geometry and in the theory of random sets (and spatial statistics) there exist various approaches for defining such an 'expectation', sometimes also related to the Hausdorff distance (1). Molchanov (1997) presents a list of different definitions, e.g.,
– the Aumann expectation (Aumann (1965)),
– the Fréchet expectation (resulting from optimality problems similar to (8),
– the Voss expectation, and the Vorob'ev expectation.
Körner (1995) defines some variance concepts, and Nordhoff (2003) investigates the properties of these definitions (e.g., convexity, monotonicity,...) in the general case and also for random rectangles.

5.2 The prototype subset for a random vector in $\mathbb{R}^p$

In the second case we assume that Y is a **random vector** in $\mathbb{R}^p$ with distribution P. We look for a **subset** $G = G(P)$ of $\mathbb{R}^p$ that that is 'most typical' for Y or P. This problem has been considered, e.g., by Pärna et al. (1999), Käärik (2000, 2005), and Käärik and Pärna (2003). These approaches relate the definition of $G(P)$ to the 'optimum approximation of a distribution P by a set', i.e. the problem of finding a subset G of $\mathbb{R}^p$ that minimizes the approximation criterion

$$W(G;P) := \int_{\mathbb{R}^p} \psi(d_H(y,G))dP(y) = \int_{y \notin G} \psi(d_H(y,G))dP(y) \;\to\; \min_{G \in \mathcal{G}} \quad (12)$$

Here $d_H(y,G) := inf_{x \in G}\{||y-x||\}$ is the Hausdorff distance between a point $y \in I\!R^p$ and the set G, $\mathcal{G}$ is a given family of subsets (e.g., all bounded closed sets, all rectangles, all spheres in $I\!R^p$), and ψ is a given isotone scaling function on $I\!R_+$ with $\psi(0) = 0$ such as $\psi(s) = s$ or $\psi(s) = s^2$.

Määrik (2005) has derived very general conditions (for P, ψ, and $\mathcal{G}$) that guarantee the existence of a solution $G^* = G(P)$ of the optimization problem (12). Unfortunately, the explicit calculation of the optimum set G^* is impossible in the case of a general P. However, Määrik has shown that a solution of (12) can be obtained by using the empirical distribution P_n of n simulated values from Y and optimizing the empirical version $W(G; P_n)$ with respect to $G \in \mathcal{G}$ (assuming that this is computationally feasible): For a large number n, the solution G_n^* of the empirical problem approximates a solution G^* of (12).

We conclude by an example in $I\!R^2$ where $Y = (Y_1, Y_2)$ has the two-dimensional standard normal distribution $P \hat{=} \mathcal{N}_2(0, I_2)$ with independent components Y_1, Y_2. $\mathcal{G}$ is the family of squares G in $I\!R^2$ that are bounded in some way (see below), and $\psi(s) = s^2$. Then (12) reads as follows:

$$W(G;P) := \int_{y \notin G} \inf_{x \in G}\{||y-x||^2\} dP(y) \;\rightarrow\; \min_{G \in \mathcal{G}}. \tag{13}$$

Since, trivially, $G = I\!R^p$ yields the minimum value $W(I\!R^p; P) = 0$, we introduce restrictions such as $vol(G) \leq c$ or $P(Y \in G) \leq c$ with some threshold $0 < c < \infty$. Under any such restriction, the optimum square will have the form $G = [-a, +a]^2$ centered at the origin and with some $a \geq 0$. The corresponding criterion value is given by

$$W([-a,+a]^2; \mathcal{N}_2) = 4 \cdot [(1+a^2)(1-\Phi(a)) - a\phi(a)] \tag{14}$$

where Φ is the standard normal distribution function in $I\!R^1$, and $\phi(a) = \Phi(a)' = (2\pi)^{-1/2} \exp^{-a^2/2}$ the corresponding density. From this formula an

a	$P(Y \in G)$	$vol(G) = 4a^2$	$W(G; \mathcal{N}_2)$
0	0	0	2
0.284	0.050	0.323	1.2427
0.407	0.100	0.664	0.9962
0.500	1.147	1.000	0.8386
0.675	0.250	1.820	0.5976
0.763	0.307	2.326	0.5000
1.000	0.466	4.000	0.3014
1.052	0.500	4.425	0.2685
1.497	0.750	8.983	0.0917
2.237	0.950	20.007	0.0113

optimum square (an optimum a) can be determined. The previous table lists some selected numerical values, e.g., for the case where the optimum prototype square should comprize only 5% (10%) of the population.

References

AUMANN, R.J. (1965): Integrals and Set-Valued Functions. *J. Math. Analysis and Appl. 12, 1-12.*

BOCK, H.-H. (2002): Clustering Methods and Kohonen Maps for Symbolic Data. *J. Japan. Soc. Comput. Statistics 15, 1-13.*

BOCK, H.-H. (2005): Visualizing Symbolic Data by Kohonen Maps. In: M. Noirhomme and E. Diday (Eds.): *Symbolic Data Analysis and the SODAS Software.* Wiley, New York. (In press.)

BOCK, H.-H. and DIDAY, E. (2000): *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data.* Studies in Classification, Data Analysis, and Knowledge Organization. Springer Verlag, Heidelberg-Berlin.

CHAVENT, M. (2004): A Hausdorff Distance Between Hyperrectangles for Clustering Interval Data. In: D. Banks, L. House, F.R. McMorris, P. Arabie, and W. Gaul (Eds.): *Classification, Clustering, and Data Mining Applications.* Studies in Classification, Data Analysis, and Knowledge Organization. Springer Verlag, Heidelberg, 2004, 333–339.

CHAVENT, M. and LECHEVALLIER, Y. (2002): Dynamical Clustering of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff Distance. In: K. Jajuga, A. Sokołowski, and H.-H. Bock (Eds.): *Classification, Clustering, and Data Analysis.* Studies in Classification, Data Analysis, and Knowledge Organization. Springer Verlag, Heidelberg, 2002, 53–60.

KÖRNER, R. (1995): *A Variance of Compact Convex Random Sets.* Fakultät für Mathematik und Informatik, Bergakademie Freiberg.

KÄÄRIK, M. (2000): Approximation of Distributions by Spheres. In: *Multivariate Statistics. New Trends in Probability and Statistics.* Vol. 5. VSP/TEV, Vilnius-Utrecht-Tokyo, 61–66.

KÄÄRIK, M. (2005): *Fitting Sets to Probability Distributions.* Doctoral thesis, Faculty of Mathematics and Computer Science, University of Tartu, Estonia.

KÄÄRIK, M. and PÄRNA, K. (2003): Fitting Parametric Sets to Probability Distributions. *Acta et Commentationes Universitatis Tartuensis de Mathematica 8, 101-112.*

MATHÉRON, G. (1975): *Random Sets and Integral Geometry.* Wiley, New York.

MOLCHANOV, I. (1997): Statistical Problems for Random sets. In: J. Goutsias (Ed.): *Random Sets: Theory and Applications. Springer, Heidelberg, 27-45.*

NORDHOFF, O. (2003): *Erwartungswerte zufälliger Quader.* Diploma thesis, Institute of Statistics, RWTH Aachen University.

PÄRNA, K., LEMBER, J., and VIIART, A. (1999): Approximating Distributions by Sets. In: W. Gaul and H. Locarek-Junge (Eds.): *Classification in the Information Age.* Studies in Classification, Data Analysis, and Konowledge Organization. Springer, Heidelberg, 215–224.

An Efficient Branch and Bound Procedure for Restricted Principal Components Analysis

Wayne S. DeSarbo[1] and Robert E. Hausman[2]

[1] Marketing Dept., Smeal College of Business, Pennsylvania State University, University Park, PA, USA 16802
[2] K5 Analytic, LLC

Abstract. Principal components analysis (PCA) is one of the foremost multivariate methods utilized in social science research for data reduction, latent variable modeling, multicollinearity resolution, etc. However, while its optimal properties make PCA solutions unique, interpreting the results of such analyses can be problematic. A plethora of rotation methods are available for such interpretive uses, but there is no theory as to which rotation method should be applied in any given social science problem. In addition, different rotational procedures typically render different interpretive results. We present restricted principal components analysis (RPCA) as introduced initially by Hausman (1982). RPCA attempts to optimally derive latent components whose coefficients are integer constrained (e.g.: {-1,0,1}, {0,1}, etc). This constraint results in solutions which are sequentially optimal, with no need for rotation. In addition, the RPCA procedure can enhance data reduction efforts since fewer raw variables define each derived component. Unfortunately, the integer programming solution proposed by Hausman can take far to long to solve even medium-sized problems. We augment his algorithm with two efficient modifications for extracting these constrained components. With such modifications, we are able to accommodate substantially larger RPCA problems. A Marketing application to luxury automobile preference analysis is also provided where traditional PCA and RPCA results are more formally compared and contrasted.

1 Introduction

The central premise behind traditional principal components analysis (PCA) is to reduce the dimensionality of a given two-way data set consisting of a large number of interrelated variables all measured on the same set of subjects, while retaining as much as possible of the variation present in the data set. This is attained by transforming to a new set of composite variates called principal components which are orthogonal and ordered in terms of the amount of variation explained in all of the original variables. The PCA formulation is set up as a constrained optimization problem and reduces to an eigenstructure analysis of the sample covariance or correlation matrix.

While traditional PCA has been very useful for a variety of different research endeavors in the social sciences, a number of issues have been noted in the literature documenting the associated difficulties of implementation and interpretation. While PCA possesses attractive optimal and uniqueness properties, the construction of principal components as linear combinations of all the measured variables means that interpretation is not always easy. One way to aid the interpretation of PCA results is to rotate the components as is done with factor loadings in factor analysis. Richman (1986), Jolliffe (1987), Rencher (1995) all provide various types of rotations, both orthogonal and oblique, that are available for use in PCA rotation. They also discuss the associated problems with such rotations in terms of the different criteria they optimize and the fact that different interpretive results are often derived. In addition, other problems have been noted (c.f. Jolliffe (2002)). Note, PCA successively maximizes variance accounted for. When rotation is utilized, the total variance within the rotated subspace remains unchanged; it is still the maximum that can be achieved overall, but it is redistributed amongst the rotated components more evenly than before rotation. This indicates, as Jolliffe (2002) notes, that information about the nature of any really dominant components may be lost, unless they somehow satisfy the criterion optimized by the particular rotation procedure utilized. Finally, the choice of the number of principal components to retain has a large effect on the results after rotation. As illustrated in Jolliffe (2002), interpreting the most important dimensions for a data set is clearly difficult if those components appear, disappear, and possibly reappear as one alters the number of principal components to retain.

To resolve these problems, Hausman (1982) proposed an integer programming solution using a branch and bound approach for optimally selecting the individual elements or coefficients for each derived principal component as integer values in a restricted set (e.g., {-1,0, +1} or {+1, 0}) akin to what DeSarbo et al. (1982) proposed for canonical correlation analysis. Successive restricted integer valued principal components are extracted sequentially, each optimizing a variance accounted for measure. However, the procedure is limited for small to medium sized problems due to the computational effort involved. This manuscript provides computational improvements for simplifying principal components based on restricting the coefficients to integer values as originally proposed by Hausman (1982). These proposed improvements increase the efficiency of the initial branch and bound algorithm, thus enabling the analysis of substantially larger datasets.

2 Restricted principal components analysis - branch and bound algorithms

A. Definitions

As mentioned, principal components analysis (PCA) is a technique used to reduce the dimensionality of the data $x = [x_1, \ldots, x_p]$ while retaining as much information as possible. More specifically, the first principal component is traditionally defined as that linear combination of the random variables, $y_1 =$

$a_1^T x$, that has maximum variance, subject to the standardizing constraint $a_1^T a_1 = 1$. The coefficient vector a_1 can be obtained as the first characteristic vector corresponding to the largest characteristic root of Σ, the covariance matrix of x. The variance of $a_1^T x$ is that largest characteristic root.

We prefer an alternative, but equivalent, definition provided by Rao (1964), Okamoto (1969), Hausman (1982), and others. The first principal component is defined as that linear combination that maximizes:

$$\phi_1(a_1) = \sum_{i=1}^{n} \sigma_i^2 R^2(y_1; x_i), \tag{1}$$

where $R^2(y_1; x_i)$ is the squared correlation between y_1 and x_i, and σ_i^2 is the variance of x_i. It is not difficult to show that this definition is equivalent to the more traditional one.

$\phi_1(a_1)$ is the variance explained by the first principal component. It is also useful to note that $\phi_1(a_1)$ may be written as the difference between the traces of the original covariance matrix, Σ, and the partial covariance matrix of x given y_1, which we denote as Σ_1. Thus, the first principal component is found by maximizing:

$$\begin{aligned} \phi_1(a_1) &= tr\{\Sigma\} - tr\{\Sigma_1\} \\ &= tr\left\{\Sigma - \left(\Sigma - \frac{\Sigma a_1 a_1^T \Sigma}{a_1^T \Sigma a_1}\right)\right\} \\ &= \frac{a_1^T \Sigma^2 a_1}{a_1^T \Sigma a_1}. \end{aligned} \tag{2}$$

After the first component is obtained, the second is defined as the linear combination of the variables that explains the most variation in Σ_1. It may be computed as the first characteristic vector of Σ_1, or equivalently, the second characteristic vector of Σ. Additional components are defined in a similar manner.

B. The first restricted principal component

In Restricted Principal Components Analysis (RPCA), the same $\phi_1(a_1)$ is maximized, but with additional constraints on the elements of a_1. Specifically, these elements are required to belong to a small pre-specified integer set, Θ. The objective is to render the components more easily interpreted. Toward that end, Hausman (1982) found two sets of particular use: {-1, 0, 1} and {0, 1}. With the first of these, the components become simple sums and differences of the elements of x. With the second, the components are simply sums of subsets of the elements of x. Of course, any other restricted set could be utilized as well.

If the number of variables, p, is small, each RPCA component can be obtained by merely examining all allowable vectors a_1. However, as p increases, the number of possible vectors rapidly becomes too large. In general there are $|\Theta|^p$ possible vectors. (Although in the case of $\Theta = \{-1, 0, 1\}$, only $\frac{3^p+1}{2}$ vectors need to be tried since a_1 and $-a_1$ are equivalent.) In order to overcome this problem, Hausman (1982) proposed a branch and bound (B&B) algorithm which we summarize below.

Consider a solution search tree for RPCA defined in terms of the restricted integer values permitted. Each node in the tree corresponds to a particular optimization problem. The problem at the root node is the PCA problem with no constraints. At the next level, each node corresponds to the same problem, but with the first element of $|\Theta|$ constrained to some fixed value. Corresponding to the possible values, we have $|\Theta|$ nodes. At each subsequent level, one more coefficient is constrained, so that at level $p+1$ all the coefficients are fixed. The value of each node is the maximal value of $\phi_1(a_1)$ obtained for that node's problem. For node i, denote that value as ϕ_{1i}. The RPCA solution is then identified by the leaf (level $p+1$) node with the greatest ϕ_{1i}.

If one had to evaluate all the problems in the tree, there would be no advantage to creating this tree. But note that the value at each node is an upper bound on the value of any node below it since as one moves down the tree constraints are only added. This fact allows large portions of the tree to remain unevaluated. For example, suppose in the course of evaluating nodes, one finds a final node A that has a value of, say, 2. And suppose there is another node B somewhere in the tree that has already been evaluated and found to have the value 1.9. Then there is no need to evaluate any descendents of B since none of them can possibly improve upon node A.

This leads to the following algorithm for finding the optimal final node:

1. Evaluate the root node.
2. Among all evaluated nodes having no evaluated children, find the node with the greatest value.
3. If this node is a leaf node, then it is the solution. Stop.
4. Otherwise, evaluate the children of that node and go to Step 2.

The remaining issue is how to efficiently evaluate ϕ_{1i} for each node i. Note first that $\phi_1(a_1)$ is invariant under scale transformations of a_1. Thus, rather than constraining the first k elements of a_1 to be specific members of Θ, we can instead require that they be proportional to those specific members of Θ. That is, we can simply require that $a_1 = T\nu$ for some ν where T has the form:

$$T = \begin{bmatrix} t & 0 \\ 0 & I \end{bmatrix}. \tag{3}$$

The k-vector t specifies the constrained values of the first k elements of a_1, I is a $(p-k) \times (p-k)$ identity matrix, and ν is a $(p-k+1)$ vector which will be chosen to maximize $\phi_1(T\nu)$. Thus, ϕ_{1i}, the value of node i, is the solution to:

$$\max_{\nu}\left\{\phi_{1i}=\frac{\nu^T T^T \Sigma^2 T\nu}{\nu^T T^T \Sigma T\nu}\right\}. \tag{4}$$

The solution to this problem is the largest characteristic root of $T^T\Sigma^2 T$ with respect to $T^T\Sigma T$, that is, the largest root of the determinantal equation:

$$\left|T^T\Sigma^2 T - \phi_{1i}T^T\Sigma T\right| = 0. \tag{5}$$

C. Additional RPCA components

The first RPCA component is obtained by executing the algorithm specified above. A second RPCA component may be obtained as in standard PCA, by computing Σ_1, the partial covariance matrix of x given the first RPCA component, $y_1 = a_1^T x$, and then applying the above algorithm to Σ_1. This process may be repeated until p RPCA components have been obtained that account for all the variance in the system.

There are two drawbacks to this approach. First, unlike standard PCA analyses, there is no guarantee that the components will be orthogonal. This may make it quite difficult for the analyst to provide an interpretation to the components. Second, after several components have been extracted, there are often many potential candidate components that are equivalent in that they all account for nearly the same amount of variance. For these reasons, it is often useful to add a constraint that each RPCA component have a coefficient vector orthogonal to those of the previous RPCA components. While the addition of this constraint can only limit even further the variance explained, in our tests with various datasets we have never found this decrease to surpass 2%.

The orthogonality constraint for the k+1$^{\text{st}}$ RPCA component can be written as $Aa_{k+1} = 0$, where A is the $k \times p$ matrix whose rows are the first k RPCA components, a_{k+1} is the k+1$^{\text{st}}$ RPCA component, and 0 is a k-vector of zeros. In the subproblem for a particular node in the B&B tree we have $a_{k+1} = T\nu$, and so the orthogonality constraint is incorporated into that sub-problem by adding the constraint $AT\nu = 0$. Thus, the sub-problem is now:

$$\max_{\nu}\left\{\phi_{k+1,i}=\frac{\nu^T T^T \Sigma_k^2 T\nu}{\nu^T T^T \Sigma_k T\nu}\right\} \tag{6}$$
$$s.t. AT\nu = 0,$$

where Σ_k is the partial covariance matrix of x given the first k RPCA components.

Now suppose AT is $k \times m$ and has rank r. If $r = m$, then ν must be the null vector and so the value of the node is zero. Otherwise, $r < m$, let $(AT)^*$ be any $m \times (m - r)$ matrix of full column rank $m - r$ such that:

$$(AT)(AT)^* = 0 \tag{7}$$

Then

$$AT\nu = 0 \tag{8}$$

and

$$(\exists z)\nu = (AT)^* z \tag{9}$$

both define $p - r$ dimensional linear subspaces for ν. Furthermore, since:

$$\nu = (AT)^* z \tag{10}$$

implies:

$$AT\nu = AT(AT)^* z = 0, \tag{11}$$

they must both define the same subspace.

Thus, the sub-problem with the orthogonality constraint can be rewritten as:

$$\begin{aligned} &\max_z \left\{ \phi_{k+1,i} = \frac{\nu^T T^T \Sigma_k^2 T \nu}{\nu^T T^T \Sigma_k T \nu} \right\} \\ &s.t.\ \nu = (AT)^* z. \end{aligned} \tag{12}$$

or equivalently:

$$\max_\nu \left\{ \phi_{k+1,i} = \frac{z^T (AT)^{*T} \Sigma_k^2 T (AT)^* z}{z^T (AT)^{*T} \Sigma_k T (AT)^* z} \right\} \tag{13}$$

The maximal value of $\phi_{k+1,i}$ (the value of node i) is then the largest root of the determinental equation:

$$\left| (AT)^{*T} T^T \Sigma^2 T (AT)^* - \phi_{k+1,i} (AT)^{*T} T^T \Sigma T (AT)^* \right| = 0. \tag{14}$$

3 Efficiency issues with the RPCA branch and bound algorithm

The branch and bound algorithm described above adapted from Hausman (1982) works fine for small to medium sized problems, but the tree can grow far too large for efficient solution when performing RPCA with larger numbers of variables. For these situations, we have found two additional techniques that help to obtain solutions in a reasonable period of time. Both techniques work to keep the tree relatively thin so that we reach the leaf nodes more quickly. These techniques can be used individually or in tandem.

1 Adding depth bias.

In the B&B tree, the value of each node is an upper bound on the values of all nodes below it. At each step, we select the leaf node in the currently evaluated tree having the greatest value. If that node is a leaf node in the complete tree (that is, all coefficients have fixed values in Θ), then we have found our solution. If not, we expand the tree by creating and evaluating the immediate children of that node.

The problem that can arise is that if there are, as is usually the case, many final solutions that are similar in terms of their variance explained, then the evaluated tree can be very bushy with a large number of nodes at each level examined before proceeding to the next level. In order to minimize this behavior, we propose adding a slight bias toward looking at nodes further down the tree rather than widening the tree at a higher level.

In practice, we add a small amount $n\alpha$ to the value of each node, where n is the number of levels that node is removed from the root, and α is typically on the order of 0.001. We have found that while this can lead to non-optimal solutions, the variation explained by those solutions is still well within 1% of the variation explained by the optimal RPCA.

2 Randomizing B&B ordering.

Another issue that can cause the tree to be bushy is the ordering of the variables. The algorithm as described above splits first on the first variable, then on the second, and so on. With a good ordering, the tree may expand almost exclusively downward, perhaps solving a 50 variable problem by evaluating well under 1000 nodes. With a poor ordering for the same problem, the algorithm may evaluate several million nodes without arriving at a final solution. At first, we experimented with various heuristics for identifying a good ordering and then solved the problem using that ordering. Sometimes the number of evaluated nodes was greatly decreased, but in other cases the opposite was true. Since there were often several orders of magnitude difference in the resources required depending upon the ordering, we decided to take a different approach. Instead of deciding on a particular ordering up front, we randomly order the variables and then try to solve for the RPC in a reasonable time. ("Reasonable" is defined by a user-specified maximum number of node evaluations.) If the final solution is not found, the variable ordering is re-randomized and we try once more for the solution. This continues until either the solution is found or a pre-specified number of randomizations have been attempted.

4 A Marketing application to luxury automobile preference analysis

A. Study background

A major U.S. automobile manufacturer sponsored research to conduct personal interviews with $N = 240$ consumers who stated that they were intending to purchase a luxury automobile within the next six months. These customers were demographically screened a priori to represent the target market segment of interest. The study was conduced in a number of automobile clinics occurring at different geographical locations in the U.S. One section of the questionnaire asked the respondent to check off from a list of ten luxury cars, specified a priori by this manufacturer and thought to compete in the same market segment at that time (based on prior research), which brands they would consider purchasing as a replacement vehicle after recalling their perceptions of expected benefits and costs of each brand. Afterwards, the respondents were asked to use a ten-point scale to indicate the intensity of their purchase consideration for the vehicles initially checked as in their respective consideration sets. The ten nameplates tested were (firms that manufacture them in parentheses): Lincoln Continental (FORD), Cadillac Seville (GM), Buick Riviera (GM), Oldsmobile Ninety-Eight (GM), Lincoln Town Car (FORD), Mercedes 300E (DAIMER/CHRYSLER), BMW 325i (BMW), Volvo 740 (FORD now but not at the time of the study), Jaguar XJ6 (FORD), and Acura Legend (HONDA). The vast majority of respondents' elicited consideration/choice sets in the range of 2 - 6 automobiles from the list of ten. See DeSarbo and Jedidi (1995) for further study details. As in Hauser and Koppelman (1979) and Holbrook and Huber (1979), we will use PCA here to generate product spaces

B. Traditional PCA results

Table 1 presents the traditional PCA results from an analysis of the resulting correlation matrix generated from this data. As shown, using the traditional "eigenvalue > 1.00 rule", some three PCA components result which account for 57.27% of the variance. The first PCA component clearly separates the domestic vs. the import automobiles, but the remaining two factors are difficult to interpret without some form of rotation employed.

C. RPCA results

Table 2 presents the RPCA analysis for this same data set using binary loadings without orthogonality constraints (based on inspection of a variety of RPCA analyses with various types of constraints imposed). Here, the first component clearly distinguishes the domestic cars from the imports. The second component separates the Japanese cars from the non-Japanese cars.

PCA Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.972	29.722	29.722	2.972	29.722	29.722	2.266	22.657	22.657
2	1.536	15.359	45.080	1.536	15.359	45.080	1.751	17.510	40.166
3	1.219	12.191	57.271	1.219	12.191	57.271	1.710	17.105	57.271
4	.941	9.412	66.683						
5	.763	7.634	74.317						
6	.682	6.824	81.141						
7	.604	6.042	87.183						
8	.526	5.262	92.444						
9	.389	3.886	96.330						
10	.367	3.670	100.000						

PCA Component	1	2	3
Lincoln Continental	-.371	.622	-.444
Cadillac Seville	-.411	.567	.181
Buick Riviera	-.517	.128	.518
Oldsmobile Ninety-Eight	-.538	.163	.566
Lincoln Town Car	-.607	.463	-.335
Mercedes 300E	.628	.430	.228
BMW 325i	.644	.453	.326
Volvo 740	.607	.313	.060
Jaguar XJ6	.367	.275	-.354
Acura Legend	.654	.071	.028

Table 1. Traditional PCA Results for Luxury Automobile Preference Analysis: Total Variance Explained (upper table) and Component Matrix (lower table)

PCA Component	1	2	3
Root	2.579903	1.838475	1.172538
% of Var.	25.8%	18.4%	11.7%
Cum %	25.8%	44.2%	55.9%
Nodes	249	29	93
Lincoln Continental	0	1	1
Cadillac Seville	0	1	0
Buick Riviera	0	1	0
Oldsmobile Ninety-Eight	0	1	0
Lincoln Town Car	0	1	1
Mercedes 300E	1	1	0
BMW 325i	1	1	0
Volvo 740	1	1	0
Jaguar XJ6	1	1	1
Acura Legend	1	0	0

Table 2. RPCA Results for Luxury Automobile Preference Analysis

And the third component discriminates the Ford brands from the non-Ford brands (Volvo was not purchased by Ford at the time of this study). Note that the total variance explained is 55.9% - one only loses less than 1.4% in explained variance in terms of this restricted solution which is imminently more interpretable than the traditional PCA solution presented in Table 1.

References

DESARBO, W.S., HAUSMAN, R.E., LIN, S., and THOMPSON, W. (1982): Constrained canonical correlation. *Psychometrika, 47, 489–516.*

DESARBO, W.S. and JEDIDI, K. (1995): The spatial representation of heterogeneous consideration sets. *Marketing Science, 14, 326–342.*

HAUSER, J.R. and KOPPELMAN, F.S. (1979): Alternative perceptual mapping techniques: Relative accuracy and usefulness. *Journal of Marketing Research, 16, 495–506.*

HAUSMAN, R. (1982): Constrained multivariate analysis. In H. Zanakis and J.S. Rustagi (Eds.): *TIMS/Studies in the management sciences.* Vol. 19, North-Holland Publishing Company, Amsterdam, 137–151.

HOLBROOK, M.B. and HUBER, J. (1979): Separating perceptual dimensions from affective overtones: An application to consumer aesthetics. *Journal of Consumer Research, 5, 272–283.*

JOLLIFFE, I.T. (1987): Rotation of principal components: Some comments. *Journal of Climatology, 7, 507–510.*

JOLLIFFE, I.T. (2002): *Principal component analysis* (2nd Edition). Springer-Verlag, New York.

OKAMATO, M. (1969): Optimality of principal components. In P.R. Kishnaiah (Ed.): *Multivariate Analysis II.* Academic Press, New York.

RAO, C. R. (1964): The use and interpretation of principal component analysis in applied research. *Sankhya, A 26, 329–359.*

RENCHER, A.C. (1995): *Methods of multivariate analysis.* Wiley, New York.

RICHMAN, M.B. (1986): Rotation of principal components. *Journal of Climatology, 6, 293–335.*

A Tree Structured Classifier for Symbolic Class Description

Edwin Diday[1], M. Mehdi Limam[1], and Suzanne Winsberg[2]

[1] LISE-CEREMADE, Université Paris IX Dauphine,
Place du Maréchal de lattre de Tassigny, 75775 Paris, France
[2] IRCAM, 1 Place Igor Stravinsky, 75004 Paris, France

Abstract. We have a class of statistical units from a population, for which the data table may contain symbolic data; that is rather than having a single value for an observed variable, an observed value for the aggregated statistical units we treat here may be multivalued. Our aim is to describe a partition of this class of statistical units by further partitioning it, where each class of the partition is described by a conjunction of characteristic properties. We use a stepwise top-down binary tree method. At each step we select the best variable and its optimal splitting to optimize simultaneously a discrimination criterion given by a prior partition and a homogeneity criterion; we also aim to insure that the descriptions obtained describe the units in the class to describe and not the rest of the population. We present a real example.

1 Introduction

Suppose we want to describe a class, C, from a set or population of statistical units. A good way to do so would be to find the properties that characterize the class, and one way to attain that goal is to partition the class. Clustering methods are often designed to split a class of statistical units, yielding a partition into L subclasses, or clusters, where each cluster may then be described by a conjunction of properties. Partitioning methods generally fall into one of two types namely: clustering methods which optimize an intra-class homogeneity criterion, and decision trees, which optimize an inter-class criterion. Our method combines both approaches.

To attain our goal we partition the class using a top-down binary divisive method. It is of prime importance that these subdivisions of the original class to be described, C, be homogenous with respect to the selected or available group of variables found in the data base. However, if in addition, we need to explain an external criterion, which gives rise to an a priori partition of the population, or some part of it, which englobes C the class to describe, we need also to consider a discrimination criterion based on that a priori partition into say S categories. Our technique arrives at a description of C by producing a partition of C into subclasses or clusters $P_1, ..., P_l, ..., P_L$ where P_l satisfies both a homogeneity criterion and a discrimination criterion with respect to an a priori partition. So unlike other classification methods which are either unsupervised or supervised, this method has both unsupervised and supervised aspects.

We use a divisive top-down monothetic approach such that each split is carried out using only one variable, as it provides a clearer interpretation of the obtained clusters. Divisive methods of this type are often referred to as tree structured classifiers with acronyms such as CART and ID3 (see Breiman et al.(1984), Quinlan (1986)). Not only does our paper combine the two approaches: supervised and nonsupervised learning, to obtain a description induced by the synthesis of the two methods, which is in itself an innovation, but it can deal with interval type and histogram data, ie data in which the entries of the data table are intervals and weighted categorical or ordinal variables, respectively. These data are inherently richer, possessing potentially more information than the data previously considered in the classical algorithms mentioned above. We encounter this type of data when dealing with more complex, aggregated statistical units found when analyzing very large data sets. Moreover, it may be more interesting to deal with aggregated units such as towns rather than with the individual inhabitants of the towns. Then the resulting data set, after the aggregation will most likely contain symbolic data rather than classical data values. By symbolic data we mean that rather than having a specific single value for an observed variable, an observed value for an aggregated statistical unit may be multivalued. For example, the observed value may be an interval or a histogram. For a detailed description of symbolic data analysis see Bock and Diday (2000). Naturally, classical data are a special case of the interval and histogram type of data considered here. This procedure works interval or histogram data as well as for classical numerical or nominal data, or any combination of the above. Others have developed divisive algorithms for data types encountered when dealing with symbolic data, considering either a homogeneity criterion or a discrimination criterion, but not both simultaneously; Chavent (1997) has done so for unsupervised learning, while Périnel (1999) has done so for supervised learning.

2 The method

Five inputs are required for this method: 1) the data, consisting of n statistical units, each described by K histogram variables; 2) the prior partition of the population; 3) the class, C, the user aims to describe; and 4) a coefficient α, which gives more or less importance to the discriminatory power of the prior partition or to the homogeneity of the description of the given class, C. Alternatively, instead of specifying this last coefficient, the user may choose to determine an optimum value of this coefficient, using this algorithm; 5) the choice of a stopping rule, including the overflow rate.

The algorithm always generates two kinds of output. The first is a graphical representation, in which the class to describe, C, is represented by a binary tree. The final leaves are the clusters constituting the class and each branch represents a cutting (y, c). The second is a description: each final leaf is described by the conjunction of the cutting values from the top of the tree to this final leaf. The class, C, is then described by a disjunction of these conjunctions. If the user wishes to choose an optimal value of α using our data driven method, a graphical representation enabling this choice is also generated as output.

Let $H(N)$ and $h(N_1; N_2)$ be respectively the homogeneity criterion of a node N and of a couple of nodes $(N_1; N_2)$. Then we define $\Delta H(N) = H(N) - h(N_1; N_2)$. Similarly we define $\Delta D(N) = D(N) - d(N_1; N_2)$ for the discrimination criterion. The quality Q of a node N (respectively q of a couple of nodes $(N_1; N_2)$) is the weighted sum of the two criteria, namely $Q(N) = \alpha H(N) + \beta D(N)$ (respectively $q(N_1; N_2) = \alpha h(N_1; N_2) + \beta d(N_1; N_2)$) where $\alpha + \beta = 1$. So the quality variation induced by the splitting of N into $(N_1; N_2)$ is $\Delta Q(N) = Q(N) - q(N_1; N_2)$. We maximize $\Delta Q(N)$. Note that since we are optimizing two criteria the criteria must be normalized. The user can modulate the values of α and β so as to weight the importance that he gives to each criterion. To determine the cutting $(y; c)$ and the node to cut: first, for each node N select the cutting variable and its cutting value minimizing $q(N_1; N_2)$; second, select and split the node N which maximizes the difference between the quality before the cutting and the quality after the cutting, $max\Delta Q(N) = max[\alpha\Delta H(N) + \beta D(N)]$.

We recall that we are working with interval and histogram variables. We must define what constitutes a cutting for this type of data and what constitutes a cutting value. For an interval variable y_i, we define the cutting point using the mean of the interval. We order the means of the intervals for all units, the cutting point is then the mean of two consecutive means of intervals. For histogram type variables the cutting value is defined on the value of the frequency of just one category, or on the value of the sum of the frequencies of several categories. So for each subset of the categories of y_i, the cutting value is chosen as the mean of any two sums of the frequencies of these categories. See Vrac et al. (2002) and Limam et al. (2003) for a detailed explanation with examples.

The clustering or homogeneity criterion we use is an inertia criterion. This criterion is used in Chavent (1997). The inertia of a class N is

$$H(N) = \sum_{w_i \in N} \sum_{w_j \in N} \frac{p_i p_j}{2\mu} \delta^2(w_i, w_j),$$

and

$$h(N_1, N_2) = H(N_1) + H(N_2);$$

where p_i = the weight of individual w_i, and $\mu = \sum_{w_i \in N} p_i$ = the weight of class N, and δ is a distance between individuals. For histograms with weighted categorical variables, we choose δ, defined as,

$$\delta(w_i;w_j) = \sum_{k=1}^{K} \sum_{m=1}^{mod[k]} |y_k^m(w_i) - y_k^m(w_j)|^2,$$

where $y_k^m(w)$ is the value of the category m of the variable k for the individual w, $mod[k]$ is the number of categories of the variable k, and K is the number of variables. This distance must be normalized. We normalize it to fall in the interval [0,1]. So δ must be divided by K to make it fall in the interval [0,1].

For interval quantitative variables, we choose δ^2, defined as,

$$\delta^2(w_i, w_k) = \left| \frac{\frac{y_j^{\min}(w_i)+y_j^{\max}(w_i)}{2} - \frac{y_j^{\min}(w_k)+y_j^{\max}(w_k)}{2}}{m_j} \right|^2,$$

where $[y_j^{\min}(w_i), y_j^{\max}(w_i)]$ is the interval value of the variable y_j for the unit w_i and $m_j = \left|\max_{wi} y_j^{\max} - \min_{wi} y_j^{\min}\right|$ which represent the maximum area of the variable y_j. We remark that δ^2 fall in the interval [0, 1].

The discrimination criterion we choose is an impurity criterion, Gini's index. Gini's index, which we denote as D, was introduced by Breiman et al. (1984) and measures the impurity of a node N with respect to the prior partition $G_1, G_2, ..., G_J$ by

$$D(N) = \sum_{l \neq j} p_l p_j = 1 - \sum_{j=1,...,J} p_j^2,$$

with $p_j = n_j/n$, $n_j = card(N \cap G_j)$ and $n = card(N)$ in the classical case. In our case n_j = the number of individuals from G_j such that their characteristics verify the current description of the node N. To normalize $D(N)$ we multiply it by $J/(J-1)$; where J is the number of prior classes; it then lies in the interval [0,1].

We have two types of affectation rules according to whether the variable y_j is quantitative, or qualitative.

If y_j is a quantitative variable that is an interval multivalued continuous variable, we must define a rule to assign to N_1 and N_2 (respectively, the left and the right node of N) an interval with respect to a cutting point c. We define p_{wi} the probability to assign a unit w_i to N_1 with $y_j(w_i) = [y_j^{\min}(w_i), y_j^{\max}(w_i)]$ by :

$$\mathrm{p}_{wi} = \begin{array}{ll} \frac{c - y_j^{\min}(w_i)}{y_j^{\max}(w_i) - y_j^{\min}(w_i)} & \text{if } c \in [y_j^{\min}(w_i), y_j^{\max}(w_i)] \\ 0 & \text{if } c < y_j^{\min}(w_i) \\ 1 & \text{if } c > y_j^{\max}(w_i) \end{array}$$

then w_i belongs to N_1 if $\mathrm{p}_{wi} \geq \frac{1}{2}$ else it belongs to N_2.

If y_j is a qualitative multivalued weighted categorical (histogram type) variable, we define a rule to assign to N_1 and N_2 (respectively, the left and the right node of N) a multivalued weighted categorical description with respect to a cutting categorical set V and a cutting frequency value c_V. We define p_{wi} the probability to assign a unit w_i to N_1 :

$\mathrm{p}_{wi} \quad = \sum_{m \in V} y_j^m(w_i)$

with $y_j^m(w_i)$ the frequency of the category m of the variable y_j for the unit w_i. then w_i belongs to N_1 if $\mathrm{p}_{wi} \geq c_V$ else it belongs to N_2.

The user may choose to optimize the value of the coefficient α. To do so, one must fix the stopping rule. The influence of the coefficient α can be determinant both in the construction of the tree and in its prediction qualities. This variation influences splitting and consequently results in different terminal nodes. We need to find the inertia of the terminal nodes and the rate of misclassification as we vary α. Then we can determine the value of α which optimizes both the inertia and the rate of misclassification ie the homogeneity and discrimination simultaneously.

For each terminal node t of the tree T associated with class c_s we can calculate the corresponding misclassification rate $R(s/t) = \sum_{r=1}^{L} P(r/t)$ where $r \neq s$ and $P(r/t) = \frac{n_r(t)}{n_t}$ is the proportion of the individuals of the node t allocated to the class c_s but belonging to the class c_r. The misclassification MR of the tree T is the sum over all terminal nodes ie $MR(A) = \sum_{t \in A} \frac{n_i}{n} R(s/t) = \sum_{t \in A} \sum_{r=1}^{L} \frac{n_r(t)}{n}$, where $r \neq s$. For each terminal node of the tree T we can calculate the corresponding inertia, $H(t)$ and we can calculate the total inertia by summing over all the terminal nodes. So, $H(t) = \frac{1}{2n|t|} \sum_{w_i \in t} \sum_{w_j \in t} \delta(W_i, W_j)$ with $|t| = card(t)$, and the total inertia of T, $I(A) = \sum_{t \in T} H(t)$. For each value of α we build trees from many bootstrap samples and then calculate the inertia and misclassification rate. For each sample and for each value of α in a given number of steps between 0 and 1, we build a tree; then we calculate the bootstrap mean our two parameters (inertia and misclassification rate) for each value of α. In order to visualize the variation of the two criteria, we display a curve showing the inertia and a curve showing the misclassification rate as a function of α. The optimal value of α is the one which minimizes the sum of the two parameters.

We now consider stopping rules. Our aim here is to produce a description of a class C coming from a population of κ units. Naturally, the description includes all the units of the class C because we induce this description from all the units of this class. However, units not belonging to this class but included in this description should be minimized. For example, consider the class C to describe is the districts of Great Britain containing a number of inhabitants greater than $500k$ and the population is the districts of Great Britain. It is desirable that the description should include as little as possible districts with a number of inhabitants less than $500k$. So it is of interest to consider the overflow of the class to describe in making a stopping rule.

We define the overflow rate of a node $N : OR(N) = \frac{n(\overline{C}_N)}{n_N}$ with $n(\overline{C}_N) =$ number of units belonging to the complement $\overline{C}$ of the class C which verify the current description of the node N and n_N the total number of units verifying the description of the node N.

A node N is considered as terminal (a leaf) if: the variation of its overflow $\Delta OR(N)$ overflow is less than a threshold fixed by the user or a default value say 10%; the difference between the discrimination before the cutting and the quality after the cutting $\Delta D(N)$ is less than a threshold fixed by the user or a default value say 10%; its size is not large enough (value < 2); it creates two empty son nodes (value < 1). We stop the division when all the nodes are terminal or alternatively if we reach a number of divisions set by the user.

3 Example

Our real data example deals with real symbolic data with a population Ω of u =18 units. Each unit represent a soccer team in the French league 2000/2001. The class to describe C gathers the teams with a large proportion of French players ($\geq$ 70%) and induced by a nominal variable Y_C with two categories large (La) and small (Sm). We have u_c =16 units in this class. Because we have aggregated the data for the players of each team, we are not dealing with classical data with a single value for each variable for each statistical unit, here the team. Each unit is described by K = 3 explanatory variables, two of them are interval variables : the age, (AGE), and weight, (POIDS), of the players of each team, these interval variables describe the variation among all the players of each team. The third variable is a histogram-type variable which describes the distribution of the positions of the players in each team, (POS), the categories of this histogram type variable are : attack(1), goal keeper(2), defence(3) and midfield (4). We also have an a priori partition of Ω with two classes (G_1, G_2) : G_1 = *best teams:HAUT* (the ten best teams of the championship) and G_2 = *worst teams:FAIBLE.* Our aim is to explain the factors which discriminate the best from the worst teams in the class of teams with a large (La) proportion of French players. But we also wish to have good descriptors of the resultant clusters due to their homogeneity. The class to describe, C, contains 16 teams, $\overline{C}$ contains the rest of the teams.

We stopped the division of a node if its $\Delta OR(N)$ and $\Delta D(N)$ is less than 10% or if we reached 5 terminal nodes. We show the results for three values of α, $\alpha = 1$, $\alpha = 0$ and α optimized with data driven method $\alpha = 0.4$. The results are shown in Table 1; $OR(T)$ is the overflow rate aver all terminal nodes.

At $\alpha = 0.4$ the rate of misclassification is equal to 6.25% the same as that for $\alpha = 0$ and less than for $\alpha = 1$. The inertia only slightly increased over that for $\alpha = 1$, which is the best rate. If we choose $\alpha = 0.4$, we obtain five terminal nodes and an overflow rate equal to 6.25% which is good; we have a good misclassification rate and a much better class description, than that which we obtain when considering only a discrimination criterion; and

α	$I(T)$	$MR(T)\%$	$I + MR_{Norm}$	$OR(T)\%$
1	0.156	25	1.82	12.5
0	0.667	6.25	0.979	12.5
0.4	0.388	6.25	0.546	6.25

Table 1. Results of the example

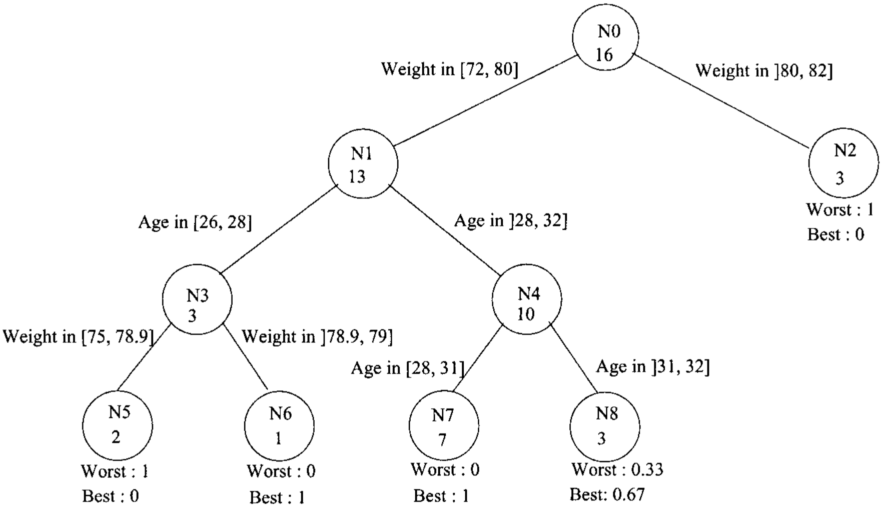


Fig. 1. Graphical representation of the tree with alpha=0.4

we have a much better misclassification rate than that which we obtain when considering only a homogeneity criterion.

From the tree in Figure 1, we derive the descriptions presented in Table 2, each of which corresponds to a terminal node.

Some variables are found in the description more than once because they are selected two times in the cutting and others are missing because they are not selected. Each of these descriptions describes a homogenous and well discriminated node. On examination of the resultant tree we remark that the poorer teams are those who have either the heaviest players or those that have the lightest and the youngest ie the most inexperienced players. The class to describe C (the units with a large proportion of French players) is described by the disjunction of the all descriptions presented in the table below $desc(C) = D_1 \vee D_2 \vee D_3 \vee D_4 \vee D_5$.

4 Conclusion

In this chapter we present a new approach to get a description of a class. Our approach is based on a divisive top-down tree method, restricted to recursive binary partitions, until a suitable stopping rule prevents further

Description	N_i
D_1: $[Weight \in [72,79]] \wedge$ $[Age \in [27,32]] \wedge$ $[Age \in [31,32]]$	N_8
D_2: $[Weight \in [72,79]] \wedge$ $[Age \in [27,32]] \wedge$ $[Age \in [28,30]]$	N_7
D_3: $[Weight \in [72,79]] \wedge$ $[Age \in [26,27]] \wedge$ $[Weight \in]74.5,75.5]]$	N_6
D_4: $[Weight \in [72,79]] \wedge$ $[Weight \in [75.5,78.5]] \wedge$ $[Age \in [26,27]]$	N_5
D_5: $[Weight \in [80,82]]$	N_2

Table 2. The descriptions of the class

divisions. This method is applicable to most types of data, that is, classical numerical and categorical data, symbolic data, including interval type data and histogram type data, and any mixture of these types of data. The idea is to combine a homogeneity criterion and a discrimination criterion to describe a class and explain an a priori partition. The class to describe can be a class from a prior partition, the whole population or any class from the population. Having chosen this class, the interest of the method is that the user can choose the weights α and thus $\beta = 1 - \alpha$ he/she wants to put on the homogeneity and discrimination criteria respectively, depending on the importance of these criteria to reach a desired goal. Alternatively, the user can optimize both criteria simultaneously, choosing a data driven value of α. A data driven choice can yield an almost optimal discrimination, while improving homogeneity, leading to improved class description. In addition, to obtain the class description for the chosen class, the user may select to use a stopping rule yielding a description which overflows the class as little as possible and which is as pure as possible.

References

BOCK, H.H. and DIDAY, E. (2000): *Analysis of Symbolic Data.* Springer, Heidelberg.

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A., and STONE, C.J. (1984): *Classification and Regression Trees.* Wadsworth, Belmont, California.

CHAVENT, M. (1997): *Analyse de Données Symboliques, Une Méthode Divisive de Classification.* Thèse de Doctorat, Université Paris IX Dauphine.

DIDAY, E. (2000): Symbolic Data Analysis and the SODAS Project: Purpose, History, and Perspective In: H.H. Bock and E. Diday (Eds.): *Analysis of Symbolic Data.* Springer, Heidelberg, 1–23.

LIMAM, M., DIDAY, E., and WINSBERG, S. (2003): Symbolic class description with interval data. *Journal of Symbolic Data Analysis, 1.* [available online: http://www.jsda.unina2.it/newjsda/volumes/index.htm.]

PÉRINEL, E. (1999): Construire un arbre de discrimination binaire à partir de données imprécises. *Revue de Statistique Appliquée, 47, 5–30.*

QUINLAN, J.R. (1986): Induction of decision trees. *Machine Learning, 1, 81–106.*

VRAC, M., LIMAM, M., DIDAY, E., and WINSBERG, S. (2002): Symbolic class description In: K. Jajuga, A. Sokolowski, and H.H. Bock (Eds.): *Classification, Clustering, and Data Analysis.* Springer, Heidelberg, 329–337.

A Diversity Measure for Tree-Based Classifier Ensembles

Eugeniusz Gatnar

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

Abstract. Combining multiple classifiers into an ensemble has proved to be very successful in the past decade. The key of this success is the diversity of the component classifiers, because many experiments showed that unrelated members form an ensemble of high accuracy.

In this paper we propose a new pairwise measure of diversity for classifier ensembles based on Hamann's similarity coefficient.

1 Introduction

The development of aggregation methods is one of the recent major advances in classification and regression. Multiple models are built independently and then aggregated into an ensemble or a committee to achieve optimal accuracy.

Classifier ensembles proved to be high performance classification systems in numerous applications, e.g. pattern recognition, document analysis, personal identification, data mining etc. Their high accuracy is usually achieved if the members of an ensemble are "weak" and diverse. The term "weak" refers to classifiers that have high variance, e.g. classification trees, nearest neighbours, and neural nets.

Diversity among classifiers means that they are different from each other, i.e. they misclassify different examples. This is usually obtained by using different training subsets, assigning different weights to instances or selecting different subsets of features (subspaces).

Tumer and Ghosh (1996) have shown that the ensemble error decreases with the reduction in correlation between component classifiers. Then Breiman (2001) introduced an upper bound for the error of the ensemble of classification trees and found it depends on the averaged pairwise correlation between members of the ensemble.

Several measures of dependence between the base classifiers have been proposed in the literature. For example Kuncheva and Whitaker (2003) have explored ten measures of diversity taken from numerical taxonomy.

This paper is organised as follows. In Section 2 we give a short description of the methods for combining classifiers. Section 3 contains a discussion on classifier dependency and relations between the diversity and accuracy of classifier ensembles.

In Section 4 we review several measures of classifier dependence already proposed in the literature. The new measure of diversity based on Hamann similarity coefficient is introduced in Section 5. Section 6 contains a brief description of the results of our experiments. The last section contains a short summary.

2 Combining methods

Given a set of training examples: $T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, we form a set of subsets: $T_1, T_2, \ldots, T_K$ and a classifier C is fitted to each of them, resulting in a set of base classifiers: $C_1, C_2, \ldots, C_K$. Then they are combined in some way to produce the ensemble C^*.

Several variants of aggregation methods have been developed in the past decade. They differ in two aspects: the way that the subsets used to train component classifiers are formed, and the method according to which the base classifiers are combined.

Generally, there are three approaches to obtain the training subsets:

- Manipulating training examples, e.g. *Bagging* (Breiman (1996)); *Boosting* (Freund and Shapire (1997)) and *Arcing* (Breiman (1998)).
- Manipulating input features, e.g. *Random subspaces* (Ho (1998)) and *Random forests* (Breiman (2001)).
- Manipulating output values: *Error-correcting output coding* (Dietterich and Bakiri (1995)).

Having a set of classifiers, they can be combined using one of the following methods:

- *Averaging methods*, e.g. average vote and weighted vote.
- *Non-linear methods*, e.g. majority vote (the component classifiers vote for the most frequent class as the predicted class), maximum vote, Borda Count method, etc.
- *Stacked generalisation* developed by Wolpert (1992).

3 Diversity of component classifiers

In many experiments one has observed that improvement of classification accuracy depends on the correlation between the ensemble members. Formal proof of this result has been given in two papers.

Tumer and Ghosh (1996) provided a decomposition of the ensemble prediction error into two factors (using simple averaging combination rule):

$$e(C^*) = e^B(C^*) + \frac{1 + \delta(K-1)}{K} e(C_k) \qquad (1)$$

where $e^B(C^*)$ is the error achieved by Bayes rule, δ is a correlation coefficient between errors of component classifiers, K is the number of base classifiers and $e(C_k)$ is an error of individual classifier. We can learn from the equation (1) that the ensemble error decreases with the decrease of the correlation between members of the ensemble.

Breiman (2001) developed an upper bound for the error of the ensemble:

$$e(C^*) \leq \bar{\rho}\frac{1 - s^2}{s^2}, \tag{2}$$

where $\bar{\rho}$ is the averaged pairwise correlation between component classifiers and s is the "strength" of the ensemble, i.e. the expected size of the margin by which the ensemble obtains the correct prediction.

The conclusion from the above equations is obvious: the stronger the correlation between members of the ensemble, the higher the ensemble classification error. Therefore, the independence or diversity of the base classifiers should be maximized.

Sharkey and Sharkey (1997) distinguished four levels of diversity :

- Level 1 - no more than one classifier is wrong on each example.
- Level 2 - up to half of classifiers could be wrong for each example (majority vote is alway correct).
- Level 3 - at least one classifier is correct for each example.
- Level 4 - none of the classifiers is correct for some examples.

The level of diversity among candidate classifiers determines the method by which they should be combined. For example, the majority vote is good for the classifiers that exhibit the level 1 an 2 diversity. Otherwise some more complex methods, e.g. stacked generalization, are more appropriate.

Several combining methods that take into account the diversity of classifiers have been proposed in the literature. Rosen (1996) has presented a combination method that incorporates an error-decorrelation penalty term. It allows component classifiers to make errors which are uncorrelated. Hashem (1999) has proposed the use of relative accuracy of component classifiers as weights in linear combinations of the members of an ensemble. Oza and Tumer (1999) developed a method named *input decimation* that eliminates features low correlated with the class. Zenobi and Cunnigham (2001) have used the hill-climbing search for feature subsets that is guided by diversity measure.

Recently, Melville and Mooney (2003) have developed a method called *DECORATE* that reduces the classification error of an ensemble by increasing diversity. It adds artificial examples to the original training set, i.e. examples that have been oppositely labelled. They have also observed that the Spearman's rank correlation between the diversity of ensembles and the reduction of their classification error is 0.66 ("fairly strong").

4 Measures of diversity

Let $[\hat{C}(\mathbf{x}_1), \hat{C}(\mathbf{x}_2), \ldots, \hat{C}(\mathbf{x}_N)]$ be a vector of predictions of the classifier C for the set of examples $V = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$. The relationship between a pair of classifiers C_i and C_j can be shown in the form of the 2×2 contingency table (Table 1). In order to use this table for any number of classes the "oracle" labels are applied. We define the "oracle" output (R_i) of the classifier C_i as:

$$R_i(\mathbf{x}_n) = \begin{cases} 1 \; if \; \hat{C}_i(\mathbf{x}_n) = y_n \\ 0 \; if \; \hat{C}_i(\mathbf{x}_n) \neq y_n \end{cases}. \tag{3}$$

In other words, the value of $R_i = 1$ means that the classifier C_i is correct, i.e. it recognizes the true class (y_n) of the example $\mathbf{x}_n$, and $R_i = 0$ means that the classifier is wrong.

Classifiers	$R_j = 1$	$R_j = 0$
$R_i = 1$	a	b
$R_i = 0$	c	d

Table 1. A 2×2 contingency table for the "oracle" outputs.

Kuncheva et al. (2000) proposed the Yule's Q statistics to evaluate the diversity of all possible component classifier pairs. The Yule's Q statistics is the original measure of dichotomous agreement, designed to be analogous to the Pearson's correlation:

$$Q_{ij} = \frac{ad - bc}{ad + bc}. \tag{4}$$

This measure is pairwise and symmetric, and varies between -1 and 1. A value of 0 indicates statistical independence of classifiers, positive values mean that the classifiers have recognized the same examples correctly and negative values – that the classifiers commit errors on different examples:

The Q statistics for an ensemble is calculated by averaging the values from all possible pairs of the ensemble members $(\bar{Q})$. Kuncheva et al. (2000) have reported that negatively related component classifier results in more accurate ensembles than the independent ones.

The most famous diversity measure is the binary version of the Pearson's correlation coefficient:

$$r_{ij} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}. \tag{5}$$

Its values range from -1 to 1 with 0 indicating independence of the two classifiers. It can be also shown that $|r_{ij}| \leq |Q_{ij}|$.

Giacinto and Roli (2001) have introduced a measure based on the compound error probability for the two classifiers and named *compound diversity*:

$$CD_{ij} = \frac{d}{a+b+c+d}. \tag{6}$$

This measure is also named "double-fault measure" because it is the proportion of the examples that have been misclassified by both classifiers.

Partridge and Yates (1996), and Margineantu and Diettrich (1997) have used a measure named *within-set generalization diversity*. This measure is simply the κ statistics developed by Fleiss (1981) as a *measure of interrater reliability*. It measures the level of agreement between two classifiers with the correction for chance. The pairwise version of the κ is calculated as:

$$\kappa_{ij} = \frac{2(ac-bd)}{(a+b)(c+d)+(a+c)(b+d)}. \tag{7}$$

Skalak (1996) has used the *disagreement measure* to characterize the diversity between base classifiers:

$$DM_{ij} = \frac{b+c}{a+b+c+d}. \tag{8}$$

This is the ratio between the number of examples on which one classifier is correct and the other is wrong to the total number of examples.

Several non-pairwise measures have also been developed. For example, Hansen and Salamon (1990) proposed the *measure of difficulty θ*. It is in fact the variance of a variable representing the proportion of classifiers that correctly classify an example $\mathbf{x}_n$ chosen at random.

Partridge and Krzanowski (1997) have introduced the *generalized diversity measure*, and Cunnigham and Carney (2000) suggested using the *entropy function* to assess the dependence between classifiers.

5 Hamann's measure

The Yule's statistics is very useful to assess the similarity between classifiers C_i and C_j, but in some cases its value may be undefined, e.g. when $a = 0$ and $b = 0$ etc. Moreover, the Q statistics cannot distinguish between different distributions of classifier outputs as shown in Table 2. For both tables, the Q value is the same and equals -1, indicating the strongest possible correlation, although the classifiers in the left table are not related at all.

To overcome the drawbacks of the Yule's Q statistics, we propose to measure the diversity between two classifiers using the Hamann's coefficient.

Hamann (1961) has proposed a binary similarity coefficient that is simply the difference between the matches and mismatches as a proportion of the total number of entries:

$$H_{ij} = \frac{(a+d)-(b+c)}{a+b+c+d}. \tag{9}$$

	$R_j = 1$	$R_j = 0$
$R_i = 1$	49	1
$R_i = 0$	50	0

	$R_j = 1$	$R_j = 0$
$R_i = 1$	1	49
$R_i = 0$	50	0

Table 2. Two distributions of classifier outputs for $K = 100$.

It ranges from -1 to 1. A value of 0 indicates an equal number of matches to mismatches, -1 represents perfect disagreement, and 1 – perfect agreement.

The Hamann's measure can distinguish between the two distributions observed in Table 2. For the right table, it has the value of $H_{ij} = -0.98$ that correctly indicates the strong negative dependence, and for the left one, the value of $H_{ij} = -0.02$ means no dependence between classifiers C_i and C_j.

From the probabilistic point of view, the Hamann's coefficient is the probability that a randomly chosen example will score the same on both classifiers minus the probability it will score differently.

For a classifier ensemble, the Hamann's coefficient is calculated as an average over all H_{ij} values:

$$\bar{H} = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} H_{ij}. \tag{10}$$

6 Experiments

In order to compare the performance of the Hamann's coefficient to the Yule's statistics we followed the synthetic experiment described by Kuncheva et al. (2000). We used a set of 10 examples ($N = 10$) and an ensemble of tree classifiers: C_1, C_2, C_3, each of them having the same classification error $e(C_i) = 0.4$ (6 out of 10 examples are recognized correctly). That gives the total number of 28 different classifier ensembles.

We have used the Hamann's coefficient to measure the diversity between the members of each ensemble (Table 3) and majority vote combining.

The Pearson correlation between the Hamann's and Yule's coefficients is 0.972 which indicates strong concordance of the two measures of diversity. Also, their correlations with the ensemble classification error are similar, i.e. it is 0.462 for the Hamann's coefficient, and 0.431 for the Yule's Q.

The diversity measure can also be used to guide the clustering process in the *cluster and select* approach to classifier combining. In this approach, a large set of candidate classifiers is divided into K disjoint subsets, i.e. classifiers with similar outputs are in the same cluster, and classifiers with different predicted class labels are assigned to different clusters. Then, a member of each cluster is selected, e.g. the one with the highest accuracy or the one that exhibits the maximum average distance from all other cluster centers. Finally, the selected classifiers are combined by majority voting.

Ensemble	H_{12}	H_{13}	H_{23}	H	Q	$e(D^*)$
1	0.2	-0.6	-0.6	-0.33	-0.56	0.2
2	-0.2	-0.6	-0.2	-0.33	-0.67	0.2
3	-0.2	-0.2	-0.2	-0.20	-0.50	0.1
4	0.6	-0.6	-0.6	-0.20	-0.37	0.3
5	0.2	-0.6	-0.2	-0.20	-0.39	0.3
6	-0.2	-0.2	-0.2	-0.20	-0.50	0.3
7	0.2	-0.2	-0.2	-0.07	-0.22	0.2
8	1.0	-0.6	-0.6	-0.07	-0.33	0.4
9	0.6	-0.6	-0.2	-0.07	-0.21	0.4
10	0.2	-0.6	0.2	-0.07	-0.11	0.4
11	0.2	-0.2	-0.2	-0.07	-0.22	0.4
12	0.6	-0.2	-0.2	0.07	-0.04	0.3
13	0.2	-0.2	0.2	0.07	0.05	0.3
14	0.2	0.2	0.2	0.20	0.33	0.2
15	0.6	-0.2	-0.2	0.07	-0.04	0.5
16	1.0	-0.2	-0.2	0.20	0.00	0.4
17	0.2	-0.2	0.2	0.07	0.05	0.5
18	0.6	-0.2	0.2	0.20	0.24	0.4
19	0.2	0.2	0.2	0.20	0.33	0.4
20	0.6	0.2	0.2	0.33	0.51	0.3
21	0.2	0.2	0.2	0.20	0.33	0.6
22	0.6	0.2	0.2	0.33	0.51	0.5
23	1.0	0.2	0.2	0.47	0.55	0.4
24	0.6	0.2	0.6	0.47	0.70	0.4
25	0.6	0.6	0.6	0.60	0.88	0.3
26	0.6	0.6	0.6	0.60	0.88	0.5
27	1.0	0.6	0.6	0.73	0.92	0.4
28	1.0	1.0	1.0	1.00	1.00	0.4

Table 3. The H and Q coefficients and errors for 28 classifier ensembles.

We have applied the *cluster and select* approach to 5 benchmark datasets from the Machine Learning Repository at the UCI (Blake et al. (1998)). For each dataset we formed 500 training subsets using the Random Subpace Method developed by Ho (1998), and trained 500 candidate classifiers[1]). Then they have been divided, using Yule's and Hamann's measures (as similarity measures), into $K = 30$ clusters. Next, for each cluster, one classifier with the highest accuracy has been selected to form the committee. Finally, the aggregated classifier has been built using the majority vote and its classification error estimated on appropriate test sets.

The preliminary results from this experiment have shown that the ensembles combined using the Hamann's measure are more accurate than those built with the Yule's measure of diversity (Table 4).

[1] We used classification trees grown with the *Rpart* procedure written by Therneau and Atkinson (1997) for the R environment.

Data set	Single tree (Rpart)	Yule's measure	Hamann's measure
DNA	6.40%	5.27%	5.11%
Letter	14.00%	10.89%	10.84%
Satellite	13.80%	11.07%	11.32%
Segmentation	3.70%	2.97%	2.41%
Soybean	8.00%	7.34%	6.89%

Table 4. Ensemble classification errors for two diversity measures.

7 Summary

In this paper we have proposed a new measure of diversity for classifier ensembles. It is based on the Hamann's similarity coefficient and is more flexible than the Yule's Q coefficient. It ranges from -1 to 1 and its value can always be determined.

Preliminary results of experiments based on the *cluster and select* approach suggested that Hamann's measure leads to more accurate aggregated models than those built with the Yule's Q.

References

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

BREIMAN, L. (1998): Arcing classifiers. *Annals of Statistics, 26, 801–849.*

BREIMAN, L. (2001): Random Forests. *Machine Learning 45, 5–32.*

CUNNIGHAM, P. and CARNEY, J. (2000): Diversity versus quality in classification ensembles based on feature selection. In: *Proceedings of European Conference on Machine Learning,* LNCS, vol. 1810, Springer, Berlin, 109–116.

DIETTERICH, T. and BAKIRI, G. (1995): Solving multiclass learning problem via error-correcting output codes. *Journal of Artificial Intelligence Research, 2, 263–286.*

FLEISS, J.L. (1981): *Statistical Methods for Rates and Proportions.* John Wiley and Sons, New York.

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.*

GIACINTO, G. and ROLI, F. (2001): Design of effective neural network ensembles for image classification processes. *Image Vision and Computing Journal, 19, 699–707.*

HAMANN, U. (1961): Merkmalsbestand und Verwandtschafsbeziehungen der farinosae. Ein Beitrag zum System der Monokotyledonen. *Willdenowia, 2, 639–768.*

HANSEN, L.K. and SALAMON, P. (1990): Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, 993–1001.*

HASHEM, S. (1999): Treating harmful collinearity in neural network ensembles. In: A.J. Sharkey (Ed.): *Combining Atrificial Neural Nets*, Springer-Verlag, London, 101–125.

HO, T.K. (1998): The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 832–844.*

KUNCHEVA, L., WHITAKER, C., SHIPP, D., and DUIN, R. (2000): Is independence good for combining classifiers. In: Proceedingd of the 15th International Conference on Pattern Recognition, Barcelona, Spain, 168–171.

KUNCHEVA, L. and WHITAKER, C. (2003): Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning, 51, 181–207.*

MARGINEANTU, M.M. and DIETTERICH, T.G. (1997): Pruning adaptive boosting. In: *Proceedings of the 14th International Conference on Machine Learning*, Morgan Kaufmann, San Mateo, 211–218.

MELVILLE and MOONEY (2003): Constructing diverse classifier ensembles using artificial training examples. In: *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, San Mateo, 505–510.

OZA, N.C. and TUMER, K. (1999): Dimensionality reduction through classifier ensembles. *Technical Report NASA-ARC-IS1999-126*, NASA Ames Labs.

PARTRIDGE, D. and KRZANOWSKI, W.J. (1997): Software diversity: practical statistics for its measurement and exploitation. *Information and Software Technology, 39, 707–717.*

PARTRIDGE, D. and YATES, W.B. (1996): Engineering multiversion neural-net systems. *Neural Computation, 8, 869–893.*

ROSEN, B.E. (1996): Ensemble learning using decorrelated neural networks. *Connection Science, 8, 373–383.*

SHARKEY, A. and SHARKEY, N. (1997): Diversity, selection, and ensembles of artificial neural nets. In: *Neural Networksand their applications, NEURAP-97, 205–212.*

SKALAK, D.B. (1996): The sources of increased accuracy for two proposed boosting algorithms. In: *Proceedings of the American Association for Artificial Intelligence AAAI-96*, Morgan Kaufmann, San Mateo.

THERNEAU, T.M. and ATKINSON, E.J. (1997): *An introduction to recursive partitioning using the RPART routines.* Mayo Foundation, Rochester.

TUMER, K. and GHOSH, J. (1996): Analysis of decision boundaries in linearly combined neural classifiers. *Pattern Recognition, 29, 341–348.*

WOLPERT, D. (1992): Stacked generalization. *Neural Networks 5, 241–259.*

ZENOBI, G. and CUNNINGHAM, P. (2001): Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error. *Lecture Notes in Computer Science 2167, 576–587.*

Repeated Confidence Intervals in Self–Organizing Studies

Joachim Hartung and Guido Knapp

Fachbereich Statistik,
Universität Dortmund, D-44221 Dortmund, Germany

Abstract. By using the generalized inverse chi–square method of transforming p–values in self–organizing studies as introduced by Hartung and Knapp (2003), one can construct a two–sided confidence interval for the parameter of interest after each realized sequential stage. Provided that in each sequential stage an unbiased estimate of the parameter of interest is available, the repeated confidence intervals always hold the nominal confidence level even in case of early stopping the trial. Moreover, the repeated confidence intervals are always nested, that is, an interval calculated after a performed stage is completely included in all the previous calculated intervals. The use of p–values in the present approach allows the calculation of confidence intervals for various parameters given a suitable test statistic and an unbiased estimate in each sequential stage.

1 Introduction

In the self–designing approach of group sequential trials as introduced by Fisher (1998), one decides adaptively after each interim–analysis during the course of the study whether exactly one or at least two further study stages will be performed by use of the unblinded results of all the already conducted interim–analyses. Consequently, the number of stages that will be performed is not fixed in advance. The self–designing trial ends when the (finite) variance of an a priori fixed final test statistic is used up. The sample sizes of the various stages can also be chosen adaptively. For normally distributed variables with known variances, Shen and Fisher (1999) present the explicit choice of the sequential weights but they fix a priori a sequence of sample sizes. By using p–values of the sequential test statistics, the distributional restriction on the response variable can be lifted, see Hartung (2001).

Hartung (2001) derives completely self–designing rules where the p–values of the sequential test statistics are combined using the weighted inverse normal method. In Hartung and Knapp (2003), completely self–designing algorithms are presented using the generalized inverse chi–square method for the combination of the sequential results so that early stopping in favor of the alternative is possible. In both approaches, simultaneously the weights and the sample sizes can be chosen adaptively so that the trials are becoming self–organizing studies based on fully automatic learning algorithms, cf. Hartung (2001), Hartung and Knapp (2003). Now, we will derive confidence intervals for the parameter of interest in the context of such self–organizing studies.

In classical group sequential trials, the repeated confidence interval approach by Jennison and Turnbull (1989) may be applied. This approach is also considered in Lehmacher and Wassmer (1999) where the authors consider an adaptive version of a Pocock (1977) group sequential trial procedure with a fixed number of stages. Recently, Brannath et al. (2003) derive improved repeated one–sided confidence bounds in trials with a maximal goal and a prefixed finite number of sequential stages.

The outline of the present paper is as follows: In Section 2, the method of combining independent test statistics using the generalized inverse chi–square method will be described. Section 3 contains essential ideas of adaptive designing already introduced in Hartung and Knapp (2003). In Section 4, we will present the nested repeated confidence intervals for the adaptive designs. The proposed repeated confidence intervals will be demonstrated by a real data example in Section 5. In Section 6, the results of a simulation study will be presented concerning the actual confidence coefficients of the final confidence interval. At the end, some final remarks are given where also the construction of point estimates is briefly addressed in the self–designing setting.

2 Generalized inverse chi–square method

For some real valued parameter θ, we consider the one–sided test problem

$$\mathrm{H}_0 : \theta = 0 \qquad \text{versus} \qquad \mathrm{H}_1 : \theta > 0\,. \tag{1}$$

We assume that the study is formally divided into K several disjoint study parts denoted by $stp(k)$, $k = 1, \ldots, K$. In $stp(k)$, let T_k be a test statistic for testing H_0 versus H_1, and large values of T_k lead to a rejection of H_0. We assume that, under H_0, the test statistic T_k has a continuous distribution function, say $F_{k,0}$. Then, the p–value $p_k = 1 - F_{k,0}(T_k)$, $k = 1, \ldots, K$, is uniformly distributed on the interval $(0, 1)$ under H_0.

Let G_{ν_k} denote the distribution function of a χ^2–variate with ν_k degrees of freedom and $G_{\nu_k}^{-1}$ the corresponding inverse, then, under H_0, the transformed p–value $q_k(\nu_k) = G_{\nu_k}^{-1}(1 - p_k)$, $k = 1, \ldots, K$, is distributed as a χ^2–variate with ν_k degrees of freedom.

Let us consider a sequence of degrees of freedom, say $\nu_1, \ldots, \nu_K$, which are fixed in advance. The disjoint study parts are assumed to be independent. So, after each $stp(k)$, $k = 1, \ldots, K$, we can calculate the combination statistic $S_k = \sum_{j=1}^{k} q_j(\nu_j)$. Under H_0, the statistic S_k is distributed as a χ^2–variate with $\nu_\Sigma(k) = \sum_{j=1}^{k} \nu_j$ degrees of freedom, $k = 1, \ldots, K$. So, in particular, when $S_{k^*} \geq \chi^2_{\nu_G\,;\,1-\alpha}$ for some $k^* \in \{1, \ldots, K\}$ and $\nu_G \geq \nu_\Sigma(K)$, the null hypothesis H_0 is rejected at level α since $S_K \geq S_{k^*}$, where $\chi^2_{\nu_G\,;\,1-\alpha}$ denotes the $(1 - \alpha)$–quantile of the χ^2–variate with ν_G degrees of freedom.

In the general context of combining p–values, the combination procedure using the test statistics S_k is asymptotically optimal in the sense of Bahadur efficiency, see Berk and Cohen (1979). This means that the combination test S_k gives the 'most significant' results in large samples. Furthermore, the combination statistic S_k is neither insensitive nor fragile for all $\alpha \in (0,1)$ as characterized by Marden (1991).

3 Adaptive designing

In this section, we put together the essential characteristics of the adaptive designing which have been presented in more detail in Hartung and Knapp (2003). At the onset of the study, we define the total available degrees of freedom, say $\nu_\Sigma(K) = \sum_{j=1}^{K} \nu_j$, and K is the maximum number of study parts. The choice of $\nu_\Sigma(K)$ determines the global critical value as $cv(\alpha) = \chi^2_{\nu_\Sigma(K);1-\alpha}$. Further, we define a minimum number of degrees of freedom, say ν_{min}, that will be assigned at least to each realized study part. For ease of presentation, we put $\nu_{min} = 1$ and $\nu_\Sigma(K) = K$ in the following. So, the global critical value is $cv(\alpha) = \chi^2_{K;1-\alpha}$.

In the first step, we divide the available degrees of freedom, K, into $1 \leq \nu_1 < K$ and $\nu_2^* = K - \nu_1 \geq 1$ degrees of freedom so that $q_1(\nu_1) + q_2(\nu_2^*)$ is χ^2–distributed with K degrees of freedom under H_0. After the first stage of the trial, $stp(1)$, we obtain $q_1(\nu_1)$ with the a priori associated non–random degrees of freedom ν_1. If $S_1 = q_1(\nu_1) \geq cv(\alpha)$, the hypothesis H_0 will be rejected at level α and the trial stops. Otherwise, the trial will continue.

When the non–random degrees of freedom ν_1 in the first stage are larger than $K - 2$, then it holds $\nu_2^* < 2$. This implies that all the still available degrees of freedom ν_2^* have to be used in the next study part because of $\nu_{min} = 1$, and the trial will stop definitively after the second stage. If $\nu_2^* \geq 2$, we can divide the a priori fixed value ν_2^* into two parts, say $\nu_2 \geq 1$ and $\nu_3^* = \nu_2^* - \nu_2$, so that $q_1(\nu_1) + \{q_2(\nu_2) + q_3(\nu_3^*)\}$ is χ^2–distributed with K degrees of freedom under H_0. With ν_2 degrees of freedom assigned to $stp(2)$, we observe $S_2 = q_1(\nu_1) + q_2(\nu_2)$. Then the trial will stop if S_2 is not less than $cv(\alpha)$, otherwise the next stage, $stp(3)$, will be performed.

If $\nu_3^* < 2$, the study part $stp(3)$ will be the last one in the trial. Otherwise, we can again divide the available degrees of freedom into two parts and assign the weight ν_3, $\nu_3 < \nu_3^*$, to $stp(3)$, and so on.

Up to stage k, that is, after $stp(k-1)$, we attain the following combination statistic

$$\sum_{j=1}^{k-1} q_j(\nu_j) + q_k\left(K - \sum_{j=1}^{k-1} \nu_j\right)$$

which is, under H_0, a χ^2–variate with K degrees of freedom. Note that $\nu_j \geq 1$, $j = 1, \ldots, k-1$, and $K - \sum_{j=1}^{k-1} \nu_j \geq 1$. The degrees of freedom ν_j are chosen based on the knowledge of the previous study parts, $stp(1)$ up to $stp(j-1)$. Let $stp(0)$ denote a priori information, this choice can be expressed as

$$\nu_j = \hat{\nu}\{j-1\} = \hat{\nu}\{stp(0), \ldots, stp(j-1)\}\,.$$

The procedure will stop after the study part $stp(k^*)$ in any case, if we decide to take all of the still available degrees of freedom $\nu_{k^*} = K - \sum_{j=1}^{k^*-1} \nu_j$ before the beginning of $stp(k^*)$.

The trial will be stopped early at stage k^*, that is, for $\nu_\Sigma(k^*) < K$, with a rejection of the null hypothesis, if $S_{k^*} = \sum_{j=1}^{k^*} q_j(\nu_j) \geq cv(\alpha)$.

We start with an a priori fixed sample size n_1 in $stp(1)$, but then the sample sizes n_j of the further study parts can be determined upon the knowledge of the previous study parts. Under the null hypothesis, the distribution of the p–values p_j and the independence of p_{j_1} and p_{j_2}, $j_1 \neq j_2$, still hold provided that the distribution of the test statistics under H_0 is continuous. Thus, the procedure remains valid when we allow:

$$n_j = \hat{n}\{j-1\} = \hat{n}\{stp(0), \ldots, stp(j-1)\}\,.$$

When the distribution of the test statistics under H_0 is not continuous, the independence of p_{j_1} and p_{j_2} does not necessarily hold. In this case, the distribution of p_1 and the conditional distributions of p_j given $(p_1, \ldots, p_{j-1})$ have to be stochastically larger than the uniform distribution to ensure that the combination test does not exceed the pre–chosen type I error rate; for a detailed discussion on this topic, see Brannath et al. (2002).

For explicit learning algorithms to determine automatically the sample sizes and the degrees of freedom in the adaptive designing, let us refer to Hartung and Knapp (2003).

4 Repeated confidence intervals

We assume that each $stp(k)$ provides an estimator of the parameter of interest, say $\hat{\theta}_k$, which should be unbiased for θ. The test statistic T_k depends on $\hat{\theta}_k$ in each $stp(k)$, that is, $T_k = T_k(\hat{\theta}_k)$, and T_k is a (strictly) monotone increasing function in $\hat{\theta}_k$. Let us now consider the function $T_k(\hat{\theta}_k - \theta)$ which is monotone decreasing in θ. Based on the statistics $T_k(\hat{\theta}_k - \theta)$ from each $stp(k)$, we will deduce confidence intervals for θ.

By using the generalized inverse chi–square method in the adaptive designing described in Section 3, we can test the null hypothesis H_0 after each performed stage of the trial. So, we are also able to construct a (two–sided) confidence interval for θ after each performed stage.

For each realized stage $stp(j)$, $j = 1, 2, \ldots$, let us first consider the function

$$Q_{L,j}(\theta) = \sum_{i=1}^{j} G_{\nu_i}^{-1}\left[F_{i,0}(T_i(\hat{\theta}_i - \theta))\right]. \tag{2}$$

All the involved functions on the right hand side of (2) are increasing functions in their arguments. Consequently, the function $Q_{L,j}(\theta)$ is a monotone decreasing function in θ. Furthermore, when the two functions $Q_{L,j}(\theta(j))$

and $Q_{L,j^*}(\theta(j^*))$ provide the same value with $j < j^*$, that is, $Q_{L,j}(\theta(j)) = Q_{L,j^*}(\theta(j^*))$, then it holds $\theta(j) < \theta(j^*)$.

If all the available degrees of freedom ν_G are used up in the last realized stage, say *stp*(k^*), then, for the true value θ,

$$Q_{L,k^*}(\theta) = \sum_{i=1}^{k^*} G_{\nu_i}^{-1}\left[F_{i,0}(T_i(\hat{\theta}_i - \theta))\right] \tag{3}$$

will be a χ^2–variate with ν_G degrees of freedom.

As we would like to construct a two–sided confidence interval for θ, we have to consider the two–sided test problem, $H_0 : \theta = 0$ versus $H_1^* : \theta \neq 0$. The two–sided test problem may be checked by two appropriate one–sided tests in the adaptive designing.

When we consider the one–sided test problem, $H_0 : \theta = 0$ versus $H_1^\dagger : \theta < 0$, then small values of the test statistics T_i may support $H_1^\dagger$. Consequently, we can build the following statistic in analogy to (2)

$$Q_{U,j}(\theta) = \sum_{i=1}^{j} G_{\nu_i}^{-1}\left[1 - F_{i,0}(T_i(\hat{\theta}_i - \theta))\right]. \tag{4}$$

By using similar arguments as for (2), the function $Q_{U,j}(\theta)$ is monotone increasing in θ as $1 - F_{i,0}(\cdot)$ is a decreasing function in its argument $(\cdot)$. Furthermore, when the two functions $Q_{U,j}(\theta(j))$ and $Q_{U,j^*}(\theta(j^*))$ provide the same value with $j < j^*$, that is, $Q_{U,j}(\theta(j)) = Q_{U,j^*}(\theta(j^*))$, then it holds now $\theta(j) > \theta(j^*)$.

Again, if all the available degrees of freedom ν_G are used up in the last realized stage, say *stp*(k^*), then, for the true value θ,

$$Q_{U,k^*}(\theta) = \sum_{i=1}^{k^*} G_{\nu_i}^{-1}\left[1 - F_{i,0}(T_i(\hat{\theta}_i - \theta))\right] \tag{5}$$

will be a χ^2–variate with ν_G degrees of freedom.

Based on these considerations the following assertions can be easily proven.

Theorem

(i) For each realized stage *stp*(j) of the trial, a confidence interval on θ with a confidence coefficient of at least $(1 - 2\alpha)$, $0 < \alpha < 0.5$, is given as

$$CI_j = [\,\theta_{Lj}\ ;\ \theta_{Uj}\,]\,,$$

where the bounds θ_{Lj} and θ_{Uj} are the solutions of the equations

$$Q_{L,j}(\theta_{Lj}) = \chi^2_{\nu_G\,;\,1-\alpha} \qquad \text{and} \qquad Q_{U,j}(\theta_{Uj}) = \chi^2_{\nu_G\,;\,1-\alpha}.$$

(ii) The confidence interval CI_j will be an exact $(1-2\alpha)$–confidence interval if all the available degrees of freedom are used up in $stp(j)$, that is, if $\nu_j = \nu_G - \sum_{i=1}^{j-1} \nu_i$.

(iii) The confidence intervals are nested, that is, it holds

$$CI_1 \supset CI_2 \supset \cdots \supset CI_{k^*}, \tag{6}$$

where k^* denotes the last realized stage of the trial.

5 A real data example

The real data example is taken from Lehmacher and Wassmer (1999). In a randomized, placebo–controlled, double–blind study involving patients with acne papulo pustulosa (Plewig's grade II–III), the effect of treatment under a combination of 1% chloramphenicol (CAS 56–75–7) and 0.5% pale sulfonated shale oil versus the alcoholic vehicle (placebo) was investigated. After 6 weeks of treatment, the reduction of bacteria from baseline, examined on agar plates (log CFU/cm^2; CFU: colony forming units), was assessed for the active group as well as for the placebo group. By using the study description and the parameter estimates from Lehmacher and Wassmer (1999), Hartung and Knapp (2003) illustrate their self–designing algorithm with one–sided t–tests on each realized stage. They use $\chi^2_{10,0.99} = 23.21$ as global critical value. The test results are put together in Table 1, where $\hat{\theta}_k$ is the observed difference between the active group and the placebo group in stage k and $\hat{\sigma}_k$ the corresponding standard error. The trial stops with the rejection of the null hypothesis after stage 2. Note that the total available degrees of freedom, $\nu_G = K = 10$, are not used up because of the early stop of the trial.

stage k	sample size per group $n_{1k} = n_{2k}$	treatment difference $\hat{\theta}_k$	standard error $\hat{\sigma}_k$	p–value p_k	degrees of freedom ν_k	combined statistic S_k
1	12	1.549	0.537	0.0043	5	17.106
2	6	1.580	0.850	0.0463	4	26.777

Table 1. Results for a clinical trial with normal response

In both stages, the estimated treatment difference $\hat{\theta}_k$ is unbiased for the true treatment difference θ and the required test statistic is

$$T_k(\hat{\theta}_k - \theta) = \frac{(\hat{\theta}_k - \theta)}{\hat{\sigma}_k}$$

which is t–distributed with $n_{1k} + n_{2k} - 2$ degrees of freedom.

Let F_{t_j} denote the cumulative distribution function of a t–variate with $n_{1j} + n_{2j} - 2$ degrees of freedom in $stp(j)$. Then, the two functions needed to determine the bounds of the confidence intervals after $stp(j)$, $j = 1, 2$, can be constructed using (2) and (4), respectively, where the required degrees of freedom ν_i, $i = 1, 2$, are given in Table I, and $F_{j,0} = F_{t_j}$.

By solving the equations $Q_{L,j}(\theta) = \chi^2_{10,0.99}$ and $Q_{U,j}(\theta) = \chi^2_{10,0.99}$ for θ, $j = 1, 2$, the resulting two–sided nested repeated confidence intervals are $CI_1 : [\,-0.59446\,;\,3.69246\,]$ and $CI_2 : [\,0.24464\,;\,2.87300\,]$ and both intervals have a confidence coefficient of at least $1 - 2\alpha = 0.98$.

6 Simulation results

In this section, we briefly report some simulation results on the actual confidence coefficients of the final repeated confidence interval according to (6). We strictly follow the adaptive designing rules used in the simulation study in Hartung and Knapp (2003) for the comparison of two normal variates with known variances equal to one.

Hartung and Knapp (2003) consider two explicit strategies of their learning algorithm with different external restrictions on the sample sizes. Here, we consider only one strategy. This strategy has the following parameters: $\nu_\Sigma(K) = 10$, $\nu_1 = 2$, $n_1 = 34$, and $n_{min} = 18$. As an upper bound for the sample size in each stage, we use $n_{max} = 138$. The resulting test attains an empirical power of 90% at $\theta = 0.5$, where θ denotes the difference of the two involved means.

In the small simulation study, we estimate the actual confidence coefficients of the final nested confidence interval for different true values of θ, namely $\theta = 0$, 0.025, 0.050, 0.075, …, 0.75. Each estimated confidence coefficient is based on 1.000.000 simulation runs where the nominal confidence coefficient is chosen as $(1 - 2\alpha)\,100\% = 95\%$.

In the chosen strategy, the estimated confidence coefficient starts at 95% for $\theta = 0$ and then monotonously increases for growing θ. The increase is nearly linear over the range of the considered values of θ. Still at $\theta = 0.5$ the estimated confidence coefficients are smaller than 96%. The maximal estimated confidence coefficient is 96.6% at $\theta = 0.75$.

The conservatism of the final confidence intervals is a consequence of the possibility of early rejecting the null hypothesis, that is, before all the total available degrees of freedom are used up the trial stops in favor of the alternative, and the chance for an early stop increases with θ.

7 Final remarks

Hartung (2001) considers self–designing rules where the inverse normal method is used for the transformation of the p–values. In that approach, no early stopping in favor of the alternative is possible and the significance test will be only carried out at the end of the trial if the total available weights are used up. Confidence intervals can be determined in an analogue manner in that approach as described in this paper. But only the final confidence interval always holds the prefixed confidence level. As the total available weights are definitely used up at the end of the trial, the final confidence interval is always an exact $(1-2\alpha)$–interval.

Finally, we would like to make some remarks on point estimates in the present context of self–organizing trials with the generalized inverse chi–square method. The construction of the repeated confidence intervals may lead to certain proposals of determination of a point estimate which may be valid for various parameters of interest. One proposal may be to use the centre of the final confidence interval. Another proposal would be to use the statistics (2) and (4). When the total available degrees of freedom are used up then the statistics (3) and (5) are χ^2–distributed with ν_G degrees of freedom. By equating (3) and (5) to the mode of a χ^2–variate with ν_G degrees of freedom, that is, to $\nu_G - 2$ and by solving these equations for θ would yield two, probably different, maximum likelihood estimates. The arithmetic mean of these two estimates could be used as a unique estimate, also in case of early stopping.

References

BERK, R.H. and COHEN, A. (1979): Asymptotically Optimal Methods of Combining Tests. *Journal of the American Statistical Association, 74, 812–814.*

BRANNATH, W., KÖNIG, F., and BAUER, P. (2003): Improved Repeated Confidence Bounds in Trials with a Maximal Goal. *Biometrical Journal, 45, 311–324.*

BRANNATH, W., POSCH, M., and BAUER, P. (2002): Recursive Combination Tests. *Journal of the American Statistical Association, 97, 236–244.*

FISHER, L. (1998): Self–Designing Clinical Trials. *Statistics in Medicine, 17, 1551–1562.*

HARTUNG, J. (2001): A Self–Designing Rule for Clinical Trials with Arbitrary Response Variables. *Controlled Clinical Trials, 22, 111–116.*

HARTUNG, J. and KNAPP, G. (2003): A New Class of Completely Self–Designing Clinical Trials. *Biometrical Journal, 45, 3–19.*

JENNISON, C. and TURNBULL, B. (1989): Interim analysis: The repeated confidence interval approach. *Journal of the Royal Statistical Society, Series B, 51, 305–361 (with discussion).*

LEHMACHER, W. and WASSMER, G. (1999): Adaptive Sample Size Calculations in Group Sequential Trials. *Biometrics, 55, 1286–1290.*

MARDEN, J.C. (1991): Sensitive and Sturdy p–Values. *The Annals of Statistics, 19, 918–934.*

POCOCK, S.J. (1977): Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika, 64, 191–199.*

SHEN, Y. and FISHER, L. (1999): Statistical Inference for Self–Designing Clinical Trials with a One–Sided Hypothesis. *Biometrics, 55, 190–197.*

Fuzzy and Crisp Mahalanobis Fixed Point Clusters

Christian Hennig[1,2]

[1] Fachbereich Mathematik - SPST,
Universität Hamburg, D-20146 Hamburg, Germany,
[2] Department of Statistical Science,
University College London, London WC1e 6BT, United Kingdom

Abstract. Fixed point clusters (FPCs) are based on the idea that local optima of redescending M-estimators can be used to locate clusters. FPCs satisfy a fixed point condition which means that they are data subsets that do not contain any outlier and with respect to that all other points in the data set are outliers. In this paper, outliers are defined in terms of the Mahalanobis distance. Crisp FPCs (where outlyingness is defined with a 0-1 weight function) are compared to fuzzy FPCs where outliers are smoothly downweighted. An algorithm to find substantial crisp and fuzzy FPCs is proposed, the results of a simulation study and a data example are discussed.

1 Introduction

Fixed point clusters (FPCs) are based on the idea that local optima of redescending M-estimators can be used to locate clusters. In this paper, FPCs for p-dimensional data based on the Mahalanobis distance are treated. Assume $p = 1$ for a moment to illustrate the principle. An M-estimator T_M for location in a real-valued data set $\mathbf{x} = (x_1, \ldots, x_n)$ is defined by

$$\sum_{i=1}^{n} \rho\left(\frac{x_i - T_M}{S}\right) \overset{!}{=} \min \text{ or } \sum_{i=1}^{n} \psi\left(\frac{x_i - T_M}{S}\right) \overset{!}{=} 0, \tag{1}$$

where S is a scale estimator, ρ is a positive function with $\rho(0) = 0$ (usually nondecreasing in $|x|$) and $\psi = \rho'$ (possibly piecewise). It can be shown easily (Huber (1981), p. 183 ff.) that the second equation in (1) is equivalent to

$$T_M \overset{!}{=} \frac{\sum_{i=1}^{n} \omega\left(\frac{(x_i - T_M)^2}{S^2}\right) x_i}{\sum_{i=1}^{n} \omega\left(\frac{(x_i - T_M)^2}{S^2}\right)} \text{ with } \omega(y) = \frac{\psi(\sqrt{y})}{\sqrt{y}}, \tag{2}$$

i.e., T_M is a weighted mean of the observations with weights depending on T_M itself. An M-estimator is called "redescending" if $\psi(x) = 0$ for x outside some interval $[-c, c]$. The interesting feature for cluster analysis is the invariance of (2) against changes in the data in the areas where $\omega = 0$. If T_M solves (2) and the points with $\omega > 0$ are well separated, i.e., they form a cluster, the solution is independent of the structure of the data far away from this cluster. Thuss, clusters of a certain shape (compatible with ω) can be found if there are similar clusters elsewhere in the data (in which case (2) will have multiple solutions), but also if the rest of the data consists of subsets with different shapes, is heterogeneous or contains extreme outliers.

The scale estimator S is often chosen as a preliminary robust estimator such as the MAD, but this is not adequate for cluster analysis, because such an estimator depends always on more than 50% of the data points and cannot yield a valid estimate for the scale of smaller clusters. The use of multiple local solutions for cluster analysis requires a scale estimation that adapts itself to the local cluster. $w_i = \omega\left(\frac{(x_i - T_M)^2}{S^2}\right)$ can then be interpreted as a cluster membership indicator. The most natural way to define a cluster-adapted scale is

$$S^2 \overset{!}{=} \frac{\sum_{i=1}^n \omega\left(\frac{(x_i-T_M)^2}{S^2}\right)(x_i - T_M)^2}{\sum_{i=1}^n \omega\left(\frac{(x_i-T_M)^2}{S^2}\right)}. \tag{3}$$

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a data set consisting of p-dimensional points. With $n(\mathbf{w}) = \sum_{i=1}^n w_i$, $T(\mathbf{w}) = \frac{1}{n(\mathbf{w})}\sum_{i=1}^n w_i\mathbf{x}_i$ and $\mathbf{S}(\mathbf{w}) = \frac{1}{n(\mathbf{w})}\sum_{i=1}^n w_i(\mathbf{x}_i - T(\mathbf{w}))(\mathbf{x}_i - T(\mathbf{w}))'$ for any weight vector $\mathbf{w} = (w_1, \ldots, w_n) \in [0,1]^n$, the fixed point condition (2)/(3) can be written as a fixed point condition for the cluster indicator:

Definition 1. An indicator vector $\mathbf{w} \in [0,1]^n$ is called **Mahalanobis FPC indicator** w.r.t. a decreasing weight function ω with $\omega(0) = 1$, $y \geq c \Rightarrow \omega(y) = 0$ for some $\infty > c > 0$, if for $i = 1, \ldots, n$:

$$\mathbf{w} = \left(\omega\left[(\mathbf{x}_i - T(\mathbf{w}))'\mathbf{S}(\mathbf{w})^{-1}(\mathbf{x}_i - T(\mathbf{w}))\right]\right)_{i=1,\ldots,n}. \tag{4}$$

$\frac{(x_i-T_M)^2}{S^2}$ has been replaced by the Mahalanobis distance $(\mathbf{x}_i - T)'\mathbf{S}^{-1}(\mathbf{x}_i - T)$ for $p > 1$. The connection between weight functions and ψ-functions for multivariate M-estimators of location and scatter is a bit more complicated than for $p = 1$, but leads to an analogous fixed point equation, see Hampel et al. (1986, p. 289). The resulting estimators are affine equivariant and the cluster indicators are invariant under linear transformations (Hampel et al. (1986, p. 283)).

FPC analysis has been introduced for linear regression data by Hennig (1997, 2002, 2003). Mahalanobis FPCs have been used by Hennig and Christlieb (2002) to validate the findings of a graphical cluster search. This idea of using local optima of redescending M-estimators for clustering goes back to Hampel (1975) and has become somewhat popular in recent literature on unsupervised pattern recognition, see, e.g., Müller and Garlipp (2005).

The FPC concept is a local cluster concept. The FPC property of a data subset depends on its shape and on its separateness, but not on data points far away from it. FPCs may overlap, and a fixed point clustering does not need to be exhaustive. It may be interesting to compare the FPC concept with more classical concepts of overlapping clustering, which are a research topic of Professor Gaul (e.g., Gaul and Schader (1994), Baier, Gaul and Schader (1997)).

The present paper focuses on two aspects:

- The choice of the weight function (especially if it should be chosen crisp or fuzzy) is discussed in Section 2.
- An algorithm is proposed to find all relevant Mahalanobis FPCs in a data set (Section 3).

In Section 4, the results of a simulation study are given where crisp FPCs have been compared with fuzzy FPCs. In Section 5, the methods are applied to a real data set.

2 Crisp and fuzzy weight functions

In previous papers (Hennig (1997, 2002, 2003), Hennig and Christlieb (2002)), the weight function ω has been chosen as 0-1-valued:

$$\omega_\alpha(y) = 1(y \leq \chi^2_{p;1-\alpha}), \tag{5}$$

$\chi^2_{p;1-\alpha}$ denoting the $1-\alpha$-quantile of the χ^2-distribution with p degrees of freedom. This corresponds to the definition of an α-outlier region w.r.t. the $\mathcal{N}_p(\mathbf{a}, \mathbf{\Sigma})$ (normal)-distribution by Becker and Gather (1999). α-outliers w.r.t. some distribution are defined as points lying in low density regions so that the probability of a point to be an outlier under the distribution is not larger than a small α:

$$A_\alpha = \{\mathbf{x} : \ (\mathbf{x}-\mathbf{a})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{a}) > \chi^2_{p;1-\alpha}\}.$$

If ML estimators are plugged into this definition, crisp FPCs are exactly the data subsets that do not contain any outlier and with respect to that all other points in the data set are outliers. This formalizes two properties necessary to call a data subset a "cluster", namely homogeneity (no outliers included) and separation (all other points are outliers). In terms of statistical models, an FPC can then be interpreted as a central part of a well separated normal subpopulation, which can be said to be the "cluster prototype model" for this method. The corresponding estimators work well for more general subpopulations that are at least approximately symmetric and unimodal. Differences in tail areas do not matter because of $\omega = 0$.

The choice (5) for the (1-dimensional) location estimator T_M, corresponds to the skipped mean, which is among the optimal B-robust redescending M-estimators (minimum variance in the normal model under a bound on the gross error sensitivity, Hampel et al. (1986), pp. 87, 158). Nevertheless, it

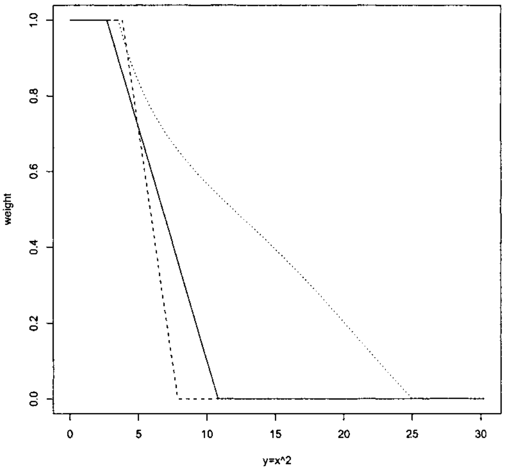


Fig. 1. Weight functions $\omega_{1,0.9,0.999}$ (solid), $\omega_{1,0.95,0.995}$ (dashed) and corresponding to a tanh-estimator (dotted) recommended by Hampel et al. (1986).

has been demonstrated by theoretical considerations and extensive simulations (Andrews et al. (1972), Hampel et al. (1986), p. 166, Hampel (2002)) that smooth rejection of outliers as done with weight functions that decrease continuously to 0 leads to more stable and efficient location estimators.

FPC analysis may therefore be improved by replacing the crisp weight function ω_α by a continuously decreasing weight function. This is recommended by Hampel (2002). Nevertheless, the situation is more complicated in cluster analysis than for the estimation of the parameters of a homogeneous majority of the data. Good choices of ω for the estimation problem of a single population have a wide region of decrease of the weight function to include information not only from central points, but also from points in a more or less large distance from the center. The dotted line in Figure 1 is the weight function corresponding to one of the recommended ("optimal V-robust") M-estimators in Hampel et al. (1986). But if the center a_1 of a cluster should be estimated in the presence of another cluster with a center a_2 with, say, $|a_1 - a_2|^2 = 25S^2$, S^2 being the within-cluster variance for both clusters, about half of the points of the second cluster affect the estimation of the first one if the shown tanh-estimator is used, and the two clusters are not even separable by the fixed point condition (4). Therefore, the separation of clusters that are not extremely far away from each other requires much faster decreasing weight functions. Weight functions of the form

$$\omega_{p,\beta,\gamma}(y) = \begin{cases} 1 & : y < c_1 = \chi^2_{p;1-\beta} \\ 0 & : y > c_2 = \chi^2_{p;1-\gamma} \\ \frac{c_2 - y}{c_2 - c_1} & : \quad c_1 \leq y \leq c_2 \end{cases} \tag{6}$$

are a simple alternative. $\omega_{1,0.9,0.999}$ and $\omega_{1,0.95,0.995}$ are shown in Figure 1. The simulations summarized in Section 4 show that these are reasonable choices, and that the decrease of $\omega_{1,0.9,0.999}$, though much faster than that of the tanh-estimator, is still a bit too slow to separate close, but clearly visible clusters.

3 An algorithm for Mahalanobis fixed point clusters

It is impossible to check (4) for all possible $\mathbf{w}$. But FPCs can be found by a fixed point algorithm defined by

$$\mathbf{w}^{k+1} = \left(\omega\left[(\mathbf{x}_i - T(\mathbf{w}^k))'\mathbf{S}(\mathbf{w}^k)^{-1}(\mathbf{x}_i - T(\mathbf{w}^k))\right]\right)_{i=1,\dots,n}. \tag{7}$$

This algorithm is shown to converge toward an FPC in a finite number of steps for crisp Mahalanobis FPCs in Hennig and Christlieb (2002). Up to now, no such proof exists for fuzzy FPCs. However, using a weight function of the form (6), I have carried out about a million of runs of this algorithm on real and simulated data sets, and I have never seen any convergence problems, though I know of artificial counterexamples with a particular different continuous weight function. The remainder of this section applies to crisp as well as to fuzzy Mahalanobis FPCs.

The main problem is the choice of reasonable starting configurations $\mathbf{w}^0$. While there are many very small FPCs, which are not very meaningful (e.g., all sets of p or fewer points yield FPCs), FPC analysis aims at finding all substantial FPCs, where "substantial" means all FPCs corresponding to well separated, not too small data subsets which give rise to an adequate description of the data as a central part of a normal (or at least symmetric and unimodal) distribution. For clusterwise regression, this problem is discussed in depth in Hennig (2002). For Mahalanobis FPCs, the following strategy is suggested.

For every point of the dataset, one initial configuration is chosen, so that there are n runs of the algorithm (7). Initial configurations are always chosen as crisp indicators. For every point, the p nearest points in terms of the Mahalanobis distance w.r.t. $\mathbf{S}(1,\dots,1)$ are added, so that there are $p+1$ points. Because such configurations often lead to too small clusters, the initial configuration is enlarged to contain n_{start} points. To obtain the $(p+2)$nd to the n_{start}th point, the covariance matrix of the current configuration is computed (new for every added point) and the nearest point in terms of the new Mahalanobis distance is added.

The stability or "significance" of an FPC $\mathbf{w}$ can be measured by $r(\mathbf{w}) = t(\mathbf{w})/n(\mathbf{w})$, where $t(\mathbf{w})$ is the number of times this FPC has been found with the algorithm described above (for fuzzy FPCs, there has to be some numerical convergence criterion and an FPC can only be reproduced up to a prespecified accuracy). Usually some of the FPCs are unstable, and more than one FPC may correspond to the same significant pattern in the data. Therefore the number of FPCs is reduced. A similarity matrix is computed between FPCs. Similarity between two FPC indicators $\mathbf{v}, \mathbf{w}$ is defined as follows:

$$s(\mathbf{v}, \mathbf{w}) = \frac{n(i(\mathbf{v}, \mathbf{w}))}{n(u(\mathbf{v}, \mathbf{w}))},$$

where $i(\mathbf{v}, \mathbf{w})$ denotes the indicator of the (fuzzy) intersection (elementwise minimum) and $u(\mathbf{v}, \mathbf{w})$ denotes the indicator of the (fuzzy) union (elementwise maximum). In the literature, s is often called "Jaccard coefficient". Then, the single linkage groups of FPCs of level s_0 are computed, i.e. the connectivity components of the graph where edges are drawn between FPCs with similarity larger than s_0. This is the most easily computable clustering based on similarities. Because groups of FPCs usually seem to be well separated in terms of s, the choice of the clustering method is not too crucial for this task. Groups of FPCs whose members are found often enough (which means that the sum of the $r(\mathbf{w})$ of the members of the group is larger than some cutoff value r_0) are considered as "stable". A representative FPC is chosen for every stable group of FPCs by maximizing $r(\mathbf{w})$.

Some tuning constants have to be chosen for this algorithm: α, β and γ for the weight functions, n_{start}, s_0, and r_0. Suggestions for these constants have been assessed by the simulation study that is described in the next section. The algorithm is implemented in the add-on package "fpc" for the statistical software system R (`www.R-project.org`).

4 Simulation results

A simulation study has been carried out to compare crisp and fuzzy FPCs and different choices of the tuning constants.

The algorithm described above has been applied to a setup with $p = 6$ and four simulated subpopulations. The separation of the clusters has been governed by a parameter d, by which all cluster centers have been multiplied while the covariance structure remained unchanged. d has been chosen as 1.5, 1, 0.85, 0.7. Variable 5 has been homogeneous normal noise, variable 6 has been homogeneous noise from a t_2-distribution. With respect to variable 1-4, subpopulation 1 consisted of 100 points from an $N_4(\mathbf{a}, \mathbf{\Sigma})$-distribution with $\mathbf{a} = d * (0, 2, 0, 2)$ and $\mathbf{\Sigma} = 0.1\mathbf{I}_4$ ($\mathbf{I}_4$ is the unit matrix). Subpopulation 2 consisted of 200 points from an $N_4(\mathbf{a}, \mathbf{\Sigma})$-distribution with $\mathbf{a} = d * (3, 3, 3, 3)$. The variances have been 0.5, the off-diagonal entries of $\mathbf{\Sigma}$ have been 0.25.

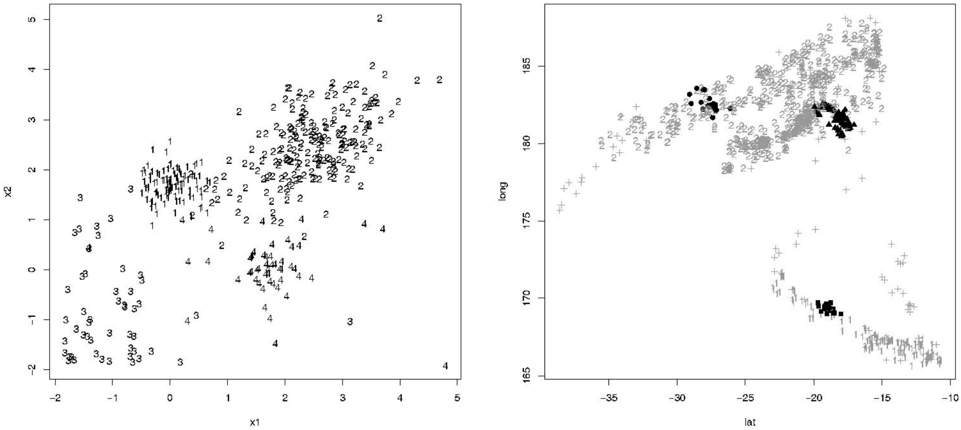


Fig. 2. Left: first two dimensions of simulated data set ($d = 0.85$). Numbers denote cluster memberships. Right: first two dimensions of quakes data with fuzzy FPCs ($\beta = 0.9$, $\gamma = 0.999$). Points denoted by "+" are not part of any FPC. The black squares, triangles and circles denote FPCs no. 3, 4 and 5, which are subsets of FPC no. 1 or 2.

Subpopulation 3 (50 points) has been generated independently from 4 exponential distributions with $\lambda = 1$ shifted to a center of $-d$. Subpopulation 4 (50 points) has been generated by a multivariate t_2-distribution with center $d * (2, 0, 2, 0)$ and the same scatter matrix as for cluster 1. Variables 1 and 2 of such a data set ($d = 0.85$) are shown on the left side of Figure 2.

A desirable solution for a simulated data set consists of 2 FPCs for the two normal clusters no. 1 and 2 and one FPC containing the non-outlying points (but not all) of the t_2-cluster no. 4. There may or may not be an FPC corresponding to some core points of subpopulation 3, of which the shape deviates strongly from the cluster prototype model (subsamples with non-normal shape occur often in complex data), and there should not be other FPCs except of reasonable unions of the clusters.

Additionally there have been simulations of two different types of homogeneous data sets, namely from a single $N_p(\mathbf{0}, \mathbf{I}_p)$-distribution and a single multivariate t_2-distribution with $p = 2, 5, 10$ and $n = 50, 200, 500$ ($n = 100$ instead of $n = 50$ has been used for $p = 10$). Simulations on homogeneous data show the tendency of the methods to generate meaningless FPCs (there should be always only one FPC in these situations). Because FPCs are invariant against addition and deletion of points in areas where the weight function is 0, the results generalize to data sets with further points and subpopulations that are outlying w.r.t. the original FPC. For every combination of parameters and tuning constants there have been 100 simulation runs.

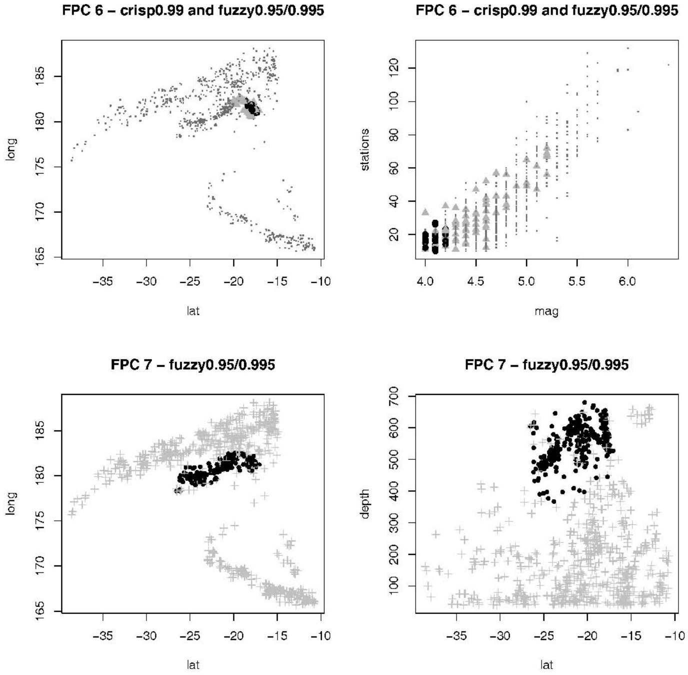


Fig. 3. Above: scatterplots with FPC no. 6 found by crisp0.99 and by fuzzy0.95/0.995 (black). Gray triangles denote FPC no. 4 from Fig. 2, of which FPC no. 6 is a subset. Below: scatterplots with FPC no. 7 only found by fuzzy0.95/0.995.

α for crisp FPCs has been chosen as 0.95 and 0.99 ($\alpha = 0.95$ yielded always too many FPCs). (β, γ) for fuzzy FPCs have been chosen as $(0.9, 0.999)$, $(0.9, 0.995)$, $(0.9, 0.99)$, $(0.95, 0.995)$ (only the first and the last choice did not almost always find too many FPCs). r_0 has been chosen as 0.1 and 0.3 ($r_0 = 0.3$ proved necessary to find a reasonable small number of FPCs). n_{start} has been chosen as $6 + 2p, 12 + 2p, 18 + p$ (the latter choice delivered always the best results). s_0 has been chosen as $17/23 = 0.739$ always, which means that two sets with 20 points both are regarded as "similar" if they have at least 17 points in common.

The full simulation results will be published elsewhere. Here are the most interesting results with the best three methods ("crisp0.99", "fuzzy0.95/0.995" and "fuzzy0.9/0.999" with $r_0 = 0.3$, $n_{start} = 18 + p$):

- For the homogeneous populations, the number of found FPCs is sometimes too high for $p = 2$: fuzzy0.9/0.999 delivers on average 1.02 FPCs for $n = 50$, 1.05 FPCs for $n = 200$ and 1.13 FPCs for $n = 500$ (normal distribution; the t_2 results are very similar). For $n = 500$, this is by far the best value: crisp0.99 yields 2.10 FPCs and fuzzy0.95/0.995 leads to

1.67 FPCs. For $p = 5$ and $p = 10$, the fuzzy methods work well, but crisp0.99 delivers still too many FPCs for $p = 5, n < 500$ and often not even one of them for $p = 10, n < 500$. It may be worthwhile to choose α, β and γ dependent of p or n (namely larger than used here in the cases where problems arose).

- In the clustered data set, the quality of the recovery of the clusters has been measured by the maximum Jaccard similarity and the minimum mean squared error of the center and covariance matrix estimation among the found FPCs. It turned out that all three methods were able to find cluster 1 for $d = 1.5, 1, 0.85$, but for $d = 0.7$, fuzzy0.9/0.999, which requires the strongest simulation, has been seriously biased. Cluster 2 has been successfully separated from the rest of the data for $d = 1.5$ by all methods (fuzzy0.9/0.999 yields the best results if the patterns are separated enough) and for $d = 1$ by crisp0.99 and fuzzy0.95/0.995. For cluster 4, all methods work for $d = 1.5, 1$. For $d = 0.85$, fuzzy0.9/0.999 fails in most cases while the other methods often work and crisp0.99 is a bit better than fuzzy0.95/0.995.

As a conclusion, fuzzy0.9/0.999 is the most stable method, but requires the strongest separation. In terms of the ability to find clusters that are not strongly separated, crisp0.99 is a bit better than fuzzy0.95/0.995, but the latter fuzzy method seems to be a good compromise because it delivers clearly better and more stable parameter estimators on well enough separated data sets, as could be expected from the experience with redescending M-estimators.

5 A data example

The methods crisp0.99, fuzzy0.9/0.999 and fuzzy0.95/0.995 have been applied to a 5-dimensional data set of 1000 seismic events on Fiji ("quakes"). The variables are latitude, longitude, depth, Richter magnitude and number of stations reporting. The data set comes with the R base package.

The right side of Figure 2 shows a scatterplot of latitude and longitude, where some clear patterns are visible. The scatterplot shows the five FPCs that were found by fuzzy0.9/0.999 (points with $w_i > 0.5$ are indicated). The data seem to be separated into two clearly visible parts at longitude= 175. The cores of these two parts form the FPCs no. 1 and 2. The further three FPCs are more concentrated subsets of these clusters. Patterns which are not identical, but have a Jaccard similarity of $s > 0.9$ with these FPCs, are found by crisp0.99 and fuzzy0.95/0.995 as well. The latter two methods find a further FPC, which is plotted on the top line of Figure 3. It is a subset of FPC no. 4 of fuzzy0.9/0.999, but it is much more homogeneous in terms of the magnitude and the number of stations reported (right scatterplot). fuzzy0.95/0.995 even yields a seventh FPC, which can be seen on the bottom line of Figure 3. This FPC corresponds to a pattern that is clearly visible in the latitude/longitude scatterplot, is fairly homogeneous also in terms of depth (right scatterplot), and is geographically meaningful. This pattern has a nonlinear structure and deviates strongly from the normal cluster prototype. fuzzy0.9/0.999 seems to converge to the full cluster no. 2 if the algorithm is started from most of the member points of this pattern, while crisp0.99 seems to deliver one of the subclusters no. 4 and 6 in most cases. Thus, in this data set, fuzzy0.95/0.995 yields the most interesting results, at least from an exploratory point of view.

References

ANDREWS, D.F., BICKEL, P.J., HAMPEL, F.R., HUBER, P.J., ROGERS, W.H., and TUKEY, J.W. (1972): *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton NJ.

BAIER, D., GAUL, W., and SCHADER, M. (1997): Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring. In: R. Klar and O. Opitz (Eds.): *Classification and Knowledge Organization.* Springer, Berlin, 557–566.

BECKER, C. and GATHER, U. (1999): The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association, 94, 947-955.*

GAUL, W. and SCHADER, M. (1994): Pyramidal Classification Based on Incomplete Dissimilarity Data. *Journal of Classification, 11, 171–193.*

HAMPEL, F.R. (1975): Beyond location parameters: Robust concepts and methods. *Bulletin of the International Statistical Institute 46 - Proceedings of the 40th Session*, 375–382.

HAMPEL, F.R. (2002): Some Thoughts about Classification. In: K. Jajuga, A. Sokolowski, and H.-H. Bock (Eds.): *Classification, Clustering, and Data Analysis.* Springer, Berlin, 5–26.

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. and STAHEL, W.A. (1986): *Robust Statistics.* 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 Organization,* Springer, Berlin, 20–28.

HENNIG, C. (2002): Fixed point clusters for linear regression: computation and comparison. *Journal of Classification, 19, 249–276.*

HENNIG, C. (2003): Clusters, Outliers, and Regression: Fixed Point Clusters. *Journal of Multivariate Analysis, 86, 183–212.*

HENNIG, C. and CHRISTLIEB, N. (2002): Validating visual clusters in large datasets: fixed point clusters of spectral features. *Computational Statistics and Data Analysis, 40, 723–739.*

HUBER, P.J. (1981): *Robust Statistics.* Wiley, New York.

MÜLLER, C.H. and GARLIPP, T. (2005): Simple consistent cluster methods based on redescending M-estimators with an application to edge identification in images. *Journal of Multivariate Analysis, 92, 359–385.*

Interpretation Aids for Multilayer Perceptron Neural Nets

Harald Hruschka

Department of Marketing, University of Regensburg, Universitätsstraße 31,
D-93053 Regensburg, Germany

Abstract. Neural nets of the multilayer perceptron (MLP) type possess excellent approximation and forecasting capabilities as many empirical business studies demonstrate. On the other hand, MLPs are often criticized for their black box character because as a rule no single parameter of a MLP indicates the direction of the effect of any predictor. By this property MLPs differ from (general) linear models. We distinguish MLPs for regression analysis, market share analysis and choice modeling. We suggest two different interpretation aids allowing to gain insight into predictors' effects which both require that parameters of the MLP have already been estimated.

1 Introduction

Mathematical proofs show that neural nets (NNs) of the multilayer perceptron (MLP) type possess excellent approximation capabilities compared to other flexible methods for nolinear modeling like polynomial expansions, splines or kernels (e.g. Hornik et al. (1989), Barron (1993), Jones (1992)). These proofs are confirmed by many empirical studies. Moreover, MLPs often attain good results not only w.r.t. to model fit, but also w.r.t. forecasting performance if model complexity is restricted. This explains why neural nets have found many applications in business, above all in finance (Qi (1999), Donaldson and Kamstra (1997), Garcia and Gencay (2000)), accounting (Baetge and Uthoff (1998), Luther (1998), Anders and Szczesny (1998)) and marketing (Agrawal and Schorling (1996), Heimel et al. (1998), Hruschka (1993, 1999, 2001), Hruschka et al. (2002, 2004), Hruschka and Natter (1999), Mazanec (1999), West et al. (1997)).

On the other hand, MLPs are often critized for their black box character. As a rule, the sign of any single parameter of a MLP does not indicate whether the effect of a predictor is positive or negative, i.e. that the dependent variable increases or decreases with higher values of the predictor. By this property MLPs differ from linear models (e.g. linear regression models) or general linear models (e.g. logit oder probit models with linear deterministic utility), for which, e.g. a negative price coefficient tells that sales or deterministic utilities decrease if price is raised.

The neural net literature does not pay much attention to the interpretation of MLP results. That is why we suggest two interpretation aids aimed at gaining insight into effects of predictors. These interpretation aids require that the parameters of the MLP have already been estimated (details on estimation procedures for MLPs may be found in Bishop (1995).

2 Multilayer perceptrons

We only consider MLPs with one layer of Q hidden units all of which have the binary logistic as activation function. The potential z_j of a hidden unit j is formed as linear combination of predictors (input variables), its output value is computed by putting z_j into the (binary) logistic function $g(z_j) = 1/(1+\exp(-z_j))$. Output values of the logistic function lie between zero and one, its first derivative is $g(z_j)(1-g(z_j))$. Increasing (decreasing) the potential z_j causes output values to approach one (zero).

Input variables of MLPs are often standardized or z-transformed (Bishop (1995) to avoid numerical problems during estimation which are caused by different value ranges:

$$\tilde{x}_{pit} = \begin{cases} x_{pit}/s_p & : \text{standardization} \\ (x_{pit} - \bar{x}_p)/s_p & : \text{z-transformation} \end{cases} \tag{1}$$

x_{pit} denotes the p-th predictor of brand i in period t, $\bar{x}_p, s_p$ denote arithmetic mean and standard deviation of predictor p, respectively.

In the following we present three types of MLPs, which are generalizations of regression models, multinomial logit choice models and attraction market share models. A regression-MLP (see Bishop (1995)) consists of a constant β_0, a linear part $\sum_p \beta_{1p} x_{pt}$ (x_{pt} denotes predictor p in period t) and a nonlinear part formed by a linear combination of Q hidden units with potentials z_j:

$$y_t = \beta_0 + \sum_p \beta_{1p} x_{pt} + \sum_{j=1}^{Q} \beta_{3j} g(z_j) \quad \text{with}\, z_j = \beta_{2j0} + \sum_p \beta_{2jp} x_{pt} \tag{2}$$

The first derivative of a regression-MLP w.r.t. predictor x_{pt} is:

$$\frac{\partial y_t}{\partial x_{pt}} = \beta_{1p} + \sum_{j=1}^{Q} \beta_{3j} \beta_{2jp} g(z_j)(1-g(z_j)) \tag{3}$$

The basic equation of the multinomial logit (MNL) model makes choice probability (market share) of brand i for purchase occasion (period) t dependent on (deterministic) utilities of all brands including brand i for purchase occasion (or period) t in the following way (McFadden (1973), Cooper and Nakanishi (1988)):

$$P_{it} = \frac{\exp(V_{it})}{\sum_j \exp(V_{jt})} \tag{4}$$

For the neural net-multinomial logit (NN-MNL) choice model (Hruschka et al. (2002, 2004)) utility of a brand consists of a brand constant β_{0i}, a linear part (a linear combination of predictors) and a nonlinear part which is formed as linear combination of Q hidden units :

$$V_{it} = \beta_{0i} + \sum_p \beta_{1p} x_{pit} + \sum_{j=1}^{Q} \beta_{3j} g(z_j) \quad \text{with}\, z_j = \beta_{2j0} + \sum_p \beta_{2jp} x_{pit} \quad (5)$$

Coefficients of the NN-MNL model do not vary across brands with the exception of brand constants.

The artificial neural net attraction (ANNAM) model (Hruschka (2001)) is similar to the NN-MNL model, but has brand-specific coefficients and brand-specific numbers of hidden units Q_i:

$$V_{it} = \beta_{0i} + \sum_p \beta_{1ip} x_{pit} + \sum_{j=1}^{Q_i} \beta_{3ij} g(z_{ij}) \text{ with}\, z_{ij} = \beta_{2ij0} + \sum_p \beta_{2ijp} x_{pit} \, (6)$$

The first derivatives of the NN-MNL and ANNAM models w.r.t. predictor x_{pit} are:

$$\frac{\partial V_{it}}{\partial x_{pit}} = = \begin{cases} \beta_{1p} + \sum\limits_{j=1}^{Q} \beta_{3j} \beta_{2jp} g(z_j)(1 - g(z_j)) & : \text{NN-MNL} \\ \beta_{1ip} + \sum\limits_{j=1}^{Q_i} \beta_{3ij} \beta_{2ijp} g(z_{ij})(1 - g(z_{ij})) & : \text{ANNAM} \end{cases} \quad (7)$$

To demonstrate the use of the interpretation aids we introduce an example for the NN-MNL model with reference price, gain (i.e. max(reference price − observed price, 0)), loss (i.e. max(observed price − reference price, 0), loyalty, display (binary) and feature (binary) as predictors affecting utility. These predictors vary over households, brands and purchase occasions. Display (i.e. POS advertising) and feature (i.e. newspaper advertising) can be obtained directly from purchase data. Loyalties and reference prices have to be estimated.

Reference prices constitute internal prices to which households compare observed prices (Winer (1988)). Observed prices below reference price (which households perceive as gains) stimulate purchases, i.e. increase choice probability. Observed prices above reference price (which households perceive as losses) may deter from purchasing and therefore decrease choice probability. Prospect theory defines a value function with gains and losses as arguments. Prospect theory postulates that losses have more effect than gains of the same size. Reference prices, price gains and losses are z-transformed (mean 112.04 and standard deviation 25.33), gains and losses are standardized (standard deviation 25.33). Following the seminal paper of Guadagni and Little (1983) brand-specific loyalty values are first-order exponentially smoothed past purchases for each household.

	Hidden Units			
	1	2	3	4
β_{2j0}	0.257	-2.466	-2.238	-2.033
β_{3j}	5.476	10.856	9.527	-12.516
β_{2jp}				
Reference Price	0.988	-0.466	-0.722	0.235
Gain	0.193	0.103	1.565	0.541
Loss	-0.099	-0.073	-0.065	3.129
Loyalty	1.425	1.055	1.021	-0.100
Display	0.633	-0.019	0.002	0.116
Feature	1.022	-0.110	-0.291	-0.043
brand constants β_{0i}				
Brand 1	0.000			
Brand 2	-0.087			
Brand 3	-0.234			
Brand 4	-0.665			
Brand 5	-0.921			
Brand 6	-1.219			

Table 1. Parameters of NN-MNL Model (example)

Table 1 contains the parameter values of this model which has no linear part (i.e. for all predictors $\beta_{1p} = 0$). To ensure identification the brand constant of the first brand is set to zero. Therefore the other brand constants measure utility difference w.r.t. the first brand.

3 Elasticities and probability differences of predictors

In economics and marketing effects of predictors on dependent variables are often measured by means of elasticities (Hanssens et al. (1990), Hruschka (1996)). Elasticities indicate the relative change of a dependent variable given a relative change of the respective predictor with values of other predictors set to constant values (e.g. average or most frequent values). Elasticities have the nice property that they do not depend on the measurement unit used (e.g. it does not matter whether a price is measured in Euro or US Dollars).

For a regression-MLP the point elasticity of predictor x_{pt} is given by:

$$\frac{\partial y_t}{\partial x_{pt}} \frac{x_{pt}}{y_t} = \frac{x_{pt}}{y_t}[\beta_{1p} + \sum_{j=1}^{Q} \beta_{3j}\beta_{2jp} g(z_j)(1 - g(z_j))] \tag{8}$$

The point elasticity of predictor x_{piz} for the NN-MNL choice model is:

$$\frac{\partial P_{it}}{\partial V_{it}} \frac{\partial V_{it}}{\partial x_{pit}} \frac{x_{pit}}{P_{it}} = (1 - P_{it}) x_{pit} [\beta_{1p} + \sum_{j=1}^{Q} \beta_{3j}\beta_{2jp} g(z_j)(1 - g(z_j))] \tag{9}$$

For the ANNAM market share model we obtain the following expression for the point elasticity of predictor x_{pit}:

$$\frac{\partial P_{it}}{\partial V_{it}} \frac{\partial V_{it}}{\partial x_{pit}} \frac{x_{pit}}{P_{it}} = (1 - P_{it}) x_{pit} [\beta_{1ip} + \sum_{j=1}^{Q_i} \beta_{3ij} \beta_{2ijp} g(z_{ij})(1 - g(z_{ij}))] (10)$$

For a transformed predictor we obtain the elasticity referring to the raw (i.e. non-transformed) value by multiplying the elasticity obtained for the transformed value by the first derivative of the transformed predictor w.r.t. the original predictor as the following equalities for the regression-MLP show:

$$\frac{\partial y_t}{\partial x_{pt}} \frac{x_{pt}}{y_t} = \frac{\partial y_t}{\partial \tilde{x}_{pt}} \frac{\partial \tilde{x}_{pt}}{\partial x_{pt}} \frac{x_{pt}}{y_t} = \frac{\partial y_t}{\partial \tilde{x}_{pt}} \frac{1}{s_p} \frac{x_{pt}}{y_t} \tag{11}$$

For the two transformations (z-transformation, standardization) considered here (see equation 1) these means multiplying by the inverse of the standard deviation $1/s_p$ of the predictor (this result is also valid for NN-MNL and ANNAM models).

For a binary predictor we may consider the difference of sales, choice probabilities or market shares caused by changing the value of a predictor from zero to one. Values of the dependent variable for a predictor value of zero as well as of one are simply obtained by inserting these values into equations 2, 4, 5 or 6 (holding values of the other predictors constant).

Continuing the NN-MNL example of Table 1 we determine choice elasticities of brand 1 w.r.t. reference price, loss and loyalty given raw values of predictors as shown in Table 2.

Predictor	Brands					
	1	2	3	4	5	6
Reference Price	130	110	110	110	110	110
Gain	0	0	0	0	0	0
Loss	5	0	0	0	0	0
Loyalty	0.35	0.2	0.15	0.1	0.1	0.1
Display	0	0	1	0	1	0
Feature	0	0	1	0	1	0

Table 2. Predictor Values for Elasticity Computations

Using equation 9 (and multiplying by the inverse of standard deviation for the transformed variables reference price and loss) we obtain elasticities of −5.4326, −0.7212 and 1.2135 for reference price, loss and loyalty, respectively. For the values considered increasing any of these predictors by 1% leads to a decrease of choice probability by −5.43% or −0.72%, an increase of choice probability by 1.21%.

If gain is set to 5 (i.e. observed price is lower than expected price by 5) we obtain elasticities for brand 1 of −3.9518, 0.1943 and 0.8828 for reference price, gain and loyalty, respectively. Note that in absolute values relative effects of reference price and loyalty are smaller compared to the situation with a loss (i.e. observed price higher than expected price). Moreover, the absolute effect of a gain is lower than the effect of a loss of the same size which is in accordance with prospect theory.

On the basis of the values given in Table 2 the choice probability of brand 1 amounts to 0.1354. This probability increases by 0.0603 (0.1008) if brand 1 is displayed (featured) and by 0.1245 if brand 1 is both displayed and featured. These results show that features have more effect than displays and that the total effect of both instruments is subadditive (i.e. lower than the sum of their individual effects).

Predictor	Graphs			
	1	2	3	4
Utility vs. Gain				
Reference Price	0.00	0.00	-1.50	-1.50
Loyalty	0.30	0.30	0.30	0.30
Feature	0	1	0	1
Display	0	1	0	1

Table 3. Input Values of Predictors Used in 2D Graphs

4 Two-dimensional graphs

To illustrate predictors' effects modeled by MLPs we generate graphs which show values of the output variable (e.g. sales, utility) obtained by varying the value of a predictor given selected combinations of the values of the other predictors. Values of the dependent variable are computed by plugging values of all predictors into the relevant equations 2,4,5,6. Given parameter estimates such graphs can easily be generated by spreadsheet software. Table 3 gives values of the remaining predictors (z-transformed values for reference prices, standardized values for gain and loss) used in the following two-dimensional graphs. The effect of gain (i.e. observed price below reference price) on utility for the NN-MNL example of Table 1 can be seen in Figure 1 for four different constellations characterized by low vs. high reference price (corresponding to the two upper and two lower curves, respectively) and use vs. non-use of advertising. Utility increases with gain and follows an S-shape. The change of this effect becomes very small at high gain values. While the positive effect of advertising on utility seems to be constant at lower reference prices, this effect becomes practically zero for somewhat higher gains for the high reference price constellation.

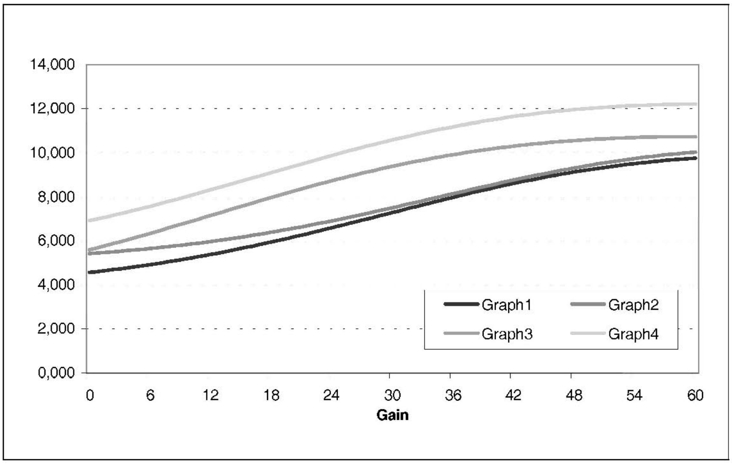


Fig. 1. Utility vs. Gain

References

AGRAWAL, D. and SCHORLING, C. (1996): Market Share Forecasting. An Empirical Comparison of Artificial Neural Networks and Multinomial Logit Model. *Journal of Retailing, 72, 383-407.*

ANDERS, U. und SCZESNY, A. (1998): Prognose von Insolvenzwahrscheinlichkeiten mit Hilfe logistischer neuronaler Netzwerke. *Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung, 892–915.*

BAETGE, J. and UTHOFF, C. (1998): Development of a Credit-Standing-Indicator for Companies Based on Financial Statements and Business Information with Backpropagation-Networks. In: G. Bol (Ed.): *Risk Measurement, Econometrics and Neural Networks.* Physica, Heidelberg, 17–38.

BARRON, A.R. (1993): Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transasctions on Information Theory, 39, 930–945.*

BISHOP, C.M. (1995): *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford.

COOPER, L.G. and NAKANISHI, M. (1988): *Market-Share Analysis.* Kluwer Academic Publishers, Boston.

DONALDSON, R.G. and KAMSTRA, M. (1997): An Artificial Neural Network-GARCH Model for International Stock Return Volatility. *Journal of Empirical Finance, 4, 17–46.*

GARCIA, R. and GENCAY, R. (2000): Pricing and Hedging Derivative Securities with Neural Networks and a Homogeneity Hint. *Journal of Econometrics, 94, 93–115.*

GUADAGNI, P.M. and LITTLE, J.D.C. (1983): A Logit Model of Brand Choice Calibrated on Scanner Data. *Marketing Science, 2, 203–238.*

HANSSENS, D.M., PARSONS, L.J., and SCHULTZ, R.L. (1990): *Market Response Models. Econometric and Time Series Analysis.* Kluwer Academic Publishers, Boston.

HEIMEL, J.P., HRUSCHKA, H., NATTER, M., and TAUDES, A. (1998): Konnexionistische Kaufakt- und Markenwahlmodelle. *Zeitschrift für betriebswirtschaftliche Forschung, 596–613.*

HORNIK, K., STINCHCOMBE, M., and WHITE, H.(1989): Multilayer Feedforward Networks are Universal Approximators. *Neural Networks, 3, 359–366.*

HRUSCHKA, H. (1993): Determining Market Response Functions by Neural Network Modeling. A Comparison to Econometric Techniques. *European Journal of Operational Research, 66, 27–35.*

HRUSCHKA, H. (1996): *Marketing-Entscheidungen.* Vahlen, München.

HRUSCHKA, H. (1999): Neuronale Netze. In: A. Herrmann and C. Homburg (Eds.): *Marktforschung.* Gabler, Wiesbaden, 661–683.

HRUSCHKA, H. (2001): An Artificial Neural Net Attraction Model (ANNAM) to Analyze Market Share Effects of Marketing Instruments. *Schmalenbach Business Review-zfbf, 27–40.*

HRUSCHKA, H. and NATTER, M. (1999): Comparing Performance of Feedforward Neural Nets and K-Means for Cluster-Based Market Segmentation. *European Journal of Operational Research, 114, 346–353.*

HRUSCHKA, H., FETTES, W., und PROBST, M. (2002): Die Bewährung von Ankerpreismodellen bei der Erklärung der Markenwahl. *Zeitschrift für betriebswirtschaftliche Forschung, 426–441.*

HRUSCHKA, H., FETTES, W., und PROBST, M. (2004): An Empirical Comparison of the Validity of a Neural Net Based Multinomial Logit Choice Model to Alternative Model Specifications. *European Journal of Operational Research, 159, 166–180.*

JONES, L.K. (1992): A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit and Neural Network Training *Annals of Statistics, 20, 608–613.*

LUTHER, R.K. (1998): An Artificial Neural Network Approach to Predicting the Outcome of Chapter 11 Bankruptcy. *Journal of Business and Economic Studies, 4, 57– 73.*

MAZANEC, J.A. (1999): Marketing Applications of Neural Networks. *Journal of Retailing and Consumer Services, 6, 182–261.*

MCFADDEN, D. (1973): Conditional Logit Analysis of Qualitative Choice Behavior. In: P. Zarembka (Ed.): *Frontiers in Econometrics.* Academic Press, New York, 105–142.

QI, M. (1999): Nonlinear Predictability of Stock Returns Using Financial and Economic variables. *Journal of Business & Economic Statistics, 17, 419–429.*

WEST, P.M., BROCKETT, P.L., and GOLDEN, L.L. (1997): A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice. *Marketing Science, 16, 370–391.*

WINER, R.S. (1988): Behavioral Perspective on Pricing, In: T.M. Devinney (Ed.): *Issues in Pricing.* Lexington Books, Lexington, MA, 35–57.

An Unfolding Scaling Model for Aggregated Preferential Choice Data

Tadashi Imaizumi

School of Management & Information Sciences, Tama University,
4-1-1 Hijirigaoka, Tama-shi, Tokyo 206-0022, Japan

Abstract. The unfolding model has been widely used as a model for preferential choice data. This model is treated as the special case of multidimensional scaling with so-called "ideal" points. In this model, the distance between an "ideal" point and object points are related to the degree of individual preferential choice data for objects. However, the unfolding model has some difficulties, degeneracies, indeterminacies and multidimensionality problems in application to real data. In this paper, we propose a parametric unfolding model for aggregated choice data by introducing the attractiveness of objects.

1 Introduction

Preferential choice data is often gathered in marketing research for exploring market structures, market segments, positionings of established and new products in future markets and so on. A product positioning map is useful to get hold of the facts in the data. So, e.g., we can use the results obtained to decide which advertising is more effective or whether we introduce an new product into a market. MultiDimensional Scaling (MDS) models and methods have been used to explore the hidden market structure in these data. The similarity between two objects or the preference to an object for each individual are related to the distance between two points, which represent two objects, or an object and an individual, respectively. The MultiDimensional Unfolding (MDU) models, such as the ideal point model or the ideal vector model, are models for analyzing preferential choice data. We can interpret this MDU as a special case of MDS models, in which we try to analyze the rectangular sub-matrix of similarity matrix of $n + m$ objects. On the other hand, these MDU models also are derived as a generalization of the models for paired-comparison data such as BTL models,

$$p_{jk|i} = p_{j|i}/(p_{j|i} + p_{k|i}). \qquad (1)$$

The MDU model assumes that n individuals share the common space of preferential choice. And each individual is represented as an "ideal" point in the t-dimensional space as the distances from the ideal point to the object point are related to the preference value to the objects for that individual. These unfolding models have also the same characteristics as ordinary MDS models and incorporated with the model based on a choice axiom. However, the MDU model has some difficulties in application to real data set, specially, degeneracies, indeterminacies and multidimensionality problems. The degeneracies and the indeterminacies problems may be overcome by technical procedures partially. The problem of multidimensionality is related to the problem of the case that they are essentially embedded into a subspace, though, all of individuals and objects are embedded into the multidimensional space. Wang et al. (1975), Borg and Groenen (2001, pp. 251 and pp. 265-269) discussed this problem. Their analysis indicated that individuals evaluate the preferential choice on the dimensions embedded in a subspace. This problem can mislead us in some decision making. One main cause is the assumption that all individuals are embedded as points of space of same dimensionality.

To relax this assumption and to find the latent structure of the preferential choice, the weighted MDU model is a valuable one to be applied since the dimension weights define the subspace of the preferential choice.

It is assumed in the MDU model that preferential choice is determinated only by the distance between "ideal" point and object point. However, this model has some difficulties in application. When someone analyze the goods purchasing data matrix of which row is corresponding to the resident city and column is corresponding to the the city purchasing goods, the results of MDU will indicate the map based on the preferential choice data. This map is different from the usual geometrical map. Both of map, the preferential map and geometrical map are important to investigate the consumer purchasing behavior. This indicates that MDU models only based on distances may be improper for some choice data. In this case, the geometrical distance and the attractiveness of the city are two important factors in the data. We also propose the model in which the preferential choice data is represented by distance parameter and attractiveness parameter.

2 Model

MDU models have been proposed as the models for analyzing individual preferential choice behavior.

$$pref_{ij} = Mono(d_{ij}), \tag{2}$$

where $pref_{ij}$ is the preference value to object o_j for individual I_i and

$$d_{ij} = \{\sum_{p=1}^{t} \|y_{ip} - x_{jp}\|^r\}^{(}1/r), r \geq 1. \tag{3}$$

As the number of parameters for individual differences MDU model increases as number of individuals, it need some investigation for modeling. And we focus on analyzing the aggregated preferential choice data instead of individual choice data in this paper.

Let $P(O = o_j|G = g)$ be the probability of object o_j chosen for group g,then this probability is expressed as follows:

$$p_{j|g} = P(O = o_j|G = g) = P(G = g, O = o_j)/P(G = g). \tag{4}$$

where $P(G = g, O = o_j)$ is a joint probability of G and O, then we need to express this joint probability using distance. A rough approach is

$$P(G = g, O = o_j) \propto 1/d_{gj}, \tag{5}$$

This has a serious defect that this probability goes up to the infinity as the object point come near to the "ideal" point. And we adopt other formation. The simple MDU model will be

$$P(G = g, O = o_j) \propto e^{-d_{gj}^2}, \tag{6}$$

and

$$d_{gj}^2 = [\sum_{p=1}^{t} |y_{gp} - x_{jp}|^2], \tag{7}$$

where $(y_g : g = 1, 2, \ldots, n)$ and $(x_j : j = 1, 2, \ldots, m)$ are points in R^t-dimensional Euclidean space and

$$P(G = g) = \sum_{j=1}^{m} P(G = g, O = o_j). \tag{8}$$

2.1 The weighted MDU model

The weighted MDU model can be expressed as follows:

$$d_{gj}^2 = [\sum_{p=1}^{t} w_{gp}|y_{gp} - x_{jp}|^2], w_{gp} \geq 0. \tag{9}$$

As the dimension weights (w_{gp}) define the iso-preference contours for each individual. This weighted MDU model assumes that each individual of a group evaluates the preferential choice on the unique dimension which is common to all groups, and that preferential choice behavior is explained in a subspace of this common space. One purpose of analyzing these preferential choice data is to find the latent factors or dimensions on which each individual expresses his/her degree of preferential choice. And this property is one of the important property we keep. The number of parameters to be estimated in this weighted MDU model is $n \times t + (n + m - 1) \times p$. MDU model has several difficulties in general as mentioned, and the more constrained model will be useful. As one parsimonious weighted MDU model, we assume that the iso-preference contours for n groups are classified from 1 contour to n contours. This is done by clustering of the dimension weight of each group.

$$d_{gj}^2 = [\sum_{p=1}^{t} \tilde{w}_{cp}|y_{gp} - x_{jp}|^2], \tilde{w}_{cp} \geq 0, c = 1, 2, \cdots, n. \tag{10}$$

2.2 The attractiveness of object

The model with only distance term will not be the proper one. And we propose the model with the attractiveness parameter to analyze the aggregated data. Let F_{gj} is observed preferential frequency to object o_j for group g.
A well-known model in which incorporated the attractiveness of objects is a gravity model,

$$F_{gj} = \beta S_g^{\alpha} M_j^{\lambda} / d_{gj}^2 \tag{11}$$

where F_{gj} is the degree of preferential choice to object o_j of group g, S_g^{α} is such as consumer mass of group g,and M_j is brand mass of object o_j, α is mass parameter for group and λ is mass parameter for object.
This model is an interesting one, but, has some defect mentioned. And we propose the different one. The attractiveness of object is independent with distance, and we relate the F_{gj} to $P(G = g, O = o_j)$ in this paper,

$$p_{gj} = P(G = g, O = o_j) \propto e^{-d_{gj}^2} e^{-1/\tilde{M}_j}, \tilde{M}_j > 0. \tag{12}$$

3 Optimization

We assumed that the obtained data is aggregated choice data, in which each sample of group chose one out of n objects. To estimate the model parameter $(x_{jp}), (y_{gp}), (u_p), (w_g)$ and $\tilde{M}_j$ for given t, we find the values by maximizing

$$l(\theta, F) = \prod_{g}^{n} \prod_{j}^{m} p_{g|j}^{f_{gj}} (1 - p_{j|g})^{(f_{g\cdot} - f_{gj})}, \tag{13}$$

where $f_{g\cdot} = \sum_{j=1}^{m} f_{gj}$. By taking logarithm of $l(\theta, F)$, we obtain

$$L(\theta, F) = \sum_{g=1}^{n} \sum_{j=1}^{m} f_{gj}(-d_{gj}^2 - 1/\tilde{M}_j) + (f_{g\cdot} - f_{gj}) ln(1 - e^{-d_{gj}^2 - 1/\tilde{M}_j}) + c_i(\theta). \tag{14}$$

4 Algorithm

4.1 Initial configuration

An initial configuration of $X^{(0)} = (x_j)$ and $Y^{(0)} = (y_g)$ is obtained by SVD of the observed P matrix.

$$P = (p_{j|g}) = (f_{j|g}/f_{g\cdot}), U\Lambda V' = P, Y^{(0)} = U\Lambda^{1/2}, X^{(0)} = V\Lambda^{1/2}, \tag{15}$$

we set w_{gp} to 1, and M_j to 1.

4.2 Iteration process

For given the number of dimensions t, we repeat the following process (1) to (5) until some convergence criteria are satisfied.
(1) Compute $L(\theta^{(s)}, F)$ at the iteration s,
(2) Update $X^{(s)}$ and $Y^{(s)}$, with the gradient method

$$x_{jp}^{(s+1)} = x_{jp}^{(s)} + \alpha^{(s)}\partial L(\theta, F)/\partial x_{jp}, y_{gp}^{(s+1)} = y_{gp}^{(s)} + \alpha^{(s)}\partial L(\theta, F)/\partial y_{gp}. \quad (16)$$

(3) Apply the K-means clustering method for$W^{(s)}$ as $K = 1, 2, \ldots, n$, and choose k which attained the minimum of RMS. Replace all w_{gj} with the cluster mean of which group g belonged.
(4) Update $W^{(s)}$ which minimizes $L(\theta^{(s)}, F)$ using $\tilde{W}^{(s)}$,

$$w_{gp}^{(s+1)} = w_{gp}^{(s)} + \beta^{(s)}\partial L(\theta, F)/\partial w_{gp}. \quad (17)$$

(5) Update $M^{(s)}$

$$\tilde{M}_j^{(s+1)} = \tilde{M}_j^{(s)} + \gamma^{(s)}\partial L(\theta, F)/\partial \tilde{M}_j. \quad (18)$$

5 Application

Toyoda (2004) gathered preferential choice data on a bottled Japanese tea from 284 university students. We formed five groups by his/her frequency of drinking tea, from heavy to light.

Group	No. of Students
Heavy	138
less Heavy	68
Middle	25
less Light	30
Light	23

Table 1. Number of students in each group

In Table 2, we show the normalized preferential choice data.

Group	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
Heavy	.02	.01	.07	.01	.05	.02	.06	.05	.08	.05	.05	.03	.11	.35	.01	.03	.00	.00	.00
less Heavy	.01	.06	.07	.00	.00	.03	.15	.01	.00	.09	.00	.06	.13	.31	.00	.01	.04	.01	.00
Middle	.00	.04	.08	.00	.00	.00	.12	.04	.00	.12	.00	.00	.28	.28	.00	.04	.00	.00	.00
less Light	.07	.00	.10	.00	.07	.07	.10	.00	.03	.00	.00	.03	.17	.30	.00	.07	.00	.00	.00
Light	.04	.00	.09	.00	.04	.00	.13	.00	.00	.00	.00	.00	.09	.48	.00	.00	.00	.13	.00

* row sum is rounded. Letter A,B, ·, S represents each of the bottled Japanese tea.

Table 2. Normalized Frequency (row sum to 1.)

Tea Brands N and M are dominated over all groups, Brand N is a Chinese Oolong tea, and Brand M is mixed Japanese tea. And many other tea is variety of Chinese tea and Japanese tea. Before obtaining configuration, we done the SVD of normalized preferential choice data in Table 2. The singular values were $0.930, 0.234, 0.142, 0.111$ and 0.098, and we set $t = 2$ for this data. By applying an optimization process to this data, the final configuration was obtained, and is shown in Figure 1.

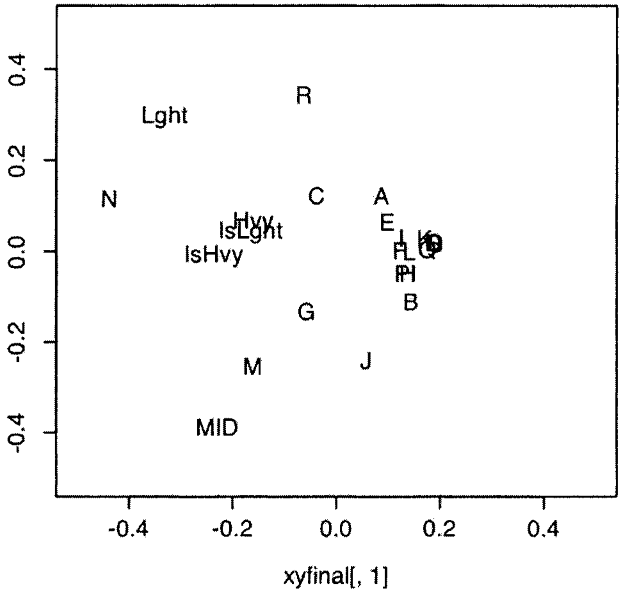


Fig. 1. The Obtained Joint Configuration

Group	Cluster ID.	w_{c1}	w_{c2}
Heavy	1	1.163	0.943
less Heavy	1	1.163	0.943
Middle	1	1.163	0.943
less Light	1	1.163	0.943
Light	2	1.364	0.816

Table 3. Dimension Weights for each group with number of clusters = 2

Figure 1 indicates Tea N and M have different properties.

Table 4 and Figure 2 show the brand attractiveness for each tea. Tea N is dominated, and Tea M is half of it.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
.061	.057	.232	.006	.075	.058	.262	.044	.067	.131	.057	.076	.603	1.000	.001	.082	.019	.017	.000

* we normalized the maximum value to 1.

Table 4. Normalized attractiveness of each object

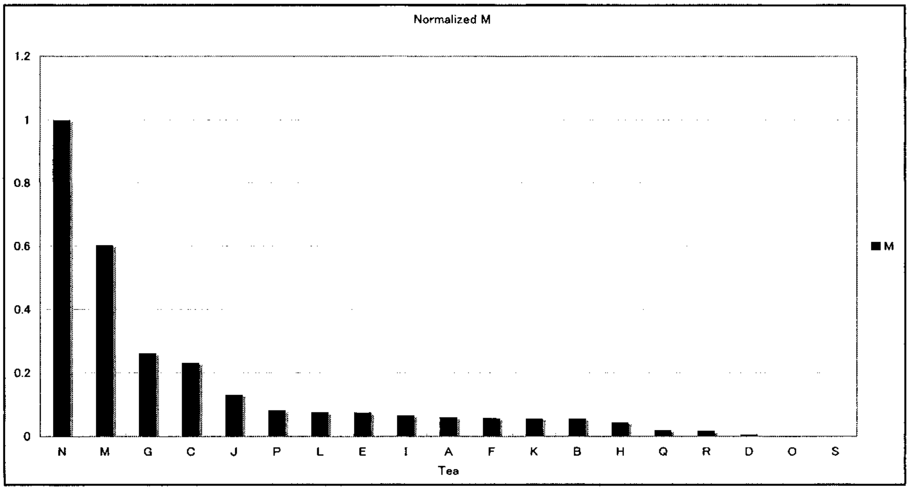


Fig. 2. The Obtained attractiveness of object(Normalized)

6 Discussion

Application to real data indicated that we could get many valuable information from a single resource. As there are only two clusters for the dimension weights, one is four of five groups, and the other only one group, we guess that the dimension weights between cluster is not differ from and all five groups belong to one cluster.

And it also indicated that the estimation procedure of $\tilde{w}_{cp}$ need to improve. The estimation procedure is complex one. And more sophisticated algorithm must be developed.

In this paper, we proposed a simple weighted MDU model which focuses on the finding the latent structure and supporting decision making by adopting a parsimonious model. The results of real data indicates that this MDU model and method is against the degeneracies. We used the Euclidean distance model in this paper. The Minkowski distance model,

$$d_{gj} = [\sum_{p=1}^{t} \tilde{w}_{cp}|y_{gp} - x_{jp}|^r]^{1/r}, \tilde{w}_{cp} \geq 0, c = 1, 2, \cdots, n, \qquad (19)$$

is more interesting one. However, the final configuration of the model with $r \neq 2$ has some difficulties in visual representation. And the relationship between the joint probability and distance has ambiguous, for example,

$$P(G = g, O = o_j) \propto e^{-d_{gj}}, \tag{20}$$

is alternative one.

References

BORG, I. and GROENEN, P. (1997): *Modern Multidimensional Scaling.* Springer-Verlag, New York.

WANG, M. M., SCHÖNEMANN, P. H., and RUSH, J. G. (1975): A Conjugate Gradient Algorithm for the Multidimensional Analysis of Preference Data. *Multivariate Behavioral Research, 10, 45–80.*

TOYODA, Y. (2004): A Characteristic of Free Call Data. Research paper presented at the Faculty Research Conference 2004 of Tama University.

Model-Based Clustering – Discussion on Some Approaches

Krzysztof Jajuga

Department of Financial Investments and Insurance,
Wroclaw University of Economics,
ul. Komandorska 118/120, 53-345 Wroclaw, Poland

Abstract. One of the most well grounded approaches in clustering is model-based clustering, where one assumes particular multivariate distribution for each class. Most results in model-based clustering were obtained under multivariate normal distribution. In the paper we propose to adopt other approach, namely copula analysis in model-based clustering. Two possible stochastic approaches, namely classification approach and mixture approach, are considered as the framework to apply copula analysis. In the paper iterative algorithms are proposed to find optimal solution of clustering problem.

1 Introduction

Clustering is one of the most important areas of multivariate statistical data analysis. It is characterized by the extensive diversity of the proposed methods and approaches. On the other hand, clustering proved its usefulness in very many fields of applications. In this paper we would like to provide some discussion as well as to give some new proposals in the area of the so called **model-based clustering**. The term "model-based clustering" was popularized by Banfield and Raftery (1993). This approach is one of the most advanced approaches in clustering.

Here it is assumed that multivariate data is a sample from a population which consists of a number of subpopulations and a particular multivariate distribution is a model for each subpopulation. Clearly, such an assumption is very appropriate one to build the framework in which one can apply clustering methods to identify classes corresponding to the subpopulations and then to estimate the parameters of the models for these classes.

This type of clustering approach, where the underlying model for classes is defined, is different from the other approach, where the classes are identified by taking into account some relations (for example distances) between different multivariate observations, instead of defining the model for classes. Such an alternative approach is used, for example, in agglomerative hierarchical clustering methods (e.g. Gordon (1999)), where classes are formed to meet the following condition: within-class distances are small and between-class distances are large.

Model-based clustering approach usually assumes stochastic approach and the appropriate model is simply the particular multivariate distribution, for example multivariate normal distribution. In our opinion, the term "model-based clustering" can be extended to non-stochastic, data-analytic approach, provided that the model is precisely defined. For example, as a model for a class we can take location vector, scatter matrix, regression, principal components. In this case the assumption of stochastic framework is not necessary. Such a "data-analytic" model-based clustering consists in the minimization of the goodness-of-fit function. If the model for class is location vector and scatter matrix, we get the minimization of the following function:

$$\sum_{j=1}^{K} \sum_{i \in C_j} (\mathbf{x}_i - \mathbf{v}_j)^{\mathrm{T}} \mathbf{M}_j (\mathbf{x}_i - \mathbf{v}_j)$$

$$|\mathbf{M}_j| = b_j$$

Here:

$\mathbf{v}$ – location vector,

$\mathbf{M}$ – scatter matrix (scaled to keep the volumes of classes at appropriate level).

In the stochastic model-based clustering one can apply one of the two possible approaches, namely:

- classification likelihood (shortly: classification) approach;
- mixture approach.

The classification approach was described for the first time by Scott and Symons (1971). Here it is assumed that the observations

$$\mathbf{x}_1, , \mathbf{x}_2, , ..., \mathbf{x}_n$$

come from one of Ksubpopulations:

$$\Pi_1, \Pi_2, ..., \Pi_K$$

Then the likelihood function is given as:

$$L(\theta\,|\mathbf{x}_1, , \mathbf{x}_2, , ..., \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i\,|\theta) = \prod_{i=1}^{n} f_{\gamma_i}(\mathbf{x}_i\,|\theta_{\gamma_i})$$

$$\gamma_i = j \Leftrightarrow \mathbf{x}_i \in \Pi_j$$

Here *gamma* parameters correspond to observations (so there are n such parameters) – these are indicator parameters showing number (label) of a class the respective observation belongs to. If we assume that the number of parameters of the model for each class is equal to s, then the total number of parameters is equal to $Ks+n$.

In practice, to find the classification that maximizes the likelihood, the following E-M type algorithm is used:

1. Start from initial classification.
2. In each iteration:
   (a) Estimate parameters of the distribution for each class.
   (b) Calculate density for each observation given this observation belongs to respective class.
   (c) Make new classification by assigning each observation to the class of highest density.
3. Iterate step 2 until classification does not change.

The particular variant of classification approach depends on the particular multivariate distribution being the model assumed for the class. As expected, most results were obtained in the case of multivariate normal distribution.

The mixture approach is presented for example by Wolfe (1970). Here it is assumed that the observations

$$\mathbf{x}_1, , \mathbf{x}_2, , ..., \mathbf{x}_n$$

come from the population consisting of Ksubpopulations:

$$\Pi_1, \Pi_2, ..., \Pi_K$$

and the distribution is a mixture of Kdistributions corresponding to respective subpopulations. Then the likelihood function is given as:

$$L(\theta \,|\mathbf{x}_1, , \mathbf{x}_2, , ..., \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i \,|\theta) =$$

$$= \prod_{i=1}^{n} \left(\sum_{j=1}^{K} P_j f_j(\mathbf{x}_i \,|\theta_j) \right)$$

Here instead of assigning observations to classes through indicator parameters (like in the classification approach) the other parameters are introduced. These are (prior) probabilities of belonging to class. If we assume that the number of parameters of the model for each class is equal to s, then the total number of parameters is equal to Ks+K-1.

It can be proved that maximum likelihood estimates in a mixture approach can be obtained through the following equations (after taking derivatives of log-likelihood function):

$$\hat{P}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{p}(j \,|\mathbf{x}_i)$$

$$\sum_{i=1}^{n} \hat{p}(j \,|\mathbf{x}_i) \; \nabla \hat{\theta}_j [\log f_j(\mathbf{x}_i \left| \hat{\theta}_j \right.)] = 0$$

$$\hat{p}(j\,|x_i) = \frac{\hat{P}_j f_j(\mathbf{x}_i \left| \hat{\theta}_j \right.)}{\sum\limits_{l=1}^{K} \hat{P}_l f_l(\mathbf{x}_i \left| \hat{\theta}_l \right.)}$$

As one can see, here posterior probabilities are introduced, showing for each observation the probability of belonging to each class, given prior probabilities and estimates of the parameters of the distribution for this class.

In practice, to find the maximum likelihood solutions, the following E-M type algorithm is used:

1. Start from initial posterior probabilities.
2. In each iteration:
   (a) Estimate parameters of the distribution for each class.
   (b) Calculate new posterior probabilities density for each observation given each class.
3. Iterate step 2 until posterior probabilities do not change significantly.

Of course, after completing estimation process, the classification can be easily obtained by assigning each observation to the class for which the posterior probability is the largest.

The particular variant of classification approach depends on the particular multivariate distribution being the model assumed for the class. As expected, the results were obtained in the case of multivariate normal distribution.

Banfield and Raftery (1993) discuss model-based clustering in the classification likelihood approach, where the underlying model for each class is multivariate normal distribution. They provide very useful (in terms of application and interpretation) framework that encompasses many other proposals given in the literature. This framework is based on the eigenvalue decomposition of covariance matrix, given as:

$$\mathbf{\Sigma} = \lambda \mathbf{DAD}^{\mathbf{T}}$$

where:

lambda coefficient reflects the size of class;

matrix D reflects the orientation of class;

matrix A reflects the shape of class.

In this case it is assumed that classes have (at least approximately) elliptical shapes.

2 Copula analysis – non-classical tool to analyze multivariate data

Most of the methods developed for multivariate statistical analysis, including model-based clustering, consider covariance matrix (or more generally, scatter matrix) as main "driving force" of analysis. Therefore it is assumed that most (or all) information about dependence between the components of the random vector is contained in covariance matrix.

We propose here an alternative approach, where instead of analyzing jointly scale parameters and dependence parameters, given in the covariance matrix, the analysis is performed separately for scale parameters (through univariate analysis), and for dependence parameters. This approach is based on the so-called copula analysis.

The main idea of copula analysis lies in the decomposition of the multivariate distribution into two components. The first component – it is marginal distributions. The second component is the function linking these marginal distributions to get a multivariate distribution. This function reflects the structure of the dependence between the components of the random vector. Therefore the analysis of multivariate distribution function is conducted by „separating" univariate distribution from the dependence.

This idea is reflected in Sklar theorem (Sklar (1959)), given as:

$$F(x_1,...,x_m) = C(F_1(x_1),...,F_n(x_m))$$

Where:

F – the multivariate distribution function;

F_i – the distribution function of the i-th marginal distribution;

C – copula function.

So copula function is simply the distribution function of the multivariate uniform distribution. The multivariate distribution function is given as the function of the univariate (marginal) distribution functions. This function is called copula function and it reflects the dependence between the univariate components.

The other notion strictly connected to copula function is copula density. It is given as:

$$c(u_1,...,u_m) = \partial^m C(u_1,...,u_m)$$

$$f(x_1,...,x_m) = c(F_1(x_1),...,F_m(x_m)) \cdot f_1(x_1) \cdot ... \cdot f_m(x_m)$$

Where:

f – the multivariate density function,

f_i – the univariate density function,

c– copula density.

As we can see, the information about the dependence contained in the multivariate distribution is „retained" in copula function. The advantage of copula analysis lies in the fact that it is suited for many possible shapes of classes, also non-elliptical shapes. The important feature is that for higher dimensions there is less parameters to be estimated, than in the classical (covariance-based) approach, since the number of parameters grows linearly with the growth of dimension of random vector.

Of course, one can get very many possible and suitable distributions, depending on the choice of marginal distributions and the choice of copula function. The classical approach of multivariate normal distribution can be put in the framework of copula analysis, by assuming univariate normal distribution as marginal distribution and choosing the so-called normal (Gaussian) copula, given through distribution functions of univariate and multivariate normal distribution as:

$$C(u_1, ..., u_m) = \Phi^m(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_m))$$

By applying normal copula to non-normal univariate marginal distributions we get different model. On the other hand, we can apply other than normal copula to univariate normal distributions. Of course, non-normal univariate distributions and non-normal copula function can be also combined. The detailed description of copula functions is given in Nelsen (1999) and Joe (1997).

The importance of copula analysis comes from its usefulness of modeling the dependence, not necessarily linear dependence. This is possible because of some properties of the copula function. The most important are the following properties:

- for independent variables we have:

$$C(u_1, ..., u_m) = C^{\perp}(u_1, ..., u_m) = u_1 u_2 ... u_m$$

- the lower limit for copula function is:

$$C^-(u_1, ..., u_m) = \max(u_1 + ... + u_m - m + 1; 0)$$

- the upper limit for copula function is:

$$C^+(u_1, ..., u_m) = \min(u_1, ..., u_m)$$

The lower and upper limits for the copula function have important consequences for the modeling of dependence. Suppose that we have two variables, X and Y, and there exists function (not necessarily a linear one), which links these two variables. We have the so-called total positive dependence between $X$and Y, when $Y = T(X)$ and $T$is the increasing function. Similarly, we have the so-called total negative dependence between $X$and Y, when $Y = T(X)$ and $T$is the decreasing function. Then:

- in the case of total positive dependence the following relation holds:

$$C(u_1, u_2) = C^+(u_1, u_2) = \min(u_1, u_2)$$

- in the case of total negative dependence the following relation holds:

$$C(u_1, u_2) = C^-(u_1, u_2) = \max(u_1 + u_2 - 1; 0)$$

This leads to the natural ordering of the multivariate distributions with respect to the strength and the direction of the dependence. It is given by the following order:

$$C_1(u_1, ..., u_m) \leq C_2(u_1, ..., u_m) \Rightarrow C_1 \prec C_2$$

and then we have:

$$C^{-} \prec C^{\neg} \prec C^{+}$$

The presented considerations are valid for any type of the dependence, including classical linear dependence.

Similarly as in the classical approach, the very important problem in copula approach is the estimation of the parameters of the model by maximum likelihood method. The log-likelihood function is given as:

$$l(\theta) = \sum_{i=1}^{n} \log c(F_1(x_{i1}), ..., F_m(x_{im})) + \sum_{i=1}^{n} \sum_{j=1}^{m} \log(f_j(x_{ij}))$$

This estimation is performed in two steps:

1. First step is maximum likelihood estimation of the parameters of the marginal distributions (for each j), thorough the maximization of the following function:
$$l(\theta_j) = \sum_{i=1}^{n} \log(f_j(x_{ij}))$$
2. Second step is maximum likelihood estimation of the parameters of copula function (given estimates obtained in the first step), through the maximization of the following function:
$$l(\alpha) = \sum_{i=1}^{n} \log c(F_1(x_{i1}), ..., F_m(x_{im}))$$

Clearly, such an approach may sometimes lead to local optimum problem.

3 Copula function in model-based clustering

In this section we propose to apply copula approach in model-based clustering. The main idea here is to replace classical, covariance based analysis, by more general approach, allowing for different shapes of classes, not necessarily elliptical one. As a model for each class we still consider multivariate distribution, however now it is represented as the linking of marginal distributions through copula function, like in Sklar theorem. We will consider here both stochastic approaches, namely classification approach and mixture approach

3.1 Copula function in classification model-based clustering

Here it is assumed that the observations

$$\mathbf{x}_1, , \mathbf{x}_2, , ..., \mathbf{x}_n$$

come from one of Ksubpopulations:

$$\Pi_1, \Pi_2, ..., \Pi_K$$

Then the likelihood function is given as:

$$L(\theta\,|\mathbf{x}_1, , \mathbf{x}_2, , ..., \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i\,|\theta) = \prod_{i=1}^{n} f_{\gamma_i}(\mathbf{x}_i\,|\theta_{\gamma_i})$$

$$\gamma_i = j \Leftrightarrow \mathbf{x}_i \in \Pi_j$$

And:

$$f_j(x_1, ..., x_m) = c_j(F_{j1}(x_1), ..., F_{jm}(x_m)) \cdot f_{j1}(x_1) \cdot ... \cdot f_{jm}(x_m)$$

To perform clustering in practice, an iterative algorithm should be used. This algorithm works in the same way as in the classical approach with exception that (in each iteration) maximum likelihood estimation is done separately for the parameter of marginal distributions and for copula parameters.

3.2 Copula function in mixture model-based clustering

Here it is assumed that the observations

$$\mathbf{x}_1, , \mathbf{x}_2, , ..., \mathbf{x}_n$$

come from the population consisting of Ksubpopulations:

$$\Pi_1, \Pi_2, ..., \Pi_K$$

and the distribution is a mixture of Kdistributions corresponding to respective subpopulations. Then the likelihood function is given as:

$$L(\theta\,|\mathbf{x}_1, , \mathbf{x}_2, , ..., \mathbf{x}_n) = \prod_{i=1}^{n} f(\mathbf{x}_i\,|\theta) =$$

$$= \prod_{i=1}^{n} \left(\sum_{j=1}^{K} P_j f_j(\mathbf{x}_i\,|\theta_j) \right)$$

And

$$f_j(x_1, ..., x_m) = c_j(F_{j1}(x_1), ..., F_{jm}(x_m)) \cdot f_{j1}(x_1) \cdot ... \cdot f_{jm}(x_m)$$

To perform clustering in practice, an iterative algorithm should be used. This algorithm works in the same way as in the classical approach with exception that (in each iteration) maximum likelihood estimation is done separately for the parameter of marginal distributions and for copula parameters.

Certainly, the new proposed approach needs extensive studies to show its theoretical quality and practical usefulness. Some simulation studies as well as the application for financial market data displayed encouraging results. Among the most important problems related to the proposed approach we should mention the following ones:

- avoiding local optima due to E-M type of iterative algorithms;
- evaluation of the performance of the approach comparing to classical approaches;
- choice of the appropriate copula among many possible copulas.

References

BANFIELD, J.D and RAFTERY, A.E. (1993): Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics, 49, 803–828.*

GORDON, A.D. (1999): *Classification. 2nd Edition.* Chapman and Hall, London.

JOE, H. (1997): *Multivariate Models and Dependence Concepts.* Chapman and Hall, London.

NELSEN, R.B. (1999): *An Introduction to Copulas.* Springer, New York.

SCOTT, A.J. and SYMONS, M.J. (1971): Clustering Methods Based on Likelihood Ratio Criteria. *Biometrics, 27, 387–397.*

SKLAR, A. (1959): Fonctions de repartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris, 8, 229–231.*

WOLFE, J.H. (1970): Pattern Clustering by Multivariate Mixture Analysis. *Multivariate Behavioral Research, 5, 329–350.*

Three–Way Multidimensional Scaling: Formal Properties and Relationships Between Scaling Methods

Sabine Krolak–Schwerdt

Department of Psychology, Saarland University, D-66041 Saarbrücken, Germany

Abstract. This paper is concerned with methods of three-way two-mode multidimensional scaling which were developed for the joint analysis of a number of promities matrices. The classification of these methods into trilinear and quadrilinear models (Kruskal (1983)) is outlined, and it is shown, that a number of specific properties and interpretations are associated with this classification the methods within each class have in common. Finally, relationships of the methods within and between the two model classes are outlined.

1 Introduction

Models and methods of three–way two–mode multidimensional scaling were developed for the analysis of individual differences in the representation of proximities. In the following, the scalar product form of the models will be outlined. The input data consist of a three–way data matrix $X = (x_{ijj'})$, $i = 1, \ldots, I$, $j, j' = 1, \ldots, J$, that can be thought of as comprising a set of $I (\geq 2) J \times J$ scalar products matrices. X_i, a slice of the three–way matrix, consists of estimated scalar products between objects j, j' and they represent the co–occurrence of objects under the point–of–view of an individual or a condition i. Obtained proximities estimated by a specific individual i may be transformed into scalar products by the procedures described by Torgerson (1958).

Three–way scaling models may be classified into those which rest on a trilinear or a quadrilinear decomposition of the data (Kruskal (1983)). Associated with the type of decomposition are different assumptions on the nature of individual differences.

2 Trilinear scaling methods

Trilinear approaches assume that there is a common space which underlies the objects in general. Thus, different subjects are presumed to perceive or judge stimuli on common sets of dimensions. On the basis of the common object space, it is assumed that individuals may distort the object space in their perception by attaching different importance or weights to the object dimensions than others do. The basic equation is given by

$$x_{ijj'} = \sum_{f=1}^{F} a_{if} b_{jf} b_{j'f} + e_{ijj'} \tag{1}$$

where F denotes the number of dimensions, or, equivalently,

$$X = A_F I(B'_F \otimes B'_F) + E \tag{2}$$

where I denotes the three–way identity matrix, A_F is a $I \times F$ matrix with the weights or saliences of the subjects on the F dimensions, reflecting the subject space, and B_F is a $J \times F$ matrix representing the common object space. The $e_{ijj'}$ are the errors of approximation collected in the three–way matrix E and $\otimes$ denotes the Kronecker product.

A very compact representation of trilinear models is obtained if the values of a subject i in the subject dimensions are arranged in a diagonal $F \times F$ matrix A_i. This yields the expression

$$X_i = B_F A_i B'_F + E_i$$

where A_i contains the dimensional weights under the point–of–view of subject i. Applying the individual weights to the common object space B_F yields the object space specific to the individual, $Y_i = B_F A_i^{\frac{1}{2}}$.

The most prominent trilinear model is the INDSCAL approach proposed by Carroll and Chang (1970). To fit the model to a given three–way data matrix, an alternating–least–squares algorithm is used.

Recently, two methods termed "SUMM–ID 1" and "SUMM–ID 2" (Krolak-Schwerdt (in press); Krolak-Schwerdt (1991)) were introduced where SUMM–ID 1 rests on the same model equation as INDSCAL, but uses another method to decompose the data matrix according to Equation (1). Basically, the method uses a three–way generalization of the centroid approach such that dimensions correspond to the centroids of the data. The central feature of the method is to base the object dimension b_f on the introduction of sign vectors z_f for the objects j, $z_{jf} \in \{-1, 1\}$, and, in an analogous way, to base the subject dimension a_f on sign vectors s_f for individuals i, $s_{if} \in \{-1, 1\}$, where

$$\sum_i \sum_j \sum_{j'} s_{if} z_{jf} z_{j'f} x_{ijj'} = \gamma_f := max \ . \tag{3}$$

The vectors z_f and s_f are the basis for the determination of dimension a_f and b_f in the following way:

$$\begin{array}{llll} a_{if} = \frac{1}{\sqrt[3]{\gamma_f}^2} & u_{if} \ , & \text{where} & u_{if} = \sum_j \sum_{j'} z_{jf} z_{j'f} x_{ijj'} \ , \\ b_{jf} = \frac{1}{\sqrt[3]{\gamma_f}^2} & q_{jf} \ , & \text{where} & q_{jf} = \sum_i \sum_{j'} s_{if} z_{j'f} x_{ijj'} \ , \\ & & \text{and} & \gamma_f = \sum_i s_{if} u_{if} = \sum_j z_{jf} q_{jf} \ . \end{array} \tag{4}$$

In the equations, the scalar γ_f is a normalizing factor. After the determination of a_f and b_f in this way, the procedure continues in computing the residual data $x^*_{ijj'} = x_{ijj'} - a_{if} b_{jf} b_{j'f}$ and repeating the extraction of dimensions according to (3) and (4) on the residual data until a sufficient amount of the variation in the data is accounted for by the representation.

As Equation (3) shows, the values of the sign vectors z_f and s_f are determined in such a way that the sum of the data values collected in a dimension will be maximized. To compute the sign vectors, an algorithm is used that alternates between the subject and the object mode. That is, given the sign vector of one mode, multiplying the data matrix (the residual data matrix, respectively) with this sign vector yields the signs of the other vector. This is due to the cyclical relation between the modes

$$\begin{aligned}\textstyle\sum_i \sum_j \sum_{j'} s_{if} z_{jf} z_{j'f} x_{ijj'} &= \textstyle\sum_i s_{if} u_{if} = \sum_i \mid u_{if} \mid \\ &= \textstyle\sum_j z_{jf} q_{jf} = \sum_j \mid q_{jf} \mid = \gamma_f \ ,\end{aligned}$$

which also motivates the definition of the normalizing factor γ_f in Equation (4). The relation between the modes is used in the SUMM–ID algorithm by iteratively fixating one vector and estimating the other vector until the values of both sign vectors have stabilized. The procedure was first introduced by Orlik (1980).

3 Quadrilinear scaling models

Quadrilinear models share the assumption of the common object space with trilinear approaches. Furthermore, they assume that individual representations of the object space are distortions of the common object space. However, these distortions are more complex than in trilinear models. Besides the introduction of differential weights attached to the object dimensions, it is assumed that individual representations may be rotated versions of the common object space in which independent dimensions may become correlated.

The basic equation of the approaches can be expressed as

$$x_{ijj'} = \sum_{m=1}^{M} \sum_{p=1}^{P} \sum_{p'=1}^{P} a_{im} b_{jp} b_{j'p'} g_{mpp'} + e_{ijj'}. \tag{5}$$

A matrix formulation of the model is

$$X = AG(B' \otimes B') + E. \tag{6}$$

The coefficients a_{im} and b_{jp} are the elements of matrices A and B, the $g_{mpp'}$ are the elements of a three–way core matrix G (cf. Tucker (1972)). A is a $I \times M$ matrix with the coefficients of the subjects on M dimensions. B is a $J \times P$ matrix specifying an object space which is common to all subjects.

A more compact matrix representation is

$$X_i = BH_iB' + E_i \tag{7}$$

where the H_i, termed 'individual characteristic matrix' (cf. Tucker (1972)), is a linear combination of the M frontal slices, G_m, of the core matrix

$$H_i = \sum_{m=1}^{M} a_{im} G_m \tag{8}$$

Thus, H_i is a $P \times P$ symmetric matrix designating the nature of individual $i's$ distortion of the object dimensions. Diagonal elements h_{ppi} of H_i correspond to weights applied to the object dimensions by individual i, while off–diagonal elements $h_{pp'i}$ indicate the perceived relationships among the object dimensions p and p' under the point–of–view of individual i. As Equation (7) shows, the matrix H_i transforms the common object space into the individual representation.

The most prominent model of this type is the Tucker (1972) model which uses normalized principal components of scalar products of the data to derive the dimensions both in the subject space and in the object space. More precisely, writing the three–way data matrix as an ordinary two–way matrix $X_{(I)}$, $X_{(I)} \in R^{I \times JJ}$, by making use of combination modes (Tucker (1966)), A consists of the eigenvectors of $X_{(I)}X'_{(I)}$. In an analogous way, B consists of the eigenvectors of $X_{(J)}X'_{(J)}$, $X_{(J)} \in R^{J \times IJ}$. Kroonenberg and De Leeuw (1980) developed an alternating–least–squares algorithm to fit the Tucker model to a given three–way data matrix.

SUMM–ID 2 also rests on Equation (5). The model derives from its trilinear counterpart SUMM–ID 1 in the following way. It involves a rotation of the common object space

$$B = B_F T \ , \tag{9}$$

where the orthonormal transformation matrix T evolves from the singular-value decomposition $B_F = K\Delta_B^{\frac{1}{2}}T'$.

Analogously, the subject space is rotated

$$A = A_F V \tag{10}$$

with the orthonormal transformation matrix V deriving from the singular–value decomposition $A_F = L\Delta_A^{\frac{1}{2}}V'$. Inserting the rotated matrices into the model equation from SUMM–ID 1, that is

$$\begin{array}{rlccl} X & = A_F & I & (B'_F \otimes B'_F) & +E \\ & = A & V'I(T \otimes T) & (B' \otimes B') & +E \\ & = A & G & (B' \otimes B') & +E \end{array}$$

yields the final representation that was introduced in Equation (6).

As a very general approach, IDIOSCAL (Carroll and Chang (1970), Carroll and Wish (1974)) introduces a symmetric positive definite matrix C_i into the model equation $X_i = B\ C_i B'$ in order to allow for idiosyncratic rotations of the object space for different individuals. The central question within the model is how to decompose C_i. Two different ways of decomposing C_i have been proposed (cf. Carroll and Wish (1974)).

One procedure is given by $C_i = T_i \Delta_i T'_i$, with T_i orthonormal and Δ_i a diagonal matrix. Geometrically, the decomposition consists of an orthogonal rotation to a new coordinate system for subject i and a rescaling or weighting of dimensions of this new coordinate system by the diagonal entries of Δ_i. By this procedure, IDIOSCAL becomes equivalent to the SUMM–ID 2 formulation with the extra constraint in IDIOSCAL that individual spaces for subjects consist of orthogonal object dimensions. The second procedure is the one introduced by Tucker (1972). In this case, IDIOSCAL becomes an identical account to the Tucker model (cf. Kroonenberg (1994), Carroll and Wish (1974)).

4 Properties of trilinear and quadrilinear models

Associated with the classification into tri- and quadrilinear decompositions are specific properties the models within each class have in common. Perhaps the most important feature of trilinear models which is common to INDSCAL and SUMM–ID 1 is the 'intrinsic axis property' (Kruskal (1983)). That is, the dimensions of the object space B_F and of the subject space A_F are uniquely determined up to a joint permutation of the columns of the two matrices, and up to a scaling of the columns of the two matrices. Thus, both accounts provide solutions where the rotational position of dimensions is fixed.

Another special feature of trilinear models is that the number of dimensions in the subject space and the object space must be the same. In other words, one set of dimensions is extracted from the data matrix which then specifies the dimensionality F in both spaces.

Furthermore, both in INDSCAL and in SUMM–ID 1 the dimensionality of A_F and B_F is very high. For data forming a three–way array of size $I \times J \times J$ the number of dimensions necessary to reproduce the data is larger than $min(I, J)$ (cf. Kruskal (1976, 1983)). This implies that the dimensions are generally oblique and may become even linear dependent.

With respect to formal characteristics, quadrilinear models are different in nature. As has already been stated, these models introduce additional parameters by means of the core matrix G which have to be estimated along with the other unknown parameters in A and B. Thus, these models are more general formulations than trilinear approaches. As a consequence of using singular–value–decomposition either to extract dimensions or to rotate spaces, quadrilinear models retain rank properties of one–mode methods. That is, the number of dimensions does not exceed the number of subjects or objects and dimensions are always orthogonal.

Furthermore, as another consequence of the introduction of the core matrix, both the subject space and the object space are subject to rotations. Specifically, postmultiplication of the matrices A and B by orthonormal transformation matrices does not affect the model estimates provided that the core matrix is counter–rotated. Thus, these models are not uniquely identified. Another property due to the introduction of the core matrix is that the dimensionality M of the subject space may differ from the one of the object space which is P.

5 Connections between three–way scaling methods

As has been outlined above, trilinear and quadrilinear methods have very different properties. Thus, connections between these models have been found under rather restrictive conditions only. That is, the Tucker model becomes algebraically equivalent to INDSCAL if the core matrix can be diagonalized (Kroonenberg (1983), Carroll and Wish (1974)). However, there exist more general relations in that the quadrilinear models can be derived from trilinear methods by the SUMM–ID approach even in the case of unconstrained core matrices. These model connections will be outlined in the following.

SUMM–ID has some very close model relations to INDSCAL and the Tucker approach. The relationship to INDSCAL refers to SUMM–ID 1 and will be presented first in the following. Subsequently, the connections of SUMM–ID 2 to the Tucker model will be outlined.

To present the relationship between SUMM–ID 1 and INDSCAL, the triple product of the matrices A_F and B_F will be denoted in a more compact manner (cf. Kruskal (1976)) by $[A_F, B_F, B_F]$, that is $A_F I(B'_F \otimes B'_F) = [A_F, B_F, B_F]$. In the following, the SUMM–ID 1 will be termed $[A_F, B_F, B_F]$ and the INDSCAL representation will be referred to by $[\overline{A}, \overline{B}, \overline{B}]$. Theorem 1 states the relationship between the two models (cf. Krolak-Schwerdt (1991)).

Theorem 1. Suppose that $X = [\overline{A}, \overline{B}, \overline{B}]$ is of minimum rank R such that $R = rank(X)$. Then $R = F$, and there exist two permutation matrices P_A and P_B and two non–singular diagonal matrices D_A and D_B such that

$$\overline{A} = A_F P_A D_A \quad \text{and} \quad \overline{B} = B_F P_B D_B \quad \text{where}$$

$$[P_A D_A, P_B D_B, P_B D_B] = I.$$

As Theorem 1 states, the number of dimensions F of the SUMM–ID 1 representation is the same as the dimensionality of the INDSCAL representation. Furthermore, from the SUMM–ID 1 configuration the INDSCAL dimensions in both the object and the subject space may be derived by simply permuting and rescaling the dimensions where the effects of permutation and rescaling in the two modes compensate each other.

If the standard procedure of INDSCAL to scale the dimensions (that is, to normalize the object space so that the variances of projections of objects on the several dimensions are equal to one and to compensate the normalization in the object space by multiplying the weights in the subject space by the reciprocal scaling factors) is applied to the SUMM–ID 1 representation in this way, then SUMM–ID 1 and INDSCAL simply differ in the sequence in which the dimensions are extracted. Thus, both accounts create equivalent representations although they use a very different rationale and method of analysis. Stated in other words, INDSCAL may be derived from SUMM–ID 1 by two classes of transformations, permutation and rescaling of dimensions.

If the classes of permissible transformations on SUMM–ID are extended to orthogonal rotations of the subject and the object space, then the Tucker (1972) model derives from SUMM–ID 2 as will be shown in the following.

The Tucker model will be denoted by $X = \overline{A}\,\overline{G}(\overline{B'} \otimes \overline{B'})$. Writing the three–way data matrix as an ordinary two–way matrix $X_{(I)}$, $X_{(I)} \in R^{I\times JJ}$, $\overline{A}$ thus consists of the eigenvectors of $X_{(I)}X'_{(I)}$ and $\overline{B}$ consists of the eigenvectors of $X_{(J)}X'_{(J)}$, $X_{(J)} \in R^{J\times IJ}$ (cf. Section 3).

For the SUMM–ID 2 model $X = AG(B' \otimes B')$, the three–way core matrix G will be represented as an ordinary two–way matrix in the two different forms $G_{(M)}$, $G_{(M)} \in R^{M\times PP}$, and $G_{(P)}$, $G_{(P)} \in R^{P\times MP}$.

To present the relationship between SUMM–ID 2 and the Tucker model, the SUMM–ID 2 representation is rescaled according to the reasoning of the Tucker model. That is, the object space matrix $B = K\Delta_B^{\frac{1}{2}}$ as well as the subject space matrix $A = L\Delta_A^{\frac{1}{2}}$ are normalized so that the variances of projections of objects (of subjects, respectively) on the dimensions are equal to one. This yields K in the object space and L in the subject space. To compensate the normalization, the core matrix is rescaled in a reciprocal manner, that is, $\Delta_A^{\frac{1}{2}}G_{(M)}(\Delta_B^{\frac{1}{2}} \otimes \Delta_B^{\frac{1}{2}}) := Z_{(M)}$ or $\Delta_B^{\frac{1}{2}}G_{(P)}(\Delta_A^{\frac{1}{2}} \otimes \Delta_B^{\frac{1}{2}}) := Z_{(P)}$. With this normalization, Theorem 2 states the relationship between SUMM–ID 2 and the Tucker model (cf. Krolak-Schwerdt (1991)).

Theorem 2. Suppose W_A consists of the eigenvectors of $Z_{(M)}Z'_{(M)}$ and W_B consists of the eigenvectors of $Z_{(P)}Z'_{(P)}$. Then

$$\overline{A} = L\,W_A\ , \quad \overline{B} = K\,W_B \quad \text{and}$$

$$\overline{G}_{(M)} = W'_A\,Z_{(M)}\,(W_B \otimes W_B).$$

As Theorem 2 states, the Tucker model derives from SUMM–ID 2 by two transformations which consist of rescaling the SUMM–ID configuration and subsequently rotating the configuration orthogonally. The rescaling step involves normalizing the length of dimensions in the subject and in the object space as well as weighting the elements of the core matrix with the corresponding variances of the dimensions. The eigenvectors of the weighted core matrix constitute orthogonal rotation matrices which map the SUMM–ID 2 dimensions onto the Tucker configuration in both modes. Furthermore, the Tucker core matrix is obtained from orthogonally counter–rotating the rescaled SUMM–ID 2 core matrix.

6 Concluding remarks

Although differing in formal properties, there exist some general relationships between trilinear and quadrilinear models for three–way data analysis. The latter methods can be derived from the trilinear class without restricting the core matrix to a diagonal form.

In sum, SUMM–ID appears as a unifying account which establishes these connections and which may easily emulate other three–way multidimensional scaling representations, each by two transformations. To derive INDSCAL, the SUMM–ID 1 representation must be rescaled and dimensions must be permutated. To obtain the Tucker (1972) model, the rescaling of the SUMM–ID 2 representation must be followed by orthogonal rotations of the configuration. From this it is directly evident, that IDIOSCAL may be obtained from SUMM–ID 2.

Finally, it should be noted on more practical grounds that the estimates from SUMM–ID are computationally quite efficient and fast to obtain in terms of CPU time as compared to alternating–least–squares algorithms.

Another aspect of importance in applied research concerns the interpretation of the core matrices. Within the Tucker model the eigenvalues of dimensions obtained in both modes are connected with the core matrix which makes the interpretations of the values of the core matrix rather difficult (cf. Kroonenberg (1983)). In contrast, the core values in SUMM–ID 2 were defined as

$$g_{mpp'} = v'_m I(t_p \otimes t_{p'}) = [v_m t_p t_{p'}]$$

with unit length vectors v_m, t_p and $t_{p'}$. This corresponds to the three–way generalization of the standard scalar product. Thus, in SUMM–ID 2 values of the core matrix have a much more specific meaning in terms of interrelations between dimensions which makes the model parameters more easy to interpret.

References

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

KROLAK-SCHWERDT, S. (1991): Modelle der dreimodalen Faktorenanalyse: formale Eigenschaften, theoretische Zusammenhänge und ihre Implikationen für das Konzept individueller Differenzen. *Psychologische Beiträge, 133, 314–346.*

KROLAK-SCHWERDT, S. (in press): A three-way multidimensional scaling approach to the analysis of judgments about persons. *Studies in Classification, Data Analysis, and Knowledge Organisation, Vol. 27 Classification: The ubiqitous challenge.* Springer, Berlin.

KROONENBERG, P.M. (1983): *Three–mode principal component analysis.* DSWO Press, Leiden.

KROONENBERG, P.M. (1994): The TUCKALS line. A suite of programs for three–way data analysis. *Computational statistics and data analysis, 18, 73–96.*

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

KRUSKAL, J.B. (1976): More factors than subjects, tests and treatments: An indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika, 41, 281–293.*

KRUSKAL, J.B. (1983): Multilinear methods. In R. Gnanadesikan (Ed.): *Proceedings of Symposia in Applied Mathematics* (Vol. 28, pp. 75-104). Providence: American Mathematical Society.

ORLIK, P. (1980): Das SUMMAX–Modell der dreimodalen Faktorenanalyse mit interpretierbarer Kernmatrix. *Archiv für Psychologie, 133, 189–218.*

TORGERSON, W.S. (1958): *Theory and methods of scaling.* Wiley, New York.

TUCKER, L.R. (1966): Some mathematical notes on three-mode factor analysis. *Psychometrika, 31, 279–311.*

TUCKER, L.R. (1972): Relations between multidimensional scaling and three–mode factor analysis. *Psychometrika, 37, 3–27.*

Empirical Approach as a Scientific Framework for Data Analysis

Shizuhiko Nishisato

The Ontario Institute for Studies in Education, University of Toronto,
252 Bloor Street West, Toronto, Ontario, Canada M5S 1V6
email: snishisato@oise.utoronto.ca

Abstract. The traditional statistical procedure is typically based on the notion that data are a random sample from the normal population. Knowing that the bivariate normal distribution has only three parameters, mean, variance and linear correlation, the use of the normal distribution as an analytical framework leads to what we call linear analysis. This paper starts with discarding the normal distribution assumption, and then advocates total reliance on data in hand. In the social sciences, we face the majority of data to be categorical. To explain the data exhaustively, it is almost necessary to employ an empirical approach without any prior assumptions such as the distribution, levels of measurement or models for data. This framework opens up the possibility of capturing nonlinear information and multidimensionality in data, as well as of cautionary wisdom for the popular data reduction approach to data analysis.

1 Introduction

The main statistical procedure today is based on the assumption that data are a random sample from the normal population. When this setup is reasonable, there exists a powerful inferential procedure of statistics available for data analysis. However, as we know very well, the above setup is not always appropriate for data in hand or for what we want to do in data analysis.

To lay the basis for our discussion, we should be re minded that the assumption of the normal distribution has very important implications for data analysis. First, data must be continuous. This restriction excludes a large number of data sets we face in the social sciences. Second, the multivariate normal distribution means that the relation between any variables is linear. Thus, the adoption of the normal distribution provides an exclusive situation in which an elegant procedure via Pearsonian correlation suffices to investigate inter-variable relations exhaustively. In other words, so long as the multivariate normal distribution is assumed in analysis, we must ignore the possibility of nonlinear relations between variables.

Since there exists a widespread belief that normally distributed variables can be found everywhere or variables can be easily transformed into normal distributions, it seems that the plea for nonlinear analysis often receives little or no attention. Or, rather the fact that the bivariate normal distribution does not contain a nonlinear-relation parameter is totally forgotten. To open up our eyes to the exhaustive explanation of data as an object of data analysis, we would like to demonstrate that data we typically analyze contain a substantial amount of nonlinear relations and that the linear restriction, imposed by the normal distribution assumption, can be detrimental to the understanding of data in hand.

2 Correlation: Empirically speaking

Pearsonian correlation between variables X and Y, say, is 1 if $Y = aX + b$ for some non-zero value of a. In other words, when we plot two variables with the variables as axes, the graph must show a straight line to attain the correlation of 1. In the social sciences, we often assign integers 1, 2, 3, 4 and 5, for example, to the five categories of an ordered categorical variable (e.g., never, rarely, sometimes, often, always). For two such variables to attain perfect correlation, the intervals between these assigned numbers must be in some sense accurate. If the responses to these variables are summarized in a 5x5 joint frequency table (i.e., contingency table), the product-moment correlation of 1 is attained only when all the responses fall in the main diagonal positions. Any other distributions of responses in the twenty five cells are departures from the linear relation, thus reducing the correlation to a smaller value than 1.

In the social sciences, there have been a number of discussions on the importance of the assumption of bivariate normality in using Pearsonian correlation. The general argument has been that one can calculate Pearsonian correlation without any assumption of normality, but the normal distribution assumption is necessary for significance testing of the correlation. Such argument, however, seems to ignore a more important question of what Pearsonian correlation represents.

Pearsonian correlation is an indicator of linear relationship between two variables. If we express the two variables in the orthogonal coordinate system, each variable can be represented as an axis. The value of correlation then is the cosine of the angle between the two axes, representing the variables. As we noted earlier, Pearsonian correlation captures the entire linear relation between two normally distributed variables. But, when variables are nonlinearly related, Pearsonian correlation becomes an uninterpretable statistic. Or, at best, it reflects the linearity of the relationship between two variables, without specifying what nonlinear relation might be involved.

When the bivariate normal distribution assumption holds, Pearsonian correlation between X and Y of 0.3, for example, tells us a substantial number of responses do not lie on a straight line if the graph is constructed with X and Y as the two axes. In the orthogonal coordinate system, the correlation of 0.3 indicates that the angle between the two variables is about 73 degrees.

Thus, the message here is that under the bivariate normal distribution the value of Pearsonian correlation has a definite, thus interpretable, meaning.

When the variables are nonlinearly related, however, Pearsonian correlation of 0.3 may mean a great number of nonlinear relations such as polynomial relations of different degrees or wide dispersions of responses. In the case of the 5x5 contingency tables, Pearsonian correlation of 0.3 can be generated by a wide variety of distributions of responses over 25 cells, many of which are indicative of nonlinear relations. In other words, when the distribution is not bivariate normal, that is, when there exists nonlinear relation between variables, the correlation of 0.3 does not tell us anything about the actual relation.

3 Numerical example

With this much discussion, we can draw a conclusion that multidimensional decomposition of a Pearsonian correlation matrix is far from being called data analysis if nonlinear relations are involved in data. Factor analysis, for example, is then reduced to the analysis of only linear relations in data, ignoring probably abandunt nonlinear information in the data. This conclusion suggests that if we were to talk about 'data analysis,' we need an alternative procedure that also taps into nonlinear relations, that is, a procedure that analyzes a data set itself irrespective of intervariable relations being linear or nonlinear.

Let us use a numerical example to make a point. The following six questions and the corresponding data have been used in several papers and are from Nishisato (2000).

1. How would you rate your blood pressure?...(Low, Medium, High): coded 1, 2, 3
2. Do you get migraines?...(Rarely, Sometimes, Often): 1, 2, 3 (as above)
3. What is your age group?...(20-34; 35-49; 50-65): 1, 2, 3
4. How would you rate your daily level of anxiety?...(Low, Medium, High): 1, 2, 3
5. How would you rate your weight?...(Light, Medium, Heavy): 1, 2, 3
6. What about your height?...(Short, Medium, Tall): 1, 2, 3

The data collected from 15 subjects (see Nishisato (2000)) can be presented in two distinct formats:
Format 1. Assign 1, 2 and 3 to the three ordered categories of each item, that is, the so-called Likert-type scores (Likert (1932)). The data matrix is 15x6.
Format 2. Represent the data in the form of (1,0) response-patterns of each item, that is, (100) if the subject chooses the first option, (010) for the choice of the second option, and (001) for the choice of the third option of each item. The data matrix is 15x12.

BP	1.00	-.06	.66	.18	.17	-.21
Mig	-.06	1.00	.23	.21	-.58	.10
Age	.66	.23	1.00	.22	-.02	-.30
Anx	.18	.21	.22	1.00	.26	-.23
Wgt	.17	-.58	-.02	.26	1.00	-.31
Hgt	-.21	.10	-.30	-.23	-.31	1.00

Table 1. Correlation Matrix of Likert-Type Scores

Using Format 1 and treating the data as continuous, calculate the matrix of Pearsonian correlation (Table 1).

Notice something quite revealing. Take a look at the correlation between blood pressure and age (0.66) and the correlation between blood pressure and migraines (-0.06), together with the corresponding data expressed in the form of contingency tables (Table 2).

	Age				Migraine		
Blood	20-34	35-49	50-65	Blood	Rarely	Sometimes	Often
High BP	0	0	4	High BP	0	0	4
Mid BP	1	4	1	Mid BP	3	3	0
Low BP	3	1	1	Low BP	0	0	5

Table 2. Blood Pressure, Migraines and Age

We can immediately tell that blood pressure is correlated linearly to age, hence the correlation of 0.66, while blood pressure is not linearly correlated to migraines, thus the correlation of -0.06. It is clear, however, that blood pressure is nonlinearly related to migraines in such a way that frequent migraines occur when blood pressure is either low or high. Furthermore, we can see that this data set contains a substantial amount of nonlinear information (see Nishisato (2000)). Thus, if we subject the above correlation matrix to factor analysis, the results would be based on only a small amount of linearity in the data, and thus are not appropriate to be called comprehensive analysis of this data set. We may go on to ask if in this situation factor analysis is meaningful at all, or if we should even try to interpret the results. We should ask if it is meaningful to factor analyze practically non-existent linearity between blood pressure and migraines.

We also wonder if the scores 1, 2 and 3, assigned to the three ordered categories, were appropriate. To answer this question, we can determine the three scores of each variable, subject to the weak order constraint on them, that is, we can adjust the interval between 1 and 2 and that between 2 and 3, while maintaining the ordinal information of the categories. The correlation matrix obtained under the weak order constraint is given in table 3.

BP	1.00	.49	.74	.40	.30	.21
Mig	.49	1.00	.39	.40	.33	.29
Age	.74	.39	1.00	.54	.08	.29
Anx	.40	.40	.54	1.00	-.36	.26
Wgt	.30	.33	.08	-.36	1.00	.30
Hgt	.21	.29	.29	.26	.30	1.00

Table 3. Correlation Matrix of Monotone Regression

The correlation between blood pressure and migraines is now 0.49. However, when we examined the 3x3 contingency table, we saw that the two variables were almost perfectly correlated in a nonlinear fashion. Thus, factor analysis of the above correlation matrix again would not qualify as a method of data analysis, since it still leaves a lot of nonlinear information in data out of analysis.

4 Empirical approach to correlation

Without any assumptions on the scores given to three categories of each item, we can determine the scores in the light of the data. This is an empirical approach. Using Format 2, we may then subject the 15x12 data to quantification, the procedure being called by many different names such as dual scaling and multiple correspondence analysis. Here we place our entire trust in the data in hand. We may typically know how data are generated and we may therefore say that the rejection of our prior knowledge is disadvantageous. But, one may also argue against such a belief as it seems often the case that our 'knowledge' is hindrance for discerning the detailed information inbedded in data. Dual scaling of the response-patterns yields twelve correlation matrices, corresponding to the twelve components. In quantification theory, we determine a nonlinear transformation of each set of categories of a variable in such a way that the average of Pearsonian correlation of $n(n-1)/2$ pairs of n categorical variables be a maximum. This procedure typically yields (m-n) components, where m is the total number of categories of n variables. The correlation matrix based on the first component is summarized in Table 5.

BP	1.00	.99	.60	.47	.43	.56
Mig	.99	1.00	.58	.52	.39	.57
Age	.60	.58	1.00	.67	.08	.13
Anx	.47	.52	.67	1.00	-.33	.19
Wgt	.43	.39	.08	-.33	1.00	.20
Hgt	.56	.57	.13	.19	.20	1.00

Table 4. First Correlation Matrix from Dual Scaling

Notice that the correlation between blood pressure and migraines is now 0.99. What is certain now is that this approach captures whatever relations, linear or nonlinear, in the data without any computational difficulty. This correlation matrix, however, is only one of the twelve (i.e., 18-6) matrices, and as such it does not represent all the information in data, which we seek to analyze.

5 Multidimensionality and correlation

Nishisato (2004) provided a framework for categorical variables. A variable with three categories is given coordinates (1,0,0), (0,1,0) and (0,0,1), corresponding to the three possible choices. The variable can be therefore represented as a triangle if we connect the three possible responses. Once data are collected, those 1's in the three coordinates are replaced with respective response frequencies, and in quantification theory the final positions of the three vertices are determined in such a way that the distance between the centroid to each vertex times the frequency of the vertex is equal to a prespecified constant and this holds true for all three vertices. Thus, each variable with three categories occupy two-dimensional space, and since typically variables are not perfectly correlated to each other the six variables with three categories in each in our example occupy twelve-dimensional space. In other words, six triangles are floating in twelve-dimensional space with the same centroid but different orientations. Nishisato (2004) thought that correlation between two variables could be conceived as the projection of one triangle onto the other triangle. However, he retreated this concept when Greenacre (2004) presented a counter example.

In quantification theory, it is known that the 3x3 contingency table typically yields two non-trivial eigenvalues. Using forced classification of dual scaling (Nishisato (1984), Nishisato and Gaul (1990), Nishisato and Baba (1999)), Nishisato (2005) has shown the following: the sum of squared correlations between one categorical variable q and the other variable q' when the latter is projected onto the first variable, indicated by A, is equal to that when the former is projected to the latter, and it is also equal to the sum of non-trivial eigenvalues from the contingency table of the two variables, indicated by B. In terms of formulas,

$$A = \sum_{k=1}^{J_q-1} r^2_{q't(k)} = \sum_{k'=1}^{J_{q'}-1} r^2_{qt(k')} \tag{1}$$

where k means the k-th component, J_q and $J_{q'}$ are the number of categories of variable q and that of variable q', respectively, $r^2_{q't(k)}$ and $r^2_{qt(k')}$ are the squared item-component (i.e., k-th component) correlations of the non-criterion variables associated with the proper forced classification solutions,

$$B = \sum_{k=1}^{p-1} \lambda_k \tag{2}$$

where p is the smaller number of categories of the two variables, that is, $\min(J_q, J_{q'})$, and λ_k is the eigenvalue of the k-th nontrivial component of the contingency table. He defined the correlation between two categorical variables ν as the positive square root of the average of A or that of B,

$$\nu = \sqrt{\frac{A}{p-1}} = \sqrt{\frac{B}{p-1}} \tag{3}$$

He has also shown that his proposed coefficient is equal to Cramér's coefficient of association V (Cramér (1946)), given by

$$V = \sqrt{\frac{\chi^2}{f_t(p-1)}} = \nu \tag{4}$$

where f_t is the total number of responses. Using this measure, we obtain Table 5, which includes linear and nonlinear relations, that is, the entire information in data, calculated in multidimensional space.

BP	1.00	.71	.63	.44	.37	.46
Mig	.71	1.00	.45	.56	.50	.45
Age	.63	.45	1.00	.55	.40	.25
Anx	.44	.56	.55	1.00	.31	.20
Wgt	.37	.50	.40	.31	1.00	.40
Hgt	.46	.45	.25	.20	.40	1.00

Table 5. Matrix of Cramér's V

Notice that the sum of both linear and nonlinear correlation in the above table tends to be smaller than the corresponding value from the first component of dual scaling (Table 4). Why?

6 Warnings on dimension reduction

When we consider all variables together, the fact that our sample data yield twelve correlation matrices presents an interesting problem to consider. We often talk about dimension reduction and employ a practice to interpret data in terms of a few major components. Our numerical example indicates a serious pitfall of the practice for data analysis.

Recall that two continuous variables can be represented as axes in the orthogonal coordinate system and that the correlation is defined as the cosine of the angle between two variables. If two variables span in multidimensional space as a typical case, however, the angle between two axes can never become smaller in reduced space. For instance, the angle of two vectors in three dimensional space is likely to become smaller if projected onto two dimensional space, thus increasing its cosine value, that is, correlation. In other words, the correlation becomes larger in smaller space than in original multidimensional space. Recall that our sample data occupy 12-dimensional space. Yet, we typically interpret the data in two-dimensional space. The above argument indicates that in looking at data in two dimensions we are exaggerating cohesion of the data clouds, thus adopting an easier framework for interpreting data. This practice seems to tell us that we are interpreting data by not strictly adhering to data structure. With inflated correlation, the variables look more tightly clustered, leading to a description of data with a simpler structure than the data structure itself. What we are then interpreting may be a conveniently distorted structure. If this should be the case, the problem is very serious: we are interpreting non-existent or overly simplified data configuration. This leads to the conclusion that interpreting data in reduced space involves a danger of over-evaluating correlation, hence easier interpretation of the relations between variables. A better alternative to the interpretation of data via dimension reduction, therefore, seems to be to calculate inter-variable distances in multidimensional space and analyze those distances into components. This alternative will identify data structure which is closer to data themselves than dimension reduction will. This problem, however, will be left for further investigation.

7 Concluding remarks

There are several messages from this paper:
(1) The assumption of the normal distribution precludes nonlinear relations in data. Most procedures based on normal distribution assumptions are therefore variations of linear analysis.
(2) There are many circumstances in which nonlinear relations are introduced into data, in particular when data are categorical. Likert-type scores are a good example to show that intervals between assigned integers are not always as indicated by the assigned scores. Nonlinear relations can be easily tapped into by the quantification method, known by many familiar names. For a quantification method to work, however, we must desensitize or categorize variables first. Adherence to the high level measurement makes it difficult to look into nonlinear relations.
(3) Once data are categorized, even a set of perfectly correlated variables can no longer be explained by a single component, but multiple components become necessary to understand the data.

(4) Whether data are continuous or categorical, correlation in multidimensional space tends to be inflated if data are viewed in reduced space, a hidden danger that exists in the so-called dimension reduction framework.

(5) Finally, it is the contention of this paper that we should not adhere to the normal distribution assumption or any statistical model, but rather we should be open-minded to allow any relationships among variables in the light of information contained in the sample data. In this way, we will find comprehensive analysis of information in data. An old fashioned empirical approach to data analysis seems viable for comprehensive analysis of data, that is, for the objective of data analysis.

References

CRAMÉR, H. (1946): *Mathematical Methods of Statistics.* Princeton University Press, Princeton.

GREENACRE, M.J. (2004): *Personal communication.* Dortmund, Germany.

LIKERT, R. (1932): A technique forthe measurement of attitudes. *Archives of Psychology, 140, 44–53.*

NISHISATO, S. (1984): Forced classification: A simple application of a quantification method. *Psychometrika, 49, 25–36.*

NISHISATO, S. (2000): Data types and information: Beyond the current practice of data analysis. In R. Decker and W. Gaul (Eds.) *Classification and Information Processing at the Turn of the Millennium.* Springer, Heidelberg, 40–51.

NISHISATO, S. (2004): A unified framework for multidimensional data analysis from dual scaling perspectives: Another look and some suggestions. *Japanese Journal of Sensory Evaluation, 8, 4–10* (in Japanese).

NISHISATO, S. (2005): Correlational structure of multiple-choice data as viewed from dual scaling. A revised paper submitted for a book by J. Blasius and M.J. Greenacre (Eds.): Proceedings of a conference held in Barcelona, 2003 (the title of the book is not decided yet).

NISHISATO, S. and BABA, Y. (1999); On contingency, projection and forced classification of dual scaling. *Behaviormetrika, 26, 207–219.*

NISHISATO, S. and GAUL, W. (1990). An approach to marketing data analysis: Forced classification procedure of dual scaling. *Journal of Marketing Research, 27, 354–360.*

Asymmetric Multidimensional Scaling of Relationships Among Managers of a Firm

Akinori Okada[1], Tadashi Imaizumi[2], and Hiroshi Inoue[3]

[1] School of Social Relations, Rikkyo (St. Paul's) University
171-8501 Toshima-ku, Tokyo, Japan
[2] School of Management and Information Sciences, Tama University
206-0022 Tama, Tokyo, Japan
[3] Department of Human Sciences, Kyushu Institute of Technology
804-8550 Tobata, Kitakyushu, Japan

Abstract. Relationships among managers of a firm were analyzed. Each manager responded who goes to whom for help or advice for work-related problems. Resulting responses constitute a set of matrices, each comes from a manager. Each matrix is transformed to a matrix of geodesic distances from row to column managers. The set of geodesic distance matrices was analyzed by the asymmetric multidimensional scaling. The result represents the hierarchical structure of the firm. The dimensions for symmetric relationships represent differences among departments, and those for asymmetric relationships represent differences within and between supervisors.

1 Introduction

The present study deals with help or advice relationships at work among managers of a small high-tech firm (Krackhart (1987)) by using asymmetric multidimensional scaling (MDS)(Zielman and Heiser (1996)). Each manager responded who goes to whom for asking help or advice for work-related problems in the firm including the manager oneself. The obtained relationships among managers are described by a set of matrices, where each matrix comes from a manager or a perceiver. The set of matrices is two-mode three-way (object×object×source) proximities. Krackhart (1987) used three procedures to analyze the three-way data on the social structure among managers. One of the three procedures analyzed each matrix or two-way data separately. But it is difficult to disclose differences among perceivers, because the result derived by the procedure does not have any commonality over perceivers. The other two procedures tried to aggregate three-way data into two-way data or to extract two-way relationships from three-way data. The differences between the two procedures lie in the way of the aggregation. Thus they cannot disclose differences among perceivers neither.

Several procedures dealing with three-way data on a social structure have been introduced. Batchelder et al. (1997) introduced a statistical method to analyze three-way data. While the method was able to assess the reliability of a social network (Batchelder et al. (1997), p. 56), it aggregated three-way data into two-way relationships. The method cannot disclose differences among perceivers. Borgatti and Everett (1992) introduced block models to deal with multiway social structures. Although the result showed ties or relationships among the obtained blocks, it cannot represent differences among perceivers. Kumbasar et al. (1994) analyzed three-way data on relationships among members of a professional group at a computer company. They rearranged three-way data into a two-way matrix by stacking matrices, and analyzed the stacked matrices or a super matrix by correspondence analysis. The differences among perceivers are not fully represented in a derived multidimensional space. Nakao and Romney (1993) analyzed three-way data of a social structure among university students (Nordlie (1958)). They used INDSCAL (Carroll and Chang (1970))to analyze the data. The analysis disclosed differences among perceivers. While some researches aggregated three-way social structures into two-way structures, some researches did not aggregate three-way social structures that makes it possible to represent differences among perceivers. The social structure perceived by a person might vary from the one perceiver to another (Krackhart (1987), p. 114), and it seems very important to deal with differences among perceivers.

Another important aspect of relationships among managers is that they are not always symmetric or a matrix for a manager is not always symmetric. While the three procedures of Krackhart (1987) deal with the asymmetry, the asymmetry is represented by the derived sociogram, which seems not practical for disclosing the asymmetry because of its complexity and difficulty in representing the degree of the asymmetry. Kumbasar et al. (1994) deals with the asymmetry by representing each object as a mean row point and as a column point in a configuration derived by the correspondence analysis of the stacked matrices. This seems to be a flaw in representing the asymmetry, because the asymmetry can also be dealt with by representing each object as a row point and as a mean column point. In any event, the configuration shows the asymmetry among objects which is common to all perceivers, and the differences of the asymmetry among perceivers are not shown. The asymmetry was ignored in Nakao and Romney (1993), because INDSCAL cannot deal with asymmetric relationships. The purpose of the present study is to analyze the social structure data of Krackhart (1987) by using two-mode three-way asymmetric MDS (Okada and Imaizumi (2002)) and to disclose the aspects of the social structure which cannot be disclosed from the analyses by using procedures described above; (a) representing differences among perceivers, and (b) disclosing the asymmetry in the relationships among managers which varies from one perceiver to another (Okada et al. (2003)).

2 The data

The data analyzed in the present study are relationships among 21 managers of a small high-tech firm employing approximately 100 people shown in

Krackhardt (1987, pp. 129-132). Twenty-one managers consist of a president, four vice-presidents each of whom heads up a department, and 16 supervisors of the four departments:
President; 7,
Department 1; 21 (vice-president), 6, 8, 12, and 17,
Department 2; 14 (vice-president), 3, 5, 9, 13, 15, 19 and 20,
Department 3; 18 (vice-president), 10, and 11, and
Department 4; 2 (vice-president), 1, 4, and 16.

Each of the 21 managers responded who goes to whom for help or advice for work-related problems including the respondent oneself. A 21×21 matrix of relationships is given by a respondent or a perceiver. The (j,k) element of the i-th matrix is dichotomous (the row manager go or does not go for help or advice to the column manager), and it represents whether manager j goes to k or not for help or advice perceived by manager i.

A matrix of geodesic distances among 21 managers was calculated from each matrix. A set of 21 matrices of geodesic distances among managers is derived. The (j,k) element of the i-th matrix represents the geodesic distance from manager j to k perceived by manager i. The set is two-mode three-way asymmetric proximities, because each matrix is not necessarily symmetric. The reason why the matrix of geodesic distances was derived is twofold; (a) one is that relationships not only directly connected but also indirectly connected any two managers can be dealt with, and (b) another is that the data were on relationships among managers at an organization having a hierarchical structure of superiors and subordinates (president - vice-president - supervisor). When a manager is unreachable from another manager, the geodesic distance from the latter to the former is defined as 21. We only use the fact that the geodesic distance 21 is larger than any geodesic distance which is smaller than 21 in the analysis, because the algorithm of the method is nonmetric, and does not depend on the numerical value itself.

3 The method

The set of 21 geodesic distance matrices was analyzed by two-mode three-way asymmetric MDS which allows different orientations of dimensions for symmetric and asymmetric relationships (Okada and Imaizumi (2002)). While in Okada et al. (2002) symmetric and asymmetric relationships were represented by the same set of dimensions, in the present study they are represented by two different sets of dimensions having different orientations.

The two-mode three-way asymmetric MDS used in the present study (Okada and Imaizumi (2002)) was extended from Okada and Imaizumi (1997). The present MDS gives three configurations; the common object configuration of objects, the symmetry weight configuration of perceivers, and the asymmetry weight configuration of perceivers. In the common object config-

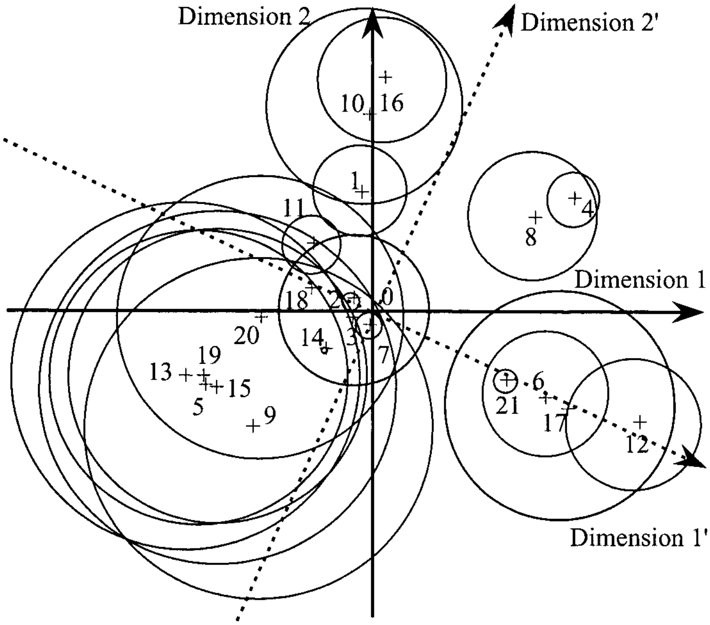


Fig. 1. Common object configuration of 21 managers.

uration, each manager is represented as a point and a circle (sphere, hypersphere) centered at that point in a multidimensional Euclidean space, which shows the relationships among managers like the joint stimulus space of the INDSCAL model (Carroll and Chang (1970)). Interpoint distances represent symmetric relationships among managers, and the radii represent asymmetric relationships among managers. In the symmetry weight configuration each perceiver is represented as a point in a multidimensional space. The coordinate of a perceiver represents the salience of the symmetric component of the relationships among managers along each dimension for the perceiver like the weight configuration of the INDSCAL model. The asymmetry weight configuration is based on dimensions derived by orthogonally rotating the dimensions of the symmetry weight configuration. Similar to the symmetry weight configuration, each perceiver is represented as a point in the asymmetry weight configuration, and the coordinate of a perceiver represents the salience of the asymmetric component of the relationships along each dimension for the perceiver. In the symmetry or asymmetry weight configurations, all perceivers are on a line passing through the origin, because the symmetry or asymmetry weight for a perceiver can be multiplicatively decomposed into two terms; the symmetry or asymmetry weight for a dimension, and the symmetry or asymmetry weight for a perceiver (Okada and Imaizumi (2000)).

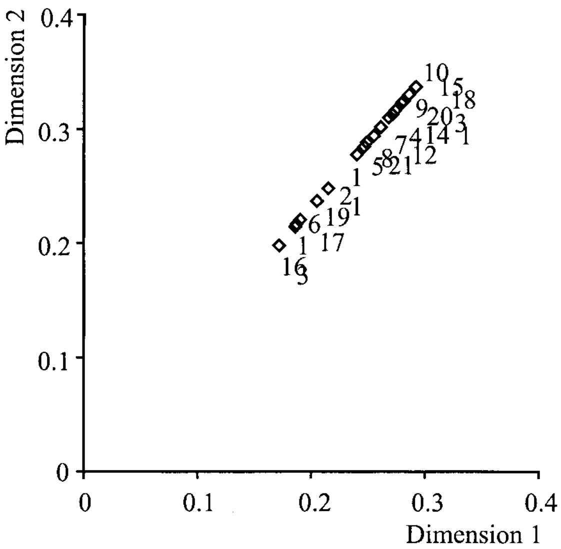


Fig. 2. Symmetry weight configuration of managers (perceivers).

4 Result

The analysis was done by using the maximum dimensionality of five through nine and using the minimum dimensionality of one. Among five kinds of results using different maximum dimensionalities (initial configurations), the result having the largest VAF was chosen at each of five- through unidimensional spaces. The two-dimensional result was chosen as the solution (Okada et al. (2003)). The obtained common object configuration of managers shown in Figure 1 represents the relationships among the managers in the firm which is common to all perceivers. Each manager is represented as a point and a circle centered at that point. There are two sets of dimensions in the configuration; one is the dimensions for the symmetry weight configuration drawn by solid lines, and the other is the dimensions for the asymmetry weight configuration drawn by dashed lines. The orientation of dimensions of both sets of dimensions are uniquely determined up to reflections and permutations, because of the introduction of the symmetry and asymmetry weights for dimensions.

Figures 2 and 3 shows the symmetry weight configuration and the asymmetry weight configuration respectively. Each manager or perceiver is represented as a point in the two-dimensional configuration. Dimensions of the symmetry weight configuration are dimensions 1 and 2 drawn by solid lines in Figure 1 for symmetric relationships among managers. Dimensions of the asymmetry weight configuration are dimensions 1' and 2' drawn by dashed lines in Figure 1 for asymmetric relationships. In both figures perceivers are on a lines passing the origin respectively.

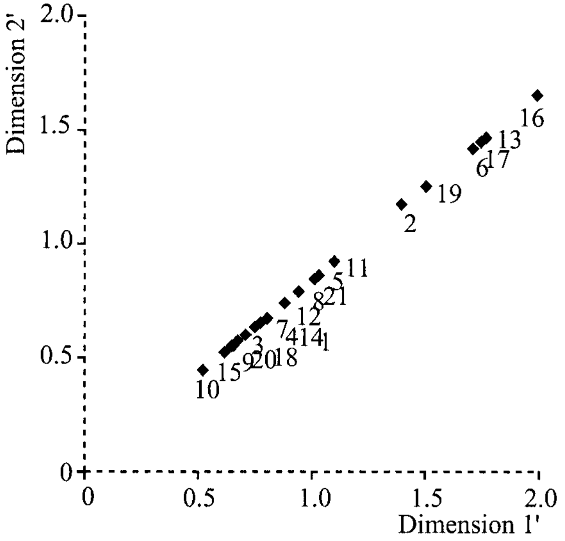


Fig. 3. Asymmetry weight configuration of managers (perceivers).

5 Discussion

In Figure 1 manager 7 (president) is located almost at the center of the configuration. Managers of a same department are located closely. Four vice-presidents are located between the president and supervisors of the department each heads up. Thus the configuration clearly depicts the hierarchical structure of managers of the firm. Manager 18 has the smallest radius (zero by the normalization), and manager 15 has the largest radius.

In the present application, the larger radius means that the corresponding manager has the larger tendency of going to the other managers for help or advice and has the smaller tendency of being asked by the other managers. Manager 18 (vice-president) has the largest tendency of being asked help or advice by the other managers and the smallest tendency of going to the other managers, and manager 15 has the largest tendency of going to the other managers and the smallest tendency of being asked by the other managers. The president (manager 7) and four vice-presidents (managers 2, 14, 18, and 21) have the five smallest radius among the 21 managers. The correlation coefficient between the radius and the level of the manager (Wasserman and Faust (1994), p. 273); 1 (president), 2 (vice-president), and 3 (supervisor of a department), is 0.61. And the correlation coefficient between the distance from the origin (centroid of the configuration) to the manager and the level of the manager is 0.64. These figures tell that the higher the level of a manager is, the smaller radius the manager has and the closer to the centroid the manager is located in the configuration. Manager 21, who is the vice-president of Department 1, is located farther from the centroid as well as from manager

7 (president) than the other three vice-presidents (managers 2, 14, and 18) are, and has the largest radius among the four vice-presidents. This suggests that manager 21 has weaker ties with the president as well as with the other vice-presidents, and has a smaller tendency of being asked help or advice by other managers than the other vice-presidents have. This validates manager 21 loses his prominence (Krackhardt (1987), p. 119). The hierarchical structure represented in Figure 1 is compatible with Figures 1 and 2 of Krackhardt (1987, pp. 120-121), and represents the degree of asymmetry of relationships among managers which is common to all perceivers.

The horizontal dimension for symmetric relationships differentiates three groups of managers; those of Department 1, of Departments 3 and 4, and of Department 2, and the vertical dimension differentiates two groups of managers; those of Departments 3 and 4, and of Departments 1 and 2. As described below, dimensions 1' and 2' for asymmetric relationships differentiate managers within a same level and between different levels.

The president and the vice-presidents (except manager 2) have larger symmetry weights and smaller asymmetry weights, suggesting they perceived the relationships among managers more symmetrically than departmental supervisors did. The symmetry weight for dimension 1 is smaller than that for dimension 2, because points representing perceivers in Figure 2 are above the 45 degree line passing through the origin. Thus dimension 2 is more salient than dimension 1 for these managers. The asymmetry weight for dimension 1' is larger than that for dimension 2', because points representing perceivers in Figure 3 are below the 45 degree line passing through the origin. Thus dimension 1' is more salient than dimension 2'.

It is expected that dimensions 1 and 2 correspond to fundamental processes of perceiving symmetric relationships of help or advice among managers, and that dimensions 1' and 2' correspond to fundamental processes of perceiving asymmetric relationships (Arabie et al. (1987), p. 21). As mentioned earlier, dimensions 1 and 2 correspond to the processes of perceiving symmetric relationships among departments. In Figure 1, points representing the president and four vice-presidents are located along a line almost parallel with dimension 1'. This suggests that dimension 1' corresponds to the process of perceiving asymmetric relationships within (vice-)presidents and within supervisors, and that dimension 2' corresponds to the processes of perceiving asymmetric relationships between (vice-)presidents and supervisors. It seems that (a) symmetric relationships are based on the interaction among departments and (b) asymmetric relationships are based on the vertical and horizontal aspects of the hierarchical structure of managers at the firm.

References

ARABIE, P., CARROLL, J.D., and DeSARBO, W.S. (1987): *Three-way Scaling and Clustering.* Sage Publications, Newburry Park, CA.

BATCHELDER, W.H., KUMBASAR, E., and BOYD, J. (1997): Consensus Analysis of Three-way Social Network Data. *Journal of Mathematical Sociology, 22, 29–58.*

BORGATTI, S.P. and EVERETT, M. G. (1992): Regular Block Models of Multiway, Multimode Matrices. *Social Networks, 14, 91–120.*

CARROLL, J.D. and CHANG, J.J. (1970): Analysis of Individual Differences in Multidimensional Scaling. *Psychometrika, 35, 283–319.*

KRACKHARDT, D. (1987): Cognitive Social Structures. *Social Networks, 9, 109–134.*

KUMBASAR, E., ROMNEY, A.K., and BATCHELDER, W.H. (1994): Systematic Biases in Social Perception. *Social Networks, 15, 109–131.*

NAKAO, K. and ROMNEY, A.K. (1993): Longitudinal Approach to Subgroup Formation: Reanalysis of Newcomb's Fraternity Data. *Social Networks, 15, 109–131.*

NORDLIE, P.G. (1958): *A Longitudinal Study of Interpersonal Attraction in a Natural Group Setting.* Doctoral dissertation, University of Michigan.

OKADA, A. and IMAIZUMI, T. (1997): Asymmetric Multidimensional Scaling of Two-mode Three-way Proximities. *Journal of Classification, 14, 195–224.*

OKADA, A. and IMAIZUMI, T. (2000): Two-mode Three-way Asymmetric Multidimensional Scaling with Constraints on Asymmetry. In: R. Decker and W. Gaul (Eds.): *Classification and Information Processing at the Turn of the Millennium.* Springer, Berlin, 52–59.

OKADA, A. and IMAIZUMI, T. (2002): Multidimensional Scaling with Different Orientations of Dimensions for Symmetric and Asymmetric Relationships. In: S. Nishisato, Y. Baba, H. Bozdogan, and K. Kanefuji (Eds.): *Measurement and Multivariate Analysis.* Springer, Tokyo, 97–106.

OKADA, A., IMAIZUMI, T., and INOUE, H. (2003): Asymmetric Multidimensional Scaling of Relationships among Managers. *Proceedings of the 13th International Meeting and the 68th Annual American Meeting of the Psychometic Society.*

OKADA, A., INOUE, H., and IMAIZUMI, T. (2002): An Analysis of Social Network Data by Asymmetric Multidimensional Scaling. *Proceedings of the 8th Conference of the International Federation of Classification Societies, 140.*

WASSERMAN, S. and FAUST, K. (1994): *Social Network Analysis.* Cambridge University Press, Cambridge.

ZIELMAN, B. and HEISER, W. (1996): Models for Asymmetric Proximities. *British Journal of Mathematical and Statistical Psychology, 49, 127–146.*

Aggregation of Ordinal Judgements Based on Condorcet's Majority Rule

Otto Opitz and Henning Paul

Lehrstuhl für Mathematische Methoden der Wirtschaftswissenschaften,
Universität Augsburg, D-86135 Augsburg, Germany

Abstract. This paper deals with the aggregation of ordinal judgements using Condorcet's majority rule. Since this rule achieves not necessarily a transitive relation, some authors formulate appropriate constraints to avoid this disadvantage. But also certain distance functions assigning appropriate dissimilarity values to pairs of relations lead to the majority rule with constraints. In comparison with Borda's scoring method, which directly generates an aggregation being a complete preorder, the results are sometimes different.

1 Introduction

There are many attempts to aggregate different ordinal judgements for getting a compromise ranking. Two of the most famous rules are Borda's "scoring method" and Condorcet's "majority rule". This majority rule is used by many authors to find an aggregating preorder which leads to a best approximation of the given different judgements (see for example: Gaul and Schader (1988), Marcotorchino and Michaud (1979), Schader and Tüshaus (1988), Tüshaus (1983)).

Let $C = \{c_1, \ldots, c_n\}$ be the set of cases and denote by $R_1, \ldots, R_m$ complete preorders on C. So R_k is a binary relation on C which is reflexive, transitive and complete. With the binary variable $r_{ijk} \in \{0, 1\}$ we define

$$(c_i, c_j) \in R_k \Leftrightarrow c_i \underset{k}{\preceq} c_j \Leftrightarrow r_{ijk} = 1 \quad (i, j \in \{1, \ldots, n\}; k \in \{1, \ldots, m\}). \quad (1)$$

The problem is to find a complete preorder R as an aggregation which minimizes an appropriate distance $\sum_{k=1}^{m} d(R_k, R)$.

In this paper we consider some well-known approaches in comparison with Borda's rule and we show some similarities as well as differences. Finally we discuss an example using data published by Gaul and Schader (1988), Gaul and Baier (1993).

2 Majority based rules

Using (1) Borda's scoring method (1784) is defined by

$$(c_i, c_j) \in R \iff \sum_k \sum_{c \in C} r_{cik} \leq \sum_k \sum_{c \in C} r_{cjk}. \tag{2}$$

Because the sum $\sum_k \sum_{c \in C} r_{cik}$ is equal to the number of all worse cases c in relation to c_i for all R_k, the relation R is a complete preorder.

The majority rule of Condorcet (1785) is based on paired comparisons of the following form:

$$(c_i, c_j) \in R \iff \sum_k r_{ijk} \geq \sum_k r_{jik} \tag{3}$$

This aggregation is reflexive and complete but not necessarily transitive. Marcotorchino and Michaud (1979), Michaud (1983) discussed this problem and analysed the binary optimization problem

$$\max \left(\sum_k \sum_{i \neq j} (r_{ijk} - r_{jik})\, x_{ij} \right) \tag{4}$$

with the constraints

$$\begin{aligned} & x_{uv} + x_{vw} - x_{uw} \leq 1 && (4.1)\\ & x_{uv} + x_{vu} \geq 1 && (4.2)\\ & x_{uv} \in \{0,1\}; \quad u, v, w \in \{1, \ldots, n\}. \end{aligned}$$

The objective function is increasing if $x_{ij} = 1$ for $\sum_k r_{ijk} > \sum_k r_{jik}$. Additionally transitivity and completeness are ensured by (4.1) and (4.2). Therefore the optimal solution $(x_{ij})_{n,n}$ represents a complete preorder.

Obviously the results of (2) and (4) are not identical. For the set of cases $C = \{c_1, c_2, c_3, c_4\}$ and the complete preorders

R_1 with $c_1 \prec c_2 \prec c_3 \prec c_4$,
R_2 with $c_2 \prec c_3 \prec c_4 \prec c_1$,
R_3 with $c_3 \prec c_4 \prec c_1 \prec c_2$,

we obtain by the method (2): $c_3 \prec c_2 \prec c_1 \prec c_4$,
respectively by the method (4): $c_2 \prec c_3 \prec c_4 \prec c_1$.

Even if (3) leads to a complete preorder, the results of (2) and (3) may be different. For $C = \{c_1, c_2, c_3, c_4\}$ and the complete preorders

R_1 with $c_1 \prec c_2 \prec c_3 \prec c_4$,
R_2 with $c_1 \prec c_2 \prec c_3 \prec c_4$,
R_3 with $c_4 \prec c_2 \prec c_1 \prec c_3$,

we obtain by the method (2): $c_1 \prec c_2 \prec c_4 \prec c_3$,
respectively by the method (4): $c_1 \prec c_2 \prec c_3 \prec c_4$.

Obviously in (2) the R_k–specific ranking positions of all c_i are included for calculation of R, while in (3) and (4) only the paired comparisons (c_i, c_j) of all R_k are important not depending on how much cases are ranked between c_i and c_j.

Gaul and Schader (1988) as well as Tüshaus (1983) use the distance function

$$\begin{aligned} d(R_k, R) = \mid \{(c_i, c_j) \in C \times C : ((c_i, c_j) \in R_k, (c_i, c_j) \notin R) \\ \vee ((c_i, c_j) \notin R_k, (c_i, c_j) \in R)\} \mid \end{aligned} \tag{5}$$

summarizing the "symmetric differences" between R_k and R. The corresponding optimization problem to (4) is

$$\min \sum_k d(R_k, R) \tag{6}$$

with the constraints (4.1), (4.2) and $x_{ij} \in \{0, 1\}$.

For the objective function we can write using (1) and assuming that no ties ($r_{ijk} = 1 \Leftrightarrow r_{jik} = 0$) exist

$$\sum_k d(R_k, R) = \sum_k \sum_{i \neq j} (r_{ijk} - x_{ij})^2 = \sum_k \sum_{i \neq j} (r_{ijk} + x_{ij} - 2r_{ijk}x_{ij}) \tag{7}$$

$$= \sum_k \sum_{i \neq j} r_{ijk} + \sum_{i \neq j} (m - 2\sum_k r_{ijk}) x_{ij} = c + \sum_{i \neq j} \left(\sum_k r_{jik} - \sum_k r_{ijk} \right) x_{ij}$$

$$\text{with} \quad \sum_k \sum_{i \neq j} r_{ijk} = c, \quad m - \sum_k r_{ijk} = \sum_k r_{jik}.$$

Therefore the minimization problem (6) is equivalent to the maximization problem

$$\max \left(\sum_k \sum_{i \neq j} (r_{ijk} - r_{jik}) x_{ij} - c \right) \tag{8}$$

with the constraints (4.1), (4.2) and $x_{ij} \in \{0, 1\}$,

and the solutions of (4) and (6) are identical. Further the optimization problem (6), with the symmetric distance $\sum_k d(R_k, R)$ as the objective function which has to be minimized, can be interpreted as the transitive approximation to the classical majority solution (3).

3 An example

For the aggregation of ordinal data we use a data set of Gaul and Schader (1988). By paired comparisons 69 persons had to judge ten cognac print ads. The participants have been asked to decide for 45 pairs, which of two ads is more appealing. To get an opportunity of comparison with the results of Gaul and Schader (1988) we use their partition of persons in two classes M_1, M_2 with 43 and 26 elements.

Characterizing the cognac print ads by c_i $(i = 1, \ldots, 10)$ we apply

- Borda's scoring method (2),
- Michaud's majority rule with constraints (4),
- the distance function with constraints (6).

Using a branch and bound algorithm we obtain the same complete preorder as Gaul and Schader (1988) for the class M_1:

$$c_9 \prec c_2 \prec c_4 \prec c_6 \prec c_7 \prec c_1 \prec c_8 \prec c_3 \prec c_{10} \prec c_5$$

For the class M_2 we get concording results

$$c_2 \prec c_6 \prec c_7 \prec c_8 \prec c_3 \prec c_{10} \prec c_1 \prec c_5 \prec c_4 \prec c_9$$

for all methods with the exception of Borda's rule, which shows two inversions $c_3 \prec c_8$ instead of $c_8 \prec c_3$ and $c_5 \prec c_1$ instead of $c_1 \prec c_5$.

4 Conclusion

The discussed problem of aggregating ordinal judgements is important for many applications, for instance multiattribute or multiobjective data analysis or collective choice analysis. Generally the classical majority rule with constraints (4.1), (4.2) seems to be more appropriate for an aggregation than the scoring method (2). Obviously an incontestable solution cannot exist. Nevertheless the confidence in this approach will increase the more similar the results will be. This is given in the example of Gaul and Schader (1988).

References

BARTHELEMY, J.P. and MONJARDET, B. (1981): The Median Procedure in Cluster Analysis and Social Choice Theory. *Mathematical Social Sciences, 1, 235–267.*

BORDA, J.-C. (1784): *Mémoir sur les Élections au Scrutins.* Academie Royale des Sciences, Paris.

CONDORCET, MARQUIS DE (1785): *Essai sur l'Application de l'Analyse à la Probabilité des Décisions rendues à la Pluralité des Voix.* Imprimerie Royale, Paris.

GAUL, W. and BAIER, D. (1993): *Marktforschung und Marketing Management.* Oldenbourg, München, Wien.

GAUL, W. and SCHADER, M. (1988): Clusterwise aggregation of Relations. *Applied Stochastic Models and Data Analysis, 4, 273–282.*

MARCOTORCHINO, J.-F. and MICHAUD, P. (1979): *Optimisation en Analyse Ordinale des Données.* Masson, Paris, 1979.

MICHAUD, P. (1983): Opinions Aggregation. In: J. Janssen, J.-F. Marcotorchino, and J.M. Proth (Eds.): *New Trends in Data Analysis and Applications.* North Holland, Amsterdam, 6–27.

SCHADER, M. and TÜSHAUS, U. (1988): Analysis of Qualitative Data: A Heuristic for Finding a Complete Preorder. In: H.H. Bock (Ed.): *Classification and Related Methods of Data Analysis.* North Holland, Amsterdam, 341–346.

TÜSHAUS, U. (1983): *Aggregation binärer Relationen in der qualitativen Datenanalyse.* Athenäum-Hain-Hanstein, Königstein/Taunus.

ANOVA Models with Generalized Inverses

Wolfgang Polasek[1] and Shuangzhe Liu[2]

[1] Institute of Advanced Studies, Stumpergasse 56, A-1060 Vienna, Austria
[2] Centre for Mathematics and its Applications, Mathematical Sciences Institute,
Australian National University, Canberra, ACT 0200, Australia

Abstract. The 2-way ANOVA model is analyzed with generalized matrix or g-inverses. We derive the co-called OLS^{-} and OLS^{+} estimators of the rank deficient ANOVA model. The new g-inverses lead to two simple effects in a two-way ANOVA model: column means and adjusted row means or vice versa: row means and adjusted column means. For the F- Test this parameterizations is invariant.

1 Introduction

ANOVA models are widely used in applied statistics and are a basic tool for many designed experiments. The theory of the classical ANOVA model can be found in e.g. Scheffé (1959), Rao (1973) and Seber (1984). The ANOVA model can be formulated as a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where $\mathbf{X}$ consists only of 0/1-elements and reflects the design of the analysis. If $\mathbf{X}$ is rank deficient then the usual OLS estimator does not exist (uniquely) since $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is singular. Theil (1971, pp. 75-81) studied a (general) normal regression model. For classification models with arbitrary patterns (in their notation), Dodge and Majumdar (1979) introduced an algorithm to find the least square generalized inverse of $\mathbf{X}$. Bayesian methods for ANOVA models can be found in Press (1989), variance component models in Searle et al. (1992), a hierarchical analysis in Smith (1973) and robustness with respect to the prior covariance matrix in Polasek and Poetzelberger (1994). Note that the multicollinearity problem is less important in an informative Bayesian analysis, because the covariance matrix can be made invertible with informative prior information. The usual solution to overcome the multicollinearity problem in classical estimation is to incorporate an intercept for estimating the overall (or grand) mean and to specify the level of the factors by dummy variables except for one. All the group effects have to be interpreted relative to the overall mean and the left out effect (dummy variable). Such parameterisation is not necessary from a Bayesian point of view and the parameterisation matters for the elicitation of prior information. Therefore the goal of the paper is to investigate the relationship of the Bayesian and classical ANOVA model when we incorporate all the column and row effects (e.g. into the 2-way ANOVA model) without deleting any level and using a full informative prior information. In section 2 to 4 we describe the rank deficient OLS estimates based on matrix g-inverses. Section 5 discusses invariance properties and Section 6 concludes.

1.1 The 2-way ANOVA model

For the 2-way ANOVA with balanced design (one observation per cell) the $n \times q$ matrix $\mathbf{Y}$ is decomposed as

$$y_{ij} = \alpha_i + \beta_j + u_{ij}, \quad i = 1, \ldots, q; \quad j = 1, \ldots, n. \tag{1}$$

Using Kronecker products we obtain the linear regression model

$$\mathbf{y} = (\mathbf{I}_q \otimes \mathbf{1}_n)\boldsymbol{\beta} + (\mathbf{1}_q \otimes \mathbf{I}_n)\boldsymbol{\beta} + \mathbf{u} = \tilde{\mathbf{X}}\boldsymbol{\gamma} + \mathbf{u}, \tag{2}$$

where

$$\begin{aligned}
\mathbf{y} &= vec\,\mathbf{Y} = (y_{11}, ..., y_{1n}, ..., y_{q1}, ..., y_{qn})^{\top} = (\mathbf{y}_1^{\top}, ..., \mathbf{y}_q^{\top}),\\
\boldsymbol{\alpha} &= (\alpha_1, ..., \alpha_q)^{\top}, \qquad \boldsymbol{\beta} = (\beta_1, ..., \beta_n)^{\top},\\
\tilde{\mathbf{X}} &= (\mathbf{X}_1, \mathbf{X}_2) = (\mathbf{I}_q \otimes \mathbf{1}_n, \mathbf{1}_q \otimes \mathbf{I}_n), \qquad \boldsymbol{\gamma} = (\boldsymbol{\alpha}^{\top}, \boldsymbol{\beta}^{\top})^{\top},\\
\mathbf{u} &= (u_{11}, ..., u_{1n}, ..., u_{q1}, ..., u_{qn})^{\top}.
\end{aligned}$$

$\tilde{\mathbf{X}}$ is the design matrix of a 2-way ANOVA model and $\mathbf{X}_1$ is the block of the design corresponding to the column effects while $\mathbf{X}_2$ builds up the row effects. Note that $\tilde{\mathbf{X}}$ is singular with rank $n + q - 1$.

1.2 g-inverses

The matrix g-inverses of $\mathbf{A}$ are defined by (some of) the four equations: (a) $\mathbf{ABA} = \mathbf{A}$, (b) $\mathbf{BAB} = \mathbf{B}$, (c) $(\mathbf{AB})^{\top} = \mathbf{AB}$ and (d) $(\mathbf{BA})^{\top} = \mathbf{BA}$.

A non-unique g-inverse $\mathbf{B} = \mathbf{A}^-$ satisfies only (a). The unique Moore-Penrose inverse $\mathbf{B} = \mathbf{A}^+$ satisfies (a) through (d), see e.g. Rao (1973) or Magnus and Neudecker (1988). We derive two rank deficient OLS$^-$ estimators using the two types of g-inverses, $(\mathbf{X}^{\top}\mathbf{X})^-$ and the unique OLS$^+$ estimator for the Moore-Penrose inverse $(\mathbf{X}^{\top}\mathbf{X})^+$. We show that the R^2 measure and the F statistic are not affected by the choice of the g-inverse.

2 The OLS$^-$ estimators

Consider the 2-way ANOVA model $\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\gamma} + \mathbf{u}$ in (2). Because $\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}}$ is singular, we propose to compute the OLS$^-$ estimator $\hat{\boldsymbol{\gamma}} = (\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}})^-\tilde{\mathbf{X}}^{\top} vec\,\mathbf{Y}$ using the following two matrix g-inverses:

$$(a) \quad (\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}})^- = \begin{pmatrix} \frac{1}{n}\mathbf{M}_q & 0 \\ 0 & \frac{1}{q}\mathbf{I}_n \end{pmatrix}, (b) \quad (\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}})^- = \begin{pmatrix} \frac{1}{n}\mathbf{I}_q & 0 \\ 0 & \frac{1}{q}\mathbf{M}_n \end{pmatrix}.$$

Both g-inverses are block diagonal matrices where one block is proportional to the identity matrix and the other block is proportional to the column or row projection matrix, respectively.

Theorem 21 *OLS^{-} estimation*
The OLS^{-} estimator of 2-way model according to the two g-inverses is obtained from the $n \times p$ matrix $\mathbf{Y}$ by appropriate row and column averages. In case (a) the estimates are

$$\hat{\boldsymbol{\alpha}} = \bar{\bar{\mathbf{y}}}_{col} = \mathbf{M}_q \mathbf{Y}^{\top} \mathbf{1}_n / n = \tilde{\mathbf{Y}}_{col}^{\top} \mathbf{1}_n / n \text{ and } \hat{\boldsymbol{\beta}} = \bar{\mathbf{y}}_{row} = \mathbf{Y}\mathbf{1}_q / q, \tag{3}$$

and in case (b)

$$\hat{\boldsymbol{\alpha}} = \bar{\mathbf{y}}_{col} = vec\, \mathbf{1}_n^{\top} \mathbf{Y}/n \text{ and } \hat{\boldsymbol{\beta}} = \bar{\bar{\mathbf{y}}}_{row} = vec\, \mathbf{M}_n \mathbf{Y} \mathbf{1}_q / q = vec\, \tilde{\mathbf{Y}}_{row} \mathbf{1}_q / q. \tag{4}$$

The column and row averages are calculated from the row or column adjusted data matrices, respectively: $\tilde{\mathbf{Y}}_{col} = \mathbf{Y}\mathbf{M}_q$, and $\tilde{\mathbf{Y}}_{row} = \mathbf{M}_n \mathbf{Y}$, where the projectors are given by $\mathbf{M}_q = \mathbf{I}_q - \frac{1}{q}\mathbf{1}_q \mathbf{1}_q^{\top}$ and $\mathbf{M}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n^{\top}$. The variance of the OLS^{-} estimator is for the estimates in (3)

$$Var(\hat{\boldsymbol{\alpha}}) = Var(\bar{\bar{\mathbf{y}}}_{col}) = \frac{1}{n}\mathbf{M}_q \quad \text{and} \quad Var(\hat{\boldsymbol{\beta}}) = Var(\bar{\mathbf{y}}_{row}) = \frac{1}{q}\mathbf{I}_n \tag{5}$$

and for the estimates in (4)

$$Var(\hat{\boldsymbol{\alpha}}) = Var(\bar{\mathbf{y}}_{col}) = \frac{\sigma^2}{n}\mathbf{M}_q \quad \text{and} \quad Var(\hat{\hat{\boldsymbol{\beta}}}) = Var(\bar{\mathbf{y}}_{row}) = \frac{\sigma^2}{q}\mathbf{I}_n \tag{6}$$

Proof. The OLS^{-} estimator in (3) is

$$\hat{\gamma} = \begin{pmatrix} \frac{1}{n}\mathbf{M}_q & 0 \\ 0 & \frac{1}{q}\mathbf{I}_n \end{pmatrix} \begin{pmatrix} \mathbf{I}_q \otimes \mathbf{1}_n^{\top} \\ \mathbf{1}_q^{\top} \otimes \mathbf{I}_n \end{pmatrix} vec\, \mathbf{Y} = \begin{pmatrix} vec\, \mathbf{1}_n^{\top} \mathbf{Y} \mathbf{M}_q / n \\ vec\, \mathbf{Y}\mathbf{1}_q / q \end{pmatrix} = \begin{pmatrix} \bar{\bar{\mathbf{y}}}_{col} \\ \bar{\mathbf{y}}_{row} \end{pmatrix}.$$

The OLS^{-} estimator in (4) is

$$\hat{\gamma} = \begin{pmatrix} \frac{1}{n}\mathbf{I}_q & 0 \\ 0 & \frac{1}{q}\mathbf{M}_n \end{pmatrix} \begin{pmatrix} \mathbf{I}_q \otimes \mathbf{1}_n^{\top} \\ \mathbf{1}_q^{\top} \otimes \mathbf{I}_n \end{pmatrix} vec\, \mathbf{Y} = \begin{pmatrix} \frac{1}{n} vec\, \mathbf{1}_n^{\top} \mathbf{Y} \\ \frac{1}{q}\mathbf{M}_n vec\, \mathbf{Y}\mathbf{1}_q \end{pmatrix} = \begin{pmatrix} \bar{\mathbf{y}}_{col} \\ \bar{\bar{\mathbf{y}}}_{row} \end{pmatrix}.$$

For the variance of the OLS^{-} estimator we have $Var(\hat{\gamma}) = \sigma^2 (\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}})^{-1}$ from where (3) and (4) follow by substitution.

Remarks:

(a) The two g-inverses can be checked to be two reflexive g-inverses as defined in Dodge and Majumdar (1979).

(b) It is interesting to note that the two solutions of g-inverses have some similarity with the two solutions of the median-polish method of 2-way tables (Tukey, 1977).

(c) The OLS$^-$ estimators presented in (3) and (4) can be written elementwise in case (a) as

$$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \text{ and } \hat{\beta}_j = \bar{y}_{\cdot j}, \tag{7}$$

and in case (b) as

$$\hat{\alpha}_i = \bar{y}_{i\cdot} \quad \text{and} \quad \hat{\beta}_j = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}, \quad i = 1, \ldots, q; \quad j = 1, \ldots, n. \tag{8}$$

The variance of the OLS estimators are given elementwise in case (a) as

$$Var(\bar{\bar{y}}_{col,i}) = Var(\bar{y}_{i.} - \bar{\bar{y}}_{..}) = \frac{q-1}{q} \cdot \frac{\sigma^2}{n} \quad \text{and} \quad Var(\bar{y}_{.j}) = \frac{\sigma^2}{n} \tag{9}$$

and in case (b) as

$$Var(\bar{y}_{i.}) = \frac{\sigma^2}{q} \quad \text{and} \quad Var(\bar{\bar{y}}_{row,j}) = Var(\bar{y}_{.j} - \bar{\bar{y}}_{..}) = \frac{n-1}{n} \cdot \frac{\sigma^2}{q} \tag{10}$$

which follows form (5) and (6).

(d) If we specify the ANOVA model with a grand mean μ then the linear model becomes

$$\mathbf{y} = (\mathbf{1}_q \otimes \mathbf{1}_n)\mu + (\mathbf{I}_q \otimes \mathbf{1}_n)\boldsymbol{\alpha} + (\mathbf{1}_q \otimes \mathbf{I}_n)\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{u} \tag{11}$$

with constraints $\mathbf{1}_q^\top\boldsymbol{\alpha} = 0$, $\mathbf{1}_n^\top\boldsymbol{\beta} = 0$, and the model matrix $\mathbf{X} = (\mathbf{1}_q \otimes \mathbf{1}_n,\ \mathbf{I}_q \otimes \mathbf{1}_n,\ \mathbf{1}_q \otimes \mathbf{I}_n)$, the OLS estimators are

$$\hat{\mu} = \bar{y}_{\cdot\cdot}, \quad \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}, \quad \hat{\beta}_j = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}. \tag{12}$$

Based on (12) the estimators of α_i and $\mu + \beta_j$ correspond to the results in (7), and the estimators of $\mu + \alpha_i$ and β_j to the results in (8). Searle et al. (1992, sections 4.3 and 4.9) discuss this model for the unbalanced case. The OLS estimator of the linear model in (11) may be written as

$$\hat{\boldsymbol{\gamma}} = \mathbf{G}\mathbf{X}^\top\mathbf{y},$$

where $\mathbf{G}$ is a g-inverse of $\mathbf{X}^\top\mathbf{X}$:

$$\mathbf{G} = \begin{pmatrix} 1/qn & 0 & 0 \\ -\mathbf{1}_q/qn & \mathbf{I}_q/n & 0 \\ -\mathbf{1}_n/qn & 0 & \mathbf{I}_n/q \end{pmatrix}.$$

Note that the covariance matrix is $Var(\hat{\boldsymbol{\gamma}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^-$ and has the simpler structure for the centered row and column effects:

$$Var(\hat{\alpha}_i) = \frac{\sigma^2}{n} \quad \text{and} \quad Var(\hat{\beta}_j) = \frac{\sigma^2}{q}.$$

3 Stepwise OLS$^-$ estimation

The 2-step procedure for rank deficient OLS estimators is formulated in the next theorem and generalizes the result of Seber (1988, p. 464). The resulting estimators are called 2-step OLS$^-$ estimators.

Theorem 31 *For the linear regression model with design matrices* $\mathbf{X}_1$ *and* $\mathbf{X}_2$*:*

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\alpha} + \mathbf{X}_2\boldsymbol{\beta} + \mathbf{u}, \tag{13}$$

the OLS$^-$ estimators are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_2^\top\mathbf{R}\mathbf{X}_2)^-\mathbf{X}_2^\top\mathbf{R}\mathbf{y} \quad \text{with} \quad \mathbf{R} = \mathbf{I}_n - \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^-\mathbf{X}_1^\top$$

and

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}_1^\top\mathbf{X}_1)^-\mathbf{X}_1^\top(\mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}). \tag{14}$$

If $Var(\mathbf{u}) = \sigma^2\mathbf{I}$*, the covariance matrix of the OLS$^-$ estimators is*

$$\text{Var}\begin{pmatrix}\hat{\boldsymbol{\alpha}}\\ \hat{\boldsymbol{\beta}}\end{pmatrix} = \sigma^2\begin{bmatrix}\mathbf{P} + \hat{\mathbf{X}}_2\mathbf{Q}\hat{\mathbf{X}}_2^\top & -\hat{\mathbf{X}}_2\mathbf{Q}\\ -\mathbf{Q}\hat{\mathbf{X}}_2^\top & \mathbf{Q}\end{bmatrix}$$

with $\mathbf{P} = (\mathbf{X}_1^\top\mathbf{X}_1)^-\mathbf{X}_1^\top\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{\top-}$, $\hat{\mathbf{X}}_2 = (\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\mathbf{X}_2$ *and* $\mathbf{Q} = (\mathbf{X}_2^\top\mathbf{R}\mathbf{X}_2)^-\mathbf{X}_2^\top\mathbf{R}\mathbf{X}_2(\mathbf{X}_2^\top\mathbf{R}\mathbf{X}_2)^{\top-}$.

Proof. We premultiply (13) by $\mathbf{R}$ to obtain

$$\mathbf{R}\mathbf{y} = \mathbf{R}\mathbf{X}_1\boldsymbol{\alpha} + \mathbf{R}\mathbf{X}_2\boldsymbol{\beta} + \mathbf{R}\mathbf{u} = \mathbf{R}\mathbf{X}_2\boldsymbol{\beta} + \tilde{\mathbf{u}}, \quad \text{with} \quad \tilde{\mathbf{u}} = \mathbf{R}\mathbf{u}.$$

The OLS$^-$ estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}_2^\top\mathbf{R}^\top\mathbf{R}\mathbf{X}_2)^-\mathbf{X}_2^\top\mathbf{R}^\top\mathbf{R}\mathbf{y} = (\mathbf{X}_2^\top\mathbf{R}\mathbf{X}_2)^-\mathbf{X}_2^\top\mathbf{R}\mathbf{y}$. Since (13) can be written as $\mathbf{y} - \mathbf{X}_2\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\alpha} + \mathbf{u}$, the OLS$^-$ estimator for $\boldsymbol{\alpha}$ is (14). The variance-covariance matrix of the OLS$^-$ estimator follows from direct insertion.

Theorem 32 *The 2-step OLS estimators for the 2-way ANOVA model.*
In the ANOVA model (1) we estimate the parameters by the 2-step OLS$^-$ method:
(a) Row-wise start (we choose $\mathbf{X}_1 = \mathbf{I}_q \otimes \mathbf{1}_n$*, and* $\mathbf{X}_2 = \mathbf{1}_q \otimes \mathbf{I}_n$*)*

$$\hat{\boldsymbol{\beta}} = \bar{\bar{\mathbf{y}}}_{row} \quad \textit{and} \quad \hat{\boldsymbol{\alpha}} = \bar{\mathbf{y}}_{col}.$$

(b) Column-wise start (we choose $\mathbf{X}_1 = \mathbf{1}_q \otimes \mathbf{I}_n$*, and* $\mathbf{X}_2 = \mathbf{I}_q \otimes \mathbf{1}_n$*)*

$$\hat{\boldsymbol{\beta}} = \bar{\mathbf{y}}_{row} \quad \textit{and} \quad \hat{\boldsymbol{\alpha}} = \bar{\bar{\mathbf{y}}}_{col}.$$

Proof. (a) Since the 2 components of the design matrix $\mathbf{X}$ are $\mathbf{X}_1 = \mathbf{I}_q \otimes \mathbf{1}_n$ and $\mathbf{X}_2 = \mathbf{1}_q \otimes \mathbf{I}_n$, we find $(\mathbf{X}_1^\top\mathbf{X}_1)^{-1} = \mathbf{I}_q/n$ and $\mathbf{R} = \mathbf{I}_q \otimes \mathbf{M}_n$, where $\mathbf{M}_n = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\top/n$ with $\mathbf{M}_n\mathbf{M}_n^-\mathbf{M}_n = \mathbf{M}_n$.
Now it follows from theorem 31 that the OLS$^-$ estimator for $\boldsymbol{\beta}$ is

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \left[(\mathbf{1}_q^\top \otimes \mathbf{I}_n)(\mathbf{I}_q \otimes \mathbf{M}_n)(\mathbf{1}_q \otimes \mathbf{I}_n)\right]^- (\mathbf{1}_q^\top \otimes \mathbf{I}_n)(\mathbf{I}_q \otimes \mathbf{M}_n) vec\,\mathbf{Y}\\ &= vec\,\mathbf{M}_n\mathbf{Y}\mathbf{1}_q/q = \bar{\bar{\mathbf{y}}}_{row}\end{aligned}$$

The two-step ANOVA estimator for $\boldsymbol\alpha$ using equation (14) with $\hat{\boldsymbol\beta} = \bar{\tilde{\mathbf{y}}}_{row}$ is

$$\hat{\boldsymbol\alpha} = (\mathbf{I}_q \otimes \mathbf{1}_n^\top \mathbf{1}_n)^{-1}(\mathbf{I}_q \otimes \mathbf{1}_n^\top)[vec\,\mathbf{Y} - (\mathbf{1}_q \otimes \mathbf{I}_n)\bar{\tilde{\mathbf{y}}}_{row}] = vec\,\mathbf{1}_n^\top \mathbf{Y}/n = \bar{\mathbf{y}}_{col}.$$

(b) Using $\mathbf{X}_1 = \mathbf{1}_q \otimes \mathbf{I}_n$, $\mathbf{X}_2 = \mathbf{I}_q \otimes \mathbf{1}_n$ and $\mathbf{R} = \mathbf{M}_q \otimes \mathbf{I}_n$, we find the OLS$^-$ estimator for $\boldsymbol\alpha$ (associated with $\mathbf{X}_2 = \mathbf{I}_q \otimes \mathbf{1}_n$ in the model)

$$\hat{\boldsymbol\alpha} = [(\mathbf{I}_q^\top \otimes \mathbf{1}_n^\top)(\mathbf{M}_q \otimes \mathbf{I}_n)(\mathbf{I}_q \otimes \mathbf{1}_n)]^-(\mathbf{I}_q \otimes \mathbf{1}_n^\top)(\mathbf{M}_q \otimes \mathbf{I}_n)vec\,\mathbf{Y} = \bar{\tilde{\mathbf{y}}}_{col}.$$

Using the above result we obtain the OLS$^-$ estimator for $\boldsymbol\beta$

$$\hat{\boldsymbol\beta} = (\mathbf{I}_n/q)(\mathbf{1}_q \otimes \mathbf{I}_n)[\,vec\,\mathbf{Y} - (\mathbf{I}_q \otimes \mathbf{1}_n)\hat{\boldsymbol\alpha}] = \bar{\mathbf{y}}_{row}.$$

4 The OLS$^+$ estimator

The OLS$^+$ estimator for the ANOVA model (2) is given by the unique solution of the Moore-Penrose inverse $(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^+$.

Theorem 41 *OLS$^+$ estimation*
The OLS$^+$ estimators of γ and $\tilde{\mathbf{X}}\gamma$ for model (2) are given by

$$\hat\gamma = (\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^+\tilde{\mathbf{X}}^\top\mathbf{Y} = \tilde{\mathbf{X}}^+\mathbf{y}, \tag{15}$$

$$\tilde{\mathbf{X}}\hat\gamma = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^+\mathbf{y} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^-\tilde{\mathbf{X}}^\top\mathbf{y}, \tag{16}$$

where

$$\tilde{\mathbf{X}}^+ = \begin{pmatrix} \mathbf{M}_q \otimes \mathbf{1}_n^\top/n + \mathbf{N}_q n/(q+n) \otimes \mathbf{1}_n^\top/n \\ \mathbf{1}_q^\top/q \otimes \mathbf{M}_n + \mathbf{1}_q^\top/q \otimes \mathbf{N}_n q/(q+n) \end{pmatrix}, \tag{17}$$

$$\tilde{\mathbf{X}}\tilde{\mathbf{X}}^+ = \mathbf{N}_q \otimes \mathbf{I}_n + \mathbf{M}_q \otimes \mathbf{N}_n = \mathbf{I}_q \otimes \mathbf{N}_n + \mathbf{N}_q \otimes \mathbf{M}_n, \tag{18}$$

and the projection matrices

$$\mathbf{M}_q = \mathbf{I}_q - \mathbf{1}_q\mathbf{1}_q^\top/q, \quad \mathbf{N}_q = \mathbf{1}_q\mathbf{1}_q^\top/q,$$
$$\mathbf{M}_n = \mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n^\top/n \text{ and } \mathbf{N}_n = \mathbf{1}_n\mathbf{1}_n^\top/n.$$

Proof. It is easy to verify (17) and (18). From (15), (17), (18) and the properties of vec-operations we find for the OLS$^+$ estimator

$$\begin{aligned}\hat\gamma &= \tilde{\mathbf{X}}^+\mathbf{y} \\ &= \begin{pmatrix} \mathbf{M}_q \otimes \mathbf{1}_n^\top/n + \mathbf{N}_q n/(q+n) \otimes \mathbf{1}_n^\top/n \\ \mathbf{1}_q^\top/q \otimes \mathbf{M}_n + \mathbf{1}_q^\top/q \otimes \mathbf{N}_n q/(q+n) \end{pmatrix} vec\,\mathbf{Y} \\ &= \begin{pmatrix} vec\,(\mathbf{1}_n^\top/n\mathbf{Y}\mathbf{M}_q + \mathbf{1}_n^\top\mathbf{Y}\mathbf{1}_q\mathbf{1}_q^\top/q(q+n)) \\ vec\,(\mathbf{M}_n\mathbf{Y}\mathbf{1}_q/q + \mathbf{1}_n\mathbf{1}_n^\top\mathbf{Y}\mathbf{1}_q/n(q+n)) \end{pmatrix} \\ &= \begin{pmatrix} \bar{\tilde{\mathbf{y}}}_{col} + \bar{\bar{y}}\mathbf{1}_q n/(q+n) \\ \bar{\tilde{\mathbf{y}}}_{row} + \bar{\bar{y}}\mathbf{1}_n q/(q+n) \end{pmatrix},\end{aligned} \tag{19}$$

$$\begin{aligned}\tilde{\mathbf{X}}\hat{\gamma} &= \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{+}\mathbf{y} = (\mathbf{N}_q \otimes \mathbf{I}_n + \mathbf{M}_q \otimes \mathbf{N}_n)vec\,\mathbf{Y} \\ &= vec\,\mathbf{Y}\mathbf{N}_q + vec\,\mathbf{N}_n\mathbf{Y}\mathbf{M}_q = vec\,\bar{\mathbf{y}}_{row}\mathbf{1}_q^{\top} + vec\,\mathbf{1}_n\bar{\bar{\mathbf{y}}}_{col}^{\top} \\ &= (\mathbf{I}_q \otimes \mathbf{1}_n, \mathbf{1}_q \otimes \mathbf{I}_n)\begin{pmatrix}\bar{\bar{\mathbf{y}}}_{col} \\ \bar{\mathbf{y}}_{row}\end{pmatrix} = \tilde{\mathbf{X}}\begin{pmatrix}\bar{\bar{\mathbf{y}}}_{col} \\ \bar{\mathbf{y}}_{row}\end{pmatrix}. \qquad (20)\end{aligned}$$

Similarly, using the properties of projection matrices in (18), we get

$$\begin{aligned}\tilde{\mathbf{X}}\hat{\gamma} &= \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{+}\mathbf{y} = (\mathbf{I}_q \otimes \mathbf{N}_n + \mathbf{N}_q \otimes \mathbf{M}_n)vec\,\mathbf{Y} \\ &= \tilde{\mathbf{X}}\begin{pmatrix}\bar{\mathbf{y}}_{col} \\ \bar{\bar{\mathbf{y}}}_{row}\end{pmatrix}, \qquad (21)\end{aligned}$$

where $\bar{\bar{\mathbf{y}}}_{col} = \mathbf{M}_q\mathbf{Y}^{\top}\mathbf{1}_n/n, \bar{\bar{\mathbf{y}}}_{row} = \mathbf{M}_n\mathbf{Y}\mathbf{1}_q/q, \bar{\bar{y}} = \mathbf{1}_n^{\top}\mathbf{Y}\mathbf{1}_q/nq, \bar{\mathbf{y}}_{col} = \mathbf{Y}^{\top}\mathbf{1}_n/n$, and $\bar{\mathbf{y}}_{row} = \mathbf{Y}\mathbf{1}_q/q$. We see that the OLS$^+$ estimator $\hat{\gamma}$ in (19) is a weighted average between the overall mean $\bar{\bar{\mathbf{y}}}$ and the centered estimators $\bar{\bar{\mathbf{y}}}_{row}$ and $\bar{\bar{\mathbf{y}}}_{col}$, respectively. Note that $\tilde{\mathbf{X}}\hat{\gamma}$ is unique in the sense that $\hat{\gamma}$ can be an OLS estimator derived from any $(\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}})^{-}$ including our two chosen g-inverses, and $(\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}})^{+}$, because $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{+}\mathbf{y} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^{\top}\tilde{\mathbf{X}})^{-}\tilde{\mathbf{X}}^{\top}\mathbf{y}$.

For programming purposes the following result for design matrices with switched design blocks is useful. If we use the alternative parameterisation of the linear model (8), i.e.

$$\check{\mathbf{y}} = (\mathbf{1}_n \otimes \mathbf{I}_q)\boldsymbol{\alpha} + (\mathbf{I}_n \otimes \mathbf{1}_q)\boldsymbol{\beta} + \check{\mathbf{u}} = \check{\mathbf{X}}\gamma + \check{\mathbf{u}}, \qquad (22)$$

where

$$\begin{aligned}\check{\mathbf{y}}^{\top} &= (y_{11}, ..., y_{q1}, ..., y_{1n}, ..., y_{qn})^{\top} = (\mathbf{y}_{(1)}^{\top}, ..., \mathbf{y}_{(n)}^{\top}), \\ \boldsymbol{\alpha} &= (\alpha_1, ..., \alpha_q)^{\top}, \quad \boldsymbol{\beta} = (\beta_1, ..., \beta_n)^{\top}, \\ \check{\mathbf{X}} &= (\mathbf{1}_n \otimes \mathbf{I}_q, \mathbf{I}_n \otimes \mathbf{1}_q), \qquad \gamma = (\boldsymbol{\alpha}^{\top}, \boldsymbol{\beta}^{\top})^{\top}, \\ \check{\mathbf{u}} &= (u_{11}, ..., u_{q1}, ..., u_{1n}, ..., u_{qn})^{\top},\end{aligned}$$

then we find

$$\hat{\gamma} = \check{\mathbf{X}}^{+}\check{\mathbf{y}}, \qquad (23)$$

$$\check{\mathbf{X}}\hat{\gamma} = \check{\mathbf{X}}\check{\mathbf{X}}^{+}\check{\mathbf{y}} = \check{\mathbf{X}}(\check{\mathbf{X}}^{\top}\check{\mathbf{X}})^{-}\check{\mathbf{X}}^{\top}\check{\mathbf{y}}. \qquad (24)$$

We obtain the unique OLS$^+$ estimator $\hat{\gamma}$ based on (15) or (22), i.e.,

$$\hat{\gamma} = \check{\mathbf{X}}^{+}\check{\mathbf{y}} = \tilde{\mathbf{X}}^{+}\mathbf{y}, \qquad (25)$$

because the following identities hold

$$\begin{gathered}\mathbf{1}_n \otimes \mathbf{I}_q = \mathbf{K}_{nq}(\mathbf{I}_q \otimes \mathbf{1}_n), \quad \mathbf{I}_n \otimes \mathbf{1}_q = \mathbf{K}_{nq}(\mathbf{1}_q \otimes \mathbf{I}_n), \\ \check{\mathbf{X}} = \mathbf{K}_{nq}\tilde{\mathbf{X}}, \quad \check{\mathbf{X}}^{+} = \tilde{\mathbf{X}}^{+}\mathbf{K}_{qn}, \quad \check{\mathbf{y}} = \mathbf{K}_{nq}\mathbf{y},\end{gathered}$$

and $\mathbf{K}_{nq}$ is the $nq \times nq$ commutation matrix (for details see Magnus and Neudecker, 1991). Finally we find $\check{\mathbf{X}}^{+}\check{\mathbf{y}} = \tilde{\mathbf{X}}^{+}\mathbf{K}_{qn}\mathbf{K}_{nq}\mathbf{y} = \tilde{\mathbf{X}}^{+}\mathbf{y}$.

5 Invariance properties

The rank deficient OLS estimators have some useful invariance properties. Several results are invariant for any g-inverse $(\mathbf{X}^{\top}\mathbf{X})^{-}$, because $\mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-}\mathbf{X}^{\top}$ is invariant. For the full rank and the rank deficient linear model (8), i.e. $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{u}$, we have the well-known (centered) sum of squares decomposition

$$\mathbf{y}^{\top}\mathbf{M}_1\mathbf{y} = \mathbf{y}^{\top}\mathbf{M}_{\mathbf{X}}\mathbf{y} + \mathbf{y}^{\top}(\mathbf{M}_1 - \mathbf{M}_{\mathbf{X}})\mathbf{y}, \tag{26}$$

or the uncentered decomposition

$$\mathbf{y}^{\top}\mathbf{y} = \mathbf{y}^{\top}(\mathbf{1}_N\mathbf{1}_N^{\top}/N)\mathbf{y} + \mathbf{y}^{\top}\mathbf{M}_{\mathbf{X}}\mathbf{y} + \mathbf{y}^{\top}(\mathbf{M}_1 - \mathbf{M}_{\mathbf{X}})\mathbf{y}, \tag{27}$$

where the $N \times p$ model matrix $\mathbf{X}$ has rank $k \leq p$. The projection matrices in (33) are $\mathbf{M}_1 = \mathbf{I}_N - \mathbf{1}_N\mathbf{1}_N^{\top}/N$ and $\mathbf{M}_{\mathbf{X}} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}^{\top}\mathbf{X})^{-}\mathbf{X}^{\top}$. These decomposition are the basis for the following invariance result.

Theorem 51 *The R^2 and $F-$statistic are invariant for any g-inverse* $(\mathbf{X}^{\top}\mathbf{X})^{-}$

Proof. The coefficient of determination can be expressed as a ratio of sum of squares in the centered decomposition:

$$R^2 = \frac{\mathbf{y}^{\top}(\mathbf{M}_{\mathbf{X}})\mathbf{y}}{\mathbf{y}^{\top}\mathbf{M}_1\mathbf{y}},$$

and dividing by the degrees of freedom which are given by the ranks of the projection matrices we obtain the classical $F-$statistic

$$F = \frac{\mathbf{y}^{\top}(\mathbf{M}_1 - \mathbf{M}_{\mathbf{X}})\mathbf{y}/(k-1)}{\mathbf{y}^{\top}\mathbf{M}_{\mathbf{X}}\mathbf{y}/(N-k)}.$$

By an application of the Cochran theorem, see e.g. in Rao (1973, pp. 185-189) or Graybill (1976, chap. 13), we find that the quadratic forms are independently distributed. Note that the proof relies on $\mathbf{M}_1(\mathbf{M}_1 - \mathbf{M}_{\mathbf{X}}) = \mathbf{M}_1 - \mathbf{M}_{\mathbf{X}}$, and the following orthogonality properties of projection matrices:
$\mathbf{M}_{\mathbf{X}}(\mathbf{M}_1 - \mathbf{M}_{\mathbf{X}}) = \mathbf{0}$, $\mathbf{M}_{\mathbf{X}}\mathbf{1}_N\mathbf{1}_N^{\top}/N = \mathbf{0}$, $(\mathbf{M}_1 - \mathbf{M}_{\mathbf{X}})\mathbf{1}_N\mathbf{1}_N^{\top}/N = \mathbf{0}$.
Furthermore, the ranks and traces of the projection matrices are given by
$r\ (\mathbf{1}_N\mathbf{1}_N^{\top}/N) = tr\ \mathbf{1}_N\mathbf{1}_N^{\top}/N = 1$, $r\ (\mathbf{M}_1 - \mathbf{M}_{\mathbf{X}}) = tr\ (\mathbf{M}_1 - \mathbf{M}_{\mathbf{X}}) = k-1$, and $r\ (\mathbf{M}_{\mathbf{X}}) = tr\ \mathbf{M}_{\mathbf{X}} = N-k$.

The invariance result means that we can use the output of a computer program which computes rank-deficient OLS estimates for regression models for the purpose of ANOVA models. Then the R^2 and $F-$statistic can be used in the usual way since the numerical values are invariant for any g-inverse. Furthermore we conclude that any Bayesian linear regression program which allows the specification of a normal-gamma prior distribution can be used for a Bayesian one or two-way ANOVA analysis. Design matrices can be generated also for more complex designs if an interface exists to statistical matrix languages with Kronecker products (like e.g. R, S-Plus, Matlab).

6 Conclusions

For the 2-way ANOVA model we proposed two matrix g-inverses and the Moore-Penrose inverse which are different from Dodge and Majumdar's (1979) least square g-inverse of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$. This leads to the definition of OLS^{-} and OLS^{+} estimators for the rank deficient ANOVA model. We showed that the rank deficient 2-way analysis can be obtained from the 1-way results in a sequential 2-step procedure. Furthermore, the results for the 1- and 2-way ANOVA models can be generalized to the MANOVA case (see Polasek and Liu (1997)) and are also useful in a Bayesian model.

References

DODGE, Y. and MAJUMDAR, D. (1979): An algorithm for finding least square generalized inverses for classification models with arbitrary patterns. *J. Statist. Comput. Simul., 9, 1–17.*

GRAYBILL, F.A. (1976): *Theory and Application of the Linear Model*, Wadsworth & Brooks/Cole, Pacific Grove, California.

MAGNUS, J.R. and NEUDECKER, H. (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, Chichester.

POIRIER, D.J. (1995): *Intermediate Statistics and Econometrics*, The MIT Press, Cambridge, Mass.

POLASEK, W. and LIU, S. (1997): MANOVA models: A Bayesian analysis, University of Basel, mimeo.

POLASEK, W. and POETZELBERGER, K. (1994): Robust Bayesian methods in simple ANOVA models. *J. Statist. Planning and Inference, 40, 295–311.*

PRESS, S.J. (1989): *Bayesian Statistics: Principles, Models and Applications*, Wiley, New York.

RAO, C.R. (1973): *Linear Statistical Inference and Applications*, 2nd edn., Wiley, New York.

SCHEFFÉ, H. (1959): *The Analysis of Variance*, Wiley, New York.

SEARLE, S.R., CASELLA, G., and MCCULLOCH, C.E. (1992): *Variance Components*, Wiley, New York.

SEBER, G.A.F. (1984): *Multivariate Observations*, Wiley, New York.

SMITH, A.F.M. (1973): A general Bayesian linear model. *J. Royal Stat. Soc. B., 35, 67–75.*

THEIL (1971): *Principle fo Econometrics.* Wiley, New York.

TUKEY, J.W. (1977): *Exploratory Data Analysis*, Addison-Wesley, Reading, Mass.

Patterns in Search Queries

Nadine Schmidt-Mänz[1] and Martina Koch[2]

[1] Institut für Entscheidungstheorie und Unternehmensforschung,
Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

[2] Institut für Technische Informatik,
Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

Abstract. To optimize the performance of search engines, it is important to understand (human) online searching behavior.

We examine search queries in four different German (meta) search engines to describe some general patterns of search queries such as the percentage of operators used in search queries, length of queries and number of queries per time unit (days or hours). We also analyze the frequencies of queries and terms respectively. In addition, we describe the functionality of the vocabulary growth. Further, we try to find differences between search queries in meta and general search engines.

With this information in mind, we give some managerial implications for search engines.

1 Introduction

Although there are several studies of search queries conducted in English there is a lack of such studies of queries which are mostly transmitted by German-speaking people. To close this gap, we examined four different German search engines during a period of up to five months. We try to find general patterns in those queries, compare the results with former studies, and juxtapose the (human) searching behavior in meta and the searching behavior in general search engines.

In the second section, we give descriptions and definitions, after that we give a literature review, and explain our web data used. In section 3, we present our results such as the number of search queries and terms, complexity of queries, topics searched for, distribution of search queries, and the vocabulary growth.

In section 4, we summarize our results, compare general and meta search engines to give managerial implications for search engines in section 5.

2 Analyzing search queries

In this section, we shortly explain the different kinds of search engines and introduce some expressions typically used in this field of research, such as 'search query' or 'term'.

Afterwards, we give a brief summary of former studies and describe the data used for our sample.

2.1 Descriptions and definitions

General search engines (e.g., Google) have their own index of documents and web pages which are automatically generated by a web crawler. Meta search engines (e.g., Metager) do not have an own index but use the indices of different search engines simultaneously. Some search engines have a 'live ticker' or 'live search' enabling one to see the current search queries of other users. This possibility is also often called 'spy function'.
A (search) term is an uninterrupted sequence of characters (e.g., letters or numbers) without a space in between. A non-empty (search) query is the combination of terms a person (or maybe a robot) has typed in a search engine's search interface. A (search) topic is a label which combines different terms of the same subject.
The observation period is $[0;T]$ and t is one equidistant interval in this period (hours, days, or weeks). $\boldsymbol{V}$ is the N-dimensional vector where every dimension $n = 1, \ldots, N$ represents a term that has appeared at least once in the search queries observed. The values $v_n \in \mathbb{N}$ stored in each dimension are the number of occurrences of the term n until the end of the observation period.
$||\boldsymbol{V}||_1 = \sum_{n=1}^{N} v_n$ is referred to as the gross number of term occurrences and N is the net number of terms at time point T. $\boldsymbol{S}$ is the M-dimensional vector of search queries and $w_m \in \mathbb{N}$ the number of occurrences of search query $m = 1, \ldots, M$. The gross and net numbers of search queries are calculated equivalently to the numbers of term occurrences.
Operators (e.g., " ", OR, AND or NOT) are logic constructs to combine terms and are not seen as terms. To give an example, the search for an exact phrase is enfolded by double quotes.

2.2 Review of literature

There are several studies of interest in this field of research of which we briefly summarize the most important ones. Silverstein et al. (1999) analyzed a query log of the AltaVista search engine with 840 Mio. non-empty search queries of one month. The net number of search queries was 150 Mio., whereas 63,7% were only used once. The average number of terms in a query was 2.35. Operators were used in 20,4% of the search queries. One half of the top 25 search queries belonged to a topic which could be labeled with 'erotic'.
Beitzel et al. (2004) examined a query log of America Online of one week. They examined the distribution of query occurrences over the hours of the days. The "rush hour" of search queries is between 9:00 p.m. and 10:00 p.m., and between 5:00 a.m. and 6:00 a.m., the smallest amount of queries reached the engine. Concrete numbers of term and operator usage were not reported.
Jansen et al. (2000), Spink and Xu (2000), Spink et al. (2001, 2002, and 2000a), and Wolfram et al. (2001) analyzed query logs of the Excite search engine. They compared partial logs of days in different years. They concluded, e.g., that the topics searched for changed from 'erotic' to 'business'.

Hölscher and Strube (2000) analyzed a query log of the Fireball engine which consisted of 16 Mio. search queries and 27 Million terms. The average number of terms per query was 1.7, whereas 54.6% of all non-empty queries contained only one term.

The only study based on an observation of a live ticker is reported in Zien et al. (2000). Among other results, they describe the vocabulary growth of search queries and terms. Those distributions follow an exponential function. All in all, search queries are found to consist of less than three terms. The usage of operators is different from engine to engine. The number of queries and topics of interest changes during the hourly periods of a day.

2.3 Our web data used

We examined four different German (meta) search engines. We observed the live tickers of Metaspinner, Fireball, and Lycos. Metager sent us a daily list of the top 4,000 search queries and also the query logs. Table 1 shows the observation periods and the number of search queries corresponding to the search engines tracked by our program. In the live tickers observed, the list of search queries could be updated by refreshing those web pages. This circumstance was used to automatically collect a nearly complete list of search queries performed on these engines during the observation period. For every request our program sent to an engine, the list of search queries in the live ticker and the timestamp of the request were saved. The frequency of requests sent to Fireball, e.g., was 2.5 times per second (Lycos: 0.9, Mataspinner: 1.3). The varying interval lengths are caused by the different response times and volumes of search queries shown by the tickers. For every search engine tracked this huge amount of data was aggregated to one list containing all non-redundant search queries and timestamps.

Search Engine	Observation Period	Number of Months	Gross Number of Search Queries
Fireball	08/2004 - 01/2005	5	58.442.652
Lycos	08/2004 - 01/2005	5	76.343.559
Metager (top 4,000)	11/2004 - 01/2005	2	943.833
Metaspinner	09/2004 - 01/2005	4	2.174.823

Table 1. Time period and size of data collected

To give an idea of the degree of completeness of our lists, we performed 100,000 unique search queries on every engine observed. In the live search of Fireball 98,068 of these queries appeared again in our list (Lycos: 98,963, Metaspinner: 99,968). After this performance test those search queries were deleted from our lists for not tampering the results.

3 Composition of search queries

In this section we present our results and discuss typical patterns of search queries in both meta and general search engines.

3.1 Number of queries and terms

The gross and net numbers of queries and terms observed are shown in Table 2. The average number of appearance of a search query in Fireball (Lycos) is 7.0 (5.8) times. The average number of search query appearance in Metager is 6.4 and in Metaspinner 3.3, but the Metager data set contains only the top 4,000 search queries of a day during the observation period. 4.9 Mio. (8.4% of gross number) queries appeared only once and 1.37 Mio. (4.7%) twice in the Fireball query list. The distribution of queries in the Lycos list shows nearly the same characteristics (10.5% once and 5.5% twice).

Engine	$\|\|S\|\|_1$	M	$\|\|V\|\|_1$	N
Fireball	58.442.652	$8,349,176$	$104,940,963$	$3,114,105$
Lycos	76.343.559	$13,235,975$	$131,178,306$	$5,614,278$
Metager (top 4,000)	943.833	$146,566$	$1,529,383$	$121,039$
Metaspinner	2.174.823	$654,664$	$3,837,051$	$355,164$

Table 2. Numbers of queries and terms

The average appearance of search terms is much higher than that of queries, e.g., 33.7 times in the Fireball list (23.4 Lycos). The average number of terms in the meta search engines is nearly the half with 12.6 times in the Metager and 10.8 times in the Metaspinner list. This means that it is more probable that a term occurs a second time than a search query. In the Lycos list, 2.4% of the terms occurred only once and 0.6% twice (0.2% three times). Thus, there are a few terms which have a very large number of occurrences.
The average length of all queries ($||S||_1/||V||_1$) collected is 1.8 (Fireball), 1.7 (Lycos), 1.6 (Metager), and 1.8 (Metaspinner). Silverstein et al. (1999) calculated an average length of 2.4.
The average length of distinct search queries is a little bit longer, e.g., 2.4 (Fireball), 2.3 (Metager), and 2.2 (Lycos/Metaspinner). In the Excite studies, the length increased from 1.5 (1996) to 2.4 (1999).

3.2 Complexity of queries

All in all, operators were sparsely used in search queries. Less than 3% of all search queries in every search engine observed contained correctly applied operators. In the Excite studies, comparable figures were reported whereas Hölscher and Strube (2000) mentioned the use of operators in 25.3% of all queries submitted to Fireball. We explain this discrepancy to our findings with the fact that we differentiated between operators generated automatically by the search engines and those which were typed in by the users themselves. We only take into account those manually generated operators. The most popular operator used is the phrase search. In the Fireball (Lycos) list, 1,247,851 (1,320,517) queries represented a search for an exact phrase. This means that between 1.7% and 2.1% use this option to refine their search queries.

3.3 Topics and categories in search queries

The most favored topics searched for in general search engines could be labeled with 'erotic'. In the top 10 search queries of fireball eight referred to this topic. On the first place is 'sex'. In Lycos the top 10 queries could be categorized into the topics 'erotic' and 'search engine'. The topic 'search engine' contains queries such as 'lycos', 'ebay', or 'google'. The topics occurring in the top 10 of the meta search engines are more general. One finds here search queries with the German words for 'hotel', 'employee's car', or 'phone book'. The topic 'erotic' is only represented once with the query 'sex' on the fourth place (Metager).

The assumption of Spink et al. (2002) that people search rather for business topics than for erotic ones can't be approved but there is a difference between searches in meta and general search engines.

The Fireball and Metaspinner live search also show the categories as 'picture search', 'German search only', 'international search', or 'no special search area' which were selected to refine the search area. The 'German search only' option is the standard adjustment in both engines. A quarter of the search queries observed in the Fireball ticker were searches for pictures, an additional 66.1% showed the 'German search only' option (Metaspinner: 88.9%). This means that most users only search with the standard configuration settings of a search engine.

3.4 Vocabulary growth

Zien et al. (2000) discuss the vocabulary growth of search queries and terms depending on time according to the so-called *Heap's law* (Baeza-Yates and Ribeiro-Neto (1999)). This law describes the functionality between the number of words in a document and the document size. Here, the functionalities are the following:

$$M = \alpha_1 * ||\boldsymbol{S}||_1^{\beta_1} \tag{1}$$

and

$$N = \alpha_2 * ||\boldsymbol{V}||_1^{\beta_2}, \tag{2}$$

with $\alpha_1, \alpha_2 \in \mathbb{R}$ and $\beta_1, \beta_2 \in [0; 1]$. Thus, we plot $||\boldsymbol{S}||_1$ and $||\boldsymbol{V}||_1$ on the x-axis and M and N on the y-axis over time. In Figure 1, both functions of the Lycos vocabulary growth and their linear fitting functions are shown on a log_{10}/log_{10}-scale with the upper line being the bisector. We estimated $\beta_1 = 0.8848$ (middle curve) and $\beta_2 = 0.8136$ (lower curve). Zien et al. (2000) specify β_1 with 0.95 and β_2 with 0.69. This means that the number of terms doesn't increase as rapidly as the number of search queries. The fitting functions of the other search engines show almost the same characteristics.

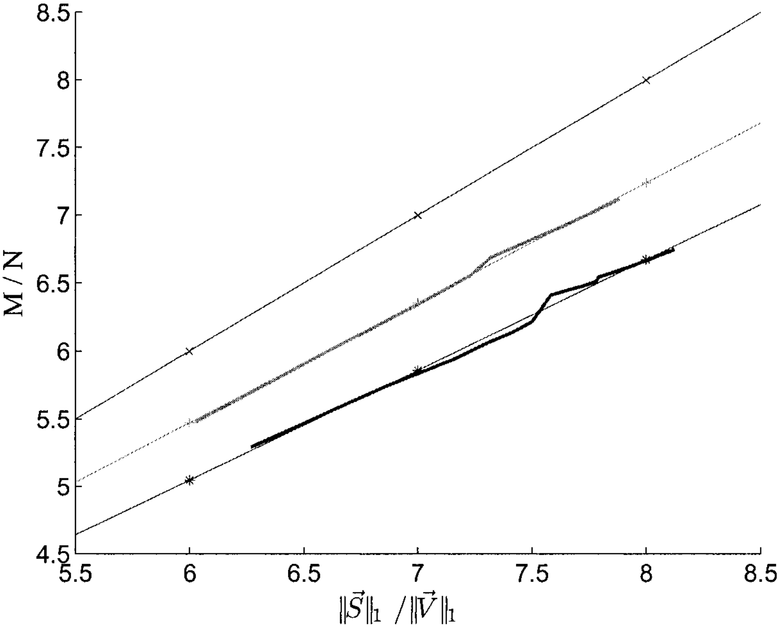


Fig. 1. Vocabulary growth of search queries and terms

3.5 Distribution of queries over time

In Figure 2, the average number of search queries per hour in percent of the average number of search queries per day are shown. One sees that all four search engine show the same behavior. The "quiet period" is between 5:00 a.m. and 6:00 a.m., and the traffic hours are between 11:00 a.m. and 9:00 p.m. The "rush hour" is between 2:00 p.m. and 3:00 p.m., right after lunch time.

Considering data with respect to weekdays, it has to be noted that after a peak on Monday, the number of search queries per day decreases until Saturday to rise again. We refer to this as the "Monday Effect". If one Monday is a public holiday there is no peak on that special day. For an Easter Monday, e.g., the peak appears on the day after.

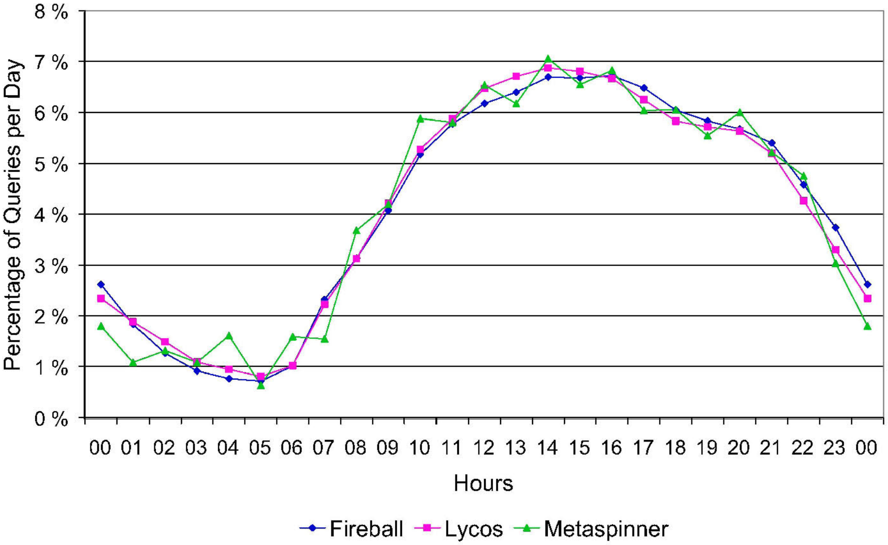


Fig. 2. Average number of search queries per hour

4 Summary of results

We closed the gap of knowledge with respect to German searching behavior by collecting search queries in live tickers of four different German (meta) search engines.

Recapitulating, search queries are very short and are not complex. The favored operator used by a searching person is the phrase search. The topic which appears most frequently in the top 10 of general search engines is 'erotic' whereas in the top 10 of meta search engines, more general topics are important. The 'erotic' topic is represented only once by the query 'sex'. Most search engine users conduct searches with the standard configuration settings.

By regarding the functionality between distinct terms/search queries and gross number of search queries per time, we see that the number of terms increases more slowly than the number of search queries.

The "quiet period" of search traffic is in the early morning and the "rush hour" is at noon with a peak after lunch time. On Mondays the volume of search queries conducted is higher than on other days with the minimum on Saturdays.

5 Managerial implications

Since search engine users do not deploy operators or special configuration settings, the search engine interface should be designed in a very simple way with a special section for experts. The preferred topic is 'erotic' so, it would be advisable to build an 'erotic' cache to serve such searches directly without affecting the 'normal' cache. Additionally, it is possible to cache the most frequently used terms to recommend the top web sites corresponding to these searches. In consideration of "rush hours" and "high frequency" days, administrative work on search engines should be done on early Saturday morning. Special events or search features such as 'news' should be realized on Mondays.

References

BAEZA-YATES, R. and RIBEIRO-NETO, B. (1999): *Modern Information Retrieval*, Addison Wesley, New York, NY.

BEITZEL, S.M., JENSEN, E.C., GROSSMAN, D., and FRIEDER, O. (2004): Hourly Analysis of a Very Large Topically Categorized Web Query Log. In: M. Sanderson, K. Jrvelin, J. Allan, and P. Bruza (Eds.): *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, ACM, 321–328.

HÖLSCHER, CH. and STRUBE, G. (2000): Web Search Behavior of Internet Experts and Newbies. In: H. Maurer and R.G. Olson (Eds.): *Proceedings of the Ninth International World Wide Web Conference*, North-Holland Publishing Co, Amsterdam, 337–346 [available online: http://www.www9.org/w9cdrom/81/81.html].

JANSEN, B., SPINK, A., and SARACEVIC, T. (2000): Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing and Management, 36(2), 207–227.*

SILVERSTEIN, C., HENZINGER, M., MARAIS, H., and MORICZ, M. (1999): Analysis of a Very Large Web Search Engine Query Log. *ACM SIGIR Forum, 33(1), 6–12.*

SPINK, A., JANSEN, B.J., WOLFRAM, D., and SARACEVIC, T. (2002): From E-Sex to E-Commerce: Web Search Changes. *IEEE Computer, 35(3), 107–109.*

SPINK, A., OZMUTLU, S., OZMUTLU, H.C., and JANSEN, B.J. (2002a): U.S. Versus European Web Searching Trends. *SIGIR Forum, 36(2), 32–38.*

SPINK, A., WOLFRAM, D., JANSEN, B.J., and SARACEVIC, T. (2001): Searching the Web: The Public and Their Queries. *Journal of the American Society for information science and technology, 52(3), 226–234.*

SPINK, A. and XU, J.L. (2000): Selected results from a large study of Web searching: the Excite study. *Information Research, 6(1)* [available online: http://InformationR.net/ir/6-1/paper90.html].

WOLFRAM, D., SPINK, A., JANSEN, B.J., and SARACEVIC, T. (2001): Vox Populi: The Public Searching of the Web. *Journal of the American Society for Information Science and Technology, 52(12), 1073–1074.*

ZIEN, J., MEYER, J., TOMLIN, J., and LIU, J. (2000): *Web Query Characteristics and their Implications on Search Engines.* Almaden Research Center, Research Report, RJ 10199 (95073).

Performance Drivers for Depth-First Frequent Pattern Mining

Lars Schmidt-Thieme[1] and Martin Schader[2]

[1] Institute for Computer Science, University of Freiburg,
Georges-Köhler-Allee 51, D-79 110 Freiburg, Germany

[2] Department of Information Systems, University of Mannheim,
Schloß, D-68 131 Mannheim, Germany

Abstract. Fast algorithms for mining frequent itemsets nowadays are highly optimized and specialized. Often they resemble the basic algorithms as Apriori and Eclat only faintly. But algorithms for other pattern domains as sequences etc. typically are built on top of the basic algorithms and thus cannot participate in improvements for highly specialized algorithms for itemsets.

Therefore, we would like to investigate different properties of a basic depth-first search algorithm, Eclat, and identify its performance drivers. We view Eclat as a basic algorithm and a bundle of optional algorithmic features that are taken partly from other algorithms like lcm and Apriori, partly new ones. We evaluate the performance impact of these different features and identify the best configuration of Eclat.

1 Introduction

Algorithms for mining frequent itemsets often are presented in a monolithic way and labeled with a fancy name for marketing. Careful inspection often reveals similarities with other mining algorithms that allow the transfer from a smart solution of a specific (detail) problem in one algorithm to another one. We would like to go one step further and view such mining algorithms as a basic algorithm and a bundle of algorithmic features.

Basically, there are only two large families of mining algorithms, Apriori (Agrawal and Srikant (1994)) and Eclat (Zaki (2000)) (counting fpgrowth (Han et al. (2000)) in the Eclat family what might be arguable). As the basic computation schemes of both of these algorithms are quite simple, one might get the impression, that nowadays they are good only as examples how mining algorithms work in principle for textbooks, but in practice more sophisticated algorithms have to be applied to get good performance results: for example, the four best-performing algorithms of the FIMI-03 workshop, patricia, kdci, lcm, and fpgrowth* (see Pietracaprina and Zandolin (2003), Orlando et al. (2003), Uno et al. (2003), Grahne and Zhu (2003)), for the implementations and Goethals and Zaki (2003b) for a performance evaluation of these algorithms, respectively) do use candidate generation procedures and data structures quite different from those usually associated with the basic algorithms.

Highly specialized algorithms typically cannot be transferred to other pattern domains as sequences, sequences with wildcards etc. Therefore, most algorithms for pattern domains other than itemsets have been developed starting from one of the basic algorithms, see e.g., Gaul and Schmidt-Thieme (2001) and Schmidt-Thieme and Gaul (2001). Consequently, these algorithms cannot participate in improvements for the itemsets domain.

In this paper, we would like to go one step towards a challenge of that point of view by investigating different performance drivers for Eclat and come up with a configuration of the Eclat algorithm that is much faster than other configurations at least for dense datasets.

We will start with a formal outline of Eclat algorithm in section 2. In section 3 we investigate several algorithmic features of Eclat, partly gathered from other algorithms as lcm, fpgrowth, and Apriori, partly new ones, review their usefulness in Eclat and shortly discuss their possible performance impact along with possible reasons thereof. In section 4 we present an empirical evaluation of that impact. – We will stick to Eclat. See Borgelt (2003) for an excellent discussion and evaluation of different features of Apriori.

Let us fix notations for the frequent itemset mining problem in the rest of this section. Let A be a set, called **set of items** or **alphabet**. Any subset $X \in \mathcal{P}(A)$ of A is called an **itemset**. Let $\mathcal{T} \subseteq \mathcal{P}(A)$ be a multiset of itemsets, called **transaction database**, and its elements $T \in \mathcal{T}$ called **transactions**. For a given itemset $X \in \mathcal{P}(A)$, the set of transactions that contain X

$$\mathcal{T}(X) := \{T \in \mathcal{T} \mid X \subseteq T\}$$

is called **(transaction) cover of X in $\mathcal{T}$** and its cardinality

$$\sup_{\mathcal{T}}(X) := |\mathcal{T}(X)|$$

(absolute) support of X in $\mathcal{T}$. An **(all) frequent itemset mining task** is specified by a dataset $\mathcal{T}$ and a lower bound minsup $\in \mathbb{N}$ on support, called **minimum support**, and asks for enumerating all itemsets with support at least minsup, called **frequent** or **(frequent) patterns**.

2 Basic Eclat algorithm

Most frequent itemset mining algorithms as Apriori (Agrawal and Srikant (1994)) and Eclat (Zaki (2000)) use a total order on the items A of the alphabet and the itemsets $\mathcal{P}(A)$ to prevent that the same itemset, called **candidate**, is checked twice for frequency. Items orderings $\leq$ are in one-to-one-correspondence with **item codings**, i.e., bijective maps $o : A \to \{1, \ldots, n\}$ via natural ordering on $\mathbb{N}$. – For itemsets $X, Y \in \mathcal{P}(A)$ one defines their **prefix** as

$$\begin{aligned}\text{prefix}(X,Y) :=& \{\{x \in X \mid x \leq z\} \mid \text{maximal } z \in X \cap Y : \\ & \{x \in X \mid x \leq z\} = \{y \in Y \mid y \leq z\}\}\end{aligned}$$

Any order on A uniquely determines a total order on $\mathcal{P}(A)$, called **lexicographic order**, by

$$X < Y :\Leftrightarrow \min(X \setminus \text{prefix}(X,Y)) < \min(Y \setminus \text{prefix}(X,Y))$$

For an itemset $X \in \mathcal{P}(A)$ an itemset $Y \in \mathcal{P}(A)$ with $X \subset Y$ and $X < Y$ is called an **extension of** X. An extension Y of X with $Y = X \cup \{y\}$ (and thus $y > \max X$) is called an **1-item-extension of** X. The extension relation organizes all itemsets in a tree, called **extension tree** or **search tree**.

Eclat starts with the empty prefix and the item-transaction incidence matrix $C_\emptyset$, shortly called **incidence matrix** in the following, and stored sparsely as list of item covers: $C_\emptyset := \{(x, \mathcal{T}(\{x\})) \,|\, x \in A\}$. The incidence matrix is filtered to only contain frequent items by

$$\text{freq}(C) := \{(x, \mathcal{T}_x) \,|\, (x, \mathcal{T}_x) \in C, |\mathcal{T}_x| \geq \text{minsup}\}.$$

that represent frequent 1-item-extensions of the prefix. For any prefix $p \in \mathcal{P}(A)$ and incidence matrix C of frequent 1-item-extensions of p one can compute the incidence matrix C_x of 1-item-extensions of $p \cup \{x\}$ by intersection rows:

$$C_x := \{(y, \mathcal{T}_x \cap \mathcal{T}_y) \,|\, (y, \mathcal{T}_y) \in C, y > x\}$$

where $(x, \mathcal{T}_x) \in C$ is the row representing $p \cup \{x\}$. C_x has to be filtered to get all frequent 1-item-extensions of $p \cup \{x\}$ and then this procedure is recursively iterated until the resulting incidence matrix C_x is empty, signaling that there are no further frequent 1-item-extensions of the prefix. See alg. 1 for an exact description of the Eclat algorithm.

3 Features of Eclat

The formal description of the Eclat algorithm in the last section allows us to point to several algorithmic features that this algorithm may have. These sometimes are described as implementation details, sometimes as extensions of Eclat, and sometimes as new algorithms.

3.1 Transaction recoding

Before the first incidence matrix $C_\emptyset$ is built, it is usually beneficial 1) to remove infrequent items from the transactions, 2) to recode the items in the transaction database s.t. they are sorted in a specific order, and 3) to sort the transaction in that order. As implementations usually use the natural order on item codes, item recoding affects the order in which candidates are checked. There are several recodings used in the literature and in existing implementations of Eclat and other algorithms as Apriori (see e.g., Borgelt (2003)). The most common codings are coding by increasing frequency and

Algorithm 1 Basic Eclat algorithm.

```
input: alphabet $A$ with ordering $\leq$,
  multiset $\mathcal{T} \subseteq \mathcal{P}(A)$ of sets of items,
  minimum support value minsup $\in \mathbb{N}$.
output: set $F$ of frequent itemsets and their support counts.
  $F := \{(\emptyset, |\mathcal{T}|)\}$.
  $C_\emptyset := \{(x, \mathcal{T}(\{x\})) \mid x \in A\}$.
  $C'_\emptyset := \mathrm{freq}(C_\emptyset) := \{(x, \mathcal{T}_x) \mid (x, \mathcal{T}_x) \in C_\emptyset,$
                                   $|\mathcal{T}_x| \geq \mathrm{minsup}\}$.
  $F := \{\emptyset\}$.
  addFrequentSupersets($\emptyset, C'_\emptyset$).
```

```
function addFrequentSupersets():
input: frequent itemset $p \in \mathcal{P}(A)$ called prefix,
  incidence matrix $C$ of frequent 1-item-extensions of $p$.
output: add all frequent extensions of $p$ to global variable $F$.
  for $(x, \mathcal{T}_x) \in C$ do
    $q := p \cup \{x\}$.
    $C_q := \{(y, \mathcal{T}_x \cap \mathcal{T}_y) \mid (y, \mathcal{T}_y) \in C, y > x\}$.
    $C'_q := \mathrm{freq}(C_q) := \{(y, \mathcal{T}_y) \mid (y, \mathcal{T}_y) \in C_q,$
                                   $|\mathcal{T}_y| \geq \mathrm{minsup}\}$.
    if $C'_q \neq \emptyset$ then
      addFrequentSupersets($q, C'_q$).
    end if
    $F := F \cup \{(q, |\mathcal{T}_x|)\}$.
  end for
```

coding by decreasing frequency. For Eclat in most cases recoding items by increasing frequency turns out to give better performance. Increasing frequency means that the length of the rows of the (initial) incidence matrix $C_\emptyset$ grows with increasing index. Let there be f_1 frequent items. As a row at index i is used $f_1 - i$ times at left side (x in the formulas above) and $i - 1$ times at right side (y in the formulas above) of the intersection operator, the order of rows is not important from the point of view of total usage in intersections. But assume the data is gray, i.e., the mining task does not contain any surprising associative patterns, where surprisingness of an itemset X is defined in terms of lift:

$$\mathrm{lift}(X) := \frac{\sup(X)}{|\mathcal{T}|} / \prod_{x \in X} \frac{\sup(\{x\})}{|\mathcal{T}|} /$$

$\mathrm{lift}(X) = 1$ means that X is found in the data exactly as often as expected from the frequencies of its items, $\mathrm{lift}(X) > 1$ or $\mathrm{lift}(X) < 1$ means that there is an associative or dissociative effect, i.e., it is observed more often or less often than expected. Now, if lift ≈ 1 for all or most patterns, as it is typically for benchmark datasets, then the best chances we have to identify a pattern

X as infrequent before we actually have counted its support, is to check its subpattern made up from its least frequent items. And that is exactly what recoding by increasing frequency does.

3.2 Types of incidence structures: Covers vs. diffsets

One of the major early improvements of Eclat algorithms has been the replacement of item covers in incidence matrices by their relative complement in its superpattern, so called **diffsets**, see Zaki and Gouda (2001). Instead of keeping track of $\mathcal{T}(q)$ for a pattern q, we keep track of $\mathcal{T}(p) \setminus \mathcal{T}(q)$ for its superpattern p, i.e., $q := p \cup \{x\}$ for an item $x > \max(p)$. $\mathcal{T}(p) \setminus \mathcal{T}(q)$ are those transactions we loose if we extend p to q, i.e., its additional **defect** relative to p. From an incidence matrix C of item covers and one of the 1-item-extensions $(x, \mathcal{T}_x) \in C$ of its prefix we can derive the incidence matrix D of item defects of this extension by

$$D_x := \{(y, T_x \setminus T_y) \mid (y, T_y) \in C, y > x\}$$

From an incidence matrix D of item defects and one of its 1-item-extensions $(x, \mathcal{T}_x) \in D$ of its prefix we can derive the incidence matrix D_x of item defects of this extension by

$$D_x := \{(y, T_y \setminus T_x) \mid (y, T_y) \in D, y > x\}$$

If we expand first by x and then by y in the second step, we loose transactions that not contain y unless we have lost them before as they did not contain x.

Defects computed from covers may have at most size

$$\text{maxdef}_p := |\mathcal{T}(p)| - \text{minsup},$$

those computed recursively from other defects at most size

$$\text{maxdef}_{p \cup \{x\}} := \text{maxdef}_p - |\mathcal{T}_x|$$

1-item-extensions exceeding that maximal defect are removed by a filter step:

$$\text{freq}(D) := \{(x, \mathcal{T}_x) \mid (x, \mathcal{T}_x) \in C, |\mathcal{T}_x| \leq \text{maxdef}\}.$$

Computing intersections of covers or set differences for defects are computationally equivalent complex tasks. Thus, the usage of defects can improve performance only by leading to smaller incidence matrices. For dense datasets where covers overlap considerably, intersection reduces the size of the incidence matrix only slowly, while defects cut down considerably. On the other side, for sparse data using defects may deteriorate the performance. – Common items in covers also can be removed by omitting equisupport extensions (see section 3.5).

While there is an efficient transition from covers to defects as given by the formula above, the reverse transition from defects to covers seems hard to perform efficiently as all defects on the path to the root of the search tree would have to be accumulated.

Regardless which type of incidence matrix is used, it can be stored as sparse matrix (i.e., as list of lists as discussed so far) or as dense (bit)matrix (used e.g, by Borgelt (2003)).

A third alternative for keeping track of item-transaction incidences is not to store item covers as a set of incident transaction IDs per 1-item-extension, but to store all transactions $\mathcal{T}(p)$ that contain a given prefix p in a trie (plus some index structure, known as frequent pattern tree and first used in fp-growth; see Han et al. (2000). Due to time restrictions, we will not pursue this alternative further here.

3.3 Incidence matrix derivation

For both incidence matrices, covers and defects, two different ways of computing the operator that derives an incidence matrix from a given incidence matrix recursively, i.e., intersection and set difference, respectively, can be chosen. The straightforward way is to implement both operators as set operators operating on the sets of transaction IDs.

Alternatively, intersection and difference of several sets $\mathcal{T}_y, y > x$ of transactions by another set $\mathcal{T}_x$ of transactions also can be computed in parallel using the original transaction database by counting in IDs of matching transactions (called occurrence deliver in Uno et al. (2003)). To compute $\mathcal{T}'_y := \mathcal{T}_y \cap \mathcal{T}_x$ for several $y > x$ one computes

$$\forall T \in \mathcal{T}_x \forall y \in T : \mathcal{T}'_y := \mathcal{T}'_y \cup \{T\}.$$

Similar, to compute $\mathcal{T}'_y := \mathcal{T}_x \setminus \mathcal{T}_y$ for several $y > x$ one computes

$$\forall T \in \mathcal{T}_x \forall y \notin T : \mathcal{T}'_y := \mathcal{T}'_y \cup \{T\}.$$

3.4 Initial incidence matrix

Basic Eclat first builds the incidence matrix $C_\emptyset$ of single item covers as initial incidence matrix and then recursively derives incidence matrices C_p of covers of increasing prefixes p or D_p of defects.

Obviously, one also can start with $D_\emptyset$, the matrix of item cover complements. This seems only useful for very dense datasets as it basically inverts the encoding of item occurrence and non-occurrence (dualization).

It seems more interesting to start already with incidence matrices for 1-item-prefixes, i.e., not to use Eclat computation schemes for the computation of frequent pairs, but count them directly from the transaction data. For Apriori this is a standard procedure. The cover incidence matrix $C_x = \{(y, \mathcal{T}_y)\}$

for an frequent item x, i.e., $\mathcal{T}_y = \mathcal{T}(\{x\}) \cap \mathcal{T}(\{y\})$, is computed as follows:

$$\forall T \in \mathcal{T} : \text{if } x \in T : \forall y \in T, y > x : \mathcal{T}_y := \mathcal{T}_y \cup \{T\}.$$

The test for $x \in T$ looks worse than it is in practice: if transactions are sorted, items x are processed in increasing order, and deleted from the transaction database after computation of C_x, then if x is contained in a transaction T it has to be its first item.

Similarly, a defect incidence matrix $D_x = \{(y, \mathcal{T}_y)\}$ for a frequent item x, i.e., $\mathcal{T}_y = \mathcal{T}(\{x\}) \setminus \mathcal{T}(\{y\})$, can be computed directly from the transaction database by

$$\forall T \in \mathcal{T} : \text{if } x \in T : \forall y \notin T, y > x : \mathcal{T}_y := \mathcal{T}_y \cup \{T\}.$$

If C_x or D_x is computed directly from the transaction database, then it has to be filtered afterwards to remove infrequent extensions. An additional pass over $\mathcal{T}$ in advance can count pair frequencies for all x, y in parallel, so that unnecessary creation of covers or defects of infrequent extensions can be avoided.

3.5 Omission of equisupport extensions

Whenever an extension x has the same support as its prefix p, it is contained in the closure $\bigcap \mathcal{T}(p)$ of the prefix. That means that one can add any such equisupport extension to any extension of p without changing its support; thus, one can omit to explicitly check its extensions. Equisupport extensions can be filtered out and kept in a separate list E for the active branch: whenever an itemset X is output, all its $2^{|E|}$ supersets $X' \subseteq X \cup E$ are also output.

Omission of equisupport extensions is extremely cheap to implement as it can be included in the filtering step that has to check support values anyway. For dense datasets with many equisupport extensions, the number of candidates that have to be checked and accordingly the runtime can be reduced drastically.

3.6 Interleaving incidence matrix computation and filtering

When the intersection $\mathcal{T}_x \cap T_y$ of two sets of transaction IDs is computed, we are interested in the result of this computation only if it is at least of size minsup, as otherwise it is filtered out in the next step. As the sets of transactions are sorted, intersections are computed by iterating over the lists of transaction IDs and comparing items. Once one of the tails of the lists to intersect is shorter than minsup minus the length of the intersection so far, we can stop and drop that candidate, as it never can become frequent. – For set difference of maximal length maxdef a completely analogous procedure can be used.

3.7 Omission of final incidence matrix derivation

Finally, once the incidence matrix has only two rows, the result of the next incidence matrix derivation will be an incidence matrix with a single row. As this is only checked for frequency, but its items are not used any further, we can omit to generate the list of transaction IDs and just count its length.

4 Evaluation

By evaluating different features of Eclat we wanted to answer the following question: What features will make Eclat run fastest? Especially, what is its marginal runtime improvement of each feature in a sophisticated Eclat implementation?

To answer this question about the runtime improvement of the different features, we implemented a modular version of Eclat in C++ (basically mostly plain C) that allows the flexible inclusion or exclusion of different algorithmic features. At the time of writing the following features are implemented: the incidence structure types covers and diffsets (COV, DIFF), transaction recoding (none, decreasing, increasing; NREC, RECDEC, RECINC), omission of equisupport extensions (NEE), interleaving incidence matrix computation and filtering (IFILT), and omission of final incidence matrix (NFIN). As initial incidence matrix alway covers of frequent 1-itemsets ($C_\emptyset$) was used.

To measure the marginal runtime improvement of a feature we configured a sophisticated Eclat algorithm with all features turned on (SOPH:= DIFF, RECINC, NEE+, IFILT+, NFIN+) and additionally for each feature an Eclat algorithm derived from SOPH by omitting this feature (SOPH-DIFF, SOPH-RECINC (decreasing encoding), SOPH-REC (no recoding at all), SOPH-NEE+, SOPH-IFILT+, SOPH-NFIN+).

We used several of the data sets and mining tasks that have been used in the FIMI-03 workshop (Goethals and Zaki (2003a)): accidents, chess, connect, kosarak, mushroom, pumsb, pumsbstar, retail, T10I5N1KP5KC0.25D200K, T20I10N1KP5KC0.25D200K, and T30I15N1KP5KC0.25D200K. All experiments are ran on a standard Linux box (P4/2MHz, 1.5GB RAM, SuSE 9.0). Jobs were killed if they run more than 1000 seconds and the corresponding datapoint is missing in the charts.

A sample from the results of these experiments can be seen in Figure 1 (the remaining charts can be found at http://www.informatik.uni-freiburg.de/-cgnm/papers/gaul60). One can see some common behavior across datasets and mining tasks:

- For dense mining tasks like accidents, chess, etc. SOPH is the best configuration.
- For sparse mining tasks like T20I10N1KP5KC0-25D200K etc. SOPH-diff is the best configuration, i.e., using defects harms performance here – both effects are rather distinct.

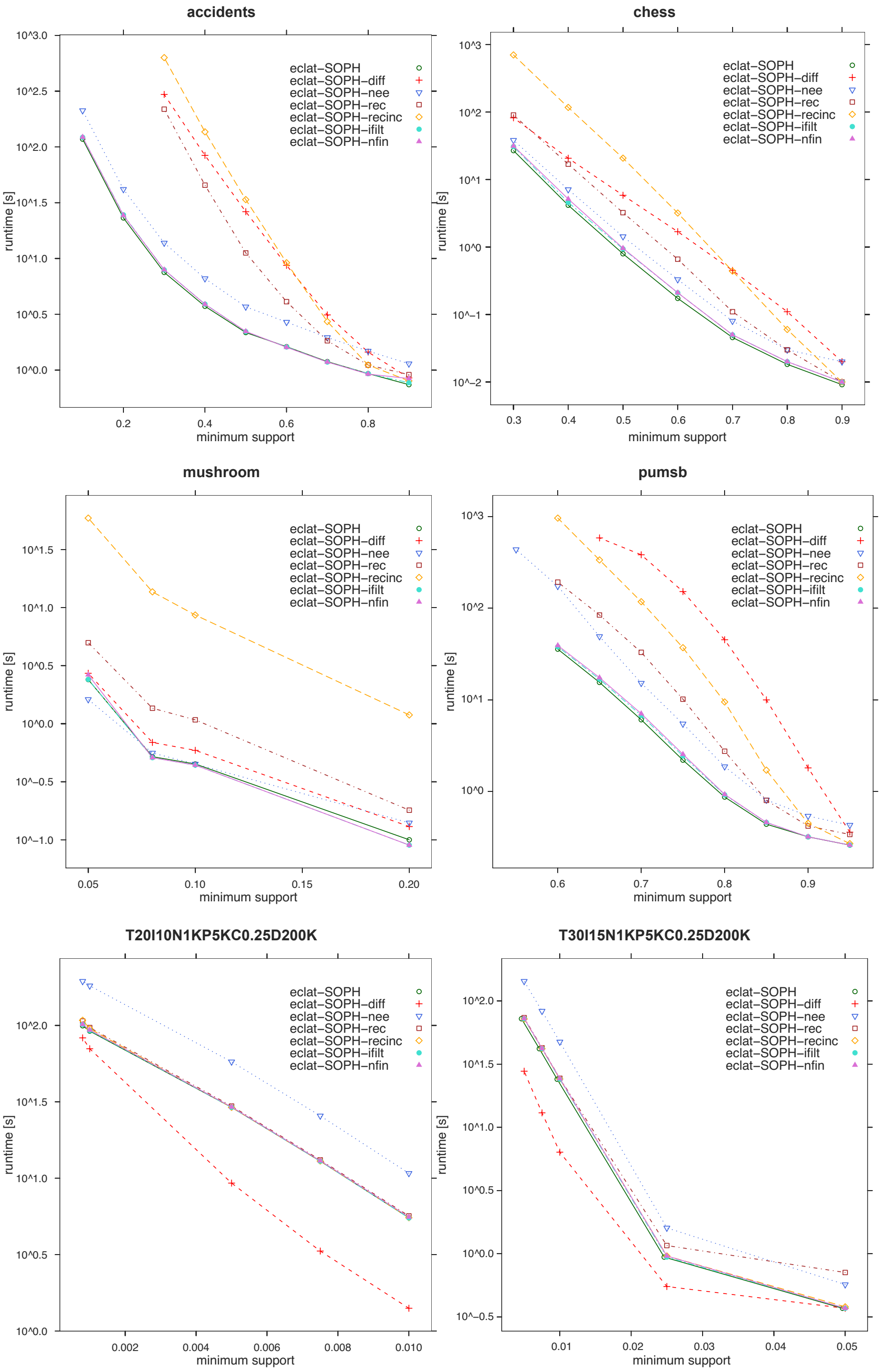


Fig. 1. Evaluation of the marginal effect of different features of Eclat on runtime.

- Recoding is important and shows a huge variety w.r.t. runtime: compare e.g., decreasing and no encoding for connect: the natural encoding is not much worse than decreasing encoding, but the curve for increasing encoding shows what harm the wrong encoding can do: note that the natural encoding is close to optimal only by mere chance and could be anywhere between increasing and decreasing!
- Omitting equisupport extensions also shows a clear benefit for most mining tasks, with exception for mushroom.
- Compared with other features, the impact of the features IFILT and NFIN is neglectible.

5 Outlook

There are at least four more features we do not have investigated yet: using tries to store the transaction covers, the method to compute the initial incidence matrix, pruning, and memory management. Our further research will try to address questions about the impact of these features.

Having identified the performance drivers for depth-first frequent pattern mining by Eclat, in a follow-up paper we will compare the best configuration of this algorithm with other algorithms in this field. We have already preliminary results that show that contrary to expectations Eclat is one of the fastest algorithms in the area if it is configured optimally.

References

AGRAWAL, R. and SRIKANT, R. (1994): Fast Algorithms for Mining Association Rules. In: J.B. Bocca, M . Jarke, and C. Zaniolo (Eds.): *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Santiago de Chile, September 12-15,* Morgan Kaufmann, 487–499.

BORGELT, Ch. (2003): Efficient Implementations of Apriori and Eclat, in Goethals and Zaki (2003a).

GAUL, W. and SCHMIDT-THIEME, L. (2001): Mining Generalized Association Rules for Sequential and Path Data. *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM), San Jose, 593–596.*

GOETHALS, B. and ZAKI, M. (eds., 2003a): *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations,* Melbourne, Florida, USA, November 19, 2003.

GOETHALS, B. and ZAKI, M. (2003b): Advances in Frequent Itemset Mining Implementations: Introduction to FIMI03, in Goethals and Zaki (2003a).

GRAHNE, G. and ZHU, J. (2003): Efficiently Using Prefix-trees in Mining Frequent Itemsets, in Goethals and Zaki (2003a).

HAN, J., PEI, J., and YIN, Y. (2000): Mining frequent patterns without candidate generation, in W. Chen, J.F. Naughton, and P.A. Bernstein (Eds.): *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data,* ACM Press, 1–12.

ORLANDO, S., LUCCHESE, C., PALMERINI, P., PEREGO, R., and SILVESTRI, F. (2003): kDCI: a Multi-Strategy Algorithm for Mining Frequent Sets, in Goethals and Zaki (2003a).

PIETRACAPRINA, A. and ZANDOLIN, D. (2003): Mining Frequent Itemsets using Patricia Tries, in Goethals and Zaki (2003a).

SCHMIDT-THIEME, L. and GAUL, W. (2001): Frequent Substructures in Web Usage Data — A Unified Approach, *Proceedings of the Web Mining Workshop, First SIAM International Conference on Data Mining 2001 (ICDM), Chicago.*

UNO, T., ASAI, T., UCHIDA, Y., and ARIMURA, H. (2003): LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets, in Goethals and Zaki (2003a).

ZAKI, M. (2000): Scalable Algorithms for Association Mining, *IEEE Transactions on Knowledge and Data Engineering 12/3, 372–390.*

ZAKI, M. and GOUDA, K. (2001): Fast Vertical Mining Using Diffsets, RPI Tech. Report. 01-1.

On the Performance of Algorithms for Two-Mode Hierarchical Cluster Analysis – Results from a Monte Carlo Simulation Study

Manfred Schwaiger[1] and Raimund Rix[2]

[1] Institut für Unternehmensentwicklung und Organisation,
Ludwig-Maximilians-Universität München, D-80539 München, Germany
[2] Fujitsu Siemens Computers,
Corporate Strategy, D-80807 München, Germany

Abstract. A variety of algorithms do exist to perform two-mode cluster analysis, normally leading to different results based on the same data matrix. In this article we evaluate hierarchical two-mode algorithms by means of a large simulation study using a dominant performance measure.

To cover the wide spectrum of possible applications, synthetic classifications are created and processed with Monte-Carlo simulations. The appropriateness of the algorithms applied is then evaluated, showing that all algorithms are pretty well able to recover the original classification, if no perturbations are applied on the data base. Otherwise, considerable performance differences can be shown.

1 Motivation

Two-mode cluster analysis has proven to be a useful tool. By simultaneously assigning not only objects (first mode[1]), but also attributes (second mode) to a cluster, the user gets a description of the resulting classification by just looking at the cluster attributes. Furthermore, during the classification procedure, these attributes that are assigned to a cluster at an early stage, are repelling objects that are less associated to them.

Several different two-mode cluster analysis algorithms have been developed, confer Schwaiger (1997) for an overview. Even if their capabilities are well documented in single examples, until now, they have been rarely used in practical applications. One reason may be that the quality of the algorithms has not been evaluated in general. Under quality we subsume to what extent an algorithm is able to reproduce the 'correct' (we will precise this later) two-mode classification. Rix and Schwaiger (2003) have shown that general criticism usually emphasizes the fact that different algorithms result in different classifications and no suitable goodness-of-fit measure is available to identify the best result.

[1] For simplicity reason we set the term 'object' equivalent for the first mode or row elements and the term 'attributes' for the second mode or column elements, knowing well that this is correct in the case of association data only.

The aim of this paper is to evaluate classifications resulting from different algorithms by means of a large simulation study. We use Monte-Carlo simulations to create synthetic classifications and apply several algorithms to these data bases. The algorithms are evaluated by the degree to which they can reproduce the original classification, that is, merge column and row elements of the data matrix correctly. The goodness-of-recovery measure we use is the Adjusted Rand Index.

From an ideal two-mode classification we would expect the objects and attributes within a single cluster to be similar to each other and the objects and attributes within two different clusters to be somewhat dissimilar. Moreover, within the cluster the objects have to be well associated to the attributes.

Between different clusters the objects of a cluster have to be dissimilar to the objects of another cluster, the attributes of a cluster have to be dissimilar to the attributes of another cluster, and the objects of one cluster have to be less associated to the attributes of another cluster. We will refer to this as the "six criteria of a two-mode classification" later on.

The objects and attributes are classified by means of the attribute values the objects show. Therefore, for every attribute A_j an attribute value x_{ij} has to exist for every object O_i, resulting in a data matrix $\mathbf{X} = (x_{ij})_{n \times m}$ containing n row elements (objects) and m column elements (attributes). In order to eliminate scale bias, the data have to be z-standardized.

2 Algorithms and goodness-of-fit measures

Two-mode classification algorithms can be divided into hierarchical and non-hierarchical methods (Arabie and Carroll (1980), Arabie et al. (1981), DeSarbo (1982), DeSarbo and De Soete (1984), Both and Gaul (1985), Arabie et al. (1988), Arabie et al. (1990), Duffy and Quiroz (1991), Baier et al. (1997)). For applications, the analysis of the fusion process provides valuable insights that one cannot get through the inspection of the resulting clusters on its own. Because this fusion process is revealed by hierarchical algorithms only, we evaluate their performance only.

The Missing Value algorithms (MV) were developed by Espejo and Gaul (1986). Like other two-mode procedures, they use a grand matrix containing distances between pairs of objects, pairs of attributes and between objects and attributes. Please refer to the given reference for formal details. Based on the grand matrix, (slightly modified) well documented one-mode cluster analysis algorithms as single, average, or complete linkage may be applied. Using the common agglomerative procedures the result is a hierarchy.

As a restraining condition we may notice, that it is not allowed to build pure object or pure attribute clusters.

As comparisons have shown a perfect agreement of MV and 2MLS (De Soete (1984), De Soete et al. (1984)) dendrogramms (Eckes and Orlik 1993), but the MV algorithms do not show estimation problems connected with the determination of the increment value within gradient-method and do not suffer from numeric problems (as they do not use a penalty-function), we prefer them to the 2MLS-algorithms and omit 2MLS in our evaluation study.

The Centroid Effect Method (CEM), an agglomerative algorithm merging at each step those two clusters whose merger minimally increases an internal heterogeneity measure, was developed by Eckes and Orlik (1991, 1993), Eckes (1991, 1993, 1995)). Elements are clustered in a way that minimizes the increase of an internal heterogeneity measure.

The only difference between the MV AL algorithm and the CEM is that the CEM calculates the difference between an object and an attribute as $(\mu - x_{ij})^2$, whereas the MV AL algorithm uses the non-squared difference $(\mu - x_{ij})$. As can be seen from the simulation results, the performances of the MV AL algorithm and the CEM are quite similar.

The last group of algorithms we'd like to test are the ESOCLUS Algorithms (Schwaiger (1996, 1997)), an extension of the MV algorithms, replacing the missing values in the grand matrix by the corresponding weighted city block distances. As these are generally of different orders of magnitude, every distance is divided through a block-specific percentile. An α-percentile p_α indicates a percentage of $\alpha\cdot$ 100% being less or equal to p_α. In the following sections the ESOCLUS algorithms using the block-specific maximum for normalization are marked as ESOCLUS M and the ones using the block-specific 50% percentile are marked as ESOCLUS P.

As opposed to CEM and MV algorithms, pure object and pure attribute clusters are allowed here.

3 Simulations

3.1 Set-up

To evaluate the performance of the two-mode hierarchical cluster analysis algorithms we did 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. One cycle of the simulation can be divided into the four steps shown in figure 1.

To test the algorithms under different conditions, the simulations were done with different presettings. For each presetting, 1 000 simulation cycles were processed and the average goodness-of-recovery value was calculated. Due to the large number of cycles, one gets a stochastically reliable mean value.

To create an artificial classification and its data matrix we started setting the total number of objects, attributes and clusters, and the range of the attribute values. The objects and attributes were assigned to the clusters in such a way that every cluster shows approximately the same number of objects and the same number of attributes.

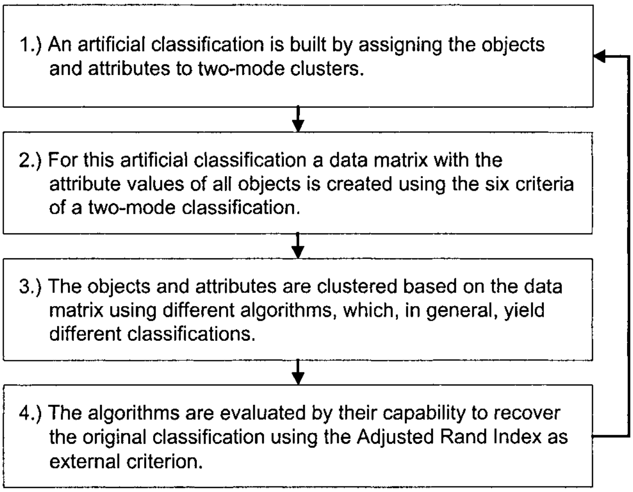

1.) An artificial classification is built by assigning the objects and attributes to two-mode clusters.
2.) For this artificial classification a data matrix with the attribute values of all objects is created using the six criteria of a two-mode classification.
3.) The objects and attributes are clustered based on the data matrix using different algorithms, which, in general, yield different classifications.
4.) The algorithms are evaluated by their capability to recover the original classification using the Adjusted Rand Index as external criterion.

Fig. 1. Four steps of a simulation cycle.

The simulation set-up is thoroughly described in Rix (2003). As a first step result we achieved ideal classification and corresponding data matrices. In a second step, perturbations were applied to get a more realistic classification and data matrix: It is unlikely for a real classification problem that all clusters might have nearly the same number of objects and attributes. To take this into consideration within the simulation, the number of objects and the number of attributes were set randomly according to a Gaussian distribution around an average cluster size with a standard deviation that is proportional to the average cluster size. It was taken into account that the sum of the number of objects and attributes respective of all clusters is equal to the total number of objects and attributes respectively.

As well, it is unlikely that in real applications all clusters must have objects and attributes simultaneously. Therefore, next to the mixed clusters, one-mode clusters C_l with objects only $(A(C_l) = \emptyset)$ or one-mode clusters C_k with attributes only $(O(C_k) = \emptyset)$ were introduced. The total number of clusters is the sum of the numbers of pure object, pure attribute, and mixed clusters.

As a further perturbation, errors in the attribute values may exist. This is considered within the simulation through an injection of random values into the data matrix.

Finally, sometimes solitary objects and attributes may exist with attribute values not fitting to the classification pattern at all. A solitary object has attribute values independent of the cluster the attribute is part of. A solitary attribute has attribute values independent of the cluster an object is part of. Therefore, in the simulation all attribute values of the solitary objects and attributes were set randomly.

For the evaluation of the result of an algorithm, the Adjusted Rand Index (ARI) by Hubert and Arabie (1986, p. 198) was used as an external goodness-of-recovery measure. It indicates how similar the result of an algorithm is compared to the original classification. Only objects and attributes, that in the original classification are part of a cluster, were taken into account. Solitary objects and attributes were not considered.

The Adjusted Rand Index has a value of one for a perfect recovery of the original classification and an average value of zero for random classifications independent on 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 1 000 simulation cycles is used.

3.2 Results

To survey the dependency of the algorithms on the different parameters, we set default values for all parameters (40 objects, 20 attributes, 5 clusters, attribute value distribution was 20% of the cluster width, cluster distance was 30% of the cluster width). Moreover, in a first stage, we changed just the value of one parameter each to find out the dependency on exactly this parameter.

Due to space limits we may not illustrate all our results. For a close look at the dependency of the performance of the algorithms on the various parameters within the separate simulation runs we may refer to Rix (2003, p. 91-121). In brief, the performance of the algorithms is nearly independent of the number of attributes, the number of objects, and the distance between clusters. Slight dependencies were found with respect to the number of clusters[2] and the standard deviation of the attribute value distribution in ratio to the cluster width[3].

Table 1 gives an overview on the performance of the algorithms. The average goodness-of-recovery value over 1 000 classifications is given, where one parameter is subject to random choice within the named range of the parameter for each classification. For every parameter, all algorithms can recover the original classification quite well. To test the performance, independent of the value of the parameters, we created 10 000 classifications. All parameters received random values out of the specified ranges, and as we can see in the last column of Table 1, similar result were obtained.

The simulations show that Espejo's and Gaul's missing value average linkage algorithm and the centroid effect method turned out to reveal superior performance when data are clean.

[2] The less clusters there are, the better the performance of the algorithms is.

[3] As expected, the performance of algorithms is slightly better using smaller deviations.

Mean ARI	Number of Attributes	Number of Objects	Number of Clusters	Range of Attribute Values	Attribute Value Distribution	Cluster Distance	All Parameters variable
ESOCLUS M SL	.9106	.9217	.9218	.9153	.9193	.9168	.9214
ESOCLUS M AL	.9507	.9589	.9604	.9525	.9533	.9526	.9553
ESOCLUS M CL	.8949	.9045	.9159	.8935	.9043	.9067	.9037
ESOCLUS P SL	.9567	.9581	.9659	.9543	.9536	.9544	.9547
ESOCLUS P AL	.9759	.9766	.9843	.9758	.9792	.9746	.9790
ESOCLUS P CL	.9502	.9543	.9595	.9448	.9569	.9496	.9542
MV SL	.9818	.9879	.9717	.9921	.9933	.9915	.9689
MV AL	**.9948**	**.9971**	**.9951**	**.9996**	**.9972**	**1**	**.9930**
MV CL	.9619	.9669	.9732	.9667	.9691	.9624	.9593
CEM	**.9948**	**.9963**	**.9951**	**.9996**	**.9964**	**.9983**	**.9924**
Max	.9974	.9991	.9983	1	1	1	.9984

Table 1. ARI-Values Achieved by Algorithms without Perturbations

We now know which algorithms to choose for ideal classification data. However, in real applications perturbations are present. Therefore, we took the same approach for all perturbations, but this time we did not use standard values for the parameters of the ideal classification. Instead, for each classification we set all parameters randomly. This enables us to reveal the dependency on a perturbation independent on the specific parameter set of the ideal classification.

Table 2 gives a survey on the performance of the algorithms when perturbations are placed. The average goodness-of-recovery values over 1 000 classifications are given with just the level of one perturbation chosen randomly. We set the level of the cluster size distribution randomly between 0 and 100%, the fraction of object clusters randomly between 0 and 60%, the fraction of attribute clusters randomly between 0 and 20%, the errors of attribute values randomly between 0 and 3%, the fraction of solitary objects randomly between 0 and 3%, and the fraction of solitary attributes randomly between 0 and 50%.

Now, the ESOCLUS P AL algorithm outperforms the other algorithms for various types of perturbations. To check if this result is still valid if different perturbations are present, we simultaneously averaged the goodness-of-recovery measure over 10 000 classifications. For each classification, we set the parameters of an ideal classification and the levels of all perturbations randomly. As you can see in the last column of Table 2, the result remains as given above.

Mean ARI	Distribution of Cluster Size	Object Clusters	Attribute Clusters	Errors of Attribute Values	Solitary Objects	Solitary Attributes	All perturbations randomly
ESOCLUS M SL	.7907	.8500	.8809	.4425	.5476	.4180	.1531
ESOCLUS M AL	.8337	.8472	.9188	.6100	.5783	.7921	.4342
ESOCLUS M CL	.8001	.7624	.8612	.5338	.5643	.6558	.4166
ESOCLUS P SL	.8683	.9201	.9112	.5313	.7135	.4368	.2009
ESOCLUS P AL	**.9429**	**.9271**	**.9413**	**.8244**	**.8521**	**.8522**	**.6710**
ESOCLUS P CL	.9266	.8472	.9072	.7462	.7888	.6275	.5980
MV SL	.7890	.4998	.8963	.4466	.5146	.2898	.0572
MV AL	.9307	.6850	.9357	.8036	.7122	.8191	.4187
MV CL	.9267	.6979	.9175	.6865	.7098	.5079	.3170
CEM	.9316	.6865	.9351	.7943	.7090	.7907	.4170
Max	.9871	.9498	.9631	.8725	.8901	.9166	.7340

Table 2. ARI-Values Achieved by Algorithms with Perturbations

4 Conclusion

In this paper the performance of hierarchical two-mode clustering algorithms was evaluated quantitatively for a wide spectrum of possible constellations. This was made possible by comparing the results of the algorithms to the original classification by means of an external goodness-of-recovery measure (ARI). The simulations have shown that without perturbations, most algorithms are quite able to well reproduce the original classification, whereby MV-algorithms have proven to outperform other methods. However, when the user may anticipate perturbations in the data, he might achieve more favourable results using the ESOCLUS P AL algorithm.

References

ARABIE, P. and CARROLL, J.D. (1980): MAPCLUS: A Mathematical Programming Approach to Fitting the ADCLUS Model. *Psychometrika, 45, 211–235.*

ARABIE, P., CARROLL, J.D., DESARBO, W., and WIND, J. (1981): Overlapping Clustering: A New Method for Product Positioning. *Journal of Marketing Research, 18, 310–317.*

ARABIE, P., HUBERT, L., and SCHLEUTERMANN, S. (1990): Blockmodels from the Bond Energy Algorithm. *Social Networks, 12, 99–126.*

ARABIE, P., SCHLEUTERMANN, S., DAWS, J., and HUBERT, L. (1988): Marketing Applications of Sequencing and Partitioning of Nonsymmetric and/or Twomode Matrices. In: W. Gaul and M. Schader (Eds.): *Data, Expert Knowledge, and Decisions.* Berlin, 215–224.

BAIER, D., GAUL, W., and SCHADER, M. (1997): Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring. In: F. Klar and O. Opitz (Eds.): *Classification and Knowledge Organisation.* Springer, Berlin et al., 557–566.

BOTH, M. and GAUL, W. (1985): PENCLUS: Penalty Clustering for Marketing Applications. Discussion Paper Nr. 82, Institut fuer Entscheidungstheorie und Unternehmensforschung, Karlsruhe University.

DE SOETE, G. (1984): Ultrametric Tree Representations of Incomplete Dissimilarity Data. *Journal of Classification, 1, 235–242.*

DE SOETE, G., DESARBO, W.S., FURNAS, G.W., and CARROLL, J.D. (1984): The Estimation of Ultrametric and Path Length Trees from Rectangular Proximity Data. *Psychometrika, 49, 289–310.*

DESARBO, W.S. (1982): GENNCLUS: New Models for General Nonhierarchical Clustering Analysis. *Psychometrika, 47, 449–475.*

DESARBO, W.S. and DE SOETE, G. (1984): On the Use of Hierarchical Clustering for the Analysis of Nonsymmetric Proximities. *Journal of Consumer Research, 11, 601–610.*

DUFFY, D.E. and QUIROZ, A.J. (1991): A Permutation Based Algorithm for Block Clustering. *Journal of Classification, 8, 65–91.*

ECKES, T. (1991): Bimodale Clusteranalyse: Methoden zur Klassifikation von Elementen zweier Mengen. *Zeitschrift fuer experimentelle und angewandte Psychologie, 38, 201–225.*

ECKES, T. (1993): Multimodale Clusteranalyse: Konzepte, Modelle, Anwendungen. In: L. Montada (Ed.): *Bericht ueber den 38. Kongress der Deutschen Gesellschaft fuer Psychologie in Trier 1992.* Hogrefe, Göettingen, 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 et al., 151–158.

ECKES, T. and ORLIK, P. (1991): An Agglomerative Method for Two-Mode Hierarchical Clustering. In: H. Bock and P. Ihm (Eds.): *Classification, Data Analysis and Knowledge Organisation.* Amsterdam, 3–8.

ECKES, T. and ORLIK, P. (1993): An Error Variance Approach To Two-Mode Hierarchical Clustering. *Journal of Classification, 10, 51–74.*

ESPEJO, E. and GAUL, W. (1986): Two-Mode Hierarchical Clustering as an Instrument for Marketing Research. In: W. Gaul and M. Schader (Eds.): *Classification as a Tool of Research.* Amsterdam, 121–128.

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

RIX, R. (2003): *Zweimodale hierarchische Clusteranalyse.* Deutscher Universitäts-Verlag, Wiesbaden.

RIX, R. and SCHWAIGER, M. (2003): Two-Mode Hierarchical Cluster Analysis - Evaluation of Goodness-of-Fit Measures, Münchner Betriebswirtschaftliche Beiträge, Heft 2003-05.

SCHWAIGER, M. (1996): Two-Mode Classification in Advertising Research. In: F. Klar and O. Opitz (Eds.) *Classification and Knowledge Organisation.* Springer, Berlin et al., 596–603.

SCHWAIGER, M. (1997): *Multivariate Werbewirkungskontrolle, Konzepte zur Auswertung von Werbetests.* Reihe neue betriebswirtschaftliche Forschung, 231, Gabler, Wiesbaden.

Clustering Including Dimensionality Reduction

Maurizio Vichi

Department of Statistics, Probability and Applied Statistics,
University "La Sapienza" of Rome, P.le Aldo Moro, 5, I-00185 Rome, Italy

Abstract. In this paper, new methodologies for clustering and dimensionality reduction of large data sets are illustrated. Two major types of data reduction methodologies are considered. The first are based on the simultaneous clustering of each mode of the observed multi-way data. The second are based on a clustering of the object mode to obtain mean profiles (centroids) and a factorial reduction of the other modes. These methodologies are described by a real application.

1 Introduction

In recent years it has become a standard practice to produce large data bases of multivariate observations involving high-dimensionality and forming multi-way arrays. To know how to mine and interpret all this information properly and quickly becomes an important issue that calls researchers to explore new methods including the reduction of 'dimensionalities' of data (e.g., objects, variables, situations, etc.), such that the relevant information is maintained as much as possible in the reduced data array.

The reduction and synthesis of a set of objects is typically achieved by a supervised or unsupervised classification algorithm or by a learning algorithm that yields classes of objects. Clustering methods can also been used for reducing the number of variables after defining specific proximity measures between variables. Even more frequently, such reduction of variables is obtained by (parametric or nonparametric) factorial techniques, e.g., a principal component analysis, and also by regression methods (including, e.g., PLS).

The dimensionality reduction is usually achieved by a procedure based on the sequential application of clustering and/or factorial techniques, thus reducing one by one each mode of the observed data sequentially after the other modes. In the case of a two-way two mode data matrix (objects by variables), a procedure called "tandem analysis" is frequently used. It starts first with a factorial technique, such as principal component analysis, to reduce the variable mode with few components. Then, on the component scores a partitioning algorithm is applied, such as k-means, to obtain a reduced set of mean profiles (centroid matrix) for the defined k clusters of the final partition of k-means. In this case the reduction of the two modes is achieved by using two different techniques: a cluster analysis for the objects and a factorial technique for the variables. These sequential procedures have the drawback to optimize two different criteria and therefore the components which maximize the total variance could no be the best linear combinations to identify the optimal partition of objects.

Alternatively a reduction both for objects and variables can be obtained by a unique clustering methodology applied twice. In fact, we could start first with the partition of the objects by using k-means to obtain a centroid matrix for the objects; then we could apply again k-means on the transpose of this centroid matrix, where these centroids should be weighted by the number of objects belonging to the classes of first k-means partition. Of course, in applying this second double k-means sequential procedure, variables expressed on the same scale are required, so that entries are comparable among both rows and columns. If this is not the case, data need to be rescaled by an appropriate standardization method. This procedure is particularly appropriate when objects and variables play a symmetric role in the data. For example, in marketing analysis, marketers are interested to know how the market can be segmented into homogeneous classes of consumers (objects) according to their preference on products (variables); at the same time, marketers may wish to know how products (variables) are clustered according to preferences of customers (objects).

In this case the proposed methodology will perform all in all a *two-mode partitioning*. The basic idea of two-mode partitioning is to identify *blocks*, i.e., sub-matrices of the observed data matrix, where objects and variables forming each block specify an *object cluster* and a *variable cluster* (Figure 1).

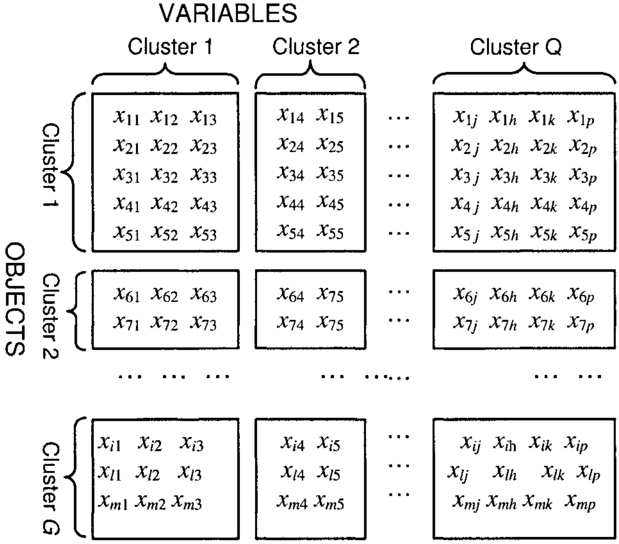

VARIABLES
Cluster 1
Cluster 2
Cluster Q
OBJECTS
Cluster 1
Cluster 2
Cluster G
x_{11} x_{12} x_{13} | x_{14} x_{15} | ... | x_{1j} x_{1h} x_{1k} x_{1p}
x_{21} x_{22} x_{23} | x_{24} x_{25} | ... | x_{2j} x_{2h} x_{2k} x_{2p}
x_{31} x_{32} x_{33} | x_{34} x_{35} | ... | x_{3j} x_{3h} x_{3k} x_{3p}
x_{41} x_{42} x_{43} | x_{44} x_{45} | ... | x_{4j} x_{4h} x_{4k} x_{4p}
x_{51} x_{52} x_{53} | x_{54} x_{55} | ... | x_{5j} x_{5h} x_{5k} x_{5p}
x_{61} x_{62} x_{63} | x_{64} x_{75} | ... | x_{6j} x_{6h} x_{6k} x_{6p}
x_{71} x_{72} x_{73} | x_{74} x_{75} | ... | x_{7j} x_{7h} x_{7k} x_{7p}
... ...
x_{i1} x_{i2} x_{i3} | x_{i4} x_{i5} | ... | x_{ij} x_{ih} x_{ik} x_{ip}
x_{l1} x_{l2} x_{l3} | x_{l4} x_{l5} | ... | x_{lj} x_{lh} x_{lk} x_{lp}
x_{m1} x_{m2} x_{m3} | x_{m4} x_{m5} | ... | x_{mj} x_{mh} x_{mk} x_{mp}

Fig. 1. Block partitioning. The data matrix **X** is partitioned into blocks

Thus, researchers would identify disjoint blocks where observations are perceived as similar to one another within each block, but most dissimilar between blocks. In the case of similarity data several models have been proposed for overlapping and non-overlapping clustering. After the paper on GENNCLUS algorithm (DeSarbo (1982)), that extended the ADCLUS model of Arabie and Carroll (1980), several very important papers have been presented by Prof. Wolfgang Gaul and his colleagues. Among others I like to remember: PENCLUS (Both and Gaul (1987)), AE algorithm (Gaul and Schader (1996)), and the new model for overlapping clustering (Baier et al. (1997)). The interested reader can find a very complete structured overview of two-mode clustering methods in Van Mechelen et al. (2004).

In this paper we show the performance of a new methodology for two mode partitioning of two way data recently proposed. In particular, the "double k-means" (Vichi (2000), Rocci and Vichi (2004)) for two-way data is discussed and compared with procedures that can be obtained by applying ordinary clustering techniques in repeated steps. Recently, Vichi and Martella (2005) have studied the maximum likelihood clustering estimation of the double k-means parameters.

As we have already observed above, clustering of objects and variables according to *double k-means* is particularly valuable when variables are not so distinct from objects as in the case above described when customers and products are considered. In this situation centroids for both objects and variables (e.g., mean profiles of customers and mean profiles of products) can be used to synthesize the observed data matrix. Another relevant case relates to genes expression microarray data where researchers are interested both in the homogeneous clustering of genes and samples of tissue.

However, for a usual multivariate data matrix a reduction of the objects is generally given by means of centroids from partitioning methodology, while a reduction of the variables is achieved by a factorial methodology as PCA, hence by means of linear combinations that give different weights to the original variables and accounts for the largest total variance of the observed data.

However, PCA, but also other factorial techniques, have often the drawback that different factors are characterized by the same original variables, so that the interpretation of these factors becomes a relevant and complex problem. In this situation it would be useful to partition objects into clusters summarized by centroids, but also to partition variables into clusters of correlated variables, summarized by linear combinations of maximum variance as it is obtained in clustering and disjoint principal component analysis (CDPCA) (Vichi and Saporta (2004)). This methodology can be seen as a generalization of the double k-means, as it will be discussed in the following sessions.

2 The clustering and dimensionality reduction model

The double k-means model (Vichi (2000)) is formally specified as follows

$$\mathbf{X} = \mathbf{U}\overline{\mathbf{Y}}\mathbf{V}' + \mathbf{E}, \tag{1}$$

where $\mathbf{X}$ is a $(I \times J)$ data matrix, while matrix $\mathbf{E}$ is the error component matrix. Matrix $\mathbf{U} = [u_{ij}]$ is a $(I \times P)$ membership matrix, assuming values $\{0, 1\}$, specifying for each object i its membership to a class of the partition of objects in P classes. Matrix $\mathbf{V} = [v_{jq}]$ is a $(J \times Q)$ membership matrix, assuming values $\{0, 1\}$, specifying for each variable j its membership to a class of the partition of variables in Q classes. Matrix $\overline{\mathbf{Y}} = [\overline{y}_{pq}]$ is the $(P \times Q)$ *centroid* matrix where $\overline{y}_{pq}$ denotes the mean of values corresponding to object and variable clusters p and q, respectively. This model, designed for two mode data, can be seen as a special case of the GENNCLUS model of De Sarbo (1982), specified for similarity data.

The first term in model (1) pertains to the portion of information of $\mathbf{X}$ that can be explained by the simultaneous classification of objects and variables. Of course, matrix $\mathbf{X}$ is supposed to be column standardized if the variables are not commensurate. In the papers by Vichi (2000) and Rocci and Vichi (2004), fast alternating least-square algorithms are proposed in the case the model is estimated with the least-squares approach, while recently Vichi and Martella (2005) estimate parameters of the model according to a model-based likelihood approach.

The double k-means model can be modified to assess a partition of the objects along a set of centroids, as above, but also a partition of the variables along a set of linear combinations of maximum variance. Thus the model (1) is written (Vichi and Saporta (2004))

$$\mathbf{X} = \mathbf{U}\overline{\mathbf{Y}}\mathbf{V}'\mathbf{B} + \mathbf{E}, \tag{2}$$

where matrix $\mathbf{B}$ is a diagonal matrix defined so that $\mathbf{V}'\mathbf{B}\mathbf{B}\mathbf{V} = \mathbf{I}_Q$ and $tr(\mathbf{B}\mathbf{B}) = Q$. An efficient alternating least-square algorithm is given.

Now matrix $\mathbf{V}'\mathbf{B}$ actually represents a particular component loading matrix. If we relax the constraint of matrix $\mathbf{V}$ to be binary and row stochastic, but still maintaining $\mathbf{B}\mathbf{V} = \mathbf{A}$ to be orthonormal, model (2) is

$$\mathbf{X} = \mathbf{U}\overline{\mathbf{Y}}\mathbf{A}' + \mathbf{E} = \mathbf{U}\overline{\mathbf{X}}\mathbf{A}\mathbf{A}' + \mathbf{E} \tag{3}$$

the well known projection pursuit clustering proposed by Bock (1987). Matrix $\overline{\mathbf{X}}$ is the centroid matrix of the objects in the original J-dimensional space. In this case the centroid matrix $\overline{\mathbf{Y}}$ is projected back into the J-dimensional space in order to reconstruct the original data matrix $\mathbf{X}$.

3 Application

The short-term scenario of September 1999 on macroeconomic performance of national economies of twenty countries, members of the Organization for Economic Co-operation and Development (OECD) has been considered in the paper by Vichi and Kiers (2001) to test the ability of the factorial k-means analysis, which allows a simultaneous classification of objects and a component reduction for variables, in identifying classes of similar economies and help to understand the relationships within the set of observed economic indicators. The performance of the economies reflects the interaction of six main economic indicators: Gross Domestic Product (GDP), Leading Indicator (LI), Unemployment Rate (UR), Interest Rate (IR), Trade Balance (TB), Net National Savings (NNS). Variables have been standardized.

The classification, obtained by the tandem analysis, i.e. k-means applied on the first two principal components scores, when the number of clusters for the objects is equal to three and the number of components for the variables is equal to two, is displayed in Figure 2.

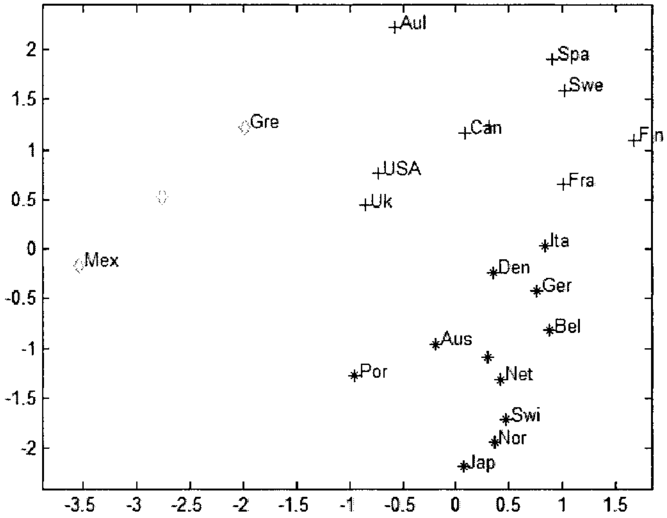


Fig. 2. K-means clustering computed on the first two PC.

The first PCA component is characterized mainly by net national savings, gross domestic product, whereas the second PCA component, by interest rate and trade balance. The unemployment rate characterizes both dimensions, as it can be observed from Table 1. The first component explains 28% of the total variance, while the second PCA dimension explains the 23%. The classification of the countries is given below:

First class: Australia, Canada, Finland, France, Spain, Sweden, United Kingdom, United States;
Second class: Greece, Mexico,
Third class: Austria, Belgium, Denmark, Germany, Italy, Japan, Portugal, Netherlands, Norway, Switzerland.

	GDP	IR	LI	UR	NNS	TB
Comp 2	-0.065	**-0.696**	**-0.229**	**0.367**	-0.092	**0.563**
Comp 1	**-0.567**	-0.175	-0.192	**-0.489**	**0.607**	0.059

Table 1. Component loadings defined by PCA

By applying Factorial k-means (Fk-means) (Vichi and Kiers (2001)) the classification of the countries is given below:

First class: Australia, Canada, Finland, France, Netherlands, Spain, Sweden, United States;
Second class: Greece, Mexico, Portugal;
Third class: Austria, Belgium, Denmark, Germany, Italy, Japan, Norway, Switzerland, United Kingdom.

	GDP	IR	LI	UR	NNS	TB
Dim 2	**0.956**	**0.309**	-0.003	**0.245**	-0.133	-0.184
Dim 1	0.175	**-0.711**	**0.297**	**0.260**	**-0.329**	**0.717**

Table 2. Correlation between dimensions of Factorial k-means and the macroeconomic variables

It can be observed that both for PCA (Table 1) and Factorial k-means (Table 2) have some variables correlated with both dimensions-components. This does not help the interpretation of the two dimensions-components.

Clustering and disjoint PCA (C&DPCA) has been applied on the same data set by fixing the number of clusters for the objects and variables equal to three and two respectively. The component loadings matrix is shown in Table 2, while the classification of the countries is given below:

First class: Australia, Canada, Denmark, Finland, France, Germany, Italy, Spain, Sweden, United Kingdom, United States;
Second class: Greece, Mexico, Portugal;
Third class: Austria, Belgium, Japan, Netherlands, Norway, Switzerland.

The first dimension of C&DPCA is still characterized mainly by net national savings, and less strongly by gross domestic product, whereas the second C&DPCA dimension by interest rate and trade balance. However, this time unemployment rate characterizes the first dimension only, as it can be observed from Table 3.

The first C&DPCA dimension explains 26% of the total variance, while the second C&DPCA dimension accounts for 21%. Thus, the loss of variance with respect to the PCA is irrelevant.

	GDP	IR	LI	UR	NNS	TB
Dim 2	0	-0.697	0.229	0	0	0.679
Dim 1	-0.383	0	0	-0.498	0.778	0

Table 3. Component loadings defined by Disjoint PCA

Comparing the two graphical representations in Figure 1 and 2, it can be observed that the C&DPCA, more clearly shows three homogeneous classes, mainly representing the same countries of the tandem analysis with some relevant differences. These are mainly due to the role on the unemployment rate in the two analyses and less strongly by the leading indicator. In C&DPCA UR characterizes the first dimension only, while it influences both dimensions in Figure 1. In Figure 2, Italy and Germany are positioned higher in the plot with respect to Figure 1 to better represent the higher unemployment rate they have.

In Figure 3 Mexico and Portugal also are located much closer because they have very similar values of GDP, LI and TB, which describe the first dimension of C&DPCA.

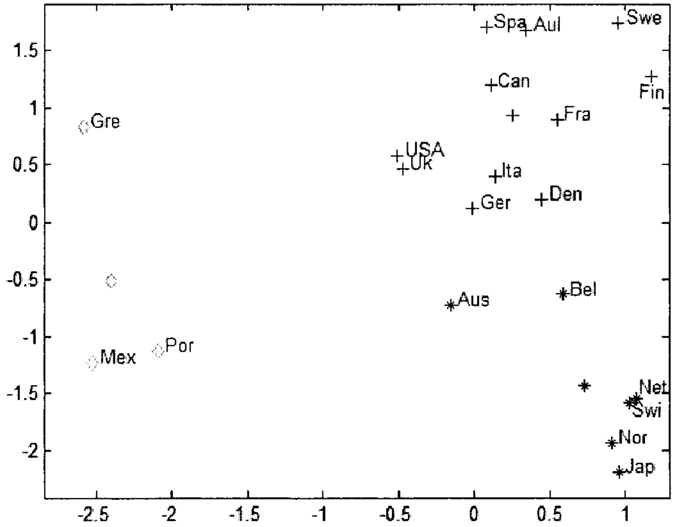


Fig. 3. Clustering and Disjoint PCA

References

ARABIE, P. and CARROLL, J.D. (1980): MAPCLUS: A Mathematical Programming Approach to Fitting the ADCLUS model. *Psychometrika, 45, 211–235.*

BAIER, D., GAUL, W., and SCHADER, M. (1997): Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring. In: R. Klar and O. Opitz (Eds.): *Classification and Knowledge Organization.* Springer, Berlin, 557–566.

BOCK, H.H. (1987): On the interface between cluster analysis, principal components, and multidimensional scaling. In: H. Bozdogan and A.J. Gupta (Eds): *Multivariate statistical modelling and data analysis*, Proceedings of Advances Symposium on Multivariate Modelling and Data Analysis, Knoxville, Tennessee, May 15-16, 1987, Reidel Publishing Co., Dordrecht, 17–34.

BOTH, M. and GAUL, W. (1987): Ein Vergleich zweimodaler Clusteranalyseverfahren. *Methods of Operations Research, 57, 593–605.*

DESARBO, W.S. (1982): GENNCLUS: New Models for General Nonhierarchical Clustering Analysis. *Psychometrika, 47, 446–449.*

GAUL, W. and SCHADER, M. (1996): A New Algorithm for Two-Mode Clustering. In: H.-H. Bock and W. Polasek (Eds.): *Data Analysis and Information Systems.* Springer, Berlin, 15–23.

ROCCI, R. and VICHI, M. (2004): Multimode partitioning, submitted.

VAN MECHELEN, I., BOCK, H.H. and DE BOECK, P. (2004): Two-mode clustering a structured overview. Statistical Methods in Medical Research, to appear.

VICHI, M. (2000): Double k-means Clustering for simultaneous classification of Objects and Variables. In: S. Borra, R. Rocchi, M. Vichi, and M. Schader (Eds): Advances in Classification and Data Analysis, 43–52, Springer, Berlin.

VICHI, M. and KIERS, H.A.L, (2001): Factorial k-means analysis for two way data. *Computational Statistics and Data Analysis, 37, 49–64.*

VICHI, M. and MARTELLA, F. (2005): Model-based clustering for block-data, submitted.

VICHI, M. and SAPORTA G. (2004): Clustering and Disjoint Principal Component Analysis, submitted.

The Number of Clusters in Market Segmentation

Ralf Wagner, Sören W. Scholz, and Reinhold Decker

Department of Business Administration and Economics,
Bielefeld University, P.O. Box 100131, D-33615 Bielefeld, Germany

Abstract. Learning the 'true' number of clusters in a given data set is a fundamental and largely unsolved problem in data analysis, which seriously affects the identification of customer segments in marketing research.

In this paper, we discuss the properties of relevant criteria commonly used to estimate the number of clusters. Moreover, we outline two adaptive clustering algorithms, a growing k-means algorithm and a growing self-organizing neural network. In the empirical part of the paper, we find that the first algorithm stops growing with exactly the number of clusters that we get when determining the optimal number of clusters by means of the $JUMP$-criterion. This cluster solution proves to be rather similar to the one we obtain by applying the neural network approach. To evaluate the clusters, we use association rules. By testing these rules, we show the differences of patterns underlying particular market segments.

1 Introduction

Segmentation is one of the main tasks in marketing research, and a crucial part of product positioning and advertising strategy development in both consumer markets, which we focus on in this paper, and industrial markets. Dibb and Stern (1995) characterize this task as a fundamental principle of modern marketing, since a single offer does not usually satisfy all the different needs of the customers within a market. Individual preferences as well as the willingness to pay are, as a rule, too heterogeneous for this. The process of market segmentation includes the understanding of how and why customers buy, how companies can, or rather should, adapt their products to customer needs, and how the customers' individual benefits can be enhanced (Lu (2003)). Segmentation, then, divides a given market into "relevant, manageable and targetable segments in order to allow better-tailored offerings to be developed" (Brassington and Pettitt (2005)). Starting from a heterogenous total market, two key functions have to be managed:

1. the identification of largely homogeneous market segments which differ from each other due to their individual profile and
2. the development of segment-specific marketing strategies.

This paper focuses on the first function. Special emphasis is put on the distinctiveness of the individual segments. This distinctiveness depends on the type of product or service considered and the circumstances prevailing in the market at the time of measurement.

Common segmentation techniques are multidimensional scaling, factor analysis and, above all, cluster analysis. A recent overview of segmentation techniques applied in marketing research is given in Wedel and Kamakura (2000). Due to the heterogeneity of most consumer markets, the determination of an appropriate number of segments or clusters is a key question in almost any market segmentation study (Boone and Roehm (2002)). Heterogeneity is expressed in a set of segmentation variables such as demographics (e.g. age and sex), psychographics (e.g. attitudes and opinions), or behavior based variables (e.g. readiness to buy, brand preferences, and benefits sought). Although the importance of correctly answering the question of how many segments exist in a certain market is emphasized in the relevant literature, there are comparatively few papers dealing with this problem in depth. The same applies to the question of overlapping segments. Sometimes the restriction to disjunctive segments is neither possible nor suitable. An example for benefit segmentation is given in Baier et al. (1997).

New developments in data production have endowed marketers with extensive information about customers and their buying habits as well as consumption behavior. The spectrum ranges from electronic micro test markets like IRI BehaviorScan and personalized customer cards to online consumer panels and web log data, which is of eminent interest for recommender systems in e-commerce (Gaul and Schmidt-Thieme (2002)). Nevertheless, the number of clusters considered in segmentation remains to be a crucial question.

The remainder of the paper is organized as follows. In section 2, we briefly discuss criteria commonly used to assess the number of clusters in the context of market segmentation and outline the computation of the $JUMP$-criterion in detail. We introduce a modified growing k-mean cluster algorithm, which controls for the number of clusters by testing the empirical distribution of the observations assigned to each cluster. Moreover, we sketch a new growing neural self-organizing network (Decker (2005)), which is closely related from a methodical point of view. For the evaluation of the clustering results, we use association rules and discuss different assessments of these rules with respect to the present task. In section 3, we demonstrate the usability of these techniques by analyzing survey data collected from a German household panel. We pay particular attention to the evaluation of the cluster solution in terms of its suitability for deducing marketing implications. The paper concludes with a brief discussion and an outlook on future research.

2 Methodology

2.1 Determination of the optimal number of clusters

In applied market segmentation, the structure of the markets under consideration is commonly assumed to be unknown. Thus, the resulting clusters have to be evaluated with respect to their reproducibility by different clustering procedures as well as to their relevance for the present marketing task (Granzin et al. (1998), Fennell et al. (2003)).

With the exception of very few applications in industrial and service marketing (e.g., Dibb and Simkin (1994), Palmer and Millier (2004)), the marketers can not refer to externally or intuitively predefined classes such as the species in Fisher's iris data set. Rather, one has to learn both the number of clusters as well as their characteristics directly from the data. For the first purpose, different criteria can be used.

Let $\boldsymbol{X} = (x)_{n=1,\dots,N;l=1,\dots,L}$ be a data matrix consisting of N observations to be grouped into k clusters by means of L features. The resulting cluster means $\boldsymbol{\theta}_h = (\theta_{h1}, \dots, \theta_{hl}, \dots, \theta_{hL}) \in \mathbb{R}^L; (h = 1, \dots, k)$ make up the $k \times L$ matrix $\boldsymbol{\Theta}_k$. The $N \times k$ assignment matrix $\boldsymbol{A}_k$ with $a_{nh} = 1$, if observation n is assigned to cluster h, and $a_{nh} = 0$, otherwise, enables the computation of the between cluster sum of squares and cross products matrix

$$\boldsymbol{B}_k = \boldsymbol{\Theta}'_k \boldsymbol{A}'_k \boldsymbol{A}_k \boldsymbol{\Theta}_k \tag{1}$$

as well as the within cluster sum of squares and cross products matrix

$$\boldsymbol{W}_k = \boldsymbol{X}'\boldsymbol{X} - \boldsymbol{B}_k \tag{2}$$

(Mardia et al. 1979). Table 1 provides an overview of common criteria for learning the optimal k^* using the data at hand.

In the table $sd(.)$ denotes the standard derivation and R equals the number of reference data sets $(r = 1, \dots, R)$ being simulated to estimate the number of clusters via the GAP statistic. These reference data are either drawn from a uniform distribution in the simple case or reflect the given data structure by the inherent principal components. Of course, this increases the computational burden as compared with other criteria depicted in the table. The criteria suggested by Calinski and Harabasz (1974) and Krzanowski and Lai (1988) are not suited to answering the very basic question of whether there is a group structure in the data at all, since these criteria are not applicable to the case $k = 1$. This becomes obvious from the specifications given in the second column of the table. The denominator of $CH(1)$ would be equal to zero and $KL(1)$ would require the computation of $\operatorname{tr} \boldsymbol{W}_0$ in the nominator. The criterium proposed by Hartigan (1985) is applicable to this fundamental case, but it found to have inferior performance in simulation studies (e.g., Milligan and Cooper (1985), Dudoit and Fridlyand (2002)). Recently, Sugar and James (2003) introduced the $JUMP$-criterion, which has proven to be

Reference	Criterion	k^*
Calinski and Harabasz (1974)	$CH(k) = \frac{\operatorname{tr} \boldsymbol{B}_k(N-k)}{\operatorname{tr} \boldsymbol{W}_k(k-1)}$	$\max_{k \geq 2} CH(k)$
Hartigan (1985)	$HA(k) = \left(\frac{\operatorname{tr} \boldsymbol{W}_k}{\operatorname{tr} \boldsymbol{W}_{k+1}} - 1\right)(N-k-1)$	$\min_{k \geq 1} HA(k)$
Krzanowski and Lai (1988)	$KL(k) = \frac{(k-1)^{\frac{2}{L}} \operatorname{tr} \boldsymbol{W}_{(k-1)} - k^{\frac{2}{L}} \operatorname{tr} \boldsymbol{W}_k}{k^{\frac{2}{L}} \operatorname{tr} \boldsymbol{W}_k - (k+1)^{\frac{2}{L}} \operatorname{tr} \boldsymbol{W}_{k+1}}$	$\max_{k \geq 2} KL(k)$
Tibshirani et al. (2001)	$GAP(k) =$ $\frac{1}{R}\sum_{r=1}^{R} \log(\operatorname{tr} \boldsymbol{W}_{rk}) - \log(\operatorname{tr} \boldsymbol{W}_k)$ and $s(k) = sd(\log(\operatorname{tr} \boldsymbol{W}_{rk}))\sqrt{1 + \frac{1}{R}}$	$\min_{k \geq 1} GAP(k) \geq$ $GAP(k+1) - s(k+1)$

Table 1. Criteria commonly used to determine the optimal number of clusters k^*

suitable for learning the number of clusters with k-means clustering, one of the most widespread clustering methods in marketing research (Green and Krieger (1995)).

Let $\boldsymbol{x}_n = (x_{n1}, \ldots, x_{nl}, \ldots, x_{nL}) \in \mathbb{R}^L$, with $n \in \{1, \ldots, N\}$, denote the input data to be analyzed, i.e. the individual profiles or feature vectors of consumer $1, 2, \ldots, n, \ldots, N$. Assuming the features used for classification to be uncorrelated (and, therefore, not to provide redundant information) the distortion $d(k)$ can be estimated by the sum of squared errors based on Euclidean distances. Since true distortion is unknown it must be estimated using the data at hand. The estimated distortion $\hat{d}(k)$ decreases with an increasing number of clusters and therefore need to be corrected by a power transformation to assess the optimal number of clusters. Figure 1 outlines the procedure used to calculate the $JUMP$-criterion by using the estimated distortion.

```
1: Group the data by assuming different numbers of clusters $k = 1, \ldots, K$.
2: Calculate $\hat{d}(k)$ for $k = 1, \ldots, K$ and define $\hat{d}(0) = 0$.
3: Compute the jumps $JUMP(k) = \hat{d}^{-y}(k) - \hat{d}^{-y}(k-1)$ with transformation power $y > 0$.
4: Determine the optimal number of clusters $k^* = \max_{k \in \{1, \ldots, K\}} JUMP(k)$.
```

Fig. 1. Calculation of the $JUMP$-criterion

The procedure also includes the case $k^* = 1$ and, therefore, explicitly checks for the absence of a group structure in the data. The transformation power y used in step three is not fixed to a certain value. Thus, we define $y = \frac{L}{2}$ according to Sugar and James (2003). The *JUMP*-criterion will be used in the empirical part of the paper for externally validating the cluster number resulting from our modification of Hamerly and Elkan (2003)'s G-means algorithm.

2.2 Modified growing *k*-means (GKM)

The basic principle of our growing k-means approach is to start with a small number of clusters and to expand this number until the consumer profiles assigned to each cluster fit a multivariate normal distribution that is centered at the cluster-centroids $\boldsymbol{\theta}_h$. The whole procedure is outlined in Figure 2.

```
1: Let $k$ be the initial number of clusters (usually $k = 1$).
2: Apply the standard $k$-means algorithm to the given data set.
3: Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{\tilde{n}}, \ldots, \boldsymbol{x}_{N_h}\}$ be the set of profiles assigned to centroid $\boldsymbol{\theta}_h$.
4: Test whether the profiles assigned to each cluster $h = 1, \ldots, k$ follow a
   multivariate normal distribution.
5: If the data meets the relevant criteria of multinormality, then stop with
   the respective optimal $k^*$, otherwise increase the number of clusters $k$
   by 1 and return to step 2.
```

Fig. 2. Modified growing k-means algorithm

To determine the optimal number of clusters, we consider the distribution of the observations (consumer profiles) assigned to each cluster. The basic idea is to assume that the output of a clustering procedure is acceptable if the centroids are providing substantial information about, at least, the majority of the cluster members (for a comprehensive discussion of probabilistic models considering homogeneity and heterogeneity in cluster analysis see Bock (1985, 1996)). Thus, it is rather intuitive to presume a cluster centroid as being not only the balance point but also the median as well as the modus of the multivariate distribution representing the cluster members.

To check whether the profiles of the members assigned to a cluster h follow a multivariate normal distribution we refer to Mardia (1974)'s measure of the standardized 4th central moment (kurtosis). The corresponding hypotheses are:

H_0 : The consumer profiles assigned to centroid $\boldsymbol{\theta}_h$ are multivariate normal-distributed.

H_1 : The consumer profiles of cluster h do not follow a multivariate normal distribution.

The measure is based on squared Mahalanobis distances, which are very helpful in detecting multivariate outliers (Mecklin and Mundfrom (2004)). According to Mardia (1974) the statistic for testing the above hypotheses is given by

$$\tilde{d}_h = \frac{\hat{d}_h - L(L+2)}{(8L(L+2)/N_h)^{1/2}} \qquad (h = 1, \dots, k) \tag{3}$$

and

$$\hat{d}_h = \frac{1}{N_h} \sum_{\tilde{n}=1}^{N_h} \left((\boldsymbol{x}_{\tilde{n}} - \boldsymbol{\theta}_h) \boldsymbol{S}_h^{-1} (\boldsymbol{x}_{\tilde{n}} - \boldsymbol{\theta}_h)' \right)^2, \qquad (h = 1, \dots, k), \tag{4}$$

where $\boldsymbol{S}_h$ denotes the cluster-specific sample covariance matrix and N_h equals the number of observations assigned to cluster h. The null hypothesis cannot be rejected for $\tilde{d}_h$ values close to 0. Therefore, we favor that k^* which meets the following condition:

$$k^* = \min_{k \in \{1, \dots, K\}} \left[\max_{h \in \{1, \dots, k\}} \tilde{d}_h \right] \tag{5}$$

The algorithm stops growing when the distribution of random vectors assigned to each cluster fits to a multivariate normal distribution. Therefore, it tends towards small numbers of clusters.

2.3 Self-controlled growing neural network (SGNN)

Let $\mathcal{B}$ be a set of units u_h with $h \in \{1, \dots, k = |\mathcal{B}|\}$, and $\mathcal{D}$ the set of connections between these units, together capturing the topological structure of an artificial neural network. Each unit is represented by a weight vector $\tilde{\boldsymbol{\theta}}_h = (\tilde{\theta}_{h1}, \dots, \tilde{\theta}_{hl}, \dots, \tilde{\theta}_{hL}) \in \mathbb{R}^L$. In fact, these weight vectors directly correspond to the centroids considered above. The general idea of the algorithm used in this paper can be sketched as follows (for a more detailed description see Decker (2005)).

First, an initial network with two non-connected units u_1, u_2 (i.e. $\mathcal{B} = \{u_1, u_2\}$ and $\mathcal{D} = \emptyset$) and corresponding weight vectors $\tilde{\boldsymbol{\theta}}_1$, $\tilde{\boldsymbol{\theta}}_2$ is created. Furthermore, two learning rates ϵ_{Best} and ϵ_{Second} to control the network adaptation process are specified. Then, an adaptation step proceeds as follows: An input vector $\boldsymbol{x}_n$ is selected at random from the database and the distances between $\boldsymbol{x}_n$ and the currently available units of the neural network (or rather weight vectors) are calculated. Then the best and the second best matching unit $u_{h_{Best}}$ and $u_{h_{Second}}$ are selected. Both will be connected, if this is not yet the case. Next, the activity $v_{h_{Best}}$ of the best matching unit, i.e. the extent to which $\tilde{\boldsymbol{\theta}}_{h_{Best}}$ fits $\boldsymbol{x}_n$, is computed. Furthermore, thresholds v_{Thres} and w_{Thres} for both the activity ($v_{h_{Best}}$) and the training requirement ($w_{h_{Best}}$) of $u_{h_{Best}}$ are computed. If either $v_{h_{Best}}$ or $w_{h_{Best}}$ exceeds the corresponding thresholds v_{Thres} and w_{Thres} respectively, then $u_{h_{Best}}$ and all those units u_{h_i} to which it is connected are adapted to $\boldsymbol{x}_n$ according to ϵ_{Best} and ϵ_{Second} as well as the current training requirements. Otherwise, a new unit is added to the neural network in-between $\tilde{\boldsymbol{\theta}}_{h_{Best}}$ and $\boldsymbol{x}_n$. Subsequently, the age of all edges emanating from $u_{h_{Best}}$ is increased by one, the internal control parameters are updated, and all the connections of the neural network that are 'older' than a user-defined maximum age of edges (a_{Max}) are removed. Furthermore, all units with no connection to any other unit and low relevance are removed as well. Finally, the adaptation step counter is updated, a stopping criterion is checked, and, if necessary, the whole adaptation step is repeated. As a result of this adaptation process, we get the connectivity matrix $\boldsymbol{D}$, which describes the final topological structure of the neural network. Each unit, or rather the corresponding weight vector $\tilde{\boldsymbol{\theta}}_h$, represents a frequent pattern in the available survey data and is referred to as a prototype.

2.4 Cluster evaluation

When the 'true' number of clusters is unknown in empirical data analysis, the estimated number of clusters can be correct, too high, or too low. Thus, a reliable cluster evaluation is needed. On the one hand, statistical techniques, such as cluster stability indices are proposed (see Roth et al. (2002) for a recent discussion). These criteria are comparable to those depicted in Table 1 and feature similar methodological characteristics. On the other hand, Hair et al. (1998, p. 500) advise that "the researcher should take great care in validating and ensuring the practical significance of the final cluster solution." Unfortunately, there is no single method for this purpose. In order to ensure both the methodical straightness and the practical interpretability of the clustering solution with respect to the underlying segmentation task, we apply

- canonical discriminant analysis,
- parallel coordinates visualization,
- correlation analysis, and
- association rules

to the given problem. Extracting association rules assists the interpretation of patterns revealed by the cluster analysis and, therefore, proves to be useful when high-dimensional feature spaces are considered, e.g., in market segmentation. In this context, a rule $\mathcal{A} \Rightarrow \mathcal{C}$ describes the relation between two sets of features (e.g. items or statements expressing individual consumption or nutrition preferences). In the following, a feature is assumed to be relevant to a respondent if its measured value is equal to or higher than the third quartile of all feature values of this respondent. Thus, each respondent (indexed by $\tilde{n}$) assigned to cluster h is characterized by a set of relevant features $\mathcal{I}_{h\tilde{n}}$, which we call the item set. All respondents assigned to cluster h, or rather the respective item sets, are represented by set $\tilde{\mathcal{N}}_h$. This allows us to characterize each cluster by different rules. However, to evaluate an existing partition, the relevance of each rule for all clusters must be assessed.

For $0 < sup(\mathcal{A} \cup \mathcal{C}) \leq sup(\mathcal{A}) \wedge sup(\mathcal{C})$ the (cluster-wise) support of a rule can be defined as follows

$$sup_h(\mathcal{A} \Rightarrow \mathcal{C}) = \frac{|\mathcal{I}_{h\tilde{n}} \in \tilde{\mathcal{N}}_h|(\mathcal{A} \cup \mathcal{C}) \subseteq \mathcal{I}_{h\tilde{n}}|}{|\tilde{\mathcal{N}}_h|} \qquad (h = 1, \ldots, k). \tag{6}$$

Accordingly, the confidence and the lift of a rule is given by

$$conf_h(\mathcal{A} \Rightarrow \mathcal{C}) = \frac{|\mathcal{I}_{h\tilde{n}} \in \tilde{\mathcal{N}}_h|(\mathcal{A} \cup \mathcal{C}) \subseteq \mathcal{I}_{h\tilde{n}}|}{|\mathcal{I}_{h\tilde{n}} \in \tilde{\mathcal{N}}_h|\mathcal{A} \in \tilde{\mathcal{N}}_h|} \qquad (h = 1, \ldots, k) \tag{7}$$

and

$$lift_h(\mathcal{A} \Rightarrow \mathcal{C}) = \frac{sup_h(\mathcal{A} \Rightarrow \mathcal{C})}{sup_h(\mathcal{A}) \cdot sup_h(\mathcal{C})} \qquad (h = 1, \ldots, k), \tag{8}$$

where

$$sup_h(\mathcal{A}) = \frac{|\mathcal{I}_{h\tilde{n}} \in \tilde{\mathcal{N}}_h|\mathcal{A} \subseteq \mathcal{I}_{h\tilde{n}}|}{|\tilde{\mathcal{N}}_h|}.$$

Since each rule reveals a particular pattern in the available data and since the above algorithms aim at separating the observations with respect to these patterns, there may be some clusters that do not include any observation satisfying the antecedent of the relevant rules. Therefore, the measures defined in the equations (7) and (8) are applicable only to rules with positive support for the antecedent and the implication in a cluster h.

Although the lift is the most common measure of interestingness, it has been the subject of criticism (see Hilderman and Hamilton (2001) for an overview). One shortcoming of this measure is its symmetry, i.e. $lift_h(\mathcal{A} \Rightarrow \mathcal{C}) = lift_h(\mathcal{C} \Rightarrow \mathcal{A})$. Alternatively Brin et al. (1997) introduced the conviction

$$conv_h(\mathcal{A} \Rightarrow \mathcal{C}) = \frac{sup_h(\mathcal{A}) \cdot sup_h(\neg\mathcal{C})}{sup_h(\mathcal{A} \Rightarrow \neg\mathcal{C})} \qquad (h = 1, \ldots, k), \tag{9}$$

whereas Wagner (2005) proposed the following new measure of interestingness:

$$int_h(\mathcal{A} \Rightarrow \mathcal{C}) = \frac{sup_h(\mathcal{A}) + sup_h(\mathcal{C}) - 2 \cdot sup_h(\mathcal{A} \cup \mathcal{C})}{sup_h(\mathcal{A} \cup \mathcal{C})} + 1 \tag{10}$$
$$(h = 1, \ldots, k; (\mathcal{A}, \mathcal{C})|sup_h(\mathcal{A} \cup \mathcal{C}) \neq 0)$$

In the case of maximum interestingness of a rule $int_h(.)$ equals 1. To simplify the interpretation of this measure the fraction term is increased by 1. If $int_h(.) = z$ then every z^{th} occurrence of $\mathcal{A}$ is a co-occurrence with $\mathcal{C}$.

3 Empirical application

3.1 The data

The data set to be analyzed has been provided by the German ZUMA institute, and it is a sub-sample of the 1995 GfK ConsumerScan Household Panel (for a detailed description see Papastefanou et al. 1999). It comprises socio-economic and demographic characteristics of several thousands of households as well as attitudes towards nutrition (e.g., slimness orientation, plain fare, and branded goods), aspects of daily life (e.g., traditional living, convenience-oriented cooking, and mistrust towards new products), environment (e.g., ecological awareness, mobility, and industry), and shopping (e.g., tendency to purchase new products, price consciousness, and preference for small retail stores). A considerable number of these items are related to individual nutrition behavior. Typical items are "Multivitamin juices are an important supplement to daily nutrition." or "I eat vegetarian food only." We exemplarily consider the attitudes or opinions of 4,266 households measured by means of 81 mostly Likert-scaled items (with 1 = 'I definitely disagree.', ..., 5 = 'I definitely agree.'). The scale is assumed to be equidistant and, therefore, the data are treated as metric.

3.2 Market segmentation by means of growing clustering algorithms

According to the *JUMP*-criterion the optimal number of clusters for the data at hand is 37. Applying the GKM-algorithm also results in a 37-cluster solution, where sets of 174 to 394 individual consumer (profiles) are assigned to 15 meaningful clusters with centroids $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{15}$. Moreover, 22 consumer (profiles) are assigned to 22 single-object clusters. These outliers will be ignored in the following discussion. Figure 3 shows a projection of the available cluster solution in a two-dimensional discriminant space.

As obvious from the figure, the clusters are overlapping. This is not surprising since the clustering is done in an 81-dimensional feature space. In Figure 3, those four clusters with the highest and the second highest distances of centroids are highlighted. The highest distance is found between GKM-cluster 10 ($\boldsymbol{\theta}_{10}$: ∘) and GKM-cluster 14 ($\boldsymbol{\theta}_{14}$: •). Both clusters are separated with respect to the vertical axis (1^{st} canonical factor). Furthermore, we will focus on the differences between cluster 3 ($\boldsymbol{\theta}_3$: □) and cluster 9 ($\boldsymbol{\theta}_9$: △) to evaluate the corresponding market segmentation in view of its validity and practical interpretability.

Applying the SGNN algorithm with $\epsilon_{Best} = 0.975$, $\epsilon_{Second} = 0.025$, and $a_{Max} = 81 (= L)$ results in a neural network that comprises $H = 14$ prototypes. Fixing ϵ_{Best} close to 1 causes a strong compression of the data and leads to an easy-to-grasp number of consumption and nutrition style prototypes. The numbers of consumers that are represented by the weight vectors $\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_{14}$ range between 199 and 400. In fact, SGNN leads to comparable results, thus we can use it as an object of comparison.

Table 2 gives a brief description of those survey items that cause the above-mentioned highest distances between GKM-cluster centroids $\boldsymbol{\theta}_3$ and $\boldsymbol{\theta}_9$ as well as between $\boldsymbol{\theta}_{10}$ and $\boldsymbol{\theta}_{14}$). The table also displays the weights resulting from applying the SGNN algorithm. Obviously, the cluster profiles

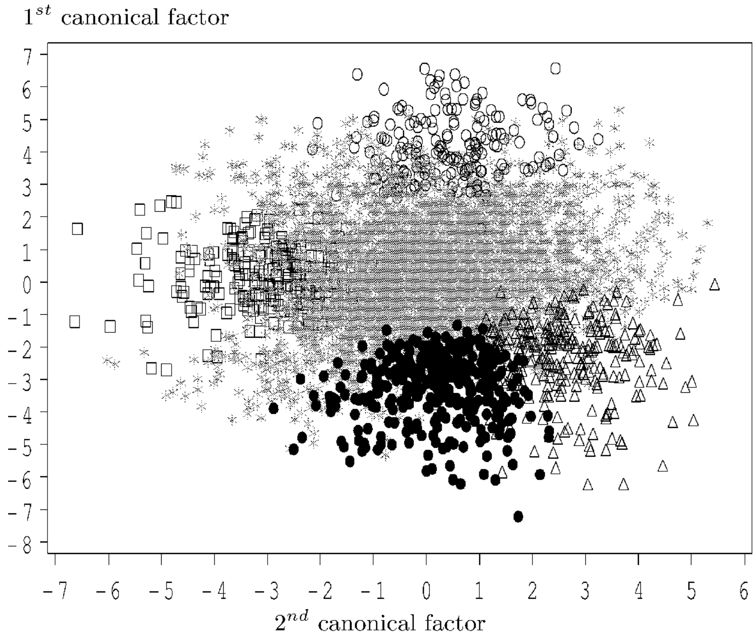


Fig. 3. Projection of the GKM solution in a two-dimensional discriminant space

represented by the respective weights match the four clusters obtained from the GKM algorithm to a great extent. As indicated by the Pearson correlations quoted in the bottom line of the table, both approaches yield very similar results with respect to the selected set of items. To emphasize the content validity of the clustering solution computed with GKM, we utilize parallel coordinates. To facilitate visual inspections, we restrict our illustration to the comparison of two cluster profiles.

In Figure 4 the profiles of GKM-cluster 10 and 14 ($\boldsymbol{\theta}_{10}$ and $\boldsymbol{\theta}_{14}$) are displayed. Both clusters can be discriminated in terms of the importance given to health-oriented nutrition. When looking at the largest distances of singular items from the survey, we find that the members of GKM-cluster 10, on average, are highly involved in food consumption. The quality of food products is of accentuated importance in comparison to GKM-cluster 14 (see, e.g., item 19 and 27). This, as a consequence, implies a strong preference of exclusive and fancy foods (item 33) as well as a strong preference for branded products and a distinct mistrust of no-name products, respectively. These consumers are characterized by health-oriented nutrition styles such as fat free foods (items 29 and 43), whole foods, or even vegan diets. The motivation for this behavior seems to stem from hedonistic rather than well-founded health concerns (item 3). Their primary focus is on maintaining physical attractiveness

l	Brief item description	$\boldsymbol{\theta}_3$	$\tilde{\boldsymbol{\theta}}_7$	$\boldsymbol{\theta}_9$	$\tilde{\boldsymbol{\theta}}_4$	$\boldsymbol{\theta}_{10}$	$\tilde{\boldsymbol{\theta}}_5$	$\boldsymbol{\theta}_{14}$	$\tilde{\boldsymbol{\theta}}_2$
2	Likes the atmosphere in small shops	3.82	3.65	3.17	3.27	4.26	4.18	3.29	3.29
3	Likes to have company	4.28	4.31	3.65	3.64	4.65	4.54	3.90	3.96
6	Spends spare time very actively	3.97	4.10	3.18	3.17	4.29	4.17	3.68	3.69
15	Enjoys life to the fullest	3.15	3.62	2.68	3.26	3.72	4.26	3.59	3.22
19	Regards food quality as important	3.57	3.61	2.78	4.31	4.45	4.41	2.54	3.85
21	Prefers traditional lifestyle	2.60	3.51	3.62	3.93	4.46	3.96	3.17	3.70
27	Quality-oriented food purchase	3.13	2.03	2.64	3.55	4.37	3.45	2.40	2.26
29	Watches his/her figure when eating and drinking	3.52	3.72	2.67	3.12	2.82	4.43	2.81	2.69
33	Prefers fancy drinks and food	3.59	3.15	1.70	1.83	4.18	3.75	1.90	2.22
34	Prefers hearty plain fare	2.89	2.91	4.40	4.14	4.55	4.19	3.23	3.50
35	Prefers whole foods nutrition	2.31	2.28	1.63	1.89	3.65	3.37	1.53	1.52
39	Uses multivitamin juices as a supplement to daily nutrition	2.82	3.28	3.08	3.24	4.36	4.03	3.06	2.94
42	Pays attention to anti-allergic food	2.91	2.85	2.12	2.62	4.45	4.11	1.96	1.94
43	Perceives himself/herself as slimness-oriented	2.53	2.82	1.68	2.12	3.68	3.49	1.93	1.76
44	Purchases foods without additives	3.47	3.18	2.21	2.66	4.49	4.17	2.09	2.06
48	Eats vegetarian food only	1.32	1.37	1.10	1.19	1.82	1.65	1.22	1.10
49	Indulges oneself with good meals	4.37	4.11	2.90	2.90	4.36	4.10	3.36	3.69
59	Counts calories when eating	1.86	2.18	1.41	1.72	3.50	3.21	1.40	1.38
60	Prefers to apply tried and tested recipes	2.13	2.52	4.31	4.14	4.50	4.06	3.08	3.06
	Pearson's correlation coefficient r	0.86*		0.90*		0.91*		0.92*	

* highly significant ($p = 0.01$)

Table 2. Comparability of consumption and nutrition style prototypes

and fitness (items 43, 59, and 69), which is a premise for their active and sociable lifestyle. They watch their figure, count calories, often use vitamin supplements and purchase fresh instead of canned foods. This distinct health orientation is only rudimentarily reflected in the individual cooking and nutrition preferences. The respective consumers favor home cooking, especially plain fare, but they are also using convenience foods when cooking. The consumers are demanding in terms of the services and atmosphere of stores where they do their daily shopping (item 2). As a result of these characterizations,

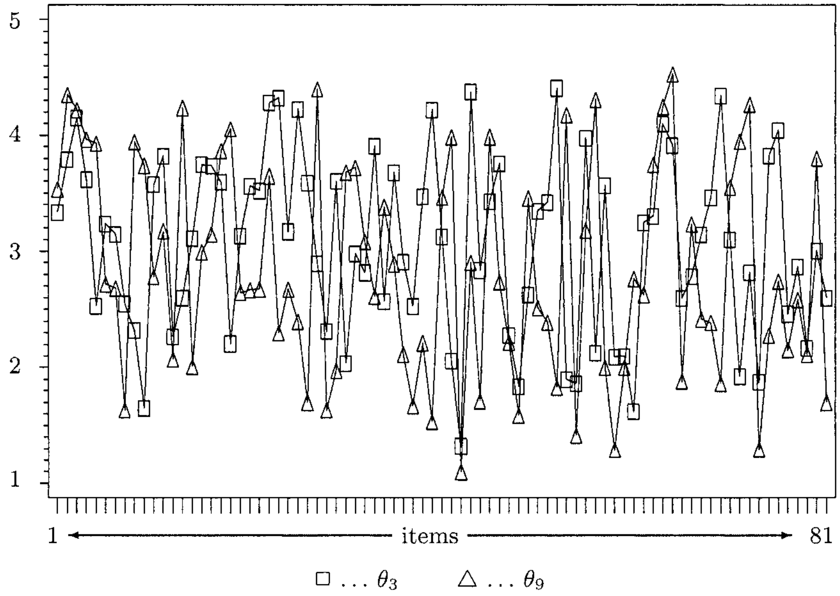


Fig. 4. Parallel coordinates for cluster profiles $\boldsymbol{\theta}_{10}$ and $\boldsymbol{\theta}_{14}$

we can refer to the respondents represented by cluster centroid $\boldsymbol{\theta}_{10}$ as the *hedonistic consumers.*

In contrast to this, the consumers making up cluster 14 (represented by $\boldsymbol{\theta}_{14}$) pay less attention to nutrition. Their diet consists almost completely of convenience food. Accordingly, the consumers assigned to this cluster are less concerned about food products with additives or preservatives (item 44). Health-oriented food consumption behavior is less important in their lives. Likewise, they strongly prefer low prices. The origin and quality of the purchased foods is secondary (item 27, 38, and 51). Consequently, we can label this cluster as the *uninvolved consumers.*

Figure 5 illustrates the profiles corresponding with cluster centroids $\boldsymbol{\theta}_3$ and $\boldsymbol{\theta}_9$. Although the maximum distances between these profiles are smaller than those depicted in Figure 4, these clusters are still distinguishable due to their nutrition and consumption behavior. Cluster 3 (represented by $\boldsymbol{\theta}_3$) comprises consumers who have a strong preference for natural, unprocessed foods (items 35, 42, and 44). In contrast to the cluster represented by $\boldsymbol{\theta}_9$, they favor shopping in small stores (item 2). This behavior is accompanied by a distinctive fondness for exclusive and, partly, even fancy foods that seem to be hardly an option for members of cluster 9 (item 33). The members of cluster 3 are less motivated by hedonistic reasons such as maintaining physical attractiveness (items 43, 59, and 69), but primarily aim to avoid the

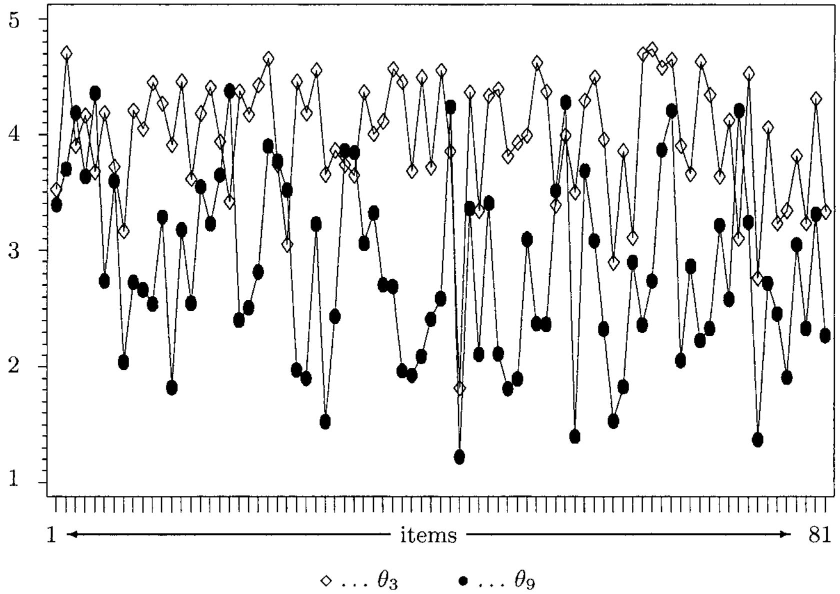


Fig. 5. Parallel coordinates for cluster profiles $\boldsymbol{\theta}_3$ and $\boldsymbol{\theta}_9$

consumption of any kind of non-natural food ingredients. Ostensible ways of promoting good health (as favored by the aforementioned hedonistic consumers to a certain degree), like using vitamin supplements, are not very popular in this cluster. Therefore, in opposition to cluster 9 (cf. $\boldsymbol{\theta}_9$), radical forms of nutrition behavior such as vegan or whole food diets are prevalent (items 35 and 48). Nutrition and cooking play an important role in their life, so they take nutrition tips very seriously. However, there is no susceptibility to social-oriented nutrition trends (e.g., item 43). Concluding, we refer to these consumers as *health-oriented consumers.*

Cluster centroid $\boldsymbol{\theta}_9$ represents consumers who are critical of social or technical progress in their personal environment. This cluster stands out due to its traditional lifestyle (item 21). These consumers are hardly affected by 'modern' nutrition styles such as vegan or whole food diets (item 35). Cooking and nutrition do not play a central role in their lives. Consequently, they prefer dishes that are fast and easy to serve. In doing so, they show a distinctive preference for domestic foods, and they like to stick with home made cooking, especially hearty plain fare (item 34) on the basis of tried and tested recipes (item 60). Being rather uninvolved, they are price conscious and shy away from purchasing new products. Due to their distrust of advertising, these *traditional consumers* are also hardly interested in well-established brands.

As already indicated, parallel coordinates facilitate an easy interpretation of the obtained cluster solutions by means of the resulting visual profiles. However, while the aggregated cluster profiles reveal meaningful structures in the present consumer segmentation, it is impossible to assess whether all members of a cluster contribute to the associated centroid in a similar way. Therefore, conclusions regarding the within-cluster homogeneity and those about patterns being characteristic or even distinctive for the cluster's composition can not be deduced in this manner.

3.3 Exploiting patterns in consumer segments

One major challenge in market segmentation is to exploit the patterns underlying the individual segments. Utilizing association rules for this purpose leads from the basic classification problem to a combination problem. The objective is to find combinations of features that are common to the observations (consumer profiles) assigned to a cluster, but not necessarily unique to these observations (Brāzma et al. (1998)). Clearly, we are restricted to learning only from positive examples and, moreover, when extracting rules from the observations assigned to one cluster, we have no priors concerning the validity of this rule for the observations assigned to the other clusters. In Figure 6, two rules are exemplarily illustrated.

Rule 1	Rule 2
Likes the atmosphere in small shops & Watches his/her figure when eating and drinking & Considers health in nutrition ⇓ Likes to try new products & Not willing to do without service when shopping & Quality-oriented food purchase & Pays attention to low fat diet	Doesn't need new products & Favors sociability & Prefers quickly prepared meals & Prefers hearty plain fare ⇓ Prices are more important than brands & Indulges oneself with good meals

Fig. 6. Selected association rules

Rule 1 characterizes the highly involved consumers who pay attention to their physical attractiveness as well as their health. Enjoying the atmosphere of small shops hints at sophisticated shopping habits and the corresponding willingness to pay for higher quality goods and services. Marketing measures targeting these consumers should emphasize the quality of food and, if possible, the contribution to a low fat diet, both in combination with high-class individual retail services. Such consumers are more likely to show a predisposition to check out new products and, therefore, might become early adopters in the product life cycle. Moreover, they might catalyze the introduction of new products and services into their reference groups and might provide the vendors with positive word-of-mouth communication (cf. Blackwell et al. (2001) for a detailed discussion of group and personal influences related to marketing communication and buying behavior).

Rule 2 in Figure 6 circumscribes a kind of traditional–one could even say 'Bavarian'–type of consumer. These consumers prefer hefty and hearty meals of the traditional German cuisine and pay attention to the time they need for preparing the meals. They decline new nutrition concepts but are fond of company. From the rule's implication, we learn that the marketing measures targeting at this group of potential customers should emphasize the good price/performance ratio rather the brand's name. Although such consumers try to limit both their food expenditure and their personal efforts in preparing meals, they are indulging in good food. This involves a buying motivation that is suitable for communication in marketing campaigns (see Blackwell et al. (2001) for a more detailed discussion). The rather intuitive interpretability of association rules in the segmentation context encourages a deeper consideration of this approach from the core perspective of this paper, the number of required clusters.

Table 3 focuses on the validity of both rules with respect to the available GKM solution. It becomes obvious from the given data that both rules are valid to a substantial degree for the observations assigned to the clusters represented by $\boldsymbol{\theta}_4$, $\boldsymbol{\theta}_5$, and $\boldsymbol{\theta}_6$. For these clusters, we have acceptable values for $sup_h(.)$, $conf_h(.)$, and $lift_h(.)$. This is remarkable, since rule 1 implies the acceptance of higher prices by emphasizing interests in quality as well as customer-friendly retail services, whereas rule 2 explicitly refers to lower prices. Considering the values of $conv(.)$ and $int(.)$, we cannot conclude that these rules describe the patterns underlying these clusters in an outstanding manner. In fact, the respondents of cluster 10 (see $\boldsymbol{\theta}_{10}$) minimize the measure of interestingness by simultaneously maximizing the conviction. With respect to rule 2, these measures reveal two different clusters: For cluster 14 (see $\boldsymbol{\theta}_{14}$) the interestingness is minimal and for cluster 13 (see $\boldsymbol{\theta}_{13}$) the conviction is maximal. Thus, the assessment of the rules by means of $conv(.)$, $int(.)$, and $lift(.)$ does not cause redundancy in the cluster evaluation.

With respect to dynamic changes of a pattern described by a rule, Liu et al. (2001) propose dividing the data into subsets and carrying out a χ^2-test

		rule 1					rule 2				
cluster	obs	*sup*	*conf*	*lift*	*int*	*conv*	*sup*	*conf*	*lift*	*int*	*conv*
$\boldsymbol{\theta}_1$	394	.058	.250	2.141	5.000	1.178	.056	.550	1.245	8.727	1.241
$\boldsymbol{\theta}_2$	242	.182	.423	1.528	2.886	1.253	.008	.222	1.280	24.500	1.063
$\boldsymbol{\theta}_3$	238	.063	.254	1.441	5.733	1.104	.008	.500	1.608	38.000	1.378
$\boldsymbol{\theta}_4$	327	.086	.384	1.548	4.500	1.220	.050	.333	1.758	5.444	1.215
$\boldsymbol{\theta}_5$	317	.047	.161	2.045	6.867	1.098	.060	.404	1.424	6.211	1.202
$\boldsymbol{\theta}_6$	308	.065	.256	2.025	4.850	1.175	.078	.632	1.677	5.417	1.692
$\boldsymbol{\theta}_7$	216	.194	.488	1.701	2.524	1.394	.056	.316	1.392	6.250	1.130
$\boldsymbol{\theta}_8$	242	.066	.208	1.934	5.438	1.127	.062	.319	1.576	5.400	1.171
$\boldsymbol{\theta}_9$	277	.011	.214	3.957	8.667	1.204	.087	.320	1.108	5.458	1.046
$\boldsymbol{\theta}_{10}$	173	.266	.667	2.060	1.717	2.029	.069	.667	2.261	4.750	2.116
$\boldsymbol{\theta}_{11}$	353	.139	.371	1.560	3.408	1.212	.088	.585	1.386	5.516	1.392
$\boldsymbol{\theta}_{12}$	265	.075	.244	1.795	4.900	1.143	.072	.528	1.890	4.790	1.526
$\boldsymbol{\theta}_{13}$	292	.007	.091	1.659	18.000	1.040	.089	.839	1.353	7.154	2.357
$\boldsymbol{\theta}_{14}$	333	.006	.065	2.148	19.500	1.037	.153	.614	1.411	3.471	1.464
$\boldsymbol{\theta}_{15}$	267	.236	.434	1.303	2.714	1.179	.019	.455	1.957	13.600	1.408

Table 3. Evaluation of association rules

to check whether there is any difference in the proportion of observations meeting both the rule's antecedent as well as the rule's implication vis-a-vis the proportion of observations meeting the antecedent, but not the implication. If there is no difference, they argue the rule to be stable. We act on this suggestion by utilizing this approach to check whether or not there is a difference between the clusters. To become more formal, let

$$\chi^2(\mathcal{A} \Rightarrow \mathcal{C}) = \sum_{h=1}^{k} \frac{\left[sup_h(\mathcal{A}\Rightarrow\mathcal{C})N_h - sup_h(\mathcal{A})\frac{N_h}{N}\sum_{h'=1}^{k} sup_{h'}(\mathcal{A}\Rightarrow\mathcal{C})N_{h'} \right]^2}{sup_h(\mathcal{A})\frac{N_h}{N}\sum_{h'=1}^{k} sup_{h'}(\mathcal{A}\Rightarrow\mathcal{C})N_{h'}} + \sum_{h=1}^{k} \frac{\left[sup_h(\mathcal{A}\Rightarrow\neg\mathcal{C})N_h - sup_h(\mathcal{A})\frac{N_h}{N}\sum_{h'=1}^{k} sup_{h'}(\mathcal{A}\Rightarrow\neg\mathcal{C})N_{h'} \right]^2}{sup_h(\mathcal{A})\frac{N_h}{N}\sum_{h'=1}^{k} sup_{h'}(\mathcal{A}\Rightarrow\neg\mathcal{C})N_{h'}}. \qquad (11)$$

In equation 11, a $2 \times k$ contingency table is evaluated, where the first summand accounts for the squared differences of the observed and the expected numbers of observations meeting both the antecedent and the implication of the rule. The squared differences between the observed and expected numbers of observations meeting the antecedent, but not the implication is captured by the second summand. The required critical values are given by the χ^2 distribution with $k-1$ degrees of freedom.

For rule 1, we compute $\chi^2(\mathcal{A} \Rightarrow \mathcal{C}) = 99.81 > 23,68 = \chi^{2\,crit}_{(14;0.95)}$. For rule 2, the test statistic has to be calculated in the same manner, but two clusters need to be excluded since the expected number of observations

$sup_h(\mathcal{A})\frac{N_h}{N}\sum_{h'=1}^{k} sup_{h'}(\mathcal{A} \Rightarrow \neg\mathcal{C})N_{h'} < 5$. Thus, we get $\chi^2(\mathcal{A} \Rightarrow \mathcal{C}) = 47.96 > 21,03 = \chi^{2\,crit}_{(12;0.95)}$ for rule 2. The testing of these two exemplarily chosen association rules provides us with some evidence that the clusters or segments really do differ with respect to the revealed patterns.

4 Discussion and outlook

Starting with a discussion of the properties of commonly used criteria for learning the number of clusters from high-dimensional data in unsupervised classification tasks we found the *JUMP*-criterion to be suited for market segmentation because of its methodical modesty, its ability to answer the very basic question of whether there is segmented market (i.e. $k > 1$) or not (i.e. $k = 1$), and its fair performance in previous simulation studies. The criterion is used for the external validation of two innovative approaches for market segmentation, the GKM and the SGNN algorithm, outlined herein. The GKM stops growing when the assignment of the realizations of the random vector that represents the consumer profiles to the clusters fits approximately to a multivariate normal distribution. In its application to real survey data describing nutrition and consumption behavior, the GKM algorithm resulted in exactly the same number of market segments as suggested by the $JUMP(k)$-criterion. We could identify 15 substantial clusters and 22 outliers. The SGNN algorithms terminated with a 14-cluster solution without additional outliers. Most of the SGNN-clusters have very similar interpretations regarding their implications for marketing communication and planning. This remarkable congruence shows that growing clustering algorithms, such as SGNN and GKM, are not only easy-to-use tools for data mining in marketing, but also lead to statistically reliable and meaningful results.

For the evaluation of a given cluster solution, we proposed the use of association rules in connection with different measures of meaningfulness and interestingness. For the data at hand, the patterns underlying the individual clusters turn out to be different from each other, which underlines the necessity of segment-oriented marketing strategies. This result is confirmed by an adapted χ^2-test based on the aforesaid association rules.

We can conclude that, on the one hand, we have good reasons to imply that the number of clusters is not undercut, since 22 clusters representing just one single observation (consumer) could be isolated as outliers. On the other hand, we obviously did not overestimate the number of clusters since the respective nutrition and consumption patterns differ significantly between the clusters as shown by means of rules-based χ^2-testing.

Clearly, the proceeding suggested in this paper has its limitations. In addition to the application-oriented discussion of the similarity of clusters obtained by GKM and SGNN, a more theoretical consideration of the assessment approach would be useful to endorse the available results. A suitable starting point for this might be the cluster stability measure of Roth et al. (2002). Moreover, it might be useful to extend the idea of statistically testing association rules for significant differences between clusters by applying further tests such as Fisher's exact test. Finally, the integration of the presented criteria for estimating the optimal number of clusters in the SGNN approach is a challenging task that might further improve its performance in the context of market segmentation.

References

BAIER, D., GAUL, W. and SCHADER, M. (1997): Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring, in: Klar, R. and Opitz, O. (Eds.), *Classification and Knowledge Organization.* Springer, Heidelberg, 557-566.

BLACKWELL, R.D., MINIARD, P.W., and ENGEL, J.F. (2001): *Consumer Behavior*, Harcourt, Fort Worth.

BOCK, H.H. (1985): On Some Significance Tests in Cluster Analysis. *Journal of Classification, 2, 1, 77–108.*

BOCK, H.H. (1996): Probability Models in Partitional Cluster Analysis. *Computional Statistics and Data Analysis, 23, 5, 5–28.*

BOONE, D.S. and ROEHM, M. (2002): Evaluating the Appropriateness of Market Segmentation Solutions Using Artificial Neural Networks and the Membership Clustering Criterion. *Marketing Letters, 13, 4, 317–333.*

BRASSINGTON, F. and PETTITT, S. (2005): *Essentials of Marketing.* Prentice Hall, Harlow.

BRĀZMA, A., JONASSEN, I., EIDHAMMER, I., and GILBERT, D. (1998): Approaches to the Automatic Discovery of Patterns in Biosequences. *Journal of Computional Biology, 5, 2, 277–304.*

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.

CALINSKI, R. and HARABASZ, J. (1974): A Dendrite Method for Cluster Analysis. *Communications in Statistics (Series A), 3, 1, 1–27.*

DECKER, R. (2005): Market Basket Analysis by Means of a Growing Neural Network, *The International Review of Retail, Distribution and Consumer Research*, forthcoming.

DIBB, S. and SIMKIN, L. (1994): Implementation Problems in Industrial Market Segmentation. *Industrial Marketing Management, 23, 1, 55–63.*

DIBB, S. and STERN, P. (1995): Questioning the Reliability of Market Segmentation Techniques. *Omega – International Journal of Management, 3, 6, 625–636.*

DUDOIT, S. and FRIDLYAND, J. (2002): A Prediction-Based Resampling Method of Estimating the Number of Clusters in a Dataset, *Genome Biology, 3, 7, 1–21.*

FENNELL, G., ALLENBY, G.M., YANG, S., and EDWARDS, Y. (2003): The Effectiveness of Demographic and Psychographic Variables for Explaining Brand and Product Category Use. *Quantitative Marketing and Economics, 1, 2, 223–244.*

GAUL, W. and L. SCHMIDT-THIEME (2002): Recommender Systems Based on User Navigational Behavior in the Internet, *Behaviormetrika, 29, 1, 1–22.*

GREEN, P.E. and KRIEGER, A.M. (1995): Alternative Approaches to Cluster-Based Market Segmentation. *Journal of the Market Research Society, 3, 221–239.*

GRANZIN, K.L., OLSEN, J.E., and PAINTER, J.J. (1998): Marketing to Consumer Segments Using Health-Promoting Lifestyles. *Journal of Retailing and Consumer Services, 5, 3, 131–141.*

HAIR, J.F., ANDERSON, R.E., TATHAM, R.L., and BLACK, W.C. (1998): *Multivariate Data Analysis.* 5^{th} ed., Prentice Hall, Upper Saddle River.

HAMERLY, G. and ELKAN, C. (2003): Learning the k in k-means. Advances in Neural Information Processing Systems, 17, http://www.citeseer.ist.psu.edu/hamerly03learning.html.

HARTIGAN, J.A. (1985): Statistical Theory in Clustering. *Journal of Classification, 2, 1, 63–76.*

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.

KRZANOWSKI, W. and LAI, Y. (1988): A Criterion for Determining the Number of Clusters in a Dataset Using Sum of Squares Clustering. *Biometrics, 44, 1, 23–34.*

LIU, B., MA, Y., and LEE, R. (2001): Analyzing the Interestingness of Association Rules From the Temporal Dimension. *IEEE International Conference on Data Mining (ICDM-2001)*, http://www.cs.uic.edu/ liub/publications/ICDM-2001.ps.

LU, C.S. (2003): Market Segment Evaluation and International Distribution Centers. *Transportation Research Part E: Logistics and Transportation Review, 39, 1, 49–60.*

MARDIA, K.V. (1974): Applications of Some Measures of Multivariate Skewness and Kurtosis for Testing Normality and Robustness Studies. *Sankhya, 36, 115–128.*

MARDIA, K.V., KENT, J.T., and BIBBY, J.M. (1979): *Multivariate Analysis.* Academic Press, London.

MECKLIN, C.J. and MUNDFROM, D.J. (2004): An Appraisal and Bibliography of Tests for Multivariate Normality. *International Statistical Review, 72, 1, 123–138.*

MILLIGAN, G.W. and COOPER, M.C. (1985): An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika, 50, 159–179.*

PALMER, R.A. and Millier, P. (2004): Segmentation: Identification, Intuition, and Implementation. *Industrial Marketing Management, 33, 8, 779–785.*

PAPASTEFANOU, G., SCHMIDT, P., BRSCH-SUPAN, A. and LDTKE, H. and OLTERSDORF, U. (1999): *Social and Economic Research with Consumer Panel Data.* GESIS, Mannheim.

ROTH, V., LANGE, T., BRAUN, M., and BUHMANN, J. (2002): A Resampling Approach to Cluster Validation. in: W. Härdle and B. Rönz (Eds.): *Proceedings in Computational Statistics.* Physica, Heidelberg, 123–128.

SUGAR, C.A. and JAMES, G.M. (2003): Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach. *Journal of the American Statistical Society, 98, 463, 750–762.*

TIBSHIRANI, R., WALTER, G., and HASTIE, T. (2001): Estimating the Number of Clusters in a Dataset via the Gap Statistic. *Journal of the Royal Statistical Society (Series B), 63, 3, 411–423.*

WAGNER, R. (2005): Mining Promising Qualification Patterns. In: D. Baier and K.-D. Wernecke (Eds.): *Innovations in Classification, Data Science, and Information Systems.* Berlin, Springer, 249–256.

WEDEL, M. and KAMAKURA, W.A. (2000): *Market Segmentation: Conceptional and Methodological Foundations.* 2^{nd} ed., Kluwer Academic Publishers, Dordrecht.

On Variability of Optimal Policies in Markov Decision Processes

Karl-Heinz Waldmann

Institut für Wirtschaftstheorie und Operations Research,
Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

Abstract. Both the total reward criterion and the average reward criterion commonly used in Markov decision processes lead to an optimal policy which maximizes the associated expected value. The paper reviews these standard approaches and studies the distribution functions obtained by applying an optimal policy. In particular, an efficient extrapolation method is suggested resulting from the control of Markov decision models with an absorbing set.

1 Introduction

Nearly a half century has elapsed since the publication of Richard Bellman's path-breaking 1957 book, *Dynamic Programming.* In the meantime, Markov decision processes, also referred to as stochastic dynamic programs or stochastic control problems, are a key tool in analyzing and optimizing sequential decision problems under uncertainty. There is a rich field of applications including agriculture, biology, business administration, ecology, economics, engineering, and sports. E.g., the recent survey of Altman (2001) on applications in communication networks already gives nearly 300 references.

A Markov decision process (*MDP* for short) formally consists of a tuple (S, A, D, p, r, β) of objects of the following meaning: S, the state space, A, the action space A, $D(s)$ the set of admissible actions in state $s \in S$, $D := \{(s,a) \mid s \in S, a \in D(s)\}$, the constraint set, p, the transition law from D into S, $r : D \to \mathbb{R}$, the one-stage reward function, and $\beta > 0$, the discount factor.

To illustrate an *MDP*, consider a system, which is observed at discrete times $n \in \mathbb{N}_0$, where $\mathbb{N}_0$ (resp. $\mathbb{N}$) denotes the set of all nonnegative (resp. positive) integers. If the system is in state $s_n \in S$ at time n, then a decision maker selects an admissible action $a_n \in D(s_n)$. Associated with the actual state and action is a one-stage reward $r(s_n, a_n)$, which is discounted to time 0 with discount factor β. The system then moves to some state $s_{n+1} \in S$ at time $n+1$ with probability $p(s_n, a_n, s_{n+1})$.

Throughout the paper we suppose S and A to be countable, $D(s)$, $s \in S$, to be finite, and r to be bounded.

A policy $\pi = (f_0, f_1, \ldots)$ is a sequence $f_0, f_1, \ldots$ of decision rules specifying the action $a_n = f_n(s_n)$ to be taken in state s_n at time n. Let $F := \{f : S \to A \mid f(s) \in D(s)\}$ be the set of all decision rules and F^∞ be the set of all policies.

Mainly one is interested in stationary policies $\pi = (f, f, \ldots)$ for some $f \in F$, for which we also write f.

Let ζ_n denote the random variable describing the state of the process at time $n \in \mathbb{N}_0$. We refer to (ζ_n) as the state process. Further, for all policies $\pi \in F^\infty$ and all initial states $s \in S$, we use $P_{\pi s}$ to denote the distribution of (ζ_n) and $E_{\pi s}$ to denote the expectation associated with $P_{\pi s}$.

Most often policies are compared with respect (w.r.) to the total discounted reward or the average reward.

Given $\beta < 1$, let $V_\pi(s) := E_{\pi s} \sum_{n=0}^{\infty} \beta^n r(\zeta_n, f_n(\zeta_n))$ denote the expected total discounted reward starting in state s and following policy π. Then $V(s) := \sup_{\pi \in F^\infty} V_\pi(s)$, $s \in S$, is the maximal expected total discounted reward starting in state s. A policy π^* is called β–optimal if $V_{\pi^*}(s) = V(s)$ holds for all $s \in S$. We also say that a decision rule f^* is β–optimal if the associated stationary policy is β–optimal.

Further we use $G_\pi(s) := \liminf_{N \to \infty} E_{\pi s} \frac{1}{N} \sum_{n=0}^{N-1} r(\zeta_n, f_n(\zeta_n))$, $\pi \in F^\infty$, $s \in S$, to denote the expected average reward and let $G(s) := \sup_{\pi \in F^\infty} G_\pi(s)$, $s \in S$, be the maximal expected average reward. A policy π^* (resp. decision rule f^*) is called average–optimal, if $G_{\pi^*}(s) = G(s)$, $s \in S$.

Applied to these classical optimality criteria the standard results (i.e. existence of an optimality equation, optimality of a decision rule, and validity of the value iteration) are well known to hold and summarized in section 2 below.

Applying an optimal decision rule we only know that the expectation of the total discounted reward (resp. average reward) becomes maximal but we do not know how much the actual value may differ from its expected value. Therefore it may be of importance to also know the associated distribution function or at least some specified percentage points.

In section 3, applied to the total discounted reward, an extrapolation method is suggested for obtaining the distribution function in an efficient way. The approach is based on the extrapolation method derived in Hinderer and Waldmann (2004) for determining the value function of an MDP with an absorbing set. Finally, in case of the average reward criterion, the average reward is shown to be concentrated in a one-point measure and thus coincides with its expected value (a.s.).

2 The total/average reward criterion

Let $\mathfrak{V}_S$ denote the Banach space of all bounded functions on S (w.r. to the supremum norm $\|v\| := \sup_{s \in S} |v(s)|$).

On $\mathfrak{V}_S$ introduce the operator

$$Uv(s) := \max_{a \in D(s)} \left\{ r(s,a) + \beta \sum_{s' \in S} p(s,a,s')v(s') \right\}, \quad s \in S,\ v \in \mathfrak{V}_S.$$

Put $U^{n+1}v := U(U^n v)$, $v \in \mathfrak{V}_S$, $n \in \mathbb{N}_0$. Analogously for the operators to be introduced later.

Theorem 1. *Assume $\beta < 1$. Then*

(i) Value iteration works, i.e. for all $v_0 \in \mathfrak{V}_S$ it holds that $v_n := U^n v_0$ converges in norm to V as $n \to \infty$.
(ii) V is the unique bounded solution of the optimality equation $V = UV$,

$$V(s) = \max_{a \in D(s)} \left\{ r(s,a) + \beta \sum_{s' \in S} p(s,a,s')V(s') \right\}, \quad s \in S. \tag{1}$$

(iii) Each decision rule $f \in F$ that realizes the maximum in (1) is $\beta-$optimal; thus there exists an $\beta-$optimal stationary policy.

In case of the average reward criterion the situation is more complicate. To exploit the dependence of β in V, we will write V_β in place of V.

Theorem 2. *Let S be finite. Then*

(i) $G(s) = \lim_{\beta \uparrow 1}(1-\beta)V_\beta(s)$, $s \in S$.
(ii) There exists an average$-$optimal decision rule f^*, say.
(iii) f^* is $\beta-$optimal for all $\beta \geq \beta^*$ and large enough β^*.

To have an optimality equation for the average reward criterion, too, additional assumptions are necessary. We only consider the unichain case and refer to Sennott (1999) for both the multichain case and a countable state space.

(GA) For all $f \in F$, the transition matrix of the underlying Markov reward process is irreducible or at least has one recurrent class only.

Theorem 3. *Let S be finite and suppose (GA) to hold. Then there exists a constant $g \in \mathbb{R}$ and a function $h : S \to \mathbb{R}$ such that*

$$g + h(s) = \max_{a \in D(s)} \left\{ r(s,a) + \sum_{s' \in S} p(s,a,s')h(s') \right\} \tag{2}$$

and we have

(i) $G(i) = g$ for all $s \in S$.
(ii) Each $f \in F$ that realizes the maximum in (2) is average$-$optimal.

There is a rich literature on the solution of the optimality equations (1) and (2). E.g. Puterman (1994) may serve as a starting point for further investigations. Concerning the total reward criterion we also refer to Hinderer and Waldmann (2004) within the more general setting of an *MDP* with an absorbing set and/or unbounded one-stage rewards.

3 Distribution function of the total/average reward

Having a β–optimal decision rule in mind, let $f \in F$ be fixed throughout this section. Then the MDP reduces to a Markov reward process with transition matrix $(p(s,s'))_{s,s' \in S}$, bounded one-stage rewards $r(s)$, $s \in S$, and discount factor $\beta > 0$, say.

Let $c \in \mathbb{R}$. Then, given $\beta < 1$, two models with one-stage rewards $r(s)$ and $r(s) - c$, respectively, differ w. r. to the total discounted reward only by $c(1-\beta)^{-1}$. Analogously for the average reward criterion, where both models differ by c only. Therefore we suppose $r \geq 0$ without loss of generality, and, to hold the notation simple, assume $0 = r_{min} \leq r(s) \leq r_{max}$, $s \in S$, for some constant r_{max}.

For $(s,x) \in S \times \mathbb{R}$ introduce $\Phi(s,x)$ to be the probability (risk) that the actual total discounted reward, starting in s, is smaller than or equal to x,

$$\Phi(s,x) := P\left(\sum_{n=0}^{\infty} \beta^n r(\zeta_n) \leq x \mid \zeta_0 = s\right).$$

For all $s \in S$, $\Phi(s,\cdot)$ is a distribution function, i.e. a nondecreasing, right continuous function in x, fulfilling $\lim_{x \to -\infty} \Phi(s,x) = 0$ and $\lim_{x \to +\infty} \Phi(s,x) = 1$. Further, from Lemma 4 in Wu and Lin (1999) we get that Φ is a solution to the functional equation

$$\Phi(s,x) = \sum_{s' \in S} p(s,s')\Phi(s',(x-r(s))/\beta), \quad s \in S,\ x \in \mathbb{R}, \tag{3}$$

which can be obtained by the method of successive approximation, starting with $1_{[0,\infty)}(x)$, where $1_C(x) = 1$ for $x \in C$ and 0, otherwise.

Recall that $0 \leq r(s) \leq r_{max}$, $s \in S$. Thus, for $x < 0$, we additionally have $\Phi(s,x) = 0$, $s \in S$. Analogously, for $x \geq \rho := r_{max}(1-\beta)^{-1}$ it follows that $\Phi(s,x) = 1$, $s \in S$. Hence we may rewrite (3) as

$$\Phi(s,x) = 1_{(\rho,\infty)}(Y(s,x)) + \sum_{s' \in S} p(s,s')1_{[0,\rho]}(Y(s,x))\Phi(s',Y(s,x))$$

for $s \in S$, $x \in [0,\rho]$, and $Y(s,x) := (x - r(s))/\beta$ for $s \in S$, $x \in \mathbb{R}$.

¿From a practical point of view it is usually sufficient to have $\Phi(s,x)$ for a countable set X of total discounted rewards x. Together with the countable state space S we then get Φ as the solution of the optimality equation of a Markov reward process MRP' with countable state space and bounded one-stage rewards. The approach has the advantage that it allows us to apply the general results obtained in Hinderer and Waldmann (2004) for an MDP with an absorbing set in a direct way and also to exploit the finite state approximations given there.

To be more precise, let $X \subset [0,\rho]$ be a countable set such that $x' = (x - r(s))/\beta$ belongs to X, provided that $x \in X$ and $x' \in [0,\rho]$. Introduce an additional constant x_0, say, $x_0 \notin X$. Then the definition of MRP' can be completed by choice of the state space $S' = S \times (\{x_0\} \cup X)$, the transition probabilities $p'((s,x),(s',x')) = p(s,s')$ for $x' = (x - r(s))/\beta$, $(s,x),(s',x') \in S \times X$, and $p'((s,x_0),(s,x_0)) = 1$, $s \in S$, the one-stage rewards $r'(s,x) = 1_{(\rho,\infty)}(Y(s,x))$ for $(s,x) \in S \times X$, and $r'(s,x_0) = 0$, $s \in S$, and, finally, discount factor $\beta' = 1$. Note that $J_0 := S \times \{x_0\}$ is an absorbing set of MRP'. We also refer to $J := S \times X$ as the essential state space of MRP'.

For $v, v' \in \mathbb{R}^J$ we write $v \leq v'$ (resp. $v < v'$), if $v(s) \leq v'(s)$ (resp. $v(s) < v'(s)$) holds for all $s \in J$.

Let $\mathfrak{V}$ denote the Banach space of all bounded functions on J (w.r. to the supremum norm $\|v\| := \sup_{z \in J} |v(z)|$), $\mathfrak{V}_+ := \{v \in \mathfrak{V} \mid v \geq 0\}$, and $\mathfrak{W} := \{v \in \mathfrak{V} \mid \inf v > 0\}$. If $v \in \mathfrak{V}$ and $w \in \mathfrak{W}$, then $v/w \in \mathfrak{V}$ and the norm $\|v\|_w := \|v/w\|$ induced by w is equivalent to the supremum norm.

On $\mathfrak{V}$ define the operator H by

$$Hv(s,x) := \sum_{s' \in S} p(s,s') 1_X(Y(s,x)) v(s', Y(s,x)), \quad (s,x) \in J.$$

For an arbitrary operator T from $\mathfrak{V}$ into $\mathfrak{V}$ (with $Tv \in \mathfrak{V}_+$ for $v \in \mathfrak{V}_+$) we use the following definition according to Ogiwara (1995), pp. 47-49: $\lambda \in \mathbb{R}_+$ is called an eigenvalue of T if $Tv = \lambda v$ for some $v \in \mathfrak{V}_+$, $v \not\equiv 0$. Note that it is narrower than the usual definition.

We call H primitive, if for each $v \in \mathfrak{V}_+, v \not\equiv 0, v \notin \mathfrak{W}$ there exists $m = m(v) \in \mathbb{N}$ such that $H^m v \in \mathfrak{W}$. For finite sets J this definition is consistent with the usual one.

Some of our results are based on the following assumption (A0), which is shown in Hinderer and Waldmann (2004), Proposition 2.1, to be equivalent to the compactness of H.

For $z \in J$, $M \subset J$, set $p'(z,M) := \sum_{z' \in M} p'(z,z')$. Further we use $A - B$ to denote $\{a \in A \mid a \notin B\}$.

(A0) The set $\{p'(z,.), z \in J\}$ of substochastic measures on J is (uniformly) tight, i.e. for each $\varepsilon > 0$ there exists a finite set $K \subset J$ such that $p'(z, J - K) \leq \varepsilon$.

If S is finite, then (A0) reduces to the existence of a finite set $K_X \subset X$ such that the distance between K_X and X (w.r. to the usual metric on $\mathbb{R}$) is not greater than ε (which can be easily realized).

Let $e_0 := 1$ and $e_n(z) := H^n 1(z)$, $z \in J$. Obviously $\|e_n\| = \|H^n 1\|$ is an upper bound for the probability that the process has not yet entered the absorbing set J_0 at time n. The asymptotic behavior of $\|e_n\|$ plays a key role in determining the iteration schemes.

To begin with let $\lambda^* := \inf_{k \in \mathbb{N}} \|e_k\|^{1/k}$ be the spectral radius of the continuous linear operator H.

Proposition 1. *There holds*

(*i*) $0 \leq \lambda^* = \lim_{n\to\infty} \|e_n\|^{1/n} \leq \|e_1\| \leq 1$.
(*ii*) $\lambda^* = \lim_{n\to\infty}(\|H^n v\|_w)^{1/n}$ *for all* $v, w \in \mathfrak{W}$.
(*iii*) *Assume (A0). Let* $\lambda^* > 0$. *Then* λ^* *is an eigenvalue of* H. *Moreover, it is the largest one.*
(*iv*) *There are equivalent*
  (a_1) $\lambda^* < 1$.
  (a_2) $\|e_m\| < 1$ *for some* $m \in \mathbb{N}$.
  (a_3) $\|e_n\| \to 0$ *as* $n \to \infty$.
  (a_4) $\sum_{n=0}^{\infty} \|e_n\| < \infty$.
  (a_5) $\|e_n\| \leq c \cdot \delta^n$ *for all* $n \in \mathbb{N}$ *and some* $c \in \mathbb{R}_+$ *and* $\delta \in (0,1)$.
  (a_6) $\lim_{n\to\infty} \|H^n v\|_w = 0$ *for all* $v \in \mathfrak{V}$, $w \in \mathfrak{W}$.
(*v*) *Assume (A0). Then there are equivalent*
  (b_1) $\lambda^* = 1$.
  (b_2) *There exists* $z_0 \in J$ *such that* $e_n(z_0) = 1$ *for all* $n \in \mathbb{N}$.
  (b_3) *There exists* $M \subset J$, $M \neq \emptyset$, *such that* $p'(z, M) = 1$ *for all* $z \in M$.

Moreover, under (A0), (a_1) *is equivalent to the existence of a finite partition* $J_1, \ldots, J_m$ *of* J *with nonempty sets* $J_k := \{z \in J \mid e_k(z) < 1 = e_{k-1}(z)\}$, $1 \leq k \leq m$.

Proposition 1 is part of Propositions 2.2, 2.3, 2.5, and 2.6 in Hinderer and Waldmann (2004), where a large number of characterizations for the critical discount factor of an *MDP* are derived.

On $\mathfrak{V}$ define the operator L by

$$Lv(s,x) := 1_{(\rho,\infty)}(Y(s,x)) + Hv(s,x), \quad (s,x) \in J.$$

Then, from Theorem 3.1 in Hinderer and Waldmann (2004), we get

Theorem 4. *Let* $\lambda^* < 1$. *Then* Φ *is the unique solution of* $\Phi = L\Phi$ *in* $\mathfrak{V}$ *and can be obtained by value iteration, i.e.* $\Phi = \lim_{n\to\infty} L^n v_0$, *starting with any* $v_0 \in \mathfrak{V}$.

Consider a sequence (v_n) of successive approximations $v_n = Lv_{n-1}$, $n \in \mathbb{N}$, starting with some $v_0 \in \mathfrak{V}$. Further let $d_n := v_n - v_{n-1}$ denote the difference of two successive approximations. Of course, $d_n \in \mathfrak{V}$.

Value iteration is now combined with an extrapolation, giving upper and lower bounds for Φ of the form $v_n + c_n^{\pm} d_n$ at each step n of iteration.

Our extrapolation method is based on the assumption that $d_1 \geq 0$. Then $d_n \geq 0$ for all $n \geq 1$. Since $1_{(\rho,\infty)} \geq 0$, then $d_1 \geq 0$ trivially holds using $v_0 \equiv 0$.

For $n \geq 2$ let

$$\alpha_n^- := \inf_{z \in J} \{d_n(z)/d_{n-1}(z) \mid d_{n-1}(z) > 0\}$$

(with $\alpha_n^- := 0$, if $d_{n-1} \equiv 0$). Note that $0 \leq \alpha_n^- d_{n-1} \leq d_n = Hd_{n-1}$. Together with Lemma 3.1.7(iii) in Ogiwara (1995) we then get $\alpha_n^- \leq \lambda^*$. Using monotonicity of H we further get $Hd_n \geq H(\alpha_n^- d_{n-1}) = \alpha_n^-[Lv_{n-1} - Lv_{n-2}] = \alpha_n^- d_n$ and thus $d_{n+1} = Hd_n \geq \alpha_n^- d_n$, which implies $\alpha_n^- \leq \alpha_{n+1}^-$.

Next, for $k \geq 1$ select $\delta_k \in \mathfrak{V}$ such that $\hat{d}_k := d_k + \delta_k \in \mathfrak{W}$. Let

$$\hat{\alpha}_{k,m}^+ := \sup_{z \in J} \left\{ H^m \hat{d}_k(z)/\hat{d}_k(z) \right\}, \quad m \in \mathbb{N}.$$

Obviously H^m is monotone and positively homogeneous. It follows from Proposition 1(i) that H^m has the spectral radius $(\lambda^*)^m$. Using $H^m \hat{d}_k \leq \hat{\alpha}_{k,m}^+ \hat{d}_k$, we obtain $\hat{\alpha}_{k,m}^+ \geq (\lambda^*)^m$ by Lemma 3.1.7(ii) in Ogiwara (1995) for $T := H^m$. Further, by applying Proposition 1(a_6) with $v = w = \hat{d}_k$, we get $\hat{\alpha}_{k,m}^+ \to 0$ as $m \to \infty$, if $\lambda^* < 1$. In particular, $\hat{\alpha}_{k,m}^+ < 1$ for large enough m.

We are now in a position to state as a Corollary to Theorem 5.1 and Proposition 5.5 in Hinderer and Waldmann (2004)

Theorem 5. *Let $d_1 \geq 0$ and $\lambda^* < 1$. Then*

(i) For all $n > 1$ it holds that $\alpha_n^- \leq \lambda^* < 1$ and

$$\Phi \geq w_n^- := v_n + \alpha_n^-(1 - \alpha_n^-)^{-1} d_n \geq v_n.$$

(ii) Let $\hat{d}_k \in \mathfrak{W}$ for some $k \in \mathbb{N}$, and $m \in \mathbb{N}$ such that $\hat{\alpha}_{k,m}^+ < 1$. Then for all $n \geq k$

$$\Phi - v_n \leq \sup_{z \in J} \left\{ \frac{d_n(z)}{\hat{d}_k(z) - H^m \hat{d}_k(z)} \right\} \cdot \sum_{j=1}^{m} H^j \hat{d}_k \leq \frac{\|d_n/\hat{d}_k\|}{1 - \hat{\alpha}_{k,m}^+} \cdot \sum_{j=1}^{m} H^j \hat{d}_k.$$

In particular, if $d_{n-1} \in \mathfrak{W}$, and $\alpha_n^+ := \sup_{z \in J}\{d_n(z)/d_{n-1}(z)\} < 1$ hold for some $n \in \mathbb{N}$, then

$$\Phi \leq w_n^+ := v_n + \alpha_n^+(1 - \alpha_n^+)^{-1} d_n.$$

(iii) The weights $c_n^- := \alpha_n^-/(1-\alpha_n^-)$ are increasing in n, $\lim_{n\to\infty} c_n^- \leq \lambda^*/(1-\lambda^*)$. Further, the lower bounds w_n^- are increasing in n and converge in norm to Φ as $n \to \infty$.

(iv) Under the assumptions in (ii) we get $\hat{w}_{n+m,k,m}^+ \leq \hat{w}_{n,k,m}^+ := v_n + \|d_n/\hat{d}_k\| \cdot (1 - \hat{\alpha}_{k,m}^+)^{-1} \sum_{j=1}^m H^j \hat{d}_k$. Further, both upper bounds in (ii) converge in norm to Φ as $n \to \infty$.

(v) Assume that H is primitive and that H^ν is compact for some $\nu \in \mathbb{N}$. (The latter holds under (A0), in particular if J is finite.) Then the procedure either stops with $d_{n_0} \equiv 0$ for some $n_0 \in \mathbb{N}$ (hence $\Phi = v_{n_0}$) or $d_n \in \mathfrak{W}$ for $n \geq n_1$ and the weights $c_n^+ := \alpha_n^+/(1 - \alpha_n^+)$ are decreasing in $n \geq n_1$ and both $c_n^\pm$ converge to $\lambda^*/(1 - \lambda^*)$ as $n \to \infty$.

The lower bounds are easy to calculate and need $d_1 \geq 0$ only. The upper bounds are based on $\hat{d}_k \in \mathfrak{W}$ and need some additional effort due to the computation of $H^j \hat{d}_k$, $1 \leq j \leq m$, up to some $m \in \mathbb{N}$ such that $\hat{\alpha}^+_{k,m} < 1$. To reduce the computational effort, we may use these fixed objects for more than one step of iteration. In this case, we only have to adapt v_n and $\|d_n/(\hat{d}_k - H^m \hat{d}_k)\|$ (resp. $\|d_n/\hat{d}_k\|$) by switching from n to $n+1$, which can be easily realized.

We may use $\delta_1 = e_0 - d_1$ in order to obtain $\hat{d}_1 = e_0 \in \mathfrak{W}$. Then, for $m \in \mathbb{N}$ such that $\|e_m\| < 1$ it follows from Theorem 5(ii) for $n \geq 1$

$$\Phi - v_n \leq \sup_{z \in J} \left\{ \frac{d_n(z)}{1 - e_m(z)} \right\} \cdot \sum_{j=1}^{m} e_j \leq \frac{\|d_n\|}{1 - \|e_m\|} \cdot \sum_{j=1}^{m} e_j .$$

Mainly, however, we are interested in applying (ii) with $\delta_k = 0$, provided that $d_k \in \mathfrak{W}$. Extensive numerical results show that the extrapolation method works very well and, in particular, is highly superior to an extrapolation method of MacQueen type, which cannot balance out the unequal row sums of H. Finally, if $d_k \notin \mathfrak{W}$, we may use $\hat{d}_k = d_k + \varepsilon e_0$ for some $\varepsilon > 0$.

Nevertheless, the lower bounds can be handled in a much easier way. Thus, being interested in upper bounds for Φ only, we may apply Theorem 5 to $\bar{\Phi}(s,x) := 1 - \Phi(s,x)$, $(s,x) \in J$.

Theorem 6. *Let $\lambda^* < 1$. Then $\bar{\Phi}$ is the unique solution of*

$$\bar{\Phi}(s,x) = \bar{L}\bar{\Phi}(s,x) := 1_{(-\infty,0)}(Y(s,x)) + H\bar{\Phi}(s,x), \quad (s,x) \in J,$$

and can be obtained by value iteration, i.e. $\bar{\Phi} = \lim_{n\to\infty} \bar{L}^n v_0$, starting with any $v_0 \in \mathfrak{V}$.

Indeed, by applying Theorem 5 on the basis of Theorem 6, for any lower bound w, say, to $\bar{\Phi}$, we have $\Phi(s,x) = 1 - \bar{\Phi}(s,x) \leq 1 - w(s,x)$, $(s,x) \in J$.

¿From a numerical point of view, it is necessary to approximate MRP' by one with a finite state space. Such an approximation can be realized in a natural way on the basis of assumption (A0). Details of the approximation can be found in Hinderer and Waldmann (2004), section 8, together with upper and lower bounds for the resulting error.

Finally, to complete the discussion, we throw a short glance at the average reward criterion. For fixed $f \in F$, let the resulting Markov chain have (exactly) one positive recurrent class. Then, since r is bounded, we have (e.g. Norris (1997), Theorem 1.10.2)

$$P\left(\frac{1}{N} \sum_{n=0}^{N-1} r(\zeta_n) \to \sum_{s \in S} \pi(s) r(s) \quad \text{as} \quad n \to \infty \right) = 1,$$

where $(\pi(s), s \in S)$ is the unique invariant distribution of the Markov chain.

Hence, the distribution of the average reward is concentrated in a one-point measure and thus coincides with its expected value (a.s.).

References

ALTMAN, E. (2001): Applications of Markov Decision Processes in Communication Networks: a Survey. In E. Feinberg and A. Shwartz (Eds): *Markov Decision Processes, Models, Methods, Directions, and Open Problems.* Kluwer, Bosten, 488–536.

HINDERER, K. and WALDMANN, K.-H. (2004): Algorithms for countable state Markov decision models with an absorbing set. *SIAM J. Control and Optimization, to appear.*

NORRIS, J.R. (1997): *Markov Chains.* Cambridge University Press, Cambridge.

OGIWARA, T. (1995): Nonlinear Perron-Frobenius problem on an ordered Banach space. *Japan J. Math., 21, 43–103.*

PUTERMAN, M.L. (1994): *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* Wiley, New York.

SENNOTT, L.I. (1999): *Stochastic Dynamic Programming and the Control of Queueing Systems.* Wiley, New York.

WU, C. and LIN, Y. (1999): Minimizing Risk Models in Markov Decision Processes with Policies Depending on Target Values. *J. Math. Anal. Appl., 231, 47–60.*

Part II

Decision Support

Linking Quality Function Deployment and Conjoint Analysis for New Product Design

Daniel Baier and Michael Brusch

Chair of Marketing and Innovation Management, Brandenburg University of Technology, Konrad-Wachsmann-Allee 1, D-03046 Cottbus, Germany

Abstract. In this paper we develop a new approach to evaluate the importance of technical features and engineering characteristics in the eyes of the customer by combining (1) the estimated influence of these so-called product characteristics on relevant product attributes in the eyes of the customer with (2) the estimated importance of these attributes for the customer. For both estimations conjoint analysis is applied. The new approach is compared to the traditional one w.r.t. predictive validity. Two sample applications are used for demonstration. We show that the new approach outperforms the traditional approach.

1 Introduction

Quality function deployment (QFD) is a systematic process for new product development by which a multi-functional team deploys operations on the factory floor by implementing the voice of the customer (Akao (1990)). QFD was developed and first used in Japan in the mid 1960s, the first reported success story was the "Toyota rust study" in the late 1970s (Sullivan (1986)) which already showed the two main advantages of QFD compared to other new product development techniques: (1) better and earlier determination of product attributes and the product quality searched by the customer and (2) better and earlier determination of key manufacturing requirements in advance. QFD was introduced to the western world in the 1980s and has been consequently improved and modified since then. Recent surveys (Cristiano et al. (2000) on QFD usage in Japan and the U.S.) and literature reviews (Chan and Wu (2002) with 650 reviewed QFD articles) document the wide spread usage of this tool in research and practice.

Three main approaches of the work-process have been presented in the literature: the comprehensive QFD concept (Akao (1990), Mizuno and Akao (1994)), the 30 table concept (King (1989)) and the four-phase concept (ASI (1989)). All is common that along a chain of tables and graphs, the customer's requirements are transformed into technical features and engineering characteristics that lead – finally – to certain demands on production and assembly. So, e.g., in a first step the relation between

- customer's requirements described by product attributes (PAs) like "quiet' "low maintenance" in the voice of the customer and

- technical features described by engineering or product characteristics (PCs) like "sound-emission in dBA", "maintenance interval in months", or "efficiency" in the voice of the engineer

is formalized and documented in a key document, the so-called house of quality. For this parametrization, objective measurement has often been demanded in the literature (i.e. with the help of controlled experiments). However, instead, often group discussions are used with a high level of subjectivity (see, e.g., the description of the procedures in Akao (1990)).

To eliminate this potential source of error, various authors proposed the usage of regression analysis (e.g. Hauser and Simmie (1981), Yoder and Mason (1995), Askin and Dawson (2000)) for measuring the PC influence on the PAs or the integration of conjoint analysis (CA) into the approach (see, e.g., Urban and Hauser (1993), Gustafsson (1996), Pullman et al. (2002), Katz (2004)).

CA (see, e.g., Green and Srinivasan (1978, 1990)) is a wide spread group of methods in new product design (Gaul et al. (1995), Baier and Gaul (1999, 2003)) which makes it possible to measure influences on attributes, preferences or buying intentions by controlled experiments. CA seems to be a promising supplementary tool for QFD (see, e.g, Katz (2004)). However, concrete applications and comparisons with traditional QFD approaches are rare (e.g. Baier (1998)).

The present article takes this problem into account by proposing a new CA based approach for QFD (section 2). Two sample applications are used for comparisons (section 3 and 4). A discussion closes the article (section 5).

2 Linking QFD and CA

The essential advantage of CA compared to direct measurement is the "conjoint" evaluation task (see, e.g., Brusch et al. (2002)). The influence or importance of product features is measured via evaluations of synthetic feature level combinations, what forces the respondents to make trade-offs between different features and feature levels. We make use of this in our new CA based QFD approach when PA importances in the eyes of the customer (step 2 of the new approach) and when PC influences on the PAs are measured (step 4). Table 1 summarizes the five steps of the CA based approach in comparison to the traditional approach. For the application of adaptive conjoint analysis (ACA, Johnson (1987), Sawtooth Software (2002)), the most popular CA technique (e.g., Baier (1999)), more hints are given in steps 2 and 4.

First, PAs are selected. In both approaches (also true for the third step) known techniques can be used for this purpose (e.g. focus groups, rep-tests, explorative interviews). However, the CA based approach requires an attribute and level number limitation. The second step is the evaluation of the PA importance. In the traditional approach this is done on the basis of direct questioning of customers or experts, e.g., the so-called QFD team. In the CA

	Traditional approach	CA based approach
Step 1: Selecting product attributes (PAs)		
No. of PAs	arbitrary	max. 30 (ACA)
No. of PA levels	no levels specified	2 up to 9 for each PA (ACA)
Step 2: Evaluating PA importances		
Respondents	customers, QFD Team	customers, QFD Team
Data collection and analysis	direct ratings of PAs, e.g. 0(unimp.), ..., 4(important)	ratings of stimuli, analyzed by OLS imp. estimation (ACA)
Results	PA (importance) shares	PA (importance) shares
Step 3: Selecting product characteristics (PCs)		
No. of PCs	arbitrary	max. 30 (ACA)
No. of PC levels	no levels specified	2 up to 9 for each PC (ACA)
Step 4: Evaluating the influence of PCs on PAs		
Respondents	experts, QFD Team	experts, QFD Team
Data collection and analysis	group discussion w.r.t. each PA resulting in 0, 1(=$\triangle$), 3(=$\bigcirc$) or 9(=$\odot$) as weights for each PC	for each PA: ratings of stimuli, analyzed by OLS importance estimation (ACA)
Results	PC (imp.) shares per PA	PC (imp.) shares per PA
Step 5: Computing PC importances in the eyes of the customer		
Calculation	sum of PC shares per PA weighted by PA shares	sum of PC shares per PA weighted by PA shares

Table 1. Traditional vs. CA based approach for measuring the importance of product characteristics in the eyes of the customer (ACA=Adaptive Conjoint Analysis, OLS=Ordinary Least Squares)

based approach a conjoint study is used for this purpose. For the evaluation of the PC influence on the PAs in step 4, the CA based approach makes again use of conjoint studies. For each PA, members of the QFD team are questioned. The PCs are evaluated on how strongly they contribute to the fulfilment of the respective PA. The determined influences are standardized (similar to step 2) so that the influences of the PC on each PA add up to 1. A similar standardization should also be executed in the traditional approach where weights are determined in QFD team discussions. The calculation of the PC importance in view of the customer is implemented in a fifth step.

3 Sample application 1: Laptops for students

High quality laptops for students were selected as design object in our first sample application. The selected object has the advantage that the customers possess a considerable expert knowledge w.r.t. potential PAs and PCs.

For the execution of the study (especially for the implemented group discussions), a small QFD team was formed including students who were particularly experienced in technical matters and who have been using laptops for a long time. As outlined in section 2, the PAs were determined by the QFD team in a first step. Supporting information was collected and prepared through retailers as well as research in related journals and the internet. The result was a list of seven PAs important to the customers (PA1-PA7): "speed", "multimedia", "display", "transport", "handling", "mobile usage", and "low price". Assumed favorable and unfavorable levels of this PAs (e.g. "low" and "high" "speed") were also determined.

In a similar manner, the PCs were determined. Altogether 20 were selected (PC1-PC20): "processor", "memory", "cache" and so on. For each PC, relevant favorable and unfavorable outcomes were defined. So, e.g., for "memory" "16 MB" and "32 MB" were chosen. Additionally, probable relations between PCs and each PA were assumed to reduce the complexity of data collection. So, e.g., it was assumed that "speed" is only influenced by "processor", "memory", "cache", and "hard drive".

As already indicated, an advantage of the object chosen was the possibility to ignore differences between customers and experts. In the course of the CA based approach the respondents were asked – as experts – to assess both the PA importance and the PC influence on each PA. Altogether 40 respondents were questioned. A total of eight ACA interviews had to be performed by each respondent: Six interviews for the relation between PAs and PCs (ACA1-ACA6), one for the relation between PAs and preferences (ACA7), and one for the relation between PCs and preferences (ACA8). ACA8 could only be performed since the respondents where customers and experts. ACA8 is not needed for step 1 to 5 but for deriving additional data w.r.t. predictive validity by providing "true" evaluations of altogether 32 further PC described laptops w.r.t. a pre-determined orthogonal design. Additionally, purchase intentions were collected for four fictitious laptops, the so-called holdout stimuli.

The computer-supported interview lasted on average 42 minutes per respondent and was judged by most respondents as interesting and diversified. A high internal validity of the individually determined importances and influences was observed in the different data collection phases (not further discussed here for space restrictions).

Fig. 1 summarizes the results of the CA based approach. The mean values and standard deviations of the PA importances as well as the influences of the PCs on each PA are available. E.g., PA1 "speed" shows the highest average importance with 18.7% and is influenced above all from the PCs "memory" (mean value 33.6%), "processor" (31.4%), and "cache" (29.3%). The determined influence of PC4 "hard drive" is on average unimportant. However, the standard deviations indicate differentiated evaluations across respondents. The PC importances in Figure 1 result - as described in section 2 - from the weighted sums of the PC influences on the PAs. The estimation was carried out separately for each respondent so that mean values and standard deviations can be presented in Figure 1 again. One recognizes that above all the PCs price, display type, color and processor have highest importance.

Product characteristics			Product attributes								
			PA1 Speed	PA2 Multimedia	PA3 Display	PA4 Transport	PA5 Handling	PA6 Mobile usage	PA7 Price	Unforable / forable configuration	Importance
PC20	Price	→							1.00	4.000 DM / 2.500 DM	.149 *.059*
PC19	Network interface	○					.345 *.130*			without / With	.047 *.015*
PC18	Pointer	○					.162 *.095*			Trackball / Touchpad	.022 *.007*
PC17	Keyboard	←					.380 *.113*			Small / large keys	.052 *.016*
PC16	Transport bag	○				.194 *.128*	.113 *.082*	.078 *.059*		No / yes	.047 *.012*
PC15	Size	→				.285 *.102*		.136 *.063*		Big / small	.050 *.018*
PC14	Weight	→				.286 *.122*		.109 *.059*		4 kg / 2 kg	.046 *.017*
PC13	Removable battery	○						.206 *.064*		No / yes	.025 *.013*
PC12	Operating time	←						.258 *.053*		1:15 h / 2:30 h	.031 *.016*
PC11	Speakers	○		.100 *.057*						Not integrated / integrated	.011 *.007*
PC10	Soundboard	○		.192 *.055*						without / With	.021 *.013*
PC9	CD-Rom	○		.187 *.071*		.234 *.132*				Not integrated / integrated	.047 *.015*
PC8	Color	○		.227 *.049*	.337 *.077*					Monochrom / Color	.086 *.022*
PC7	Display Size	←			.140 *.079*					10,6" / 12,1"	.026 *.008*
PC6	Resolution	←			.229 *.093*					640x480 / 1024x768	.042 *.012*
PC5	Display type	○		.168 *.064*	.294 *.090*			.211 *.079*		DSTN / TFT	.097 *.019*
PC4	Hard drive	←	.057 *.050*							810 MB / 1310 MB	.010 *.003*
PC3	Cache	←	.293 *.084*							no / 256 kB	.054 *.015*
PC2	Memory	←	.336 *.094*							16 MB / 32 MB	.062 *.018*
PC1	Processor	←	.314 *.091*	.126 *.067*						P100 / P166 MMX	.072 *.018*
	Importance		.187 *.053*	.109 *.067*	.184 *.053*	.117 *.053*	.136 *.043*	.120 *.064*	.149 *.059*		

Average value
standard deviation

Fig. 1. PA and PC importances as well as PC influences on the PAs in the CA based approach

Besides the CA based approach, the traditional approach has also been carried out. The PA importance could be taken from the first phase of the computer-supported examination. The PC influences on each PA were determined by the QFD team through a group discussion. For the assessment of the predictive validity of the two approaches, the four evaluations of the holdout stimuli and the 32 derived evaluations were used (as already mentioned). For each respondent and each group of stimuli, predictions were estimated on basis of the traditional and the CA based approaches and correlated with the observed values per respondent. Table 2 shows that the CA based approach outperforms the traditional approach in each case. Wilcoxon sign-rank-tests and t-Tests indicate significance.

	Predictive validity w.r.t. the stimuli			
	of the holdout task		of the ACA8 task	
Number of respondents with...	traditional approach	CA based approach	traditional approach	CA based approach
... $r \in$[1,00;0,97)	9	10	0	0
... $r \in$[0,97;0,94)	7	11	0	0
... $r \in$[0,94;0,91)	6	3	0	6
... $r \in$[0,91;0,88)	3	6	1	6
... $r \in$[0,88;0,85)	4	1	6	7
... $r \in$[0,85;0,82)	0	1	6	9
... $r \in$[0,82;0,79)	1	2	10	9
... $r \in$[0,79;0,76)	1	1	6	2
... $r \in$[0,76;0,73)	0	1	3	0
... $r \in$[0,73;0,70)	1	2	4	0
... $r \in$[0,70;0,00)	8	2	4	1
Average value regarding r	0,843	0,898	0,789	0,848
p-value in the sign rank test	0,003		<0,001	
t-value (p-value) in the t-test	2,855 (0,007)		6,483 (<0,001)	

Table 2. Predictive validity in the traditional and the CA based approach (r: Pearson's correlation coefficient)

4 Sample application 2: Luxury purses for men

The advantages of the new approach were analyzed empirically in a second application. Men's purses manufactured by a well-known Italian manufacturer have been selected as design object. The sales figures of these products had significantly decreased in the last few years as a result of changing customer preferences and demands as well as waning price-willingness. The manufacturer has therefore decided to adjust his products to the market.

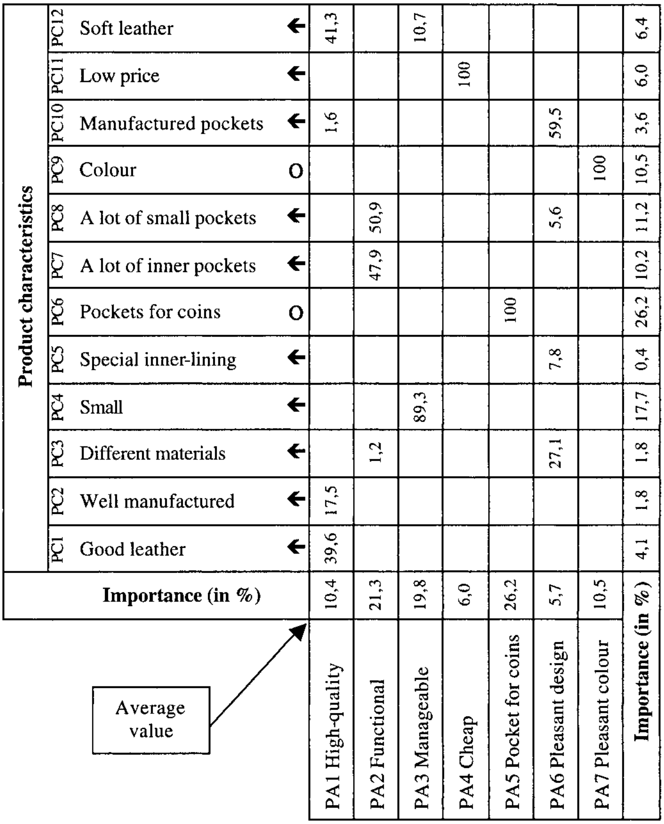
First, a pre-test was carried out. Using one-to-one interviews a sample of target customers were questioned about important PAs for their buying decision. Additionally, a workshop with employees from the company's production and design department was performed. Basing on the pre-test results and the available expert knowledge in the workshop, a list of seven most important PAs was developed (PA1-PA7): “high-quality”, “functional”, “manageable”, “cheap”, “with a pocket for coins”, “pleasant design” and “pleasant color”. Other important attributes like manufacturer’s name and brand name were omitted since they couldn’t be influenced in the design process.

Also in this workshop, for each PA favorable and unfavorable levels from a customer’s point of view were defined. So, e.g., for realizing differences w.r.t. high-quality, three different materials (A, B and C) were selected which in terms of quality, processing and softness could easily be distinguished from the customer. In the case of “functional” as well as “manageable”, differentiations like “weak”, “normal” and “strong” on the basis of the number of small inner pockets as well as whether or not they could fit in a trousers pocket were used. It should be pointed out that for the later customer questioning, precise reference products were available for all characteristics in each case, which helped in the clarification of PA differences. So, e.g., for “cheap” (PA4), exact retail purse prices were used.

In a similar way, the 12 PCs were defined in a workshop with company experts (PC1-PC12): "good leather", "well manufactured", "different materials", "small", "special inner-lining", "pockets for coins", "a lot of inner pockets", "a lot of small pockets", "color", "manufactured pockets", "low price", and "soft leather".

Regarding assumed influences of the PCs on the PAs altogether fifteen experts were asked to assess 20 real products. They rated them in terms of the PAs and in terms of the assumed influencing PCs on a 7 point rating scale with 1 = “PA (PC) doesn’t apply at all.” and 7 = “PA (PC) applies fully and completely.”. Fig. 2 shows the relationship determined from these data by multivariate regression analysis. The rows of the matrix reflect the PC influence on each PA as a percentage. One recognizes, e.g., that “high-quality” is assumed to be influenced above all by the PCs “soft leather” and “good leather”.

For deriving PA importances, a customer survey was carried out in a German outlet of the manufacturer. Altogether, 40 potential buyers participated. First, the goal and content of the survey were explained. Then, 16 stimuli from an orthogonal design w.r.t. the seven PAs and their 2-3 levels were chosen for preference data collection. In order to bring the survey as close to reality as possible, real men’s purses were provided from the manufacturer. These purses could be chosen or supplemented as to correspond to the 16 stimuli. Supplemented means: if a purse didn’t exactly correspond to the affiliated stimulus, the respondent was advised about this orally and was requested to provide a suitable modification for this purse. E.g., “Please assume this purse not in black but in this bordeaux red.” or “Please assume purse 1 with the same leather as purse 8.” A buying intention scale from 1 (“I wouldn’t buy this purse by any means.”) to 10 (“I would quite certainly buy this purse.”) was used. Additionally, for generating holdout data, the



	Product characteristics												
	PC1 Good leather ←	PC2 Well manufactured ←	PC3 Different materials ←	PC4 Small ←	PC5 Special inner-lining ←	PC6 Pockets for coins O	PC7 A lot of inner pockets ←	PC8 A lot of small pockets ←	PC9 Colour O	PC10 Manufactured pockets ←	PC11 Low price ←	PC12 Soft leather ←	Importance (in %)
PA1 High-quality	39,6	17,5								1,6		41,3	10,4
PA2 Functional			1,2				47,9	50,9					21,3
PA3 Manageable				89,3								10,7	19,8
PA4 Cheap											100		6,0
PA5 Pocket for coins						100							26,2
PA6 Pleasant design			27,1		7,8			5,6		59,5			5,7
PA7 Pleasant colour									100				10,5
Importance (in %)	4,1	1,8	1,8	17,7	0,4	26,2	10,2	11,2	10,5	3,6	6,0	6,4	

Average value

Fig. 2. PA and PC importances as well as PC influences on the PAs in the CA based approach

respondents were asked to evaluate further five purses. Altogether the survey lasted, according to the respondents, between seven and ten minutes.

The collected preference data were analyzed using the TRANSREG procedure of the program-package SAS. For each respondent, the individual PA importances were estimated as the difference between the part worthes for the most and less preferred levels. Additionally, using the above derived influences of the PCs on the PAs, the individual PC importances were calculated (see Figure 2). One recognizes that the PAs "with pocket for coins" and "functional" as well as the PCs "pockets for coins" and "small" are most important. Of little importance are, e.g., the PCs "special inner-lining", "well manufactured" and "different materials".

The predictive validity was verified w.r.t. to observed and estimated buying intentions for the four holdout stimuli using Pearson's correlation coefficient (see Table 3). One recognizes an overall high predictive validity. This certainly is due to the fact that the assessment of the respondents was de-

Number of respondents with...	Predictive validity
... $r \in [1{,}00;0{,}97)$	19
... $r \in [0{,}97;0{,}94)$	4
... $r \in [0{,}94;0{,}91)$	8
... $r \in [0{,}91;0{,}88)$	3
... $r \in [0{,}88;0{,}85)$	2
... $r \in [0{,}85;0{,}82)$	2
... $r \in [0{,}82;0{,}79)$	1
... $r \in [0{,}79;0{,}76)$	1
... $r \in [0{,}76;0{,}00)$	0

Table 3. Predictive validity w.r.t. the holdout stimuli (r: Pearson's correlation coefficient)

pendent on real purses, and not on descriptions or pictures. The stimuli were thus relatively simple for the respondents to assess.

5 Conclusion and outlook

The introduced and applied CA based approach for QFD shows a number of advantages in comparison to the traditional approach. PA importances as well as PC influences on PAs are measured "conjoint" resp. simultaneously. Furthermore, the calculated weights are more precise (real valued instead of 0-, 1-, 3-, or 9-values) which resulted – in the sample applications in a higher predictive validity. However, further clarifications of the strengths and weaknesses using more applications are of course necessary.

References

AKAO, Y. (1990): *QFD, Integrating Customer Requirements into Product Design.* Productivity Press, Cambridge, MA.

ASI (1989): *Quality Function Deployment.* American Supplier Institute Inc., Dearborn, Michigan.

ASKIN, R.G. and DAWSON, D. (2000): Maximizing Customer Satisfaction by Optimal Specification of Engineering Characteristics. *IIE Transactions, 32, 9–20.*

BAIER, D. (1998): Conjointanalytische Lösungsansätze zur Parametrisierung des House of Quality. In: VDI-GSP (Ed.): *QFD – Produkte und Dienstleistungen marktgerecht gestalten.* VDI-Verlag, Düsseldorf.

BAIER, D. (1999): Methoden der Conjointanalyse in der Marktforschungs- und Marketingpraxis. In: W. Gaul and M. Schader (Eds.): *Mathematische Methoden der Wirtschaftswissenschaften.* Physica, Heidelberg, 197–206.

BAIER, D. and GAUL, W. (1999): Optimal Product Positioning Based on Paired Comparison Data. *Journal of Econometrics, 89, 365–392.*

BAIER, D. and GAUL, W. (2003): Market Simulation Using a Probabilistic Ideal Vector Model for Conjoint Data. In: A. Gustafsson, A. Herrmann, and F. Huber (Eds.): *Conjoint Measurement - Methods and Applications.* 3rd ed., Springer, Berlin, 97–120.

BRUSCH, M., BAIER, D., and TREPPA, A. (2002): Conjoint Analysis and Stimulus Presentation: a Comparison of Alternative Methods. In: K. Jajuga, A. Sokolowski, and H.-H. Bock (Eds.): *Classification, Clustering, and Analysis.* Springer, Berlin, 203–210.

CHAN, L.-K. and WU, M.-L. (2002): Quality Function Deployment: A Literature Review. *European Journal of Operational Research, 143, 463–497.*

CRISTIANO, J.J., LIKER, J.K., and WHITE, C.C. (2000): Customer-Driven Product Development Through Quality Function Deployment in the U.S. and Japan. *Journal of Product Innovation Management, 17, 286–308.*

GAUL, W., AUST, E., and BAIER, D. (1995): Gewinnorientierte Produktliniengestaltung unter Berücksichtigung des Kundennutzens. *Zeitschrift für Betriebswirtschaft, 65, 835–854.*

GREEN, P.E. and SRINIVASAN, V. (1978): Conjoint Analysis in Consumer Research: Issues and Outlook. *Journal of Consumer Research, 5, 103–123.*

GREEN, P.E. and SRINIVASAN, V. (1990): Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice. *Journal of Marketing, 54, 3–19.*

GUSTAFSSON, A. (1996): *Customer Focused Product Development by Conjoint Analysis and Quality Function Deployment.* Linköping University Press, Linköping.

HAUSER, J.R. and SIMMIE, P. (1981): Profit Maximizing Perceptual Positions: An Integrated Theory For The Selection of Product Features And Price. *Management Science, 27, 33–56.*

JOHNSON, R.M. (1987): Adaptive Conjoint Analysis. In: *Proceedings of the Sawtooth Software on Perceptual Mapping.* Ketchum, ID, 253–265.

KATZ, G.M. (2004): A Response to Pullman et al.'s (2002) Comparison of Quality Function Deployment versus Conjoint Analysis. *Journal of Product Innovation Management, 21, 61–63.*

KING, B. (1989): *Better Designs in Half the Time: Implementing QFD Quality Function Deployment in America.* 3rd ed., GOAL/QPC, Methuen, Mass.

MIZUNO, S. and AKAO, Y. (1994): *QFD The Customer-Driven Approach to Quality Planning and Deployment.* Asian Productivity Center, Tokyo.

PULLMAN, M.E., MOORE, W.L., and WARDELL, D.G. (2002): A Comparison of Quality Function Deployment and Conjoint Analysis in New Product Design. *Journal of Product Innovation Management, 19, 354–364.*

SAWTOOTH SOFTWARE (2002): *ACA System Adaptive Conjoint Analysis Version 5.0.* Sawtooth Software Inc., Evanston, IL.

SULLIVAN, L.P. (1986): Quality Function Deployment. *Quality Progress, June, 39–50.*

URBAN, G.L. and HAUSER, J.-R. (1993): *Design and Marketing of New Products.* Prentice Hall, Englewood Cliffs, NJ.

YODER, B. and MASON, D. (1995): Evaluating QFD Relationships Through the Use of Regression Analsis. In: *Proceedings of the Seventh Symposium on Quality Function Deployment, ASI&GOAL/QPC.* American Supplier Institute, Livonia, MI, 239–249.

Financial Management in an International Company: An OR-Based Approach for a Logistics Service Provider

Ingo Böckenholt and Herbert Geys

Dachser GmbH & Co. KG, Memminger Str. 140, D-87439 Kempten, Germany

Abstract. The article gives an overview over financial management in the family business group Dachser, a global player in Europe-logistics. It is discussed how a pragmatic approach in strategic investment planning helps to maintain and develop one of the most significant logistics networks in Europe which is referred to as the branch's top of the form and is cited as benchmark in branch analyses.

1 Introduction

Dachser as a a family-owned company with a 75 years old history today belongs to the global players in Europe-logistics. The business model consists of three pillars: Land Transport Europe (share of turnover 70%), Air & Sea Freight (15%) as well as Food Logistics (15%). The company's backbone is a dense network of several branches and partners all over Europe. All branches and partners are connected with each other by regular services via which the costumers are daily served with purchasing and distribution. Furthermore, contract logistics offers warehousing and value-added services. Dachser contributes to its costumers' appreciation by redesigning logistic procedures. This results in an improved cost and performance position of their costumers. Dachser's mission is to improve their costumers' logistics balance sheet. Dachser's logistic solutions are characterised by an integration of all sub-processes reaching the whole process chain. Dachser's driving force is the homogeneous network of commodity flows, information, transport operators and people.

2 Logistics needs a network

Being a system service provider Dachser demands to create an integrated network and to operate whilst permanently considering improvement. In addition, specific solutions for the costumers' individual requirements are planned and realised within the scope of contract logistics. Dachser has been pursuing a consequent strategy of internationalisation for years without overstreching the process mastery. Acquisitions have to remain controllable by the organisation. A cross-border process standardisation is completed by an informational chain anticipating the physical commodity flow. The expansion of the integrated Europe-network is marked by the construction of a goods distribution centre in the "Saarland" (one of Germany's federal states in the west of the country) for the whole of Europe and the introduction of the product family "entargo". "entargo" represents a Europe-wide logistics solution defined at a high standard and according to uniform guidelines, whose physical requirements of assured transit time are provided for by a Europe-wide dense network.

It is not so long since analysts mainly preferred only those companies listed which invested little capital in their network structure. The investors' attitude has been analogous. Dachser had to stand the question asked by banking circles why we invest about 100 million euros each year in the construction of our branches. It would make more sense to buy services from transport networks.

Today this point of view changed radically. The market showed that there is no getting around with an own infrastructure. Almost unnoticed by the public Dachser has created one of the most significant logistics networks in Europe. Today Dachser is referred to as the branch's top of the form and is cited as benchmark in branch analyses.

3 A challenge for strategic investment planning

Persons form companies. People become entrepreneurs and undertake projects on their own responsibility to create values for costumers. In order to judge and to lead these companies, organisations need strategies, targets and guidance systems so that the respective undertaking does not end up in an instrument flying.

Above all, a concrete idea of the future is essential. Dachser's vision is to be the world's best provider of Europe-logistics. The basic principle of a successful company is, beside this vision of the future and the idea for a product, a sustainable management and control structure, a corporate culture shared by all employees as well as the professional management of financial resources.

The existing network represents an essential part of the group's enterprise-value. That is why the planning of this network is of central importance which is reflected in the permanent modernisation of single locations and the creation of additional capacities.

Investment in the network is planned in the long-term over a period of 10 years. At the same time a simulation model based on realistic return and growth assumptions and standards of aimed at balance sheet ratios shows scope for additional growth and investment. By a sensitivity analysis the management in the managing board as well as in the board of directors is given the necessary certainty.

The 10-year planning, currently extended to 2012, includes an investment volume of about 1 billion euros. Within the scope of planning premises there are still considerable reserves which can be used for further development of the European network and for additional growth.

4 The planning model

Being a family-owned company that has the aim to remain autonomous and self-dependent within the next 10 years, which is postulated by the proprietors, Dachser has chosen a simple, pragmatic approach for demonstrating their strategic investment planning (Süchting (1995)). This approach dispenses with a simultaneous improvement of investment and financial planning (Hanssmann (1990)). Equity capital guarantees independence. A too high equity ratio does not make sense for business management reasons. A too low equity ratio puts independence at risk. That is why the proprietors defined a lower limit for the equity ratio. As German Commercial Law does not comprise leasing in the balance sheet but as this does essentially bother the company's risk by long-term commitments, the leasing volume is included in the determination of the relevant equity ratio (so-called "leasing-equity-ratio"). If at the same time access to "outside" equity capital is refused a priori, the compliance with a lower limit for equity capital does in a positive way automatically restrict growth.

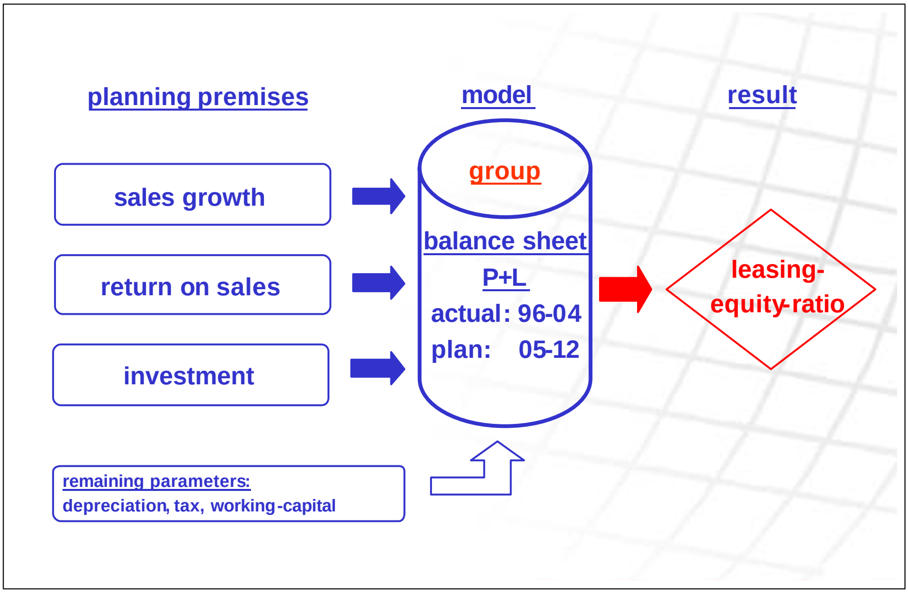


Fig. 1. Strategic financial planning at Dachser

Planning horizon:

a) The planning model comprises a planning horizon of 10 years. Every 5 years the planning is extended by another 5 years. For the first time the planning 1997 was set up for the period from 1998 to 2007. In the meantime the planning has been extended to 2012.
b) The **Target value** of the planning model is the long-term maintenance of a fixed equity ratio, defined as so-called "leasing-equity ratio". A short-term forecasting accuracy is not necessary. Short and medium-term financial management is improved by a separate control mechanism.

Control parameter:

a) **Sales growth:** Based on the values of the past and currently defined goals and objectives a sales growth is predetermined for the planning years which mainly considers organic growth. Growth by acquisitions is taken into consideration ad hoc.
b) **Return on sales:** Here, in a first step, a target return based on experience is determined as well.
c) **Investment:** Most of the annual investment volume flows into the development of the physical network. Based on differentiated capacity plannings and target-volume growth, expansion schemes and new buildings can be defined per location with a relatively high accuracy as far as size and time is concerned. Additional investment for the expansion of the locational network and warehouse capacities scheduled within the scope of contract logistics are taken into consideration as well as planned acquisitions. For investment in furniture and office equipment - according to amount of secondary importance - growth-orientated standard values are defined.

Within the scope of the sensitivity analysis the classic scenarios best, medium, worst are checked for all three control parameters according to their effect on the leasing-equity ratio. In a further step, the investment volume is raised to such an extent that the lower limit for the leasing-equity ratio, stipulated by the proprietors, is achieved at the end of the planning horizon. That shows the maximum scope possible for investment depending on the assumed values for sales growth and sales return and serves as basis for the coordination of the long-term strategy between the board of management and the board of directors.

Operative parameters

a) **Depreciation ratio:** Annual depreciation is defined in % of the opening capital of fixed assets. Even though, as far as accounting is concerned, depreciation refers to the cost of acquisition or production of the single assets and liabilities and even if a different useful economic life has to be rated for the single assets and liabilities, the simplification included in this model averages in a correct result. This applies as long as no structural modifications of fixed assets occur. However, in order to consider that as well, the model "learns" from the development of this ratio in the past.

b) **Tax- and withdrawal-ratio:** The income-tax burden of a concern results from a mixture of different national legislations and different legal forms (partnerships and private limited companies). The parent company (at the same time largest part of the group) is a partnership. The proprietors' income tax directly charges the equity capital of the company. In addition, the proprietors' withdrawals have to be considered. Combined as one, this is taken into consideration by a tax- and withdrawal-ratio which is permanently adjusted to current fiscal development.

c) **Growth working capital:** For the development of working capital mainly the terms of payment on the debit and credit side are taken into consideration.

5 Success creates future

Independently from all planning models profit is and remains the decisive measure for success. Profit is essential to finance growth and investment in the long-term if an extensive independence from the capital market is aimed at. Only profit creates future. Dachser never made the mistake, as recently was practised to some extent in the sign of "new economy", to declare, as far as performance evaluation is concerned, a profit as such if no outside interests (EBIT) or even depreciation (EBITDA) was earned. The return for the equity capital provided by the proprietors has to be rated as well considerably higher than that of risk-free assets. Only then the management has been successful, only then the creation of additional enterprise-value begins.

References

HANSSMANN, Friedrich (1990): *Quantitative Betriebswirtschaftslehre.* Oldenbourg, München, Wien.

SÜCHTING, Joachim (1995): *Finanzmanagement - Theorie und Politik der Unternehmensfinanzierung.* Gabler, Wiesbaden.

Development of a Long-Term Strategy for the Moscow Urban Transport System

Martin Both

Dornier Consulting, D-88003 Friedrichshafen, Germany

Abstract. This paper reports on a study for the development of a long-term strategy for the transport system in the City of Moscow. A traffic simulation model was developed in order to integrate all modes of transport and to test alternative strategies and economic development scenarios with respect to selected key performance indicators. The model was used to support and rationalize the decision process of the various stakeholders involved. A "balanced" strategy was recommended and accepted by the decision makers as a long-term framework for implementation.

1 Background and introduction

The Moscow City Transport Strategy Study was commissioned by the United Nations Development Programme (UNDP) and the Moscow City Government (MCG) with a view to preparing a viable 20 year transport strategy for Moscow that:

- meets the challenge of increasing car ownership in that period;
- supports the development objectives of the city;
- respects the financial constraints of the city;
- is genuinely implementable.

The **Long Term Strategy** (LTS) for Moscow City Government was developed after the exhaustive analysis of transport policy options and testing using a specific traffic simulation model. The LTS is designed to provide for the mobility and transportation requirements of Moscow residents and businesses up to 2020 and uses the best predictions of economic growth, population, employment and land use available as the basis for plan development.

Moscow transport context

The City of Moscow covers an area of 1,091 km and had a resident population (in 2000) of 8.64 million people. The Moscow oblast is an administrative division of the Russian Federation, outside of Moscow city. The oblast covers a much larger area than the city of Moscow itself. In 2000, the population of the oblast, outside Moscow city, was some 6.3 million. The oblast is governed by the Moscow Regional Administration, with its own Duma.

Moscow City and its transport system are at an intermediate development stage, which is typical of many large cities during a period of rapid motorisation (Moscow Government (1998)). The most distinctive features of Moscow are the high density of population and high (about 80%) current level of use of public transport for journeys around the city. The capacity of much of the road network has already been reached, particularly within central areas. Substantial further growth of motorisation is expected to 2020. Without any effective control measures this will lead to traffic paralysis.

The Moscow city transport system comprises **4 major networks**, i.e.:

- Roads,
- Metro,
- Surface public transport (SPT), including buses, trolley-buses and trams,
- Surface rail.

The extent of the **road network** is estimated at 4600 km, carrying approximately 1.2 million passenger trips per hour.

Buses, trolley-buses, trams and minibuses all operate in Moscow. There are two main types of operator, the municipally owned Mosgortrans, which operates buses, trolley-buses and trams, and various private companies, who mostly operate minibuses. Some of the private companies also operate taxis. Mosgortrans currently has a fleet of about 5,300 vehicles. The major private operator, Autoline, has a fleet of about 1,500 minibuses.

The main characteristics of the **metro system** in 2000 were:

- 11 lines operated,
- 262 track kilometres,
- 160 stations,
- 9000 trains per day,
- 31,300 employees,
- 3.24 billion passengers carried.

The annual **mobility rate** per resident in Moscow amounted to 1,140 trips in 1990 and to 1,200 trips in 2000. For the year 2020 a trip rate of 1,300 to 1,500 is forecast. Broken down to an average day, the mobility was 3.1 trips per resident in 1990 and 3.3 in 2000. This includes walking, which has, however, to be considered as of minor importance due to the size of the city and the spatial separation of residential and working areas. The 2020 forecast corresponds to a range from 3.6 to 4.1 trips per resident per day.

Figure 1 overleaf shows that the core business district of Moscow experiences a hugh daily influx due to the concentration of work places and other traffic attractors.

2 Traffic simulation model

A peak hour computerised traffic simulation model was developed in cooperation with the Institute of System Anaylsis in Moscow to assist in the evaluation of long term strategies (Dornier Consulting, GIBB (2001)). This

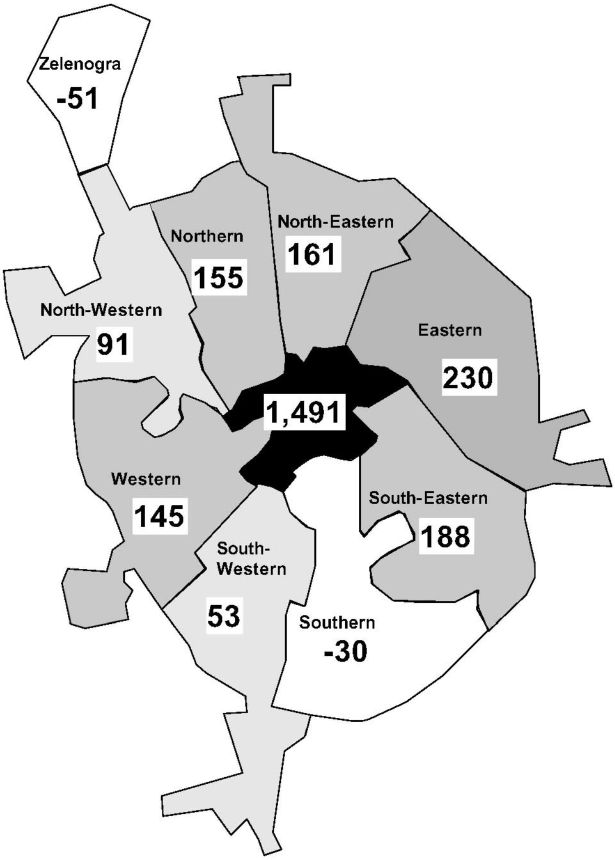


Fig. 1. Daily influx into the City of Moscow (in thousand people)

model contains the necessary equations to accurately simulate future traffic conditions. In January 2000 all required data (the road, rail and metro networks) was input in this model. The model has a sufficiently large number of traffic zones to distribute trips. Given the convenient interface to input data, the model could be developed, calibrated and successfully used to test transport options and still meet the study schedule.

Model calibration

The data on traffic generation and attraction received from Moscow City Government was first checked for general plausibility. A precondition for the use of the model is that traffic generation (traffic which has its origin in a specific zone) and traffic attraction (traffic, which has its destination in a specific zone) are balanced. Minor alterations were made to ensure that this was the case.

In total, eight calibration model runs were necessary in order to develop a network with reasonable and logical characteristics. Table 1 summarises the output from the final calibration run.

Indicator	Unit	Total	Share (%)
Road traffic turnover	**Veh km**	**4,975,000**	**100**
- of which are private cars	Veh km	4,692,000	94.3
- of which are buses	Veh km	283,000	5.7
Public transport turnover	**Pass km**	**27,533,000**	**100**
- of which are SPT	Pass km	5,662,000	20.6
- of which are metro	Pass km	15,575,000	56.6
- of which are rail trains	Pass km	6,296,000	22.9
Travel times (total)	**Pass hr**	**1,250,000**	**100**
- in private car	Pass hr	319,000	25.5
- in public transport	**Pass hr**	**931,000**	**74.5**
- of which by SPT	Pass hr	402,000	32.2
- of which by metro	Pass hr	371,000	29.7
- of which by rail trains	Pass hr	157,000	12.6

Table 1. Output from final model calibration run - AM peak hour

This final traffic model iteration confirmed that calibration had been successfully completed. This traffic assignment is the 2000 base network from which all new networks and assignments have been derived.

3 Strategy testing

The steps undertaken for strategy testing were as follows:

- description of the traffic and infrastructure situation to be expected in 2020. Only projects which are committed and likely to be realised were included in the "reference" case;
- definition of general policy strategies to be followed in the future;
- definition of policy measures and projects which are necessary to realise the strategies;
- prediction of the traffic volumes to be expected in 2020;
- estimation of effects on key performance indicators (traffic performance, congestion, traffic safety, environmental effects);
- comparison of effects and evaluation relative to the "reference" case;
- recommendation of a strategy most suitable and effective for the further development of the City of Moscow.

Two different **economic development scenarios** were tested. These scenarios are expressed in the model as different rates of growth in motorisation over the next 20 year period.

Economic **Scenario A** implies conservative economic development. Annual growth rates will be low, which will cause only a moderate increase in motorisation. The annual growth rate used in the model is 2.5%.

Economic **Scenario B** envisages a faster growing economy with increasing incomes of the population. Car ownership rates could follow from the trend observed in Moscow in the early 1990's. This would cause a dramatic increase in traffic volumes which could become an important obstacle to the development of Moscow. The annual growth rate of motorisation assumed in the model is 4.5%.

Four main **transport system strategy options** were tested as follows:

1. A baseline option, which implements transport and highway projects already approved by MCG. This option was incorporated in all other strategies tested.
2. A highway based strategy which emphasises major and secondary highway construction and with several radial toll roads.
3. A public transport based strategy which adds new metro lines and LRT routes and provides for bus priority on major roads.
4. A demand management strategy which uses traffic restraint mechanisms to control congestion in the core of the city.

This approach is standard for such studies in western European cities. The baseline strategy includes transport elements already built, being under construction and approved for construction. A programme of transport projects was developed for each of the options. These were based on international experience of schemes likely to give most benefit to each scenario and from local recommendation.

These projects were then added to the existing road and public transport networks and tested using the traffic simulation model. The model was used to forecast transport conditions in 2020 using two vehicles ownership assumptions. The "low" car ownership assumption (scenario A) gave good results but was rejected as being too restrictive. The "high" mobility assumption (scenario B) is consistent with information gathered from other urban areas in Europe today. This predicts that 80% of families in the future will own a vehicle by 2020. This assumption was used in the traffic forecasts to compare strategy alternatives.

4 Results and recommendations

For various reasons, each of these pure strategies was deficient in some way when compared to the agreed five primary objectives of:

- reducing congestion;
- assisting Master plan implementation;
- reducing air pollution;

- improving public transport;
- reducing subsidies.

This led to the development of a compromise strategy known as the "balanced" option. A roads based strategy promotes greater car usage and does not eliminate congestion. A public transport strategy builds up metro and SPT traffic but congestion is still unacceptable. A demand management strategy solves the congestion problem in the dense core, but at a high cost to car drivers in the outer areas of the city. The "balanced" strategy is a synthesis of these three options, which in overall congestion cost and emission reduction terms is better than the other three options. The results of the traffic model tests for 2020, are shown in Table 2.

Performance Index	**Baseline (S1)**	**Roads based (S2)**	**Public trans-port (S3)**	**Demand mana-gement (S4)**	**"Balan-ced" (S5)**	**Exis-ting**
Cars/taxi usage (a) (1.4 person/ (vehicle)	9.9	10.8	9.5	8.9	8.8	6.6
Metro usage (a)	13.5	12.5	14.3	12.8	14.5	15.6
SPT usage (a)	4.2	4.6	4.1	6.8	5.5	5.6
Rail usage (a)	5.4	5.0	5.1	4.7	4.2	6.3
Vehicle speed (km/h)	19.2	21.8	20.3	23.5	23.6	25.0
SPT speed (km/h)	13.6	16.9	14.8	17.0	18.0	14.1
Overloaded roads (%)	75	65	72	61	60	56
Car congestion cost ($m per year)	7,874	8,892	7,678	7,562	7,498	na
Pollutants (tonnes/day)	816	815	762	649	< 625	886
Mode split (% public)	66	63	68	71	72	80
Passenger time (million hours/ year)	1.39	1.33	1.31	1.29	1.24	1.25

Notes: (a) million passenger km in the peak hour
(b) na = not applicable
(c) all results are based on economic scenario B.

Table 2. Comparison of strategies

In general, public transport usage is highest and car usage is lowest in those strategies catering for public transport, and vice-versa in road-based strategies. This relationship is not always clearcut because of, for example, the positive effects of demand management on SPT as well as cars.

The "balanced" strategy option is superior in 8 of the 11 categories used to compare the strategic options tests, including environmental policies. While this does not prove that the "balanced" option is the optimal strategy for Moscow, it does show that strategically it is better than options of building more roads or even building more metro and railway lines.

Referring back to Table 2, the "balanced" strategy is expected to reduce traffic on congested links (percentage of overloaded roads in Table 2) by 15% relative to the baseline strategy. This is a significant impact in the context of an estimated growth of 46% in car ownership, giving 80% of the population access to a car in the year 2020. These figures relate the overall conditions in Moscow. Within the Garden Ring, the situation is rather different. Congestion, and hence average speeds, are already lower than 19.2 km/h (the average speed throughout Moscow in the baseline strategy in 2020). In all strategies the serious congestion already experienced within the Garden Ring will rapidly deteriorate without some form of restraint.

Overall congestion is measured by average traffic speeds. Referring again to Table 2 and without any new transport measures, road speeds will decline from an estimated average speed in the peak hour of 25.0 km/h today to 19.2 km/h by 2020. In the "balanced" strategy the average speed is 23.6 km/h, a significant improvement on the baseline.

Restricting the use of vehicles in the central area, inside the Garden Ring, can take many forms. The recommended approach is a combined policy of:

- high capacity main arterial routes;
- area traffic management and junction control;
- calming in residential areas by restricting through movements;
- charging for use of the roads during peak travel periods.

Charging for the use of the roads would provide revenues for funding public transport and other traffic management projects. With these traffic management measures, average speeds on and within the Garden Ring are predicted to increase relative to current conditions and would be considerably better than was the case for the other road or public transport scenarios tested.

With respect to public transport usage and speeds, the "balanced" strategy is the most effective in maintaining the usage of the metro and SPT modes, especially for trips to the city centre. The mode split is estimated at 72% using public transport. It is less successful at increasing suburban rail usage due to the competition of the private car once vehicle ownership increases. SPT's share of usage would decline by 2% but, relative to the base case, the "balanced" strategy is predicted to generate a 34% increase in patronage. Metro's share of traffic is expected to decline by 7%, but relative to the base case, absolute patronage would increase by 7%. Surface modes would be able to operate at higher speeds due to lower congestion levels. A 4 km/h increase in peak hour average bus and trolley speeds to 18 km/h is predicted under the "balanced" strategy.

Finally, the "balanced" strategy gives scope for increasing public transport patronage and thus should permit the metro, rail and surface modes to collect the revenues that they are supposed to receive. The reduction of subsidies, while an important objective, is complicated by the high proportion of users who are eligible for concessionary and "free" fares. Notwithstanding the problems of enforcement, both rail and metro modes are now collecting a greater proportion of revenues by introducing better ticketing systems. This experience needs to be transferred to the SPT modes where less than 20% of the users pay for their tickets; in the early 1990's over 50% of users paid a fare.

Realising the objectives of the "balanced" transport strategy involves compromises between modes, between individuals and between low and high density areas of the city. A balance has to be struck between a number of transport and development policy strands:

- greater car usage outside the most congested areas, offset by public transport improvements to increase accessibility to the metro and SPT networks;
- regeneration stimulated by a focussed programme of road investments with complementary parking and public transport programmes;
- restraints on the private car in the congested city centre where it is not an efficient mode of transport. Access will have to be restrained in the future, as public transport services are improved to handle the trips diverted.

The analysis and recommendations from the study were reviewed by a panel of experts nominated by the Moscow City Government. As a result of the review process the recommended "balanced" strategy was adopted by the Moscow City Government as the long-term framework for the development of the urban transport system. On this basis a short-term implementation program was developed in another stage of the overall study and an international investors conference was conducted in order to attract private investors and operators.

References

DORNIER CONSULTING, GIBB (2001): Moscow City Transport Strategy - Long/Short Term Investment and Implementation Programme, Final Report.

MOSCOW GOVERNMENT - COMMITTEE FOR ARCHITECTURE AND URBAN PLANNING (1998): Moscow City Transport Strategy - Short/Medium Term Investment and Action Programme, Internal Report.

The Importance of E-Commerce in China and Russia – An Empirical Comparison

Reinhold Decker, Antonia Hermelbracht, and Frank Kroll

Department of Business Administration and Economics,
Bielefeld University, P.O. Box 100131, D-33615 Bielefeld, Germany

Abstract. This paper presents selected results of a comparative study focussing on the political, social and economic conditions, as well as the current status quo and future prospects of e-commerce in China and Russia. The empirical findings are based on data collected by two independent online-surveys covering Chinese and Russian companies from different branches as well as management consultancies engaged in these countries. The surveys provide information about the current importance of e-commerce in both countries and work out both potentials and challenges for market development.

1 Introduction

In the last decade the increasing relevance of e-commerce has initiated numerous empirical studies referring to industrial countries such as Germany and USA. In contrast to this, empirical studies that have thoroughly investigated the status quo and future prospects of e-commerce in emerging markets like China and Russia are still rare. The existing ones are mostly restricted to partial aspects of e-commerce (cf. Hawk (2004)). But in particular these two markets are highly interesting from an economical point of view, since both the evolution of China towards a future economic power and the relevance of Russia for the East-European market expansion are undisputable. The potential future demand in these markets is a great attractor for foreign companies. Due to their large geographical size the application of e-commerce solutions suggests itself as an alternative to conventional market development techniques. First results of the Russian study have been discussed in Decker et al. (2003) and have shown the necessity of cross national comparisons to enable generalizations of the empirical findings. Doing so we can detect several parallels as well as interesting differences between both countries that suggest specific e-commerce strategies.

A review of the relevant empirical literature reveals the heterogeneity of the foci of the respective studies. A short paper by the BDI (Bundesverband der Deutschen Industrie e.V.), e.g., deals with strategic and economical hindrances of German companies when engaging in the Chinese and East-European area (cf. BDI (2004)). The general conditions of e-commerce in Asia, and in particular in China, are the subject of discussions in various up-to-date studies. Haley (2002), for instance, considered the economic infrastructure in more detail, whereas Wong et al. (2004) analyzed the factors that have a significant impact on the development of e-commerce in China. The goal of a paper by Tan and Ouyang (2004) was to investigate the diffusion and impacts of internet and e-commerce in China. A comparative study on the growth of e-commerce between China and Canada has been published by Xiao (2004). In contrast to China, the number of freely available e-commerce studies for Russia is still very small. The contributions of Perminov (2000) and Fey and Doern (2002) can be mentioned here exemplarily. Current empirical studies comparing e-commerce in China and Russia do not yet exist to our knowledge.

2 General conditions of e-commerce in China and Russia

China and Russia show some common characteristics that are relevant for e-commerce. Both countries, for example, rank among the geographically largest countries in the world with huge populations including different ethnic groups. In both countries the average income is low, which has a negative influence on the success of web-based B2C activities. But it can be assumed that this situation will change in the near future, since it is to be expected that the average income will increase, due to the dynamic developments in different economic fields.

An essential prerequisite of e-commerce is a well-functioning telecommunication infrastructure (cf. Hollensen (2004)). This, together with a sufficient availability of personal computers, enables access to the internet for the broad population. Unfortunately, according to the existing studies, neither of these prerequisites is given to a satisfactory degree in Russia nor in China at the moment. But both countries show large growth rates regarding the availability of private computers and the use of the internet for privat and commercial purposes. Special programs like "Electronic Russia 2002-2010", which supports the computerization of companies and private households, and the Chinese "1110-Project", which focuses on the security of IT-systems, are supporting and promoting the observable positive trends. However, in both countries a strong regional discrepancy concerning the availability and use of the internet exists. In Russia the internet is used mainly in the large cities and in the western part of the country. An east-west slope can also be observed in China. Here we have a concentration of internet usage in the eastern areas. Another crucial prerequisite for the success of e-commerce activities is the existence of a comprehensive legal framework. In both countries the lack of jurisdiction for e-commerce was a major handicap for a long time. But recently both governments have ratified various new laws to establish such a legal framework.

While in Russia an operative logistics system is available in most of the relevant regions, the same only partially applies to China. In view of the

Olympic Games in 2008 substantial improvements for the existing logistics system have been decided by the Chinese government.

3 Empirical study and comparisons

3.1 Description of the data

In the following we are going to outline selected results of two recent internet-based surveys comprising Russian and Chinese companies as well as Russian and Chinese subsidiaries of German companies. Moreover, the assessments of numerous management consultants (called "experts" in the following) can be used. For Russia altogether 93 completed questionnaires are available. 48 have been filled in by Russian companies, 20 by Russian subsidiaries of German companies and 25 by Russian experts. The sample includes commercial enterprises (44%), industrial enterprises (36%), service enterprises (19%) and others (1%).

In the same way a total of 91 questionnaires have been collected in China. 37 respondents belong to Chinese companies, 30 to Chinese subsidiaries of German companies, and 19 answers came from Chinese experts. 5 respondents did not provide the respective information. In contrast to the Russian study the distribution of the covered branches is as follows: 6% of the respondents belong to commercial enterprises, 60% to industrial enterprises, 25% to service enterprises. For the remaining respondents the respective answer is not available. Due to the heterogeneous structure of both samples we carried out a chi-square test of homogeneity to check whether the answers of the companies and the experts can be pooled for each country. The assumption could not be rejected at level $\alpha = 0.05$.

3.2 Selected results

Starting from a review of the current pertinent literature we formulated hypotheses and theses to be examined by the data at hand. The first one focuses on the general interest of the companies in e-commerce and the extent of their individual involvement. To deal with this point we asked questions on the status quo as well as the intended implementation of different e-commerce applications. The hypothesis to be examined reads: "The companies in China and Russia are increasingly involved in e-commerce." Figure 1 reflects the relevance of different internet activities and e-commerce applications for industrial, commercial, and service enterprises.

Both in China and Russia the internet is already intensively used for information search and online marketing activities (with a special focus on online advertising). Furthermore, most of the companies have their own internet homepage. However differences occur with respect to homepages including an order function, which seem to be more relevant for Russian companies. In contrast to this the use of the intranet is much more popular in China.

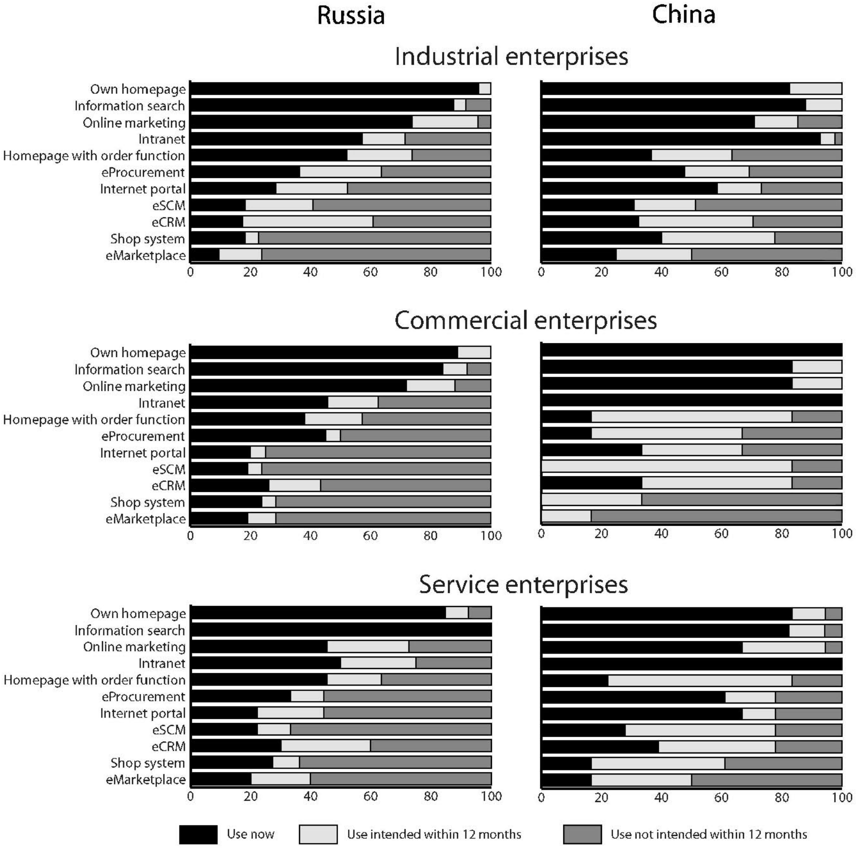


Fig. 1. Internet applications in China and Russia

By means of a chi-square test we checked which of the eleven applications can be assumed to increase significantly within the twelve months following the date of data collection. In the case of Russia such an increase can only be stated for the use of eCRM (electronic customer relationship management). In China the items "homepage with order function" , "shop system", "eCRM", "eSCM" (electronic supply chain management) and "eMarketplace" are promising a significant change in the near future. Obviously both countries are increasingly involved in e-commerce activities but the developments in China show greater dynamics.

A further block in our questionnaire refers to the effects of e-commerce and internet as whole on the company's business processes and, as a consequence thereof, on the creation of lasting competitive advantages. To deal with this aspect we investigated, among other things, the impact of e-commerce activities on a company's value chain and the impact of the internet on a company's general popularity. The following hypotheses reflect the question of interest:

"By means of e-commerce the companies can enhance the efficiency of their business processes and thus increase their competitiveness." and "Including the internet in business processes increases the national and international level of awareness and thus generates competitive advantages."

Both countries are characterized by an extensive and largely identical perception of the presumptive competitive advantages. After all 69% of the Russian and 95% of the Chinese companies agree that e-commerce induces an increase in efficiency regarding the internal processes, 61% and 92% respectively agree about the potential increase in personnel productivity. E-commerce solutions are considered to particularly affect the company's popularity and image as well as the sustainability of customer orientation and the acquisition of new customers. While the Russian companies notice that e-commerce has a more powerful effect on the acquisition of new domestic customers (70% versus 49% regarding foreign customers), Chinese companies assess this effect almost equally to both markets (home market: 78%, foreign countries: 83%). The experts largely agree with this assessment. Concerning the realization of competitive advantages with respect to cost reduction and time saving the experts tend to show a higher level of approval. Altogether there is seemingly a greater enthusiasm for e-commerce in China. Both the Chinese companies and the experts show a more positive estimation of the economic advantages resulting from e-commerce then the Russians.

To enable a deeper understanding of the given answers regarding the managerial and economic effects of the internet usage and e-commerce activities, an explorative factor analysis with a subsequent varimax rotation was carried out. The items included in the analysis could be put down to the latent factors depicted in Table 1 and 2.

In both samples the first factor can be labelled "growth orientation" and explains about 35% of the total variance. In the Russian sample the factor loads highly on the "increase of popularity on the market" and the "acquisition of new domestic costumers". In contrast to this the first factor of the Chinese sample shows the highest loadings on the "reduction of business process time" and the "acquisition of new foreign costumers". This example alone demonstrates the differences in the country-specific strategies. Russia concentrates more strongly on the domestic market while China puts the main focus on international markets. The distribution of the economic sectors in the sample might be one explanation for this pattern. While in the Russian sample the respondents stem primarily from commercial enterprises, those of the Chinese sample are predominantly members of industrial enterprises. This, by the way, hints at the actual availability of the concerning internet technologies, since the questionnaire was made accessible via internet. The second factor can be labelled "efficiency" in both samples. It explains 12% of the total variance in Russia and 32% in China. Thus efficiency has a higher value in China. While in China an increase of the efficiency in the business processes is supposed to be achievable, among other things, by optimizing

Characteristics	Factor loadings
Factor 1: Growth orientation	
Increase in popularity on the market	0.72
Acquisition of new domestic costumers	0.72
Better communication with costumers/suppliers	0.61
Open-mindedness regarding technology	0.58
Increase of sales	0.57
Factor 2: Efficiency	
Reduction of business process time	0.80
Increase in efficiency of the internal processes	0.73
Open-mindedness regarding technology	0.52
Factor 3: Continuance	
Acquisition of new foreign costumers	0.78
Standardization of business processes	0.77
Intensification of costumer orientation	0.61
Better communication with costumers/suppliers	0.46
Factor 4: Cost savings	
Reduction of procurement costs	0.87
Reduction of the stock	0.75
Simplification of the payment modalities	0.65
Factor 5: Additional benefits	
Availability of an additional supply channel	0.86
Temporal competitive advantages	0.58

Table 1. Factor analysis of the Russian sample

the procurement processes, this is attempted in Russia primarily by means of reducing the business process time. The remaining three, respectively two factors can be interpreted analogously and explain 24% or 33% of the total variance.

Considering the different factors and the respective loadings we can conclude that e-commerce is differently used and assessed in Russia and China. While it seems to become an essential component of the production process in China, it is primarily used as an additional but promising marketing channel in Russia. Both developments are plausible against the available sample distribution. In both countries e-commerce may play an essential role in the future with regard to the growth of the companies and the efficiency of the internal business processes. By means of analysis of variance it can be shown that those items which significantly effect e-commerce are mainly assigned to factor 2 ("efficiency") and factor 3 ("continuance") in the case of Russia and to factor 1 ("growth orientation") and factor 2 ("efficiency") in the Chinese sample.

According to some general trends it is to be expected that both in Russia and China B2B will dominate B2C in the near future. The corresponding hypothesis reads as follows: "Those companies which are involved in e-commerce are mainly active in B2B domains." This assumption could be confirmed by

Characteristics	Factor loadings
Factor 1: Growth orientation	
Reduction of business process time	0.86
Acquisition of new foreign costumers	0.82
Better competitive position and effectivity	0.79
Simplification of the payment modalities	0.72
Increase in efficiency of the business processes in general	0.65
Open-mindedness regarding technology	0.64
Better communication with customers/suppliers	0.64
Higher degree of business process automation	0.62
Standardization of business processes	0.62
Factor 2: Efficiency	
Reduction of procurement time	0.90
Reduction of procurement costs	0.86
Reduction of the stock	0.74
Increase in popularity on the market	0.70
Increase in productivity of the personnel	0.69
Creation of new business practices	0.67
Increase in efficiency of the internal processes	0.62
Additional distribution opportunities	0.53
Factor 3: Distribution advantages	
Reduction of distribution costs	0.75
Acquisition of new domestic costumers	0.66
Factor 4: Capitalization	
Faster capitalization	0.87
Increase of sales	0.73

Table 2. Factor analysis of the Chinese sample

means of chi-square testing (at $\alpha = 0.05$) on the basis of information provided by the companies about their business environment. In Russia 60% of the companies are active in B2B - as opposed to 37.5% in B2C (a company can be active in both fields). In China the discrepancy is even more obvious (66.3% - B2B and 15% - B2C). eMarketplaces have a high significance considering the transactions in B2B e-commerce. While in Russia the importance of eMarketplaces is circumstantial and no relevant increase is expected in the near future, in China the usage of eMarketplaces is already of particular importance. Transactions in this area will even gain in importance according to the Chinese enterprises.

In the final part of the questionnaires the companies and experts were asked to evaluate the external and internal conditions for an e-commerce commitment. In fact we put a special focus on the following two theses: a) "The required external general conditions for the inclusion of e-commerce in business processes are given." and b) "The companies are prepared for the use of e-commerce."

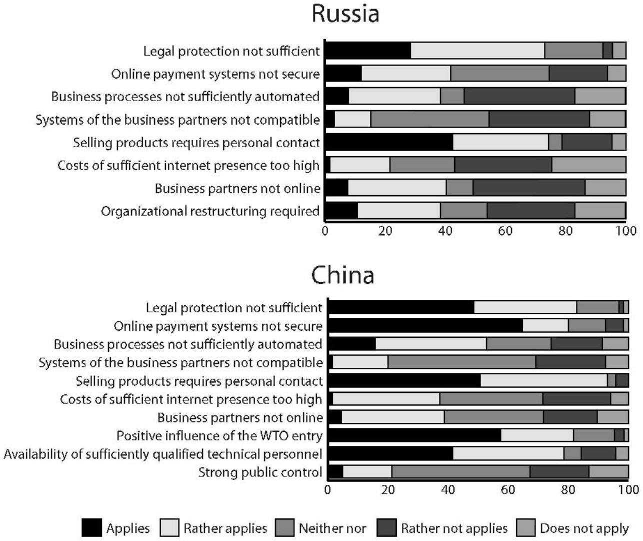


Fig. 2. General conditions for an e-commerce commitment

The opinions of the companies concerning selected general conditions are given in Figure 2. The assessments of the experts show only marginal differences. In fact they are slightly more critical than the companies, both in Russia and China. In both countries the assessment for the legal protection of online businesses is rather bad. The insufficiently developed telecommunication network and the corresponding inadequate diffusion of the internet are criticized in both countries, whereas clear differences occur with regard to the online payment systems. While criticism in Russia is rather moderate we detected a clearly negative assessment of this important aspect of e-commerce in China. Furthermore, the Chinese companies put less emphasis on the degree of public control than the Russians. This must be seen in contrast to the historical background of many Chinese enterprises, especially the large-scale ones, which dominate our sample and which are partly successors of originally state-owned companies. Therefore such dimensions can be seen slightly differently than through West European eyes.

A closer look at the answers of the experts shows their criticism regarding the inadequacy of the logistics system as well as the lack of confidence between business partners in both countries. The latter is also reflected in the emphasis on personal contact in selling processes. While the opinions of the experts and the companies strongly differ with respect to the degree of automation of business processes in Russia, there is a basic agreement regarding this point in China. A different assessment can also be concluded for the qualifications of the involved personnel. While both the Chinese companies and the Chinese experts rate staff qualifications as positive, the Russians are not in agreement with this point. The Russian experts clearly complain about the existing lack of qualified personnel while the companies are seemingly not aware of this problem. The WTO entry of China is assumed to have a positive impact on e-commerce by the Chinese experts as well as the Chinese companies.

All in all the study made it clear that the required external general conditions were neither totally fulfilled in China nor in Russia at the date of data collection. Instead interesting differences can be identified regarding the existing deficits. As a consequence thereof the above-mentioned theses can not be confirmed at the moment.

4 Conclusions

The importance of e-commerce is nowadays largely undisputed. In both countries considered in this paper a continuous diffusion of internet and e-commerce can be observed, although the development seems to be slightly faster in China. Different reasons might be responsible for this asymmetry, such as, the more positive attitude towards e-commerce or the better staff qualifications, for example. Most of the e-commerce activities are concentrated on B2B, where the companies expect to realize increasing efficiency and sustainable competitive advantages. Neither in Russia nor in China are the external general conditions optimal, and this has caused the respective governments to initiate measures focussing on the improvement of the legal and the infrastructural environment of e-commerce and internet usage in general. Nevertheless, the differences between Western Europe and the USA on the one side and Russia and China on the other side will still continue to exist, at least in the near future. Therefore, the existing and partly consolidated e-commerce strategies of western companies cannot be applied without country-specific adjustments in either emerging market. The development of such country-specific e-commerce strategies could be the subject of future research in e-marketing.

References

BDI (2004): Globalisiserungsstrategien deutscher Unternehmen. *BDI-Aussenwirtschaftsbarometer, October 2004, 10–12.*

DECKER, R., HERMELBRACHT, A., and KHOROUNJAIA, T. (2003): E-Commerce als Instrument zur Bearbeitung des russischen Marktes, Status Quo und Entwicklungsperspektiven. *Journal für Betriebswirtschaft, 53, 5-6, 190–207.*

FEY, C.F. and DOERN, R. (2002): The Role of External and Internal Factors in Creating Value Using eCommerce: The Case of Russia. *Working Paper No. 02-102, Stockholm School of Economics in St. Petersburg.*

HALEY, G.T. (2002). E-commerce in China - Changing business as we know it. *Industrial Marketing Management, 31, 2, 119–124.*

HAWK, S. (2004): A Comparison of B2C E-Commerce in Developing Countries. *Electronic Commerce Research, 4, 181–199.*

HOLLENSEN, S. (2004): *Global Marketing – A Decision-Oriented Approach*, 3rd ed., Prentice Hall, Harlow.

PERMINOV, S.B. (2000): The Dissemination of E-Commerce Technologies in Russia. *Emergo : Journal of Transforming Economies and Societies, 7, 4, 83–94.*

TAN, Z. and OUYANG, W. (2004): Diffusion and Impacts of the Internet and E-Commerce in China. *Electronic Markets, 14, 1, 25–35.*

WONG, X., YEN, D.C., and FANG, X. (2004): E-Commerce Development in China and its Implication for Business. *Asia Pacific Journal of Marketing and Logistics, 16, 3, 68–83.*

XIAO, H. (2004): A Comparative Analysis of E-Commerce Growth between China and Canada. *International Economics and Trade Research, 20, 2, 31–34.*

Analyzing Trading Behavior in Transaction Data of Electronic Election Markets

Markus Franke, Andreas Geyer-Schulz, and Bettina Hoser

Information Services and Electronic Markets,
Institute for Information Engineering and Management,
Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

Abstract. In this article we apply the analysis of eigensystems in Hilbert space for analyzing transaction data in real-time double auction markets. While this method is well known in quantum physics, its application for the analysis of financial markets is new. We show that transaction data from a properly designed financial accounting system of a market place completely reflect all market information and that this transaction data can be represented as Hermitian adjacency matrices without information loss.

In this article we apply the analysis of the resulting eigensystem to detect and investigate market-making behavior. We show how some of the stylized facts about trading behavior can be recognized in the eigensystem of the market. We demonstrate the method in a small case study for a political stock market for the 2004 elections for the European Parliament in Germany.

1 Introduction

In a recent article by Hoser and Geyer-Schulz (2005) the analysis of eigensystems of Hermitian matrices in Hilbert space was introduced for directed asymmetric communication structures and applied to the analysis of social networks. However, directed asymmetric communication structures can be interpreted in a managerial context as the flow of operational accounting records over a graph of accounts (the chart of accounts of a market place) as suggested in Franke et al. (2005).

The analysis of directed asymmetric communication structures has a rich and multi-disciplinary tradition e.g. in military signal intelligence in the 1st and 2nd World War, in H. A. Simons's analysis of the behavioral aspects of organizations (Simon (2000)), in social network analysis (Wassermann (1994)), and in Internet search engines for relevance ranking e.g. Page et al. (1998) and Kleinberg (1999).

This article consists of two parts. In the first we give a short introduction to the eigensystem analysis of asymmetric directed transaction streams in markets. In the second we present a small case study with an application to the analysis of trading behavior in a political stock market for the 2004 elections for the European Parliament in Germany.

2 The eigensystem analysis of asymmetric directed transaction streams

The notational conventions of this article are: $x \in \mathbb{C}$ is a complex number in algebraic or in exponential form $x = a + ib = |x|e^{i\phi}$. $Re(x) = a$ is the real, $Im(x) = b$ the imaginary part of x. The absolute value of x is $|x| = \sqrt{a^2 + b^2}$, and the phase $\phi = \arccos \frac{Re(x)}{|x|}$, $0 \leq \phi \leq \pi$, where $i = \sqrt{-1}$, its complex conjugate is $\overline{x} = a - ib$. A column vector is denoted in bold $\mathbf{x}$, its components are $x_j, j = 1 \ldots n$. The vector space is defined by $\mathbf{V} = \mathbb{C}^n$. Matrices are printed as capital letters A. a_{kl} represents the entry in the k-th row and the l-th column. Greek letters denote eigenvalues. λ_k represents the k-th eigenvalue. The complex conjugate transpose of a vector $\mathbf{x}$ is defined as $\mathbf{x}^*$. The transpose of a vector $\mathbf{x}$ is $\mathbf{x}^t$. The outer product of two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$\mathbf{x}\mathbf{y}^* = \begin{pmatrix} x_1\overline{y_1} & \ldots & x_1\overline{y_n} \\ \ldots & \ldots & \ldots \\ x_n\overline{y_1} & \ldots & x_n\overline{y_n} \end{pmatrix} \tag{1}$$

A Hilbert space is a complete normed inner product space as defined by Eqs.(2) - (7) (see e.g. M. H. Stone (1932)):

$$\text{Inner product.} \quad \langle \mathbf{x} \mid \mathbf{y} \rangle = \mathbf{x}^*\mathbf{y} = \sum_{k=1}^{n} \overline{x_k} y_k \tag{2}$$

$$\langle \mathbf{x} \mid \mathbf{x} \rangle \geq 0 \qquad \text{with } \langle \mathbf{x} \mid \mathbf{x} \rangle = 0 \text{ if and only if } \mathbf{x} = 0 \tag{3}$$

$$\langle a\mathbf{x} \mid \mathbf{y} \rangle = \overline{a}\langle \mathbf{x} \mid \mathbf{y} \rangle; \quad \langle \mathbf{x} \mid a\mathbf{y} \rangle = a\langle \mathbf{x} \mid \mathbf{y} \rangle \tag{4}$$

$$\langle \mathbf{x} + \mathbf{y} \mid \mathbf{z} \rangle = \langle \mathbf{x} \mid \mathbf{z} \rangle + \langle \mathbf{y} \mid \mathbf{z} \rangle \tag{5}$$

$$\langle \mathbf{x} \mid \mathbf{y} \rangle = \overline{\langle \mathbf{y} \mid \mathbf{x} \rangle} \tag{6}$$

$$\text{Norm.} \quad \sqrt{\langle \mathbf{x} \mid \mathbf{x} \rangle} = \| \mathbf{x} \| \tag{7}$$

The adjoint space $\mathbf{X}^*$ of $\mathbf{X}$ is given by the set of all semilinear forms (Eq.(4)) on $\mathbf{X}$ (Kato (1995, p.11)). A Hermitian operator is selfadjoint and linear. A matrix H for which $H^* = H$ with $h_{lk} = \overline{h_{kl}}$ holds is called Hermitian. Hermitian matrices are normal: $HH^* = H^*H$. The eigenvalue equation $H\mathbf{x} = \lambda\mathbf{x}$ of a complex Hermitian matrix H can be represented (due to its complete orthonormal eigenvector system) as a Fourier sum representation:

$$H = \sum_{k=1}^{n} \lambda_k P_k; \quad P_k = \mathbf{x}_k\mathbf{x}_k^* \tag{8}$$

$\mathbf{x}_k$ is the k-th eigenvector and P_k the k-th orthogonal projector. Note, that $\sum_{k=1}^{n} P_k = I$, $P_k^* = P_k$, $P_k^2 = P_k$. The spectrum is the set of all eigenvalues.

Since the basis of the eigensystem is orthogonal, it can be chosen to be orthonormal and this holds under arbitrary rotation of the eigenvectors. Hermitian matrices have full rank and, therefore all eigenvalues are simple and real (see Meyer (2000, p. 548)): $\lambda_k \in \mathbb{R} \forall k$. This has the advantage that the interpretation of real eigenvalues is more intuitive than the interpretation of complex eigenvalues of non-symmetric real matrices.

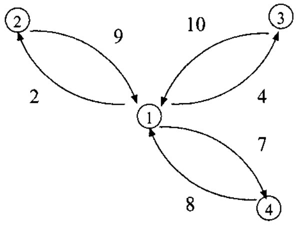
For a complex Hermitian matrix with $tr(H) = 0$ (e.g. no self-reference transactions) some eigenvalues have to be negative due to the fact that $tr(H) = \sum_{k=1}^{n} h_{kk} = \sum_{k=1}^{n} \lambda_k = 0$. A bipartite graph can be represented by a matrix B of order $l = n + m$ with

$$B = \begin{pmatrix} 0_{n \times n} & A \\ A^* & 0_{m \times m} \end{pmatrix} \tag{9}$$

with A representing a n by m matrix. For $n = 1$ and $m = l - 1$ this matrix is a directed, weighted star graph. The spectrum of that system is given by:

$$\sigma(B) = \{+\lambda_1, -\lambda_1, \ldots, +\lambda_{\frac{n+m}{2}}, -\lambda_{\frac{n+m}{2}}\} \tag{10}$$

(see Meyer (2000, p. 555)). For example, consider the directed and weighted star graph with 4 members and its eigensystem shown in table 1.



λ_k		x_{kl}			
1	abs(z)	0.71	0.37	0.43	0.42
	arg(z)	0	-0.57	-0.40	-0.07
-1	abs(z)	0.71	0.37	0.43	0.42
	arg(z)	0	2.6	2.7	3.1
0	abs(z)	0	0.60	0.24	0.78
	arg(z)	undefined	2.8	2.8	0
0	abs(z)	0	0.61	0.76	0.24
	arg(z)	undefined	3.0	0	-2.8

Table 1. Stargraph MStar4 and its eigensystem with $z = |z|e^{i\phi}$

Table 1 shows two non-zero eigenvalues of the same absolute value but with different sign. Equation (10) predicts this as the characteristic of a star graph adjacency matrix. Furthermore, the eigenvectors belonging to the two eigenvalues are the same in absolute values but differ approximately by π in phase. Note, that the trader with ID 1 is the center of the star graph and is indicated as such by the highest absolute value of the eigenvector component.

The matrix H from which the eigensystem in table 1 is computed from the weighted graph shown in table 1 by constructing a real-valued adjacency matrix A from the graph is constructed by $H = (A + iA^T)e^{-i\frac{\pi}{4}}$. This transformation preserves all order properties and leads to a Hermitian matrix as was shown in Hoser and Geyer-Schulz (2005).

The graph in table 1 shows the aggregated transaction stream for the six transactions between the 4 accounts of the traders shown in Figure 1. Figure 1 shows a small financial accounting example with six trades of the form trader k sells to trader l shares and receives an amount of money s. The T-accounts of traders 1, 2, 3, and 4 are shown with the 6 trades recorded according to the conventions of financial accounting. (For an introduction to financial accounting see Geyer-Schulz (1991)). The journal entries of the records numbered (1) to (6) are shown next to the T-accounts. The graph in table 1 shows the monetary flow between the traders. That is the edge is weighted by the money flown from trader k to trader l. In the example, the money flow on the edges of the graph in the observation period is generated by a single transaction. In general and in the rest of this paper, the edge weight is the sum of all monetary flows s from trader k to trader l. And we see that the monetary flow between accounts can be represented as an asymmetric directed communication structure. Note that the weighting function of the transaction stream can be chosen as required by the intended application.

(1) Record trader 1 trader 2, 9 Euro
(2) Record trader 1 trader 3, 10 Euro
(3) Record trader 3 trader 1, 4 Euro
(4) Record trader 4 trader 1, 7 Euro
(5) Record trader 1 trader 4, 8 Euro
(6) Record trader 2 trader 1, 2 Euro

Trader 1

(1) 9	(3) 4
(2) 10	(4) 7
(5) 8	(6) 2

Trader 2

(6) 2	(1) 9

Trader 3

(3) 4	(2) 10

Trader 4

(4) 7	(5) 8

Fig. 1. 4 Accounts and 6 Records

Interpreting this graph as a transaction stream in the accounting system of a market we see that the trader with ID 1 is the only counterparty for all trades. Depending on the strength of the asymmetry of the flows in the graph the trader with ID 1 can be identified either as a monopolist (only seller, more inflows of money,), a monopsonist (only buyer, more outflows of money), or as a market maker (liquidity provider, symmetric flow of money). Hakansson et al. (1985) characterize market makers as liquidity providers for markets which offer bid/ask spreads around the last market transaction price with minimal involvement. They investigate the performance of several decision-making strategies of a market maker in a simulated market environment. The eigensystem of a completely balanced market maker has a real solution of the form characteristic for a star with all trader accounts having a phase shift of π (as theoretically expected). The eigensystems of a monopolist/monopsonist are conjugate complex and exhibit the typical star pattern which we expect from theory. A monopolist has an inbound star pattern, a monopsonist an outbound star pattern. However, the example indicates the limits of the method: without additional contextual information (who has market making obligations which are available e.g. on the Vienna stock exchange (Wiener Börse AG (2002)) a monopolist or a monopsonist star pattern might also indicate an unsuccessful market maker caught in a position.

An election market generates a specific structure for the transactions occurring in it. Since the transactions take place in an anonymous setting (unknown counterparty), the visible transaction partner for a participant is the respective market for a share or the portfolio market. Thus, the transaction graph is a bipartite graph with the traders on one side and the markets on the other as shown in Fig. 2 and represented in equation 9.

As a consequence, the basic communication pattern is that of a star: For each trader, the communication matrix shows a star with him as center. Equally, for each share and the portfolio, there is a star pattern with the market as center.

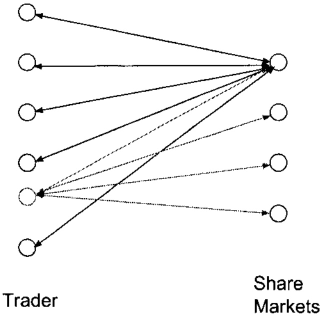


Fig. 2. Star graphs build the transaction structure of the market

3 Analyzing Trading Behavior in an Experimental Forecasting Market

Recent interest in experimental forecasting markets is due to the fact that such markets are seen as marketing research tools comparable to opinion polls (e.g. Spann and Skiera (2003, 2004)) which are more cost effective. Forecasting markets have been pioneered by Forsythe, Nelson, Neumann and Wright with the Iowa Election Market which successfully predicted the US presidential elections in 1988 and 1992 and outperformed the polls (Forsythe et al. (1992)).

As a case study, we investigate transaction data from a political stock market for the elections of the European parliament in Germany in 2004. In an election market the following design is typical: On the primary market, portfolios containing one of each share present in the market can be bought or sold at a fixed price. Furthermore, for each party exists a secondary market where the shares of this party can be traded in a continuous double auction. After the close of the market, the shares in the portfolio of each trader are bought back by the market operator at a price that reflects the election's result. Depending on the design of the market, the participants receive the value of their portfolio in cash or receive a price according to rank or are simply ranked according to that value.

This market was open from beginning of March 5th, 2004 till the closing of the polling stations on June 13th, 2004. 58 traders submitted 2230 offers for the six shares (SPD, CDU, Grüne, FDP, PDS, and others), resulting in a total of 1014 transactions. There was no monetary incentive for participation in the market, instead the rank of the players was used for motivation purposes.

For the analysis of this market transactions were weighted with the current market price of the transaction. Offers to buy a share are weights of arcs going from markets to traders, offers to sell a share are weights of arcs going from traders to markets.

Figure 3 shows the symmetry of the spectrum as predicted by equation 10. In Figure 4 we see that the first two subspaces already cover approximately 75% of the data variance. The first four subspaces cover almost 90%.

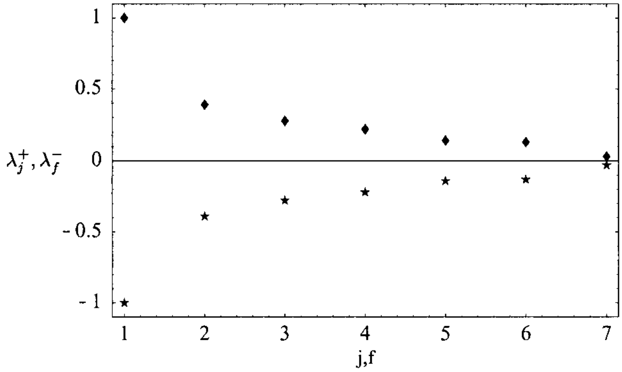


Fig. 3. Eigenvalues for the elections to the EU parliament in Germany

Figures 5 and 6 show the real and the imaginary part of the projectors defined in equation 8 of the third subspace. This subspace (and the fourth which shows a rotation of approximately π and which we do not show, because of space restrictions) show a star pattern dominated by trader 52 which is the most active trader in this market. A closer inspection of the projector (and of his trading transactions) show that this trader did not trade any portfolios. This eliminates the possibility of arbitrage trades. For the shares of CDU, SPD, and FDP he proved to be the most important trader, although his position in these shares was almost balanced (except for FDP where he got caught in a position at the close of the market). His final position was 50% in the money with positions in Grüne, FDP, PDS and others. His total profit

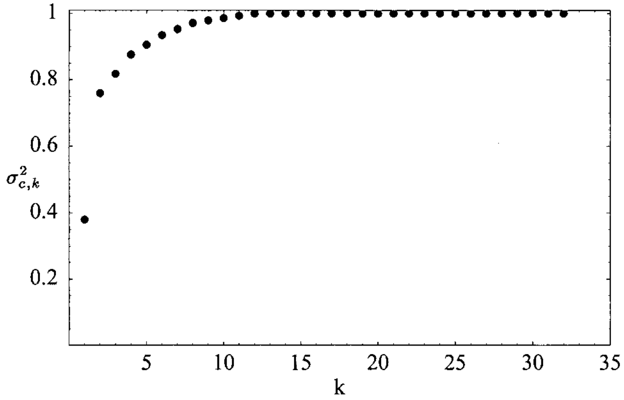


Fig. 4. Cumulated Variance Covered by the Eigenspaces

was 70010.31 which made him not only the most active but also the most successful trader in the market. Furthermore, Figures 5 and 6 show a band for this trader indicating trading activity with many other traders. A closer inspection of his transaction data revealed that he traded with 25 counterparties which makes it unlikely that he tried to manipulate the market and which verifies the pattern suggested by the band in the projector. Furthermore, the eigenvalues indicate that this trader accounts for approximately 25% of the market activity.

In this experiment, the market was not efficient in predicting the outcome of the election as shown by prediction errors for SPD and others in the magnitude of 5.5% and 5.37%. However, given that trader 52 accounts for 25% of the market activity, this is not a surprise: In election markets identification of the power structure is an indirect indication of market inefficiency.

4 Conclusion

In this article we have shown that the eigensystem analysis in a Hilbert space is a promising tool for the analysis of market transaction data. We applied the method for analysing trading behavior in an experimental election market for the 2004 elections to the European Parliament in Germany. Further research is required in order to systematically investigate stylized facts about trading behavior from market microstructure analysis and their patterns in the eigensystem of market transactions.

Acknowledgement. We gratefully acknowledge the funding of the project “SESAM” by the Bundesministerium für Bildung und Forschung within the scope of the research initiative “Internetökonomie”.

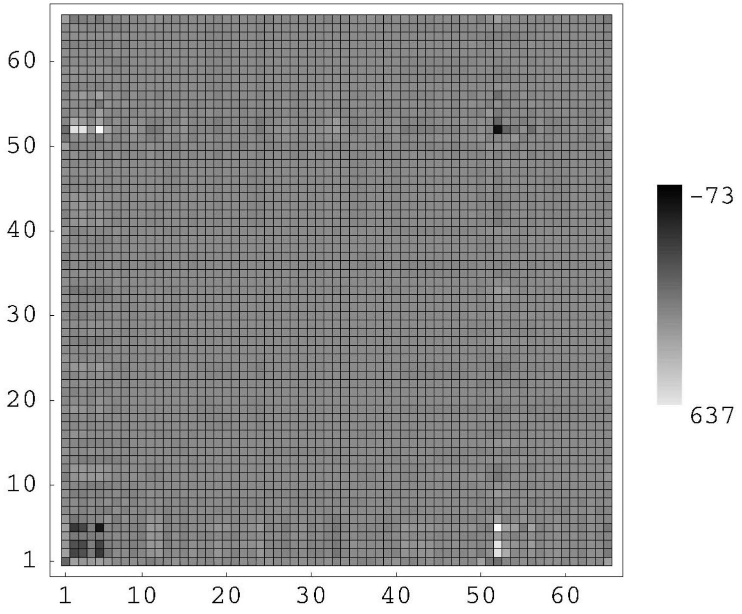


Fig. 5. *Re*(P_3)

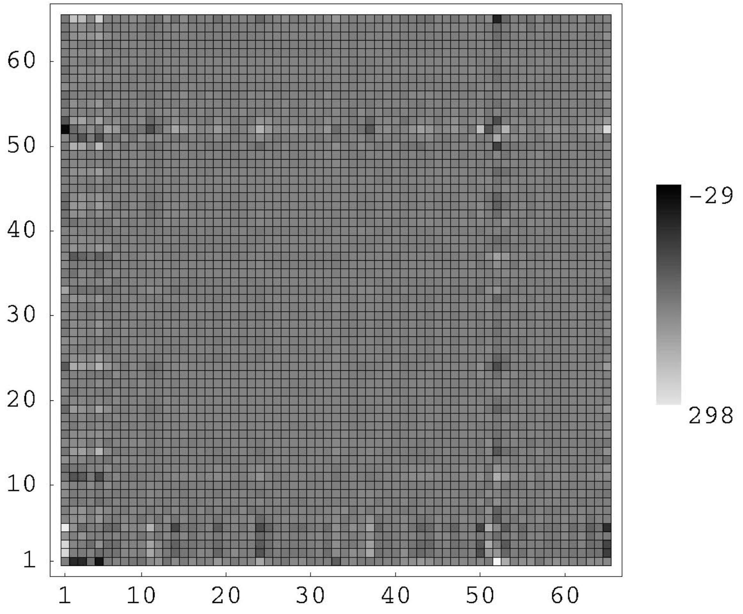


Fig. 6. *Im*(P_3)

References

FORSYTHE, R., NELSON, F., NEUMANN, G. R., and WRIGHT, J. (1992): Anatomy of an experimental political stock market. *The American Economic Review, 82(5), 1142–1161.*

FRANKE, M., GEYER-SCHULZ, A., and HOSER, B. (2005): On the analysis of asymmetric directed communication structures in electronic election markets. In: T. Fent, A, Prskawetz, F. Billari, and J. Scheffran (Eds.): *Proc. Agent-Based Computational Modelling in Demography 2003.* Submitted.

GEYER-SCHULZ, A. (1991): An Introduction to Financial Accounting with APL2. Technical report, ACM SIGAPL, New York, N.Y.

HAKANSSON, N. H., BEJA, A., and KALE, J. (1985): On the feasibility of automated market making by a programmed specialist. *The Journal of Finance, 40(1), 1–20.*

HOSER, B. and GEYER-SCHULZ, A. (2005): Eigenspectralanalysis of hermitian adjacency matrices for the analysis of group substructures. *Journal of Mathematical Sociology.* In press.

KATO, T. (1995): *Perturbation Theory for Linear Operators.* Springer, New York, 2 edition.

KLEINBERG, J. M. (1999).: Authoritative sources in a hyperlinked environment. *JACM, 46(5), 604–632.*

MEYER, C. D. (2000): *Matrix Analysis and Applied Linear Algebra.* Society for Industrial and Applied Mathematics, Philadelphia.

PAGE, L., BRIN, S., MOTWANI, R., and WINOGRAD, T. (1998): The Page Rank Citation Ranking: Bringing Order to the Web. Technical report, Computer Science Department, Stanford University.

SIMON, H. A. (2000): *Administrative Behavior : A Study of Decision-Making Processes in Administrative Organizations.* Free Press, New York, 4 edition.

SPANN, M. and SKIERA, B. (2003): Internet-based virtual stock markets for business forecasting. *Management Science, 49(10), 1310–1326.*

SPANN, M. and SKIERA, B. (2004). Einsatzmöglichkeiten virtueller Börsen in der Marktforschung. *ZfB – Zeitschrift für Betriebswirtschaft (Ergänzungsheft), (2), 25–48.*

STONE, M. H. (1932): *Linear Transformations in Hilbert Space and their Applications to Analysis*, volume 15 of *American Mathematical Society, Colloquium Publications.* American Mathematical Society, New York.

WASSERMANN, S. and FAUST, K. (1994): *Social Network Analysis, Methods and Applications.* Cambridge University Press, Cambridge.

WIENER BÖRSE AG (2002): Das Xetra Marktmodell. Technical report, Wiener Börse AG.

Critical Success Factors for Data Mining Projects

Andreas Hilbert

Fakultät Wirtschaftswissenschaften,
Universität Dresden, D-01062 Dresden, Germany

Abstract. Due to the strategic reorientation of many companies in recent years data mining, as a tool for the analytical customer relation management, became more and more important. Because, however, the success of data mining is not always guaranteed, this paper wants to explain whether respectively under which conditions the investment in data mining projects could be profitable. Using the theoretical background of critical success factors, data mining and some related topics, a model to explain the success of a data mining project in a company has been developed. The derived hypotheses have been tested in an empirical study of German companies. As a result the following critical success factors could be proofed: the *commitment* of the top management, the existence of a *change management*, a fixed budget for the project, a good integration of the data mining process in the *IT landscape* as well as a high *quality of the used data.*

1 Introduction

Since the end of the 80th a clear intensification of the competition could be noticed in a lot of markets (Raab and Lorbacher (2002), p. 11). In order to ensure the own competitive advantages, more and more companies increasingly moved the customer into the mid-point of their activities. Thus, a new management philosophy was born: the so called customer relationship management or CRM (Homburg and Bruhn (2000), p. 7). This new style of management attempts – as a consequence of the customer orientation – to give each customer his product at the right moment on the suitable channel for the right price. To do this, all information about the customers are needed; information, which have to be suitably stored, evaluated and interpreted. Fortunately, different very powerful and (relatively) cheap new information technologies, as for example data warehouses, online analytical processing (OLAP) or data mining techniques, were introduced at the same time as CRM came up. Thus, a immense potential of applications was predicted for customer relationship management and its components. But unfortunately, different studies report that the same structural problems could be noticed at the implementation of CRM systems as years before at the introduction of management information systems (MIS), decision support systems (DSS) or executive information systems (EIS). A hype came up, each company would and should install a CRM system, whether justified or not. Unfortunately, however, the implementation of such IT based solutions is normally very expensive, causes additionally more expensive experts and can only deliver success, if (sufficient) information on the customers are available. This directly implies the question, when and under which conditions it is worth to install a CRM system in a company. A variety of studies considers this problem, but the most only discuss the topic CRM system in general (Wilde and Hippner (2000) or Holzer et al. (2000)), from a very practical point-of-view or analyze only the technical point-of-view of some single components as for example of data warehouses or data mining tools (Wilde and Hippner (2002)). A more theoretically orientated but descriptive study only exists for data warehousing project from 1996 (see Dittmar (1999)). Some recent studies or studies which are theoretically based, empirically evaluated as well as confirmatorical in their design are not mentioned in the literature. Therefor, the present paper wants to fill this gap and discuss which factors – also called critical success factors (CSF) – can affect the success of the application of data mining projects in a company.

2 A framework for CSF research

The starting point of the CSF research was a paper of Daniel (1961) who analyzed some critical factors in context of management information systems. The concept itself can be defined as follows: *A* **success factor** *is a factor which has a sustainable and positive effect on the success of a company. By using these factors a competitive advantage could be realized.*

Because, however, there are many different potential success factors, the academic research in this field is only interested in the most critical ones. These factors are called CSF (or key factors) and can be classified in three groups: The first group is the subset of the so called endogenous CSF which can be directly controlled by the management of a company, e.g. the marketing or the business strategy. The second class contains the so called exogenous CSF which are not directly manageable. Examples are the market or the competitors. The third group is the class of moderator variables which have the task to mediate between the 'real' success factors and the success values. A moderator variable is a kind of sub-ordinate target, a means to an end. An example is the market share as a sub-ordinate target and a determinate of the return on investment (ROI).

Considering all these groups of CSF, a very general model for the causal relationship within the CSF research could be defined in Figure 1 (see also Hildebrandt and Trommsdorf (1989), p. 16 and Steffenhagen (1998), p. 327). This model can be used for all CSF research, and thus also for the analysis of critical success factors for data mining projects.

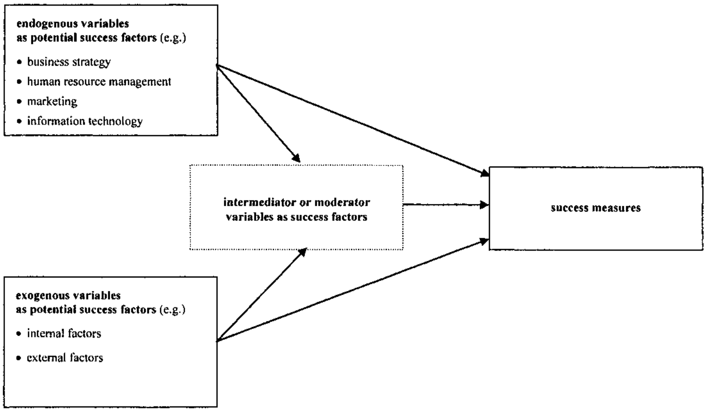

endogenous variables
as potential success factors (e.g.)
• business strategy
• human resource management
• marketing
• information technology
intermediator or moderator
variables as success factors
success measures
exogenous variables
as potential success factors (e.g.)
• internal factors
• external factors

Fig. 1. Basic framework for the empirical CSF research

3 Some hypotheses for CSF research

Based on the above mentioned framework (see Figure 1) and the underlying theory (for details see Hilbert (2002)), in the following some research hypotheses are formulated which should be verified in an empirical study. The first hypothesis considers the area of human culture in a company, also called corporate culture. In general, it could be expected that the openness for innovations and the open-mindedness of the top management have a positive effect on the success of such projects. This positive corporate culture advantages new technologies like data mining and prohibits the build-up of barriers. Thus, the hypothesis about the potential success factor *corporate culture* (F1) reads as follows:

H1: The openness for innovations and the open-mindedness of the top management have a positive effect on the success of data mining projects.

Empirical studies support the hypothesis that the status of the MIS responsible – and thus the top management commitment for the project – is very closely related to the success of a MIS system (Weitzendorf (2000), p. 114). Accordingly, the hypothesis about the potential success factor *commitment* (F2) reads as follows:

H2: The top management commitment directly effects the engagement and the sensibility of the employees regarding the importance of data mining projects and causes in this way the success of such projects.

Furthermore, the organization(al structure) is very important for the success of data mining projects. Especially, some studies (see for example Wilde (2001)) support the hypothesis that a company with a well working change management – which motivates and qualifies the employees for the new tasks and which modifies the work flow early enough – could be very successful. As a consequence the hypothesis for the potential success factor *change management* (F3) is the following:

> *H3: The existence of a change management positively effects the integration and application of data mining results, and thus the success of such projects.*

Another hypothesis for the success factor *organizational guarantees* (or short: organization) (F4) is the following:

> *H4: The better the guarantees for the data mining project – for example in form of fixed financial budgets – are, the more successful the project is.*

The last hypothesis out of the organizational area considers the integration and application of the results of the data mining projects and could be formulated as a hypothesis regarding a intermediator (as moderator variable). Accordingly, the hypothesis about the potential success factor *integration of the results* (F5) reads as follows:

> *H5: The more open-minded the operating departments are in respect of the data mining results and the more often the data mining results are applied, the more successful the data mining projects are.*

The existence of an intelligent IT landscape can be seen as a minimum requirement for the success of data mining projects. However, it is assumed that this condition is (almost) always fulfilled. Thus, the corresponding hypothesis out of the IT area for the potential success factor *information technology* (F6) aims at the cooperation of the different components:

> *H6: A good cooperation of all soft- and hardware components in the periphery of a data mining project as well as the high quality of the customer data are critical for the success of a data mining project.*

Finally, the area of external resources should be considered. Here, the following hypothesis for the success factor *external resources* (F7) can be noticed:

> *H7: The use of external resources like for example consultants or external data has no effect on the success of data mining projects.*

Beside these hypotheses regarding the effects of endogenous variables one hypothesis for a exogenous variable should be formulated. This hypothesis considers the *market and the competitors* (success factor F8) of a company – which can not be controlled by the company – and postulates an independency in the following sense:

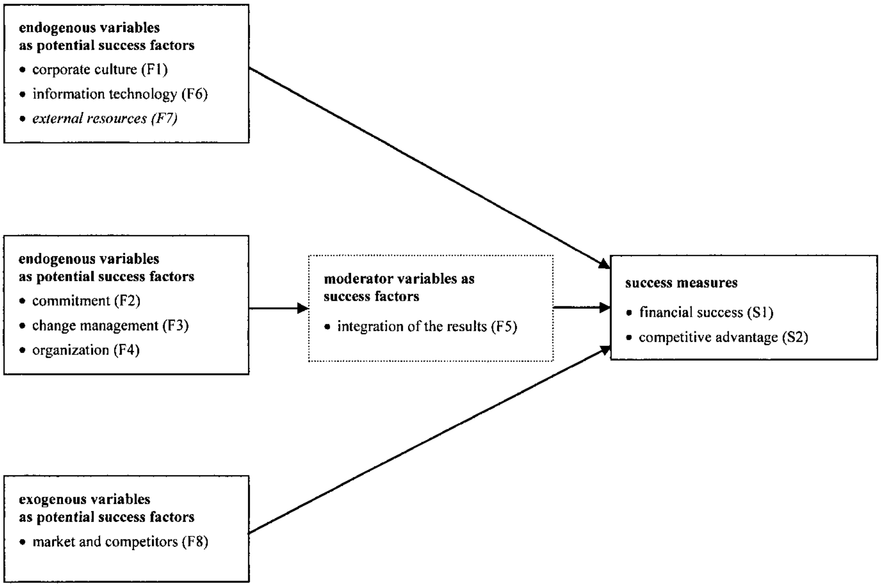

endogenous variables
as potential success factors
• corporate culture (F1)
• information technology (F6)
• *external resources (F7)*
endogenous variables
as potential success factors
• commitment (F2)
• change management (F3)
• organization (F4)
moderator variables as
success factors
• integration of the results (F5)
success measures
• financial success (S1)
• competitive advantage (S2)
exogenous variables
as potential success factors
• market and competitors (F8)

Fig. 2. Framework for the CSF research for data mining projects

H8: The market and the competitor situation have no effect on the success of data mining projects.

If these hypotheses H1 - H8 with the corresponding success factors (F1 to F8) would be integrated in one model using the framework of Figure 1, a concrete model for the CSF research for data mining projects follows (see Figure 2). This model could be used to identify and/or analyze the most important critical success factors, but without demand for completeness. Finally, it should be figured out that this derived model is similar in its basic structure to other models presented in the CSF research (Müller (1999), p. 140, Daniel (1996), p. 5 or Weitzendorf (2000), p. 112). The main difference is, however, the focussed view on details of the information technology area.

4 The empirical evaluation

The objective of this empirical study follows the tradition of the CSF research and shall identify critical success factors for data mining projects. To achieve this objective it is necessary to operationalize at first the derived but latent success factors F1 to F8 as well as the latent construct *success of data mining projects*, called construct S. The operationalization which only serves for the indirect measurement of the not directly observable factors can be solved by a well-drawn definition of some directly measurable indicators (Homburg and Giering (1996), p. 6). These indicators themselves are ordinary variables which should have a strong relationship to the correspondent factor F1 to F8 as well as to construct S and should be selected under consideration of different aspects like objectivity, reliability and validity. An intensive discussion of this step can be read in Hilbert (2002).

After this operationalization step the 'real' empirical part of the study has to be done. A sample of 145 proper German companies was drawn and the above mentioned indicators as well as some other information were observed (see also Hilbert (2002)). After that, the different indicator-factor models were analyzed to get an idea of the empirical fitness of the operationalization, which means how well can the indicators describe the corresponding factor. It could be noticed that all models for F1 to F8 and the factor models for S are quite good, so that the combination of all factors – using the corresponding indicators – to an overall (structural equation) model could be verified.

The used structure within this overall model bases on the above mentioned hypotheses and the conceptional framework in Figure 2. But it should be noticed that – due to the need for a model which is reduced in its complexity – no direct connections, also called paths, between all success factors and the one success measurement exist. In fact, a new factor, the *overall success*, is defined which summarizes the subsets of *financial orientated* as well as *competition orientated success* indicators. An analysis based on an explorative and a confirmatorical factor analysis as well as some content based considerations approves the adequacy of this approach. Finally, the link of the different parts of the model, one part for the success factors and one part for the success measurement, is realized by this new factor. The resulting path model, without the indicators, is described in Figure 3. Additionally, some selected fit measures are listed in Table 1.

	GFI	AGFI	RMR	χ^2/df	RMSEA	CFI	NFI
overall model	0.955	0.937	0.074	1.259	0.048	0.999	0.929
expected level	≥ 0.90	≥ 0.80	≤ 0.10	≤ 2.50	≤ 0.05	≥ 0.90	≥ 0.90

Table 1. Selected global fit measurements

Interpreting the fit measures in Table 1 it could be figured out that the model with 164 degrees of freedom is very good. The standardized path coefficients result from an ULS estimation and already converge after 60 iterations. Highly significant coefficients ($\alpha < 0.01$) are marked up with ***, significant values ($\alpha < 0.05$) with ** and weak significant path coefficients ($\alpha < 0.10$) with *. All unmarked paths are not significant to a level of $\alpha > 0.10$ (n.s.).

Analyzing these empirical results it can be recognized that the endogenous factor *corporate culture* (F1) as well as the exogenous factor *market and competitors* (F8) have no significant effect on the success of data mining

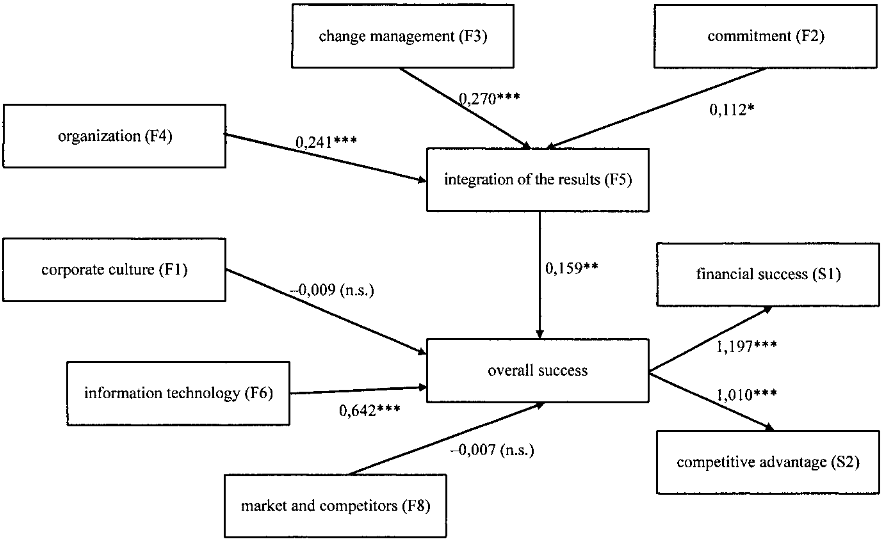


Fig. 3. The path model of critical success factors for data mining projects

projects in the analyzed companies. First, this seems to be very astonishing but taking a closer look that could be explained. Not only companies which are present in rather decreasing or stagnating markets can use data mining tools with success, also companies on expanding markets can do so. Thus, considering all different types of markets and competitor situations, no effects can be observed. The result for factor F1 is indeed rather surprising, particularly if one compares the literature and the stressed importance of the corporate culture for such projects. But taking a closer look it should also be clear that not the general way of thinking, established in the corporate culture, but the way of acting is important for the success of new projects. And if there is now a (slight) difference between the way of thinking and the way of acting – especially for 'conservative' companies which have recognized that a change management is necessary for being a competitive company – no differences between modern and conservative companies can be analyzed, i.e. no significant effects of corporate culture on the success of data mining projects exist. Here, further research should analyze this problem in detail.

Interpreting the local fit measures of the different factor models the following could be figured out: (Almost) all indicator reliabilities are higher than the minimum level of 0.4. The factor reliabilities are always higher than 0.6 and with two exceptions (F2, F6) even higher than 0.7. Finally, the average explained variance has two exceptions (F2, F6 again), too, which do not fulfil the minimum level of 0.7.

The fit of the structural model itself is also (very) good. The squared multiple correlation (smc) which describes – like the coefficient of determination in the framework of linear regression analysis – the share of the variance of a factor by using a model with all the other factors, varies between 0.501 for the factor *overall success* and 0.962 for the factor *financial orientated success.* Only the factor *integration of the results* which has to be treated as an moderator variable and which is not the main focus of this study has a value of 0.226 for the smc. Furthermore, it is worth to be mentioned that the reduced model only with the significant factors, i.e. without factor F1 and F8, has similar path coefficients and similar local as well as global fit measures. Thus, the presented model can be labeled as well fitted and stable.

A substantial interpretation of the model is as follows: The *organizational background*, the *top management commitment*, an existing *change management*, the *information technology* as well as the *integration of the results* in the different departments of a company have clearly significant, direct or indirect effects on the success of data mining projects. Thus, the above mentioned hypotheses H2 to H6 can be supported. The business environment (*market and competitors*) – expressed in form of hypothesis H8 – has no significant effect on such projects. A possible explanation was given above. The same is true for the *corporate culture* H1. Finally, it has to be mentioned that only the hypothesis H7 about the effect of *external resources* on the success of data mining projects could not be analyzed due to the poorly fitted indicator-factor model in respect to the reliability and validity.

5 Conclusion

This study shows that the 'characteristic' of a company could positively affect the success of data mining projects. Particularly in that cases, in which the top management has the right feeling for the importance of such projects, data mining can contribute to the achievement of both - *financial* as well as *competitive* orientated - aims. Other important positive assumptions can be summarized as follows:

- An existing *change management* give the organizational conditions for the integration of data mining projects into the companies, for example by the adoption of the business work flow.
- Certain *organizational guarantees* exist, e.g. some fixed financial budgets.
- The *information technology* enables the integration of data mining into the IT landscape.
- The *quality of the underlying data* is very high.

While these results are not very surprising, it is also coherent and theory-compliant that the *business environment* and the strength of the *competitors* have no (significant) effects on the success of data mining projects. On the other hand, another interesting result of the study is the fact that the *corporate culture*, i.e. the openness for innovations and the open-mindedness of the management, has also no effect on the success of such projects. First of all, this seems to be astonishing, but taking a closer look it becomes clear.

Whereas the corporate culture is rather a general way of thinking, the operative implementation can be very different, even if the way of thinking is identical. So, a company can apply data mining very successfully, even if its management is rather conservative – assumed, a modern change management is implemented and/or the commitment of the top management is guaranteed. By this, it is clear that not the way of thinking but the way of acting is critical for the success of such and similar projects.

Finally, the following short recommendations can be derived: To ensure the success of a data mining project a company has to fulfil the right qualifications in the areas of human being (*commitment*), organization (*change management, fixed budgets*) and IT (*integration, data quality*).

References

DANIEL, R. (1961): Management Information Crisis. *Harvard Business Review, September/October, 111–121.*

DANIEL, D.R. (1996): *Erfolgsfaktoren für die Einführung von Informationstechnologien in internationalen Unternehmen in Indonesien.* Dissertation, Freiburg.

DITTMAR, C. (1999): *Erfolgsfaktoren für Data Warehouse Projekte - eine empirische Studie aus Sicht der Anwendungsunternehmen.* Institut für Unternehmensführung und Unternehmensforschung, Arbeitsbericht Nr. 78, Universität Bochum.

HILBERT, A. (2002): *Data Mining Projekte im unternehmerischen Umfeld: Eine empirische Studie deutscher Unternehmen.* Arbeitspapiere zur Mathematischen Wirtschaftsforschung, Universität Augsburg, Heft 183.

HILDEBRANDT, L. and TROMMSDORF, V. (1989): Anwendungen der Erfolgsfaktorenanalyse im Handel. In: V. Trommsdorff (Ed.): *Handelsforschung 1989.* Gabler, Wiesbaden, 15-26.

HOLZER, M., BÖLSCHER, A., and GRÄTHER, M. (2000): *Welchen Stellenwert hat CRM? Ein Vergleich des Einsatzes von Customer Relationship Management-Systemen bei deutschen Autobanken und Kfz-Leasinggesellschaften.* Managementteam Unternehmensberatung, Wiesbaden.

HOMBURG, C. and BRUHN, M. (2000): Kundenbindungsmanagement - Eine Einführung. In: M. Bruhn and C. Homburg (Eds.): *Handbuch Kundenbindungsmanagement.* Gabler, Wiesbaden.

HOMBURG, C. and GIERING, A. (1996): Konzeptionalisierung und Operationalisierung komplexer Konstrukte: Ein Leitfaden für die Marketingforschung. *Marketing ZFP, 1, 5–24.*

MÜLLER, R. (1999): *Erfolgsfaktoren schnell wachsender Software-Startups: eine lebenszyklische Untersuchung von Softwareunternehmen des Produktionsgeschäfts.* Lang, Frankfurt.

RAAB, G. and LORBACHER, N. (2002): *Customer Relationship Management: Aufbau dauerhafter und profitabler Kundenbeziehungen.* Sauer, Heidelberg.

STEFFENHAGEN, H. (1998): Erfolgsfaktorenforschung für die Werbung - Bisherige Ansätze und deren Beurteilung. In: M. Bruhn and H. Steffenhagen (Eds.): *Marktorientierte Unternehmensführung.* Gabler, Wiesbaden, 323–350.

WEITZENDORF, T. (2000): *Der Mehrwert der Informationstechnologie: eine empirische Studie der wesentlichen Erfolgsfaktoren auf den Unternehmenserfolg.* Gabler, Wiesbaden.

WILDE, K.D. (2001): Return on Investment: Erfolgsfaktoren im Customer Relationship Management. *absatzwirtschaft online, Oktober 2001.*

WILDE, K.D. and HIPPNER, H. (2000): *CRM2000 - Customer Relationship Management.* Verlagsgruppe Handelsblatt, Düsseldorf.

WILDE, K.D. and HIPPNER, H. (2002): *Data Mining - Mehr Gewinn aus Ihren Kundendaten.* Verlagsgruppe Handelsblatt, Düsseldorf.

Equity Analysis by Functional Approach

Thomas Kämpke and Franz Josef Radermacher

Forschungsinstitut für anwendungsorientierte Wissensverarbeitung FAW/n,
Helmholtzstr. 16, D-89081 Ulm, Germany
{kaempke, radermacher}@faw.uni-ulm.de

Abstract. A notion of relative poverty leads to different types of one-parametric Lorenz curves. These curves allow to compare any individual income to the average of all larger or all smaller incomes. The underlying least square fittings are effectively computable.

1 Introduction

Quantitative analysis of income distribution has regained attention in the last years in the context of sustainable pathways to the future. In the following, the EU definition of poverty and the EU principle to co-funding regional development lead to a differential equation approach. This approach has already allowed some major insights and has been publicly debated, see Radermacher (2002a, 2002b) and Kämpke et al. (2002). It is a major basis for the Global Marshall Plan/planetary contract initiative (Möller et al. (2004), Radermacher (2004a)) and – in particular – it helped to demonstrate that market-fundamentalism makes countries poorer than they need to be, see Radermacher (2002b, 2004b, 2004c).

The poverty notion of the European Union considers an individual of a nation to be poor whenever he falls short of 50% of the average per capita income. This view and variations of it allow to derive several one-parametric classes of Lorenz curves. While this derivation can be traced completely, its empirical counterpart can less be so. Measured income data adhere to different concepts and discourses. Income may be measured before and after taxes, state subsidies may be included or not, prices may be internal or dollar adjusted, the black market and illegal labour force may be included or not, and income data may be tuned by government or not. Some countries do not have or provide data at all. So, availability and comparability of data for a world income distribution are important issues, comp. Sala-i-Martin (2002a) and (2002b).

The transition from empirical data to parameter identification of Lorenz curves will be facilitated by regression. Regressions are based on minimum square error fits which here are not computable in closed form. Nevertheless, the error function of each fit appears to have a unique minimum and heuristics will provide for approximations of that minimum.

2 Framework

2.1 Differential equation approach

The cumulative distribution of income or consumption within a nation can be described by the nation's Lorenz curve $F(x)$, $x \in [0,1]$. Here, no distinction is made between "income", "consumption" and "welfare". According to the poverty notion of the European Union, an individual of a nation is considered to be poor if his income falls short of 50% of the average per capita income of that nation (European Parliament (1999) and Finland (2000)). Instead of considering only the poorest, any individual's income can be compared to the average income of all richer individuals. Moreover, the actual fraction of individual vs. average income need not be 50% but some other, unknown value. This value will be estimated.

Comparing the income of an individual to the income of a group leads to a differential equation. The rationale is as follows. Assume that quantile x of the population has received its cumulative income $F(x)$. This leaves an income $1 - F(x)$ to be distributed among the remaining fraction of the population which is $1 - x$. The average income of all richer individuals thus equals $\frac{1-F(x)}{1-x}$. The income at the top level of any quantile x is now assumed to be a constant fraction ε, $0 < \varepsilon < 1$, thereof. This results in the ordinary differential equation (ODE)

$$F'(x) = \varepsilon \cdot \frac{1 - F(x)}{1 - x},$$

comp. Kämpke et al. (2002), Radermacher (2001, 2002b). Solutions of this linear inhomogeneous ODE that also satisfy the normalization conditions $F(0) = 0$ and $F(1) = 1$ are given by the manifold $F_\varepsilon(x) = 1 - (1 - x)^\varepsilon$. A curve of type F_ε is sketched in Figure 1.

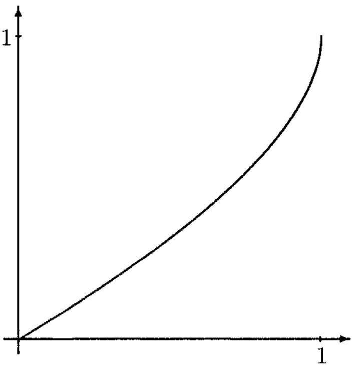


Fig. 1. Sketch of Lorenz curve $F_{0.588}(x)$.

Distributional inequity decreases with the parameter which means that a larger parameter ε corresponds to less inequity. Curves of the type F_ε Lorenz dominate each other which means that they do not intersect except at the endpoints.

The Gini index as measure of the inequality of an income distribution in terms of integral distance from the equal distribution can be computed in closed form here by $\int_0^1 x - F_\varepsilon(x)\, dx = \frac{1}{\varepsilon+1} - \frac{1}{2}$. The Gini index belonging to functions of type F_ε as well as to other parametric Lorenz curves is called functional Gini index.

2.2 Related work

The current type of Lorenz curve has also been proposed by Rasche et al. (1980) in the more general form $F(x) = (1-(1-x)^\alpha)^{\frac{1}{\beta}}$, $0 < \alpha < 1$, $\beta \leq 1$. This type of curve apparently was motivated by insufficient curvature features of other types of Lorenz curves. Driven by empirical motivation, β-distributions $F(x) = x - \vartheta x^\gamma (1-x)^\delta$ and quadratic income distributions $F(x) = \frac{1}{2}(bx + e + \sqrt{mx^2 + nx + e^2})$ were considered, see Datt (1998). An overview on parametric Lorenz curves is provided in Chotiikapanich and Griffiths (1999). Quite a few parametric Lorenz curves were adopted from probability distribution functions, see Ryu and Slotje (1999).

Variations of the Rasche curves like $F(x) = x^\alpha(1-(1-x)^\beta)$, $\alpha > 0$, $0 < \beta \leq 1$, and the exponential curves $F(x) = \frac{e^{\kappa\, x}-1}{e^\kappa - 1}$, $\kappa > 0$, have also been proposed, see Cheong (2002). Parametric Lorenz curves are complemented by non-parametric curves such as kernel estimates and by quantile ratios, see Ginther (1995).

Poverty measures such as poverty lines, poverty gaps (degree of shortfall of poverty lines), the Lorenz family of inequality measures (Aaberge (2000)), and the Foster-Greer-Thorbecke measures (Foster et al. (1984)) are not considered here. These let poverty appear as a problem of only the lower part of the income distribution. Inequality indices requiring other than income or consumption data like entropy measures and the Atkinson index (Litchfield (1999)), are also not considered because they need external parameters which are not obvious to set. A survey of inequity measurement can be found in Silber (1999).

A recent network approach (Bouchard and Mezard (2000)) for a finite set of economical agents concludes that the heavy tail of absolute wealth is approximately Pareto-distributed. This result from "econophysics" is coherent with the present assumptions since the Lorenz curves of type F_ε can be shown to be equivalent to the absolute income distribution being of the Pareto type. This means it obeys a so-called power law. Here, the equivalence to the power law solution is exact and it holds for the complete domain of wealth values instead of large values only.

3 Extension by inequalities

Convexity of an arbitrary differentiable Lorenz curve implies that the tangent slope at $(x, F(x))$ can be sandwiched by the slope of the secant through $(0,0)$ and $(x, F(x))$, and by the slope of the secant through $(x, F(x))$ and $(1,1)$, comp. figure 2. The slope of the lower secant is proportional to the per capita income of the segment $[0,x]$ and the slope of the upper secant is proportional to the per capita income of the segment $[1-x,1]$. The slopes are related by

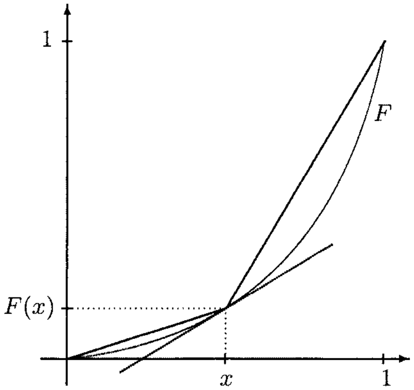


Fig. 2. Tangent at x compared to two secants.

$$\frac{F(x)}{x} \leq F'(x) \leq \frac{1-F(x)}{1-x}.$$

Requiring that the inequalities become equations leads to the trivial solution $F(x) = x$. Relaxations of the equations are obtained by the parameter ε and by an additional parameter $b \geq 1$ such that

$$\frac{F(x)}{x} \cdot b = F'(x) = \varepsilon \cdot \frac{1-F(x)}{1-x}.$$

The right ODE is solved by $F_\varepsilon(x) = 1-(1-x)^\varepsilon$ and the left ODE is solved by $F_b(x) = x^b$. Instead of focussing on the poorest of all richest, the left ODE focusses on the richest of all poorest, stating that his income exceeds the average of all lower incomes by a constant multiple.

Both foregoing ODEs form a system of overconstrained differential equations (Kämpke (2002)) which means that they cannot be fulfilled simultaneously. A relaxation leaving out the differential amounts to the pure functional equation $\frac{F(x)}{x} \cdot b = \varepsilon \cdot \frac{1-F(x)}{1-x}$ with solution $F_{\varepsilon,b}(x) = \frac{x}{b/\varepsilon-(b/\varepsilon-1)x}$ which is a valid Lorenz curve. Replacing b/ε by a single value m shows that this class of Lorenz curves also is one-parametric.

4 Application

The one-parametric Lorenz curves derived in section 2.1 and in section 3 were used for empirical investigations. This required to solve three fitting problems. Any finite point set $\{(x_i, y_i) \mid i = 1, \ldots, n\}$ can be used as support set for the one-parametric regression problems

$$\min_{0\leq\varepsilon\leq1} \sum_{i=1}^{n} (F_\varepsilon(x_i) - y_i)^2, \min_{b\geq1} \sum_{i=1}^{n} (F_b(x_i) - y_i)^2, \min_{m\geq1} \sum_{i=1}^{n} (F_m(x_i) - y_i)^2$$

with $F_\varepsilon(x) = 1 - (1-x)^\varepsilon, F_b(x) = x^b$ and $F_m(x) = \frac{x}{m-(m-1)x}$.

None of the regression problems is solvable in closed form but each fitting function has a unique minimum. Approximate solutions are found by selecting a finite parameter domain, enumerating all fitting errors and selecting the parameter which attains the minimum error.

The regressions data stem from the so-called world development indicators as given in Worldbank (2001). Best fit values and corresponding errors are given in Table 1. The regression error is consistently minimized by the intermediate function type.

For parameter illustration, an individual with 40% of the population having a lower and 60% having a higher income is considered. For $\varepsilon = 0.5525$ (Canada), the income of the individual approximately is 55% of the average income of the population ranked above him. For parameter $b = 1.96$, the income of the individual approximately is $1.96 \cdot 100\% = 196\%$ of the average income of the population ranked below him. According to the intermediate functions, the income of the individual approximately is $1/(m-(m-1)\,x) \cdot 100\% = 1/(2.62 - 1.62 \cdot 0.4) \cdot 100\% \approx 51\%$ of the average of all larger incomes and it is $m/(m-(m-1)\,x) \cdot 100\% = 2.62/(2.62 - 1.62 \cdot 0.4) \cdot 100\% \approx 132\%$ of the average of all smaller incomes. While the difference in comparison to the upper segment is small (51% instead of 55%), the difference to the lower segment is significant (132% instead of 196%).

The functional Gini indices of the best fit curves are compared to the empirical Gini index as given by Worldbank (2001) in Table 2.

Bold face functional Gini indices indicate least absolute difference to the empirical Gini index. In 25 of the 30 cases, the minimum regression error curve is also best approximating the Gini index. In the remaining five cases, the best approximation of the Gini index is attained by functions of type F_b.

All empirical Gini indices exceed the corresponding functional Gini indices for functions of type F_ε. Lorenz curves of the other types either overestimate or underestimate the empirical Gini index.

5 Conclusion

Parametric Lorenz curves were derived from a "first principle" and empirically evaluated. Modified curves rather than the original curve stemming

Nation	ε	*err*	b	*err*	m	*err*
Austria	0.6493	.00378	1.61	.00417	2.01	.00032
Brazil	0.2778	.00902	4.65	.02053	7.53	.00095
Canada	0.5525	.00677	1.96	.00708	2.62	.00061
China	0.4464	.00842	2.57	.01299	3.70	.00065
Czech Rep.	0.6173	.00206	1.69	.00878	2.17	.00139
Denmark	0.6289	.00423	1.66	.00486	2.11	.00070
Finland	0.6135	.00301	1.71	.00729	2.20	.00069
France	0.5376	.00643	2.02	.00836	2.73	.00085
Germany	0.5882	.00441	1.81	.00719	2.36	.00058
Gr. Britain	0.5025	.00693	2.20	.01046	3.06	.00079
Greece	0.5376	.00586	2.02	.00908	2.74	.00076
Hungary	0.5225	.00371	1.94	.01084	2.61	.00084
India	0.4673	.00897	2.31	.02899	3.36	.00661
Italy	0.5988	.00516	1.77	.00569	2.30	.00063
Japan	0.6211	.00209	1.67	.00826	2.15	.00093
Korean Rep.	0.5525	.00627	1.96	.00761	2.63	.00079
Mexico	0.3279	.00668	3.77	.02381	5.91	.00185
Netherlands	0.5405	.00618	2.00	.00848	2.71	.00091
Nigeria	0.3546	.00496	3.40	.02613	5.26	.00273
Norway	0.6211	.00328	1.69	.00649	2.16	.00063
Poland	0.5319	.00449	2.03	.01138	2.77	.00130
Portugal	0.5000	.00439	2.20	.01425	3.07	.00140
Russian Fed.	0.3731	.00584	3.18	.02304	4.85	.00235
S. Africa	0.2809	.01070	4.64	.01784	7.42	.00015
Slovakia	0.6897	.00206	1.48	.00399	1.80	.00043
Spain	0.5405	.00584	2.01	.00895	2.72	.00087
Sweden	0.6289	.00481	1.68	.00443	2.13	.00042
Switzerland	0.5376	.00693	2.03	.00803	2.74	.00098
USA	0.4673	.01197	2.43	.00813	3.44	.00100
Venezuela	0.3788	.00942	3.15	.01694	4.77	.00130

Table 1. Best fit values and corresponding errors

from the EU poverty notion, gave best fitting results. Thus, normative and descriptive inequity differ significantly.

Parameters of all considered Lorenz curves are constants while extended theory suggests that they may differ with population segments or that significantly better fits can be obtained by "slightly varying" equity functions.

Nation	functional Gini indices			empirical Gini
	F_ε	F_b	F_m	index
Austria	.1063	**.1169**	.1145	.116
Brazil	.2826	.3230	**.2966**	.300
Canada	.1441	.1622	**.1557**	.158
China	.1913	.2199	**.2063**	.202
Czech Rep.	.1183	.1283	**.1266**	.127
Denmark	.1139	**.1241**	.1222	.124
Finland	.1198	.1309	**.1287**	.128
France	.1503	.1689	**.1619**	.164
Germany	.1296	**.1441**	.1397	.150
Gr. Britain	.1656	.1875	**.1789**	.181
Greece	.1503	.1689	**.1625**	.164
Hungary	.1441	.1599	**.1551**	.154
India	.1815	.1979	**.1926**	.189
Italy	.1255	.1389	**.1357**	.137
Japan	.1168	.1255	**.1251**	.125
Korean Rep.	.1441	.1621	**.1563**	.158
Mexico	.2531	.2904	**.2681**	.269
Netherlands	.1491	.1667	**.1608**	.163
Nigeria	.2382	.2727	**.2536**	.253
Norway	.1168	**.1283**	.1259	.129
Poland	.1528	.1699	**.1641**	.165
Portugal	.1667	.1875	**.1794**	.178
Russian Fed.	.2282	.2607	**.2431**	.244
S. Africa	.2807	.3227	**.2949**	.297
Slovakia	.0918	.0968	**.0969**	.098
Spain	.1491	.1678	**.1614**	.163
Sweden	.1139	.1269	**.1237**	.125
Switzerland	.1503	.1699	**.1625**	.166
USA	.1815	**.2085**	.1959	.204
Venezuela	.2253	.2590	**.2409**	.244

Table 2. Functional Gini indices and empirical Gini index from Worldbank (2001)

References

AABERGE, R. (2000): Characterizations of Lorenz curves and income distributions. *Social Choice and Welfare, 17, 639–653.*

BOUCHARD, J.-P. and MEZARD, M. (2000): Wealth condensation in a simple model of economy. *Physica A, 282, 536ff.*

CHEONG, K.S. (2002): An empirical comparison of alternative functional forms for the Lorenz curve. *Applied Economics Letters, 9, 171–176.*

CHOTIKAPANICH, D. and GRIFFITHS, W.E. (1999): Estimating Lorenz curves using a Dirichlet Distribution. Working paper 110, University of New England, Armidale.

DATT, G. (1998): Computational tools for poverty measurement and analysis. Discussion paper 50, International Food Policy Research Institute, Washington.

EUROPEAN PARLIAMENT (1999): The fight against poverty in the ACP countries and in the European Union. Working Document, Strassbourg.

FINLAND (2000): Ministry of Social Affairs and Health. Finland, Internet presentation. www.stm.fi/english/tao/publicat/poverty/definit.htm.

FOSTER, J.E., GREER, J., and THORBECKE, E. (1984): A class of decomposable poverty measures. *Econometrica, 52, 761–766.*

GINTHER, D.K. (1995): A nonparametric analysis of the US earnings distributions. Discussion Paper 1067-95, University of Wisconsin, Madison.

KÄMPKE, T. (2002): Overconstrained ordinary differential equations. *International Mathematical Journal, 2, 1125–1139.*

KÄMPKE, T., PESTEL, R., and RADERMACHER, F.J. (2002): A computational concept for normative equity. *European Journal of Law and Economics,15, 129–163.*

LITCHFIELD, J.A. (1999): Inequality: methods and tools. Discussion paper, Worldbank, Washington.

MÖLLER, U., RADERMACHER, F.J., RIEGLER, J., SOEKADAR, S.R., and SPIEGEL. P. (2004): *Global Marshall Plan. Mit einem Planetary Contract für eine Ökosoziale Marktwirtschaft weltweit Frieden, Freiheit und nachhaltigen Wohlstand ermöglichen.* Horizonte, Stuttgart.

RADERMACHER, F.J. (2001): Balance or Destruction: Ein Plädoyer für eine weltweite öko-soziale Marktwirtschaft. Ökoeffizienz, weltweiter sozialer Ausgleich und geordnete weltweite Wachstumsprozesse als Schlüssel zu einer nachhaltigen Entwicklung. Proceedings *Nachhaltigkeit als Geschäftsfeld, Natur, Macht, Märkte, Wuppertal.*

RADERMACHER, F.J. (2002a): 10 $\rightsquigarrow$ 4:34 – Die Formel für Wachstum und Gerechtigkeit. *Bild der Wissenschaft, 4, 78–86.*

RADERMACHER, F.J. (2002b): Balance oder Zerstörung – Ökosoziale Marktwirtschaft als Schlüssel zu einer nachhaltigen Entwicklung. *Ökosoziales Forum Österreich, Wien.*

RADERMACHER, F.J. (2004a): *Global Marshall Plan.* Global Marshall Plan Initiative, Hamburg.

RADERMACHER, F.J. (2004b): Ökosoziale Grundlagen für Nachhaltigkeitspfade. *GAIA, 170–175.*

RADERMACHER, F.J. (2004c): Was macht Gesellschaften reich? In: R. Loske and R. Schaffer (Eds.): *Infrastruktur, Metropolis* (to appear).

RASCHE, R.H., GAFFNEY, J., KOO, A.Y.C., and OBST, N. (1980): Functional forms for estimating the Lorenz curve. *Econometrica, 48, 1061–1062.*

RYU, H.K. and SLOTJE, D.J. (1999): Parametric approximations of the Lorenz curve. In: J. Silber (Ed.): *Handbook of income inequality measurement.* Kluwer, Boston, 291–312.

SALA-I-MARTIN, X. (2002a): The disturbing "rise" of global income inequality. Working paper 8904, National Bureau of Economic Research, Cambridge, MA.

SALA-I-MARTIN, X. (2002b): The world distribution of income. Working paper 8933, National Bureau of Economic Research, Cambridge, MA.

SILBER, J. (1999): *Handbook of income inequality measurement.* Kluwer, Boston.

WORLDBANK (2001): World Development Report 2000/2001. Worldbank, Washington.

A Multidimensional Approach to Country of Origin Effects in the Automobile Market

Michael Löffler and Ulrich Lutz

Dr. Ing. h.c. F. Porsche AG, Porschestr. 15-19, D-71634 Ludwigsburg, Germany

Abstract. Research on Country of Origin effects has a long tradition. In the course of time, an increasing number of new aspects and additional dimensions have been taken into account. With regard to an international brand management it is here of interest whether and, if applicable, how specific country images vary from one target market to another. This paper contributes to answering this question by using the automobile market as an example and draws conclusions which involve managerial implications.

1 Introduction

Consumer behaviour has changed significantly in recent decades. Demographic changes and advances in major consumer-related technological areas, including Internet and e-Business, have resulted in new patterns of consumer behaviour. Low-dimensional and classical types of consumer segmentation are becoming less meaningful. Purchasing patterns are increasingly determined by today's hybrid and multi-optional consumers in an international context. Country-specific brand images and country of origin effects are key topics in this area.

The following article contributes to this field of research with a focus on the multi-facetted character of country of origin effects in the automotive sector. A short overview highlights the increasing multi-dimensionality of this topic in the literature and its relevance to management as well. Own empirical results extend already existing findings and delineate managerial implications.

2 Research on country of origin effects

Country of origin effects are an area of interest under constant research, yet the research topics themselves are becoming more and more differentiated. The following section reviews in brief the main steps of this development:

Early contributions deal with the general implications of "made in"-labels on purchasing behaviour. Also, explanation factors of "made in"-labels were researched in depth from a behavioural science point of view. For an early overview on literature on country of origin effects see Bilkey and Nes (1982).

Subsequent research activities have lead to a more differentiated view on the country of origin phenomenon: A substantial amount of research has contributed to the differentiated evaluation of country of origin effects with respect to specific product categories. Successful product-country combinations have been scrutinized in more detail, including combinations like German automobiles, Swiss watches, French perfume, Italian fashion, Japanese consumer electronics, etc. For an overview on literature and its results on successful product-country combinations see Lutz (1994). More recent publications about selected product and service areas refer, for example, to automobiles, specific consumer goods, cruise lines or cover several product categories at a time (Haubl (1996), Knight (1999), Ahmed et al. (2002), Balabanis and Diamantopoulos (2004)).

Further research activities have shown an increasing differentiation with regard to country of origin effects. The product evaluation based on country of origin in general has been supplemented by research on individual product attributes, like technical quality, design, etc. The country of origin effect as a complete construction was disaggregated as a composition of many overlapping perceptions, see e.g. Shimp et al. (1993) or Chao (1998).

Furthermore, different approaches were chosen in order to break down the phenomenon of country of origin effects into various dimensions. Dimensions which were scrutinized in this context include (e.g. Gaul et al. (1994), Gürhan-Canli and Maheswaran (2000), Watson and Wright (2000), Klein (2002), Parameswaran and Pisharodi (2002), Löffler (2005)):

- Different target segments,
- particular market areas,
- cultural influences within a region,
- influence of domestic versus non-domestic origin in particular,
- consideration of theories from behavioural science, and also
- the development of images in the course of time.

Parallel to this, from the practical marketing point of view, an increasing globalisation and internationalisation in brand management has taken place. Local or national brand management is increasingly determined centrally and standardised by globally acting companies in particular. Only thus can an internationally consistent brand image be developed and continually improved. As a result, the combination of country of origin with individual target markets is considered as an additional dimension in this field of analysis. Marketing research repeatedly leads us to the conclusion that country of origin effects with regard to individual product attributes should be evaluated against the background of consumer predispositions in the target country. Thus, country of origin effects on the level of product attributes overlap with country-specific evaluation patterns from a consumer point of view.

This paper expands on existing findings in the automobile market and indicates corresponding implications for management. Brand perception was analysed on the basis of individual image dimensions with respect to important continental European countries and regarding the most important automobile makes. Market-specific differences of perception were then established.

3 Methodology and data

Consumers mostly base purchase decisions of automobiles on a few key characteristics, which reflect beyond the transaction price in particular aspects of overall car quality as well as emotional aspects. When discussing the evaluation of brands on a disaggregated level it is necessary to separate the influences of these two dimensions. For convenience, in the following, these dimensions are referred to as "Quality" and "Emotions". Figure 1 depicts various situations of brand image perceptions in domestic versus foreign markets.

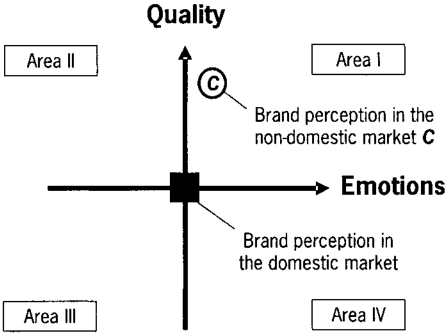

Quality
Area II
Area I
C
Brand perception in the non-domestic market C
Emotions
Brand perception in the domestic market
Area III
Area IV

Fig. 1. Brand perception of brands in their country of origin and in a non-domestic market.

Taking the brand positioning in the domestic market as neutral point, four different areas for consumer specific perceptions should be differentiated: Area I indicates the perception of a brand facing an overall positioning advantage in the foreign market: The perception of product attributes related to "Emotions" (for example design, sportiness) is above the corresponding perception in the domestic market. In addition the perception of product attributes related to "Quality" (for example high reliability, good craftsmanship, advanced technology) go beyond the perception in the domestic market. Therefore, on an aggregated level the brand has an overall positioning advantage in the foreign market compared to its domestic market. Area III indicates the situation where an automotive brand is perceived less favourably with respect to both dimensions "Quality" and "Emotions" in the foreign market. A mixed brand image perception can be found in the Areas II and IV. The brand is judged superior with respect to one dimension in the foreign market, but inferior with respect to the other one.

There are relatively few empirical findings with regard to differences in perception of automobile brands on non-domestic markets in comparison with their domestic market. In particular there are very few results regarding the question to what extent it can be inferred, that combined effects of the brand dimensions "Quality" and "Emotions" valid in one market are also relevant in another market, and whether these brand dimensions, despite possible contradictory effects, result in a total brand perception which deviates from the brand image in the domestic market: For example, perception advantages in the dimension "Quality" are partly or completely compensated by positioning advantages regarding the dimension "Emotions". Therefore, it is analysed empirically,

- if such differences of perception exist despite the fact that the automobile industry functions globally in general in todays transparent continental European market
- and how any given differences in perception affect the individual dimensions "Quality" and "Emotions".

Moreover, not only from a marketing management perspective it seems necessary to differentiate these effects for individual automobile segments (see Chao and Gupta (1995)). In particular the affordability of an automobile make and, thus, the general exclusivity of a brand are significant factors which have to be considered in evaluation of country of origin effects. The empirical analysis of the effects described above was therefore carried out separately for the low/medium price segment and the luxury automobile segment. For a comprehensive description of the hypothesis used for the analysis see Löffler (2001). Factor and Variance analysis were applied.

Data cover 13 European countries including Germany, France, Italy and Spain. In all countries the same method of a self- administered survey was used. For the German part of the survey Figure 2 illustrates the high congruence between the actual registrations of brands and cars driven by respondents. In order to guarantee confidentiality, letters were randomly assigned to the different brands.

Within the survey participants had to indicate makes that best fit statements like "good looks/styling", "makes sporty cars", "very reliable cars" and "well made". It was possible to name more than one make. A comprehensive listing of all nameplates being offered in the countries mentioned above was part of the questionnaire.

The ordering of the brands was altered in order to avoid biasing effects. A total of 383,000 readers of automobile magazines participated in the survey.

4 Empirical results

Data were carefully checked before being analysed by factor analysis: based on the large sample sizes it is adequate to assume the data to be normally distributed and therefore key requirements for Bartlett's test of sphericity are

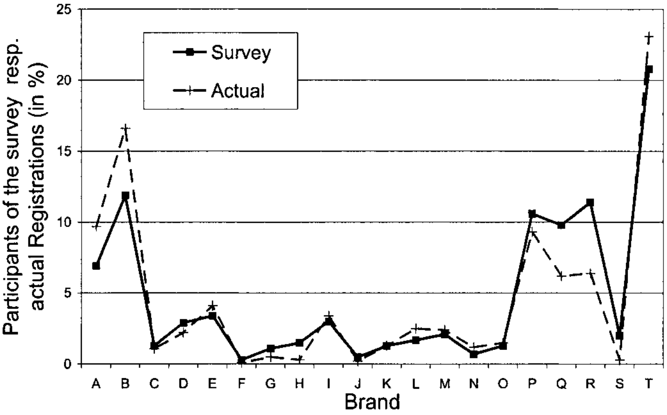


Fig. 2. Correspondence between total registrations and makes driven by survey participants (Germany).

met. For the present data set of the low/medium car segment, Bartlett's test statistic $B = 190,03$ is highly significant ($p < 0.001$); similar results hold true for the luxury segment. Following Basilevsky (1994) the correlation matrix is not orthogonal. The anti-image covariance matrix consists of off-diagonal elements, which do not indicate any meaningful deviation from diagonal matrix. The Kaiser-Meyer-Olkin measure of sampling adequacy MSA indicated appropriateness for factor analysis. For the low/medium as well as the luxury segment factor analysis resulted in a two factor solution. For both segments a scree plot as well as selection on Kaiser eigenvalue criterion came to a number of two factors. The rotated factor scores as well as the eigenvalues are summarized in Table 1. The factors explained 89.3% of variance for the low/medium segment and 88.4% of variance for the luxury segment.

For the low/medium segment variables 1-3 clearly correspond to the first factor consisting mainly of emotional perceptions like "good looks, styling", "I like this make". For factor 2, high scores were found at variables 5 and 6. Both variables reflect the cognitive components of brand perception.

An assignment of variable "advanced technology" failed. This might be due to the fact that drivers may have aspects like reduced fuel consumption (cognitive components) in mind as well as sporty acceleration figures (affective components).

Based on their individual factor values brands were positioned within the axes "Emotions" (factor 1) and "Quality" (factor 2).

4.1 Results for the low/medium segment

Even if the results for the individual automobile brands were different, country-related perception patterns can be established which are valid for several au-

		Segment "low/medium"		Segment "luxury"	
Nr.	Variable	Factor 1 ("Emotions")	Factor 2 ("Quality")	Factor 1 ("Emotions")	Factor 2 ("Quality")
1	good looks/styling	**0.97003**	0.07605	**0.92981**	0.05083
2	makes sporty cars	**0.94216**	-0.19112	**0.92342**	0.27564
3	I like this make	**0.81954**	0.43393	**0.92023**	0.20000
4	advanced technology	0.70457	0.56196	0.08557	**0.92884**
5	very reliable cars	0.03831	**0.96702**	n.a.	n.a.
6	well made	0.08632	**0.93343**	-0.07211	**0.92726**
	eigenvalue 1	3.51		2.58	
	eigenvalue 2	1.85		1.84	

Table 1. Factor scores and eigenvalues.

tomobile brands. These more general perception patterns can be summarised as follows:

Hypothesis testing based on *t*-tests as well as Wilcoxon matched-pairs signed-rank test provided the following results: In France, Germany, Italy and Spain foreign automotive brands have a positioning disadvantage with respect to quality-related product attributes when lacking the "domestic make"-label and compared with their domestic judgements. "Quality" is perceived lower in foreign markets (see Figure 3).

In Germany, the lack of the "domestic make"-label results in positioning advantages with regards to the affective brand perception ("Emotions"). Foreign brands are in general judged to be more vivid, fresh and fascinating, compared to their perception in the corresponding domestic markets. The results of mixed perceptions in foreign markets extend findings of Knight (1999) and Bilkey and Nes (1982). They summarized parts of their findings that consumers tend to evaluate their own country's products more favourably. According to our results, in Germany perceptions of foreign brands are a mixture of positive as well as negative country of origin effects.

In Italy and Spain foreign automotive brands face overall positioning disadvantages. Non-domestic makes are judged to be poorer quality as well as have less sporty brand characteristics compared with the evaluation of these brands in their domestic markets.

4.2 Results for the luxury segment

The brands in the luxury segment have in general a more differentiated perception than brands in the low/medium segment. From a management perspective, a generalization of the results is less adequate than for the low/medium segment. We therefore report on main patterns detected: For hypothesis testing, only German luxury brands were available; sporty premium cars could not be included due to an insufficient statistical basis.

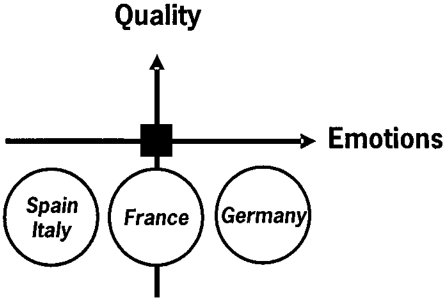


Fig. 3. Perception of brands in foreign markets Spain, Italy, France, Germany with reference to their perception in the domestic market (low/medium segment).

The findings were mixed: In some European countries the quality of German brands is perceived to be higher than in Germany, not in others. In general, luxury brands are perceived as less emotional than in their domestic market Germany.

5 Conclusions

The paper contrasts brand perceptions in foreign against domestic markets. Several nameplates in the low/medium as well as luxury segment were examined. In general, consumers in France, Germany, Italy and Spain tend to judge the quality of foreign makes less favourably compared to those on the domestic markets. The results indicate differentiated effects for Germany, Italy and Spain in the perception of emotional brand attributes of non-domestic nameplates.

Even in a Europe which is growing together brands are evaluated differently in the various countries. German cars are perceived to be of higher quality, but often they are seen as less emotional and fascinating. Possible managerial implications, with respect to marketing communication, indicate that advertisements for German makes should highlight emotional factors, whereas advertisements for foreign makes should stress the quality related image dimensions more. A brand with an undoubtedly sporty image would clearly gain when manufactured or assembled in Germany. A perfect combination of a sporty image and the country of origin effect of a German production location ("High Quality") leads to advantages in overall brand and product perception.

In Italy and Spain, in the low/medium segment, foreign brands have a positioning disadvantage with respect to both dimensions "Quality" and "Emotions". Therefore, non-domestic brands should strengthen their perceived image with regard to both dimensions: Foreign brands have to convince consumers in Italy and Spain concerning product quality as well as sportiness and good design of the car.

References

AHMED, Z.U., JOHNSON, J.P., LING, C.P., FANG, T.W., and HUI, A.K. (2002): Country-of-Origin and Brand Effects on Consumers' Evaluations of Cruise Lines. *International Marketing Review, 19, 3, 279–302.*

BALABANIS, G. and DIAMANTOPOULOS, A. (2004): Domestic Country Bias, Country-of-Origin Effects, and Consumer Ethnocentrism: A Multidimensional Unfolding Approach. *Journal of the Academy of Marketing Science, 32, 1, 80–95.*

BASILEVSKY, A. (1994): *Statistical Factor Analysis and Related Methods*, Wiley, New York, NY.

BILKEY, W.J. and NES, E. (1982): Country-of-Origin Effects on Product Evaluation. *Journal of International Business Studies, 13, Spring/Summer, 89-99.*

CHAO, P. (1998): Impact of Country of Origin Dimensions on Product Quality and Design Quality Perceptions. *Journal of Business Research, 42, 1–6.*

CHAO, P. and GUPTA, P. (1995): Information Search and Efficiency of Consumer Choices of New Cars. *International Marketing Review, 12, 6, 47–59.*

GAUL, W., LUTZ, U., and AUST, W. (1994): Goodwill Towards Domestic Products as Segmentation Criterion: An Empirical Study within the Scope of Research on Country of Origin-Effects. In: H.H. Bock, W. Lenski, and M.M. Richter (Eds.): *Information Systems and Data Analysis. Prospects – Foundations – Applications.* Springer, Heidelberg, 415–424.

GÜRHAN-CANLI, Z. and MAHESWARAN, D. (2000): Cultural Variations in Country of Origin Effects. *Journal of Marketing Research, 37, 3, 309–318.*

HAUBL, G. (1996): A Cross-national Investigation on Effects of Country of Origin and Brand Name on the Evaluation of a New Car. *International Marketing Review, 13, 5, 76–97.*

KLEIN, J.G. (2002): Us versus Them, or Us versus Everyone? Delineating Consumer Aversion to Foreign Goods. *Journal of International Business Studies, 33, 2, 345–363.*

KNIGHT, G. (1999): Consumer Preferences for Foreign and Domestic Products. *Journal of Consumer Marketing, 16, 2, 151–162.*

LÖFFLER, M. (2001): A Multinational Examination of the "(Non-) Domestic Product" Effect. *International Marketing Review, 19, 5, 482–498.*

LÖFFLER, M. (2005): Automobilimages im Dekadenvergleich. To appear in: *ZfbF - Zeitschrift für betriebswirtschaftliche Forschung und Praxis.*

LUTZ, U. (1994): *Preispolitik im Internationalen Marketing und westeuropäische Integration.* Lang, Fankfurt.

PARAMESWARAN, R. and PISHARODI, R.M. (2002): Assimilation Effects in Country Image Research. *International Marketing Review, 19, 3, 259–278.*

SHIMP, T., SAMIEE, S., and MADDEN, T. (1993): Countries and their Products: A Cognitive Structure Perspective. *Journal of the Academy of Marketing Science, 21, 4, 323–330.*

WATSON, J.J. and WRIGHT, K. (2000): Consumer Ethnocentrism and Attitudes Toward Domestic and Foreign Products. *European Journal of Marketing, 34, 9/10, 1149–1166.*

Loyalty Programs and Their Impact on Repeat Purchase Behaviour: An Extension on the "Single Source" Panel *BehaviorScan*[⋆]

Lars Meyer-Waarden

Université Toulouse Paul Sabatier - IUT GEA Ponsan,
115 Route de Narbonne, 31077 Toulouse Cedex, France
email: lars.meyer-waarden@iut-tlse3.fr

Abstract. The purpose of this research is to contribute to a better theoretical knowledge about the sources of efficiency in loyalty programs, in the retail sector. It is based on the BehaviorScan single-source panel which has been crossed with the store data base of a French retailer. We implemented the multinomial Dirichlet model, in order to test the impact of loyalty programs on the general market structure. The double jeopardy phenomenon is present and loyalty programs do not substantially change market structures. When all companies have loyalty programs, the market is characterized by an absence of change of the competitive situation.

1 Introduction

Loyalty programs, in particular those employed by distributors, are currently regarded as fundamental by many companies. They lie within the scope of rather defensive customer retention strategies, being based on the double conviction that retaining customers is less expensive than conquering new ones, and that the best customers are the most profitable (Bolton and Drew (1994), Reichheld (1996)). Despite the practitioners' strong interest in loyalty programs, there is only scarce empirical evidence about their potential impacts. While anecdotal evidence seems to be plentiful, certain academic authors worry about their effectiveness (Nako (1997), Dowling and Uncles (1997), Sharp and Sharp (1997), Benavent et al. (2000), Meyer-Waarden (2002, 2004), Mägi (2003)). This report is all the more surprising as many companies developed them during previous years. Indeed, if one looks at the retail sector in Europe, the costs associated with the management of loyalty cards were estimated in 1999 at 2.5 billion dollars for 350 million emitted cards (Wall Street Journal June 19th 2000). This is why, English retailers, such as *Safeway*, decided to give up their loyalty programs. Indeed, *Safeway* considers savings 75 million $ per annum. However, other retailers, such as E. Leclerc in France, still reinforce their marketing expenditures by devoting

[⋆] The author would like to express his warmest thanks to *MarketingScan* company and a French retailer wishing to remain anonymous for kindly providing their data.

approximately 18 million of their marketing budget to the animation and management of their program.

Consequently, there seems to be a need for more rigorous empirical evidence on the efficiency of loyalty programs. We shall therefore make explore the impact of four retailing loyalty programs situated in the *BehaviorScan* test market in *Angers* (France) on repeat purchase behaviour.

2 Theoretical background

Loyalty programs absorb considerable resources and it should be noticed that in order to pay off those tremendous costs, most managers are probably going to consider a program as efficient if it increases sales, penetration and market share. In this context, a major issue arises. In saturated and stationary markets most marketing mix tools - including loyalty programs - do not raise well established brands' market shares. As Ehrenberg (1997) notices, they only maintain their positions in a defensive perspective, which is in opposition with the managers' viewpoints. This stationarity and the lack of effectiveness is certainly due to the fact that in a competitive market, the initiator of such campaigns will certainly be imitated, so that the total result will be a return to the former situation and will consist in an increase in marketing costs.

Considering these issues, Sharp and Sharp (1997) suggest to complete the measurement of loyalty programs' efficiency by using loyalty indicators, in particular the repeat purchase rate (*purchase loyalty*), and/or the decrease in the sensibility for competitors' offers (*differentiation loyalty*). In this paper, we shall rather focus on the impact of loyalty programs on repeat purchase behaviour (*purchase loyalty*, i.e. average purchase frequency, penetration, share of requirement) for several reasons - First of all, these judgement criteria correspond to their expectations in terms of expected results. Secondly, in most markets, (particularly those more characterized by imitation than differentiation), it is slightly likely that a decrease in the sensibility for competitors' offers may occur without being followed by a rise in repeat purchase behaviour.

Loyalty programs' impact on the "normal" Dirichlet market

If a number of individual customers do change their behaviour in a similar way, one can imagine that changes in aggregated market structures, that is to say store levels, may occur. In concrete terms, we might suppose that certain stores with a loyalty program succeed in creating a niche position by possessing subgroups of customers, who are probably card holders and are more prone to loyalty which would be exemplified by higher repeat purchase rates (Kahn et al. (1988)).

In order to test this idea, we shall use the Dirichlet model which became famous in marketing by Goodhardt et al. (1984) thanks to its simplicity and some empirical regularities providing theoretical benchmarks.

The Dirichlet model incorporates the assumption that markets are stationary (unaffected by the marketing mix), that consumers' purchase patterns remain stable but it does not in any way imply that their purchases are identical from one period to another. It is a zero-order static stochastic model which is composed of four probability distributions used to specify individual purchasing structures (Poisson distribution, Gamma distribution, zero-order multinomial distribution, multivariate Beta or Dirichlet multinomial distribution). According to repeat purchase theory, purchasers demonstrate regular tendencies in their purchase behaviour which vary throughout the population according to these distributions.

The Dirichlet methodology thus estimates theoretical "norms" for evaluating the performance of brands/stores with indicators as average purchase/visit frequency, share of category requirement (SOR), share of sole buyers with a given penetration level in the category, which represents the market in a stationary situation, i.e. unaffected by the marketing mix (loyalty cards for the purposes of our study). Discrepancies between these theoretical values and actual observed values will show how the marketing mix can "disturb" the stationary market (Ehrenberg et al. (2004)), making the model very useful for understanding repeat purchase patterns.

Empirical generalizations of the Dirichlet model

The empirical regularities outlined and described by the NBD/Dirichlet model have been observed by Ehrenberg and his colleagues (Ehrenberg et al. (2004)) for over 50 product categories (from cosmetics and foodstuffs to cars, as well as store selection) in a large number of different sectors and countries (Europe, the USA, Asia, and Australia). They are regarded as fundamental assumptions for purchasers' behaviour (double jeopardy effect, purchase duplication law, etc.). They will be explained hereafter and used to establish our research hypotheses and comparison norms.

The double jeopardy effect

Ehrenberg (1988) highlighted a fundamental concept, the double jeopardy phenomenon, which can be found in a number of markets. To a great extent, the success of brands with high market shares is due to the fact that they have more customers (higher penetration) than their smaller competitors. The slight variations in loyalty levels between brands also follow the double jeopardy principle. Thus, the higher the penetration of a store (or brand) is, the higher the visit frequency (or purchase frequency) as well as the share of category requirement and vice versa. More specifically, brands or stores which are less popular receive a double penalty, since they not only attract fewer customers, they are also bought or visited less frequently by their customers. The "double jeopardy" line in the following diagram (Figure 1) shows that for brands or retail outlets in the same category an increase in penetration leads to a constant increase in average purchase (visit) frequency. It thus suggests that most marketing instruments, such as sales promotion for example, mainly affect market share, followed by penetration, i.e. recruiting new purchasers (Ehrenberg et al. (1990)).

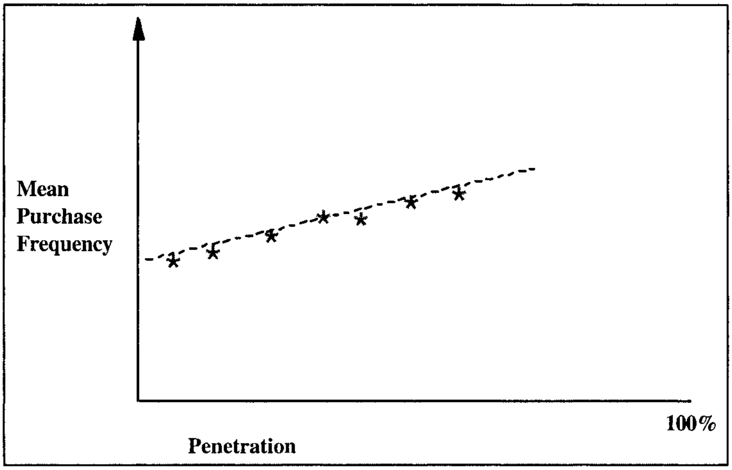


Fig. 1. Double jeopardy effect

On the other hand, loyalty programmes should provoke a different type of effect, linked to the long-term oriented, cumulative nature of rewards which should thus provide the company's loyal and heavy purchasers with more benefits (i.e. probability of receiving a reward, value of rewards) and less costs (i.e. changing purchase behaviour, concentrating purchasing at one retail outlet, reducing variety-seeking, increased switching costs, etc.). It is thus likely that such programmes have a greater effect on the visit frequency of existing customers than on penetration. Indeed, since rewards are accumulated on a long-term basis they may put off promotion-hunters who would thus reject a company with a loyalty scheme in favour of retail outlets which continue to use a promotion policy with immediate bonuses. This could even lead to a decline in the store's penetration, which does not mean that the outlet is losing customers, although this may be the case.

Our research hypothesis can be formulated on the basis of these conclusions. It refers to the manner in which the loyalty programme may "disturb" the "normal" Dirichlet market. Thus we do not expect this company to undergo a linear "shift" along the "double jeopardy" line, but rather a vertical one, characterized by Sharp and Sharp (1997) as "excessive loyalty" (see Figure 2).

The store would then be situated above the double jeopardy line and it has gained "too much loyalty" (i.e. average visit frequency, higher share

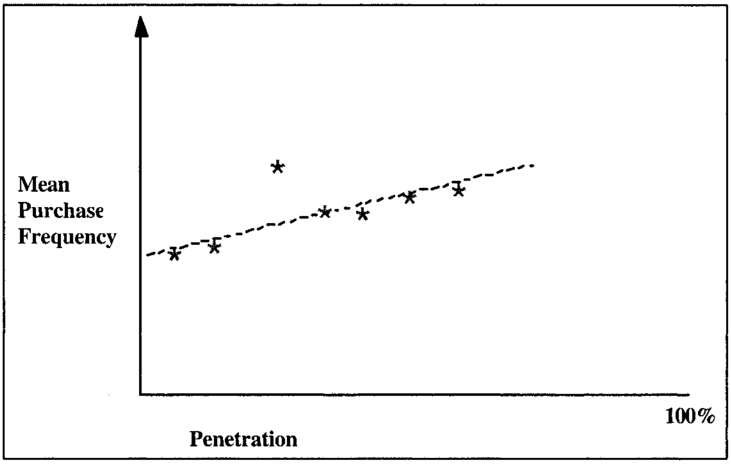


Fig. 2. Excess loyalty and double jeopardy effect

of category requirement) when compared against its penetration level. The company is frequented more often by the same consumers than predicted by the Dirichlet model and thus occupies a niche position (Kahn et al. (1988)). "Excessive loyalty" is therefore a deviation from one of the basic assumptions of Ehrenberg's empirical generalizations, that of non-segmented populations, which leads us to believe that certain sub-segments are more favourable to loyalty than others (Fader and Schmittlein (1993)). We can therefore formulate the following hypothesis:

For the population of loyalty cardholders, the purchase frequency, the share of category requirement, the percentage of 100% loyal buyers in the store with the loyalty programme is proportionally higher in comparison to their penetration than for the purchaser without loyalty card. Those stores are situated above the "double jeopardy" line and develop "excessive loyalty".

3 Methodology

In order to test our hypotheses, we shall resort to the "single source" *BehaviorScan* panel in *Angers*, France (five hypermarkets respectively named S1 to S5 and two supermarkets known as S6 and S7). These stores represent 95% of the consumer goods sales of the area. It is a closed test market, which allows to follow nearly all the purchases per customer in the seven stores over our period of time. We also have at our disposal some information about the possession of one of the four loyalty cards of the seven stores present in the market. The only retailer that has not got one is S6. S1 and S2 belong to the same retailing chain and the loyalty card can be used in both stores. S3 and S4 are as well part of the same retailing chain with a loyalty program valid in both outlets.

From this panel, we have extracted a total of 50.000 purchase acts coming from 2.476 individuals who are holders or non-holders of the four loyalty cards over a 24 week period (week 28/2000 to week 52/2000). A 28 week duration has been chosen because it corresponds to the lapses of time that Ehrenberg (1988) uses for the estimation of the Dirichlet model.

The procedure for calculating performance indicators for stores or brands requires complex arithmetical procedures. We will not embark on a detailed discussion here since this would render the article disproportionately lengthy and complex. Full details can be found in Annex C in Ehrenberg's book (1988). We use the "*DIRICHLET*" software, developed by Kearns et al. (1998), to estimate parameters on the basis of company penetration values and their visit frequency (aggregated to the level). It uses the marginal method of moments.

Theoretical values (visit frequency, share of category requirement, rate of sole buyers) estimated using the Dirichlet model are compared against actual values observed for each of the samples, which will enable us to examine the general adjustment of the model as well as detecting deviations which demonstrate how a marketing tool can "disturb" the market structure. Generally speaking, according to the Dirichlet modelling approach (Goodhardt et al. (1984)), stores are experiencing significant "excessive loyalty" when:

- observed penetration $\leq 3\%$ theoretical penetration,
- observed visit frequency ≥ 0.3 theoretical visit frequency,
- observed share of category requirement $\geq 3\%$ theoretical share of category requirement.

If the deviations are insignificant, the model is well adjusted, which means that the market is approximately stationary, characterized by repeat purchases. If there are significant deviations, they could result from a change in observation conditions (loyalty cards) which affects the market. In this case we might assume that loyalty cards create sub-segments of more loyal consumers. This would be a violation of the basic assumption for the Dirichlet distribution.

4 Results

The estimation of the Dirichlet model brings out the nearly universal double jeopardy phenomenon. Stores with the highest penetration (S4, S1, S3) also have higher purchase frequencies than those with low penetration rates (S5, S6, S7).

One can, however, notice significant deviations concerning three outlets (S2, S4 and S7) between the predicted theoretical values and those which can actually be observed. S2 and S4, both having a loyalty program, exhibit "excess loyalty". Are the loyalty cards responsible for this comfortable situation by catching and isolating individuals from competitors' actions ? The

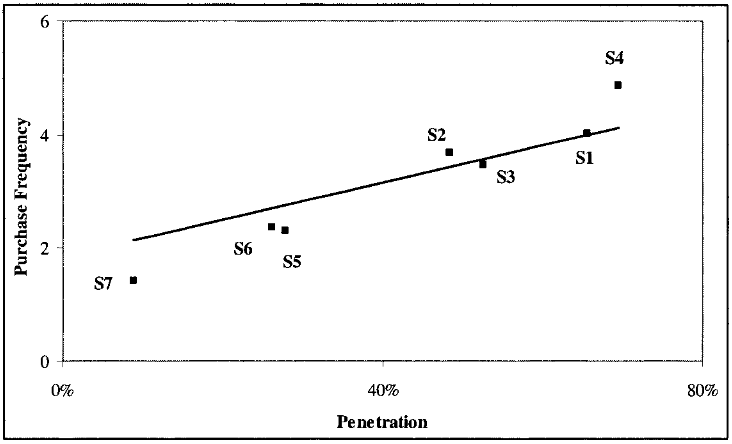


Fig. 3. The Double jeopardy line (2.476 holders and non-holders of the 4 loyalty cards

answer is not obvious because S1 has indeed the same program as S2 and S3 offers S4's card as well. This does not mean though that S1's and S3's clients develop a loyalty level proportionally superior to their penetration rates as they are in the Dirichlet norm. Nonetheless, S5 and S7 which as well possess loyalty programs, have got values under the norm as for visit frequency but above it as for penetration. This means that both stores proportionally have more clients than their sizes would theoretically allow, but these customers are not loyal and do not go shopping here in a regular way. Kahn et al. (1988) call this a "change-of-pace" situation. Finally, only S6 which does not have a loyalty program is perfectly in the norm. This can be explained by its location, very close to the town centre, and its small size. This outlet is probably perceived as a convenience store where to do one's every day shopping.

In order to test if S2's and S4's excess loyalty is linked to their respective loyalty programs, we shall again estimate the Dirichlet model, first only on the panelists holding at least one of the four available loyalty cards (N=1,646) and then on those having none (N=830). The results can be seen in Tables 1 and 2.

The category leader S4 possesses "excess loyalty" regarding both holders and non-holders of its loyalty card, since the respective observed purchase frequencies observed are higher than the theoretical values (+0.4 points for card holders and + 0.5 points for non-holders of the loyalty card which can be viewed as significant according to the Dirichlet norms), i.e. those which one would expect to find for a given penetration. The same is true for share of category requirement (+11 points for card holders and +8 points for non-holders of the loyalty card in comparison to the Dirichlet model forecasts). On the other hand, the penetration level observed for the retail outlet is lower than theoretical values (-6 points for card holders and -4 points for non-

Store	Penetration		Purchase Frequency		Share of Requirement		100% Loyal	
	O	T	O	T	O	T	O	T
S 1	69%	67%	4.2	4.4	15%	14%	1.7%	1.2%
S 4	66%	72%	5.3	4.9*	25%	14%*	1.5%	1.2%
S 3	55%	59%	3.9	3.7	17%	15%	1.5%	1.3%
S 2	37%	39%	2.8	2.7	10%	10%	0.7%	0.9%
S 5	37%	39%	2.6	2.7	9%	10%	0.9%	0.9%
S 6	26%	21%	1.8	2.2*	6%	7%	0.1%	0.7%
S 7	19%	19%	2.1	2.2	5%	6%	0.6%	0.6%

O = Observed Value, T = Theoretical Value,* : significant deviation from Dirichlet

Table 1. Results of the Dirichlet model (card holders all programs N=1.646, 24 weeks)

Store	Penetration		Purchase Frequency		Share of Requirement		100% Loyal	
	O	T	O	T	O	T	O	T
S 1	61%	59%	3.7	3.9	15%	14%	1.2%	1.2%
S 4	55%	59%	4.1	3.6*	22%	14%*	3.4%	1.2%
S 3	46%	42%	2.7	2.9	12%	10%	2.6%	0.9%
S 2	51%	63%	5.1	4.1*	21%	15%	3.0%	1.3%
S 5	25%	16%	2.1	2.3	7%	7%	0.3%	0.7%
S 6	34%	39%	3.3	2.9*	14%	10%*	1.4%	0.9%
S 7	7%	1%	1.6	1.8	5%	6%	0.1%	0.6%

O = Observed Value, T = Theoretical Value,* : significant deviation from Dirichlet

Table 2. Results of the Dirichlet model (non-card holders N=830, 24 weeks)

holders of the loyalty card which is again significant according to the Dirichlet norms). Moreover, S4 also has a greater number of 100%-loyal customers for loyalty card holders than the norm (+0.5).

On the other side, we can notice that stores with weaker penetration as well as market shares do not have "excess loyalty" for loyalty card holders. S1, S3, S5 and S7 observed values for purchase frequencies, share of category requirements are under theoretical values but within the Dirichlet norm. S6 is clearly disadvantaged.

Regarding non-holders of the loyalty card, S2 has got "excess loyalty". S1, S3, S5 as well as S7 are in the norm. Moreover, regarding the share of requirements and in particular the rates of sole-buyers, we can say that non-holders possess superior values to holders which lets us suppose that loyalty is not generated by the loyalty program (because S6 has not got one) but by other factors, such as competitive position, sale surface, proximity, comfort, product variety (i.e. S4), or the relative isolation from other stores (i.e. S2).

5 Conclusion

Our research has supported previous findings (Sharp and Sharp (1997)), and has given the data context and meaning. It suggests that it would be fruitful for grocery retailers to employ these generalizations in order to analyse the loyalty program's performance in light of these patterns.

Thus, our main finding is that without substantially modifying the competitive structure of the market (market shares, penetration, double jeopardy) loyalty programs' impact on visit frequency, purchase duplication and sole-loyalty is quite weak. Programs thus seem to be playing a defensive role, slightly reducing the scattering of purchases among several shops. This can be explained by the fact that thanks to the rewards cumulative characteristics over the long term, loyalty programs are mainly interesting for the heavy buyers of the retailing chain or the category (Meyer-Waarden (2002)) who in general have a broader purchase repertory, including the store in question. Thus there is a significant probability for these customers to buy in the store before subscribing to the program. Indeed, if we look at our sample, we can notice that the majority of cardholders have already been clients (88%) before subscribing to the card, which is comparable with the results found by Ehrenberg (1988). Loyalty cards therefore mainly attract existing customers and affect their visit frequency more than penetration (the recruiting of new clients).

One explanation of this could be linked to the fact that the long-term cumulative nature of rewards means that loyalty programmes mainly appeal to a company or possibly category's heaviest purchasers, who generally have a wider purchase portfolio which includes the store in question. Conversely, light purchasers in the category, for whom the likelihood of achieving a reward is low, are probably attracted to instant sales promotions offered by competitors, since loyalty cards are less beneficial to them. We can thus assume that the choice of retail outlet as well as loyalty are guided by elements other than the loyalty programme, in particular the competitive position, proximity, inertia, comfort, choice, product variety, store size, sales promotions and the store's relative isolation from other retail outlets (i.e. S2). Another explanation could, at least in part, be a valid justification for the lack of excessive loyalty demonstrated by S1: its loyalty programme is a multi-sponsor programme (since the loyalty card is valid in all the company's retail outlets and in a network of external partner stores). This implies that it is easier for the consumer to obtain points without any major changes to their purchase behaviour. Thus new, light purchasers may be attracted, which would partially cancel the aggregated effects of excessive loyalty, by counterbalancing the increase in existing customer's average purchase frequency through the recruitment of a significant number of new light purchasers, with lower purchase behaviour. Finally it should also be noted that S1 suffers from S4, the largest store in the area that is located alongside.

Another observation strengthens an interesting theoretical and managerial point of view: loyalty is an additional product of market share and penetration, and retention strategies seem to be consequently more efficient for market leaders and small firms are twice penalized (McGahan and Ghemawat (1994)). The most important implication for managers is therefore to be realistic as small retailers cannot be expected to have loyalty rates equal to those of their larger competitors.

An important theoretical and managerial contribution is the confirmation of the hypothesis of Dowling and Uncles (1997): in the case of the simultaneous presence of several loyalty programs, the market is characterized by stationarity. Indeed, there more seems to be an effect of imitation than innovation and the result of marketing actions is a certain stability, even a return to the previous situation, before the existence of loyalty programs. This observation leads us to question the efficiency of S6's recent loyalty program's launching (S6 indeed had not got yet a loyalty program during our period of observation).

However, some conditions (proximity, products offers) are more prone to attract individuals and to create excess loyalty with the help of a loyalty program. This phenomenon, underlines the defensive feature of loyalty programs and distinguishes them from the other marketing mix variables. It is therefore important to notice that these first conclusions should not though encourage to give up all loyalty programs because this could lead to a disadvantage in comparison with competitors. The keeping of loyalty systems is all the more necessary than that of other marketing tools, because, to quote Ehrenberg (1997): "*Marketers must work hard to stand still*".

This empirical investigation presents some limits and many questions remain open which would enable our work to be developed further. We shall first recommend a certain level of prudence with regard to the external validity of our findings, coming from a test market in a banal food distribution context. Furthermore it should be underlined that our analysis focuses on repetitive purchase behaviour alone and does not integrate financial data. More research and replications in other business domains are therefore necessary to respond to this issue even if Harris (2003) found similar results in the airline sector. Purchase behaviour analysis has shown that loyalty cards result in very little behavioural changes. Is it the lack of individualized management of loyalty systems which means that loyalty cards have little impact on purchase behaviour? It is likely, as demonstrated in literature on sales promotion (Chintagunta et al. (1991)), that certain segments of consumers are more likely to be influenced by loyalty schemes than others, since this is linked to customer heterogeneity phenomena. Individualized reward systems therefore have a central role to play. Though experimental research exists on this issue, no evidence from field data is available. Therefore, such a comparison of the effectiveness of loyalty programmes' designs should be done. It is thus clear that studies in the area of encouraging loyalty and specific programmes are rare and incomplete, since the majority of studies are norm-referenced and not validated empirically. There is therefore no lack of scope for investigation and many questions of a diverse nature remain open to exploration.

References

BENAVENT, C., CRIÉ, D., and MEYER-WAARDEN, L. (2000): Analysis of the Efficiency of Loyalty Programs. *The 3rd AFM French-German Conference about Retailing and Distribution in Europe, St. Malo, June, 120–135.*

BOLTON, R.N. and DREW, J.H. (1994): Linking Customer Satisfaction to Services Operations and Outcomes. In: R.T. Rust and R.L. Oliver (Eds.): *Service Quality: New Directions in Theory and Practice.* Thousand Oaks, Sage, CA, 173–200.

CHINTAGUNTA, P., JAIN, D., and VILCASSIM, N. (1991): Investigating heterogeneity in brand preferences in logit models for panel data. *Journal of Marketing Research, 28, 4, 417–28.*

DOWLING, G.R. and UNCLES, M. (1997): Do Customer Loyalty Programs Really Work? *Sloan Management Review, Summer, 71–82.*

EHRENBERG, A.S.C. (1988): *Repeat Buying, Facts, Theory and Applications.* Charles Griffin & Company Limited, London.

EHRENBERG, A.S.C. (1997): Description and Prescription. *Journal of Advertising Research, Nov/Dec, 17–22.*

EHRENBERG, A.S.C., GOODHARDT, G.J., and BARWISE, T. P. (1990): Double Jeopardy Revisited. *Journal of Marketing 54(July), 82–91.*

EHRENBERG, A.S.C., UNCLES, M., and GOODHARDT, G. (2004): Understanding brand performance measures: Using Dirichlet benchmarks. *Journal of Business Research, 57, 12, 1307–1325.*

FADER, P.S. and SCHMITTLEIN, D.C. (1993): Excess Behavioral Loyalty for High-Share Brands: Deviations from the Dirishlet Model for Repeat Purchasing. *Journal of Marketing Research, 30 (November), 478–493.*

GOODHARDT, G.J., EHRENBERG, A.S.C., and Chatfield C. (1984): The Dirichlet: A Comprehensive Model of Buying Behaviour. *Journal of the Royal Statistical Society, A 147, 621–55.*

HARRIS, J. (2003): *An investigation of brand choice in repeat purchase markets: the case of business airline travel.* PHD Dissertation, University New South Wales, Sydney.

KAHN, B., KALWANI, M., and MORRISON, D. (1988): Niching Versus Change-of-Pace Brands: Using Purchase Frequencies and Penetration Rates to Infer Brand Positionings. *Journal of Marketing Research, 15, 4,384–90.*

KEARNS, Z., BOUN, J., and GOODHARDT, G. (1998): *DIRICHLET.* Southbank University.

KENG, K. and EHRENBERG, A.S.C.(1984): Patterns of store choice. *Journal of Marketing Research, 21, 399–409.*

MÄGI, A.W. (2003): Share of wallet in retailing: the effects of customer satisfaction, loyalty cards and shopper characteristics. *Journal of Retailing, 109,2, 1–11.*

MCGAHAN, A.M. and GHEMAWAT, P. (1994): Competition to Retain Customers. *Marketing Science, 13, 2, 165–176.*

MEYER-WAARDEN, L. (2002): Les sources d'éfficacit des programmes de fidélisation: une investigation empirique sur un panel single source [The sources of efficiency of loyalty programmes - an investigation based upon a single-source panel]. PHD Dissertation, University Pau and Institut für Entscheidungstheorie und Unternemehmensforschung, Karlsruhe.

MEYER-WAARDEN, L. (2004): La fidélisation client: stratégies, pratiques et efficacité des outils du marketing relationnel [Customer retention: strategies, practice and efficiency of customer relationship management tools]. Vuibert Eds., Paris.

NAKO, S.M. (1997): Frequent flyer programs and business travelers: an empirical investigation. *Logistics and Transportation Review, 28, 4, 395- 410.*

REICHHELD, F. (1996): *The loyalty effect: the hidden force behind growth, profits, and lasting value.* Harvard Business School Press, Boston, MA.

SHARP, B. and SHARP, A. (1997): Loyalty Programs and Their Impact on Repeat-Purchase Loyalty Patterns. *International Journal of Research in Marketing, 14, 473-486.*

An Empirical Examination of Daily Stock Return Distributions for U.S. Stocks*

Svetlozar T. Rachev[1,2], Stoyan V. Stoyanov[3], Almira Biglova[1], and Frank J. Fabozzi[4]

[1] Department of Econometrics, Statistics and Mathematical Finance, University of Karlsruhe, D-76128 Karlsruhe, Germany
[2] Department of Statistics and Applied Probability, University of California Santa Barbara, CA 93106, USA
[3] FinAnalytica, Inc., Seattle, USA
[4] Frederick Frank Adjunct Professor of Finance, Yale University, School of Management, P.O. Box 208200, New Haven, CT 06520-8200, USA

Abstract. This article investigates whether the Gaussian distribution hypothesis holds 382 U.S. stocks and compares it to the stable Paretian hypothesis. The daily returns are examined in the framework of two probability models - the homoskedastic independent, identical distributed model and the conditional heteroskedastic ARMA-GARCH model. Consistent with other studies, we strongly reject the Gaussian hypothesis for both models. We find out that the stable Paretian hypothesis better explains the tails and the central part of the return distribution.

1 Introduction

The cornerstone theories in finance such as mean-variance model for portfolio selection and asset pricing models that have been developed rest upon the assumption that asset returns follow a normal distribution. Yet, there is little, if any, credible empirical evidence that supports this assumption for financial assets traded in most markets throughout the world. Moreover, the evidence is clear that financial return series are heavy-tailed and, possibly, skewed. Fortunately, several papers have analyzed the consequences of relaxing the normality assumption and developed generalizations of prevalent concepts in financial theory that can accommodate heavy-tailed returns (see Rachev and Mittnik (2000) and Rachev (2003) and references therein).

Mandelbrot (1963) strongly rejected normality as a distributional model for asset returns, conjecturing that financial return processes behave like non-Gaussian stable processes. To distinguish between Gaussian and non-

* Prof. Rachev gratefully acknowledges research support by grants from Division of Mathematical, Life and Physical Sciences, College of Letters and Science, University of California, Santa Barbara, the Deutschen Forschungsgemeinschaft and the Deutscher Akademischer Austausch Dienst.

Gaussian stable distributions. The latter are commonly referred to as "stable Paretian" distributions or "Levy stable" distributions [1].

While there have been several studies in the 1960s that have extended Mandelbrot's investigation of financial return processes, probably, the most notable is Fama (1963, 1965). His work and others led to a consolidation of the stable Paretian hypothesis. In the 1970s, however, closer empirical scrutiny of the "stability" of fitted stable Paretian distributions also produced evidence that was not consistent with the stable Paretian hypothesis. Specifically, it was often reported that fitted characteristic exponents (or tail-indices) did not remain constant under temporal aggregation[2]. Partly in response to these empirical "inconsistencies", various alternatives to the stable law were proposed in the literature, including fat-tailed distributions being only in the domain of attraction of a stable Paretian law, finite mixtures of normal distributions, the Student t-distribution, and the hyperbolic distribution.

A major drawback of all these alternative models is their lack of stability. As has been stressed by Mandelbrot and argued by Rachev and Mittnik (2000), among others, the stability property is highly desirable for asset returns. This is particularly evident in the context of portfolio analysis and risk management. Only for stable distributed returns does one obtain the property that linear combinations of different return series (e.g., portfolios) follow again a stable distribution. Indeed, the Gaussian law shares this feature, but it is only one particular member of a large and flexible class of distributions, which also allows for skewness and heavy-tailedness.

Recent attacks on Mandelbrot's stable Paretian hypothesis focus on the claim that empirical asset return distributions are not as heavy-tailed as the non-Gaussian stable law suggests. Studies that come to such conclusions are typically based on tail-index estimates obtained with the Hill estimator. Because sample sizes beyond 100,000 are required to obtain reasonably accurate estimates, the Hill estimator is highly unreliable for testing the stable hypothesis. More importantly, the Mandelbrot's stable Paretian hypothesis is interpreted too narrowly, if one focuses solely on the *marginal* distribution of return processes. The hypothesis involves more than simply fitting marginal asset return distributions. Stable Paretian laws describe the fundamental "building blocks" (e.g., innovations) that drive asset return processes. In addition to describing these "building blocks," a complete model should be rich enough to encompass relevant stylized facts, such as (1) non-Gaussian, heavy-tailed and skewed distributions, (2) volatility clustering (ARCH-effects), (3)

[1] Stable Paretian is used to emphasize that the tails of the non-Gaussian stable density have Pareto power-type decay "Levy stable" is used in recognition of the seminal work of Paul Levy's introduction and characterization of the class of non-Gaussian stable laws.

[2] For a more recent study, see Akgiray and Booth (1988) and Akgiray and Lamoureux (1989).

temporal dependence of the tail behavior, and (4) short- and long-range dependence.

An attractive feature of stable models — not shared by other distributional models — is that they allow us to generalize Gaussian-based financial theories and, thus, to build a coherent and more general framework for financial modeling. The generalizations are only possible because of specific probabilistic properties that are unique to (Gaussian and non-Gaussian) stable laws, namely, the stability property, the Central Limit Theorem and the Invariance Principle for stable processes [3].

In this paper we present additional empirical evidence comparing normal and stable Paretian for a large sample of U.S. stocks. Our empirical analyses go beyond those typically found in the literature wherein the focus is almost exclusively on the *unconditional* distribution of equity returns. Here we also investigate conditional homoskedastic (i.e., constant-conditional-volatility) and heteroskedastic (i.e., varying-conditional-volatility) distributions. It is because asset returns typically exhibit temporal dependence that the *conditional* distributions are of interest. If asset return embed information on past market movements, it is not the *unconditional* return distribution which is of interest, but the *conditional* distribution, which is conditioned on information contained in past return data, or a more general information set.

We believe our study is the first one investigating the stable Paretian distribution for equity returns that includes a large sample of companies (382). Most other studies have been limited to stock indexes. In studies where individual stock returns have been analyzed, the samples have been small, typically limited to the constituent components of the Dow Jones Industrial Average or a non-U.S. stock index with no more than 40 stocks. Most likely the reason for the limitation of other researchers to use a large sample of companies to analyze the stable Paretian distribution is the computational time involved to calculate the maximum likelihood estimate of the parameters of the stable distribution. For our 382 companies, for example, it took approximately 3.5 hours on PC 2.4 GHz Intel Pentium IV, 1Gb of RAM to compute the stable parameters for both models that we estimate in this study. All calculations were done in MATLAB. The same amount of calculations would have taken about 17.5 hours in 1999 and about 68 hours in 1996. The RAM requirements would have made such a study very hard to organize in 1990 and practically impossible in 1985 on an average PC [4].

[3] Detailed accounts of properties of stable distributed random variables can be found in Samorodnitsky and Taqqu (1994) and Janicki and Weron (1994).

[4] DuMouchel (1971) was the first to study maximum likelihood estimation of the parameters of stable distributions in his dissertation in the beginning of the 1970s. He did his calculations on a CDC 6400 computer at the University of California, Berkeley. The computers of this class were among the fastest computers in the world at the end of the 1960s. Assuming infinite amount of computer memory available, he would have needed more than 20 days to perform only the parameter fitting of the stable distributions for our sample. This extremely rough estimate

The paper is organized as follows. The methodology employed is explained in Section 2 followed by a description of our sample in Section 3. The empirical results are reported in Section 4 and a summary of our major conclusions are presented in Section 5.

2 Methodology

As noted in Section 1, because asset returns typically exhibit temporal dependence, the focus of the analysis should be on *conditional* distributions. The class of autoregressive moving average (ARMA) models is a natural candidate for conditioning on the past of a return series. These models have the property that the conditional distribution is homoskedastic. However, because financial markets frequently exhibit volatility clusters, the homoskedasticity assumption may be too restrictive. As a consequence, conditional heteroskedastic models, such as that proposed by autoregressive conditional heteroskedastic (ARCH) models as proposed Engle (1982) and the generalized GARCH proposed by Bollerslev (1986), possibly in combination with an ARMA model, referred to as an ARMA-GARCH model, are common in empirical finance. It turns out that ARCH-type models driven by normally distributed innovations imply unconditional distributions which themselves possess heavier tails. However, many studies have shown that GARCH-filtered residuals are themselves heavy-tailed, so that stable Paretian distributed innovations ("building blocks") would be a reasonable distributed assumption.

In the general case, no closed-form expressions are known for the probability density and distribution functions of stable distributions. They are described by four parameters: α, called the index of stability, which determines the tail weight or density's kurtosis with $0 < \alpha \leq 2$, β, called the skewness parameter, which determines the density's skewness with $-1 \leq \beta \leq 1$, $\sigma > 0$ which is a scale parameter, and μ which is a location parameter. Stable distributions allow for skewed distributions when $\beta \neq 0$ and when β is zero, the distribution is symmetric around μ. Stable Paretian laws have fat tails meaning that extreme events have high probability relative to the normal distribution when $\alpha < 2$. The Gaussian distribution is a stable distribution with $\alpha = 2$. (For more details on the properties of stable distributions see Samorodnitsky and Taqqu (1994).) Of the four parameters, α and β are most important as they identify two fundamental properties that are untypical of the normal distribution- heavy tails and asymmetry.

3 Description of the data

The sample of companies used in this study was obtained as follows. We began with all the companies that were included in the S&P500 index over the

was determined by us based on the calculation time in DuMouchel's dissertation and the relative speed of his algorithm compared to ours.

12-year time period January 1, 1992 to December 12, 2003. The constituent companies in the S&P 500 are determined by a selection committee of the Standard & Poor's Corporation which periodically adds and removes companies from the index. Over the 12-year time period, there were more than 800 companies that had been included in the index. We then selected from these companies those that had a complete return history (3,026 observations). The rest of the companies that were included in S&P500 stocks but not in our sample have shorter historical series with unequal histories.

Daily returns were calculated as $r(t) = \log(S(t)/S(t-1))$, where $S(t)$ is the stock value at t (the stocks are adjusted for dividends).

4 Tests and results

We employ two tests in our investigation. In the first test, we assume that daily return observations are independent and identically distributed (iid); in the second test, the daily return observations are assumed to follow a ARMA(1,1)-GARCH(1,1) process. The first test concerns the unconditional, homoskedastic distribution model while the second one belongs to the class of conditional heteroskedastic models.

For both tests, we verify whether the Gaussian hypothesis holds. The normality tests employed are based on the Kolmogorov distance (KD) and computed according to

$$KD = \sup_{x \in \mathbb{R}} |F_s(x) - F(x)|$$

where $F_s(x)$ is the empirical sample distribution and $F(x)$ is the cumulative distribution function (cdf) of the estimated parametric density and emphasizes the deviations around the median of the distribution.

For both the iid and the ARMA-GARCH tests, we compare the goodness-of-fit in the case of the Gaussian hypothesis and the more general stable Paretian hypothesis. We use two goodness-of-fit measures for this purpose, the KD-statistic and the Anderson-Darling (AD) statistic. The AD-statistic is computed as follows:

$$AD = \sup_{x \in \mathbb{R}} |F_s(x) - F(x)|$$

The AD-statistic accentuates the discrepancies in the tails. Since in the calculation of the AD statistic the extreme observations are most important, we repeat the calculations assuming that the observations below the 0.1% quantile and above 99.9% quantile result from errors in the data. We replace them with the average of the two adjacent observations.

4.1 The iid model

In the simple setting of the iid (independent identically distributed returns) model, we have estimated the values for the four parameters of the stable Paretian distribution using the method of maximum likelihood. Figure 1 shows a scatter plot of the estimated pairs (α, β) for all stocks.

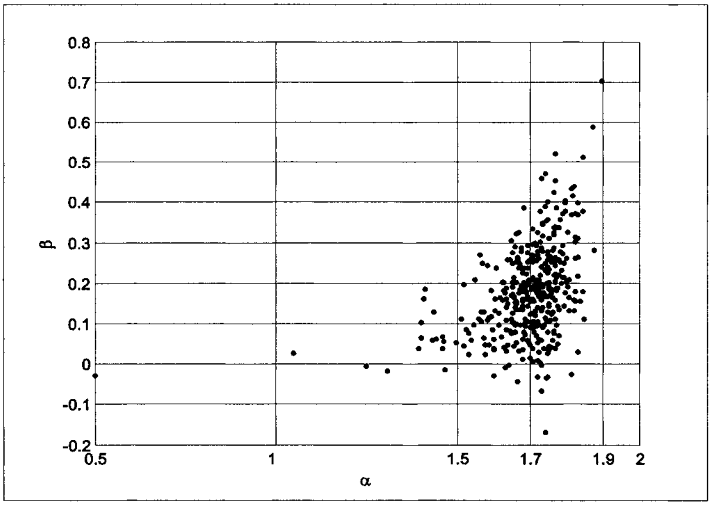


Fig. 1. Scatter plot of the index of stability and the skewness parameter for the daily returns of 382 stocks, iid model.

We find that for the return distribution of every stock in our sample that (1) the estimated values of the index of stability are below 2 and (2) there is asymmetry ($\beta \neq 0$). These two facts alone would strongly suggest that the Gaussian assumption is not a proper theoretical distribution model for describing the return distribution of stocks. Additional support for the stable Paretian hypothesis is contained in Tables 1 and 2. The two tables show that we can reject the normality using the standard Kolmogorov-Smirnov test for (1) more than 95% of the stocks at the extremely high confidence level of 99.9% and (2) 100% of the stocks at the traditional levels of 95% and 99%. In contrast, the stable-Paretian hypothesis is rejected in much fewer cases.

The superiority of the stable Paretian assumption over the Gaussian assumption is clearly seen by examining Figures 4 and 5. The figures show the computed KD and AD statistics for all stocks under the two distributional assumptions. For every stock in our sample, the KD-statistic in the stable Paretian case is below the KD-statistic in the Gaussian case. The same is true for the AD-statistic. The KD-statistic implies that for our sample firms there a better fit of the stable Paretian model around the center of the distribution

Confidence level	95%	99%	99.50%	99.90%
Original data	100.00%	100.00%	99.74%	99.21%
Truncated data	100.00%	99.48%	98.43%	96.34%

Table 1. Percentage of stocks for which normality is rejected at different confidence levels using Kolmogorov distance in the iid model.

while the AD-statistic implies a better fit in the tails. The huge difference between the AD-statistic computed for the stable Paretian model relative to the Gaussian model strongly suggests a much better ability for the stable Paretian model to forecast extreme events and confirms an already noticed phenomenon — the Gaussian distribution fails to describe observed large downward or upward asset price shifts, that is in reality extreme events have larger probability than predicted by the Gaussian distribution.

Confidence level	95%	99%	99.50%	99.90%
Original data	28.54%	15.71%	11.78%	7.07%
Truncated data	33.51%	16.49%	12.83%	7.85%

Table 2. Percentage of stocks for which the Stable Paretian hypothesis is rejected at different confidence levels using Kolmogorov distance in the iid model.

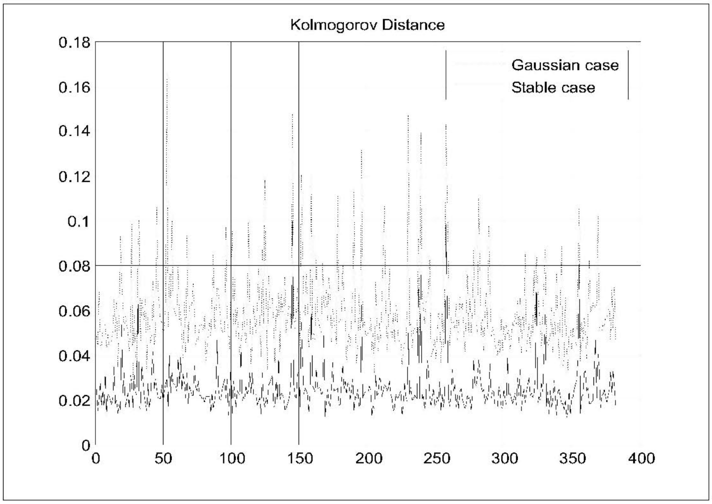


Fig. 2. Kolmogorov Distance in both stable Paretian and Gaussian cases for all 382 stocks, iid model.

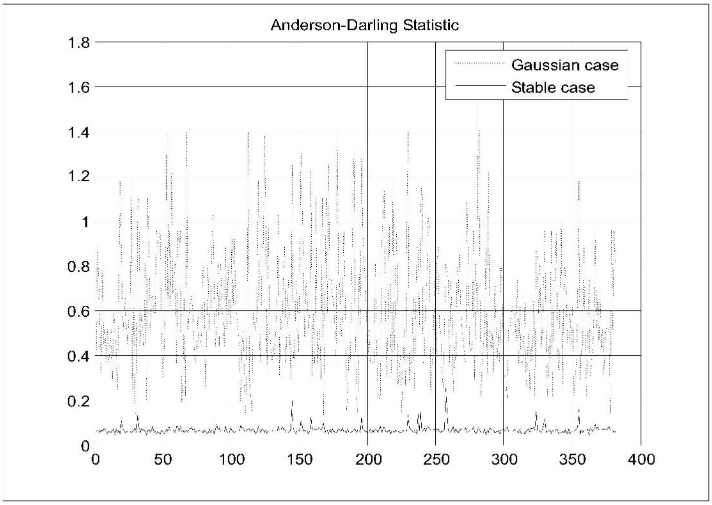


Fig. 3. Anderson-Darling statistic in both stable Paretian and Gaussian cases for all 382 stocks, iid model.

Summary statistics of the various statistical tests and parameter estimates for the entire sample are provided in Table 3. Truncating the data improves the AD statistic for the Gaussian model but it remains more than 8 times larger than the AD statistic for the stable Paretian model. In the case of the non-truncated data, the former is more than 10 times larger than the latter.

		α	β	KD Normal	KD Stable	AD Normal	AD Stable
	mean	1.697	0.18	0.0718	0.0241	0.6575	0.0612
Original	median	1.715	0.18	0.0638	0.0217	0.5867	0.0577
Data	25% quantile	1.661	0.1	0.0543	0.0184	0.5128	0.0517
	75% quantile	1.761	0.25	0.0769	0.0257	0.8084	0.0639
	mean	1.72	0.2	0.0588	0.0248	0.5977	0.0669
Truncated	median	1.739	0.19	0.0543	0.0223	0.5128	0.064
Data	25% quantile	1.683	0.11	0.0471	0.0193	0.365	0.0581
	75% quantile	1.786	0.27	0.0633	0.0268	0.8084	0.0696

Table 3. Summary statistics for the entire sample of 382 stocks, iid model.

4.2 ARMA-GARCH model

Since the simple iid model does not account for the clustering of the volatility phenomenon, as a second test, we consider the more advanced ARMA-

GARCH model. The general form of the ARMA(p,q)-GARCH(r,s) model is:

$$
\begin{aligned}
r_t &= C + \sum_{i=1}^{p} a_i r_{t-i} + \sum_{i=1}^{q} b_i \varepsilon_{t-i} + \varepsilon_t \\
\varepsilon_t &= \sigma_t \delta_t \\
\sigma_t^2 &= K + \sum_{i=1}^{r} w_i \varepsilon_{t-i}^2 + \sum_{i=1}^{s} \nu_i \sigma_{t-i}^2
\end{aligned}
\tag{1}
$$

where a_i, $i = 1, \ldots, p$, b_i, $i = 1, \ldots, q$, w_i, $i = 1, \ldots, r$, ν_i, $i = 1, \ldots, s$, C and K are the model parameters. δ_t's are called the *innovations process* and are assumed to be iid random variables which we additionally assume to be either Gaussianl or stable Paretian. An attractive property of the ARMA-GARCH process is that it allows a time-varying volatility via the last equation.

The model in which $p = q = r = s = 1$ proved appropriate for the 382 stock returns time series that we consider because the serial correlation in the residuals disappeared. The model parameters are estimated using the method of maximum likelihood assuming the normal distribution for the innovations. In this way, we maintain strongly consistent estimators of the model parameters under the stable Paretian hypothesis since the index of stability of the innovations is greater than 1 (see Rachev and Mittnik (2000) and references therein).

After estimating the ARMA(1,1)-GARCH(1,1) parameters, we computed the model residuals and verified which distributional assumption is more appropriate. Figure 4 shows a scatter-plot of the estimated (α, β) pairs. For every return distribution in our sample, the estimated index of stability is greater than 1. Even though the estimated values of α are closer to 2 than in the iid model, they are still significantly different from 2.

Comparing the results reported in Table 4 to those reported in Table 1, we observe that the Gaussian model is rejected in fewer cases in the ARMA-GARCH model than in the simple iid model; nevertheless, the Gaussian assumption is rejected for more than 82% of the stocks at the 99% confidence level using the truncated data. The stable Paretian assumption is rejected in only about 6% of the stocks at the same confidence level (see Table 5).

A summary of the computed statistics for the residuals of the ARMA-GARCH model is reported in Table 6. Generally, the results imply that the stable Paretian assumption is more adequate as a probabilistic model for the innovations compared to the Gaussian assumption. As in the iid model, truncating the data improves the AD-statistic in the Gaussian case but still it remains about seven times larger than the AD-statistic in the stable Paretian case.

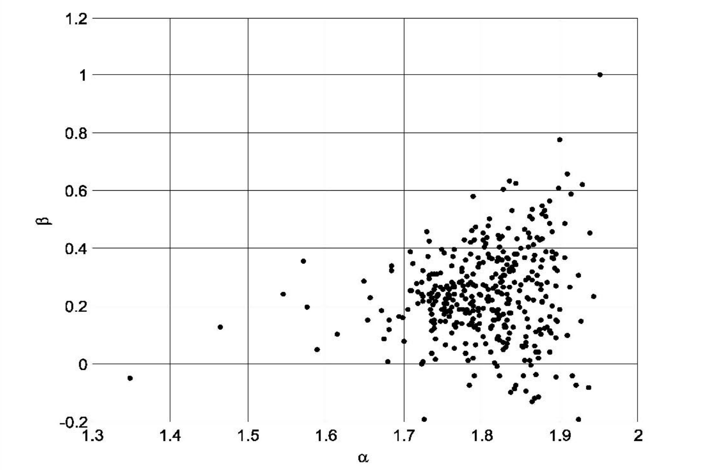


Fig. 4. Scatter plot of the index of stability and the skewness parameter for the residuals in the ARMA-GARCH model for all 382 stocks.

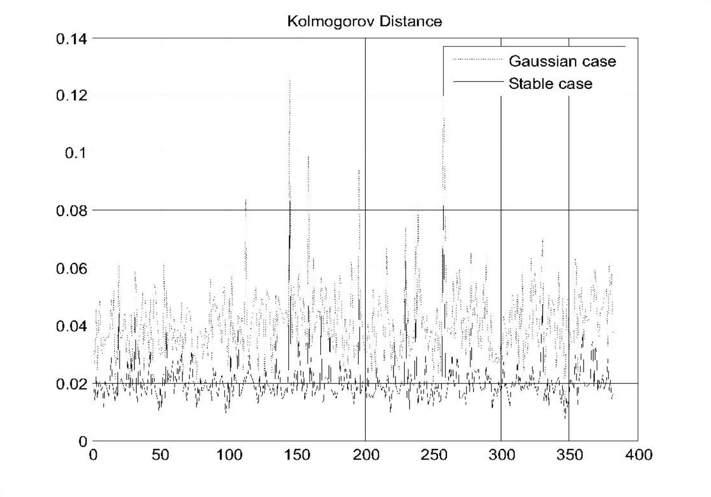


Fig. 5. Kolmogorov Distance in both stable Paretian and Gaussian cases for the residuals in the ARMA-GARCH model.

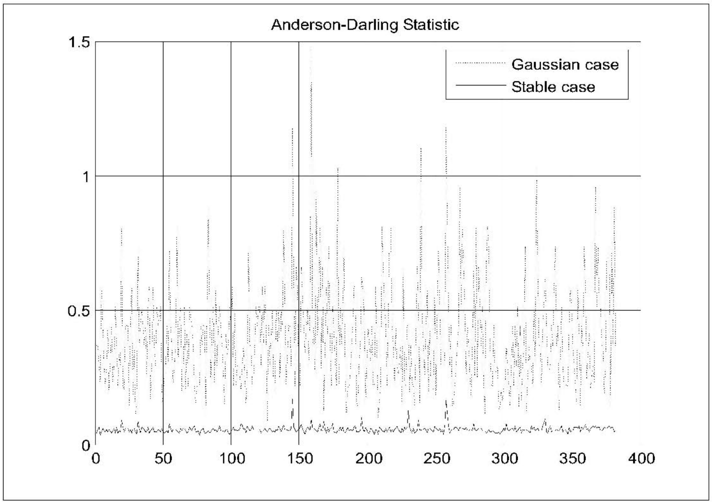


Fig. 6. Anderson-Darling statistic in both stable Paretian and Gaussian cases for the residuals in the ARMA-GARCH model.

Confidence level	95%	99%	99.50%	99.90%
Original data	97.38%	92.94%	89.79%	79.58%
Truncated data	93.46%	82.46%	75.92%	61.78%

Table 4. Percentage of stocks for which normality is rejected at different confidence levels using Kolmogorov distance in the ARMA-GARCH model.

Confidence level	95%	99%	99.50%	99.90%
Original data	12.05%	5.76%	4.98%	2.62%
Truncated data	13.87%	6.28%	4.50%	3.40%

Table 5. Percentage of stocks for which the Stable Paretian hypothesis is rejected at different confidence levels using Kolmogorov distance in the ARMA-GARCH model.

5 A back-testing example

We perform a back-testing analysis for the time series of the equity ILN, which is one of the 382 equities considered in the previous analyses. The model parameters are estimated using a moving time window of 1000 observations and the back-testing period is 500 and 1500 days. We compare the performance of the more simple GARCH(1,1) model from the ARMA-GARCH family with stable and normal innovations with respect to risk estimation in the cases of Value-at-risk (VaR) and Expected tail loss (ETL) risk measures at 99.5% and 99.9% confidence levels. The performance is compared in terms of the

		α	β	KD Normal	KD Stable	AD Normal	AD Stable
	mean	1.805	0.24	0.0502	0.0198	0.4859	0.0525
Original	median	1.813	0.23	0.0464	0.0184	0.4389	0.0512
Data	25% quantile	1.763	0.14	0.0381	0.0159	0.365	0.0465
	75% quantile	1.856	0.33	0.0566	0.0219	0.5867	0.0566
	mean	1.834	0.28	0.041	0.0205	0.4061	0.0548
Truncated	median	1.842	0.26	0.0391	0.0192	0.365	0.0533
Data	25% quantile	1.794	0.16	0.033	0.0164	0.2549	0.0474
	75% quantile	1.883	0.39	0.0466	0.0227	0.5128	0.0595

Table 6. Summary statistics for the entire sample of 382 stocks, ARMA-GARCH model.

number of exceedances for the VaR measure, that is how many times the forecast of the VaR is above the realized asset return. We verify if the number of exceedances is in the 95% confidence interval for the corresponding back-testing period.

Confidence level	Stable VaR	Normal VaR	95% Confidence interval
99.5% (500 days)	4	10	[0, 5]
99.9% (1500 days)	2	10	[0, 3]

Table 7. The number of exceedances of the VaR at a given confidence level in the back-testing period, GARCH(1,1) model.

Confidence level	Stable average ETL	Normal average ETL
99.5% (500 days)	0.1878	0.1231
99.9% (1500 days)	0.3142	0.1433

Table 8. he average ETL in the back-testing period of 1500 days produced by both GARCH(1,1) models.

Exactly as expected, the exceedances of the Normal model are not in the 95% confidence interval, in contrast to the exceedances of the Stable model (see Table 7). Therefore the Stable model better approximates the tail of the empirical equity return distribution. The average index of stability of the residuals for the entire back-testing period of 1500 days is 1.83.

The ETL produced by the stable model is more conservative than the corresponding figure of the normal model (see Table 8). The VaR analysis suggests that the Stable ETL is more realistic than the Normal one.

6 Conclusions

We have studied the daily equity returns distribution of 382 U.S. companies comparing the Gaussian and the stable Paretian hypotheses in the context of two assumptions — independent and identical distribution of the daily stock returns and the ARMA-GARCH model. For both models, we strongly reject the normality assumption in favor of the stable Paretian Hypothesis.

References

AKGIRAY, A. and BOOTH, G.G. (1988): The stable-law model of stock returns. *Journal of Business and Economic Statistics, 6, 51–57.*

AKGIRAY, V. and LAMOUREUX, G.G. (1989): Estimation of stable-law parameters: A comparative study. *Journal of Business and Economic Statistics, 7, 85-93.*

BOLLERSLEV, T. (1986): Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics, 31, 307–327.*

DUMOUCHEL, W.H. (1971): Stable Distributions in Statistical Inference. *A dissertation Presented to the Faculty of the Graduate School of Yale University in Candidacy for the Degree of Doctor of Philosophy Yale University.*

ENGLE, R.F. (1982): Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica, 50, 276–287.*

FAMA, E. (1963): Mandelbrot and the stable Paretian hypothesis. *Journal of Business 36, 420–429.*

FAMA, E. (1965): The behavior of stock market prices. *Journal of Business 38, 34–105.*

JANICKI, A. and WERON, A. (1994): *Simulation and Chaotic Behavior of Alpha-Stable Stochastic Processes.* Marcel Dekker, New York.

MANDELBROT, B. (1963): The variation of certain speculative prices. *Journal of Business, 26, 394–419.*

RACHEV, S.T. and MITTNIK, S. (2000): *Stable Paretian Models in Finance.* John Wiley & Sons, Chichester.

RACHEV, S.T. (Ed.) (2003): *Handbook of Heavy Tailed Distributions in Finance.* Elsevier/North-Holland, Amsterdam.

SAMORODNITSKY, G. and TAQQU, M. (1994): *Stable Non-Gaussian Random Processes, Stochastic models with Infinite Variance.* Chapman and Hall, New York.

Stages, Gates, and Conflicts in New Product Development: A Classification Approach*

Alexandra Rese[1], Daniel Baier[1], and Ralf Woll[2]

[1] Chair of Marketing and Innovation Management, Brandenburg University of Technology, Konrad-Wachsmann-Allee 1, D-03046 Cottbus, Germany
[2] Chair of Quality Management, Brandenburg University of Technology, Konrad-Wachsmann-Allee 1, D-03046 Cottbus, Germany

Abstract. Formal management processes for new product development (NPD) make use of multi-functional integration with cross-functional team meetings at so-called communication points. These meetings are of different type and reach from information exchange up to choice decisions or "sign offs" of concepts, products, or production processes. For the outcome of the NPD the recognition and adequate management of conflicts play an important role. We concentrate on information exchange and choice decisions in cross-functional integration using stage-gate approaches. Communication points and conflict potentials are identified. Finally, a conflict classification is derived.

1 Introduction

During the last 30 years various approaches and methods have been proposed to better integrate business functions like R&D, marketing, or production during new product development (NPD). Phase-review processes, stage-gate processes, and quality function deployment are just examples for these attempts to improve the information exchange, the harmony between integrated functional areas and individuals as well as efficiency and effectiveness of the whole NPD process (see, e.g., Urban and Hauser (1993), Gaul and Baier (1994), Griffin and Hauser (1996) for recent reviews). Most of these approaches and methods make use of cross-functional teams and communication points where team members with different background, interests, and action plans meet. There, choice decisions have to be prepared and/or made.

However, recently, the organization of this cross-functionality has also been discussed as a threat, since the underlying cultural, socio demographic, and socio emotional heterogenity of the teams tends to produce emotional tensions and conflicts that may deteriorate the team performance and/or NPD success (see, e.g., Shaw and Shaw (1998), Shaw et al. (2003)). Consequently, it is assumed that the proper recognition and adequate management of conflicts between functional areas and individuals play an important role for the NPD outcome (see, e.g., Lam and Chin (2005)).

* This work has been supported by the Deutsche Forschungsgemeinschaft, project WO 1212/1-1: Developing a Communication Concept for Product Development.

This paper tries to discuss this problem in the case of stage-gate approaches for NPD using BMW motorcycle's gateway management as a sample application. In section 2, the stage-gate approach and its realization at BMW motorcycle is shortly discussed. The communication points and their conflict potential are analyzed in section 3. In section 4 classifications for conflicts are discussed and a new one for conflicts in NPD is developed. A short summary closes the paper.

2 The stage-gate approach in NPD

Cooper's stage-gate approach (e.g. Cooper (1983) with extensions in Cooper (1994) as well as Cooper and Kleinschmidt (1986)) has influenced research and practice of NPD decisively (see, e.g., Griffin and Hauser (1996), Krishnan and Ulrich (2001), Ding and Eliashberg (2002)). It divides the NPD process into sequential phases, the so-called "stages", which are multi-functional and closed by a "gate" where the continuation of the NPD project is decided cross-functionally according to pre-specified criteria (see Fig. 1). At the same time an examination takes place whether the respective stage was carried out correctly, the required results were obtained, and the necessary information and preparations for the next stage were provided.

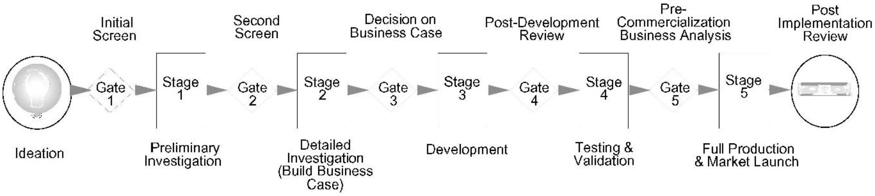

Initial Screen
Second Screen
Decision on Business Case
Post-Development Review
Pre-Commercialization Business Analysis
Post Implementation Review
Gate 1
Stage 1
Gate 2
Stage 2
Gate 3
Stage 3
Gate 4
Stage 4
Gate 5
Stage 5
Ideation
Preliminary Investigation
Detailed Investigation (Build Business Case)
Development
Testing & Validation
Full Production & Market Launch

Fig. 1. The stage-gate approach for NPD (Cooper (1983))

A comparison of the stage-gate approach with real-world NPD processes shows, that the underlying concept is wide-spread in practice. So, e.g., BMW motorcycle's NPD process for the K 1200 S (see Fig. 2 for a brief characterization) is multi-functionally organized with cross-functional gates. The so-called "gateway management" comprises altogether three major and six smaller stages and subsequent gates, where involved teams and deciders from marketing, design, construction, production, and testing have to "sign off" the concept, product, or production process for the next stage.

Besides this "sign off" at each gate, various choice decisions have to be made. So, e.g., at the subsequent gates choices with respect to (w.r.t.)

- the strategic positioning, the body variant, and the surfaces,
- the package relevant concept alternatives, the technological innovations,
- the surface concept, the suppliers, and the design,
- the outer skin, the engine suppliers,

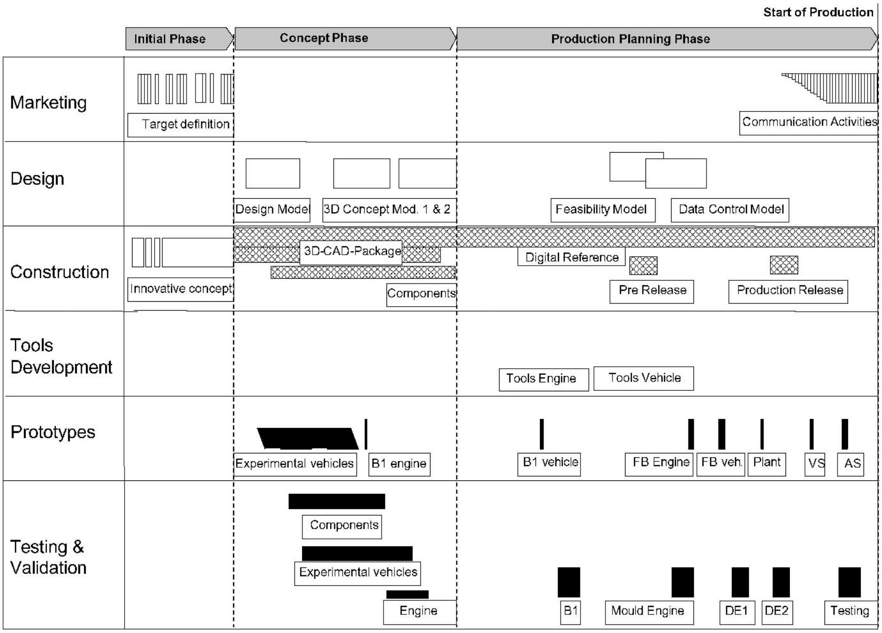


Fig. 2. BMW's NPD process for the K 1200 S (Berkmann et al. (2004), p. 635)

- the outline of the production process,
- the supplier contracts and the product quality,
- the assembly and the production process under realistic conditions,
- the market introduction plan as well as
- the marketing concept

have to be made.

In each team or rather in the weekly team meetings during the stages just like in the decision-making committees at the gates different interests and action plans meet each other and decisions nevertheless have to be prepared and made. Therefore these so-called communication points are particularly susceptible to conflicts.

3 Conflicts in NPD

In general conflicts result from perceived incompatibilities of goals and views of different parties (Boulding (1957), Deutsch (1976, 1990), Regnet (2001)). Conflicts appear so frequently at NPD or generally at innovations, that they are often described as a distinctive feature of innovations (Posch (2003), S.249). Pinto and Kharbanda (1995), e.g., estimate, that a project manager spends at least 20% of her/his time on average to clear conflicts and their aftereffects.

Badke-Schaub and Frankenberger (2004) identified in their 28-week observation of four engineering processes altogether 265 situations which were critical for the success of the engineering process. However, the interviewed engineers rated only 22 (8.3%) of these critical situations as "conflicts". For them, only unwanted work shifts (e.g. solution of ad-hoc problems outside the engineering process), outer restrictions (e.g. computer failure, supplier problems) as well as unsatisfactory work progress (e.g. the missing availability of contact persons) where conflict situations, other situations with incompatible goals and views were rated as "disturbances", "goal analysis", "solution search", "solution decisions", and the like. This observation is similar to results from surveys on the relationship between the functional areas marketing and engineering, where, e.g., only 9.1% of interviewed engineers indicated that their overall degree of conflict with marketing colleagues is high (Shaw et al. (2003)).

This leads to a central problem of understanding which critical situations shall be described and handled as conflicts in NPD (see, e.g., Högl (1998)). So numerous authors distinguish two types of conflicts:

- On one side there is the factual, task relevant discussion, which is described by Pinto and Kharbanda (1995) as a task conflict or by Jehn (1995) as a functional conflict. Several empirical examinations have shown, that such conflicts have a positive effect on the team performance particularly at increasing task complexity. So, e.g., Lechler (1997) has shown in his examination of 448 projects that frequency and intensity of conflicts influence the success of projects significantly. Such conflicts practically cannot be avoided in product development, but they are rather described as "natural conditions which can often contribute to goal attainment, by being able to show different points of view through them" (Posch (2003), p.249).
- On the other side we have "real" (as assumed by the engineers in the above surveys) conflicts, also called dysfunctional (Pinto und Kharbanda (1995)), interpersonal or task conflicts (Jehn (1995), Pelled (1996)). These conflicts are connected with emotional tensions between individuals characterized by distrust, fear, rage, frustration, and other negative feelings.

Pelled (1996) points out that task and affective conflicts are not completely independent of each other. If team members are particularly firmly convinced of a factual question (e.g. necessary actions in a NPD project), this could lead to emotionally loaded discussions and finally in affective conflicts.

Important reasons for conflicts are considered to be different personality attributes like socio economic status or personal values and norms. There is a good chance for such differences (Pelled (1996)) in case of heterogeneity with respect to

- demographic variables like age, sex, race,
- duration of team and/or organization membership or

- degree and area of education.

Besides that of course also unsatisfactory social abilities of team and organization members can be reasons for conflicts (Levine and Moreland (1990)). Particularly insufficient empathy and lacking communicative abilities are regarded as a frequent conflict source (Högl (1998), p. 49).

In connection with this, it is reasonable that particularly the cooperation over subject and department limits, e.g., in context of the work in heterogeneous teams in early NPD stages, is much conflict laden. Seidel (1996), p. 96) frequently identified the following conflict causes in an empirical examination of the communication between R&D and marketing employees (The view of the R&D employees is presented here.):

- unclear role play and insufficient knowledge of process and task,
- unsatisfactory synchronization of the marketing requirements with the regular R&D process,
- short-term thinking in the marketing collides with R&D requirements,
- unclear, unknown, incomplete and unstable market requirements, market strategies and trend statements,
- distrust and prejudices strategies partly by negative experiences in the past,
- delayed and/or insufficient flow of information,
- new product qualities are expressed for the customer by the marketing insufficiently,
- insufficient joint discussion of the target group expectations,
- lacking acceptance of technical statements of R&D in marketing,
- ex-post oriented formulation of market requirements by marketing: lack of ex-ante formulated goals and requirements.

In order to systematize the conflict sources and types, in the following, an integrative classification scheme for conflicts in NPD is developed.

4 Classifying conflicts in NPD

There already exist numerous systematization attempts for conflicts in the literature. Well-known examples are, e.g., the following:

- Deutsch (1976) distinguishes latent from manifest conflicts.
- Rüttinger (1980) speaks of perception, evaluation, distribution conflicts.
- Glasl (1980) developed the differentiation into hot and cold conflicts.
- Pinto and Kharbanda (1995) distinguish functional and dysfunctional conflicts.
- Jehn (1995) speaks in a similar way of affective and task conflicts.
- Recently, Jehn and Mannix (2001) have proposed the classification into task, relationship, and process conflicts.

Each of these classifications addresses different underlying conflict sources and appearances.

- So, e.g., as already mentioned in section 3, the main difference between functional (or task) conflicts on the one side and dysfunctional (or affective) conflicts on the other side is the presence of emotional tensions.
- Task and process conflicts differ w.r.t. the conflict object: Task conflicts address the results which should be pursued and/or the allocation of resources, whereas process conflicts address the "how to" pursue results or "how to" allocate resources.
- Perception and evaluation conflicts differ w.r.t. the way the different conflict parties see or rate the conflict objects. Differences may occur w.r.t. knowledge about the objects or w.r.t. the way the objects are evaluated.
- Latent differ from manifest conflicts w.r.t. how deeply the conflict is felt. Manifest conflicts are conflicts where the differences – at least by one party – are perceived, where conflict resolution is felt necessary and/or where conflict resolution is underway.

Fig. 3 tries to combine this systematization attempts by a conflict profile which can be used for conflict recognition. Fig. 4 shows that the new profile contains the traditional systematizations as special cases.

Conflict object:

Differences w.r.t. goals pursued ...	... are perceived.	... should ...	... will be resolved.
Differences w.r.t. resource allocation ...	... are perceived.	... should ...	... will be resolved.
Differences w.r.t. process implementation ...	... are perceived.	... should ...	... will be resolved.

Knowledge differences:

Different information status ...	... is perceived.	... should ...	... will be resolved.
Different information processing / perception ...	... is perceived.	... should ...	... will be resolved.

Evaluation differences:

Different interests / preferences ...	... are perceived.	... should ...	... will be resolved.
Different norms / values ...	... are perceived.	... should ...	... will be resolved.

Socio emotional tensions:

Different role expectations ...	... are perceived.	... should ...	... will be resolved.
Self-centered characters ...	... are perceived.	... should ...	... will be resolved.
Non-formalized confrontation patterns ...	... are perceived.	... should ...	... will be resolved.
Emotional tensions ...	... are perceived.	... should ...	... will be resolved.

Fig. 3. Profile for conflicts in NPD

	Conflict source:	Differences w.r.t. goals pursued	Differences w.r.t. resource allocation	Differences w.r.t. process implementation	Different information status	Different information processing / perception	Different interests / preferences	Different norms / values	Different role expectations	Self-centered characters	Non-formalized confrontation patterns	Emotional tensions	Reference:
Conflicts:	Latent	○	○	○	○	○	○	○	○	○	○	○	Deutsch (1976)
	Manifest	◔...●	◔...●	◔...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	
	Perception	○...●	○...●	○...●	◔...●	◔...●	○...●	○...●	○...●	○...●	○...●	○...●	Rüttinger (1980)
	Evaluation	○...●	○...●	○...●	○...●	○...●	◔...●	◔...●	○...●	○...●	○...●	○...●	
	Distribution	◔...●	◔...●	◔...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	
	Hot	○...●	○...●	○...●	○...●	○...●	○...●	◔...●	○...●	○...●	◔...●	◕...●	Glasl (1980)
	Cold	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○	○...◑	
	Functional	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○	Pinto, Kharbanda (1995)
	Dysfunctional	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	◔...●	
	Task	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○	Jehn (1995)
	Affective	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	◔...●	
	Task	◔...●	◔...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○	Jehn, Mannix (2001)
	Relationship	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	◔...●	
	Process	○...●	○...●	◔...●	○...●	○...●	○...●	○...●	○...●	○...●	○...●	○	

Scale: conflict source not present (○) ... present w.r.t. perception, resolution necessity, and intention (●).

Fig. 4. Classification scheme for conflicts in NPD

5 Summary

A new classification scheme for conflicts has been developed. It provides an integrative view on conflicts and can be used to recognize and handle conflicts in stage-gate approaches for NPD.

References

BADKE-SCHAUB, P. and FRANKENBERGER, E. (2004): *Management Kritischer Situationen: Produktentwicklung erfolgreich gestalten.* Springer, Berlin.

BERKMANN, F., BRAUNSPERGER, M., and POSCHNER, M. (2004): Die neue BMW K 1200 S. *Automobilzeitschrift (ATZ), 106, 7-8, 634–642.*

BOULDING, K. (1957): Organization and Conflict. *Journal of Conflict Resolution, 1, 122–134.*

COOPER, R.G. (1983): A Process Model for Industrial New Product Development. *IEEE Transactions on Engineering Management, 30, 1, 2–11.*

COOPER, R.G. (1994): Third Generation New Product Processes, *Journal of Product Innovation Management, 11, 1, 3–14.*

COOPER, R.G. and KLEINSCHMIDT, E.J. (1986): An Investigation into the New Product Process: Steps, Deficiencies, and Impact. *Journal of Product Innovation Management, Vol. 3, No. 3, 71-85.*

DEUTSCH, M. (1976): *The Resolution of Conflict. Constructive and Destructive Processes.* Yale University Press, New Haven.

DEUTSCH, M. (1990): Sixty years of conflict. *The International Journal of Conflict Management, 1, 237-263.*

DING, M. and ELIASHBERG, J. (2002): Structuring the New Product Pipeline. *Management Science, 48, 3, 343-363.*

GAUL, W. and BAIER, D. (1994): *Marktforschung und Marketing Management: Computerbasierte Entscheidungsunterstützung.* Oldenbourg, München.

GLASL, F. (1980): *Konfliktmanagement: ein Handbuch zur Diagnose und Behandlung von Konflikten für Organisationen.* 1st (8th 2004) ed., Huber, Bern.

GRIFFIN, A. and HAUSER, J.R. (1996): Integrating R&D and Marketing: A Review and Analysis of the Literature. *Journal of Product Innovation Management, 13, 191–215.*

HÖGL, M. (1998): *Teamarbeit in innovativen Projekten.* DUV, Wiesbaden.

JEHN, K.A. (1995): A Multidimensional Examination of the Benefits and Determinants of Intragroup Conflicts. *Administrative Science Quarterly, 40, 256–282.*

JEHN, K.A. and MANNIX, E.A. (2001): The Dynamic Nature of Conflict. *Academy of Management Journal, 44, 2, 230–251.*

KRISHNAN, V. and ULRICH, K.T. (2001): Product Development Decisions: A Review of the Literature. *: Management Science, 47, 1, 1–21.*

LAM, P.-K. and CHIN, K.-S. (2005): Conflict Bewtween Engineers and Marketers. *Industrial Marketing Management, 34, to appear.*

LECHLER, T. (1997): *Erfolgsfaktoren des Projektmanagements.* Lang, Frankfurt.

LEVINE, J. and MORELAND, R.L. (1990): Progress in Small Group Research. *Annual Review of Psychology, 41, 585–634.*

PELLED, L. H. (1996): Demographic Diversity, Conflict, and Work Group Outcomes. *Organization Science, 7, 6, 615–631.*

PINTO, J.K. and KHARBANDA, O.P. (1995): Project Management and Conflict Resolution. *Project Management Journal, 26, 4, 45–54.*

POSCH, A. (2003): Management von Innovationsprojekten. In: H. Strebel (Ed.): *Innovations- und Technologiemanagement.* WUV, Wien, 211–264.

REGNET, E. (2001): *Konflikte in Organisationen.* 2nd ed., Verlag für angewandte Psychologie, Göttingen.

RÜTTINGER, B. (1980): *Konflikt und Konfliktlösen.* Bratt, Koch.

SEIDEL, M. (1996): *Zur Steigerung der Marktorientierung der Produktentwicklung.* Difo-Druck, Bamberg.

SHAW, V. and SHAW, Ch.T. (1998): Conflict Between Engineers and Marketers: The Engineers Perspective. *Industrial Marketing Management, 27, 279–291.*

SHAW, V.,SHAW, Ch.T., and ENKE, M. (2003): Conflict Between Engineers and Marketers: The Experience of German Engineers. *Industrial Marketing Management, 32, 489–499.*

URBAN, G.L. and HAUSER, J.-R. (1993): *Design and Marketing of New Products.* Prentice Hall, Englewood Cliffs, NJ.

Analytical Lead Management in the Automotive Industry

Frank Säuberlich[1], Kevin Smith[2], and Mark Yuhn[2]

[1] Urban Science Int. GmbH,
Stresemannallee 30, D-60596 Frankfurt, Germany
[2] Urban Science Inc.,
200 Renaissance Center, Suite 1800, Detroit, MI 48243, USA

Abstract. Increasingly, vehicle buyers are using the internet as part of the purchasing process. As a service to the customer, a website can offer its visitors the opportunity to be contacted by a nearby dealer when they are ready to buy. These internet "leads" offer the ability to increase sales, enhance customer satisfaction, and develop a better understanding of buyers who use the internet. In order to take advantage of this situation, automotive manufacturers have to be able to capture website leads reliably and deliver them to the appropriate dealers directly. Furthermore, performance metrics should be implemented that enable manufacturers and dealers to improve their sales and marketing performance.
In this paper we describe an approach which is designed to maximize return on leads generated from internet websites. The core of our approach is a statistical model that is able to score each lead to determine the likelihood that it will result in a sale and prioritize each lead before sending it to the appropriate dealer. An operational lead management project for Audi of America is used for illustrative purposes.

1 Introduction

Recent studies of automotive customer behaviour show that the importance of the internet as a sales channel for cars is growing. JupiterResearch estimates that 22% of all automotive sales in 2004 were generated from the web - a number that is forecasted to grow (JupiterResearch (2004)); the percentage of new-vehicle buyers who use the internet in their vehicle shopping process remains steady at 64% (J.D. Power (2004a)).

The internet offers convenience for consumers to initiate contact with the dealership of their choice. Customer contacts initiated from the internet, so called internet "leads", are a growing part of a manufacturer's strategy to increase sales. In the course of this development, dealers are receiving more vehicle leads from manufacturers than was the case in 2003 (J.D. Power (2004b)). In contrast to this development, dealers often lack trust in internet generated customer contacts or don't have the resources to handle the increasing quantity of incoming leads efficiently. Audi of America found out in an internal study in 2002 that 48% of all leads sent to the dealers weren't followed-up at all (McGough and Säuberlich (2002)).

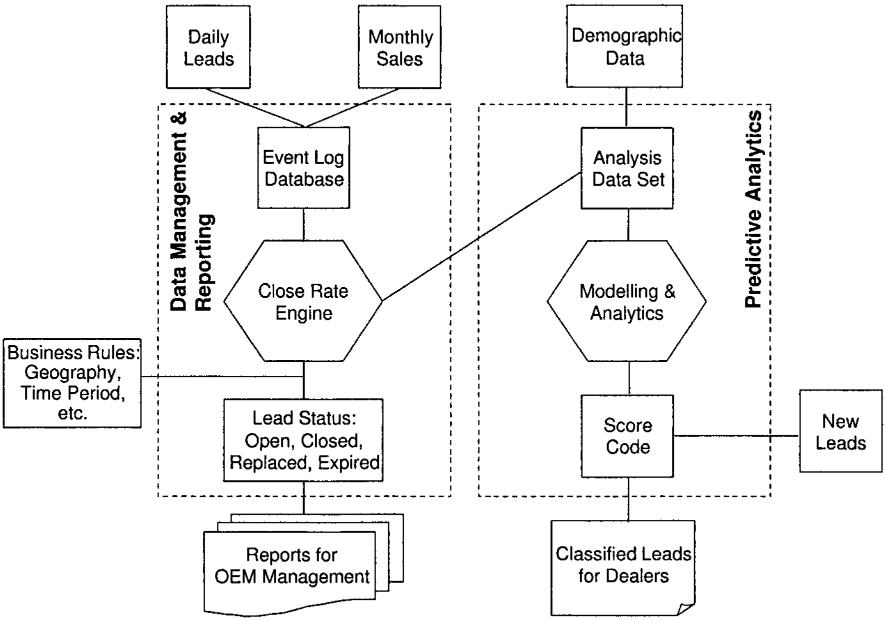

Daily Leads
Monthly Sales
Demographic Data
Data Management & Reporting
Event Log Database
Close Rate Engine
Business Rules: Geography, Time Period, etc.
Lead Status: Open, Closed, Replaced, Expired
Reports for OEM Management
Predictive Analytics
Analysis Data Set
Modelling & Analytics
Score Code
New Leads
Classified Leads for Dealers

Fig. 1. Analytical lead management approach, overview.

Internet leads can be simple requests for information or specific requests for a price quote for a vehicle that a consumer has configured on a website. They may be sourced from the manufacturer website, a dealer site, or a third party website (e.g., Edmunds, KBB). Internet leads are productive, each lead provider, manufacturer or other users track the percentage of leads resulting in a sale, the so called close rate percentage. But not all leads are created equal. In principle, each lead is important and should be pursued. However, some leads may be more productive than others. If these "best" leads can be systematically identified in advance, then this differentiation can be exploited. Shifting additional marketing efforts to the best leads may result in a higher overall close rate.

Manufacturer's key strategy to sell more vehicles is to leverage existing lead management processes. Manufacturers need to be able to take their list of leads and classify them, e.g. into two categories - high priority and normal priority leads. A high priority lead can be defined, for example, as the top 30% of the leads. Though all leads in principle are important, these high priority leads should merit specialized treatment like extra attention from a salesperson. This lead prioritization then helps dealers to focus on important leads first, structure their lead handling efforts and above all increase their trust in internet generated customer contacts.

In this article we describe an approach that automatically captures lead information, classifies each lead into classes of high or normal priority and forwards it to the corresponding dealership in real time. In addition, a multitude of lead management reports summarizing close rates at different aggregation levels are generated for dealer and manufacturer sales management. Figure 1 shows an overview of this approach. In general, it consists of two main modules: data management & close rate reporting as well as predictive analytics. In the following sections we describe these two modules in detail. In addition we summarize a lead management project done together with Audi of America showing actual operating results.

2 Data management and close rate reporting

The first step, data management, reads in the raw lead information together with sales data and combines them into a so called event log database. The second step, the close rate engine uses the event log to assign each lead's status: open, closed, replaced or expired based upon specific business rules and then close rate reports are prepared.

2.1 Data management

A lead is created by one of several means: an internet user may request a price quote while visiting an online automotive marketplace like www.edmunds.com, a customer may request a brochure from the manufacturer, a customer may visit the manufacturer website and request dealer follow-up, etc. In all these cases, information about the consumer is collected. To be able to use this lead data for computer modelling purposes, the lead is required to contain customer name and address, as well as, email and/or phone contact information. After receipt, a variety of data management steps are conducted:

- Zip code, phone number and email address supplied by the prospect are validated to identify leads that may be difficult or impossible to follow-up upon; address information is cleansed and standardized to improve the match rate with other data sources.
- Multiple leads from the same prospect are de-duped into a single prospect lead, since individuals frequently generate multiple leads in their car-shopping process, yet only intend to buy a single car; in all future considerations these cleansed leads are used.
- Previous owner status of a lead is appended by matching with the manufacturer's transaction history database to determine if the prospect is a repeat buyer.

After the appropriate time period (90 days in the case of Audi; see section 2.2), the lead file is matched with an up-to-date sales file to determine which leads have resulted in sales. This match is accomplished using customer name and address information. This combined lead and sales information then is stored in an event log database which contains a record of all events surrounding each prospects/customers lead and sales activities. The information is stored in detail, meaning one data row for each event (lead or sale, with customer/prospect key and time stamp), and is used as the basis for all following data analysis steps. Additional information can be placed in the event log database for future analytic use, such as lease expiration information, inbound or outbound communication flags indicating call center, direct mail and email contact activity.

2.2 Close rate engine

The close rate engine is the heart of the data management & reporting module. The close rate engine uses the event log database together with business rules, which are based on expert knowledge and/or preferences of the report users, to generate close rate reports. The status of each lead has to be determined and can exist as one of the following four types:

- **Open** - the prospect hasn't bought a vehicle yet and the lead generation date isn't older than the specific number of days under consideration (the number of days is one of the business rules/parameters given by the expert/user; within the project with Audi of America a time period of 90 days was used, since roughly 95% of all leads resulting in a sale occurred within 90 days after the lead was generated).
- **Closed** - the prospect has bought a vehicle within the time period under consideration.
- **Replaced** - the prospect of an open lead has generated a new lead, that replaces the old request (the day count for monitoring the closing behaviour of this lead is set to 0 again).
- **Expired** - no vehicle sale is recorded for an open lead within the time period under consideration; new requests from the same prospect are handled as new open leads.

Based on these lead types, specific close rates can be computed directly by dividing the number of closed leads of a specific time period (normally a month) by the overall number of leads generated in that period (apart from the leads replaced). Close rates can be computed for different observation levels, e.g. close rates by dealer, vehicle model, lead source, area, etc., based on the business rules/parameters used.

This close rate information then is summarized in reports for the manufacturer to help guide enterprise lead strategy. When the organization pays for leads from a lead generator, close rates identify which generators provide the greatest return on investment.

Dealer level close rates can reveal operational issues and assist field staff in improving dealer performance. For a given dealer, current and historical close rates, lead volumes, a list of open leads, and even a count of the dealer's leads that ended up buying from another same-brand dealer can be computed. By tracking lead generation across multiple generation points a profile of how prospects shop on the internet can be derived. How many leads does a prospect generate before purchasing? What is the duration of his shopping period? Are prospects using multiple websites to generate leads or do they just stay with one? Answers to these questions can help guide on-line advertising strategies.

Close rates and lead follow-up times can be examined together to assess the impact of rapid follow-up. Industry research suggests the opportunity to convert a lead into a sale drops rapidly as follow-up time grows. In summary, the information available from close rate analysis provides valuable strategic information to the sales management, e-business, marketing, product planning, and field staffs of an automotive manufacturer.

3 Predictive analytics

The data management and close rate reporting module described above provides important metrics that measure the effectiveness of sales and marketing investments by automotive manufacturers. Predictive analytics addresses specific issues that have developed between manufacturers and dealers such as: too many leads to respond to in a timely manner, and dealer uncertainty regarding the quality of internet generated customer leads. The ability to classify leads into categories, and then differentiate marketing treatment based upon this classification is one way to deal with these issues. The predictive analytics module handles this functionality.

3.1 Analysis data set

The predictive analytics step starts with generating a dataset for analysis. This dataset consists of the output of the close rate engine, using lead data processed with information about the lead stage. Only open or closed leads are used, whereas in case of a replaced lead, the newest lead information of the particular prospect is used.

Additional demographic, customer transaction history, or other data is searched for; if found, the lead is then augmented with information from those databases. By doing so, lead characteristics (meaning behavioural information of the prospect) can be used for modelling, as well as, additional appended information. The result is an analysis dataset, with one data row for each prospect and several columns of prospect characteristics. The lead status information is stored in the form of "closed (yes/no)" and will be used in the modelling step as so called dependent variable. Table 1 shows a selection of the variables used within the Audi of America project.

Category	Sub-Category	Variable Name
Behavioural	Vehicle Characteristics	Requested Vehicle Model Requested Engine Type Requested Body Style Quattro (yes/no) Power Sunroof (yes/no) Navigation System (yes/no) Sound System (yes/no) Xenon Lights (yes/no) Number of Options Exterior Gray (yes/no) Exterior Silver (yes/no) Exterior White (yes/no) Exterior User Inputted (yes/no) Interior Beige (yes/no) Interior Black (yes/no) Interior Ebony (yes/no) Interior Cloth (yes/no) Interior Leather (yes/no) Interior User Inputted (yes/no)
	Prospect/Lead Characteristics	Previous Owner (yes/no) Best Time for Contact Preferred Mode of Contact Lead Source Comment Length
Demographic	Household/Prospect Characteristics	Number of Adults in Household Number of Children in Household Estimated Household Income Education Level Prospect Occupation of Prospect Age of Prospect
	Dwelling Characteristics	Dwelling Type Home Market Value Region Length of Residence
Dependent	Lead Status	Closed (yes/no)

Table 1. Selected descriptive variables used within the Audi of America project.

3.2 Modelling and analytics

The general objective of the predictive model is to be able to classify leads into two categories - high priority and normal. The model will meet these objectives by identifying characteristics about the consumer (behavioural and demographic; see Table 1) that reliably increase the odds that the consumer will purchase a vehicle. The starting point of the analysis is the analysis dataset described in section 3.1 with the dependent variable "Closed". This analysis dataset is randomly split into a training (two thirds) and validation

CUMULATIVE RESULTS BY DECILES

SALES PROB %	% LEADS	% SALES	CLOSE RATE %	% SALES / % LEADS
28.96	10.0	21.4	33.83	2.1
23.14	20.0	37.4	29.64	1.9
19.27	30.0	51.1	27.00	1.7
16.41	40.0	61.4	24.32	1.5
13.93	50.0	71.6	22.67	1.4
11.79	60.0	79.4	20.97	1.3
9.63	70.0	86.8	19.65	1.2
7.70	80.0	92.7	18.36	1.2
5.57	90.0	97.5	17.16	1.1
0.80	100.0	100.0	15.84	1.0

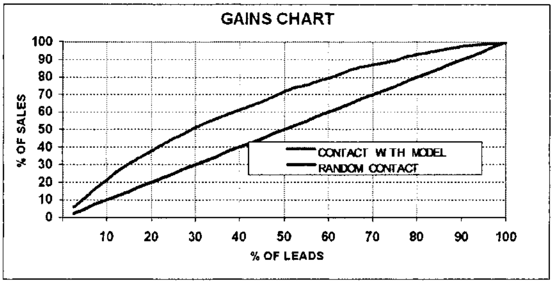


Fig. 2. Predictive accuracy on validation data file; selected results from Audi of America project.

(one third) dataset. The model is built on the training file. The validation file is used for evaluating the predictive accuracy of the model in an objective way: Once the final model is chosen, the validation file is scored and then ranked into deciles so that predicted sales can be compared to actual sales.

Based upon the modelling objective and the data availability, we chose to use logistic regression as the modelling methodology. Because of the large number of independent variables, separate blocks of variables are tested (e.g., all demographic, all behavioural variables) to identify promising variables for the final model and to sort out variables that aren't contributing predictive value. Results from these preliminary models are used to reduce the number of variables into a more manageable set. For variable selection and model calibration the data mining tool GainSmarts is used (Levin and Zahavi (2002)).

Figure 2 shows the predictive accuracy of a logistic regression model built with GainSmarts for Audi of America. The results shown are based on the validation data file. The model did well in predicting Audi sales. After scoring and ranking the validation file, 10% of the best leads included 21.4% of the (actual) sales. The 30% best leads included 51.1% of the sales. If we evaluate predictive model performance in terms of close rates, the top 10% had a close rate of 33.8% within 90 days, whereas the overall dataset had a close rate of 15.8%. The top 30% had a close rate of 27% compared to a close rate of 11.1% of the remaining 70%.

Both behavioural (e.g., purchase timing, comment length) and demographic variables (e.g., age, income) were included in the final model. Many of the variables selected were expected to be able to predict sales (e.g., car model, lead source) but a few chosen variables were a surprise (e.g., travel and entertainment card holder). By interpreting the individual parameters of the final logistic regression model, it is possible to describe a profile of high scoring vs. low scoring leads. Table 2 shows selected variables that differentiate high from normal priority leads.

The results are plausible. Audi is a luxury brand and actual buyers are more likely to be affluent. This is not only seen in household demographics (income and home value), but also in behaviour. Comment length is indicative

High Priority Leads	Normal Priority Leads
interested in an A4 or TT	interested in an A8, S4, or S6
higher household income	lower household income
purchase status of now or within two weeks	purchase status of greater than one month
lengthy open-ended comments	no or very short open-ended comments
exact name of exterior colour within comments	cloth interior as option

Table 2. Selected variables describing the profile of high vs. normal priority leads.

of interest and/or involvement with the brand and show a willingness to dialog with the dealer at a deeper, more personal level. Similarly, providing the exact name of a vehicle colour suggests greater knowledge of the Audi brand. Choice of the significant colour codes also suggests not only individual tastes but also interest, since the customer is not simply choosing the default values for these fields.

Besides these valuable insights of high priority leads vs. normal priority leads, the final deliverable of the predictive analytics step is the so called scoring code that applies the model to new incoming leads. We generate this code automatically in C#, so that it can be applied to incoming leads in real time on any hardware platform and predicts for each lead a propensity to buy probability. This probability then is used to assign the lead to one of the two priority classes (high or normal), based on a probability threshold (all leads with a propensity to buy probability greater or equal a specific threshold are assigned to the high priority class, the remaining leads are assigned to the normal priority class; for the Audi of America project, the lowest propensity to buy probability of the top 30% of the leads is used as the probability threshold). These leads together with the priority class information then are sent to the appropriate dealer.

4 Process and implementation

In this section we cover all the tasks needed to ensure that the lead is delivered to the dealer and that the response data from the participating dealers' (operational) lead management system is acquired for accurate close rate computations and predictive modelling analytics described before. This process has several components:

- **Lead Receiving** - Leads are accepted in ADF format (Auto Lead Data Format; An Industry Standard Data Format for the Export and Import of Automotive Customer Leads using XML) from manufacturer-approved lead generators such as the brand sites, dealer sites, or partner lead sites.
- **Lead Acceptance** - Leads are accepted and processed only if the source has been authorized to send leads through the system. Leads are checked

to see if the lead is a test lead or if the same prospect has generated a similar lead within the past 30 days. If so, the lead generator is notified that the manufacturer will not pay for that lead. Additionally, if it is determined that the exact same lead has already been received and processed in the past 5 minutes, then the lead is excluded and not sent to the dealer and the lead source is notified that the manufacturer will not pay for that lead.

- **Lead Validation** - Incoming data is validated (see section 2.1). The format and content is checked for accuracy. Additionally, the syntax, domain name, and user name of the email address of the prospect is validated. If the syntax check fails, a default email address is substituted.
- **Lead Enrichment** - The lead data is enhanced by appending demographic data from a data provider. To do this, a subset of the incoming data stream is posted to the data provider, who is posting back any demographic data that is found in their system for that prospect. This occurs in real time. Additionally, transaction history is added to the lead. Previous owner status with vehicle and year purchased is determined and stored with the lead info.
- **Lead Standardization** - lead data is standardized for analysis across all lead generators. Therefore, the model can accurately compare the various lead sources based on the standardized fields.
- **Lead Scoring** - The scoring code described in section 3.2 uses lead data as well as the enriched and standardized data to determine the likelihood to purchase. If the score meets a predetermine threshold, then the lead is flagged as high priority. This also occurs in real time.
- **Lead Delivery** - The appropriate dealer to send the lead to is determined. Depending on the lead source, the lead delivery has to be handled differently. If the lead source is from a dealer website, or if a preferred dealer is indicated, then that dealer is used. If the lead source is from a brand site or a partner lead source, then the zip code is used to determine the appropriate dealer. After the dealer is determined, the dealer's email address is retrieved from a dealer email table. A dealer can have multiple email addresses and a different email format for each address (either HTML or ADF). The syntax, domain name, and user name is validated for all dealer emails. If a dealer has no valid email address, a default email address is used. After the email address and format is determined, the appropriate lead information is packaged and emailed to the dealer. Additionally, other email addresses can be added to the outgoing dealer emails, e.g. field teams or corporate personal can be carbon copied on specific dealer leads.
- **Lead Storage** - The lead information, as well as the enriched and standardized data is retained for future close rate reporting and analytics as described in sections 2 and 3.

5 Summary

The results of the logistic regression model shown in section 3 are impressive. They meet the objective of classifying leads as high and normal priority. The top 30% of leads represent 51% of the Audi sales. Among the top 30%, a total of 27% purchased a new Audi vehicle within 90 days. Among the remaining 70%, a total of 11% purchased a new Audi vehicle within 90 days. These results are even more impressive because leads are consumer-initiated, meaning they are hand-raisers who are currently in market for a new vehicle. Therefore all of these leads are, theoretically, good prospects. A predictive model that can further refine this list demonstrates added value.

The model addresses a manufacturer's strategy of being able to prioritize leads. Though in principle all leads are important, there are always practical issues of time and resources. The priority classification is based on validated past close rates and therefore is scientific as opposed to an "educated guess". It helps a dealership to focus its efforts on a list of leads, spending the most time and energy on the potentially most productive leads and therefore increases the trust of dealers in internet generated customer contacts.

Besides using the scoring model to qualify leads and attach a prioritization flag to each, lead the analytical lead management approach shown also helps to evaluate dealer performance in an objective manner. The distribution of lead close rates for dealers varies; but this may be for example simply due to geographical differences. By using the more homogenous group of high priority leads, comparing close rates over different dealers is meaningful. Differences found cannot be explained by lead quality but are likely due to dealer operator characteristics. Therefore manufacturers can better identify their best dealers from those which need improvement.

In summary, the analytical lead management approach shown clearly improves the value of customer lead information, therefore improving the efficiency and effectiveness of automotive sales and marketing expenditures.

References

J.D. POWER (2004a): New Autoshopper.com Study 2004. J.D. Power & Associates, Westlake Village, California.

J.D. POWER (2004b): Dealer Satisfaction with Online Buying Services. J.D. Power & Associates, Westlake Village, California.

JUPITERRESEARCH (2004): Effectively Using the Internet to Drive Sales in the Channel. JupiterResearch, New York.

LEVIN, N. and ZAHAVI, J. (2002): GainSmarts. In: W. Klosgen and J.M. Zytkow (Eds.): *Handbook of Data Mining and Knowledge Discovery.* Oxford University Press, NY, 601–609.

MCGOUGH, J. and SÄUBERLICH, F. (2002): Demonstrating CRM ROI to Stakeholders - Case Study: Audi of America. Presentation at the Second Pan-European Automotive Forum, Automotive CRM 2002, Berlin, 23.-24. October 2002.

Die Nutzung von multivariaten statistischen Verfahren in der Praxis - Ein Erfahrungsbericht 20 Jahre danach

Karla Schiller

INFORMA Unternehmensberatung GmbH, D-75179 Pforzheim, Germany

Abstract. Basierend auf einer empirischen Studie zur tatsächliche Nutzung von multivariaten statistischen Methoden in der Praxis deutscher Unternehmen aus dem Jahre 1984 wird anhand einer aktuellen Experten-Umfrage untersucht, ob und wie sich das Methoden-Nutzungsverhalten in Unternehmen verändert hat. Es zeigt sich, dass die Relevanz statistischer Analysen deutlich zugenommen hat. Das methodische Spektrum hat sich auch verändert, wenn auch nur langsam. So bilden z.B. die sogenannten Baumverfahren inzwischen einen wesentlichen Baustein des Methodenspektrums, während sie in den 80-er Jahren noch relativ unbedeutend waren. Die deutschen Unternehmen zeigen sich also durchaus offen für innovative Methoden, wenn sie denn den Güte-Kriterien aus der Praxis gerecht werden. Hier liegt für zukünftige Entwicklungen noch ein erheblicher Verbesserungshebel.

1 Einleitung

Seit vielen Jahrzehnten stehen die Prüfung und Entwicklung von komplexen statistischen Verfahren im Fokus der Wissenschaft. Dabei beschäftigen sich nicht nur die reinen mathematischen und statistischen Bereiche mit diesen Fragestellungen, sondern auch stark methodisch-statistisch ausgerichtete Bereiche in den sogenannten Anwendungswissenschaften wie z. B. Wirtschaftswissenschaften, Sozialwissenschaften, Psychologie, Medizin (Pharmakologie, Epidemiologie,...) und weitere mehr.

Gerade bei den anwendungsorientierten Forschungen könnte es von Interesse sein, ob die im wissenschaftlichen Bereich entwickelten oder verbesserten Methoden auch Relevanz für die Anwendung in der Praxis haben oder ob es sich hier um eine rein akademische Beschäftigung handelt.

Vor diesem Hintergrund führte das Institut für Entscheidungstheorie und Unternehmensforschung an der Universität Karlsruhe 1984/85 eine empirische Studie durch, an der ca. 100 deutsche Marktforschungsinstitute oder ähnliche Abteilungen in Großunternehmen teilnahmen und über die Nutzung von Datenanalysemethoden, ihre Beweggründe und Hemmnisse Auskunft gaben (vgl. Gaul et al. (1986a, 1986b)).

Aufbauend auf diesen Ergebnissen aus den 80er Jahren wird in dieser Arbeit eine kurze Beleuchtung der aktuellen Nutzung der Datenanalysemethoden in der Praxis der Business-Anwendungen durchgeführt. Hierbei konzentrieren wir uns auf einen speziellen Ausschnitt der Business-Anwendungen:

Dem Bereich der Prognose von Kundenverhalten und Umsetzung in eine optimierte Kundensteuerung.

Im Gegensatz zur damaligen "repräsentativen Erhebung" dienen zur Bewertung der aktuellen Situation zum einen die persönliche Einschätzung der Autorin auf Basis von 20 Jahre Praxiserfahrung und eine "Blitzumfrage" unter ausgewählten Anwendungsexperten.

2 Review: Die Ergebnisse der empirischen Studie 1984/85

Die damalige Erhebung (Gaul et al. (1986a, 1986b)) beschränkte sich im wesentlichen auf Marktforschungsinstitute, da es bei diesen Unternehmen als professionelle Nutzer von Datenanalysetechniken am ehesten erwartet wurde, dass sie innovative Datenanalysetechnologien in ihrem Unternehmen einsetzen. Parallel dazu wurden auch Marktforschungsabteilungen in großen Unternehmen befragt.

Im Rahmen der Auswertung wurden die 24 erhobenen statistischen Verfahren in 5 Hauptgruppen gegliedert:

Gruppe 1: Deskriptive Verfahren – Darunter zählen Grundauszählungen, Kreuztabellen, Mittelwertstandardabweichung.

Gruppe 2: Testverfahren (zum Testen der Signifikanz eines Einzeleffektes) – Dazu zählen T-Test, Anpassungstest, nicht-parametrische Verfahren, simultane Testverfahren, F-Test.

Gruppe 3: Multivariate Standardverfahren – Hierzu zählen Korrelations-Analyse, Faktoren-Analyse, Varianz- und Regressions-Analyse, Cluster-Analyse, Diskriminanz-Analyse, MDS (Multidimensionale Skalierung).

Gruppe 4: Multivariate Spezialverfahren – Hierzu zählen: Kontingenztafel-Analyse, loglineare Modelle, AID/Tree-Analysis, Conjoint-Analysis, Manova, Kausal-Analyse.

Gruppe 5: Standardverfahren der Zeitreihen-Analyse/Prognose – Zeitreihen-Verlegung (Saison, ...), exponentielles Glätten, Verlaufskurven.

Gruppe 6: Spezialverfahren der Zeitreihen-Analyse/Prognose – Ökonometrische Modelle, Box-Jenkins-Verfahren.

Es zeigte sich, dass die deskriptiven Verfahren im wesentlichen von allen Unternehmen sehr häufig verwandt werden. Auch die "klassischen" multivariaten Verfahren der Datenanalyse, wie Varianz- und Regressions-Analyse, Faktoren-Analyse oder Korrelations-Analyse werden von über 3/4 der Befragten genutzt, jedoch nur von ca. 50-60% manchmal oder häufig. Die Testverfahren zum Testen des Einflusses von einzelnen Effekten waren bekannt

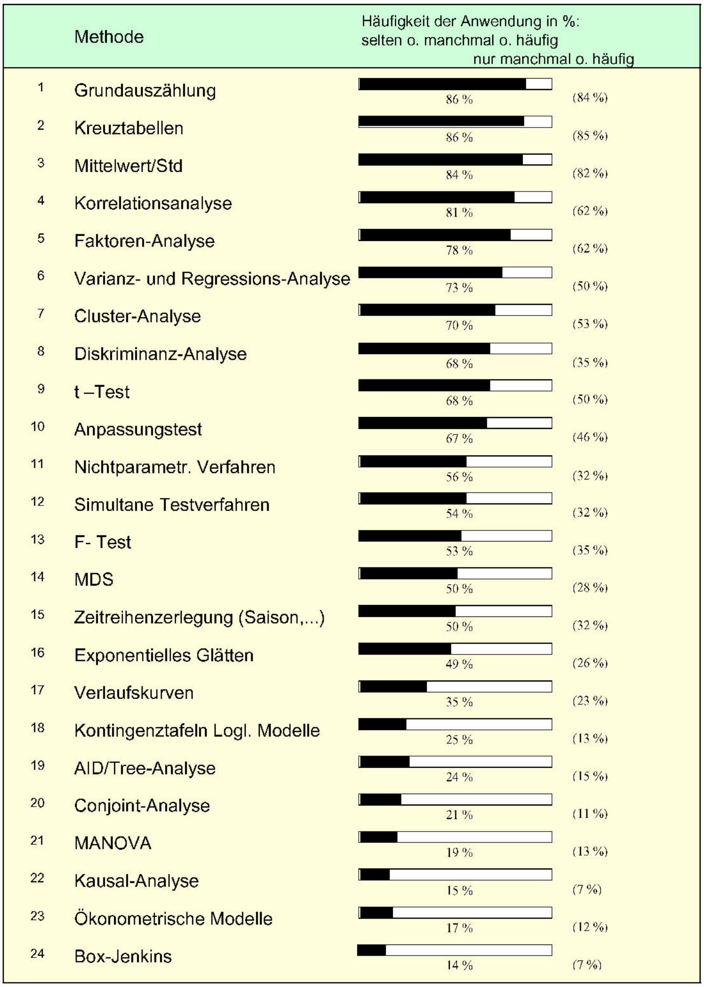


Fig. 1. Methoden-Nutzung intern vgl. Gaul et al. (1986a, 1986b)

und wurden gelegentlich genutzt (über 50%), jedoch ist schon hier ein deutlicher Abfall zur intensiveren Nutzung zu verzeichnen (ca. 30%).

Außerdem gab es noch eine Spezialanwendergruppe von Zeitreihen- und Prognoseverfahren, die auch diese Verfahren häufig bis regelmäßig nutzten, darüber hinaus war die Nutzung von neueren statistischen Verfahren im Bereich der Marktforschung in den 80er Jahren als sehr gering zu bezeichnen.

Bei den Gründen für die Nutzung bzw. Nicht-Nutzung der komplexeren Datenanalyseverfahren überwogen eindeutig die Gründe für die Nutzung. Insbesondere kann man oftmals nur mit den komplexeren Methoden Auswertungen machen, die mit einfachen Methoden nicht zu erhalten sind. Softwarebedingte Nutzungsengpässe (bzgl. Rechenintensivität, Benutzerfreundlichkeit, Dokumentation oder auch Preis) waren kaum relevant. Lediglich "externe Gründe" wurden als ausschlaggebend für die "Nicht-Nutzung" angesehen, insbesondere, dass der Kunde die Verfahren nicht kennt, nicht akzeptiert oder ihm solche Auswertungen zu teuer sind.

Als interessantes Ergebnis der Studie wurde abschließend erwähnt, dass Marktforschungsinstitute "keinerlei Gefahr sehen", dass komplexere Analysen durch ihre Kunden selber durchgeführt werden, da sie - die Marktforschungsinstitute - sich nach wie vor in der Rolle der Spezialisten in Datenanalysetechniken sehen. Diese Einschätzung wurde vor dem Hintergrund der schnellen Expansion des Hard- und Softwaremarktes gerade im Bereich der Datenanalyse von den Autoren kritisch hinterfragt.

3 Veränderungen im Markt, Trends im Bereich der statistischen Datenanalyse

In den letzten 20 Jahren hat sich, wie damals bereits prognostiziert, der Markt der statistischen Datenanalyse deutlich verändert, sowohl was die Hard- als auch die Software betrifft.

1. **Hardware-Veränderungen**
   Während in den 80er Jahren der Bereich der Datenanalyse deutlich Großrechner-orientiert war - gekennzeichnet durch nicht intuitiv bedienbare Betriebssysteme, Druckausgabe auf Endloslisten, kaum grafische Benutzeroberflächen und auch wenige und teuere grafische Ausgabemöglichkeiten - haben sich in den letzten 20 Jahren diese Anwendungsszenarien deutlich verändert:
   - Verlagerung vom Großrechner hin zu Windows-basierten PC-Lösungen, deren Anwendung auch für den Benutzer eine nicht so große Hürde darstellt.
   - Intensität der Verbreitung. Durch diese Verlagerung vom Großrechner hin zum PC steht fast jedem interessieren Anwender die Möglichkeit zur computergestützten Datenanalyse zur Verfügung.
2. **Software-Entwicklung**
   Die für den Bereich der statistischen Datenanalyse zur Verfügung stehende Software hat sich in den letzten 20 Jahren enorm erweitert. Dies betrifft zum einen die großen statistischen Softwarepakete, wie z. B. SAS und SPSS. Hier hat sich ausgehend von einem Standardpaket eine enorme Produktvielfalt entwickelt, bis hin zu großen Speziallösungen für Datenanalyse (z. B. SAS-Enterprise Miner oder SPSS Clementine). Diese sehr stark Anwendungs-orientierten Tools sind speziell für Data-Mining Anwender entwickelt worden, um hier eine in sich konsistente Benutzerführung und projekt-orientierte Bearbeitung möglich zu machen. Neben dem stark expandierenden Softwareangebot der "Allround-Anbie-

ter", die zusätzlich zu den reinen Datenanalyse-Methoden auch noch eine Vielzahl von begleitenden und vorbereitenden Möglichkeiten zur Datenvor- und aufbereitung bieten, gibt es inzwischen eine Fülle von Spezialsoftware, die für einen bestimmten Anwendungsfokus bzw. für eine bestimmte Anwendungs-Methode eine fundierte und ausgereifte Softwareumsetzung bieten. Hier sind z. B. verschiedenste Verfahren der Entscheidungsbäume (z. B. Prudsys) oder auch Visualisierungstechniken (z.B. Eudaptics) zu nennen.

3. **Tagungs- und Kongressangebot**
Begleitend zu dem stark expandierenden Softwareangebot gibt es ein stark expandierendes Tagungs- und Kongreßangebot zum Thema Datenanalyse und statistische Verfahren in der kommerziellen Nutzung. Dies wird zum einen durch die Softwareanbieter selber forciert, zum anderen aber auch durch entsprechende Berufsverbände oder Tagungsorganisatoren. Beide Entwicklungen zusammen führen dazu, dass im Laufe der vergangenen Jahre verschiedenste "Trends" bei Datenanalyse-Verfahren am Markt zu beobachten waren. Die verwendeten Labels kamen und gingen, ihre Inhalte waren zum Teil sehr ähnlich. Hier seien nur einige genannt: Neuronale Netze, Data-Mining, Analytical CRM, Business Intelligence, E-Mining, vgl. Gentsch (2002) und Martin (2002).

Alle hier beschriebenen Trends führen dazu, dass Datenanalyse-Techniken inzwischen ein anerkanntes Instrument in den Unternehmen geworden sind. In vielen Unternehmen existieren eigene Abteilungen oder werden aufgebaut, um Datenanalyse-Fragestellungen zu bearbeiten.

4 Aktueller Anwendungsfokus: Scoring-Modelle als Einsatzgebiet für komplexe statistische Verfahren

4.1 Ausgangssituation

In den letzten Jahren hat in den Business-Anwendungen die Nutzung von komplexen Verfahren der Datenanalyse auch für Marketing-Fragestellungen eine zentrale Bedeutung gewonnen. Hierbei handelt es sich vor allem um Unternehmen, die im Massenkundengeschäft arbeiten und denen ihre Endkunden (über eine vertragliche Beziehung) zumindest teilweise bekannt sind. Hierunter fallen im wesentlichen die Branchen:

- Kreditwirtschaft,
- Versandhandel,
- Telekommunikation,
- Versicherungen.

Für diese Unternehmen spielt es eine große Rolle, Kenntnis über das zukünftige Kundenverhalten mit Hilfe von statischen Prognoseinstrumenten zu erhalten, um so die Kundensteuerung deutlich zu verbessern:

- Intensivierung des Geschäftes bei ertragsstarken Kunden,
- Reduzierung des Geschäftes bei ertragsschwachen Kunden.

Durch die systematische Generierung von Erkenntnissen über das Kundenverhalten und die Umsetzung in Kundenbewertung und Steuerung kann in den Unternehmen ein erheblicher Rentabilisierungshebel generiert werden vgl. Berry und Linoff (2000).

4.2 Methodische Aufgabenstellung

Aus diesen unternehmensstrategischen Aufgabenstellungen ergibt sich eine klare methodische Herausforderung, die mit Hilfe von multivariaten statistischen Verfahren gelöst werden kann und soll.

Dabei kristallisieren sich zwei Hauptaufgabengebiete heraus:

1. Erstellen von Scoring-Modellen zur Prognose von Kundenverhalten. Je nach Geschäftsgebiet des Unternehmens können hier sehr unterschiedliche Fragestellungen bzw. Kundenverhalten im Fokus stehen:
   - Zahlungsverhalten zur Steuerung der Bonitätsprüfung,
   - Limitvergabe und Mahnsteuerung,
   - Kaufverhalten zur Steuerung der zukünftigen Werbeintensität, Werbewege, spezielle Produktangebote,
   - Stornoverhalten zur Steuerung von Kündigerprophylaxe, bzw. Kündigerrückgewinnungsaktionen.
2. Erstellung von Kundensegmentierungs-Modellen/Kundentypologien Hier geht es darum, den gesamten Kundenstamm in max. 5-10 verschiedene Kundengruppen zu strukturieren, die möglichst prägnant beschreibbar sind und sich bzgl. ihres Kundenverhaltens deutlich unterscheiden. Auf Basis von diesen Kundengruppen bzw. -typen kann dann die weitere Unternehmensplanung und -steuerung aufbauen:
   - Reportingzwecke:
     - Fortschreibung der Kundengruppenentwicklung im Rahmen der Zeit,
     - Beurteilung der Qualität der Kunden bzgl. Kaufverhalten/Zahlungsverhalten.
   - Basis für Unternehmensplanung und -entwicklung und differenzierte Analysen.

4.3 Relevanz von komplexen statistischen Methoden für diese Aufgabenstellung

In den vorgenannten unternehmerischen Aufgabenstellungen spielen komplexere statistische Datenanalysemethoden zur Bearbeitung und Lösung eine zentrale Rolle.

Gilt es doch hier, verschiedenste Einflussfaktoren bzw. Charakteristika des Kundenverhaltens zu beurteilen und zu bündeln,

- um das zukünftige Kundenverhalten möglichst gut und genau zu prognostizieren oder
- "typische" Kundengruppen zu erkennen.

Beide Aufgabenstellungen sind typische Anwendungsgebiete komplexer statistischer Verfahren.

4.4 Erfolgsfaktoren für die Nutzung statistischer Verfahren

Fragt man erfahrene Anwender nach den Faktoren für eine erfolgreiche Anwendung solcher komplexen statischen Methoden in der Business-Praxis, so ergibt sich i.d.R. folgendes Bild:

1. Wichtigster Erfolgsfaktor für die Erstellung und Anwendung von Scoring-Modellen ist die richtige **Datengrundlage**. Nur die Bereitstellung, Generierung und Aufbereitung entsprechend relevanter, prognosefähiger Daten liefert am Ende auch gute Prognose-Modelle. Diese Daten entstehen i.d.R. in den Unternehmen selbst im Rahmen der operativen Geschäftsprozesse. Sie müssen jedoch so aufbereitet werden können, dass aus ihnen leistungsfähige Indikatoren für das zukünftige Kundenverhalten generiert werden.
2. Der zweite nahezu ebenso wichtige Erfolgsfaktor ist die **Erfahrung** und das Know-how des Anwenders der Analyse-Methoden. Hierbei geht es zum einen um das Business-Know-how. Zur Erstellung von guten Prognose-Modellen zur Prognose von zukünftigen Kundenverhalten, ist es fundamental wichtig, die zugrunde liegenden Geschäftsprozesse sehr genau zu kennen. Man muss wissen, wie sich das historische Verhalten eines Kunden in den Datenströmen des Unternehmens wiederfindet, welche Steuerungsmaßnahmen das Unternehmen in der Vergangenheit ergriffen hat und welche Veränderungen das Unternehmen für die Zukunft plant. Man muss einschätzen können, wie sich diese Veränderungen auf zukünftiges Kundenverhalten auswirken können, ohne selbst dafür eine historische Datenbasis zu haben. Man muss die Einflüsse der Konkurrenz abschätzen können. Neben dem Business-Know-how des Anwenders spielt auch das methodische Know-how des Anwenders eine Rolle. Er sollte ein Grundverständnis über Prinzipien und Vorgehensweisen des statistischen Modellierens haben. Da bei solchen unzitierten Aufgabenstellungen i.d.R. ein großer Pool von erklärenden Variablen zur Verfügung steht (ca. 200 bis 1.000 erklärende Variablen) und ggf. auch eine Fülle von u. U. konkurrierenden Zielvariablen existiert, sind insbesondere Kenntnisse über die Zusammenhangsstruktur von Variablen (Interkorrelationen) und ihr gemeinsamer Einfluss auf eine Zielvariable und außerdem Kenntnisse zum Thema Variablenauswahl/Variablenreduktion zwingend notwendig.
3. **Auswahl** des tatsächlichen Analyseverfahrens zur Erstellung des statistischen Modells. I.d.R. stehen dem Anwender zur Erstellung des statistischen Prognose-Modells verschiedene konkurrierende statistische Verfahren aus dem Bereich der Varianz-/ Regressions-/Diskriminanz-Analyse

zur Verfügung. Hier unterscheiden sich kategorielle, versus-lineare, versuslogistische Regressionen, Baumverfahren, diskriminanz-analytische Verfahren, Neuronale Netze u.s.w (vgl. hierzu Hoadley (1997), Höschel und Müller (1996)).

Im Vergleich zu den beiden Erfolgsfaktoren "Datengrundlage" und "Erfahrung/Know-how des Anwenders" spielt dann zusätzlich die Wahl des richtigen methodischen Verfahrens nur noch eine untergeordnete Rolle, d.h. hier können letztendlich nur Verbesserungen im marginalen Bereich erzielt werden, während durch die ersten beiden Punkte substantielle Erfolgsfaktoren gelegt werden. Auch einige bekannte empirische Studien zeigen, dass die Wahl des Modellierungs-Verfahrens selbst bei gleicher Datengrundlage und Anwendererfahrung nur noch ein geringfügiges Verbesserungspotential bringt. Dennoch gibt es hier zuverlässigere und weniger zuverlässige Verfahren.

5 Aktuelle Nutzung komplexer statistischer Verfahren in Business Anwendungen

Um ein Bild von der Anwendungspraxis komplexer statistischer Methoden in den Business Anwendungen zu erhalten, wurden in einer Blitzumfrage ausgewählte Experten (Leiter von Datenanalysebereichen großer Unternehmen wie Versandhäuser, Versicherungen, Verlage und auch Banken) befragt. Der Anwendungsfokus liegt eindeutig im Bereich Marketing-Fragestellung. Teilweise werden in den Unternehmen jedoch auch Marketing- und Risiko-Fragestellungen integriert durch einen Datenanalysebereich bearbeitet.

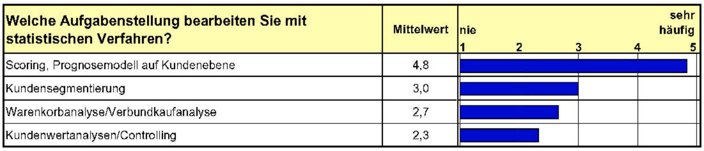
Alle Fragen wurden in einer telefonischen Befragung auf einer Skala von 1 bis 5 erhoben.

Häufigstes Anwendungsgebiet komplexer statistischer Verfahren ist die Erstellung von Prognosemodellen auf Kundenebene, zusätzlich werden jedoch auch Prognosemodelle für die Entwicklung von Produkten und auch anderen Einheiten wie z. B. Werbemitteln untersucht. Diese Anwendungen werden in nahezu allen Unternehmen mit hoher Intensität durchgeführt. Als weitere Anwendungsfelder kristallisierten sich folgende heraus:

- Kundensegmentierung/Kundentypologie,
- Kundenwertanalysen/Controllingaufgabenstellungen,
- Warenkorbanalysen/Verbundkaufanalysen.

5.1 Intensität der Verfahrensnutzung

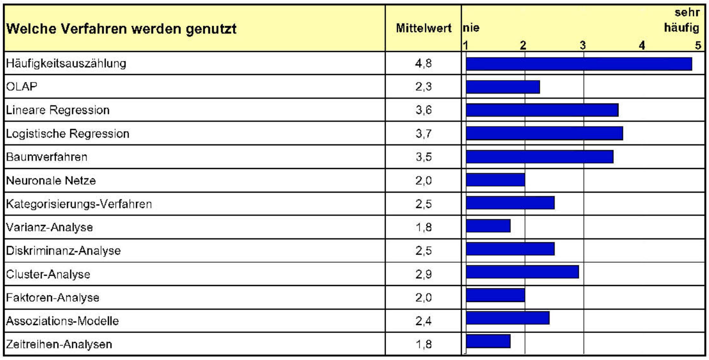
Hier ergab sich zunächst eine 100%-ige Übereinstimmung mit der Studie von 1984 in der Nutzung von einfachen Häufigkeitsauszählungen. Diese Verfahren sind nach wie vor die am meisten angewendeten Verfahren in der



Welche Aufgabenstellung bearbeiten Sie mit statistischen Verfahren?	Mittelwert
Scoring, Prognosemodell auf Kundenebene	4,8
Kundensegmentierung	3,0
Warenkorbanalyse/Verbundkaufanalyse	2,7
Kundenwertanalysen/Controlling	2,3

Fig. 2. Aktuelle Anwendungssicht

Praxis überhaupt. Erlauben sie doch, ein einfaches Grundverständnis für die Daten, Plausibiliätskontrollen und eine klare Sicht der Dinge auch für die internen Anwender. Weniger durchgesetzt haben sich bis jetzt jedoch sogenannte OLAP-Ansätze, die mit Hilfe einer datenbanknahen Spezial-Software eine komfortable mehrdimensionale Häufigkeitsdarstellung erlauben.



Welche Verfahren werden genutzt	Mittelwert
Häufigkeitsauszählung	4,8
OLAP	2,3
Lineare Regression	3,6
Logistische Regression	3,7
Baumverfahren	3,5
Neuronale Netze	2,0
Kategorisierungs-Verfahren	2,5
Varianz-Analyse	1,8
Diskriminanz-Analyse	2,5
Cluster-Analyse	2,9
Faktoren-Analyse	2,0
Assoziations-Modelle	2,4
Zeitreihen-Analysen	1,8

Fig. 3. Aktuelle Methoden-Nutzung

Entsprechend dem Anwendungsfokus stehen Verfahren zur Erstellung von Prognosemodellen bei den Anwendern besonders hoch im Kurs, dabei wird aber auf ein ganzes Verfahrensspektrum, wie lineare Regression, loglineare/logistische Regression und auch Baumverfahren zurückgegriffen.

Gerade Baumverfahren und logistische Regression waren in der Studie von 1984 noch unter den wenig genutzten Spezialverfahren angesiedelt. Inzwischen haben diese Verfahren jedoch in die Standardanwendungen bei der Erstellung von Scoring-Modellen Einzug gehalten. Dahingegen haben es die Neuronalen Netze als weiteres neueres Instrumentarium nicht geschafft hoch in den Kurs der Anwendungspraxis zu steigen. Hier gibt es jedoch einzelne Anwender, die auch diese Verfahren intensiver nutzen.

Unterstützt werden diese Verfahren durch sogenannte Kategorisierungsverfahren, die eine optimale Zerlegung des Wertebereiches einer kategoriellen oder stetigen Variable für prognostische Zwecke erlauben. Eine geringere Bedeutung haben die Diskriminanz-Analysen und insbesondere die Varianz-

Analyse, die ebenfalls zur Erstellung von Prognose-Modellen genutzt werden können.

Als weiteres Verfahren mit Bedeutung für die Business-Praxis hat sich die Cluster-Analyse gezeigt, diese wird insbesondere zur Erstellung von Kundensegmentierung und Kundentypologien genutzt. Es folgt die Faktoren-Analyse, die gelegentlich genutzt wird um den Merkmalsraum zu reduzieren und die Assoziations-Verfahren, um Warenkorb/Verbundkaufanalysen durchzuführen.

Absolut konform mit den Ergebnissen der Studie von 1984 erwies sich die geringe Nutzung von Zeitreihen-Analysemethoden.

Als allgemeiner Tenor aus der Befragung wurde deutlich, dass komplexere statistische Verfahren in der täglichen Analysepraxis der Großunternehmen eine bedeutende Rolle spielen. Die angewendeten Methoden ändern sich nur langsam, aber auch hier ist durchaus eine Erweiterung oder Verlagerung des Methodenspektrums zu erkennen. Die Anwendungsfelder verbreitern sich jedoch stetig: Die Anzahl und auch die Breite der untersuchten Fragestellungen haben in den letzten Jahren deutlich zugenommen.

Insgesamt erkennt man, dass sich auch neuere Verfahren, wie z. B. Entscheidungsbäume in der Praxis durchsetzen, wenn sie denn bestimmte Nutzungskriterien erfüllen.

5.2 Gründe für die Anwendung komplexer statistischer Verfahren

Fragt man die Anwender nach den Gründen, warum sich die Anwender bei ihrer Verfahrenswahl für das eine oder andere Verfahren entscheiden, so wird deutlich, dass sich die Anwender diese Entscheidung sehr wohl gut überlegen und viele Argumente bei dieser Entscheidung eine wichtige Rolle spielen. Es sind jedoch eindeutig 2 bis 3 Gründe von allerhöchster Priorität:

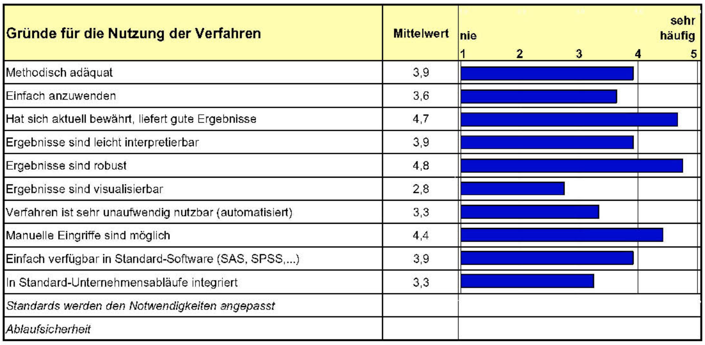
Zentrale Gründe

1. Das Verfahren soll robust sein. Das bedeutet, dass seine Ergebnisse nicht bis zum letzten Feintuning auf den Analysebestand ausgerichtet sind, sondern robust gegenüber Veränderungen sind, die zwischen der Entwicklung und der Anwendung des Modells in der Praxis immer auftreten können. Dieses Kriterium ist eindeutig für alle Praktiker sehr, sehr wichtig, aber spielt jedoch i.d.R. bei der Entwicklung und Prüfung von Datenanalyse-Methoden nur eine untergeordnete Rolle.
2. Ähnlich wichtig und inhaltlich zusammenhängend ist das 2. Argument, dass das Verfahren in der Vergangenheit und auch aktuell in dem Unternehmen immer gute Prognose/analytische Ergebnisse geliefert hat, d.h. dass es sich in der Anwendungspraxis bewährt hat. Damit ist das Verfahren für den Anwender einschätzbar und er kann auch die daraus resultierenden Ergebnisse besser bezüglich ihrer Prognoserisiken bzw. -sicherheiten einschätzen.

3. Außerdem waren sich die meisten Befragten darin einig, dass sie selbst als Anwender und Experten Steuerungsmöglichkeiten für die Anwendung der Verfahren (Parametrisierung, Variablenauswahl, ...) haben müssen. Der Anwender möchte also bis zum Ende die Ergebnisse selber kontrollieren können.

Weitere relevante Gründe

4. Weiterhin spielt die einfache Nutzbarkeit und Zugänglichkeit der Verfahren ebenso eine bedeutende, wenn auch nicht die wichtigste Rolle. Darunter fällt auch einfache Interpretierbarkeit, Integration in Standardsoftware/Statistikprogrammpakete und auch Integration in Standard-Unternehmensabläufe, stärkere Automatisierbarkeit des Verfahrens.
5. Außerdem ist es wichtig, dass das Verfahren methodisch adäquat ist.
6. Die Visualisierbarkeit spielt die geringste Rolle.

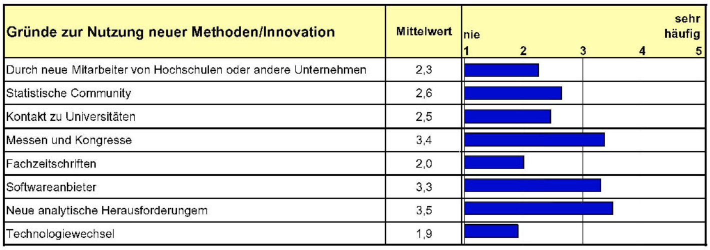


Gründe für die Nutzung der Verfahren	Mittelwert	nie sehr häufig 1 2 3 4 5
Methodisch adäquat	3,9	
Einfach anzuwenden	3,6	
Hat sich aktuell bewährt, liefert gute Ergebnisse	4,7	
Ergebnisse sind leicht interpretierbar	3,9	
Ergebnisse sind robust	4,8	
Ergebnisse sind visualisierbar	2,8	
Verfahren ist sehr unaufwendig nutzbar (automatisiert)	3,3	
Manuelle Eingriffe sind möglich	4,4	
Einfach verfügbar in Standard-Software (SAS, SPSS,...)	3,9	
In Standard-Unternehmensabläufe integriert	3,3	
Standards werden den Notwendigkeiten angepasst		
Ablaufsicherheit		

Fig. 4. Aktuelle Gründe für die Methoden-Auswahl

Insgesamt kann man sagen, dass sich die Anwender die Wahl des geeigneten statistischen Verfahrens nicht einfach machen, sondern sehr wohl überlegte, fundierte Entscheidungen treffen. Hierbei ist immer ein Kompromiss zwischen methodisch adäquat, ergebnisorientiert und Aufwand zu schließen.

Fragt man nach den Gründen wie innovative Verfahren in Unternehmen verbreitet werden, so gibt es sowohl internen Input, nämlich neue analytische Herausforderung, als auch externen Input, hierzu zählen insbesondere Messen und Kongresse und Softwareanbieter. Der Input von Universitäten, sei es durch neue Mitarbeiter oder sonstige Kontakte, spielt eher eine untergeordnete Rolle. Ganz gering wird der Input von Fachzeitschriften eingeschätzt. Auch interne Veränderungen, wie Technologiewechsel spielen nur eine untergeordnete Rolle.



Gründe zur Nutzung neuer Methoden/Innovation	Mittelwert
Durch neue Mitarbeiter von Hochschulen oder andere Unternehmen	2,3
Statistische Community	2,6
Kontakt zu Universitäten	2,5
Messen und Kongresse	3,4
Fachzeitschriften	2,0
Softwareanbieter	3,3
Neue analytische Herausforderungem	3,5
Technologiewechsel	1,9

Fig. 5. Aktuelle Innovations-Anlässe

6 Zusammenfassung und Ausblick

Die Anwendung komplexer statistischer Verfahren in der Business-Praxis ist in den letzten Jahren deutlich gestiegen, kann man doch durch datengestützte Analysen Entscheidungen deutlich verbessern.

Auch neuere analytische Verfahren finden Einzug in die Anwendungspraxis, wenn sie denn den strengen Regeln, der Praxis genügen. Sprich: Robuste, gut interpretierbare und plausible Ergebnisse hervorbringen und Aufwand und Wirkung in einem akzeptablen Verhältnis steht.

Die Universitäten als Zentrum für Methoden-Weiterentwicklungen wurden nicht sehr hoch in ihrer Innovationskraft für die Unternehmen eingeschätzt. Die Innovationskraft kommt eher von den Softwareanbietern. Es besteht jedoch der deutliche Wunsch der Business-Anwender an die Universitäten, dass in den Universitäten sowohl in der Forschung als auch insbesondere in der Lehre ein stärkerer Praxisbezug zu den Business-Anwendungen hergestellt wird.

Hier gilt es, nicht nur ein Methodenwissen zu vermitteln, sondern auch den ganzen Anwendungsprozess der erfolgreichen Datenanalyse - angefangen von der Datengenerierung bis hin zu Interpretation und Umsetzung in Entscheidungen - den zukünftigen Analysten zu vermitteln.

Alle Befragten waren einhellig der Meinung, dass hier durch ein stärkeres Zusammenrücken von Theorie und Praxis, Ausbildung und Anwendung für beide Seiten ein positiver Effekt entstehen würde und sich für die Zukunft deutliche Chancen ableiten lassen.

References

BERRY, M.J.A. and LINHOFF, G. (2000): *Mastering Data Mining.* Wiley Computer Publishing, New York, Weinheim.

GAUL, W., FÖRSTER, F., and SCHILLER, K. (1986a): Typologisierung deutscher Marktforschungsinstitute. *Marketing ZFP, 8(3), 165–172.*

GAUL, W., FÖRSTER, F., and SCHILLER, K. (1986b): Empirische Ergebnisse zur Verbreitung und Nutzung von Statistik-Software in der Marktforschung, Tagungsband zur 3. Konferenz über wiss. Anwendung von Statistik-Software, München.

GENTSCH, P. (2002): *Mit Data Mining dem "Homo Hybridicus" auf der Spur.* Data-Mining-Cup Chemnitz.

HOADLEY, B. (1997): Fair Isaac Case Study in Comparing Scoring Technologies, InterACT Barcelona.

HÖSCHEL, H.-P. and MÜLLER, R.-J. (1996): Pilotstudie Bonitätsprüfung Neckermann, Institut for International Research auf der Fachkonferenz RISK Wien.

MARTIN, W. (2002): *Data Mining - die nächste Phase.* Data-Mining-Cup Chemnitz.

Heuristic Bundling

Bernd Stauß and Volker Schlecht

Institut für Entscheidungstheorie und Unternehmensforschung,
Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

Abstract. Bundling can be used to tailor offers to the demand of consumers and helps to tackle the management of variety reduction. It is strongly based on consumer preferences which necessitates the use of elaborated data analysis techniques as, e.g., conjoint analysis.

In the following, we review a heuristic approach that goes back to a recent paper from Stauß and Gaul (2004a) and strikes for finding the most profitable bundles and respective prices. Since the algorithm starts out from an initial set of bundles, this paper focuses on the determination of such a promising initial bundle set.

1 Introduction

The well-known practice of selling two or more different products or services - so called components - together at a unique bundle price is referred to as bundling. However, despite this at first view seemingly clear definition no unique specification of bundling can be found in marketing literature. While Adams and Yellen (1976) stress the fact that the components have to be offered jointly, others like Guiltinan (1987) and Yadev and Monroe (1993) demand one special bundle price that has to be charged. And Mulhern and Leone (1991) even consider a positive cross price elasticity between two products as a kind of implicit bundling. In the very beginning of research concerning bundling, much work, s., e.g., Burstein (1960) and Stigler (1963), focused on bundling from an anti-trust law perspective, since there evolved monopolistic markets where firms attempted to expand their monopoly power into other more competitive areas by means of bundling. However, by the time the main focus changed to the determination of optimal bundling strategies (for a recent example s. Venkatesh and Kamakura (2003)). Possible motivations for bundling include the exploitation of demand complementary, the already mentioned competitive effects, and the implementation of an implicit price discrimination device by the use of bundling. As a major advantage in connection with diverging consumer tastes many authors discuss the possibility of smoothing out consumers' preferences as a nice side-effect inherent in bundling. And finally, since products often consist of optional characteristics, so called options, bundling also received growing attention in the customization literature (e.g. Swaminathan (2001)) due to its ability to tackle the problem of fast growing product variety by allowing for a more or less arbitrary mixing and matching of options (Matutes and Regibeau (1988)). This is fairly common practice, e.g., in the automobile or computer industry.

However, until recently, little work has been done concerning the development of usable decision models and appropriate algorithms for generating optimal bundles and prices, respectively. Main contributions in this area were given by the work of Hanson and Martin (1990) and Gaul and Stauß (2003).

An essential behavioristic construct frequently used in predicting the demand of potential bundles at a certain price is given by the incorporation of the reservation price concept into classical choice models. Reservation prices indicate the maximum amount of money someone is willing to pay for respective components or, equivalently, (parts of) bundles. Though its obviously of great relevance, just a few authors address the estimation of the unknown (individual) reservation prices, c.f. Kalish and Nelson (1991), Aust (1995) and more recently e.g. Jedidi et al. (2003) and Stauß and Gaul (2004b).

2 Optimizing bundle designs and prices

One of the most ingenious ways of creating promising bundles of components and determining the respective bundle prices is undoubtedly given by the use of quantitative decision support tools and corresponding optimization techniques such as, e.g., linear programming frequently in conjunction with integrality constraints with respect to some of the decision variables. Suppose, that the set of options is given by $\mathcal{J}$ and $\mathcal{B}_l \subseteq \mathcal{J}$ is the set of options contained in bundle l, $\mathcal{L}$ describes the set of all possible bundles, and $\mathcal{I}$ is the set of potential buyers (with $i = 1, \ldots, |\mathcal{I}|$, $j = 1, \ldots, |\mathcal{J}|$, and $l = 1, \ldots, |\mathcal{L}| = 2^{|\mathcal{J}|} - 1$). Obviously, an alternative interpretation of i may be that of a group of customers - so called segments - where each segment i is assigned a certain size N_i. Since each component by itself constitutes a bundle, we assume that the first $|\mathcal{J}|$ bundles in $\mathcal{L}$ correspond to the respective products. The cost c_l of bundle l is the sum of costs from the component contained in the bundle, i.e. $c_l = \sum_{j \in \mathcal{B}_l} c_j$. Every consumer i has known reservation prices r_{il} for every possible bundle l which can be offered to consumer i at a price p_{il}. Generally, it is supposed that given bundle prices, reservation prices, and the composition of the bundles, every consumer strikes for maximizing her/his surplus, i.e., the difference between reservation price and actual price of the bundle l: $r_{il} - p_{il}$. Finally, we assume that consumers cannot get any surplus if they do not purchase anything and, however, somewhat simplifying that the benefit obtained from multiple components is zero as well as there exist no unwanted components. Furthermore, in order to avoid arbitrage a resale of components remains excluded. Within a customization framework this seems to be a quite realistic assumption, since options are frequently tied with a major component, e.g., a base version of a car or house.

Consumers' choice behavior is modeled by means of binary variables $\theta_{il} \in \{0, 1\}$ with

$$\theta_{il} = \begin{cases} 1, & \text{if segment } i \text{ selects bundle } l \\ 0, & \text{otherwise.} \end{cases}$$

Although segment specific bundle prices were allowed this is done for technical reasons only (to avoid non-linearities), since an explicit price discrimination remains categorically excluded. This is assured by activating restrictions such as $p_{il} = p_l, i \in \mathcal{I}, l \in \mathcal{L}$ whenever $\theta_{il} = 1$. Obviously no restrictions concerning self-bundling have been formulated, however, one easily shows that self-bundling is implicitly excluded if the following price subadditivity condition

$$p_l \leq \sum_{k \in \mathcal{K}} p_k \quad l \in \mathcal{L}, \mathcal{K} \in \{\tilde{\mathcal{K}} | \cup_{\tilde{k} \in \tilde{\mathcal{K}}} \mathcal{B}_{\tilde{k}} = \mathcal{B}_l\}$$

holds. Furthermore since costs were additive (or in some cases even subadditive) it doesn't pay for the firm to induce self-bundling incentives, i.e., under the above assumptions there exists a profit-maximizing price schedule that maximizes firm's profit. In line with Hanson and Martin (1990) the respective optimization approach (MHM) may be formulated as follows:

MHM:

$$\max \sum_{i \in \mathcal{I}} \sum_{l \in \mathcal{L}} N_i (p_{il} - c_l \theta_{il}) \tag{1}$$

$$\sum_{\tilde{l} \in \mathcal{L}} \left(r_{i\tilde{l}} \theta_{i\tilde{l}} - p_{i\tilde{l}} \right) \geq \max\{0, r_{il} - p_l\} \quad i \in \mathcal{I}, l \in \mathcal{L} \tag{2}$$

$$p_l \leq \sum_{k \in \mathcal{K}} p_k \quad l \in \mathcal{L}, \mathcal{K} \in \{\tilde{\mathcal{K}} | \cup_{\tilde{k} \in \tilde{\mathcal{K}}} \mathcal{B}_{\tilde{k}} = \mathcal{B}_l\} \tag{3}$$

$$p_{il} \geq p_l - M(1 - \theta_{il}) \quad i \in \mathcal{I}, l \in \mathcal{L} \tag{4}$$

$$p_{il} \leq p_l \quad i \in \mathcal{I}, l \in \mathcal{L} \tag{5}$$

$$\sum_{l \in \mathcal{L}} \theta_{il} \leq 1 \quad i \in \mathcal{I} \tag{6}$$

$$p_l, p_{il} \geq 0 \quad i \in \mathcal{I}, l \in \mathcal{L} \tag{7}$$

$$\theta_{il} \in \{0, 1\} \quad i \in \mathcal{I}, l \in \mathcal{L} \tag{8}$$

where $M > 0$ is a sufficiently large constant. The constraints allow for the following interpretation: The inequalities in (3) enforce the already mentioned subadditive price schedule and so consumer choice behavior resembles a first choice situation as given by constraint (2), i.e. each consumer will buy that bundle which provides the maximum surplus. Clearly, each consumer that actually buys faces identical prices for identical bundles (because if a bundle l is selected ($\theta_{il} = 1$), (4) and (5) enforce $p_{il} = p_l$, otherwise (for $\theta_{il} = 0$) conditions (2) along with the nonnegativity of p_{il} ensure $p_{il} = 0$). Constraints (6), (7), and (8) are self-explaining. As a solution of MHM we get an optimal set of bundles $\mathcal{L}_{opt} = \{l \in \mathcal{L} | \sum_{i \in \mathcal{I}} \theta_{il} > 0\}$ that should be offered together with an optimal price schedule.

A crucial part of MHM is formed by constraint (3) that describes and guarantees for the special structure of the solution. Although, Gaul and Stauß

(2003) show that it suffices to examine price subadditivity in terms of restrictions for pairs of bundles only, instead of those given by (3) the modeling effort generally becomes immense. This is mainly due to the fact that in advance all possible bundles would have to be listed explicitly, since otherwise the appealing structure of the solution, i.e., the reduction of consumer choice behavior to an ordinary first choice situation, could not be ensured anymore. Finally, if one lacks the knowledge of a functional interrelation between r_{il} and r_{ij}, all bundles' reservation prices have to be elicited explicitly from the respective consumers which results in an exponentially growing valuation effort. This fundamental drawback drives the development of an alternative formulation of the bundle problem that is proposed by Gaul and Stauß (2003). The underlying assumptions of the new model are similar to those that were given in the beginning of this section with two slight but essential modifications. First, in order to reduce the burden of estimating parameters, additivity concerning bundles' reservation prices is supposed. Second, in the reformulated approach the price subadditivity constraint can be relaxed while additional restrictions still ensure the proposed consumer behavior. Due to the omission of the price subadditivity constraint, a crucial obstacle in developing heuristic approaches could in fact be eliminated. An application of the reformulated bundling model within the framework of a heuristic approach can be found in Stauß and Gaul (2004a). The so-called design heuristics starts from an initial bundle set $\mathcal{L}_0$ and successively adds bundles that best supplement the existing set of bundles where in each iteration prices were optimized simultaneously. In this way, a sequence of monotonically increasing lower bounds for the firm's profit can be generated. The algorithm stops if no further improvement of profit can be obtained by adding an additional bundle. If there is no prior information concerning a promising initial bundle set, usually, the set $\mathcal{L}_0$ will be empty so that probably a large number of iterations will be necessary in order to generate the set of bundles that finally should be offered. However, if reservation prices are available on a component rather than a bundle level, data analysis and in particular cluster analysis might become a valuable tool in order to identify a set of potentially promising bundles that represent the initial bundle set. This is what we will discuss in the following sections where we shortly motivate and explain the applied two-mode cluster algorithm first.

3 Creating promising bundle candidates

If bundles are derived in a way that takes the interaction of consumers and options (and thereby the preferences) into account, the resulting sets of options could be promising candidates for the initial bundle set, since ideally there is substantial support for each bundle coming from at least one segment. So the grouping might be done with two modes rather than just one, which suggests the use of two-mode clustering for which a number of different algorithms can be found in literature (e.g., DeSarbo (1982), DeSarbo et al. (1988), Gaul and Schader (1996)). If one utilizes reservation prices as input data, potential interrelations between customers (first mode elements) and options (second mode elements) may be detected by simultaneously grouping all options which are similarly valued by some customers and clustering all customers with a similar preference structure.

As already indicated, we take the options $j \in \mathcal{J}$ to be the second mode elements where the second mode clusters are to be interpreted as the respective bundles $l \in \mathcal{L}$. Furthermore, consumers constitute the first mode elements $i \in \mathcal{I}$ with the clusters corresponding to segments. Additionally, we introduce $\mathcal{N}$ as the set of segments and refer to $n \in \mathcal{N}$ as the index of these first mode clusters. Here, first mode elements and second mode elements are clustered simultaneously. Every first mode cluster is in some characteristic way associated with the second mode clusters and vice versa. Hence, every first mode cluster can be interpreted by looking at the second mode clusters it is connected with and vice versa. The following notation is used: $S = (s_{ij})$ $(\hat{S} = (\hat{s}_{ij}))$ is the observed (estimated) two-mode data matrix that contains individual reservation prices for the options. $O = (o_{in})$ $(Q = (q_{jl}))$ is the matrix which describes the cluster-membership of the first (second) mode elements with

$$o_{in}\ (q_{jl}) = \begin{cases} 1, \text{ if } i\ (j) \text{ belongs to first (second)} \\ \quad \text{mode cluster } n\ (l), \\ 0, \text{ otherwise,} \end{cases}$$

$W = (w_{nl})$ is a matrix of weights.

Two-mode clustering algorithms in general try to find the best-fitting estimator $\hat{S}$ for the given two-mode data matrix S (Gaul and Schader (1996), Baier et al. (1997)). A simple way for calculating this best-fitting estimator $\hat{S}$ is $\hat{S}=OWQ'$, where the matrices O, W, and Q have to be alternatingly determined by minimizing the objective function

$$Z = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} (s_{ij} - \hat{s}_{ij})^2,$$

where

$$\hat{s}_{ij} = \sum_{n \in \mathcal{N}} \sum_{l \in \mathcal{L}} o_{in} w_{il} q_{jl}.$$

Both overlapping and non-overlapping versions of this kind of algorithm for two-mode clustering exist. In fuzzy two-mode cluster analysis instead of binary membership indicator variables degrees of membership, i.e., $o_{in} \in [0, 1]$, $i \in \mathcal{I}$, $n \in \mathcal{N}$ and $\sum_{n \in \mathcal{N}} o_{in} = 1, \forall\ i \in \mathcal{I}$ are introduced (the restrictions for (q_{jl}) are of equivalent form). Finally, w_{nl} can be interpreted as the degree of connection between the corresponding first mode cluster n and the second mode cluster l. From the fuzzy classification several different non-fuzzy overlapping classifications can be derived by setting $q_{jl}(o_{in})$ equal to 1 if some threshold $q_{min}(o_{min})$ is exceeded and 0 otherwise. The larger the value $q_{min}(o_{min})$ is chosen, the more the resulting overlapping classification resembles a non-overlapping classification. So with $q_{min}(o_{min})$ we have additional control parameters that allow for adjusting the classification.

Starting from a non-fuzzy non-overlapping classification (e.g., Baier et al. (1997) the Delta-Method which is described in detail by Schlecht and Gaul (2004) can be used to find a fuzzy two-mode classification. In the case of fuzzy two-mode cluster analysis one alternatingly determines O, W, and Q by minimizing the objective function Z where Newton's algorithm is used for the determination of $W=(w_{nl})$. Since equations are symmetric in (o_{in}) and (q_{jl}) the idea for optimizing O and Q is the same. It is also given in detail in Schlecht and Gaul (2004). Here, we just briefly outline the general idea, which is to start from a non-fuzzy non-overlapping two-mode classification and search for fuzzy cluster-membership values which improve the objective function Z. For every first mode element i we select the highest value of $o_{in}, n \in \mathcal{N}$ and subtract $\Delta(0 < \Delta < 1)$ from it. Then we try to find the value n^* which produces the best value for the objective function Z if we add Δ to o_{in^*}. Before we move on to the $(i{+}1)$-th element we iterate this procedure until the objective function cannot be improved any further.

4 Empirical validation

4.1 Study design and application

In order to test the proposed design heuristics on real data, in particular individual reservation prices on a component level are needed. However, the direct elicitation of reservations prices, i.e., to ask consumers "how much would you pay for *this* item?", is generally known to result in downward biased estimates for the actual willingness to pay of an item (e.g., Stauß and Gaul (2004b)). But if direct elicitation is used in a conjoint analysis framework, estimates of high internal validity for the part worths usually are obtained, although the predictive validity of the part worth estimates is quite low (Kalish and Nelson (1991)). The reason for this may be attributed to the difficulty of the underlying valuation tasks that may cause some inconsistency in the data. Thus, a modification of the elicitation technique by using paired comparisons, i.e. asking for relative judgements on a monetary scale, appears to offer a potentially successful way to improve predictive validity. Despite the rather discouraging results from a study reported in Aust (1995), pp. 182, there is indeed empirical evidence that a more elaborated construction of orthogonal designs for paired comparisons on a monetary scale - so called difference designs - can actually yield more valid estimates and facilitates respondents' evaluation tasks while at the same time the potential threat of exhaustion decreases (c.f. Stauß and Gaul (2004b)). For an empirical validation, data were collected with respect to a seat system offered by a German car manufacturer. The seat system comprises several (additional) options such as a seat heating device, a memory function etc. Using previous orders seven seat options could be identified which seemed to be quite important for customers as well as the firm itself. Departing from a complete 2^7- design a fractional factorial 2^{7-4}-design was constructed which served as the initial point for the construction of the difference designs. So, after all, an experimental design was obtained that required monetary valuations based on eight paired comparisons (c.f. Stauß and Gaul (2004b)). Data collection was partly done in the firm's customer center where the respective cars are surrendered to their new owners as well as on the fringes of a firm's celebration event.

4.2 Results and discussion

Individual reservation prices were estimated from the above mentioned paired comparison data. They subsequently served as input for the two-mode cluster algorithm in order to identify a set of bundle candidates. We derived classifications for different combinations of $|\mathcal{N}|$ and $|\mathcal{L}|$. Goodness of fit for the (non-)fuzzy models is reported in table 1 by means of the variance accounted for (VAF) measure where the number in brackets correspond to the non-fuzzy solution.

	Variance accounted for (VAF)			
$\lvert\mathcal{N}\rvert$	number of option clusters $\lvert\mathcal{L}\rvert$			
	4	5	6	7
3	(0.468) 0.695	(0.419) 0.693	(0.454) 0.695	(0.484) 0.694
4	(0.455) 0.781	(0.449) 0.781	(0.457) 0.789	(0.476) 0.796
5	(0.401) 0.806	(0.469) 0.868	(0.571) 0.836	(0.594) 0.874
6	(0.594) 0.852	(0.414) 0.891	(0.630) 0.911	(0.640) 0.923
7	(0.619) 0.836	(0.592) 0.918	(0.649) 0.934	(0.707) 0.940

Table 1. Goodness of fit for (non-)fuzzy two-mode clustering

For the classification derived by non-fuzzy two-mode clustering the situation is as follows. The VAF increases very strongly for $|\mathcal{N}| = 5$ if one changes from $|\mathcal{L}| = 5$ to $|\mathcal{L}| = 6$. Another quite considerable increase in the variance accounted for can be observed if one moves to $|\mathcal{N}| = |\mathcal{L}| = 6$. However, since we are just considering 7 options, 5 of the 6 bundles will contain only one option which does not seem to be an interesting initial bundle set. On the other hand one might take a look at the VAF for $|\mathcal{N}| = 6$ and $|\mathcal{L}| = 4$, since for the non-fuzzy case the VAF is still considerably improved. With regard to fuzzy two-mode clustering the greatest improvements for $|\mathcal{N}| = 5$ are made by changing from $|\mathcal{L}| = 3$ to $|\mathcal{L}| = 4$ and from $|\mathcal{L}| = 4$ to $|\mathcal{L}| = 5$. On the other hand, for $|\mathcal{L}| = 5$ the VAF increases strongly from $|\mathcal{N}| = 3$ to $|\mathcal{N}| = 4$ and from $|\mathcal{N}| = 4$ to $|\mathcal{N}| = 5$. A further increment of the number of segments yield just slight improvements concerning the VAF.

As already mentioned, the generated clusters of options serve as the potential bundles that constitute the respective initial bundle set $\mathcal{L}_0$ for the design heuristics that was implemented within a Java routine that embedded Ilog's Cplex Solver via Ilog's Concert interface. heuristics successively adds further bundles that lead to a maximum increase of profit. The algorithm stopps whenever no further increase in profit can be attained. Since the concurrent generation of additional bundles and optimization of prices result in a rather huge mixed integer program, an exact solution within each iteration fails due to the extraordinary computational effort and so the branch&cut procedure is aborted whenever a maximum number of 10.000 nodes has been examined. The results are reported in table 2 where the third column on the r.h.s. indicates the best integer solution found, i.e., a lower bound for the objective value. In the last but one column the number of additional iterations of the design heuristics that has to be performed starting from the initial bundle set is recorded and finally computational effort is given in CPU seconds. With respect to the cluster algorithms that were used to generate the initial bundle set we distinguish between non fuzzy (nf) vs. fuzzy (f) two-mode clustering. In the non-fuzzy situation we focused on the case $|\mathcal{N}| = |\mathcal{L}| = 5$ referred to as (nf-5-5) as well as $|\mathcal{N}| = 6$, $|\mathcal{L}| = 4$ indicated by (nf-6-4). Furthermore, in the case of fuzzy two-mode clustering we looked at $|\mathcal{N}| = |\mathcal{L}| = 5$ in conjunction with different initial solutions for the cluster algorithm (A vs. B) as well as two values for $q_{min} = 0.15/0.30$ were tested and described by (f-A/B-q_{min}).

Comparison of initial bundle sets								
method	initial bundle sets					bound obj.	no. iter.	CPU sec.
	$\mathcal{B}_1$	$\mathcal{B}_2$	$\mathcal{B}_3$	$\mathcal{B}_4$	$\mathcal{B}_5$			
(nf-5-5)	$\{1,6\}$	$\{7\}$	$\{3\}$	$\{2,4\}$	$\{5\}$	99.81%	8	4760
(nf-6-4)	$\{1,3,5\}$	$\{2\}$	$\{6,7\}$	$\{4\}$	-	91.43%	6	4175
(f-A-0.15)	$\{1,4,6\}$	$\{6,7\}$	$\{3,4,5,7\}$	$\{2,4\}$	$\{3,5,6\}$	95.00%	8	5692
(f-A-0.30)	$\{1,6\}$	$\{6,7\}$	$\{3,5\}$	$\{2,4\}$	$\{5,6\}$	83.84%	5	3688
(f-B-0.15)	$\{2,4,5\}$	$\{3,4,5,6,7\}$	$\{1,4,6\}$	$\{1,2,4,7\}$	$\{6,7\}$	98.10%	6	5353
(f-B-0.30)	$\{2\}$	$\{3,5,6\}$	$\{1\}$	$\{6,7\}$		81.39%	5	3872
(d)	-	-	-	-	-	100.0%	12	7197

Table 2. Performance of initial bundle sets

Generally, the more q_{min} is increased the more the average number of options in the bundles is reduced (situation (f-A-0.30) and (f-B-0.30)) and the less iterations have to be performed. On the other hand the quality of the solution deteriorates. However, in both fuzzy situations the choice of $q_{min} = 0.15$ ((f-A-0.15) and (f-B-0.15)) yield a quite good initial bundle set, reaching almost the performance of the default situation (d). Since in the latter case the set of bundles is built up from scratch, the successively added bundles obviously match somewhat better and thus yield a slightly higher lower bound for the objective value, however, at the cost of a substantial increase of computational effort. As can be seen from table 2, the number of CPU seconds required for (f-A-0.15) and (f-B-0.15) is again higher than in the respective (f-A-0.30) and (f-B-0.30) situations. In the non-fuzzy situation the goodness of the lower bound for the objective value is quite reasonable although it becomes somewhat worse if the number of initial bundles is reduced. A rather interesting point is revealed by comparing the number of iterations and the number of initial bundle candidates. The less initial bundles are selected the less iterations have to be performed which is somewhat astonishing, since one would have supposed that more additional bundles would have to be generated in order to obtain a final set of bundles that should be offered.

5 Conclusions

In this work we reviewed some present development in formulating heuristics for designing and pricing bundles of options which is primarily based on earlier work from Gaul and Stauß (2003) and Stauß and Gaul (2004a). Though, the proposed design heuristics may start with an empty set of bundle candidates an a-priori determined set of initial candidates might also be used. For an empirical validation reservation prices have been elicited representing the monetary valuation of different options from a car's seat system. Although the results may not be generalized they indicate that there is a considerable potential in increasing computational performance of the proposed design heuristics by means of a well-chosen initial bundle set. The selected two-mode clustering thereby exploits potential relationships between options and consumers respectively and seems to constitute a valuable approach. However, as can be taken from table 2 emphasis has to be set on the term well-chosen, since for the fuzzy situation an increase in q_{min} resulted in a considerable decrease in profit. Moreover, there is no obvious interrelationship between the objective concerning two-mode clustering and firm's profit, so further research should focus in detail on the interplay between two-mode clustering and the design heuristics.

References

ADAMS, W.J. and YELLEN, J.L. (1976): Commodity Bundling and the Burden of Monopoly. *Quarterly Journal of Economics, 90 475–498.*

AUST, E. (1995): *Simultane Conjoint Analyse, Benefitsegmentierung, Produktlinien- und Preisgestaltung.* Dissertation Universität Karlsruhe.

BAIER, D., GAUL, W., and SCHADER, M. (1997): Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring. In: R. Klar and O. Opitz (Eds.): *Classification and Knowledge Organziation, Studies in Classification, Data Analysis, and Knowledge Organization.* Springer, Berlin, 557–566.

BURSTEIN, M.L. (1960): The Economics of Tie-In Sales. *Review of Economics and Statistics, 41, 68–73.*

DESARBO, W.S. (1982): GENNCLUS: New Models for General Nonhierarchical Clustering Analysis. *Psychometrika, 47, 446-449.*

DESARBO, W. S., DESOETE, G., CARROLL, J. D., and RAMASWAMY, V. (1988): A New Stochastic Ultrametric Tree Methodology for Assessing Competitive Market Structure. *Applied Stochastic Models and Data Analysis, 4, 185–204.*

GAUL, W. and SCHADER, W. (1996): A New Algorithm for Two-Mode Clustering. In: H.H. Bock and W. Polasek (Eds.): *Data Analysis and Information Systems, Studies in Classification, Data Analysis, and Knowledge Organization.* Springer, Berlin, 15–23.

GAUL, W. and STAUSS, B. (2003): Designing and Pricing new Bundles. Diskussionspapier No. 255, Institut für Entscheidungstheorie und Unternehmensforschung Universität Karlsruhe.

GUILTINAN, J.P. (1987): The Price Bundling of Services: A Normative Framework. *Journal of Marketing, 51, 74–85.*

HANSON, W. and MARTIN, R.K. (1990): Optimal Bundling. *Management Science, 36, 155–174.*

JEDIDI, K., SHARAN, J., and PUNEET, M. (2003): Measuring Heterogeneous Reservation Prices for Product Bundles. *Marketing Science, 22, 107–130.*

KALISH, K. and NELSON, P. (1991): A Comparison of Ranking, Rating and Reservation Price Measure in Conjoint Analysis. *Marketing Letters, 2, 327–335.*

MATUTES, C. and REGIBEAU, P. (1988): "Mix and Match": Product Compatibility without Network Externalities. *RAND Journal of Economics, 19, 221–234.*

MULHERN, F.J. and LEONE, R.P. (1991): Implicit Price Bundling of Retail Products: A Mulitproduct Approach to Maximizing Store Profitability. *Journal of Marketing, 55, 63–76.*

SCHLECHT, V. and GAUL, W. (2004): Fuzzy Two-Mode Clustering vs. Collaborative Filtering. *Proceedings of the GfKl 2004*, Forthcoming.

STAUSS, B. and GAUL, W. (2004a): Product Bundling as a Marketing Application. In: D. Ahr, R. Fahrion, M. Oswald, and G. Reinelt (Eds.): *Operations Research Proceedings 2003.* Springer, Berlin, 221–228.

STAUSS, B. and GAUL, W. (2004b): Estimating Reservation Prices for Product Bundles Based on Paired Comparison Data. *Proceedings of the GfKl 2004*, Forthcoming.

STIGLER, G.J. (1963): United States v. Loew's Inc.: A Note on Block-Booking. *The Supreme Court Review, 152, 152–157.*

SWAMINATHAN, J.M. (2001): Enabling Customization Using Standard Operations. *California Management Review, 43, 125–135.*

VENKATESH, R. and KAMAKURA, W. (2003): Optimal Bundling and Pricing under a Monopoly: Contrasting Complements and Substitutes from Independently Valued Products. *Journal of Business, 76, 211–231.*

YADEV, M.S. and MONROE, K.B. (1993): How Buyers Perceive Savings in a Bundle Price: An Examination of a Bundle's Transaction Value. *Journal of Marketing Research, 30, 350–358.*

The Option of No-Purchase in the Empirical Description of Brand Choice Behaviour

Udo Wagner and Heribert Reisinger

Department of Business Administration, University of Vienna,
Brünner Straße 72, A-1210 Vienna, Austria

Abstract. In this paper an analysis is conducted in order to identify empirical regularities of non-buying behaviour and it is found that switching from a no-purchase to a certain brand is proportional to the market share of this brand. Switching from a brand to the no-purchase option is observed to be different. A model is proposed to describe switching behaviour between brands including a no-purchase state from market shares only.

1 Introduction

Antecedents and realizations of purchase decisions are of central interest for research in the area of consumer behaviour. Seen from a simplified point of view, buying decisions reduce to the basic question whether to buy a certain brand or not to buy at all. Interestingly, the latter aspect did not raise an equal share of research activities as the first one, although it might be the more frequent outcome of consumers' purchase decision processes. The reason for this is probably due to the fact that no-purchase decisions are difficult to identify, especially when using observational research designs. However, by concentrating on mere buying the researcher bears the risk to neglect an important part of the picture or to abandon relevant information, e.g. why a certain product has not been purchased. The article at hand aims at contributing to this field. In particular, we focus on frequently bought, low-involvement consumer goods for which some kind of panel data (e.g. household panel or retail scanning data) are available and analyse the no-purchase option within the framework of brand choice behaviour at the aggregate level.

In the literature, extensive discussions about the basic procedures to advance the knowledge of marketing can be found (e.g. Leeflang et al. (2000), p. 28 ff.). The two lines of reasoning may be summarised by the ET versus the TE approach, i.e. whether one should first look for empirical regularities which extend over a broad variety of different market situations and subsequently build a theory (or a model) which is capable to describe these observations; on the other hand, the TE approach starts with building a theory and subsequently analyses empirical cases in order to check whether the evidence is consistent with the theory or falsifies it. As mentioned above already, little is known about no-purchase outcomes. Therefore, we decide in favour of the ET approach and will search for regularities with respect to no-purchase decisions in the context of brand choice behaviour.

The remainder of this paper is organised as follows. Recently, some authors addressed the issue at stake by extending Logit models appropriately; section 2 provides an overview of the most important findings established by the academic literature in this field. In section 3 no-purchase decisions are analysed by means of household panel data, and some regularities of no-purchase behaviour are identified. Section 4 introduces a model which builds upon these regularities. Finally, section 5 concludes the article by summarising the findings and by providing issues for further research.

2 Incorporating the no-purchase option in brand choice models

The modelling of brand choice behaviour has a long tradition in marketing science. Applying the well known Multinomial Logit (MNL) Model to the analysis of brand choice is largely influenced by the seminal paper of Guadagni and Little (1983). Many extensions and improvements of the model have been made, however, the inclusion of a no-purchase option has not received a lot of interest until recently. If scanner panel data are available, the following possibilities to accommodate the no-purchase decision are proposed in the literature (see Chib et al. (2004)).

The most obvious way to account for the no-purchase option is to include an additional alternative in the MNL Model. The standard specification that explains the brand choice in a specific product category is extended for the household's decision of not purchasing any brand in the category under study. This additional alternative is called 'outside option' (Chintagunta (2002)) or 'outside good' (Nevo (2001)). An alternative is chosen if the accompanying utility for the household exceeds the utilities of the remaining alternatives. This approach suffers from the limitation that it ignores the effects of covariates (e.g. price) on the household's decision to purchase in the product category.

The Nested Logit Model (Guadagni and Little (1998)) is another way to accommodate the no-purchase option. It has a hierarchical structure and consists of two Logit Models that are nested within each other. The first level describes the household's decision to buy or not to buy in a certain product category. The second level models brand choice, conditional on a buying occasion to take place. Therefore, a household's brand choice probability is the product of two probabilities: (1) the household's category purchase probability, and (2) the household's conditional brand choice probability. An approach similar to the Nested Logit Model is the Generalized Extreme Value Model that is derived from direct utility maximization subject to a budget constraint (Chiang (1991), Arora et al. (1998)).

A quite recent approach is the framework proposed by Chib et al. (2004). In this model, the no-purchase outcome in a product category is specified to depend on all information available on the brands in that category. Again, the approach is related to the Nested Logit Model. The key difference lies in the fact that in the Nested Logit Model all covariates must have identical effects on the no-purchase decision (because of the introduction of the 'inclusive value' which serves as a link between the two parts of the model). In the new approach the data determine the weights of the covariates and, therefore, it is possible to study which marketing variables have the highest impact on the no-purchase option.

All of the proposed methods are based to a greater or lesser extent on the Multinomial Logit formulation. The no-purchase option is accommodated for and the resulting models lead to improved fit measures. However, the decision for a no-purchase outcome is not further addressed. Contrary to the modelling of brand choice behaviour in the MNL-context we examine in this paper brand choice and repeat buying in line with the ET paradigm and the work of Ehrenberg (1972, with a second edition published 1988). Allowing for a no-purchase option and employing GfK household panel data we first analyse whether empirical generalizations of no-purchase behaviour can be identified.

3 Empirical generalizations of no-purchase behaviour

Consumer behaviour on the Austrian market for dishwashing liquids in the years 1999-2001 is studied. This product category has the advantage that usually only one package is bought at a buying occasion and, hence, the household's quantity decision is negligible. The data set at hand comprises a period of 33 months. 2416 purchases were made by 434 households. In 93.6 % of the purchases only one package was bought. Hence, subsequent analyses are based on purchase occasions. The mean purchase rate is 5.57, i.e. each household bought about twice a year.

The market is characterised by a dominating brand A with a market share of approximately 48 %. One competitor B has a market share of approximately 18 % and the remaining ones do not reach the 10 % level. We consider the five most important brands on the market, denominated as A, B, C, D, and E in descending order according to their market shares, but neglect sales of all other brands which are of minor importance only (i.e. no 'all-other' brand category is formed; thereby 227 purchases and 29 households are excluded from further analysis).

In line with Chintagunta (2002) a no-purchase state is considered as an outside option and studied at the same level as brand choice. Therefore, the no-purchase option represents an additional decision alternative for the household. From a consumer behaviour point of view it is intuitive to think of a member of a household visiting a (grocery) shop more frequently than purchasing a dishwashing detergent. Each shopping trip offers a chance to buy within this product category. The purchase probability will increase if e.g. this household's stock of dishwashing detergents is exhausted, dishwashing detergents are listed on the patron's handbill, or a promotional activity in the store catches her/his interest. Nevertheless her/his decision might still be not to purchase a dishwashing liquid.

By using interviewing techniques or other research designs based on communication one could try to establish relationships between buying cycles, consumption patterns and shopping trips (see e.g. Krishnan and Seetharaman (2000)) or to establish some association between purchasing in this particular product category and other categories (i.e. shopping basket analysis). As outlined above, we want to address this issue by simply analysing household panel data and argue that by recording purchases only, information on no-purchases have been lost. Very much in the spirit of analysing the switching between brands by means of consumer panel data we look at households' buying histories and investigate the potential occurrence of intermediate non-buying decisions (e.g. an observed sequence of buying brands A and B subsequently might originate from a decision sequence A, no-purchase, B). In particular, we assume that each household consciously makes exactly one purchase decision (eventually not to buy) within an observation period of fixed length. Individual data of purchase histories are aggregated to result in brand switching matrices. By varying the length of the observation period we hope to get some idea about the stability of the results. In order to carry out this kind of analysis it is necessary to determine the chosen alternative for each household and each observation period. The following rules are applied to assign each household to one of the states (buying brands) A, B, C, D, E or no-purchase:

a) If just one purchase has been recorded within the observation period the household is assigned to the state corresponding to the chosen brand.
b) If no purchase has been recorded within the observation period the household is assigned to the no-purchase state.
c) If several purchases of the same brand have been recorded within the observation period the household is assigned to the state corresponding to the chosen brand.
d) If purchases of different brands have been recorded within the observation period the household is not assigned to any state and excluded from the analysis for this particular period (as well as from calculating switching matrices including this period).

Finally, the length of the observation periods has to be determined. Obviously, a period of 6 months could be specified according to the mean purchase rate (i.e. each household purchases on average once per period in the product category). Nevertheless, other lengths are conceivable. We take into account three time frames and study observation periods of 6 months, 4 months, and

Case	Length of observation period	Number of observation periods	Number of switching matrices available	Percentage of households not considered[1]	Percentage of of purchases not considered	Percentage of non-buyers
1	6 months	5[2]	4	13 %	54 %	48 %
2	4 months	8[2]	7	4 %	36 %	56 %
3	3 months	11	10	1 %	27 %	64 %
Observed time span				33 months		
Number of households with at least one purchase				405		
Number of purchases				2189		

[1] because of purchasing different brands within one observation period, see assignment rule d)
[2] right censoring occurs

Table 1. Consequences of different lengths of observation periods on data base

3 months. With 33 months available in our data set we therefore have five consecutive periods of 6 months at our disposal and are able to determine four different switching matrices (see Table 1). The remaining three months are not used in this case. We consider all households with well-defined assignments and determine the corresponding switching matrices. In order to smooth out random variations the averages over these four matrices are calculated. Table 2 shows the result in terms of conditional probabilities for switching from one decision alternative to another (including loyal behaviour in the main diagonal). Similarly, we proceed with observation periods of 4 and 3 months lengths and obtain the corresponding measures in Tables 3 and 4.

Examining Table 1, two facts become apparent: (1) shorter observation periods result in higher percentages of non-buyers (column 7) which is just a consequence of the assignment rules a) - d); (2) shorter observation periods result in lower percentages of households excluded from the analysis (column 5). Once again, this follows from the employed assignment rules, the striking fact, however, are the high percentages of purchases excluded from the analysis (column 6). These purchases originate from heavy buyers or at least from consumers with a purchase frequency above the average. This segment is especially interesting from a marketing point of view. Fortunately, excluding these households does not seem to have a substantial influence on the switching matrices calculated (see below). These phenomena emphasize the inherent difficulty the researcher faces when trying to define a procedure applicable to a broad range of consumers to determine switching within a time frame of a given length.

	A	B	C	D	E	No-purchase	Rel. market shares
A	**,53**	,06	,01	,02	,01	,37	,30
B	,11	**,33**	,04	,02	,00	,50	,08
C	,06	,03	**,49**	,02	,04	,36	,07
D	,16	,06	,03	**,28**	,09	,38	,04
E	,23	,00	,05	,02	**,29**	,41	,03
No-purchase	,24	,08	,05	,05	,03	**,55**	,48

Table 2. Average conditional switching probabilities for half-yearly data (case 1)

	A	B	C	D	E	No-purchase	Rel. market shares
A	**,43**	,05	,01	,02	,01	,48	,24
B	,14	**,32**	,01	,07	,02	,44	,08
C	,05	,04	**,37**	,01	,02	,51	,05
D	,19	,08	,02	**,21**	,06	,44	,04
E	,13	,03	,03	,09	**,27**	,45	,03
No-purchase	,20	,06	,04	,04	,02	**,64**	,56

Table 3. Average conditional switching probabilities for tertially data (case 2)

	A	B	C	D	E	No-purchase	Rel. market shares
A	**,36**	,05	,00	,03	,01	,55	,20
B	,11	**,28**	,01	,05	,01	,54	,06
C	,03	,02	**,27**	,02	,01	,65	,04
D	,14	,06	,03	**,18**	,04	,55	,04
E	,11	,03	,02	,07	**,24**	,53	,02
No-purchase	,17	,05	,04	,04	,02	**,68**	,64

Table 4. Average conditional switching probabilities for quarterly data (case 3)

The last columns of Tables 2 - 4 show the average relative market shares. Since these numbers are not exclusively based on sales they have to be interpreted with care. Hundred percent represent all purchase and non-purchase occasions considered. As an example, in Table 4 the switching behaviour of approximately 400 households has been studied, each having (at least) one of these occasions per period. Thus, brand A has been purchased on average 80 times per period. In accordance with Table 1 the shares of non-buyers increase with decreasing lengths of observation periods, and therefore brand shares decrease. However, the proportions between the brand shares remain approximately the same, irrespective of the length of the observation period considered. It is interesting to note that this pattern is also observed for brand switching probabilities which is especially appealing when once again considering the substantial percentages of purchases not considered for the longer observation periods (column 6, Table 1).

When looking at the main diagonal of the switching matrices in Tables 2 - 4 brand C and to a lesser extent brand E are able to attract a greater percentage (in a relative sense with respect to their market shares) of loyal consumers. This is consistent with the fact that C and E are distributed in special retail chains only and, hence, there might also be some store loyalty effect in these cases.

Switching matrices have been analysed in the literature extensively. One important finding which extends to a broad variety of different markets is the 'switching is proportional to share' phenomenon (e.g. Ehrenberg (1988)). It results in conditional switching probabilities which are (almost) independent from the brand purchased previously, i.e. with exception of the main diagonal the elements within each column of the switching matrix are almost constant and much smaller than the repurchase probability. This phenomenon seems to hold for Austrian dishwashing liquids, too. Interestingly, this pattern also holds for the no-purchase option which shows the same regularity (last row in Tables 2 - 4). The no-purchase columns in Tables 2 - 4 behave somewhat differently: once again there is not much variation within the off-diagonal elements but they all are much larger in magnitude (1) than the off-diagonal elements of the corresponding rows and (2) relative to the main diagonal elements (in comparison with the relationships between off-diagonal and main diagonal elements of other columns). One way to interpret this finding from a consumer behaviour perspective is that inter-purchase times do not depend on brands and are on average longer than the observation interval, thus increasing the probability of a no-purchase after having bought in the preceding period. On the other hand, switching from a no-purchase to a certain brand depends on this brand's market share. Obtaining these findings, we discussed the situation with commercial market researchers who have a good knowledge about the analysed market and they confirmed us that the results attain a high level of face validity.

Hence, we conclude that the phenomenon of 'switching is proportional to share' could be extended to include a no-purchase option, and moreover the results do not depend on the length of the observation period chosen. In the next section we will include these findings in a general brand switching model.

4 A brand switching model including the no-purchase option

4.1 The model

Very much in the spirit of pioneering scholars like Ehrenberg (1959, 1988) the model tries to mirror empirically observed regularities. This rationale seems to be especially appropriate in the present case since little is known about customers' no-purchase decisions because of the specific nature of such decisions. As outlined above already, the basic assumption for the approach is that unconditional switching between two brands is proportional to the product of their market shares. When compared with models also starting from this assumption (e.g. Ehrenberg (1988), Wagner et al. (2001)) several extensions are postulated:

(i) Market shares (m_{it}) are not required to be stable but may vary between subsequent periods ($t, t{+}1$). This allows some flexibility and also seems to be more realistic.

(ii) The proportionality factor ρ is not required to be independent of the brands but may rather be brand specific, i.e. ρ_i depends on the brand purchased on the previous occasion. The motivation for this assumption comes from the idea that there might be some difference according to the sequence of buying different brands. Such a differentiation could be due to the experiences made when consuming the product or due to some marketing activities (e.g. price cuts, sales promotions) not explicitly considered here. Tables 2 - 4 also support this generalization since off-diagonal elements are not constant for a given column.

(iii) In addition to considering I brands a further state '$I{+}1$' is included which represents the no-purchase option. Thus, it is assumed that the decision to make a purchase in the product class is basically similar to the decision in favour of a certain brand. In a strict sense this conjecture is not correct because brand choice is conditioned on purchase incidence. It is, however, common practice to assume independence between these decisions mainly due to pragmatic reasoning (e.g. Gupta (1988)). For the present case this simplification is further justified because the model is formulated on the aggregate level.

(iv) The distinctiveness of the no-purchase option outlined in section 3 requires extra flexibility which is realised by introducing an additional proportionality factor ρ_{I+2} modelling switching into the no-purchase state, whereas on the other hand, ρ_{I+1} represents switching out of the no-purchase state.

Therefore, the model may be presented in mathematical terms by means of conditional switching probabilities as

$$p_{j|i} = \rho_i \cdot m_{jt+1} \qquad i = 1, \ldots, I+1 \quad i \neq j \quad j \leq I \tag{1}$$

$$p_{I+1|i} = \rho_{I+2} \cdot m_{I+1t+1} \qquad i = 1, \ldots, I \tag{2}$$

$$p_{j|j} = (1 - \rho_j) + \rho_j \cdot m_{jt+1} + m_{I+1t+1} \cdot (\rho_j - \rho_{I+2}) \qquad j \leq I \tag{3}$$

$$p_{I+1|I+1} = (1 - \rho_{I+1}) + \rho_{I+1} \cdot m_{I+1t+1} \tag{4}$$

with:	$p_{j\mid i}$	conditional probability to switch from brand i at time t to brand j at time $t+1$
	$p_{j\mid j}$	conditional probability to repurchase brand j
	ρ_i	brand specific switching constant $(0 < \rho_i < 1)$, $(i = 1, \ldots, I)$
	ρ_{I+1},ρ_{I+2}	parameters representing switching out, into, respectively the no-purchase state $(0 < \rho_{I+1}, \rho_{I+2} < 1)$
	m_{jt}, m_{jt+1}	market share of brand j for period t, $(t+1)$, respectively
	t	time index
	i, j	brand indices $(i, j = 1, \ldots, I+1)$
	I	number of brands on offer
	$I+1$	no-purchase state

The parameters ρ_i may be interpreted as indicators for the degree of competition in this market: the smaller ρ_i, the less switching and the higher the percentage of loyal customers. Equations (3) and (4) represent repurchase probabilities and ensure logical consistency of the model, i.e. $\sum_{j=1}^{I+1} p_{j|i} = 1$, $\forall i$ or written in terms of unconditional probabilities p_{ij}: $\sum_{j=1}^{I+1} p_{ij} = m_{it}$, $\forall i$. In a similar vein, logical consistency also requires

$$\sum_{i=1}^{I+1} p_{ij} = m_{jt+1} \qquad j = 1, \ldots, I+1 \tag{5}$$

Since market shares sum to 1, one of these constraints is redundant. Therefore, the $I + 2$ parameters of the model are constrained by the I equations (5). This results in a two-dimensional range of feasible solutions $\{(0 < \rho_{i^*} < 1) \times (0 < \rho_{I+2} < 1)\}$ (for convenience of presentation results are provided in terms of ρ_{i^*} and ρ_{I+2}). After some algebraic transformations (5) may be written as:

$$\rho_j = A + \frac{c_{i^*}}{c_j} \cdot (\rho_{i^*} - A) \qquad j = 1, \ldots, I \tag{6}$$

$$\rho_{I+1} = A - \frac{(1 - \rho_{I+2})}{c_{I+1} \cdot (1 - m_{I+1t+1})} \tag{7}$$

$$\text{with:} \quad A = \frac{1 - m_{I+1t+1} \cdot \rho_{I+2}}{1 - m_{I+1t+1}}$$

$$c_j = m_{jt}/m_{jt+1} \qquad \forall j$$

$$i^* \text{ such, that } c_{i^*} = \min_{i \leq I}\{c_i\}$$

In view of (6) and (7) several comments are in order:

- Constraints (6) apply for the *brand specific* probabilities. In the stationary case ($c_j = 1, \forall j \leq I$), one finds $\rho_j = \rho_{i^*}, \forall j$, and (1), (3) reduce to the basic 'switching is proportional to share' model. Further, it can be shown that for a feasible ρ_{i^*} all ρ_j will also be feasible, i.e. ($0 < \rho_i < 1$). From an interpretative point of view (6) makes sense since it basically states that a decreasing market share corresponds with an increasing tendency for switching and thus less loyalty.
- Constraint (7) describes the implied relationship between the switching probabilities into and out of the no-purchase state. In the stationary case $\rho_{I+1} = \rho_{I+2}$, but ρ_{I+2} might still be different from ρ_{i^*}. It can be shown that ρ_{I+1} will be feasible as long as $\max\{0, (1 - c_{I+1})/(1 - m_{I+1t})\} < \rho_{I+2} < 1$.
- The share of non-buyers in period $t+1$, m_{I+1t+1}, is of dominating importance when calculating the switching constants ρ_j. This underlines that the proposed model differs significantly from its nested version exclusively accounting for purchases. However, it still represents a very parsimonious, logically consistent model with two parameters only.

4.2 Estimation

It is obvious that (relative) market shares for each brand as well as the share of non-buyers for two subsequent periods are essential ingredients of the model. If additional data on e.g. switching patterns are available, a least squares approach could be employed. In section 3 it was shown that calculating switching matrices might cause problems because of the inherent difficulty with multiple purchases within one observation period. Therefore, an estimation method is proposed which uses market shares only, i.e. an entropy based procedure. Researchers of the Hendry Corporation (1970) have been the first to apply such a concept for switching matrices. The entropy principle provides ' ... a criterion for setting up probability distributions on the basis of partial knowledge ... ' and ' ... is the least biased estimate possible on the given information' (Jaynes (1957)). These characterizations are appropriate for the present situation since information on non-buying is rather limited. Therefore, ρ_{i^*} and ρ_{I+2} are estimated by maximizing the entropy H of the conditional purchase probabilities:

$$\max H(\rho_{i^*}, \rho_{I+2}) = -\sum_{i=1}^{I+1}\sum_{j=1}^{I+1} p_{j|i} \cdot ln(p_{j|i}) \tag{8}$$

Fang et al. (1997) show that H possesses several attributes (e.g. concavity) making it an attractive optimization function. The proposed estimation procedure is rather simple so that it can be implemented on standard software like EXCEL or R easily. Preliminary empirical applications produce satisfactory results (Ruhsam (2004)) but more research is needed to provide definite conclusions on the performance of this routine.

5 Summary and further research

Three aspects of this paper appear to be of major importance.

1) The procedure generally used to define brand switching and brand loyalty by means of consumer panel data is limited in scope since it is difficult to handle multiple purchases within a single observation period. If the length of this period is fixed the researcher runs the risk of having to neglect a significant number of purchases, especially originating from more frequently buying households, which in turn might generate biased results. A more flexible scheme is required to cope with this issue.
2) Analysing household panel data for the Austrian dishwashing liquids market, the well established regularity 'switching is proportional to share' was generalized to include a no-purchase option. However, more research is needed to investigate whether these regularities extend e.g. to different products, different countries, different time spans, or different data sources.
3) A simple, parsimonious model requiring quite modest input data was proposed and empirically tested to a limited extent. Validation should be carried out on a large scale basis of empirical data. If results support the proposed model one could think of quite some further applications:
   a) Since market shares only are required as input data these market shares could have been calculated from store audit scanning data. In this case switching matrices could only be inferred from the data but not compared to actual consumer switching patterns because they are unknown at the store level. Marketing managers who have scanning data at hand only could use this model as a benchmark on the extent of competition and loyalty for the analysed market (see Reisinger et al. (2003)).
   b) Some retail chains (for example the discounter Hofer in Austria) do not cooperate with market research companies collecting store audit data and sell private label brands. Because of its modest data requirements the model could still be applied if estimates for the market shares of the brands offered in these chains are available. One should keep in mind, however, that it has to be easy for consumers to switch between the outlets of the different retail chains. Otherwise, fragmented markets would arise and 'switching is proportional to share' would no longer apply.
   c) The way how we calculated the share of non-buyers is open to criticism. As an alternative one might estimate this number by just adding a third parameter to the model, e.g. a market expansion factor. Market shares based on sales would have to be transformed appropriately

before entering into the model. If an estimate for the market potential is available for the market under consideration this information should be included within the estimation procedure.

References

ARORA, N., ALLENBY, G.M., and GINTER, J.L. (1998): A Hierarchical Bayes Model of Primary and Secondary Demand. *Marketing Science, 17 (1), 29–44.*

CHIANG, J. (1991): A Simultaneous Approach to the Whether, What and How Much to Buy Questions. *Marketing Science, 10 (4), 297–315.*

CHIB, S., SEETHARAMAN, P.B., and STRIJNEV, A. (2004): Model of Brand Choice with a No-Purchase Option Calibrated to Scanner Panel Data. *Journal of Marketing Research, 41 (2), 184–196.*

CHINTAGUNTA, P.K. (2002): Investigating Category Pricing Behaviour in a Retail Chain. *Journal of Marketing Research, 39 (2), 141–154.*

EHRENBERG, A.S.C. (1959): The Pattern of Consumer Purchases. *Applied Statistics, 8, 26–41.*

EHRENBERG, A.S.C. (1988): *Repeat-Buying - Facts, Theory and Applications.* Charles Griffin, London.

FANG, S.C., RAJASEKERA, J.R., and TSAO, H.S.J. (1997): *Entropy Optimization and Mathematical Programming.* Kluwer Academic Publishers, Boston.

GUADAGNI, P.M. and LITTLE, J.D.C. (1983): A Logit Model of Brand Choice Calibrated on Scanner Data. *Marketing Science, 2 (3), 203–238.*

GUADAGNI, P.M. and LITTLE, J.D.C. (1998): When and What to Buy: A Nested Logit Model of Coffee Purchase. *Journal of Forecasting, 17, 303–326.*

GUPTA, S. (1988): Impact of Sales Promotions on When, What, and How Much to Buy. *Journal of Marketing Research, 25 (4), 342–355.*

HENDRY CORPORATION (1970): Hendro-Dynamics: Fundamental Laws of Consumer Dynamics, Chapter 1.

JAYNES, E.T. (1957): Information Theory and Statistical Mechanics. *Physical Review, 106, 620–630.*

KRISHNAN, T.V. and SEETHARAMAN, S. (2002): A Flexible Class of Purchase Incidence Models. *Review of Marketing Science*, Working Paper No. 20021136.

LEEFLANG, P.S.H., WITTINK, D.R., WEDEL, M., and NAERT, P.A. (2000): *Building Models for Marketing Decisions.* Kluwer Academic Publishers, Boston et al.

NEVO, A. (2001): Measuring Market Power in the Ready-to-Eat Cereal Industry. *Econometrica, 69 (2), 307–342.*

REISINGER, H., WAGNER, U., and SCHUSTER, M. (2003): Die Schätzung von Markentreue, Nichtkäuferanteil und Marktpotenzial aus Handelspaneldaten. In: U. Leopold-Wildburger and G. Wäscher (Eds.): *Operations Research Proceedings 2002.* Springer, Berlin et al., 127–132.

RUHSAM, M. (2004): Die Analyse des Markenwahlverhaltens auf Basis von Übergangsmatrizen mit expliziter Berücksichtigung einer Nichtkaufoption, unpublished Master thesis, University of Vienna.

WAGNER, U., REISINGER, H., and GAUSTERER, K. (2001): Die Bestimmung des Markenwechselverhaltens mit Hilfe von Querschnittsdaten. *Zeitschrift für Betriebswirtschaft, 71 (10), 1113–1130.*

klaR Analyzing German Business Cycles*

Claus Weihs, Uwe Ligges, Karsten Luebke, and Nils Raabe

Fachbereich Statistik, Universität Dortmund, D-44221 Dortmund, Germany
e-mail: weihs@statistik.uni-dortmund.de

Abstract. Decision making often asks for classification. We will present a new R package klaR including functions to build, check, tune, visualize, and compare classification rules. The software is illustrated by means of a case study of prediction of the German economy's business cycle phases.

1 Introduction

Decision making often asks for classification. In the last years the founding of classification methods explodes because of the data mining boom. What is urgently needed from our point of view to help practitioners applying classification methods is a well founded generally available software representing the state of the art not only corresponding to classification methodology but also corresponding to result assessment and visualization as well as model selection. Therefore, we will present a new R package klaR including functions to build, check, tune, visualize, and compare classification rules.

When looking for the best classification rule typical questions are:

- Which classification method should be used?
- How can classification rules be compared?
- What are the important variables?
- How can the results be interpreted?
- How can special data structure be incorporated in the model?

The new R (R Development Core Team (2004)) package klaR was developed to support practitioners in answering such questions common to many classification problems. In this paper we consider the classification analysis of the German economy's business cycle phases (Heilemann and Münch (1996)) as a case study to illuminate how the software package klaR helps to answer the above questions. Note that the package also includes some classification methods that have not been included in any other R package before.

The package is available from CRAN[1].

The data of our case study is described in Section 2, and some classification tools are applied in Section 3. Results of different methods are compared in Section 4 while variable selection is applied in Section 5. In Section 6

* This work has been supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 475. We also thank Christian Röver for his contributions.

[1] http://CRAN.R-project.org

discrimination is visualized. The incorporation of background knowledge is shown in Section 7. A new method for the visualization of data structure is given in Section 8.

2 Data: "West German Business Cycles"

Case study data consist of 13 economic variables with 157 quarterly observations from 1955/4 to 1994/4 (see Heilemann and Münch (1996)) of the German business cycle. The German business cycle is classified in a four phase scheme: upswing (class 1), upper turning point (2), downswing (3) and lower turning point (4). There were 6 complete cycles in the observed time period. Aim of the analysis is to derive a classification rule for the phases.

If prediction ability is tested by 'leave-one-cycle out' validation, i.e. each one business cycle is left out as a validation set once, the other 5 cycles are used to train the method, Weihs and Garczarek (2002) show that in general LDA is among the best classifiers for this classification task.

If klaR has been installed and loaded properly, case study data can be made available calling `data("B3")`.

3 Classification methods

Two of the functions providing classification tools that are available in package klaR are Regularized Discriminant Analysis (RDA), and Support Vector Machines (SVM). Note that we will set the parameters of these methods to sensible values, but we have not tuned them for optimal prediction, which for RDA is implemented by means of Simulated Annealing. The aim of this paper is to present features of klaR, but not to compare and judge about classification methods. In klaR tuning of classification rules can be done by optimizing the choice of meta-parameters or an appropriate pre- and post-processing of the data or results.

Regularized Discriminant Analysis, as introduced by Friedman (1989), is a generalization of LDA and QDA, the R functions of which are both taken in klaR from package MASS, see Venables and Ripley (2002). On the one hand, as an advantage, RDA is more robust against multicollinearity than QDA, but on the other hand, it is less easily interpretable.

The covariance matrices are manipulated using two parameters, γ and λ. These parameters can be fixed, e.g., by minimizing the misclassification rate (based on an internal cross-validation). Special cases are:

(γ=0, λ=0): same as QDA, individual covariances for each group.
(γ=0, λ=1): same as LDA, a common covariance matrix.
(γ=1, λ=0): Conditional independence, identical variances within class (similar to Naive Bayes).
(γ=1, λ=1): Objects are assigned to class with nearest mean (euclidean).

In the following example, the B3 data is analyzed using RDA by setting the parameters γ and λ manually:

```
> library(klaR)
> data(B3)
> train <- 12:106
> test <- 107:154
> rda_obj <- rda(PHASEN ~ ., data = B3[train, ],
+                     gamma = 0.05, lambda = 0.1)
> rda_pred <- predict(rda_obj, B3[test,])$class
```

For predictive power comparisons, one might argue that cross-validation as, e.g., leave-one-out or leave-one-cycle-out should be applied (cp., e.g., Section 5).

Here we start with a train-and-test-environment, i.e. with separation into training and test sets. The first five complete cycles (observation number (12–106)) will serve as the training set and the sixth cycle (obs. 107–154) as the test set. Note that by `PHASEN ~ .` the variable `PHASEN` of the `B3` data is set as the response variable, and that the `predict` method of `rda` returns a list with the a-posteriori class probabilities as well as the predicted class of the test data in the the `class` element.

The package klaR provides an interface to the **Support Vector Machine** implementation SVM$^{\text{light}}$ by Joachims (2004) which (in contrast to the SVM implementation in R-package 'e1071', see Dimitriadou et al. (2004)) supports loss parameters and 1-against-all classification. Moreover, it returns comparable membership scores ('posteriors'). An example using default argument setting of parameters, analogously to the RDA example above, looks as follows:

```
> svm_obj <- svmlight(PHASEN ~ ., data = B3[train, ])
```

Loss parameters and different kernel options can be passed to SVM$^{\text{light}}$ by setting parameters in `svm.options = (...)` in the `svmlight` call.

In business phase classification the cyclic structure of the B3 data classes can be used to improve classification (see Section 7).

4 Comparing classifications

Comparison of classification rules often starts with looking at the estimated error rates. A first idea of the structure of misclassifications of the RDA results using the train-and-test environment can be derived from the confusion matrix:

```
> errormatrix(B3[test, "PHASEN"], rda_pred)
          predicted
true       1  2  3  4 -SUM-
```

```
1      19 0 1  7    8
2       0 1 0  5    5
3       0 0 2  7    7
4       0 0 2  4    2
-SUM-   0 0 3 19   22
```

The matrix shows that 22 out of 48 are misclassified, worst rates when the true class is "utp" (class 2), when most misclassifications go into class "ltp" (class 4).

Also of importance are other error characteristics like the so-called confidence of the classification rule (see below) and the ability to separate the classes. For a large subset of classification methods, the so called "arg-max" classifiers, such performance measures can be compared directly after the result of the classifier has been scaled so that the size of the membership values of the different classification rules are comparable for the different classes (Garczarek and Weihs (2003)). Scaling methods map membership values to so-called scaled membership values or posteriors in $[0,1]^K$, K = number of classes, summing up to 1. Package klaR includes functions for so-called softmax-scaling and beta-scaling and a function to compare the results. Moreover, klaR enables the visualization of the membership values for 3 and 4 classes in a barycentric coordinate system.

In our example, we start by looking at posterior assignments. With RDA in the test set each observation is assigned to a class with a certain posterior probability:

```
> rda_post <-
+     predict(rda_obj, newdata = B3[test, ])$posterior
> round(rda_post, 3)
         1     2     3     4
 [1,] 0.000 0.000 0.994 0.006
 [2,] 0.000 0.000 0.999 0.001
 . . . . . . . . . .
```

The probability distribution over 4 classes may be illustrated by a point in a 3-dimensional simplex, i.e. a 'tetraeder' also called 'barycentric plot', where each corner corresponds to one class. For a point in a 'barycentric plot' the probability of a certain class is proportional to the distance to the opposite edge on the line starting at the corner of the class of interest. By calling

```
> par(mfrow = c(1, 2))
> s3d <- quadplot(rda_post, labelcol = 1, labelpch = 1:4,
+                 pch = as.numeric(B3$PHASEN[test]),
+                 main = "RDA posterior assignments")
> quadlines(centerlines(4), sp = s3d, lty = "dashed")
```

Figure 1 is generated for visualizing the RDA posteriors. The graphical parameters like `mfrow, labelcol, ...` are set in order to achieve a "nice" black-and-white print in this paper (see the help pages for klaR). SVM posteriors can be plotted analogously. Note that our implementation of SVM$^{\text{light}}$ automatically scales SVM membership values by means of softmax-scaling. Other functions in klaR not shown here include scaling of membership values so that, e.g., the confidence of a result of a SVM is more realistic. See the examples in the help page of `quadplot` for some more informative plots additionally using colors.

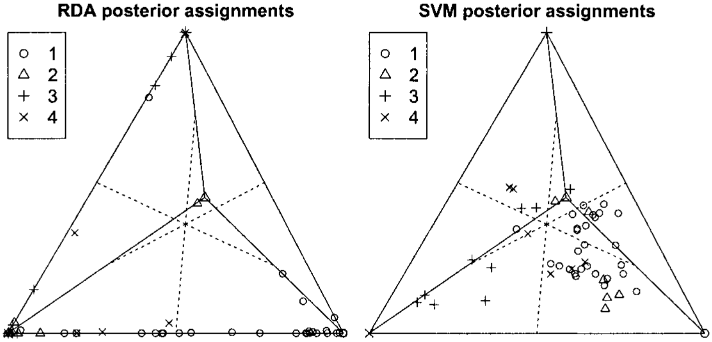


Fig. 1. Barycentric plots: posteriors of the RDA and SVM analyses.

As shown in Figure 1, RDA results in greater posterior probabilities (points on edges and corners) while the analysis using SVM$^{\text{light}}$ has more uncertainty (points inside simplex). The classification performance measures

Correctness rate: 1 - error rate
Accuracy: mean distance to 'true' corner
Ability to separate: mean distance to classified corner
Confidence: mean membership of assigned class (either by class or average)

described by Garczarek and Weihs (2003) can be derived from the scaled membership values and calculated for the RDA case by the following call:

```
> ucpm(m = rda_post, tc = B3$PHASEN[test])
```

where `m` stands for the membership values and `tc` for true classes.

	LDA	RDA	SVM
Correctness rate	0.44	0.54	0.63
Accuracy	0.07	0.18	0.09
Ability to separate	0.76	0.75	0.28
Confidence	0.84	0.84	0.46

Table 1. Classification performance measures for methods LDA, RDA, and SVM.

Table 1 confirms the impression from Figure 1 that SVM$^{\text{light}}$ is more insecure about its assignment (Confidence). On the other hand, one can see that in terms of correctness rate on the test set SVM$^{\text{light}}$ is superior to RDA which itself outperforms LDA.

5 Variable selection

The function stepclass performs "stepwise classification", i.e. it selects variables to be used in the classification rule by optimizing one of the performance measures of ucpm, e.g., by maximizing the 10-fold cross validated correctness rate. Unlike other variable selection methods, like Wilks Λ for example, stepclass does not need any distributional assumptions and is ready for general application as long as the classification method returns posteriors for new observations. Available methods for variable selection are *forward selection* (add variables to null model), *backward selection* (remove variables from full model), or *both directions* (starting with the null model).

As an example, consider stepwise selection in both directions on the B3 data using LDA:

```
> # make stepclass-internal crossval reproducible:
> set.seed(1111)
> lda_scl <- stepclass(PHASEN ~ ., data = B3[train, ],
+                      method = "lda", prior = rep(1/4, 4))
> lda_scl
method      : lda
final model : EWAJW, LSTKJW, ZINSK
correctness rate = 0.64
```

By prior=rep(1/4,4) the a-priori probabilities for every class are set to $\frac{1}{4}$. In our case the error rate based on only 3 economic input variables on the test set (B3[test,]) calculated like in Section 4 is 27%, while using all 13 variables the error rate on the test set is 56% (compare Table 1).

6 Visualization of partitions

The classification result for any two variables can be visualized by drawing the partitions of the data. By such a visualization it is possible to gain insight into the (relative) location of the classes and how any pair of variables is separating the classes.

```
> partimat(B3[train, lda_scl$model$name],
+          B3[train, "PHASEN"], method = "lda"
+          prior = rep(1/4, 4), plot.matrix = TRUE,
+          imageplot = FALSE, col.contour = "black",
+          col.wrong = "darkgrey", print.err = 1.3)
```

By using only the variables that are selected by `stepclass` which are stored in `lda_scl$model$name` a parsimonious rule is plotted. Again the other options in `partimat` are only needed to ensure a "good" black-and-white plot.

`partimat` can be used with quite a lot of classification methods including for example LDA, QDA, RDA and SVM$^{\text{light}}$. Figure 2 shows that typically the upper and lower turning points (classes 2 and 4) are separated by the other classes.

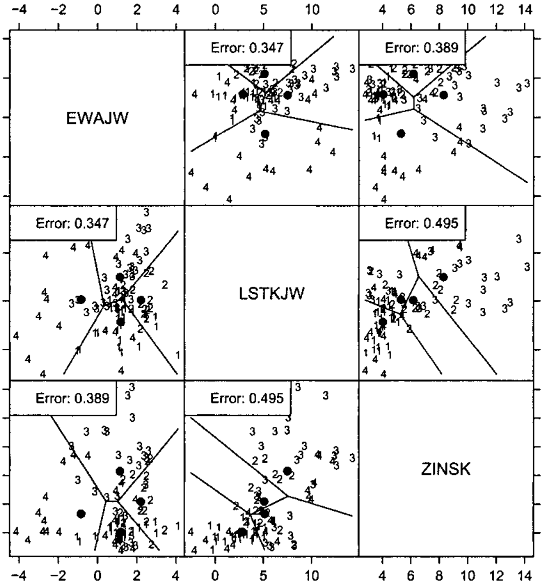


Fig. 2. Partition plot of selected variables by LDA.

7 Using background knowledge

In Business Cycles classification one can utilize the fact that the class is constant most of the time and when it changes it can only change, e.g., from upswing to upper turning point. These transition probabilities can be estimated by

```
> tm <- calc.trans(B3$PHASEN[train])
```

A posteriori probabilities of classes together with transition probabilities can be used together in a Hidden Markov Model (HMM):

```
> lda_hmm <- hmm.sop(sv = "4", trans.matrix = tm,
+                    prob.matrix = lda_scl_post)
```

where sv stands for the start class of the Markov chain (by definition a business cycle begins with a lower turning point). By applying HMM to the parsimonious rule of LDA chosen by stepclass) the error rate on the test set can be even reduced to 19%. The whole procedure (stepclass → partimat → hmm.sop → ucpm) is not limited to LDA, so a lot of classification methods can be improved by background knowledge.

8 Visualization of data structure

Finally, let us confirm that the 4 class structure of our business class data is adequate. The function EDAM (Eight Directions Arranged Map) computes a distance-based two-dimensional representation of the data in a rectangular grid known from Self-Organizing Maps (Raabe et al. (2004)). The result of EDAM for the last cycle is shown in Figure 3, visualized by the function shardsplot. In this plot the distances of the corners of the "shards" correspond to the true – in our case Euclidean – distances. Figure 3 shows a counterclockwise arrangement of the phases starting with the lower turning point starting in 1982, represented by the white shards and closing with the downswing till 1994 represented by light gray. The cycle is obvious so that the 4 class structure might be reasonable.

```
> last.cycle <- B3[107:154,]
> set.seed(1234)
> lcEDAM <- EDAM(last.cycle[,2:14], iter.max = 20,
+       classes = last.cycle[,1], standardize = TRUE)
> plot(lcEDAM, stck = FALSE, standardize = TRUE,
+       vertices = FALSE, asp = 1,
+       classes = last.cycle[,1], classcolors = "gray")
```

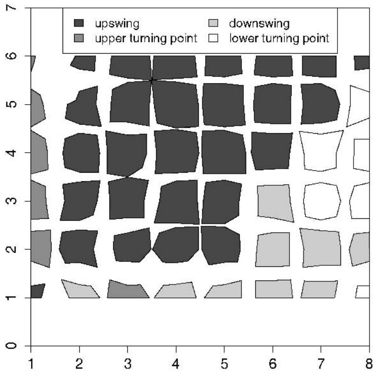


Fig. 3. Shardsplot of the last business cycle.

9 Summary

The new package klaR provides several functions which are useful for data analysis in a classification framework. In the future more new classification methods will be included in the package so the practitioner can apply state-of-the-art classifiers on his/her data.

By applying some functions of klaR it was possible to achieve a parsimonious and improved (in terms of error rate) classification rule for Business Cycle phases.

References

DIMITRIADOU, E., HORNIK, K., LEISCH, F., MEYER, D., and WEINGESSEL, A. (2005): *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.* R package version 1.5-6.

FRIEDMAN, J.H. (1989): Regularized Discriminant Analysis. *Journal of the American Statistical Association, 84, 165–175.*

GARCZAREK, U. and WEIHS, C. (2003): Standardizing the Comparison of Partitions. *Computational Statistics, 18, 143–162.*

HEILEMANN, U. and MÜNCH, J.M. (1996): West german business cycles 1963-1994: A multivariate discriminant analysis. *CIRET-Conference in Singapore, CIRET-Studien 50.*

JOACHIMS, T. (2004): SVM$^{\text{light}}$. http://svmlight.joachims.org/

R DEVELOPMENT CORE TEAM (2004): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

RAABE, N., LUEBKE, K., and WEIHS, C. (2004): KMC/EDAM: A new approach for the visualization of K-Means Clustering results. *Technical Report 65/2004, SFB 475, Universität Dortmund.*

VENABLES, W.N. and RIPLEY, B.D. (2002): *Modern Applied Statistics with S*, 4th ed. Springer, New York.

WEIHS, C. and GARCZAREK, U. (2002): Stability of multivariate representation of business cycles over time. *Technical Report 20/2002, SFB 475, Universität Dortmund.*

Index

Selected Publications of Wolfgang Gaul

Books and Proceedings

Classification as a Tool of Research (edited with M. Schader). North Holland, Amsterdam, 1986.

Data, Expert Knowledge and Decisions: An Interdisciplinary Approach with Emphasis on Marketing Applications (edited with M. Schader). Springer, Heidelberg, 1988.

Computergestütztes Marketing (Computer-Assisted Marketing) (with M. Both). Springer, Heidelberg, 1990.

Knowledge, Data and Computer-Assisted Decisions (edited with M. Schader). Springer, Heidelberg, 1990.

Operations Research Proceedings 1991 (edited with A. Bachem, W. Habenicht, W. Runge, W. Stahl). Springer, Heidelberg, 1992.

Special Issue of OR-Spektrum on *Produktionsplanung und -steuerung* (edited with G. Wäscher). *OR-Spektrum*, 1992.

Marktforschung und Marketing-Management - Computerbasierte Entscheidungsunterstützung (Market Research & Marketing-Management - Computer-Assisted Decision Support) (with D. Baier). Oldenbourg, München, 1993.

Wissensbasierte Marketing-Datenanalyse - Das WIMDAS-Projekt (Knowledge-Based Marketing Data Analysis: The WIMDAS-Project) (edited with M. Schader). Lang, Frankfurt, 1994.

From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis, and Knowledge Organization (edited with D. Pfeifer). Springer, Heidelberg, 1995.

Special Issue of Annals of Operations Research on *Data, Expert Knowledge, and Decisions: Using Knowledge to Transform Data into Information for Decision Support* (edited with F.J. Radermacher, M. Schader, D. Solte). *Annals of Operations Research*, 1995.

Operations Research Proceedings 1996 (edited with U. Zimmermann, U. Derigs, R.H. Möring, K.-P. Schuster). Springer, Heidelberg, 1997.

Classification in the Information Age (edited with H. Locarek-Junge). Springer, Heidelberg, 1999.

Mathematische Methoden der Wirtschaftswissenschaften (edited with M. Schader). Physica, Frankfurt, 1999.

Data Analysis: Scientific Modeling and Practical Application (edited with O. Opitz, M. Schader). Springer, Heidelberg, 2000.

Classification and Information Processing at the Turn of the Millennium (edited with R. Decker). Springer, Heidelberg, 2000.

Classification, Automation, and New Media (edited with G. Ritter). Springer, Heidelberg, 2002.

Between Data Science and Applied Data Analysis (edited with M. Schader, M. Vichi). Springer, Heidelberg, 2003.

Classification, Clustering, and Data Mining Applications (edited with D. Banks, L. House, F.R. McMorris, P. Arabie). Springer, Heidelberg, 2004.

Classification: The Ubiquitous Challenge (edited with C. Weihs). Springer, Heidelberg, 2005.

The Entrepreneurship - Innovation - Marketing Interface (edited with R. Würth, V. Jung). Swiridoff, Künzelsau, 2005.

Scientific Publications

On Constrained Shortest-Route Problems. In: *Optimization and Operations Research, Lecture Notes in Economics & Mathematical Systems, 117*, 1975.

Einige Aspekte in der Zuordnungstheorie (Some Aspects in Assignment Theory). *Zeitschrift für Operations Research, 19*, 1975 (with A. Heinecke).

Zur Methode der paarweisen Vergleiche und ihrer Anwendung im Marketingbereich (On the Method of Paired Comparisons and its Application to Marketing). *Methods of Operations Research, 35*, 1978.

Some Structural Properties of Project Digraphs. *Journal of Combinatorics, Information & System Sciences, 3*, 1978.

Zur Zuverlässigkeit graphischer Systeme (On the Reliability of Graphical Systems). *Operations Research Verfahren, 33*, 1978.

A Barrier Method with Arbitrary Starting Point. *Mathematische Operationsforschung und Statistik, Ser. Optimization, 10*, 1979 (with J. Hartung).

A Stochastic Flow Problem. *Journal of Information and Optimization Sciences, 1*, 1980 (with H.J. Cleef).

Bounds for the Expected Duration of a Stochastic Project Planning Model. *Journal of Information and Optimization Sciences, 2*, 1981.

Stochastische Projektplanung und Marketingprobleme (Stochastic Project Scheduling and Marketing Problems). *Methods of Operations Research, 44*, 1981.

Project Scheduling via Stochastic Programming. *Mathematische Operationsforschung und Statistik, Ser. Optimization, 13*, 1982 (with H.J. Cleef).

On Stochastic Analysis of Project Networks. In: Dempster, M.A.H. et. al. (Eds.): *Deterministic and Stochastic Scheduling.* Reidel Publishing Co., 1982.

On Reliability in Stochastic Graphs. *Networks, 12*, 1982 (with O. Frank).

Marketing-Logistik bei stochastischer Nachfrage (Marketing-Logistics with Stochastic Demand). *Operations Research Proceedings 1982.* Springer, Heidelerg, 1983.

Bounding Distributions for Random Project Completion Times. In: Beckmann, M.J., Eichhorn, W., Krelle, W. (Eds.): *Mathematische Systeme in der Ökonomie.* Athenaeum, 1983.

A Multidimensional Analysis of Consumer Preference Judgements Related to Print Ads. In: *Methodological Advances in Marketing Research in Theory and Practice*, EMAC/ESOMAR Symposium, Copenhagen, 1984 (with I. Böckenholt).

Analysis of Sales of Price Promoted Consumer Goods Using Scanner Data. In: *Methodological Advances in Marketing Research in Theory and Practice*, EMAC/ESOMAR Symposium, Copenhagen, 1984 (with M. Both).

Optimal Routes in Compound Transportation Systems. In: Hauptmann, H., Krelle W., Mosler K.C. (Eds.): *Operations Research and Economic Theory.* Springer, Heidelberg, 1984.

Financial Planning via Stochastic Programming: A Stochastic Flows with Gains Approach. In: *Risk and Capital - Lecture Notes in Economics and Mathematical Systems, 227*, 1984.

Zur mehrdimensionalen Analyse von Bildinformationen bei Printwerbung für Imagery-Produkte (Multidimensional Analysis of Print Ads Information for Imagery-Products). *Vierteljahreshefte für Mediaplanung, 4*, 1985 (with I. Böckenholt).

Reliability-Estimation in Stochastic Graphs with Time-Associated Arc-Set Reliability Performance Processes. *Annals of Discrete Mathematics, 28*, 1985.

Multistate Reliability Problems for GSP-Digraphs. *Lecture Notes in Economics and Mathematical Systems, 240*, 1985 (with J. Hartung).

Two-Mode Hierarchical Clustering as an Instrument for Marketing Research. In: Gaul, W., Schader, M. (Eds.): *Classification as a Tool of Research.* North Holland, Amsterdam, 1986 (with E. Espejo).

Analysis of Choice Behaviour via Probabilistic Ideal Point and Vector Models. *Applied Stochastic Models and Data Analysis, 2*, 1986 (with I. Böckenholt).

Typologisierung deutscher Marktforschungsinstitute: Ergebnisse einer empirischen Studie (Characterization of German Market Research Institutes: Results of an Empirical Study). *Marketing ZFP*, 1986 (with F. Förster, K. Schiller).

Zum Vergleich zweimodaler Clusteranalyseverfahren (Comparison of Two Mode Clustering Algorithms). *Methods of Operations Research, 57*, 1987 (with M. Both).

New Product Introduction Based on Pre-Test Market Data. In: *Micro and Macro Market Modelling: Research on Prices, Consumer Behaviour and Forecasting*, EMAC/ESOMAR Symposium, Tutzing, 1987 (with I. Böckenholt).

Neue probabilistische Auswahlmodelle im Marketing (New Probabilistic Choice Models in Marketing). *Operations Research Proceedings 1986*, Springer, Heidelberg, 1987 (with I. Böckenholt).

Probabilistic Choice Behavior Models and Their Combination with Additional Tools Needed for Applications to Marketing. *Communication and Cognition, 20*, 1987.

Probabilistic Multidimensional Scaling of Paired Comparisons Data. In: Bock, H.H. (Ed.): *Classification and Related Methods of Data Analysis.* North Holland, Amsterdam, 1988 (with I. Böckenholt).

PROLOG-Based Decision Support for Data Analysis in Marketing. In: Gaul, W., Schader, M. (Eds.): *Data, Expert Knowledge and Decisions: An Interdisciplinary Approach with Emphasis on Marketing Applications.* Springer, Heidelberg, 1988 (with I. Böckenholt, M. Both).

The Use of Data Analysis Techniques by German Market Research Agencies: A Causal Analysis. *Journal of Business Research, 16*, 1988 (with Ch. Homburg).

Clusterwise Aggregation of Relations. *Applied Stochastic Models and Data Analysis, 4*, 1988 (with M. Schader).

Marketing Data Analysis by Dual Scaling. *International Journal of Research in Marketing, 5*, 1988 (with S. Nishisato).

Probabilistic Choice Behavior Models and Their Combination with Additional Tools Needed for Applications to Marketing. In: De Soete, G., Feger, H., Klauer, K.C. (Eds.): *New Developments in Psychological Choice Modeling.* North Holland, Amsterdam, 1989. Revised version of a paper published in Communication & Cognition, 20, 1987.

Data Analysis and Decision Support. *Applied Stochastic Models and Data Analysis, 5*, 1989 (with M. Schader).

Generalized Latent Class Analysis: A New Methodology for Market Structure Analysis. In: Opitz, O. (Ed.): *Conceptual and Numerical Analysis of Data.* Springer, Heidelberg, 1989 (with I. Böckenholt).

Classification and Selection of Consumer Purchase Behaviour Models. In: Opitz, O. (Ed.): *Conceptual and Numerical Analysis of Data.* Springer, Heidelberg, 1989 (with R. Decker).

A Knowledge-Based System for Supporting Data Analysis Problems. *Decision Support Systems, 5*, 1989 (with I. Böckenholt, M. Both).

Computer-Assisted Market Research and Marketing Enriched by Capabilities of Knowledge-Based Systems. In: *New Ways in Marketing and Market Research*, EMAC/ESOMAR Symposium, Athens, 1990 (with D. Baier).

Einige Bemerkungen über Expertensysteme für Marketing und Marktforschung (Some Remarks about Expert Systems for Marketing and Market Research). *Marketing ZFP*, 1990 (with R. Decker).

An Approach to Marketing Data Analysis: The Forced Classification Procedure of Dual Scaling. *Journal of Marketing Research, 27*, 1990 (with S. Nishisato).

Pyramidal Clustering with Missing Values. In: Diday, E., Lechevallier, Y. (Eds.): *Symbolic-Numeric Data Analysis and Learning.* Nova Science, 1991 (with M. Schader).

Knowledge-Based Systems in Marketing and Market Research: Demonstrations and Comparisons of Own Systems. *Methods of Operations Research, 4*, 1991 (with R. Decker).

The MVL (Missing Values Linkage) Approach for Hierarchical Classification When Data are Incomplete. In: M. Schader (Ed.): *Analyzing and Modeling Data and Knowledge.* Springer, Heidelberg, 1992 (with M. Schader).

Paneuropäische Tendenzen in der Preispolitik - Eine empirische Studie. *Der Markt,* 32, 1993 (with U. Lutz).

Decision Making Concerning Product Line Design Based on Conjoint Analysis. In: *Operations Research 1993.* Physica, Frankfurt, 1994 (with E. Aust).

Pricing in International Marketing and Western European Economic Integration. *Management International Review, 34,* 1994 (with U. Lutz).

Neuronale Netze in der Kaufverhaltensforschung - Alternativ-Modell. *absatzwirtschaft, 37,* 1994 (with R. Decker, F. Wartenberg).

Analyse von Panel- und POS-Scanner-Daten mit Neuronalen Netzen. *Jahrbuch der Absatz- und Verbrauchsforschung, 40,* 1994 (with R. Decker, F. Wartenberg).

Berücksichtigung von Kaufvergangenheiten bei der Markenwahl. *Der Markt, 33,* 1994 (with R. Decker, M. Röhle).

Comparing Proposals for the Solution of Data Analysis Problems in a Knowledge-Based System. *Annals of Operations Research, 52,* 1994 (with D. Baier, F. Wartenberg).

Pyramidal Classification Based on Incomplete Dissimilarity Data. *Journal of Classification, 11,* 1994 (with M. Schader).

Positioning Analysis Using Knowledge Based Support. *Annals of Operations Research, 55,* 1995 (with D. Baier, F. Wartenberg).

Gewinnorientierte Produktliniengestaltung unter Berücksichtigung des Kundennutzens. *Zeitschrift für Betriebswirtschaft, 65,* 1995 (with E. Aust, D. Baier).

Analyzing Paired Comparisons Data Using Probabilistic Ideal Point and Probabilistic Vector Models. In: Bock, H.H., Polasek, W. (Eds.): *Data Analysis and Information Systems: Statistical and Conceptual Approaches.* Springer, Heidelberg, 1996 (with D. Baier).

A New Algorithm for Two-Mode Clustering. In: Bock, H.H., Polasek, W. (Eds.): *Data Analysis and Information Systems: Statistical and Conceptual Approaches.* Springer, Heidelberg, 1996 (with M. Schader).

Verfahren der Testmarktsimulation in Deutschland: Eine vergleichende Analyse. *Marketing ZFP,* 1996 (with D. Baier, A. Apergis).

Two-Mode Overlapping Clustering With Applications to Simultaneous Benefit Segmentation and Market Structuring. In: Klar, R., Opitz, O. (Eds.): *Classification and Knowledge Organization. Springer, Heidelberg,* 1997, 557–566 (with D. Baier, M. Schader).

Segment-Specific Aspects of Designing Online Services in the Internet. In: Balderjahn, I., Mathar, R., Schader, M. (Eds.): *Classification, Data Analysis, and Data Highways. Springer, Heidelberg,* 1998, 253–261 (with T. Klein, F. Wartenberg).

Zur Charakterisierung von Preisspielräumen. *Zeitschrift für betriebswirtschaftliche Forschung, 11,* 1999, 1056–1074 (with M. Löffler).

Optimal Product Positioning Based on Paired Comparison Data. *Journal of Econometrics, 89*, 1999, 365–392 (with D. Baier).

Data Mining: A New Label for an Old Problem. In: Gaul, W., Schader, M. (Hrsg.): *Mathematische Methoden der Wirtschaftswissenschaften.* Physica, Frankfurt, 1999, 3–14 (with M. Schader).

Market Simulation Using a Probabilistic Ideal Vector Model for Conjoint Data. In: Gustafsson, A., Herrmann, A., Huber, F. (Eds.): *Conjoint Measurement - Methods and Applications, Springer, Heidelberg*, 2000 (3nd ed. 2003), 97–120 (with D. Baier).

Methodeneinsatz zur Unterstützung erfolgreicher Produktinnovationen. *Zeitschrift für Unternehmensentwicklung und Industrial Engineering, 49*, 2000, 75–78 (with M. Volkmann).

Frequent Generalized Subsequences – A Problem From Web Mining. In: Gaul, W., Opitz, O., Schader, M. (Eds.): *Data Analysis: Scientific Modeling and Practical Application.* Springer, Heidelberg, 2000, 429–445 (with L. Schmidt-Thieme).

Decision Tree Construction by Association Rules. In: Decker, R., Gaul, W. (Eds.): *Classification and Information Processing at the Turn of the Millennium.* Springer, Heidelberg, 2000, 245–253 (with F. Säuberlich).

eMarketing mittels Recommendersystemen. *Marketing ZFP, 24*, 2002, 47–55 (with A. Geyer-Schulz, M. Hahsler, L. Schmidt-Thieme).

Mining Web Navigation Path Fragments. In: Nishisato, S., Baba, Y., Bozdogan, H., Kanefuji, K. (Eds.): *Measurement and Multivariate Analysis.* Springer, Heidelberg, 2002, 249–260 (with L. Schmidt-Thieme).

Recommender Systems Based on User Navigational Behavior in the Internet. *Behaviormetrika, 29*, 2002, 1–22 (with L. Schmidt-Thieme).

Product Bundling as a Marketing Application. In: *Operations Research Proceedings 2003.* Springer, Heidelberg, 2004, 221 - 228 (with B. Stauss).

Visualizing Recommender System Results via Multidimensional Scaling. In: *Operations Research Proceedings 2003.* Springer, Heidelberg, 2004, 189–196 (with P. Thoma, L. Schmidt-Thieme, S. van der Bergh).

Product Line Optimization as a Two Stage Problem. In: *Operations Research Proceedings 2004.* Springer, Heidelberg, 2005 (with B. Stauss).

Web Mining and Online Visibility. In: Weihs, C., Gaul, W., (Eds.): *Classification: The Ubiquitous Challenge.* Springer, Heidelberg, 2005 (with N. Schmidt-Mänz).

Estimating Reservation Prices for Product Bundles Based on Paired Comparison Data. In: Weihs, C., Gaul, W. (Eds.): *Classification: The Ubiquitous Challenge.* Springer, Heidelberg, 2005 (with B. Stauss).

Titles in the Series

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 Millenium. 2000

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

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

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

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

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

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

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

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

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