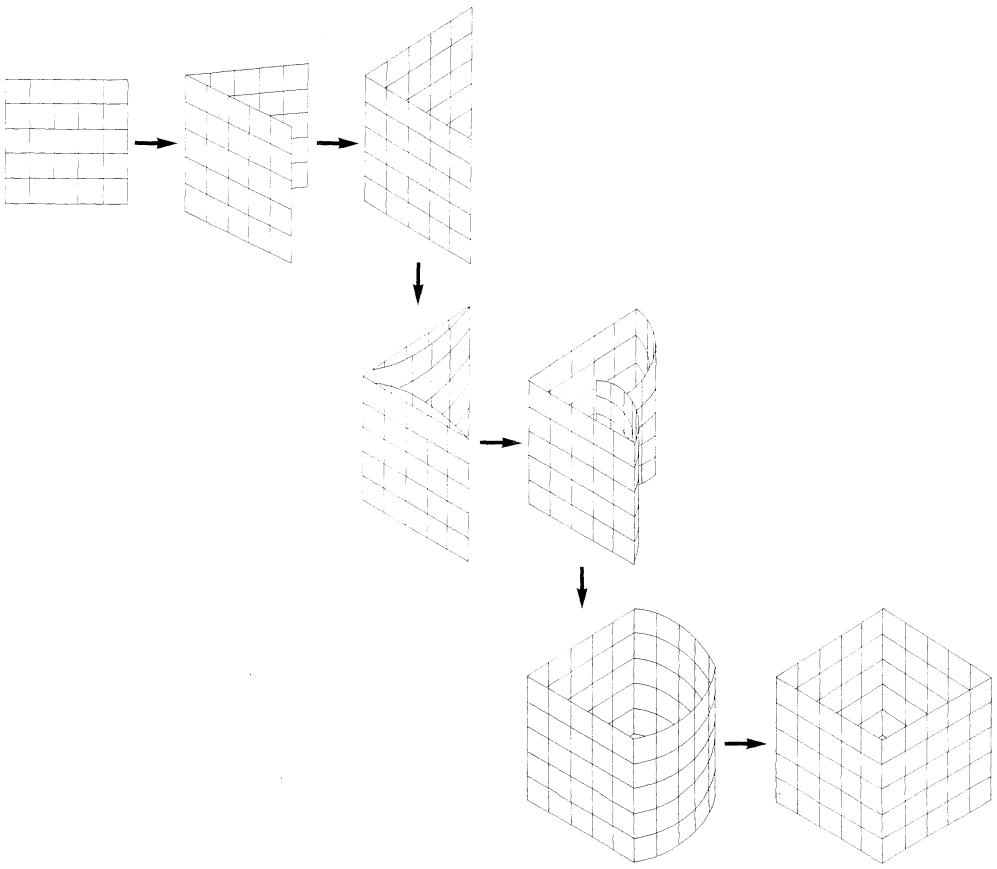
THREE-WAY METHODS FOR THE ANALYSIS OF QUALITATIVE AND QUANTITATIVE TWO-WAY DATA

Henk A. L. Kiers



DSWO PRESS

THREE-WAY METHODS FOR THE ANALYSIS OF QUALITATIVE AND QUANTITATIVE TWO-WAY DATA

M&T SERIES 15

Available from DSWO Press
Wassenaarseweg 52
2333 AK Leiden
The Netherlands
Tel. (071) 273795

Earlier publications in this series:

Jacqueline Meulman, Homogeneity analysis of incomplete data.
M&T series 1, 1982

Pieter M. Kroonenberg, Tree-mode principle component analysis.
M&T series 2, 1983, reprint 1989

Jan de Leeuw, Canonical analysis of categorical data.
M&T series 3, 1984

Ronald A. Visser, Analysis of longitudinal data in behavioural and social research
M&T series 4, 1985

John P. van de Geer, Introduction to linear multivariate data analysis.
M&T series 5, volume 1 & 2, 1986

Jacqueline Meulman, A distance approach to nonlinear multivariate analysis.
M&T series 6, 1986

Jan de Leeuw, Willem Heiser, Jacqueline Meulman, Frank Critchley (editors), Multidimensional data analysis.
M&T series 7, 1987

Peter G.M. van der Heijden, Correspondence analysis of longitudinal categorical data.
M&T series 8, 1987

Jan van Rijckevorsel, The application of fuzzy coding and horseshoes in multiple correspondence analysis.
M&T series 9, 1987

Abby Israëls, Eigenvalue techniques for qualitative data.
M&T series 10, 1987

Eeke van der Burg, Nonlinear canonical correlation and some related techniques.
M&T series 11, 1988

Kees van Montfort, Estimating in structural models with non-normal distributed variables: some alternative approaches.
M&T series 12, 1989

Jan T. A. Koster, Mathematical aspects of multiple correspondence analysis for ordinal variables.
M&T series 13, 1989

Catrien C. J. H. Bijleveld, Exploratory linear dynamic systems analysis.
M&T series 14, 1989

Henk A. L. Kiers, Three-way methods for the analysis of qualitative and quantitative two-way data.
M&T series 15, 1989

THREE-WAY METHODS FOR THE ANALYSIS OF QUALITATIVE AND QUANTITATIVE TWO-WAY DATA

Henk A. L. Kiers

Department of Psychology
University of Groningen

1989 DSWO Press, University of Leiden

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Kiers, Henk A. L.

Three-way methods for the analysis of qualitative and quantitative two-way data / Henk A. L. Kiers. – Leiden: DSWO Press. – Ill. – (M&T series; 15)
Also publ. as Thesis Groningen, 1989. With index, ref. – With summary in Dutch.
ISBN 90-6695-037-4
SISO 517.2 UDC 519.2
Subject headings: principal component analysis / multiple correspondence analysis.

Cover design Roger Busschots
Printed by 'Reprodienst, Faculteit Sociale Wetenschappen, Rijksuniversiteit Leiden'
and by Printing office 'Karstens drukkers bv, Leiden'

ISBN 90-6695-037-4

To Jeanine

"...choisir le conseilleur, c'est encore s'engager soi–même... vous êtes libre, choisissez, c'est–à–dire inventez. Aucune morale générale ne peut vous indiquer ce qu'il y a à faire; il n'y a pas de signe dans le monde. Les catholiques répondront: mais il y a des signes. Admettons–le; c'est moi–même en tout cas qui choisis le sens qu'ils ont."

Jean–Paul Sartre, "*L'existentialisme est un humanisme*", pp. 46–47. Paris: Les Éditions Nagel.

ACKNOWLEDGEMENTS

I do not know whether science would make any progress if it were undertaken by loners. What I do know is that its progress is fastened enormously by regular discussions about any new steps that might be taken. In this respect I am most indebted to Jos ten Berge, who has always had a willing ear for any new trials, good or bad, that I suggested, who helped me finding the relevant literature, who checked any new derivations with which I littered his huge blackboard, who painstakingly scrutinized all versions of the manuscript, and who continuously stimulated my research, even while many hours were needed to let me share in his almost professional knowledge of real estate.

The type of data analysis described in this study is typically developed in France and Leyden. I have learned much both from and about the French school during my stay in Montpellier with Yves Escoufier and Christine Lavit. Even more than by the French connection, however, I have benefited by the Leyden connection. Firstly, through Jan de Leeuw, who provided me with a pile of French literature that I still haven't finished, but mostly through Willem Heiser, whose great knowledge of the literature in our field helped me finding a number of overlooked references, and who critically read several versions of the manuscript. Next, I am indebted to John van de Geer, Ivo Molenaar, and Wim Krijnen for pointing out a number of errors in the manuscript, and to the Netherlands organization for scientific research (NWO) for funding this project by means of a PSYCHON–grant (560–267–011). Last but not least, I am most grateful to Jeanine for many other forms of support, but most of all for her showing me that "vous êtes libre, choisissez".

CONTENTS

PART III. ANALYSES OF EMPIRICAL DATA

1. INTRODUCTION

Principal Components Analysis (PCA) is a useful technique for the exploratory analysis of quantitative variables. It yields optimal representations of the variables and of the observation units (denoted as "objects" here) simultaneously in a limited number of dimensions.

For the exploratory analysis of qualitative data it would be desirable to have a similar method for optimally representing variables and objects simultaneously. However, one cannot handle qualitative variables in the same way as quantitative variables, because the "scores" on qualitative variables have no numerical value.

Nevertheless, several techniques have been developed for PCA of data sets in which some or all variables are qualitative. These techniques can be distinguished in two types. In the first type the relation between two qualitative variables or between a qualitative variable and a quantitative variable is expressed by means of a coefficient of association. In order to assess the association between such variables each variable is represented by a so–called "quantification" matrix. Let n be the number of objects. Then such a quantification matrix is an $n \times n$ matrix containing for all pairs of objects (including an object paired with itself) their similarity, based on the variable concerned. For example, the similarity between two objects, based on a qualitative variable, can be said to be 1 if the objects belong to the same category of that variable, and 0 otherwise. Many other definitions of similarity between objects are conceivable, and consequently many different types of quantification matrices can be used. The main idea in the first type of method is that the n^2 elements of each quantification matrix can be seen as scores on a variable, and that hence PCA can be performed on such variables. It can be shown that such a "PCA of quantification matrices" comes down to PCA on a matrix of association coefficients between variables, just as ordinary PCA can be seen as a PCA of the correlation matrix. Hence PCA of a set of such quantification matrices considered as variables analyzes and represents the association coefficients between the complete set of variables. However, it does not yield coordinates for the objects, or the categories of the variables. If one's main interest is in the representation of the variables

and one does not need any information on how the relations between variables are reflected in relations between objects and categories one might be satisfied with this (first) type of method. In practice, however, this limited amount of information is rarely satisfactory.

In contrast to the first type of techniques, the second type of techniques for PCA of qualitative variables does provide a representation for the objects and the categories. The best–known of such techniques is Multiple Correspondence Analysis (MCA), developed independently by several authors, e.g., Guttman (1941), Hayashi (1950), Benzécri et al. (1973), Nishisato (1980) and Gifi (1981), under different names, see Tenenhaus and Young (1985). When each qualitative variable is represented by means of a set of binary indicator variables, indicating for each category whether an object belongs to it (1) or not (0), then MCA can be formulated as PCA of the total set of these indicator variables with respect to some predefined metrics. This implies that MCA in fact performs a PCA on the matrix of (binary) scores of objects on all categories of all variables. Therefore, MCA is directed at optimally representing both the objects and the categories, but not necessarily the variables. As is explained in section 6.1, MCA does optimally represent some aspects of the variables, but does not take into account *all* the information of the variables.

Both techniques discussed above have not only been proposed for PCA of sets of merely qualitative variables. The techniques have been generalized to handle mixtures of qualitative and quantitative variables as well. For PCA of quantification matrices this generalization consists simply of defining quantification matrices for both qualitative and quantitative variables, and performing a PCA of these quantification matrices considered as variables. The generalization of MCA that can be used for the analysis of mixtures of qualitative and quantitative variables, called "PCAMIX" here, comes down to a PCA of the total set of indicator variables for the qualitative variables combined with the quantitative variables, as is explained in more detail in section 7.1.

Above, two types of methods for the exploratory analysis of data sets consisting (partly) of qualitative variables have been discussed. Both are incomplete in that they lack either an optimal representation of the objects (PCA of quantification matrices) or an optimal representation of the variables

(PCAMIX). A desirable property of a method seems to be that it optimally represents relations between the variables in a low–dimensional space while at the same time representing relations between object coordinates and categories.

In the first part of this study, "Three–way methods applied to quantification matrices", methods are proposed that provide a compromise between PCA of quantification matrices and PCAMIX. That is, these compromise methods provide representations of both the objects and the variables. The methods that are developed here involve the application of so–called three–way methods to quantification matrices. Three–way methods are methods for the simultaneous analysis of a number of data sets pertaining to the same entities, for example, a number of similarity matrices giving similarities between a set of objects, in a number of different instances. The idea of applying three–way methods to a set of quantification matrices directly follows what has been done (implicitly) by Saporta (1975, 1976), who first proposed what has been called here PCA of quantification matrices. In fact his method comes down to applying the three–way method STATIS–1 (see section 2.1) to a set of quantification matrices. Likewise, PCAMIX (and hence MCA) can be seen as applying SUMPCA (see section 2.4) to a set of quantification matrices. There are many other three–way methods. In principle these can all be used for the analysis of a set of qualitative data (D'Ambra & Marchetti, 1986; Coppi, 1986). This opens the possibility of generating as many alternative techniques for the analysis of qualitative variables and mixes of qualitative and quantitative variables as there are three–way methods. Some of the three–way methods available are of special interest, because they are related to each other in a very special way. In chapter 2, a number of three–way methods will be discussed and it will be shown that these form a hierarchy. Going down this hierarchy, the methods provide poorer representations of the variables, while the model becomes increasingly simple.

As has been mentioned above, three–way methods can be used to analyze quantification matrices defined for the variables. In chapter 3, the concept of a quantification matrix is explained and it is shown why quantification matrices are useful. In addition, several different choices for quantification matrices for qualitative and quantitative variables are reviewed.

Various three–way methods are available for analyzing a set of

quantification matrices, and many different choices can be made for the quantification matrices. As a consequence, one is faced with a large number of conceivable techniques. In order to facilitate the choice between several techniques, a cross–classification of these is made in chapter 4. Apart from showing which methods are *conceivable* by simply applying any of the three–way methods to quantification matrices also certain *existing* methods are identified as particular cases in the cross–classification. Finally, in section 4.4, some guidelines are provided for choosing among the abundance of available methods.

The second part of this study, "INDOQUAL and INDOMIX", focuses on one of the new methods discussed in part I. This method is INDORT (see section 2.3) applied to one particular combination of quantification matrices for qualitative and quantitative variables. First, the special case with only qualitative variables will be discussed in chapter 5. This method is called "INDOQUAL" (INDscal with Orthonormality constraints applied to quantification matrices for QUALitative variables). INDOQUAL has some interesting properties, that are similar to those of MCA. In addition, this new method can be interpreted in a number of different ways that each clarify certain differences and similarities between this method and MCA. These comparisons are discussed in chapter 6.

In chapter 7 the more general method for the analysis of mixtures of qualitative and quantitative variables, "INDOMIX" (INDscal with Orthonormality constraints applied to quantification matrices for MIXed variables), will be discussed. In chapter 8 INDOMIX is compared to a technique for simple structure rotation of PCAMIX solutions. The latter has been developed for the purpose of comparing PCAMIX and INDOMIX, and is hence described in detail first. Next, it is shown that INDOMIX can be seen as a method that also optimizes simple structure, and in fact does so to a greater extent than the simple structure rotation techniques for PCAMIX do, albeit at the cost of some inertia accounted for. Therefore, INDOMIX is not only interesting as a method for mixtures of qualitative and quantitative variables or merely qualitative variables, but also for the analysis of merely quantitative variables.

In chapter 9 a simple algorithm is provided for INDOQUAL and INDOMIX. This algorithm is a modification of an existing INDORT algorithm, and is much

simpler when the number of objects is large. This algorithm, and some variants of it, only use derived quantities, based on category frequencies, bivariate frequencies of pairs of categories from different variables, category means of quantitative variables, and correlations between quantitative variables. It follows that the method itself depends on these aggregate quantities only. As a consequence, this algorithm allows for the analysis of a number of bivariate contingency tables instead of the original data on the objects as well.

Finally, in the third part, "Analyses of empirical data", experiences with INDOQUAL and INDOMIX are reported in the form of applications to empirical data sets. Most of the INDOQUAL and INDOMIX results are compared with the results given by existing techniques. In this part, also some attention is given to the stability of the solution of INDOQUAL and INDOMIX analyses.

PART I

THREE-WAY METHODS APPLIED TO QUANTIFICATION MATRICES

2. A HIERARCHY OF THREE-WAY METHODS

Many methods have been developed for the analysis of three–way data. In the present chapter, a number of these will be discussed. The methods discussed here form a hierarchy. The methods in the hierarchy are related such that while going down the hierarchy one finds a method that represents the data by a simpler model, albeit at the cost of a poorer representation of the variables. It should be noted that this hierarchy is an extension of a similar hierarchy mentioned by Kroonenberg (1983, pp.49 ff). A different hierarchy of three–way methods was described by Carroll and Wish (1974, pp.92–96). They describe hierarchical relations between IDIOSCAL, PARAFAC2, and INDSCAL.

The three–way methods to be discussed here are all methods that can be applied to quantification matrices. Quantification matrices will be described in detail in chapter 3. For the purpose of the present chapter only some notational aspects of the quantification matrices need to be mentioned. Let m be the number of variables, n be the number of objects, and S_j, $j = 1,...,m$, be the $n \times n$ quantification matrix for variable j. As will be seen in chapter 3, quantification matrices can often be considered as similarity matrices between the objects, hence the choice S_j for the symbol to denote such matrices. It should be noted that the S_j matrices, being similarity matrices, are always symmetric. The data to be handled by the three–way methods described below consist of an $n \times n \times m$ array, that is to say, of a set of m matrices of order $n \times n$. Figure 2.1 depicts such a data array.

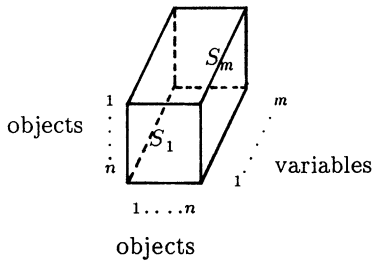


Figure 2.1. *A three–way data array for similarity matrices.*

In the next four sections methods for analyzing such data will be described. In section 2.5 it will be shown how these methods form a hierarchy.

2.1. STATIS

STATIS has been developed by L'Hermier des Plantes (1976) as a method for performing PCA on a set of quantification matrices in three steps. The first step, called STATIS–1 here, consists of performing PCA on the matrices $S_1,...,S_m$ considered as variables. That is, STATIS–1 searches linear combinations, $F_1,...,F_r$, of the matrices $S_1,...,S_m$ that optimally account for the matrices $S_1,...,S_m$. In order to see how the matrices $S_1,...,S_m$ can be considered as variables, let matrix S_j be represented by a vector $\text{Vec}(S_j)$ which contains the elements of S_j strung out row–wise, $j = 1,...,m$. Similarly, let $\text{Vec}(F_1),...,\text{Vec}(F_r)$ be the principal components of the variables $\text{Vec}(S_1),...,\text{Vec}(S_m)$. Then, STATIS–1 can be described as the method that minimizes the loss function

$$\begin{aligned} \text{STATIS–1}(F_1,...,F_r,C) &= \sum_{j=1}^{m} \| S_j - \sum_{l=1}^{r} c_{jl} F_l \|^2 \\ &= \sum_{j=1}^{m} \| \text{Vec}(S_j) - \sum_{l=1}^{r} c_{jl} \text{Vec}(F_l) \|^2 \end{aligned} \tag{1}$$

over arbitrary matrices $F_1,...,F_r$, and the $(m \times r)$ matrix C of loadings c_{jl} of the variables on the components. Collecting the variables $\text{Vec}(S_1),...,\text{Vec}(S_m)$ in the $n^2 \times m$ data matrix S, and the components $\text{Vec}(F_1),...,\text{Vec}(F_r)$ in the $n^2 \times r$ matrix F, the STATIS–1 function can be rewritten as

$$\text{PCA}(F,C) = \| S - FC' \|^2. \tag{2}$$

This description of the loss function for STATIS–1 is simply a description of the loss function for PCA (r components) on a data set S of order $n^2 \times m$, where F contains the component scores for the n^2 object–pairs, and C contains the component loadings for the variables. This method has been mentioned earlier by Tucker (1972, pp.7–8) in developing his three–mode scaling method.

The second step of STATIS consists of defining a compromise matrix as the first principal component (F_1) of the matrices $S_1,...,S_m$. That is, assuming that α_j gives the first principal component weight for matrix S_j, $j = 1,...,m$, then this compromise matrix is given by $F_1 = \sum_j \alpha_j S_j$.

The third step, which is called STATIS–3 here, consists of PCA of the

compromise matrix that has been defined in the second step. It is readily verified that PCA on matrix F_1, and thus STATIS–3, is equivalent to minimizing

$$\text{STATIS–3}(X,\Lambda) = \| \sum_{j=1}^{m} \alpha_j S_j - X\Lambda X' \|^2 \tag{3}$$

over the diagonal matrix Λ, and matrix X $(n \times r)$ subject to $X'X = I_r$. At the minimum of the STATIS–3 function, matrix X will contain the compromise component scores for the objects, and Λ will contain the corresponding eigenvalues.

The strategy of analyzing a number of matrices as if they are variables followed by a detailed analysis of the summarizing matrices has been proposed earlier by Tucker and Messick (1963) in a related context. Tucker and Messick have proposed to analyze the lower triangles of a number of *distance* matrices by means of PCA on these lower triangles strung out as vectors of order $\frac{1}{2}n(n–1)$. Typically, these components are rotated to simple structure. Next, new "distance matrices" that optimally approximate the original distance matrices are computed on the basis of each of these principal components, which are subsequently analyzed by means of classical multidimensional scaling techniques. Although the Tucker and Messick method resembles STATIS in lay–out, it clearly differs from STATIS in several respects.

2.2. TUCKALS–3

Tucker (1966) proposed various models for three–mode principal components analysis. One of these is the model which Kroonenberg and De Leeuw (1980) called the Tucker–3 model. This model represents the entries of each of the three modes by means of a smaller number of components, hence providing three sets of components. These components are related to each other by the so–called core matrix. In case this model is applied to symmetric matrices $S_1,\ldots,S_m$ it can be described as

$$\hat{s}_{ii'j} = \sum_{u=1}^{p} \sum_{v=1}^{p} \sum_{l=1}^{r} x_{iu} x_{i'v} c_{jl} h_{uvl} \ , \tag{4}$$

where $\hat{s}_{ii'j}$ is the model–description of element (i,i') of matrix S_j, x_{iu} denotes the coordinate of object i on component u, c_{jl} the loading of variable

j on component l, and h_{uvl} the element of the core matrix that relates the u^{th} and v^{th} components for the objects to the l^{th} component for the variables, $u, v = 1,...,p$, and $l = 1,...,r$.

Fitting the symmetric version of the Tucker–3 model (4) in the least squares sense (TUCKALS–3) comes down to minimizing the function

$$\text{TUCKALS–3}(X,H_1,...,H_r,C) = \sum_{j=1}^{m} \| S_j - X \sum_{l=1}^{r} c_{jl}H_lX' \|^2, \tag{5}$$

over matrices X $(n \times p)$, $H_1,...,H_r$ $(p \times p)$, and C $(m \times r)$. Kroonenberg (1983) has called this method three–mode scaling. According to Kiers (1989b) this symmetric version of TUCKALS–3 can be described as a constrained variant of PCA on the matrices $S_1,...,S_m$. In order to show this, we rearrange the elements of the matrix between $\|\ \|$ into vectors, and use the fact that $\text{Vec}(XH_lX') = (X \otimes X)\text{Vec}(H_l)$, where $\otimes$ denotes the Kronecker product. Then the loss function for TUCKALS–3 can be rewritten as

$$\begin{aligned}
\text{TUCKALS–3}(X,H_1,...,H_r,C) &= \sum_{j=1}^{m} \| \text{Vec}(S_j) - \text{Vec}(X \sum_{l=1}^{r} c_{jl}H_lX') \|^2 \\
&= \sum_{j=1}^{m} \| \text{Vec}(S_j) - \sum_{l=1}^{r} (X \otimes X)\text{Vec}(H_l)c_{jl} \|^2 \\
&= \| S - (X \otimes X)(\text{Vec } H_1 | ... | \text{Vec } H_r)C' \|^2. \qquad (6)
\end{aligned}$$

Minimizing (6) over arbitrary matrices X $(n \times p)$, $H_1,...,H_r$ $(p \times p)$, and C $(m \times r)$ is equivalent to minimizing the PCA loss function (2) over C and F, subject to the constraint that F $(n^2 \times r)$ can be written as $F = (X \otimes X)(\text{Vec } H_1 | ... | \text{Vec } H_r)$ for certain matrices X and $H_1,...,H_r$ of appropriate orders. Thus, TUCKALS–3 can be seen as constrained PCA, where matrix C yields loadings for the variables.

Whereas STATIS–1 only gives a representation of the variables, TUCKALS–3 also gives a representation of the objects (in X). In addition to these object coordinates TUCKALS–3 provides measures that indicate the interaction–relations between different components for the objects and the variables (given in the matrices $H_1,...,H_r$). The latter relations, however, are difficult to interpret, because they are relations between objects components as "viewed" by each of the components for the variables (indicated

by the subscripts of the matrices $H_1,..,H_r$).

2.3. INDSCAL and INDORT

Carroll and Chang (1970) have proposed INDSCAL for analyzing a set of distance or dissimilarity matrices. Their first step is transforming the distances into similarities by means of the Torgerson (1958) transformation. Next they propose to fit the similarities to the INDSCAL model, given by

$$\hat{s}_{ii'j} = \sum_{l=1}^{r} x_{il} x_{i'l} c_{jl} \ , \tag{7}$$

where x_{il} denotes the object coordinate of object i on component l, and c_{jl} gives the loading of variable j on component l, $l = 1,...,r$. This model is much simpler than the Tucker–3 model. The interpretational difficulties in TUCKALS–3, concerning the matrices $H_1,...,H_r$, are overcome by INDSCAL. The INDSCAL model does not consider relations between object components and variable components. In this model only one set of r components is defined. These components can be interpreted as components for objects and variables simultaneously, which makes interpretation of the results much easier than interpreting the results of a TUCKALS–3 analysis. Interpreting the TUCKALS–3 results is further complicated by the fact that the solutions of TUCKALS–3 have rotational freedom. The INDSCAL model, on the other hand, does not allow for rotation of its components. It provides unique axes.

Fitting the INDSCAL model in the least squares sense comes down to minimizing

$$\text{INDSCAL}(X,W_1,...,W_m) = \sum_{j=1}^{m} \| S_j - XW_jX' \|^2, \tag{8}$$

over an $n \times r$ matrix X of object coordinates and diagonal matrices $W_1,...,W_m$. In order to describe INDSCAL in terms of PCA of the $n^2 \times m$ matrix S, the elements of the matrix between $\| \ \|$ are again rearranged in vectors. Now it is useful to note that $\text{Vec}(XW_jX') = \text{Vec}(\sum_l \mathbf{x}_l w_{jl} \mathbf{x}_l') = \sum_l \text{Vec}(\mathbf{x}_l \mathbf{x}_l') w_{jl}$, where $\mathbf{x}_l$ is the l^{th} column of X, and w_{jl} is the l^{th} diagonal element of W_j. Let $c_{jl} \equiv w_{jl}$, then INDSCAL can be described in terms of PCA on the $n^2 \times m$ data matrix S as minimizing

$$\text{INDSCAL}(X,C) = \sum_{j=1}^{m} \| \text{Vec}(S_j) - \text{Vec}(XW_jX') \|^2$$

$$= \sum_{j=1}^{m} \| \text{Vec}(S_j) - \sum_{l=1}^{r} \text{Vec}(\mathbf{x}_l\mathbf{x}_l')c_{jl} \|^2$$

$$= \| S - (\text{Vec}(\mathbf{x}_1\mathbf{x}_1')|\ldots|\text{Vec}(\mathbf{x}_r\mathbf{x}_r'))C' \|^2. \qquad (9)$$

Minimizing (9) over arbitrary X and C is equivalent to minimizing the PCA loss function (2) over C and F, subject to the constraint that F can be written as $F = (\text{Vec}(\mathbf{x}_1\mathbf{x}_1')|\ldots|\text{Vec}(\mathbf{x}_r\mathbf{x}_r'))$ for some $n \times r$ matrix X.

In INDSCAL the constraints imposed on PCA are stronger than those for TUCKALS–3, provided that the number of components for the objects (p) is larger than or equal to the number of components for the variables (r). This can be seen by verifying that INDSCAL can be considered as TUCKALS–3 with the matrices $H_1,\ldots,H_r$ constrained such that $h_{uvl} = 1$ when $u=v=l$, and 0 otherwise, provided that $p \geq r$, (cf. Kroonenberg, 1983, p.53, where a thus constrained core matrix is called the "three–way analogue of an identity matrix"). Of course, the advantages of the stronger and simpler INDSCAL model are offset by the expected loss of fit of this more heavily constrained version of PCA.

Instead of simply minimizing (9) over arbitrary matrices X, one may minimize (9) subject to the constraint $X'X = I_r$. Kroonenberg (1983, p.118) denotes this method as "orthonormal INDSCAL". Here, it will be denoted by the acronym INDORT. Being a constrained variant of INDSCAL, INDORT is a constrained variant of PCA. This method is of special interest in the present study.

2.4. SUMPCA

Levin (1966) has developed a method for the simultaneous factor analysis of a number of data sets. His method is based on PCA of the sum of a set of matrices, $S_1,\ldots,S_m$. His method is equivalent to one of the stages in Tucker's three mode Principal Components Analysis (Tucker, 1966). As has been shown by Jaffrennou (1978), this in turn is equivalent to one of the stages of Jaffrennou's method for analyzing a three–way array. Finally, Gower (1966) proposed to analyze a dissimilarity matrix by first applying the Torgerson

transformation in order to obtain similarities, and next finding coordinates for the objects by means of PCA on the similarity matrix. The latter similarity matrix is often computed as the sum of a number of similarity matrices expressing the similarities between the same objects in terms of different variables. Then this method, "Principal Coordinates Analysis", can be seen as equivalent to the other three methods mentioned above. Because this method comes down to performing a PCA on the sum of the similarity matrices it is called "SUMPCA" here.

SUMPCA can be described mathematically as minimizing the function

$$\begin{aligned} \text{SUMPCA}(X,\Lambda) &= \| \sum_{j=1}^{m} S_j - X\Lambda X' \|^2 \\ &= m \sum_{j=1}^{m} \| S_j - X(m^{-1}\Lambda)X' \|^2 + \text{constant}, \end{aligned} \tag{10}$$

over X $(n \times r)$, subject to $X'X = I_r$, and over the diagonal matrix Λ.

From the description of SUMPCA as the method that minimizes (10) it is clear that STATIS–3, see (3), is a weighted variant of SUMPCA. When all weights $\alpha_1,\ldots,\alpha_m$ in STATIS–3 are (taken) equal, then SUMPCA and STATIS–3 coincide. Alternatively, STATIS–3 can be seen as the SUMPCA method applied to quantification matrices $\alpha_j S_j$ instead of S_j.

SUMPCA can be described as a constrained variant of PCA as follows. SUMPCA minimizes

$$\text{SUMPCA}^*(X,W) = \sum_{j=1}^{m} \| S_j - XWX' \|^2 \tag{11}$$

over X and W, where $W = m^{-1}\Lambda$, subject to the constraint that X is column–wise orthonormal, and W is diagonal. Defining $c_{jl} \equiv w_l$, hence requiring that c_{jl} be the same for all j, and making the same derivation as for (9), we have

$$\text{SUMPCA}^*(X,W) = \| S - (\text{Vec}(\mathbf{x}_1\mathbf{x}_1') | \ldots | \text{Vec}(\mathbf{x}_r\mathbf{x}_r'))C' \|^2. \tag{12}$$

SUMPCA is a constrained variant of PCA, in that it minimizes (2) over F subject to the constraint that F can be written as $F = (\text{Vec}(\mathbf{x}_1\mathbf{x}_1') | \ldots | \text{Vec}(\mathbf{x}_r\mathbf{x}_r'))$ for a certain column–wise orthonormal matrix X,

and over C subject to the constraint that all rows of C are equal. SUMPCA is a constrained variant of PCA that is even more heavily constrained than INDORT, because of the additional constraint imposed on C. This constraint implies that the components do not weight the variables differentially, contrary to INDORT and INDSCAL.

2.5. Hierarchical relations between three-way methods

As discussed above and in Kiers (1989b), all methods described here are constrained versions of STATIS–1. The methods have been treated in such an order that each method is a constrained version of its predecessor. That is, in reversed order, SUMPCA is a constrained version of INDORT, INDORT is a constrained version of INDSCAL, INDSCAL is a constrained version of TUCKALS–3 (provided that in TUCKALS–3 the number of components for the objects, p, is not smaller than r), and TUCKALS–3 is a constrained version of STATIS–1.

In Table 2.1 an overview is given of the hierarchy formed by these methods. Each method can be seen as a method that fits a model to the data. The data are represented by the $n^2 \times m$ matrix S, the model prediction by $\hat{S}$. In addition, the constraints imposed on the model parameters are given. In Table 2.1 two new symbols are introduced, H for $H \equiv (\text{Vec } H_1 | \ldots | \text{Vec } H_r)$, and $\mathbf{c}$ for the r–vector with elements w_l, $l = 1,\ldots,r$, for describing $C = \mathbf{1c}'$, with $\mathbf{1}$ an m–vector with unit elements.

Table 2.1. *A hierarchy of three – way methods.*

method	model	additional constraints
STATIS–1	$\hat{S} = FC'$	
TUCKALS–3	$\hat{S} = (X \otimes X)HC'$	
INDSCAL	$\hat{S} = (\text{Vec}(\mathbf{x}_1\mathbf{x}_1') \mid \ldots \mid \text{Vec}(\mathbf{x}_r\mathbf{x}_r'))C'$	
INDORT	$\hat{S} = (\text{Vec}(\mathbf{x}_1\mathbf{x}_1') \mid \ldots \mid \text{Vec}(\mathbf{x}_r\mathbf{x}_r'))C'$	$X'X = I_r$
SUMPCA	$\hat{S} = (\text{Vec}(\mathbf{x}_1\mathbf{x}_1') \mid \ldots \mid \text{Vec}(\mathbf{x}_r\mathbf{x}_r'))C'$	$X'X = I_r,\ C = \mathbf{1c}'$

As has been mentioned above, TUCKALS–3 only fits in the hierarchy, when $p \geq r$, because only in that case TUCKALS–3 provides a better fit for $S_1, \ldots, S_m$ than INDSCAL does. TUCKALS–3 models with $p < r$ cannot be located in this hierarchy.

The hierarchy given in Table 2.1 holds for any choice of quantification matrix. In chapter 4 an overview is given of these methods for different choices of quantification matrices for qualitative variables and for mixtures of qualitative and quantitative variables, respectively.

2.6. Suggestions for an eclectic approach to three–way analysis of a set of quantification matrices

Above, it has been shown that a number of well–known three–way methods can be ordered in a hierarchy. The higher the position a method takes in this hierarchy, the better the representation of the variables is. Simultaneously, the higher the position a method takes in the hierarchy, the more parameters are involved in the model, and hence the more complex the model is. The latter statement may not be obvious for the methods that take the highest positions in the hierarchies. That is, it may seem that STATIS–1 uses less parameters than TUCKALS–3. However, STATIS–1 in fact fits the "full" Tucker–3 model, that is, with $p = n$, as is readily verified. Thus, STATIS–1 does fit a model with more parameters than TUCKALS–3.

A larger number of parameters in a model typically makes interpretation of the results more complex. For example, TUCKALS–3 provides coordinates for the objects which are related to the variables in a complicated way by means of an extra set of "core" parameters. The coordinates for the objects provided by INDSCAL are linked in a simpler way to the components for the variables, because every component for the variables refers to exactly one component for the objects. In SUMPCA the components for the variables are trivialized in that all variables have the same loadings on each dimension. SUMPCA fits the simplest model, because the model is in fact no longer a three–way model: it does not give a differential representation of the variables.

In order to perform a PCA of the variables, one might choose from all of the methods above. No general statement as to which method is the *best* can be made. However, the hierarchy described above might be used in order to find

empirically which method is the most *useful* for describing one's data by means of a PCA of the variables. Obviously, the *best* representation of the variables is provided by STATIS–1. However, for the purpose of interpretation of the solution this method is rather poor, because it does not yield any description (coordinates) for the objects that is linked to the principal components for the variables. At the other extreme, SUMPCA only yields a good description of the relations between the variables when these are strong. A useful strategy might be to start analyzing one's data by means of the method at the bottom of the hierarchy, that is by SUMPCA. If SUMPCA yields a sufficient fit and an interpretable solution in a reasonable number of dimensions it gives the simplest possible representation of the data (in that number of dimensions). If this method does not give an adequate solution one may start "climbing" the hierarchy and analyze the data by means of the next method in the hierarchy (with the same number of dimensions). This procedure can be repeated until one finds a solution that adequately represents the variables and the objects, if such a solution is available at all. Of course, decisions about "adequate" representations, or "reasonable" numbers of dimensions will always be based on subjective evaluation to some extent.

3. THE CHOICE OF QUANTIFICATION MATRICES

3.1. Why use quantification matrices?

In the previous chapter a series of methods has been discussed for the analysis of three–way data. In chapter 1 it has been mentioned that these three–way methods can all be applied to a set of quantification matrices. However, it has not yet been explained what a quantification matrix is, nor why it should be used. These questions will be dealt with in the present section. First, the possible types of variables, for which these quantification matrices are to be defined, will be reviewed.

Variables can be distinguished according to their levels of measurement. Although many refinements in such a distinction are possible the following distinction will be made here. Variables are called "qualitative" (or "nominal") if the "scores" of the observation units (objects) on such variables do not contain any numerical information at all. Examples of such variables are a person's nationality, a person's religion, an animal's genus, a vegetable's taste, etc.

Variables are called "ordinal" if the scores of objects on the variables have a predefined ordering, possibly with ties. An object that falls in a higher category can be said to have "more" of the aspect that is measured by the variable. For example, the rank order score after a world championship soccer, or a person's preference listing of a number of food–items can be considered as ordinal variables.

Finally, variables are called "quantitative" (or "numerical" or "interval level") when the scores on the variables have a numerical meaning. That is, scores on quantitative variables do not only indicate the rank order of the objects, but also how much the objects differ in the aspect measured by the variable. A person's length, an object's volume, a tree's number of leaves, etc, are usually considered to be quantitative variables.

The above distinction in three types of variables may seem very strict. In practice, however, it is often not at all clear whether a variable can be considered to be measured at interval level or at ordinal level. Similar ambiguities may arise among qualitative and ordinal variables. Therefore, it

is important to note that the level of measurement of a variable is not a given property, but a property attributed to the variable by the practitioner. It is this person who decides at what level of measurement a variable is considered to be measured. This choice is not only based on what kind of variables one has under study, but also on what aspect of the variables one wants to analyze. It can be useful, for instance, to consider variables as "length" and "weight" as ordinal if one wants to detect nonlinear relations between such variables.

The main purpose of the present study is to describe techniques for the analysis of (mixtures of) variables that are considered qualitative or quantitative variables. As has been explained above, qualitative and quantitative variables are of a very different kind. One cannot compare scores on a qualitative variable with those on a quantitative variable, for instance, because these scores have completely different meanings. This implies that one cannot calculate (ordinary) correlation coefficients between variables of different measurement levels. It would be useful to have a means of comparing such different variables. One way of doing so is by representing each variable (qualitative or quantitative) by means of what is called a "quantification matrix" (Zegers, 1986, p.26). Here, a quantification matrix is a square matrix containing measures of similarity among the objects in terms of the variable at hand. Because these quantification matrices are of the same order and of the same kind (being similarity matrices between objects), one can compare quantification matrices for different variables, regardless whether or not they are considered at the same level of measurement. It will be explained later why such matrices can be seen as similarity matrices. First, however, it will be described for what purposes quantification matrices have been proposed and how they have been used.

The idea of using quantification matrices emerged from the wish to define "correlation coefficients" for variables of a measurement level lower than the interval level. The term "correlation coefficient" is used here in a slightly wider sense than usual. That is, the term is used for any type of association coefficient that can be seen as a scalar product between two normalized sets of scores, which do not necessarily have to be deviation scores. The definition of correlation measures for quantification matrices that represent the variables has been developed independently by different authors. Daniels

(1944) has used square matrices of order n to represent ordinal variables, and showed that for certain choices of these matrices their normalized inner products give Spearman's and Kendall's rank correlation coefficients, respectively.

Another line of research based on quantification matrices finds its origin in the work of Escoufier (1970, 1973). He proposed the notion of correlation coefficients (RV–coefficients) between so–called "operators", which are square matrices containing scalar products between certain sets of (scores) vectors. The correlation between such operators is computed simply as the normalized scalar product between the vectors containing all elements of the operators in some fixed order.

Saporta (1975) used this notion of correlation coefficients for operators (quantification matrices) in order to find correlation coefficients for qualitative variables. A qualitative variable can be considered a set of indicator variables. Each indicator variable indicates whether an object falls in a category (score 1) or not (score 0). These indicator variables are collected in an indicator matrix. For variable j this indicator matrix is denoted as G_j, of order $n \times m_j$, and $D_j \equiv G_j'G_j$, the diagonal matrix of order m_j with category frequencies on the diagonal, where n is the number of objects and m_j is the number of categories of variable j. Saporta (1975) chooses as a quantification matrix for a qualitative variable, the matrix $JG_jD_j^{-1}G_j'J$, where $J = (I-n^{-1}\mathbf{1}\mathbf{1}')$ is the centering operator, and $\mathbf{1}$ is the n–vector with unit elements. The correlation coefficient defined as the correlation between such quantification matrices, that is, $\mathrm{tr}S_j'S_l/(\mathrm{tr}S_j^{\,2})^{1/2}(\mathrm{tr}S_l^{\,2})^{1/2}$ with $S_j = JG_jD_j^{-1}G_j'J$, and $S_l = JG_lD_l^{-1}G_l'J$ for variables j and l, can be seen as a correlation between qualitative variables.

Vegelius (1973) followed the same strategy for defining correlations between variables of low measurement level. He proposed to define correlation coefficients for variables of low measurement level as correlations between quantification matrices, that is, again $\mathrm{tr}S_j'S_l/(\mathrm{tr}S_j^{\,2})^{1/2}(\mathrm{tr}S_l^{\,2})^{1/2}$, and called such correlation coefficients "E–coefficients", because they represent scalar products in Euclidean space. Together with Janson, Vegelius considered various correlation coefficients for (mixtures of) qualitative, ordinal and quantitative variables (e.g., Janson & Vegelius, 1978a, 1978b, 1982). The construction of correlation coefficients on the basis of quantification

matrices has been reviewed by Marcotorchino (1984) and Zegers (1986).

Although the use of quantification matrices was inspired by the wish to define correlation coefficients between variables of low measurement level, in this study quantification matrices are not used for determining correlations between such pairs of variables, but for techniques that simultaneously analyze a set of qualitative and quantitative variables. The idea of simultaneously analyzing a set of quantification matrices has been given by Escoufier (1973) as well as by Vegelius (1973), who both propose to analyze the matrix of correlation coefficients between quantification matrices by means of Principal Components Analysis (PCA). However, if one is not satisfied with an analysis of the variables only, but wants to have a representation of the objects as well, it is necessary to analyze these quantification matrices by means of other methods. This has motivated Cazes, Bonnefous, Baumerder and Pagès (1976) to extend Saporta's method. As is explained in section 5.1, their method does not fully succeed in simultaneously analyzing both the variables and the objects. Later on, D'Ambra and Marchetti (1986), see also Coppi (1986), have suggested to use other approaches to analyze a set of quantification matrices. In chapter 4 methods based on this suggestion are described.

Above, it has been explained why one might use quantification matrices in order to analyze (mixtures of) qualitative and quantitative variables. In the next sections, a number of possible quantification matrices will be discussed, both for qualitative and for quantitative variables.

3.2. Quantification matrices for qualitative variables

The quantification matrices, denoted by S_j, to be proposed in the present section can all be described in terms of the notation given above. A full account of all quantification matrices that have been proposed in the literature is beyond the scope of this study. The following summary of quantification matrices (Table 3.1) is based mainly on the correlation coefficients for nominal variables mentioned by Zegers (1986, pp. 50–53). Some quantification matrices have not been given explicitly in the form as they are described in Table 3.1, but only implicitly by means of the correlation coefficients that are based on them. Therefore, in addition to the

quantification matrix itself, the corresponding correlation coefficients are given, if available.

Table 3.1. *Quantification matrices for a qualitative variable.*

quantification matrix	corresponding correlation coefficient
1. G_jG_j'	
2. $JG_jG_j'J$	T–index (Janson & Vegelius, 1978a)
3. $G_jG_j'-n^{-1}\mathbf{1}\mathbf{1}'$	J–index (Janson & Vegelius, 1978a)
4. $G_jD_j^{-1}G_j'$	
5. $JG_jD_j^{-1}G_j'J$	T^2 coefficient (Tschuprow, 1939)
6. $2G_jG_j'-\mathbf{1}\mathbf{1}'-I$	Gamma coefficient (Hubert, 1977)

Some of these quantifications will now be discussed in more detail. That is, first the simplest quantification matrix, G_jG_j', will be shown to be a similarity matrix. Next, the fourth and fifth quantification matrices will be discussed, because these quantification matrices have often been adopted in practice, and will also be adopted in the second part of this study. The second, third, and sixth quantification matrices are not discussed, but can be interpreted in analogous ways.

3.2.1. The quantification matrix G_jG_j'

The elements of the quantification matrix G_jG_j' are given by

$$s_{ii'j} = \begin{cases} 1 \text{ if objects } i \text{ and } i' \text{ belong to the same category} \\ 0 \text{ if objects } i \text{ and } i' \text{ belong to different categories,} \end{cases} \quad (1)$$

Clearly, $s_{ii'j}$ is a measure of similarity between objects i and i' in terms of variable j. That is, objects in the same category are seen as similar

($s_{ii'j} = 1$) and objects in different categories as dissimilar ($s_{ii'j} = 0$). This (binary) similarity measure is very simple, and does not take into account category frequencies or numbers of categories.

3.2.2. The quantification matrices $G_jD_j^{-1}G_j'$ and $JG_jD_j^{-1}G_j'J$

The quantification matrix $G_jD_j^{-1}G_j'$ is more complicated than the one discussed in the previous section. Let the category to which object i belongs be indicated by g, then

$$s_{ii'j} = \begin{cases} f_g^{-1} \text{ if objects } i \text{ and } i' \text{ belong to the same category} \\ \\ 0 \text{ if objects } i \text{ and } i' \text{ belong to different categories,} \end{cases} \quad (2)$$

where f_g is the g^{th} diagonal element of D_j, and thus the frequency of category g of variable j. Clearly, $s_{ii'j}$ can again be seen as a similarity measure, because its value is higher when objects fall in the same category (and hence are more similar) than when they belong to different categories. In contrast to the previous similarity measure, (1), the similarity between objects that fall in the same category now does depend on the number of objects that fall in this category. The more objects belong to this category, the less similar two objects that fall in this category are considered to be. Hence this measure in a way corrects for chance, because the higher the frequency of a category, the higher the probability that two objects would fall in it if the categories were statistically independent.

Although the similarity measure (2) seems to be attractive, it leads to a quantification matrix which has certain disadvantages. That is, the correlation between two qualitative variables, $\mathrm{tr}\, S_j'S_l/(\mathrm{tr}\, S_j^2)^{1/2}(\mathrm{tr}\, S_l^2)^{1/2}$, with S_j and S_l as defined by (2) is always greater than zero, even when the variables are statistically independent. This follows from the fact that S_j and S_l are positive semi-definite (p.s.d.), hence $\mathrm{tr}\, S_j'S_l \geq 0$ with equality if and only if the column-spaces of S_j and S_l are orthogonal. However, this equality can never be attained, because matrices $G_jD_j^{-1}G_j'$ and $G_lD_l^{-1}G_l'$ both contain the vector $\mathbf{1}$ in its column-space, and hence the column-spaces of S_j and S_l can never be orthogonal, as has been pointed out for instance by

Saporta (1975, p. IV–11). Saporta proposed to remedy this by centering the quantification matrix row– and column–wise. This leads to the quantification matrix $JG_jD_j^{-1}G_j'J$. It can be verified that the correlation between the thus defined quantification matrices is Tschuprow's T^2–coefficient (Tschuprow, 1939), which is a normalized version of the χ^2 measure. It is well–known that $\chi^2 = 0$ when two variables are statistically independent. Hence, the correlation between the thus defined quantification matrices for two statistically independent variables is 0. The elements of the quantification matrix can again be seen as similarities between the objects. They are now given by

$$s_{ii'j} = \begin{cases} f_g^{-1} - n^{-1} \text{ if objects } i \text{ and } i' \text{ belong to the same category} \\ \\ -n^{-1} \text{ if objects } i \text{ and } i' \text{ belong to different categories.} \end{cases} \quad (3)$$

These similarities differ from those of (2) in that they are reduced by n^{-1}. This leads to negative similarities between objects that belong to different categories, and slightly reduced, but always positive, similarities between objects that fall into the same categories.

3.3. Quantification matrices for quantitative variables

For quantitative variables several quantification matrices can be used. Saporta (1976) and Janson and Vegelius (1982) both use the quantification matrix

$$S_j = n^{-1}\mathbf{z}_j\mathbf{z}_j' \quad (4)$$

where $\mathbf{z}_j$ is the vector of standardized scores on variable j. This quantification matrix is closely related to the scores that are ordinarily used in the analysis of quantitative variables. It can be interpreted as a similarity measure by noting that

$$s_{ii'j} = \begin{cases} n^{-1}|z_{ij}|\,|z_{i'j}| \text{ if the scores of } i \text{ and } i' \text{ have the same sign} \\ \\ -n^{-1}|z_{ij}|\,|z_{i'j}| \text{ if the scores of } i \text{ and } i' \text{ have different signs.} \end{cases} \quad (5)$$

That is, two objects that have the same sign are seen as similar to a certain extent, while two objects with different signs are seen as dissimilar to a certain extent. The degree of (dis)similarity depends on the absolute values of the scores. Clearly, this measure of similarity only partly takes into account that objects with almost equal scores have higher similarity than objects with very different scores. The following similarity measure emphasizes this aspect of the similarity between two objects.

Gower (1971) proposed the following measure of similarity between objects with respect to a quantitative variable:

$$s_{ii'j} = 1 - |h_{ij} - h_{i'j}|/\rho_j \tag{6}$$

where h_{ij} is the score of object i on quantitative variable j, and ρ_j is the range of this variable. Clearly, this measure expresses the similarity between objects i and i', that is, the smaller the difference between the scores of the objects the higher the similarity. It is of interest to note that the matrix S_j with elements given by (6) is p.s.d., as Gower (1971) has shown.

Many other quantification matrices might be chosen for quantitative variables, for instance, by taking the outer product–moment of any of the vector quantifications for quantitative variables that are mentioned by Zegers (1986, pp. 34–41), among which the outer product–moment of the vector of raw scores and that of deviation scores.

3.4. Quantification matrices for ordinal variables

Although this study focuses on the analysis of qualitative variables and of mixtures of qualitative and quantitative variables, some attention needs to be paid to the analysis of ordinal variables. Defining a quantification matrix for ordinal variables allows one to analyze ordinal variables together with qualitative and quantitative variables.

As has been mentioned earlier, Daniels (1944) proposed to use certain particular square matrices as a kind of quantification matrices for ordinal variables. He has shown that Spearman's and Kendall's rank correlation coefficients can be formulated as normalized inner products between such quantification matrices. However, the "quantification matrices" he proposed

are skew–symmetric matrices, while the methods described in chapter 2 apply to a set of symmetric matrices only. Therefore, these "quantification matrices" are not discussed here.

A very simple approach to handle ordinal variables is to treat the scores on such variables in the same way as quantitative variables. That is, one can use the quantification matrices that have been mentioned for the quantitative variables above with the ordinal scores considered as numerical scores (see Zegers, 1986, pp. 41–43). On the other hand, it is conceivable that other quantification matrices can be found that better describe the similarities between objects based on an ordinal variable. For instance, in the case of complete rank orders, that is, without ties, one may consider matrices with a simplex structure, that is, matrices with equal diagonal elements, and off–diagonal elements the size of which decreases as their distance from the diagonal increases. A particularly simple special case of such a simplex matrix is a tridiagonal matrix with unit elements on the diagonal and the two "by–diagonals", and zero elements elsewhere. This matrix could be generalized for the case of ties such that the similarity between objects in the same or neighbouring categories is set to 1, and the similarity between objects in more distant categories is set to 0. Obviously, these are only some examples of choices for quantification matrices for ordinal variables.

3.5. Normalization and weighting of quantification matrices

In methods that analyze a set of quantification matrices simultaneously, the quantification matrices may affect the solution differently. That is, some quantification matrices may affect the solution more than others, because they are "measured on a different scale". In order to prevent this one may weight the variables such that each affects the solution to the same extent. This situation parallels that of ordinary PCA, where variables can have different variances. For that reason, they are often normalized to constant sums of squares before a PCA is performed.

In order to normalize quantification matrices one needs to have an expression for how much a variable affects the solution. This can be obtained from the total sum of squares of the elements of the quantification matrix. Hence, in order to normalize a variable one may weight the quantification

matrix by the inverse of the square root of the sum of squares of its elements.

Instead of weighting the variables such that they affect the solution equally one might want to do the opposite. That is, one might want to find a solution that maximally accounts for one variable, and analyzes the other variables only in the second place. This might be achieved by giving a large weight to that one variable and small weights to the others. This procedure has been proposed by Nishisato (1984) as "forced classification", in the MCA context. Yet another weighting procedure has been proposed by Cazes et al. (1976). They proposed to analyze a set of qualitative variables by means of MCA after weighting the quantification matrices $JG_jD_j^{-1}G_j'J$ for the variables by means of the elements of the first eigenvector of the (correlation) matrix of Tschuprow's T^2 coefficients between the variables.

3.6. Conclusion

Above, a number of quantification matrices for qualitative and quantitative variables have been described, and it has been indicated how these can be modified by normalizing or weighting them. These are only some of the quantification matrices that can possibly be used. In the present study no attempt is made to find the best choice for quantification matrices. However, some considerations that can be used for making this choice are discussed in section 4.4. For the purpose of the present chapter it suffices to mention that any quantification matrix may be used to represent a variable, as long as it is a symmetric matrix of order $n \times n$, that gives, in some way, similarities between the objects with respect to the variable concerned. The choice of the quantification matrix should be suitable for the data and research question at hand. This might imply that one has to invent similarity matrices oneself, or adopt ones that have been developed and presented in the abundant literature on similarity measures. On the other hand, one might adopt the most frequently made choices for quantification matrices, as made for instance by Saporta (1976), or adhere to the (technical) advices given by Janson and Vegelius (1982) or Zegers and Ten Berge (1986) for choosing quantification matrices that result in correlation coefficients with certain technical advantages.

4. A REVIEW OF THREE-WAY METHODS FOR THE ANALYSIS OF QUALITATIVE AND QUANTITATIVE TWO-WAY DATA

In the present chapter methods will be reviewed that result from applying the three–way methods that have been discussed in chapter 2 to the quantification matrices that have been discussed in chapter 3. It will be indicated which methods have been discussed elsewhere, and which methods appear to be new.

4.1. Three–way methods applied to quantification matrices for qualitative variables

In the present section a review is given of methods that result from applying the three–way methods discussed in chapter 2 to quantification matrices for qualitative variables (discussed in section 3.2). In Table 4.1 the three–way methods are crossed with the quantification matrices. The quantification matrices have been given without the j–indices for convenience. Division by "RSSQ", that is, by the square root of the sum of squares of the elements of the quantification matrix, indicates the normalized version of a quantification matrix. The cells that pertain to methods that have been discussed in the literature are filled with the names of those methods or their references.

The cross–classification of three–way methods and quantification matrices in Table 4.1 contains many empty cells. This is a consequence of the fact that the application of three–way methods to quantification matrices has not been studied in much detail yet, and as far as it has been studied, mostly the quantification matrix $JG_jD_j^{-1}G_j'J$ has been used. The methods in the first column, STATIS–1, come down to applying PCA to the quantification matrices belonging to the rows concerned. In many of their papers, Janson and Vegelius studied some aspects of these methods. In addition, they made a comparison of PCA of J–indices and Tschuprow's T^2–coefficients by means of an example data set (Janson & Vegelius, 1978b).

As has been described in chapter 2, PCA of quantification matrices for qualitative variables has also been developed in France, following Escoufier

Table 4.1. *A cross-classification of methods for qualitative variables.*

method quant. matrix	STATIS–1	TUCKALS–3	INDSCAL	INDORT	SUMPCA
GG'		Marchetti (1988)			
$\frac{JGG'J}{\text{RSSQ}}$	PCA of T–indices	Marchetti (1988)			
$\frac{GG'-n^{-1}11'}{\text{RSSQ}}$	PCA of J–indices				
$GD^{-1}G'$		Marchetti (1988)		INDOQUAL	MCA
$JGD^{-1}G'J$	PCA of ϕ^2–coef.	Marchetti (1988)		INDOQUAL	MCA
$\frac{JGD^{-1}G'J}{\text{RSSQ}}$	PCA of T^2–coef.	Marchetti (1988)		Kiers (1989c)	
$\frac{2GG'-11'-I}{\text{RSSQ}}$	PCA of H.'s Gamma				

(1970, 1973). In fact, the methods denoted here as PCA of T–indices and PCA of Tschuprow's T^2–coefficients have been proposed independently by Saporta (1975). The PCA of ϕ^2–coefficients (product–moment correlations for 2×2 contingency tables) has been proposed by Escoufier (1980).

Applying TUCKALS–3 to quantification matrices has been considered by Marchetti (1988). He mentions explicitly the quantification matrices G_jG_j' and $G_jD_j^{-1}G_j'$. In addition, he suggests that centering these matrices row– and column–wise, as well as scaling them, might be useful. That is, indirectly, he also suggests using $JG_jG_j'J$ and $JG_jD_j^{-1}G_j'J$, and normalized versions of these.

The third column of Table 4.1 is left empty, because applications of INDSCAL to a set of quantification matrices appear not to have been considered in any detail yet, although D'Ambra and Marchetti (1986) hint at it, and Kiers (1989c) also mentions it as a potentially useful method.

The application of INDORT to a set of quantification matrices has been described by Kiers (1989c), for the normalized versions of the quantification matrices $JG_jD_j^{-1}G_j'J$. He mentions that other quantification matrices might be

chosen as well, but does not treat those in any detail. However, the application of INDORT to the (non–normalized) quantification matrices $JG_jD_j^{-1}G_j'J$ will be discussed in chapter 5, called "INDOQUAL" there. It will be noted that the INDOQUAL solution is closely related to that of INDORT applied to the quantification matrices $G_jD_j^{-1}G_j'$.

The last column contains only MCA. It is well–known that MCA can be seen as applying SUMPCA to a set of quantification matrices $JG_jD_j^{-1}G_j'J$ or $G_jD_j^{-1}G_j'$. The application of SUMPCA to quantification matrices for qualitative variables does not seem to have been considered in the literature. However, in the case of binary variables, SUMPCA has been applied to quantification matrices. Gower (1966) has discussed the application of SUMPCA to quantification matrices G_jG_j' for binary variables, albeit it not explicitly in this form. The methods for correspondence analysis of binary data discussed by Fichet (1986) and Fichet and Gbegan (1986) can be seen as particular variants of Gower's method. They have shown also that these methods are variants of ordinary MCA.

Apart from looking at the columns of Table 4.1 it is interesting to look at the rows of Table 4.1 as well. That is, a row of Table 4.1 in fact describes a hierarchy of methods for the analysis of qualitative variables, as follows from the fact that the three–way methods themselves are related hierarchically (see section 2.5). One of these hierarchies is the one for the quantification matrix $JG_jD_j^{-1}G_j'J$. That is, it can be concluded that MCA is a constrained variant of INDOQUAL (chapter 5), which is in turn a constrained variant of TUCKALS–3 applied to these quantification matrices (Marchetti, 1988). Finally, the latter method is a constrained variant of PCA of the quantification matrices $JG_jD_j^{-1}G_j'J$, that is, Escoufier's method of PCA of ϕ^2–coefficients.

Another interesting hierarchy is the one for the normalized version of the quantification matrix $JG_jD_j^{-1}G_j'J$. According to this hierarchy the method proposed by Kiers (1989c) is a constrained variant of one of the methods worked out by Marchetti (1988), which in turn is a constrained variant of PCA of Tschuprow's T^2–coefficients.

4.2. Three-way methods applied to quantification matrices for mixtures of qualitative and quantitative variables

In case one has a mixture of qualitative and quantitative variables one can again consider the application of the three-way methods from chapter 2 to the quantification matrices from chapter 3. The situation is a little more complicated than in the previous section, because now one has to choose quantification matrices both for the qualitative and the quantitative variables, which yields a large number of possible combinations. In fact one might make a three-way-table crossing quantification matrices for qualitative variables with quantification matrices for quantitative variables and with three-way methods. However, it seems that the quantification matrix $n^{-1}\mathbf{z}_j\mathbf{z}_j'$ is the most prevalent quantification matrix for quantitative variables. For this reason, only the slice of this three-way cross-classification that pertains to the quantification matrix $n^{-1}\mathbf{z}_j\mathbf{z}_j'$ for the quantitative variables, is given. The resulting two-way cross-table is given in Table 4.2. This second cross-classification can easily be related to the cross-classification given in Table 4.1, because the same three-way methods are crossed with the same quantification matrices. One might incorporate Table 4.1 as one slice of the three-way cross-table of quantification matrices for qualitative variables by quantification matrices for quantitative variables by three-way methods. Therefore, in the sequel reference is only made to one cross-classification which comprises methods for mixtures of qualitative and quantitative variables of which methods for merely qualitative variables are special cases.

The first column is again filled with a number of methods that have more or less explicitly been proposed by Janson and Vegelius (e.g., 1978a, 1982). That is, Janson and Vegelius discussed the correlation coefficients that are involved in these analyses, both for correlation between two qualitative variables, and for correlation between a qualitative and a quantitative variable. The methods are denoted here by the names Janson and Vegelius gave to the corresponding indices for correlation between a qualitative and a quantitative variable. An example of what is denoted here as "PCA of CP-indices" has been given by Janson and Vegelius (1982). Saporta (1976) also mentions the possibility of performing a PCA of mixed variables by means of PCA of what are called ZP-coefficients here.

Table 4.2. A cross-classification of methods for qualitative and quantitative variables.

method / quant. matrix	STATIS-1	TUCKALS-3	INDSCAL	INDORT	SUMPCA
GG'					
$\frac{JGG'J}{\text{RSSQ}}$	PCA of SP-indices				
$\frac{GG'-n^{-1}11'}{\text{RSSQ}}$	PCA of CP-indices				
$GD^{-1}G'$				INDOMIX	PCAMIX
$JGD^{-1}G'J$				INDOMIX	PCAMIX
$\frac{JGD^{-1}G'J}{\text{RSSQ}}$	PCA of ZP-coef.			Kiers(1988)	
$\frac{2GG'-11'-I}{\text{RSSQ}}$					

The last column contains a method called PCAMIX in the present study. This method is a straight-forward generalization of PCA and MCA, such that it can handle mixtures of qualitative and quantitative variables. It has been proposed by many different authors independently, with slight variations, under names like Partially Optimal Scaling (Nishisato, 1980, p.103-107), or "Simultaneous treatment of qualitative and quantitative variables in factor analysis" (Escofier, 1979). This method is also contained as an option in PRINCALS (De Leeuw & Van Rijckevorsel, 1980). These methods have not been presented as applications of SUMPCA to quantification matrices, but it is readily verified that they can be written as such. For more details on these methods the reader is referred to chapter 7.

The other cells in the last column of Table 4.2 are left open. However, the application of the Gower (1966) method to similarities based on mixtures of qualitative and quantitative variables is likely to have been considered for many different similarity measures. Gower (1971) proposes his general association coefficient explicitly for the purpose of analyzing mixtures of

qualitative and quantitative data. This similarity measure has been used in a discriminant analysis context by Cuadras (1989). As far as quantitative variables are concerned, both Gower and Cuadras use the quantification matrix defined by (6) in chapter 3. Therefore, these methods do not fit into the cross–classification of Table 4.2, but occur in a different slice of the three–way cross–classification that might be made.

Kiers (1988) has examined the application of INDORT to the normalized version of the quantification matrix $JG_jD_j^{-1}G_j'J$ for qualitative variables, but, in a practical example, he uses the non–normalized version. The latter method will be studied in detail in chapter 7, and is called "INDOMIX" there. Kiers (1988) mentions that other choices of quantification matrices might be made, and also that other methods might be used, like TUCKALS–3 and INDSCAL, but does not work these out.

As in Table 4.1, each row of Table 4.2 describes a hierarchy of three–way methods. That is, PCAMIX can be seen as a constrained variant of INDOMIX, which in turn is a constrained variant of PCA of the corresponding quantification matrices (that is $JG_jD_j^{-1}G_j'J$ for a qualitative variable and $n^{-1}\mathbf{z}_j\mathbf{z}_j'$ for a quantitative variable). The latter method comes down to PCA of a "correlation"–matrix of ϕ^2–coefficients for pairs of qualitative variables, squared product–moment correlations for pairs of quantitative variables, and η^2–coefficients for pairs of one qualitative and one quantitative variable.

4.3. Limitations of the given review

Above, a review has been given of methods for the analysis of mixtures of qualitative and quantitative variables, or sets of merely qualitative variables. It is by no means claimed that this review is exhaustive. The methods discussed here describe a class of methods that can be seen as applications of three–way methods to quantification matrices. One limitation of the review is that not all possible quantification matrices have been mentioned. Another limitation is that other three–way methods exist, that might be used for the analysis of a set of quantification matrices. The fact that only some three–way methods have been mentioned in Tables 4.1 and 4.2 is merely a matter of choice among the most familiar methods. This choice was partly based on the fact that the methods mentioned here could easily be seen

to form a hierarchy. Other hierarchies are available as well, like the hierarchy mentioned by Carroll and Wish (1974).

Apart from the fact that only some three–way methods have been mentioned, a more important limitation of the review given here is that many methods for the analysis of qualitative variables cannot be considered as three–way methods applied to quantification matrices. The cross–classification does not comprise, for instance, the methods proposed by Domenges and Volle (1979), Escofier (1984), Lauro and D'Ambra (1984), Ter Braak (1986), Yanai (1986), Van der Heijden (1987), and Sabatier (1987), which treat the variables asymmetrically. The applications of three–way methods to quantification matrices can treat variables asymmetrically as well, by giving some variables a larger weight than others, as discussed in section 3.5. In the case of MCA, this procedure is equivalent to Nishisato's forced classification method (Nishisato, 1984). It is not clear, however, whether or not such methods yield results comparable to those of the methods discussed above.

Apart from the fact that the above review does not comprise the methods that treat variables asymmetrically, it does not contain all "symmetric" methods that have been developed for qualitative variables either. It does not, for instance, comprise the methods for multivariate analysis of qualitative (or mixed) variables developed by Young, Takane and de Leeuw (1978), Di Ciaccio (1986), Meulman (1986), Van Rijckevorsel (1987), Greenacre (1988) and Van der Burg (1988). Apart from these, methods for the analysis of qualitative variables in the linear structure analysis approach (for instance, Muthén, 1984) do not fit into this review either. The review can be said to be more or less complete, however, in that it seems to cover all methods that are known to form a compromise between the PCA–of–variables approach as proposed in the work of Janson and Vegelius, for instance, and the PCA–of–categories–and–objects approach of PCAMIX. If one wishes to perform a PCA of qualitative or mixed variables, then one has to decide whether one wants to perform PCA of the variables or of the categories and objects, because both of them at the same time is impossible. If one does not want to settle for either of them, one can use one of the compromise methods provided here, in order to achieve both objectives partly. Then it remains to choose the quantification matrix to be used, a choice which has not received much

attention in the literature. In most cases one uses PCAMIX or related techniques for a PCA of mixtures of qualitative and quantitative variables, and MCA for a PCA of sets of qualitative variables. Thereby, one implicitly chooses one's quantification matrix. Although it is well–known that the quantification matrix used in MCA provides MCA with nice properties, it is by no means certain that using this quantification matrix yields the method adapted best to one's data.

4.4. How to choose one's method in practice

Above, a large number of methods for the analysis of qualitative and quantitative variables have been discussed. In addition, the idea of applying three–way methods to quantification matrices offers a practically unlimited number of new methods. There seems to be no ground for preferring one of these methods over all others. Empirical research might yield some comparative information on these methods, although it is not likely that conditions could be determined under which one of the methods is superior to all others. In the absence of empirical evidence the choice of the method is to be made by the practitioner, that is, on an ad hoc basis. In order to make such a choice a number of guidelines can be given that serve to clarify some implications of certain choices. These will be discussed now.

First of all the practitioner has to choose the quantification matrix that is to be used. In section 3.6 some remarks have been made concerning this choice. Having chosen one's quantification matrices, the next question to be answered is whether or not one wants to weight the variables in some way, including normalizing the variables. This question is difficult to answer, especially when one uses mixtures of qualitative and quantitative variables. A useful strategy might be based on the fact that the sum of squares of the elements of a quantification matrix indicates the (main) effect a variable has on the solution. That is, apart from the effect variables have on the solution because they are related more or less strongly (which might be called an interaction effect), there is an effect caused entirely by the size of the elements of the quantification matrix, and this is called the "main effect" here. One strategy one can adopt is to normalize the variables, thus ensuring that the main effects of the variables are equal, as is standard practice in

PCA. On the other hand, one might consider a qualitative variable to be more informative as it contains more categories. Then it seems desirable that the more categories a variable has, the more it affects the solution (in the sense of the main effect described above). Hence, in that case it would be better to use the non–normalized quantification matrices, the sums of squares of which are often directly related to the numbers of categories.

Although the above procedure may work for the analysis of merely qualitative variables, for the analysis of a mixture of qualitative and quantitative variables one cannot simply use this strategy, because one cannot compare a qualitative and a quantitative variable in terms of the numbers of categories they have. That is, when one decides to use non–normalized quantification matrices for the qualitative variables, one still has to decide what weights to attach to the quantitative variables. In one way, quantitative variables seem to be far more informative than qualitative variables, simply because of the fact that they make a finer distinction between the objects. If this interpretation of the informativeness of the variables seems appropriate for the data at hand, one may attach a very high weight to the quantitative variables, for instance, such that it has the same main effect as the qualitative variable with the highest main effect. On the other hand, one may consider a qualitative variable as a variable that has come about by combining several "latent" dimensions. For instance, one might consider a variable like "political preference" to be a combination of dimensions such as "conservative versus liberal" and "denominational versus non–denominational". In that case, it would be better to view a qualitative variable as a variable with more information than a quantitative variable, which therefore is to have more effect on the solution. A simple choice in that case would be to normalize quantification matrices for quantitative variables to unit sums of squares. This corresponds to the effect of a binary variable (that is, a qualitative variable with two categories) when it is quantified by $JG_jD_j^{-1}G_j'J$. This is one of the decision problems which might be alleviated when comparative empirical results on these two strategies are available.

Apart from the choice of a quantification matrix, one has to choose the three–way method to apply to the quantification matrices. If there is no a priori reason for using one and only one of the three–way methods from chapter

2, then a useful strategy might be to use several of them, and decide afterwards which method yields the best interpretable description of the data. In making such decisions, at least two considerations have to be borne in mind. First, it is important to note that STATIS–1 can give the best representation of the variables only at the cost of giving no representation at all of the objects. If one is not interested in the representation of the objects at all, then this might be the best method to use. However, as soon as one wants to have descriptions of both objects and variables, then STATIS–1 is no longer useful. The second consideration is that there is a trade–off in the adequacy of the description of the variables and the complexity of the model that is used. A strategy might be to choose between the different methods by choosing that method that is the lowest in the hierarchy, and hence the simplest, that still gives a reasonable representation of the variables, as has been suggested in section 2.6.

The guidelines given above only partly help someone who is to analyze qualitative or mixed variables by means of PCA. Obviously, many problems of choice between methods in the above review are yet to be investigated. It has been the aim of the present section, however, to indicate what questions can be posed, and how one might answer these.

PART II

INDOQUAL AND INDOMIX

5. INDORT FOR QUALITATIVE VARIABLES (INDOQUAL)

5.1. Introduction

In chapter 4 a review has been given of methods for the analysis of qualitative and quantitative variables, based on applying three–way methods to a set of quantification matrices for the variables. This idea has first been worked out for the particular case of the analysis of qualitative variables by Kiers (1989c). He considered the application of INDORT on quantification matrices $S_j = (m_j-1)^{-1/2}JG_jD_j^{-1}G_j'J$, $j = 1,...,m$. In the present chapter a different choice of quantification matrices is made. That is, the application of INDORT to non–normalized quantification matrices $S_j = JG_jD_j^{-1}G_j'J$ will be discussed. In the sequel this method is called "INDOQUAL" (**IND**scal with **O**rthonormality constraints applied to quantification matrices for **QUAL**itative variables). In the present chapter parts of the results given by Kiers (1989c) are repeated (adapted to the different choice of a quantification matrix), and some additional results on this method are considered.

As has been mentioned in chapter 1, two different types of methods are available for PCA of qualitative variables. The first type of method is PCA of quantification matrices, which comes down to performing PCA of the matrix of correlations between the variables. This method aims at optimally representing the variables, whereas it does not give coordinates for the objects at all. Cazes, Bonnefous, Baumerder and Pagès (1976) extended the method for PCA of qualitative variables proposed by Saporta (1975). Their method does provide object coordinates. However, the object coordinates found in their method are based only on the first principal component of the variables. Hence it can be concluded that PCA of quantification matrices does not adequately find coordinates for the objects.

The second type of method for PCA of qualitative variables is mainly represented by Multiple Correspondence Analysis (MCA). This method can be seen as PCA of the categories and objects, but does not necessarily represent the variables well. If one's main interest is in the representation of qualitative variables and one does not need any information on how the relations between variables are reflected in relations between objects and

categories one might be satisfied with the representation of the variables by the first type of method, PCA of quantification matrices. On the other hand, MCA should be used if one is mainly interested in representing the categories and the objects. Clearly, both methods are incomplete in that they lack either an optimal representation of the objects (PCA of quantification matrices), or an optimal representation of the association between the variables (MCA). In section 4.1 it has been shown that INDOQUAL can be seen as a compromise between PCA of quantification matrices (using ϕ^2–coefficients) and MCA. In the present chapter it will be shown in a different way that INDOQUAL can be seen as such a compromise, and some of its implications will be discussed. In addition, some properties that are well–known for MCA are shown to hold for INDOQUAL too. In a final section it is shown that INDOQUAL has some relations with a method proposed by Saporta (1979). First, however, INDOQUAL will be described.

5.2. INDORT for qualitative variables (INDOQUAL)

In chapter 2 the three–way method INDORT has been described as the method that finds matrices of object coordinates X $(n \times r)$ and diagonal matrices W_j $(r \times r)$ such that the loss function

$$\sigma(X,W_1,...,W_m) = \sum_{j=1}^{m} \| S_j - XW_jX' \|^2 \qquad (1)$$

is minimized over X and $W_1,...,W_m$, subject to $X'X = I_r$. The problem of minimizing (1) over X and $W_1,...,W_m$, subject to $X'X = I_r$, can be simplified as follows. In order to minimize σ over W_j for fixed X, one only has to minimize $\|S_j - XW_jX'\|^2$ over W_j, $j = 1,...,m$. This term can be expanded as

$$\| S_j - XW_jX' \|^2 = \text{tr } S_j^2 + \| X'S_jX - W_j \|^2 - \text{tr } X'S_jXX'S_jX. \qquad (2)$$

The right hand–side of (2) only contains W_j in its second term, which, is minimized by choosing W_j as $\text{Diag}(X'S_jX)$.

Having found the solution for minimizing σ over $W_1,...,W_m$ in terms of X, one can express σ as a function to be minimized over X only, by substituting $W_j = \text{Diag}(X'S_jX)$ for every j. This results in

$$\sigma(X) = \sum_{j=1}^{m} \| S_j - X\,\mathrm{Diag}(X'S_jX)\,X' \|^2$$

$$= \sum_{j=1}^{m} \mathrm{tr}\ S_j^2 - \sum_{j=1}^{m} \mathrm{tr}\ S_jX\,\mathrm{Diag}(X'S_jX)\,X'. \qquad (3)$$

Obviously, minimizing $\sigma(X)$ over X subject to $X'X = I_r$ is equivalent to maximizing

$$\mathrm{f}(X) = \sum_{j=1}^{m} \mathrm{tr}\ X'S_jX\ \mathrm{Diag}(X'S_jX) \qquad (4)$$

over X subject to $X'X = I_r$. It is useful to note that $\mathrm{f}(X)$ can equivalently be expressed as

$$\mathrm{f}(X) = \sum_{j=1}^{m} \mathrm{tr}\left[\mathrm{Diag}(X'S_jX)\right]^2$$

$$= \sum_{j=1}^{m} \sum_{l=1}^{r} (\mathbf{x}_l'S_j\mathbf{x}_l)^2, \qquad (5)$$

where $\mathbf{x}_l$ denotes the l^{th} column of X.

Kiers (1989c) uses the algorithm proposed by Ten Berge, Knol, and Kiers (1988) for maximizing $\mathrm{f}(X)$, subject to $X'X = I_r$. This algorithm becomes problematic when n is large. In chapter 9 a modification of this algorithm is described which does not depend on the size of n.

Kiers (1989c) proposed to use the normalized quantification matrix $S_j = (m_j-1)^{-1/2}JG_jD_j^{-1}G_j'J$. This has been done in order to facilitate comparing INDORT with PCA of Tschuprow's T^2–coefficients (Saporta, 1975). However, using the non–normalized quantification matrices,

$$P_j = JG_jD_j^{-1}G_j'J \qquad (6)$$

does not jeopardize the possibility of comparing INDORT and PCA of quantification matrices. This can be seen as follows. In Table 4.1, PCA of Tschuprow's T^2–coefficients and the method proposed by Kiers (1989c) are in the same row, being applications of three–way methods to the same quantification matrices. Similarly, PCA of ϕ^2–coefficients (Escoufier, 1980) and INDOQUAL are in the same row, implying that hierarchical relations between

these methods exist, just as those between PCA of Tschuprow's T^2–coefficients and the method proposed by Kiers (1989c). In addition, comparing INDOQUAL and MCA is, in fact, easier than comparing MCA and the method proposed by Kiers (1989c), because MCA and INDOQUAL are in the same row of Table 4.1.

In section 4.4 some remarks have been made about whether or not to use normalized quantification matrices. In the specific case of INDORT applied to quantification matrices for qualitative variables, using normalized quantification matrices may yield a solution that is dominated, to a certain extent, by variables with small numbers of categories, particularly, by dichotomous variables. An example and an explanation for this are given in section 10.6. It should be noted, incidentally, that normalizing the quantification matrices has no effect whatsoever on the INDORT solution when the variables have equal numbers of categories.

5.3. INDOQUAL as a compromise between MCA and PCA of φ^2–coefficients

INDOQUAL can be interpreted in a number of different ways. It will be shown here that INDOQUAL can be interpreted as a method that optimally represents the variables (as PCA of quantification matrices does) while retaining a clear link with the representation of the objects. First, it is useful to describe PCA of ϕ^2–coefficients (Escoufier, 1980) in mathematical terms.

As has been remarked in section 4.1, PCA of ϕ^2–coefficients can be seen as PCA of the quantification matrices $P_j \equiv JG_jD_j^{-1}G_j'J$ considered as variables. That is, the matrices P_j can be strung out row–wise into column–vectors $\mathrm{Vec}(P_j)$, and PCA can be performed on the resulting data matrix. It is well–known that PCA maximizes the sum of squares of loadings of the variables on the components. In general, a loading is a signed length of the projection of a variable on a component, expressing the amount of inertia of the variable accounted for by the component. Usually, the loadings in PCA are product–moment correlations. PCA of ϕ^2–coefficients can be seen as a kind of "raw" PCA, that is, it is a PCA of non–normalized variables. In this case the "loading" of variable j on component l is given by $\mathrm{tr}\, F_l'P_j$, where the $(n \times n)$ matrix F_l is obtained from $\mathrm{Vec}(F_l)$, the l^{th} principal component of the vectors $\mathrm{Vec}(P_1),\ldots,\mathrm{Vec}(P_m)$. Hence PCA of ϕ^2–coefficients can be

described as maximizing the function

$$g(F_1,...,F_r) = \sum_{j=1}^{m} \sum_{l=1}^{r} (\text{tr } F_l'P_j)^2, \tag{7}$$

over the $n \times n$ matrices F_l, $l = 1,...,r$, subject to the constraint $\text{Vec}(F_l)'\text{Vec}(F_{l'}) = \text{tr } F_l'F_{l'} = \delta_{ll'}$, where δ denotes the Kronecker symbol.

The function maximized by INDOQUAL is closely related to the function maximized by PCA of ϕ^2–coefficients, as can be seen as follows. INDOQUAL has been shown to maximize (5), which upon substitution of P_j for S_j can be rewritten as

$$f(X) = \sum_{j=1}^{m} \sum_{l=1}^{r} (\mathbf{x}_l'P_j\mathbf{x}_l)^2 = \sum_{j=1}^{m} \sum_{l=1}^{r} (\text{tr } \mathbf{x}_l\mathbf{x}_l'P_j)^2, \tag{8}$$

over matrix X, subject to $X'X = I_r$. Maximizing $g(F_1,...,F_r)$ over $F_1,...,F_r$, subject to the constraint $\text{tr } F_l'F_{l'} = \delta_{ll'}$ and subject to the additional constraint $F_l = \mathbf{x}_l\mathbf{x}_l'$, is equivalent to maximizing $f(X)$ over X, subject to the constraint $\text{tr}(\mathbf{x}_l\mathbf{x}_l'\mathbf{x}_{l'}\mathbf{x}_{l'}') = \delta_{ll'}$. The latter constraint can be reformulated as $\text{tr}(\mathbf{x}_l\mathbf{x}_l'\mathbf{x}_{l'}\mathbf{x}_{l'}') = (\mathbf{x}_l'\mathbf{x}_{l'})^2 = \delta_{ll'}$. This in turn is equivalent to $X'X = I_r$, which shows that, when $F_l = \mathbf{x}_l\mathbf{x}_l'$, the constraints $\text{tr } F_l'F_{l'} = \delta_{ll'}$ and $X'X = I_r$ are equivalent. As a consequence, maximizing $g(F_1,...,F_r)$ over $F_1,...,F_r$, subject to the constraint $\text{tr } F_l'F_{l'} = \delta_{ll'}$, for all pairs l and l', and to the additional constraint $F_l = \mathbf{x}_l\mathbf{x}_l'$ for all l, is equivalent to maximizing (8) over X, subject to $X'X = I_r$. Hence INDOQUAL can be interpreted as PCA of ϕ^2–coefficients subject to the additional constraint $F_l = \mathbf{x}_l\mathbf{x}_l'$.

An advantage of INDOQUAL over PCA of ϕ^2–coefficients is that, whereas in the latter the components that are found cannot immediately be interpreted in terms of the objects, this problem is overcome in INDOQUAL. This is due to the constraint imposed on the "components", $F_1,...,F_r$, for the qualitative variables. The constraint $F_l = \mathbf{x}_l\mathbf{x}_l'$ implies that for every component of the variables there is one vector $\mathbf{x}_l$ of coordinates for the *objects*. Therefore, in INDOQUAL the components solution is directly linked to a solution for object coordinates on the same number of dimensions as one has chosen for the components of the variables. A representation of the categories is immediately supplied by means of the centroids of the object coordinates of the objects

that score in a category.

As far as comparison of INDOQUAL and MCA is concerned, one finds that INDOQUAL yields a representation of the qualitative variables that is better than the one given by MCA, because the MCA solution satisfies the same constraints as are imposed on the components in INDOQUAL, and INDOQUAL yields the best possible representation of the variables, subject to these very constraints. This also follows immediately from the fact that MCA can be seen as SUMPCA on the quantification matrices P_j, and that SUMPCA is a constrained variant of INDORT, as has been explained in chapter 2.

In conclusion, INDOQUAL can be seen as a method that optimally represents relations between the qualitative variables, and simultaneously yields a representation of the objects and the categories that is linked to the representation of the variables. Clearly, INDOQUAL is a compromise between PCA of ϕ^2–coefficients and MCA, in that it performs a PCA of the variables (but subject to additional constraints) on the one hand, and yields coordinates for the objects (like MCA does) on the other hand.

Although INDOQUAL and MCA clearly differ in terms of the objectives they have, the methods do have certain properties in common. One of these is discussed in the next section.

5.4. Trivial solutions

A well–known property of MCA is that, when MCA is seen as finding the object coordinates as the eigenvectors of $\sum_j S_j = \sum_j G_j D_j^{-1} G_j'$, one always finds one so–called trivial axis of object coordinates (e.g., Gifi, 1981, p.94). That is, $\sum_j G_j D_j^{-1} G_j'$ always has an eigenvector $\mathbf{1}$, associated with the largest eigenvalue, regardless of the data. In order to avoid such a solution, one may obtain the object coordinates from the first r eigenvectors of $\sum_j J G_j D_j^{-1} G_j' J$, which are precisely the first r non–trivial eigenvectors of $\sum_j G_j D_j^{-1} G_j'$.

In INDOQUAL the same phenomenon is observed. That is, INDORT applied to the matrices $G_j D_j^{-1} G_j'$ yields one trivial axis, and the same (non–trivial) axes as those of INDORT applied to $J G_j D_j^{-1} G_j' J$. This phenomenon has occurred consistently in practice, but a mathematical proof is not available. It can be proven, however, that, if the trivial vector emerges, the remainder of the r+1 dimensional solution gives an r–dimensional solution for the centered case.

5.5. The interpretation of the results of an INDOQUAL analysis

The results of an INDOQUAL analysis can be interpreted in a way highly similar to that of MCA. The difference in interpretation should be that the different methods stress different aspects. That is, whereas MCA stresses optimal representation of the categories, INDOQUAL stresses optimal representation of the variables.

First of all, we have the diagonal elements of the matrices W_j. These can be interpreted as the loadings of the qualitative variables on the components. Note that these loadings are always nonnegative, due to the fact that $W_j = \text{Diag}\, X'S_jX$, and S_j is positive semi–definite. This might seem to restrict the quality of these loadings, but, considering that relations between qualitative variables cannot sensibly be expressed in terms of negative correlations, the nonnegativity of these loadings merely reflects the inappropriateness of negative correlations for qualitative variables (Janson & Vegelius, 1982).

The loading of variable j on component l is given by $\mathbf{x}_l'S_j\mathbf{x}_l$. This measure is identical to what is called "discrimination measure" by Gifi (1981). It is readily verified that this is in fact the η^2–coefficient between the qualitative variable j and the (quantitative) component l. The maximal value of η^2 is unity. The total inertia of a qualitative variable is given by $\text{tr}\, S_j^2 = (m_j–1)$, hence, when variable j has more than two categories, the total inertia of this variable is larger than 1, and it can never be accounted for completely by means of one component. This might seem to be a disadvantage, but in practice this poses no problems. The fact that a variable can never be accounted for completely by a component can be understood by noting that a variable may in fact incorporate several different relations between categories, which cannot be captured in one dimension.

For the interpretation of a component it is useful to note that the "loading" is one if and only if the component perfectly discriminates the objects in terms of the categories to which they belong. That is, it is one if and only if objects that belong to the same categories have the same scores on that component. Therefore, high loadings of certain variables on a component imply that the component discriminates the objects well in terms of the categories of the variables concerned. In order to see which categories are particularly well discriminated, it is useful to compute the mean scores of

the components in each of the categories of the variables. These are given by the elements of $Y_j = D_j^{-1} G_j' X$, $j = 1,...,m$.

In addition to the W_j, we have the object coordinates matrix X. These coordinates represent the objects in a low-dimensional space, taking into account the main relations between the qualitative variables.

Finally, we have an overall value for evaluating the quality of the solution. To this end, we use the proportion of inertia accounted for of the quantification matrices P_j. This proportion is obtained as the maximal value of f(X) (cf.(8)), divided by the total inertia. The total inertia of matrix P_j is equal to (m_j–1), hence the overall total inertia is equal to ($\sum_j m_j - m$). Therefore, the proportion of inertia accounted for is given by

$$\frac{\sum_{j=1}^{m} \sum_{l=1}^{r} (\mathbf{x}_l' P_j \mathbf{x}_l)^2}{(\sum_j m_j - m)}. \qquad (9)$$

5.6. A relation of INDOQUAL with a method proposed by Saporta

Above, INDOQUAL has been described as an alternative to MCA. This is by no means the only alternative to MCA that has been considered in the literature. One of the alternatives to MCA greatly resembles INDOQUAL, as will be shown here.

Saporta (1979) has described several techniques for analyzing qualitative variables. The objectives of these techniques are to search weights for the variables such that they yield an optimal (weighted) MCA solution, in one or more dimensions. One of his methods consists of maximizing the first eigenvalue of $\sum_j \alpha_j P_j$, subject to $\sum_j \alpha_j^2 = 1$. This is equivalent to maximizing $\mathbf{x}' \sum_j \alpha_j P_j \mathbf{x}$ over α_j and $\mathbf{x}$, subject to $\sum_j \alpha_j^2 = 1$ and $\mathbf{x}'\mathbf{x} = 1$. As he shows, the α_j, $j = 1,...,m$, that maximize this function are given by $\alpha_j = \mathbf{x}' P_j \mathbf{x} / (\sum_j (\mathbf{x}' P_j \mathbf{x})^2)^{1/2}$. As a result, his method comes down to maximizing $\sum_j (\mathbf{x}' P_j \mathbf{x})^2 / (\sum_j (\mathbf{x}' P_j \mathbf{x})^2)^{1/2} = (\sum_j (\mathbf{x}' P_j \mathbf{x})^2)^{1/2}$, subject to $\mathbf{x}'\mathbf{x} = 1$. Clearly, this is equivalent to maximizing $\sum_j (\mathbf{x}' P_j \mathbf{x})^2$, subject to $\mathbf{x}'\mathbf{x} = 1$. Hence this method is equivalent to INDOQUAL, with $r = 1$.

Saporta (1979) has generalized his method to obtain more than one dimension. For the case of more than one dimension he proposed to maximize

$\sum_l \lambda_l(\sum_j \alpha_j P_j) = \sum_l \mathbf{x}_l'(\sum_j \alpha_j P_j)\mathbf{x}_l = \text{tr } X'(\sum_j \alpha_j P_j)X$, where $\lambda_l(.)$ denotes the l^{th} eigenvalue of the matrix between brackets. The solution for α_j is now given by $\alpha_j = \text{tr } X'P_jX/(\sum_j (\text{tr } X'P_jX)^2)^{1/2}$. Then his method comes down to maximizing $\sum_j(\text{tr } X'P_jX)^2/(\sum_j (\text{tr } X'P_jX)^2)^{1/2} = (\sum_j(\text{tr } X'P_jX)^2)^{1/2}$. Clearly, maximizing this function over column–wise orthonormal X for $r > 1$ is not generally equivalent to maximizing the function $\sum_j\sum_l(\mathbf{x}_l'P_j\mathbf{x}_l)^2$, as INDOQUAL does.

The above comparison of INDOQUAL to the Saporta (1979) methods leads to another interpretation of INDOQUAL. As follows from the above, in case $r = 1$, INDOQUAL can be interpreted as the method that finds those weights for the variables that yield the best one–dimensional (weighted) MCA solution. When $r > 1$, a similar interpretation for INDOQUAL can be given. That is, INDOQUAL can be seen as the method that simultaneously finds weights w_{jl} for the variables *and* the components such that each component can be seen as a kind of one–dimensional weighted MCA solution, with the different MCA–solutions constrained such that they are mutually orthogonal. The weights for the variables differ for the different components. When the loadings are close to zero or one, INDOQUAL can be seen as a method that combines several one–dimensional MCA solutions on several subsets of variables. These subsets are defined by the variables that load high on a component. Therefore, INDOQUAL may be a particularly useful alternative to MCA for data with two or more subsets of closely related variables.

5.7. Discussion

In the present chapter, INDOQUAL has been described as a method for PCA of qualitative variables which does not only yield loadings for the variables, but also object coordinates that are linked to the loadings of the variables. It has been shown that MCA is less suitable for such a purpose than INDOQUAL. However, as has been discussed in section 4.4, the choice of a method should not only be based on its adequacy in representing the information in one's data. Another measure for the quality of a method is parsimony of the model. As described in section 2.6, there is a trade–off between parsimony of the model and fit of the model. In general, a useful strategy seems to choose the simplest model that provides a sufficient fit for the quantification matrices.

Apart from the above considerations in choosing between MCA and INDOQUAL, one might see the fact that INDORT finds "unique axes" as a particular advantage of INDOQUAL over MCA. That is, the INDOQUAL solution is determined completely, whereas the MCA solution is determined up to a rotation only. In chapter 8, however, a procedure for rotating the MCA solution so as to maximize similar criteria as maximized by INDOQUAL is discussed. In this way, "unique axes" can be determined for the MCA solution as well.

In section 5.1 it has been remarked that the object coordinates provided by the Cazes et al. method are not very useful, because they are based only on the first principal component of the variables. It is possible, however, to adapt the Cazes et al. method such that it provides object coordinates dimensions for each of the principal components. In a different context D'Alessio (1988) describes such a procedure. He also remarks that this method treats the different aspects of the data asymmetrically. In the present context this implies that the representation of the variables receives most attention, while the representation of the objects is of only secondary importance. In addition, the loadings found by the original PCA are related to the corresponding object coordinates in a less direct way than is the case in INDOQUAL.

6. SOME ADDITIONAL COMPARISONS OF MCA AND INDOQUAL

Above, a variety of methods for the analysis of qualitative variables has been discussed. Among these methods, INDOQUAL has received special attention in chapter 5. As has been mentioned in chapter 5, INDOQUAL uses the same quantification matrices as MCA does (implicitly). In chapter 5 INDOQUAL has been described as a compromise between Escoufier's method for PCA of ϕ^2–coefficients (Escoufier, 1980) and MCA. As far as the interpretation of the methods is concerned, only one difference has been discussed in chapter 5. In the present chapter, some further differences and equivalences in terms of the objectives of the methods will be discussed. That is, a number of different descriptions of the methods will be given, showing that the methods differ from each other in more than one respect.

6.1. A comparison of MCA and INDOQUAL in terms of PCA of qualitative variables

A well–known interpretation of MCA is the following. Each qualitative variable can be considered as a quantitative variable when all categories of the variable receive certain quantitative values. This process of assigning quantitative values to the categories of qualitative variables results in so–called "quantified variables". For the first MCA component, the variables are quantified such that the first eigenvalue of the matrix of correlations between the quantified variables is maximized over possible quantifications (not to be confused with quantification matrices) of the variables. The object scores corresponding to this solution are proportional to the mean of the quantified variables. MCA finds a second component by searching *other* quantifications for the variables that yield a vector of object scores (again proportional to the mean of the quantified variables) which is orthogonal to the first object scores component, and maximally accounts for the inertia of the corresponding quantified variables. Subsequent components are found in a similar way.

Gifi (1981, p.103–104) has shown that the first MCA component in fact yields the one–dimensional PCA solution for the set of optimally quantified variables. Gifi (1981) has also mentioned that subsequent MCA components do

not generally yield the one–dimensional PCA solutions for the sets of variables quantified according to these components. Therefore, it seems that MCA can be interpreted as a method for PCA of the (optimally quantified) variables only as far as the first component is concerned.

The problem of interpreting MCA as a PCA of optimally quantified variables seems to be that in MCA each single variable is represented by a number of different quantitative variables (that are seen as quantified versions of the original qualitative variables). Obviously, this takes into account the diverse information possibly present in a qualitative variable to a certain extent. However, each qualitative variable is represented by the same number (r) of quantitative variables, whereas it is conceivable that for some variables that carry more information more quantitative variables would be necessary, while for other variables representing them by r quantitative variables might be superfluous. This problem is remedied by De Leeuw and Van Rijckevorsel (1988) who propose a variant of MCA that allows for different numbers of quantified variables (called *copies*) to represent different variables. However, for both MCA and the variant of MCA proposed by De Leeuw and Van Rijckevorsel, to each quantification corresponds a different correlation matrix, and hence one cannot see these methods as methods performing just one PCA of quantified variables. In this way it becomes also rather awkward to interpret the so–called "discrimination measures", $\mathbf{x}_l' P_j \mathbf{x}_l$, $j = 1,...,m$, $l = 1,...,r$, as squared loadings of quantified variables on the components as is done in MCA (Gifi, 1981, p. 96–97). As far as only one component is concerned, this interpretation is warranted, but as soon as higher dimensional solutions are considered the "squared component loadings" cannot be compared over different components, because, in fact, different quantitative variables (that is different quantifications of qualitative variables) are involved. Therefore, MCA cannot generally be seen as a method for PCA of (optimally) quantified variables.

A method which is directed more clearly at PCA of qualitative variables (or any type of variables) has been proposed by Young, Takane & De Leeuw (1978), under the name PRINCIPALS, which is also an option of the PRINCALS program (De Leeuw and Van Rijckevorsel, 1980). In this method the qualitative variables are each replaced by one "quantified variable" and this is done in such a way that the resulting PCA solution explains as much variance as possible. Obviously, this method does allow for interpretation in terms of a

PCA of quantified variables, unlike MCA. It has a different problem, however, in that it replaces each qualitative variable by only one quantified variable. In case the qualitative variables at hand can be considered as quantitative variables that have for some reason (like measurement problems) been polytomized, then PRINCIPALS may be a useful method, assuming that it retrieves the correct quantifications for the variables (which is still a matter of doubt, because the only criterion on which the search for quantifications is based pertains to the correlations of a variable with the other variables, thus depending on which other variables are included in the analysis). However, for any variable that contains information which cannot be captured by means of one quantitative variable, the PRINCIPALS approach seems to be inappropriate.

A PCA technique that does take into account all the information present in the variables, and does not loose information by considering only one (or even several) quantification(s) of the qualitative variables is what has been called PCA of quantification matrices. As has been mentioned in chapter 5, however, this type of method can adequately represent only the variables. The objects are not represented. In chapter 5, INDOQUAL has been proposed as an alternative method for PCA of qualitative variables, which optimizes the sum of squared loadings, just as PCA does, but with respect to certain constraints in order to allow for a representation of the objects.

Neither of the methods MCA, PRINCIPALS, and INDOQUAL adequately takes into account *all* a method for PCA of qualitative variables should take into account. INDOQUAL differs essentially from MCA and PRINCIPALS, however, in that it does not first reduce the information of a qualitative variable by means of one or several optimal quantifications. INDOQUAL tries to take into account the variables as a whole, as is done in PCA of quantification matrices. The PCA objective has been relaxed in order to have component scores for the objects that are related to the components for the variables.

6.2. MCA as a method for finding an approximate solution for INDOQUAL

In the previous section a first difference in interpretation of MCA and INDOQUAL has been given, focusing on the idea that the methods should give a PCA representation of the variables. In the present section MCA is considered

as a method for finding an approximation of the solution of INDSCAL on the quantification matrices. This gives another comparison of MCA and INDOQUAL.

As has been described earlier, INDORT is a constrained variant of INDSCAL (Carroll & Chang, 1970). INDSCAL minimizes the loss function

$$\sigma(X,W_1,...,W_m) = \sum_{j=1}^{m} \| S_j - XW_jX' \|^2, \tag{1}$$

over arbitrary X and diagonal matrices $W_1,...,W_m$. That is, INDSCAL fits the INDSCAL model

$$S_j \hat{=} XW_jX', \tag{2}$$

for $j = 1,...,m$, to the data in the least squares sense, where $\hat{=}$ denotes a least squares approximation.

As mentioned by Carroll and Chang (1970), the INDSCAL model has been proposed earlier by Horan (1969). The latter, however, did not consider least squares fitting of this model. Instead, he proposed to find the object coordinates as follows. If the model fits the data perfectly, that is, if $S_j = XW_jX'$, for $j = 1,...,m$, then $\sum_j S_j = \sum_j XW_jX' = X(\sum_j W_j)X'$. It follows that, in that case, the object coordinates matrix is given by a transformation of the matrix containing the eigenvectors of $\sum_j S_j$, because these span the column– and row–spaces of $\sum_j S_j$. Hence the matrix with eigenvectors gives "the location of the points in the normal space" (Horan, 1969, p.144), where "normal space" refers to the space for the object coordinates. Horan realized that this solution is determined up to a linear transformation only. As Carroll and Chang (1970) point out, this undeterminacy is problematic in that it overlooks one of the key features of INDSCAL, that is, uniquely oriented axes. In addition, Carroll and Chang's procedure solves for the weight matrices W_j, whereas Horan ignores the estimation of the W_j matrices. Furthermore, Horan's approach is based on the assumption of perfect fit of the INDSCAL model, which does not hold in practice. Therefore, least squares fitting of the INDSCAL model, as done by Carroll and Chang (1970), is to be preferred.

Horan's approach, however, is of more than only historical interest. If the S_j matrices are the quantification matrices $P_j = JG_jD_j^{-1}G_j'J$, then Horan's approach comes down to MCA. That is, MCA can be seen as a method giving an approximate solution to INDSCAL applied to these same quantification matrices.

Because the MCA object coordinates matrix is column–wise orthonormal, this automatically provides an approximate solution for INDOQUAL (that is, INDORT applied to the P_j matrices) as well. This provides a first comparison between MCA and INDOQUAL. It is of interest to mention here that a similar result was found by Marchetti (1988) in comparing TUCKALS–3 for qualitative variables and Tucker's original approximate solution for the Tucker–3 model (Tucker, 1966).

Horan (1969) has not been the only one who used the eigenvectors of $\sum_j S_j$ as a representation of the INDSCAL object coordinates. Escofier and Pagès (1983) mention that the results of their Analyse Factorielle Multiple (AFM), which is also based on the eigenvectors of $\sum_j S_j$ and comes down to MCA when applied to quantification matrices $S_j = P_j = JG_jD_j^{-1}G_j'J$, can be interpreted as those of an INDSCAL analysis. Later on, Escofier and Pagès (1984) motivate this interpretation by considering their method as an alternative to the least squares fitting of the INDSCAL model. Instead of maximizing the *sum of squares* of certain projections, as INDSCAL does in their point of view, AFM maximizes the *sum* of these projections (nonsquared). That is, let $\mathbf{s}_j = \text{Vec}(S_j)$, where Vec(.) denotes the matrix between brackets strung out row–wise into a column, let w_{jl} be the l^{th} diagonal element of W_j, let $w_{1l},...,w_{ml}$ be collected in the vector $\mathbf{w}_j$, and let $\mathbf{x}_l$ be the l^{th} column of X. Then the INDSCAL model can be written as

$$\begin{aligned} \mathbf{s}_j &\hat{=} \text{Vec}\Big(\sum_{l=1}^{r} w_{jl}\mathbf{x}_l\mathbf{x}_l'\Big) \\ &= (\text{Vec}(\mathbf{x}_1\mathbf{x}_1')\,|\,...\,|\,\text{Vec}(\mathbf{x}_r\mathbf{x}_r'))\mathbf{w}_j. \end{aligned} \tag{3}$$

This model can be recognized as a multiple regression model. Finding $\mathbf{w}_j$ comes down to projecting $\mathbf{s}_j$ on the space spanned by the columns $\text{Vec}(\mathbf{x}_1\mathbf{x}_1'),...,\text{Vec}(\mathbf{x}_r\mathbf{x}_r')$. It can be stated that INDSCAL maximizes over X the sum of squared lengths of the projections of $\mathbf{s}_j$ on the column–space spanned by the vectors $\text{Vec}(\mathbf{x}_1\mathbf{x}_1'),...,\text{Vec}(\mathbf{x}_r\mathbf{x}_r')$.

On the other hand, AFM maximizes over column–wise orthonormal X the sum of the *nonsquared* projection lengths of the vectors $\mathbf{s}_j$ on each of the vectors $\text{Vec}(\mathbf{x}_1\mathbf{x}_1'),...,\text{Vec}(\mathbf{x}_r\mathbf{x}_r')$ *separately*. One may assume, without loss of generality, that the vectors $\mathbf{x}_1,...,\mathbf{x}_r$ are normalized to unit length. Hence the vectors $\text{Vec}(\mathbf{x}_1\mathbf{x}_1'),...,\text{Vec}(\mathbf{x}_r\mathbf{x}_r')$ have unit length also. Then the length

of the projection of $\mathbf{s}_j$ on the vector $\text{Vec}(\mathbf{x}_l\mathbf{x}_l')$ is given by $w_{jl} = \mathbf{s}_j'\text{Vec}(\mathbf{x}_l\mathbf{x}_l') = \mathbf{x}_l'S_j\mathbf{x}_l$. Thus AFM maximizes $\sum_j\sum_l w_{jl} = \mathbf{x}_l'S_j\mathbf{x}_l = \sum_j \text{tr}\, X'S_jX$.

Escofier and Pagès (1984) claim that the difference between AFM and INDSCAL is only a difference between maximizing a sum of squared projection lengths versus maximizing a sum of nonsquared projection lengths. There is an important additional difference between AFM and INDSCAL, however. In INDSCAL the sum of squared lengths of projections on the *subspace spanned by* the vectors $\text{Vec}(\mathbf{x}_1\mathbf{x}_1')$,...,$\text{Vec}(\mathbf{x}_r\mathbf{x}_r')$ is maximized. That is, apart from the difference of maximizing a sum of nonsquared versus squared projection lengths between AFM and INDSCAL, there is also a difference between the projections that are of interest in the different methods. In AFM projections on the *individual* axes are considered, whereas in INDSCAL projections on the *subspace* spanned by these axes are considered. This essential difference between INDSCAL and AFM seems to have been overlooked by Escofier and Pagès (1984). Maximizing either a sum or a sum of squares of lengths of projections on the *individual* axes (as AFM does) does not seem very interesting. When more than one axis is taken, the projection of a vector ($\mathbf{s}_j$) on the subspace as a whole should be considered for the approximation of the vector $\mathbf{s}_j$ by a vector in a subspace.

It is of interest to note that the orthonormally constrained variant of INDSCAL, IND***ORT*** can be interpreted in *both* of the above senses, that is, for INDORT the sum of squared lengths of projections on the individual axes is equal to the sum of squared lengths of the projections on the subspace spanned by the axes, due to the orthonormality of $\mathbf{x}_1$,...,$\mathbf{x}_r$, and hence of $\text{Vec}(\mathbf{x}_1\mathbf{x}_1')$,...,$\text{Vec}(\mathbf{x}_r\mathbf{x}_r')$. INDORT can hence be considered both as the method that maximizes the sum of squared lengths of projections on the *subspace*, and as the method that maximizes the sum of squared lengths of projections on the *individual axes*. It should be noted, however, that the orthonormality of the axes in AFM does not likewise guarantee that the sum of *nonsquared* lengths of projections on the individual axes is equal to the sum of *nonsquared* lengths of projections on the subspace spanned by these axes.

It can be concluded that Horan (1969) and Escofier and Pagès (1984) both propose to use the MCA coordinates as an approximate solution for the object coordinates for INDOQUAL, but that they have different motivations. In spite of their efforts to show the similarity in objectives of MCA and INDOQUAL, their papers served in fact to highlight the differences between the methods.

Nevertheless, Horan's observation that, when the INDSCAL model fits the data perfectly, the object coordinates matrix is a transformation of the matrix of eigenvectors of $\sum_j S_j$ is of considerable importance, and its implications will be discussed in the next section.

6.3. Equivalence of MCA and INDOQUAL when the INDORT model fits the quantification matrices perfectly

As has been mentioned in the previous section, when the INDSCAL model perfectly fits the matrices S_j, $j = 1,...,m$, one has

$$\sum_{j=1}^{m} S_j = X(\sum_{j=1}^{m} W_j)X'. \tag{4}$$

Now suppose that X in (4) is column–wise orthonormal. Then the matrices S_j are not only fitted perfectly by the INDSCAL model, but also by the INDORT model. When an eigendecomposition of $\sum_j S_j$ is given by $\sum_j S_j = K\Lambda K'$, then it follows from (4) that, assuming that the elements of $\sum_j W_j$ are different, $X = K$, and $\sum_j W_j = \Lambda$, up to a permutation. Hence, when the quantification matrices $P_j = JG_jD_j^{-1}G_j'J$ are fitted perfectly by the INDORT model (in r dimensions), then the object coordinates given by INDOQUAL are the same as those found by MCA (in r dimensions). It should be noted that the MCA object coordinates do not even differ a rotation from the INDOQUAL object coordinates, hence the MCA components themselves (and not a rotation of them) give the unique INDOQUAL components.

Of course, the above result in itself is useless in practice, because the INDORT model will, typically, not fit the quantification matrices in a reasonable number of dimensions. In fact, the INDORT model does not necessarily provide a perfect fit of the quantification matrices in the maximal number of dimensions. However, when the data do not fit the INDORT model perfectly, but merely to a great extent, the INDOQUAL solution and the MCA solution are still likely to be similar.

6.4. A comparison of MCA and INDOQUAL in terms of χ^2-distances

The last comparison to be made here between INDOQUAL and MCA is based

on the following useful, though not prevalent interpretation of MCA, in terms of χ^2–distances (Benzécri et al., 1973, see Gifi, 1981, p. 134). That is, Benzécri et al. (1973) describe Correspondence Analysis (CA) as the technique that yields an approximation of so–called χ^2–distances between rows (or columns, which will be disregarded here) of a contingency table. Let the h^{th} and h'^{th} rows of a contingency table of order $r \times c$ be given by $\mathbf{r}_h$ and $\mathbf{r}_{h'}$, with elements r_{hg} and $r_{h'g}$, respectively, $g = 1,...,c$. Let the row–marginals of the contingency table be given by r_h, $h = 1,...,r$, and let the column–marginals be given by f_g, $g = 1,...,c$. Then, the χ^2–distance between two rows of a contingency table can be defined by

$$d_{\chi^2}(\mathbf{r}_h,\mathbf{r}_{h'}) = \sum_{g=1}^{c} \frac{(r_{hg}/r_h - r_{h'g}/r_{h'})^2}{f_g}, \qquad (5)$$

(e.g., Gifi, p. 134), where the constant grand total n has been omitted from the usual definition. Heiser and Meulman (1983) have shown that CA approximates these χ^2–distances by means of a weighted Principal Coordinates Analysis (Gower, 1966). It is well–known (e.g., Lebart, Morineau & Tabard, 1977) that MCA can be seen as correspondence analysis of the superindicator matrix $G = (G_1|...|G_m)$. As has been shown by Meulman (1986, p. 87), MCA can be seen as a (nonweighted) Principal Coordinates Analysis on the χ^2–distances (which are equivalent to the Mahalanobis distances she mentions) defined between the rows of the super–indicator matrix. This Principal Coordinates Analysis first applies the Torgerson transformation to the matrix of distances (in the present case χ^2 distances), yielding $\sum_j JG_jD_j^{-1}G_j'J$, see Meulman (1986, p.87), where the constant m^{-1} has been dropped for convenience, and subsequently finds the best low–rank approximation of this matrix by minimizing

$$\sigma_1(X) = \| \sum_{j=1}^{m} JG_jD_j^{-1}G_j'J - XX' \|^2, \qquad (6)$$

where X is an arbitrary matrix of order $n \times r$. It is well–known (Eckart & Young, 1936) that this minimization problem is solved by finding the eigendecomposition of $\sum_j JG_jD_j^{-1}G_j'J = K\Lambda K'$, and choosing $X = K_r\Lambda_r^{1/2}$, where the subscript r indicates that only the first r eigenvalues and eigenvectors are taken. Clearly, the problem of minimizing (6) over arbitrary X is

equivalent to that of minimizing

$$\sigma_2(X,W) = \| \sum_{j=1}^{m} JG_jD_j^{-1}G_j'J - XWX' \|^2, \tag{7}$$

over diagonal matrices W, and column–wise orthonormal matrices X. This, in turn can be shown to be equivalent to minimizing

$$\sigma_3(X,W^*) = \sum_{j=1}^{m} \| JG_jD_j^{-1}G_j'J - XW^*X' \|^2, \tag{8}$$

over diagonal W^* and column–wise orthonormal X. Minimizing (8) can be interpreted in a similar way as minimizing (6) has been interpreted. That is, let the χ^2–distance based on variable j between objects i and i' be given by

$$d_j^2(i,i') = \sum_{g=1}^{m_j} \frac{(h_{igj} - h_{i'gj})^2}{f_g}, \tag{9}$$

where h_{igj} denotes the "score" of object i on the indicator variable for category g of variable j, and f_g is the frequency of category g of variable j. Then MCA can be seen as the method that performs a "simultaneous" Principal Coordinates Analysis on the χ^2–distances defined between the objects, for each of the variables. This is done by requiring that the object coordinates to be found be the same for all variables.

Now INDOQUAL can be seen as the method with almost the same objective. That is, INDOQUAL minimizes

$$\sigma(X,W_1,\ldots,W_m) = \sum_{j=1}^{m} \| JG_jD_j^{-1}G_j'J - XW_jX' \|^2, \tag{10}$$

over diagonal matrices $W_1,\ldots,W_m$, and over X subject to $X'X = I_r$. Both MCA and INDOQUAL can be seen as methods for simultaneous Principal Coordinates Analysis of the χ^2–distances between the objects, defined for all variables separately. They differ in the way they handle the simultaneousness of the two analyses. That is, in MCA the object coordinates are required to be the same for all variables, whereas in INDOQUAL the object coordinates may differ in column–scaling of the coordinates. This difference implies that INDOQUAL has a greater freedom in fitting the χ^2–distances, and hence will better approximate these distances for all variables jointly. This implies that

INDOQUAL may need fewer components to represent the distances between the objects than MCA does.

6.5. Discussion

In the previous sections some comparisons have been made between MCA and INDOQUAL. It has been emphasized which are the advantages of INDOQUAL over MCA. Obviously, MCA also has certain advantages over INDOQUAL. One of these is that the MCA solution is nested, whereas the solution of INDOQUAL is not. That is, the one–dimensional solution of INDOQUAL is not necessarily contained in the two–dimensional solution of INDOQUAL, the two–dimensional solution is not contained in the three–dimensional solution, etc. As a consequence, one cannot simply choose between an r and an r+1–dimensional INDOQUAL solution by looking at the extra dimension in the latter.

It may be seen as another advantage of MCA that it can be interpreted as an ordinary PCA of the indicator variables, that is, it gives optimal representations of the objects and the categories. The present study, however, focuses on methods that optimally represent the objects, the categories, and the variables simultaneously. Obviously, data analysis of qualitative variables may sometimes focus on the categories (and then MCA or variants of MCA seem most useful), while at other times it may focus on the variables, or on the objects, in which cases a PCA representation can be given by any of the methods described in the present study, and in particular by INDOQUAL.

Meulman (1986) has described MCA as a method for representing χ^2–distances between objects, and has offered an alternative to MCA which is better adjusted to this objective of representing distances. That is, her method aims at finding a representation of the objects in which the distances between the objects are approximated in the least squares sense. In a similar way, alternatives to INDOQUAL might be constructed. That is, one may use a method for least squares fitting of the distances given according to the INDSCAL model to the χ^2–distances between the objects, as defined by each of the variables. In doing so, however, one has to part with the objective of optimally representing the variables.

7. INDORT FOR A MIXTURE OF QUALITATIVE AND QUANTITATIVE VARIABLES (INDOMIX)

7.1. Introduction

In chapter 5 it has been mentioned that PCA of qualitative variables can be distinguished in two types, one focusing on the categories and the objects, the other on the variables. It has also been shown that INDOQUAL is a compromise between these two approaches in that it yields the best possible representation of the variables which also yields a representation of the objects. Ordinary PCA can be applied only to quantitative variables, and INDOQUAL can be applied only to data sets that consist exclusively of qualitative variables. For the exploratory analysis of mixtures of qualitative and quantitative variables a different approach is needed. Kiers (1988) has discussed some existing methods for such data and proposed a new method as well. The present chapter is based for a large part on this paper.

The exploratory analysis of a mixture of qualitative and quantitative variables seems to have received far less attention in the literature than the exploratory analysis of qualitative variables. Here three types of methods can be distinguished. The first type was proposed by Young, Takane and De Leeuw (1978), see also Tenenhaus (1977). Their method, PRINCIPALS, has already been mentioned in section 6.1 for the case where only qualitative variables are involved. In the case of a mixture of qualitative and quantitative variables each qualitative variable is "optimally quantified" by means of one quantitative variable, and an ordinary PCA is performed on the complete set of quantified qualitative variables and variables that were quantitative already. As has been explained in section 6.1, this method may work well when the qualitative variables can be fully captured by means of one quantified variable, for instance when the qualitative variables can be seen as "polytomized" quantitative variables. However, if this is not the case, much information is lost when each qualitative variable is replaced by only one quantitative variable. Because our attention goes to methods that take into account the possibility of more–dimensional information in qualitative variables, this type of methods is ignored in the present study.

The second type of methods generalizes Multiple Correspondence Analysis

(MCA) to the effect that it can handle mixtures of qualitative and quantitative variables. These generalizations of MCA have been proposed independently by many authors (De Leeuw, 1973; Escofier, 1979; Nishisato, 1980, pp.103–107; De Leeuw & Van Rijckevorsel, 1980). Although the methods slightly differ in the way in which quantitative variables are transformed, all methods essentially use the same approach to handle qualitative variables. That is, let G_j be the indicator matrix for variable j, when variable j is a qualitative variable, and let $\mathbf{h}_j$ be the vector of scores on variable j, when variable j is quantitative. Then all methods mentioned above can be described as PCA of the supermatrix containing the columns of the matrices $JG_jD_j^{-1/2}$ for qualitative variables and (transformations of) the vectors $\mathbf{h}_j$ for quantitative variables. The particular method that performs PCA of the supermatrix containing the columns of the matrices $JG_jD_j^{-1/2}$ for qualitative variables and the vectors of standard scores, $\mathbf{z}_j$, divided by $n^{1/2}$, for the quantitative variables will be denoted here as PCAMIX.

As has been mentioned in chapter 5, the very fact that MCA performs a PCA of the complete set of indicator variables for all qualitative variables causes it to yield a non–optimal representation of the qualitative variables. In fact, MCA yields an optimal representation of the *categories* of the qualitative variables, not of the variables themselves. Analogously, the generalizations of MCA for analyzing mixtures optimally represent only the categories of the qualitative variables, rather than the qualitative variables themselves, because they use the same approach for the qualitative variables as MCA does.

The third type of methods is PCA of quantification matrices for mixtures of variables. These methods have been discussed in section 4.2. Just as PCA of quantification matrices for qualitative variables, PCA of quantification matrices for mixtures does not provide a representation of the objects. On the other hand, as has been shown above, PCAMIX and its variants do not provide an optimal representation of the variables. In the present chapter it is shown that INDORT applied to certain quantification matrices yields a compromise between these two types of methods. Apart from this it is shown that this method has some interesting properties in certain special cases.

7.2. INDORT for the analysis of a mixture of qualitative and quantitative variables (INDOMIX)

In chapter 3 it has been mentioned that various quantifications can be chosen for qualitative and quantitative variables. In the present chapter the same quantification matrices are chosen for qualitative variables as in chapter 5. That is, if the j^{th} variable is a qualitative variable the quantification matrix chosen here is given by

$$P_j = JG_jD_j^{-1}G_j'J. \tag{1}$$

If the j^{th} variable is a quantitative variable the quantification matrix chosen here is given by

$$Q_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'. \tag{2}$$

It should be noted that these quantification matrices differ from those chosen by Saporta (1976) in his method for PCA of quantification matrices only in that Saporta (1976) uses the normalized version of P_j. In the sequel, INDORT applied to quantification matrices S_j chosen as P_j or Q_j, $j = 1,...,m$, will be denoted as "INDOMIX" (**IND**scal with **O**rthonormality constraints applied to quantification matrices for **MIX**ed variables).

7.3. INDOMIX as a compromise between PCA of η^2-coefficients and PCAMIX

INDOMIX can be interpreted in a number of different ways. It will be shown here that it optimally represents the variables (as a PCA technique does) while retaining a clear link with the representation of the objects. More precisely, it will be shown that INDOQUAL is a compromise between one of the methods for PCA of quantification matrices for mixed variables, "PCA of η^2-coefficients", and PCAMIX. First, PCA of η^2-coefficients will be described.

The PCA of the quantification matrices S_j taken as P_j or Q_j can be considered as PCA of a certain "correlation matrix". For a pair of qualitative variables the "correlation" is defined as the ϕ^2-coefficient, for a mixed pair the η^2-coefficient is used, and for a pair of two quantitative variables the

squared product-moment correlation is used. This method is called PCA of η^2-coefficients, named after the correlation-coefficient used for a pair of mixed variables.

PCA of η^2-coefficients maximizes the function

$$g(F_1,...,F_r) = \sum_{j=1}^{m} \sum_{l=1}^{r} (\text{tr } F_l'S_j)^2, \tag{3}$$

over the $n \times n$ matrices F_l, $l = 1,...,r$, representing "components" of the variables, subject to the constraint $\text{tr } F_l'F_{l'} = \delta_{ll'}$, where δ denotes the Kronecker symbol. As in section 5.3, $\text{tr } F_l'S_j$ can be considered as the loading of variable j on component l, and hence PCA of η^2-coefficients can be seen as the method that maximizes a sum of squared loadings.

INDOMIX is the method that maximizes (cf. chapter 5, formula (8))

$$f(X) = \sum_{j=1}^{m} \sum_{l=1}^{r} (\mathbf{x}_l'S_j\mathbf{x}_l)^2 = \sum_{j=1}^{m} \sum_{l=1}^{r} (\text{tr } \mathbf{x}_l\mathbf{x}_l'S_j)^2, \tag{4}$$

over X, subject to $X'X = I_r$, with S_j chosen as P_j or Q_j. As in section 5.3, maximizing $g(F_1,...,F_r)$ over $F_1,...,F_r$, subject to the constraint $\text{tr } F_l'F_{l'} = \delta_{ll'}$, for all pairs l and l', and to the additional constraint that $F_l = \mathbf{x}_l\mathbf{x}_l'$ is equivalent to maximizing (4) over X, subject to $X'X = I_r$. Hence INDOMIX can be interpreted as PCA of η^2-coefficients subject to the additional constraint that $F_l = \mathbf{x}_l\mathbf{x}_l'$.

As has been mentioned above, PCA of η^2-coefficients does not provide coordinates for the objects. An advantage of INDOMIX over PCA of η^2-coefficients is that it does yield coordinates for the objects. As has been shown in section 5.3, each component for the variables is directly and uniquely linked to a component for the objects.

It is well-known that the PCAMIX solution yields object coordinates for which $X'X = I_r$. That is, the PCAMIX solution satisfies the constraints imposed on the components in INDOMIX. INDOMIX yields the best possible representation of the variables, subject to these constraints. Therefore, INDOMIX yields a representation of the variables that is *better* than the one given by PCAMIX. This follows also from the fact that PCAMIX can be seen as a constrained variant of INDORT, because PCAMIX can be seen as applying SUMPCA to the quantification matrices S_j, which is a constrained variant of INDORT, as

described in chapter 2.

It can be concluded that INDOMIX is a method that optimally represents relations among mixtures of variables, and also yields a representation of the objects. Clearly, in this way, INDOMIX is a compromise between PCA of η^2–coefficients and PCAMIX, in that it consists of a (constrained) PCA of the variables and simultaneously yields coordinates for the objects (like PCAMIX does). It should be noted that this interpretation also follows from the hierarchical relations discussed in section 4.2.

7.4. The interpretation of the results of an INDOMIX analysis

Because INDOMIX is a compromise between PCA of η^2–coefficients and PCAMIX, its results partly parallel those of PCA of η^2–coefficients and partly parallel those of PCAMIX. That is, like in PCAMIX, INDOMIX provides object coordinates, collected in matrix X. These can be interpreted in the same way as in PCAMIX, but PCAMIX and INDOMIX emphasize different aspects. That is, whereas PCAMIX emphasizes optimal representation of the objects and the categories, INDOMIX aims at optimal representation of the objects and the variables. As a consequence, PCAMIX does not provide an optimal representation of the variables, and INDOMIX does not provide an optimal representation of the categories. Nevertheless, it is possible to provide category coordinates for the INDOMIX solution, by computing $Y_j = D_j^{-1} G_j' X$, the matrix of centroids of the object coordinates of the objects that fall in the category concerned, for every category of a qualitative variable.

The results of INDOMIX share with the solution of PCA of η^2–coefficients that a representation of the variables is given. This representation is provided by the diagonal matrices $W_j = \text{Diag}(X'S_jX)$. The elements of these matrices can be interpreted as the loadings of the variables on the components. As far as qualitative variables are concerned, these loadings can again (as in chapter 5) be seen as η^2–coefficients, each with a maximum of 1. For a quantitative variable, the loadings on the components are squared product–moment correlations between the variable and the components concerned. In addition to these squared correlations it is useful to inspect the nonsquared correlations, and their signs.

Finally, we have an overall value for evaluating the quality of the solution. To this end, we use the proportion of inertia accounted for of the

quantification matrices S_j. This proportion is given by the maximal value of f(X) (cf.(4)), divided by the total inertia. The total inertia of matrix S_j is equal to $\mathrm{tr}\, S_j^2$, which is (m_j–1) for a qualitative variable and 1 for a quantitative variable. Let m_a be the number of qualitative variables, and m_b the number of quantitative variables, then the overall total inertia is equal to ($\sum_j m_j - m_a + m_b$), where $\sum_j m_j$ is the number of categories of the qualitative variables. The proportion of inertia accounted for by the INDOMIX solution (IAF_I) is given by

$$\mathrm{IAF}_I = \frac{\sum_{j=1}^{m} \sum_{l=1}^{r} (\mathbf{x}_l' S_j \mathbf{x}_l)^2}{(\sum_j m_j - m_a + m_b)}. \tag{5}$$

In order to provide an indication of the quality of the INDOMIX solution, it is useful to compare this measure to the inertia of the quantification matrices that is accounted for by means of PCA of η^2–coefficients and PCAMIX, respectively. For PCA of η^2–coefficients, as in ordinary PCA, the proportion of inertia accounted for of the quantification matrices is given by the sum of the first r eigenvalues of the "correlation" matrix, divided by ($\sum_j m_j - m_a + m_b$).

If the object coordinates are computed by means of PCAMIX, one might compute the proportion of inertia accounted for by means of (5) with the PCAMIX object coordinates substituted for the INDOMIX object coordinates. It should be noted, however, that the thus computed "proportion of inertia accounted for" is the inertia accounted for by the PCAMIX object coordinates when the quantification matrices are approximated by the INDORT model. Another interesting measure for the quality of the PCAMIX solution would be a measure that is based on the model for quantification matrices that is actually fitted by PCAMIX. In section 4.2, PCAMIX has been described as the method that applies SUMPCA to the quantification matrices S_j. That is, PCAMIX fits the quantification matrices to the model

$$\hat{S}_j = XWX', \tag{6}$$

for $j = 1,\ldots,m$, where W is a diagonal matrix, and X is the orthonormal matrix of object coordinates. It has been shown in chapter 2 that this model

is a constrained variant of the INDORT model, with $W_j = W$, for all j. This model is interesting in itself, because, when it adequately represents the quantification matrices, it implies that all variables can be represented by the same coordinates in the variable space. Because for the PCAMIX solution we have $\sum_j \|S_j\|^2 = \sum_j \|\hat{S}_j - S_j\|^2 + \sum_j \|\hat{S}_j\|^2$, the inertia accounted for by the SUMPCA model (6), IAF_S, is expressed by

$$\text{IAF}_S = \frac{\sum_{j=1}^{m} \text{tr}\, \hat{S}_j^2}{\sum_{j=1}^{m} \text{tr}\, S_j^2}. \tag{7}$$

The denominator of (7) is given by $\sum_j \text{tr}\, S_j^2 = (\sum_j m_j - m_a + m_b)$. The numerator can be computed as follows. Obviously, $\text{tr}\, \hat{S}_j^2 = \text{tr}\, XWX'XWX' = \text{tr}\, W^2$. From section 2.4 it readily follows that W is given by $m^{-1}\Lambda_r$, where Λ_r contains the first r eigenvalues of $\sum_j S_j$. Hence $\text{tr}\, \hat{S}_j^2 = m^{-2}\sum_l \lambda_l^2$, where λ_l is the l^{th} eigenvalue of $\sum_j S_j$. Therefore,

$$\text{IAF}_S = \frac{\sum_{j=1}^{m} \sum_{l=1}^{r} \lambda_l^2}{m^2(\sum_j m_j - m_a + m_b)} = \frac{\sum_{l=1}^{r} \lambda_l^2}{m(\sum_j m_j - m_a + m_b)}. \tag{8}$$

Comparing IAF_I and IAF_S provides the user with a tool to choose between representing the variables by means of INDOMIX and representing the variables by means of the simpler model with poorer fit, PCAMIX.

7.5. INDOMIX applied to sets of quantitative or dichotomous variables

Above, INDOMIX has been described as a method for the analysis of a mixture of variables. One special case of this method is the case where INDOMIX is applied to qualitative variables only. In that case INDOMIX is equivalent to INDOQUAL, described in chapter 5. Another interesting special case is the case where INDOMIX is applied to quantitative variables only. Apart from this special case the case where all variables are dichotomous will also be treated here, because it turns out to be a special case of INDORT applied to merely quantitative variables.

INDORT applied to a set of quantitative variables comes down to

maximizing

$$f(X) = \sum_{j=1}^{m} \sum_{l=1}^{r} (\mathbf{x}_l' S_j \mathbf{x}_l)^2 = \sum_{j=1}^{m} \sum_{l=1}^{r} (n^{-1} \mathbf{x}_l' \mathbf{z}_j \mathbf{z}_j' \mathbf{x}_l)^2. \qquad (9)$$

Clearly, $(n^{-1}\mathbf{x}_l'\mathbf{z}_j\mathbf{z}_j'\mathbf{x}_l)$ can be rewritten as $n^{-1}(\mathbf{z}_j'\mathbf{x}_l)^2$, which is the square of the loading a_{jl} of variable j on component l. Hence INDORT applied to a set of quantitative variables can be seen as the method that maximizes the sum of fourth powers of loadings of the variables on the components, $\sum_j\sum_l a_{jl}^4$, over X. Typically, this "PCA" will not yield the same solution as ordinary PCA of quantitative variables. In section 8.5 this method is discussed a little further.

It is well–known that MCA (and hence PCAMIX) applied to a set of dichotomous (also called "binary") variables can be seen as an ordinary PCA of the dichotomous variables when the scores (zero and one) are standardized, see, for instance, De Leeuw (1973, p.56–57). This property is explained here again, and it is shown that a similar property exists for INDOQUAL.

When all variables are dichotomous, the indicator matrix for the j^{th} variable can be described by $G_j = (\mathbf{h}_j | \mathbf{h}_{j'})$, where $\mathbf{h}_j$ is the vector containing the zero–one scores on the dichotomous variable j, and $\mathbf{h}_{j'} = \mathbf{1}–\mathbf{h}_j$. It is readily verified that $J\mathbf{h}_{j'} = -J\mathbf{h}_j$, hence $JG_j = (J\mathbf{h}_j | J\mathbf{h}_{j'}) = (J\mathbf{h}_j | -J\mathbf{h}_j)$. The matrix $D_j = G_j'G_j$ has diagonal elements f_j and $(n–f_j)$, where f_j is the frequency of the unit–elements in $\mathbf{h}_j$. Then $JG_jD_j^{-1}G_j'J = \left[f_j^{-1}+(n–f_j)^{-1}\right]J\mathbf{h}_j\mathbf{h}_j'J = nf_j^{-1}(n–f_j)^{-1}J\mathbf{h}_j\mathbf{h}_j'J$. It is well–known that the variance of a dichotomous variable j is given by $n^{-2}f_j(n–f_j)$, hence, when $\mathbf{z}_j$ denotes the standardized version of $\mathbf{h}_j$, we have $JG_jD_j^{-1}G_j'J = n^{-1}\mathbf{z}_j\mathbf{z}_j'$.

From the above it follows that MCA of dichotomous variables finds object coordinates as the (normalized) eigenvectors of $n^{-1}\sum_j\mathbf{z}_j\mathbf{z}_j'$. Clearly, these object coordinates are the same as the component scores found by PCA of a matrix Z, containing as columns the vectors $\mathbf{z}_j$. INDOMIX uses the quantification matrices $S_j = JG_jD_j^{-1}G_j'J$. In case the variables are dichotomous, these quantification matrices are given by $JG_jD_j^{-1}G_j'J = n^{-1}\mathbf{z}_j\mathbf{z}_j'$, hence $S_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'$. It follows that, when INDOMIX is applied to dichotomous variables, the dichotomous variables can be considered as quantitative variables (with quantification matrices $n^{-1}\mathbf{z}_j\mathbf{z}_j'$). Furthermore, because INDOMIX maximizes $\sum_j\sum_l(\mathbf{x}_l'S_j\mathbf{x}_l)^2$ subject to $X'X = I_r$,

it can be seen that INDOMIX applied to dichotomous variables maximizes $n^{-2}\sum_j\sum_l(\mathbf{x}_l'\mathbf{z}_j\mathbf{z}_j'\mathbf{x}_l)^2 = n^{-2}\sum_j\sum_l(\mathbf{z}_j'\mathbf{x}_l)^4$, where $n^{-1/2}\mathbf{z}_j'\mathbf{x}_l$ is the point–biserial correlation between variable j and component l. This point–biserial correlation can be considered as the loading of variable j on component l. It follows that INDORT for dichotomous variables maximizes the sum of fourth powers of the loadings of the variables on the components. However, with a different interpretation of the squared point–biserial correlation, that is, as an E–correlation coefficient based on the quantification matrices $n^{-1}\mathbf{z}_j\mathbf{z}_j'$ and $\mathbf{x}_l\mathbf{x}_l'$, the "loading" of variable j on component l is given by $\text{tr}\, n^{-1}\mathbf{z}_j\mathbf{z}_j'\mathbf{x}_l\mathbf{x}_l' = n^{-1}(\mathbf{z}_j'\mathbf{x}_l)^2$, and hence INDORT for dichotomous variables can be seen as the method that maximizes the sum of squares of loadings of the variables on the components, as in ordinary PCA.

Above, it has been shown that dichotomous variables can be treated as quantitative variables by both PCAMIX and INDOMIX. This is very useful in practice, because the INDOMIX program can handle more quantitative variables than qualitative variables. However, when the INDOMIX solution is computed by treating dichotomous variables as quantitative variables, for these variables only the (point–biserial)–correlations are given, instead of the category coordinates. One can compute the category coordinates from these correlations as follows. Let y_{1jl} and y_{2jl} be the category centroids of the first and second categories, respectively, of variable j for component l. Because $\mathbf{h}_j'J\mathbf{h}_j = n^{-1}f_j(n-f_j)$, where f_j is the category frequency of the first category of variable j, and $\mathbf{x}_l = J\mathbf{x}_l$, y_{1jl} is given by $f_j^{-1}\mathbf{h}_j'\mathbf{x}_l = n^{-1/2}f_j^{-1/2}(n-f_j)^{1/2}(\mathbf{h}_j'J\mathbf{h}_j)^{-1/2}\mathbf{h}_j'J\mathbf{x}_l = n^{-1/2}f_j^{-1/2}(n-f_j)^{1/2}a_{jl}$, where a_{jl} is the point–biserial correlation between variable j and component l. Similarly, y_{2jl} can be shown to be equal to $-n^{-1/2}f_j^{1/2}(n-f_j)^{-1/2}a_{jl}$.

7.6. Discussion

In the present chapter INDOMIX has been described as a compromise between PCA of η^2–coefficients and PCAMIX. It is a compromise in that it yields a good representation of the variables (like PCA of η^2–coefficients) and at the same time it yields object coordinates (like PCAMIX). However, as has been shown in section 4.2, INDOMIX is not the only method that yields such a compromise. It has been shown there that TUCKALS–3 and unconstrained INDSCAL applied to the same quantification matrices yield other compromises

between PCA of η^2–coefficients and PCAMIX.

In the present chapter a particular choice has been made for the quantification matrices for the qualitative and quantitative variables. However, as has been said in section 4.4, it is an open question whether other choices of quantification matrices might be more useful. An important consequence of the possibility of different choices for quantification matrices is that, apart from INDORT on other quantification matrices, also alternatives for PCAMIX and PCA of η^2–coefficients can be developed. That is, alternatives of PCAMIX can be developed as methods that find object coordinates as the first r eigenvectors of the sum of the (alternative) quantification matrices. Implicitly, such an alternative method has been proposed by Gower (1971), and it is actually used by Cuadras (1989). Some alternatives for PCA of η^2–coefficients have in fact been proposed by Janson and Vegelius (1978a, 1982), by representing the associations in a set of qualitative and quantitative variables by other generalized correlation coefficients than the ones chosen by Saporta (1976).

8. SIMPLE STRUCTURE IN COMPONENTS ANALYSIS FOR MIXTURES OF QUALITATIVE AND QUANTITATIVE VARIABLES

8.1. Introduction

The present chapter focuses on two main subjects. First, a procedure for simple structure rotation for PCAMIX is considered. Next, it is shown that INDOMIX is closely related to this simple structure rotation. This relation between INDOMIX and the simple structure rotation for PCAMIX provides a useful new interpretation of INDOMIX. It explains why the loadings of the variables obtained by INDOMIX are more clearly clustered than those of PCAMIX. In addition, it leads to a better understanding of a phenomenon observed with INDOQUAL. That is, INDOQUAL does not only yield clearer clusters of variables, it also tends to yield solutions with clusters of *objects* that, per component, are more clearly separated and denser than those found in an MCA solution. Apart from giving a formal explanation of this phenomenon, it will be illustrated by means of an example data set. First, however, it will be explained why simple structure rotation of the PCAMIX solution might be useful.

In ordinary PCA, that is, PCA of quantitative variables, the solution for the components (the object scores) and the loadings is determined up to a rotation only. The purpose of so–called "simple structure" rotation is to obtain components that have a clear interpretation in terms of subsets of the variables. Simple structure criteria are usually defined in terms of optimal patterns of (in absolute sense) small and large loadings. In general, techniques for rotation to simple structure "are concerned with attaining a factor matrix with a maximum tendency to have both small and large loadings" (Kaiser, 1958, p.188). For a detailed discussion on the rationale behind simple structure rotation the reader is referred to Harman (1976).

As has been explained in chapter 7, PCAMIX can be formulated as PCA of the total set of (binary) indicator variables supplemented with the quantitative variables. Therefore, the object coordinates can be seen as component scores just as in ordinary PCA. Moreover, as in ordinary PCA, the

component scores are determined up to a rotation only. This is most easily verified by noting that PCAMIX maximizes the function $f(X) = \text{tr } X'\sum_j S_j X$, with S_j chosen as P_j or Q_j (as in chapter 7). Obviously, rotating the object coordinates matrix X by an orthonormal matrix T does not change the function value.

In practice, this rotational freedom seems not to have been used for finding "simple structure". In MCA (the special case of PCAMIX where all variables are qualitative) it is standard practice to use as components those that successively account for the maximum inertia, and ignore further rotations. In PCAMIX rotation does not seem to have been considered either. As is the case in ordinary PCA, the (unrotated) eigenvectors may yield components that are difficult to interpret. Therefore, the first purpose of the present chapter is to provide a method for rotating the component scores such that the best interpretable solution is found, according to some criterion.

As has been mentioned above, in ordinary PCA one rotates the component scores such that the loadings have optimal simple structure. That is, simple structure is expressed in terms of the loadings of the variables on the components. If one wants to optimize similar simple structure criteria by rotating the PCAMIX components, first of all one needs to define loadings, or rather squared loadings, of the variables on the PCAMIX components.

8.2. A definition of squared loadings in PCAMIX

Like PCA, PCAMIX finds component scores for the objects on several components. In ordinary PCA "loadings" of the variables on the components are given by the correlations between the variables and the components. In PCAMIX it is possible to define loadings for the quantitative variables in the same way. That is, the loading of the quantitative variable j on component l can be given by $a_{jl} = n^{-1/2}\mathbf{z}_j'\mathbf{x}_l$, the product–moment correlation between variable j and component l.

For the qualitative variables one cannot use the product–moment correlation. Instead, one has to choose another coefficient which expresses the correlation between a qualitative variable and a (quantitative) component. Such a measure in MCA is the "discrimination measure" (Gifi, 1981), which is

the contribution of a component to the inertia of a variable accounted for. This discrimination measure is given by $c_{jl} \equiv \mathbf{x}_l' P_j \mathbf{x}_l$. Gifi (1981, p.96) explains that this measure can be seen as the squared correlation between variable j when it is "optimally quantified" and component l. Another interpretation of c_{jl} is that it is the well–known correlation ratio η^2. In both interpretations the measure c_{jl} is considered as a squared correlation. Therefore, it will be considered here as the squared loading of variable j on component l.

In order to have the same notation for qualitative and quantitative variables, c_{jl} is defined for a quantitative variable as $c_{jl} = a_{jl}^2 = n^{-1}(\mathbf{z}_j'\mathbf{x}_l)^2 = \mathbf{x}_l'(n^{-1}\mathbf{z}_j\mathbf{z}_j')\mathbf{x}_l$. Hence defining S_j as P_j or Q_j, we have $c_{jl} = \mathbf{x}_l' S_j \mathbf{x}_l$ for both qualitative and quantitative variables. It is of interest to note that PCAMIX can be formulated as the method that maximizes $\sum_{jl} c_{jl}$, with c_{jl} defined as above, over X, subject to $X'X = I_r$.

Having defined squared loadings for variables on PCAMIX components, we are in a position to consider criteria that measure simple structure of the loadings. Before considering simple structure criteria for PCAMIX, some well–known simple structure criteria that are used with ordinary PCA will be discussed.

8.3. Simple structure rotations for PCA

Kaiser (1958) has described several simple structure criteria, as well as procedures to optimize these criteria over orthogonal rotations of the loading matrix. Some of these have later been included in the orthomax family of orthogonal rotations (Jennrich, 1970, Crawford & Ferguson, 1970; see Clarkson & Jennrich, 1988). The criteria which are optimized by these rotation techniques will be discussed briefly in the present section.

The orthomax family of simple structure rotations for PCA can be described as the set of techniques that maximize the orthomax criterion (denoted by the acronym ORMAX). This criterion is expressed in terms of the squared loadings of the variables on the components. Let the loading for variable j on component l be given by a_{jl}, $j = 1,...,m$, $l = 1,...,r$. Then the ORMAX criterion is given by

$$\text{ORMAX} = \sum_{j=1}^{m}\sum_{l=1}^{r} a_{jl}^4 - \frac{\gamma}{m}\sum_{l=1}^{r}\left(\sum_{j=1}^{m} a_{jl}^2\right)^2. \tag{1}$$

Although in principle γ can be any scalar, it is assumed here that $0 \leq \gamma \leq 1$. Several special choices of γ result in well–known simple structure criteria. That is, choosing $\gamma = 0$ yields the quartimax criterion (QMAX)

$$\text{QMAX} = \sum_{j=1}^{m}\sum_{l=1}^{r} a_{jl}^4, \tag{2}$$

which has originally been proposed by Ferguson (1954). On the other hand, choosing $\gamma = 1$ yields the varimax criterion (VMAX) proposed by Kaiser (1958)

$$\text{VMAX} = \sum_{j=1}^{m}\sum_{l=1}^{r} a_{jl}^4 - \frac{1}{m}\sum_{l=1}^{r}\left(\sum_{j=1}^{m} a_{jl}^2\right)^2. \tag{3}$$

In addition to quartimax and varimax, Kaiser (1958) described the following three simple structure criteria: Carroll (1953) proposed to minimize

$$\text{CROSSMIN} = \sum_{j=1}^{m}\sum_{k<l}^{r} a_{jk}^2 a_{jl}^2; \tag{4}$$

Neuhaus and Wrigley (1954) proposed to maximize the overall variance of squared loadings (OVERMAX)

$$\text{OVERMAX} = \sum_{j=1}^{m}\sum_{l=1}^{r} a_{jl}^4 - \frac{1}{mr}\left(\sum_{j=1}^{m}\sum_{l=1}^{r} a_{jl}^2\right)^2; \tag{5}$$

Saunders (1953) proposed to maximize the kurtosis of the total set of loadings combined with the set of loadings with reversed sign, which is proportional to

$$\text{KURTMAX} = \sum_{j=1}^{m}\sum_{l=1}^{r} a_{jl}^4 \Big/ \frac{1}{mr}\left(\sum_{j=1}^{m}\sum_{l=1}^{r} a_{jl}^2\right)^2. \tag{6}$$

So far, only the simple structure criteria themselves have been discussed. The techniques for optimizing these criteria will now be discussed briefly. First of all, it should be noted that optimizing these criteria over orthogonal rotations of the component scores is equivalent to optimizing these criteria over orthogonal rotations of the loadings. This follows from

the fact that, when the $m \times r$ matrix $A = n^{-1/2}Z'X$ contains the loadings of the variables (in matrix Z) on the components (in matrix X), a rotation of the components by a matrix T is paralleled by the same rotation of the matrix of loadings. That is, $AT = n^{-1/2}Z'XT$ contains the loadings of the variables on the rotated components. The techniques for optimizing the simple structure criteria mentioned above are all techniques for rotating the loading matrix.

In considering the different techniques for rotating the loading matrix, it is useful to note that the quartimax criterion and the varimax criterion are both special cases of the orthomax criterion. Moreover, as has been shown by Kaiser (1958), minimizing the CROSSMIN criterion and maximizing the OVERMAX and KURTMAX criteria over orthogonal rotations of the loading matrix is equivalent to maximizing the quartimax criterion. As a consequence, any of the optimization problems mentioned above can be subsumed under the general problem of maximizing the orthomax criterion over orthogonal rotations of the loading matrix. A description of maximizing the orthomax function, which turns out to be particularly useful in the present context (as will become clear later), was given by Ten Berge, Knol and Kiers (1988). Let A be the $m \times r$ matrix of loadings, and let $\mathbf{a}_j'$ be the j^{th} row of A. Ten Berge et al. (1988) define $E_j \equiv (\delta A'A - m\mathbf{a}_j\mathbf{a}_j')$, for $j = 1,...,m$, with δ defined such that $\gamma = \delta(2-\delta)$, and show that the problem of maximizing the orthomax function is equivalent to simultaneously diagonalizing the set of E_j matrices in the least squares sense, or equivalently, maximizing $\sum_j \text{tr (Diag } T'E_jT)^2$ over orthonormal matrices T. For this problem of simultaneously diagonalizing a set of matrices one can use an algorithm proposed by De Leeuw and Pruzansky (1978).

In the present section a number of simple structure criteria has been described and it has been pointed out that the orthogonal rotations that optimize these criteria can all be found by means of simultaneous diagonalization of the E_j matrices. In the next section, it will be shown how the same simple structure criteria can be optimized over rotations of the PCAMIX component scores solution.

8.4. Simple structure rotations for PCAMIX

In the present section, methods will be discussed for rotation of the

PCAMIX *component scores* such that the *loadings* have optimal simple structure in terms of the orthomax criteria. It should be noted that rotated component scores in PCAMIX do not correspond to a loading matrix which can be found by rotating the original loading matrix. This is a consequence of the fact that no definition of loadings for qualitative variables seems to be available such that rotating component scores corresponds to rotating the matrix of corresponding loadings. Therefore, in PCAMIX it does not suffice to express the simple structure criteria in terms of the original loadings and a rotation matrix. Instead, these criteria are to be expressed in terms of the squared loadings of the variables on the rotated components. It will now be shown how this can be accomplished.

The simple structure criteria given in (1), (2), (3), (4), (5), and (6) can be expressed in terms of the squared PCAMIX loadings by replacing a_{jl}^2 by c_{jl} for all j and l. The problem of optimizing the simple structure criteria over orthogonal rotations of the PCAMIX component scores will now be treated by giving an algorithm for maximizing the general ORMAX function (1) only. As has been explained above, the QMAX (2) and VMAX (3) functions are special cases of this function. Optimizing the CROSSMIN (4), OVERMAX (5), and KURTMAX (6) functions will be shown to be equivalent to maximizing the QMAX function.

The ORMAX criterion can be rewritten in terms of c_{jl} as

$$\text{ORMAX} = \sum_{j=1}^{m}\sum_{l=1}^{r} c_{jl}^2 - \frac{\gamma}{m}\sum_{l=1}^{r}\Big(\sum_{j=1}^{m} c_{jl}\Big)^2. \tag{7}$$

In order to write the ORMAX criterion as a function f_{or} of the component scores matrix X the loading $c_{jl} = \mathbf{x}_l' S_j \mathbf{x}_l$ is substituted for c_{jl} in (7). This gives

$$\begin{aligned} f_{or}(X) \equiv \text{ORMAX} &= \sum_{j=1}^{m}\sum_{l=1}^{r} (\mathbf{x}_l' S_j \mathbf{x}_l)^2 - \frac{\gamma}{m}\sum_{l=1}^{r}\Big(\sum_{j=1}^{m} \mathbf{x}_l' S_j \mathbf{x}_l\Big)^2 \\ &= \sum_{j=1}^{m}\sum_{l=1}^{r} (\mathbf{x}_l' S_j \mathbf{x}_l)^2 - \frac{\gamma}{m}\sum_{l=1}^{r} (\mathbf{x}_l' \textstyle\sum_j S_j \mathbf{x}_l)^2. \end{aligned} \tag{8}$$

In order to find the rotation that maximizes the ORMAX criterion we have to maximize $f_{or}(FT)$ over orthonormal matrices T, where F contains the unrotated

component scores solution for PCAMIX. The problem of finding the rotation that maximizes f_{or} will be translated here into a problem of simultaneously diagonalizing a set of matrices, just as has been done by Ten Berge et al. (1988) for the orthomax rotation of a loading matrix. Let the l^{th} column of T be given by $\mathbf{t}_l$, then $f_{or}(FT)$ is given by

$$f_{or}(FT) = \sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{t}_l'F'S_jF\mathbf{t}_l)^2 - \frac{\gamma}{m}\sum_{l=1}^{r}(\sum_j \mathbf{t}_l'F'S_jF\mathbf{t}_l)^2$$

$$= \sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{t}_l'(F'S_jF-\delta m^{-1}\textstyle\sum_k F'S_kF)\mathbf{t}_l)^2, \qquad (9)$$

with δ chosen such that $2\delta-\delta^2=\gamma$. For $\delta=\gamma=1$, the right-hand side of (9) follows at once from the fact that the variance can be written as an average squared deviation from the mean. If $\delta \neq 1$, the second equality in (9) follows from the fact that

$$\sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{t}_l'(F'S_jF-\delta m^{-1}\textstyle\sum_k F'S_kF)\mathbf{t}_l)^2 = \sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{t}_l'F'S_jF\mathbf{t}_l)^2 +$$
$$\sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{t}_l'\delta m^{-1}\textstyle\sum_k F'S_kF\mathbf{t}_l)^2 - 2\sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{t}_l'F'S_jF\mathbf{t}_l)(\mathbf{t}_l'\delta m^{-1}\textstyle\sum_k F'S_kF\mathbf{t}_l)$$

$$= \sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{t}_l'F'S_jF\mathbf{t}_l)^2 + m\sum_{l=1}^{r}\delta^2m^{-2}(\mathbf{t}_l'\textstyle\sum_j F'S_jF\mathbf{t}_l)^2$$
$$- 2\,\delta m^{-1}\sum_{l=1}^{r}(\mathbf{t}_l'F'\textstyle\sum_j S_jF\mathbf{t}_l)(\mathbf{t}_l'\sum_k F'S_kF\mathbf{t}_l)$$

$$= \sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{t}_l'F'S_jF\mathbf{t}_l)^2 + \delta^2m^{-1}\sum_{l=1}^{r}(\mathbf{t}_l'\textstyle\sum_j F'S_jF\mathbf{t}_l)^2 - 2\,\delta m^{-1}\sum_{l=1}^{r}(\mathbf{t}_l'\textstyle\sum_j F'S_jF\mathbf{t}_l)^2$$

$$= \sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{t}_l'F'S_jF\mathbf{t}_l)^2 + m^{-1}(\delta^2-2\delta)\sum_{l=1}^{r}(\mathbf{t}_l'\textstyle\sum_j F'S_jF\mathbf{t}_l)^2. \qquad (10)$$

The columns of F are eigenvectors of $\sum_j S_j$, normalized to unit sums of squares, hence $F'\sum_j S_jF = \Lambda$, where Λ is the diagonal matrix with the first r eigenvalues of $\sum_j S_j$ on its diagonal. With this result (9) can be rewritten as

$$f_{or}(FT) = \sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{t}_l'(F'S_jF-\delta m^{-1}\Lambda)\mathbf{t}_l)^2. \qquad (11)$$

Let $\tilde{E}_j$ be defined by

$$\tilde{E}_j \equiv (mF'S_jF - \delta\Lambda), \tag{12}$$

for $j = 1,...,m$. Then (11) can be rewritten as

$$f_{or}(FT) = m^{-2} \sum_{j=1}^{m} \text{tr } T'\tilde{E}_jT(\text{Diag } T'\tilde{E}_jT). \tag{13}$$

Ten Berge (1984, p.348) has shown that maximizing such a function over orthonormal matrices T is equivalent to the problem of simultaneously diagonalizing a set of matrices $\tilde{E}_1,...,\tilde{E}_m$ in the least squares sense, and that hence the algorithm for this problem, proposed by De Leeuw and Pruzansky (1978), can be used, or any other algorithm for the simultaneous diagonalization of a set of symmetric matrices. Hence maximizing the orthomax criterion by rotating the PCAMIX component scores can be done by means of an algorithm for simultaneous diagonalization of matrices $\tilde{E}_1,...,\tilde{E}_m$ with $\tilde{E}_j$ defined as in (12).

As has been mentioned above, Ten Berge et al. (1988) have shown that the problem of maximizing the orthomax function over orthogonal rotations of a PCA loading matrix is equivalent to the problem of simultaneously diagonalizing the matrices $E_j = (\delta A'A - m\mathbf{a}_j\mathbf{a}_j')$, where A is the $m \times r$ PCA loading matrix, and $\mathbf{a}_j'$ is the j^{th} row of A. It will now be shown that, in the special case where PCAMIX is applied to a set of merely quantitative variables, the procedure for the orthomax rotation of the PCAMIX solution can be seen as the simultaneous diagonalization of a set of $\tilde{E}_j$ matrices which are proportional to the E_j matrices in the Ten Berge et al. (1988) procedure. When all variables are quantitative the S_j matrices are given by $S_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'$. Hence $\tilde{E}_j = (mn^{-1}F'\mathbf{z}_j\mathbf{z}_j'F - \delta\Lambda)$, where Λ is the diagonal matrix with eigenvalues of $n^{-1}\sum_j\mathbf{z}_j\mathbf{z}_j'$. Clearly, $\mathbf{a}_j' = n^{-1/2}\mathbf{z}_j'F$, hence $mn^{-1}F'\mathbf{z}_j\mathbf{z}_j'F = m\mathbf{a}_j\mathbf{a}_j'$. In addition, matrix $A'A = F'n^{-1}\sum_j\mathbf{z}_j\mathbf{z}_j'F = \Lambda$. Therefore, $\tilde{E}_j$ can be written as $\tilde{E}_j = (m\mathbf{a}_j\mathbf{a}_j' - \delta A'A) = -E_j$. Hence the simultaneous diagonalization of the matrices $\tilde{E}_j$ and that of the matrices E_j yield the same rotation matrix. It can be concluded that the orthomax rotation procedure for ordinary PCA is a special case of the orthomax rotation procedure for PCAMIX that is suggested here.

The quartimax criterion and the varimax criterion are special cases of the orthomax criterion. Therefore, with the orthomax rotation procedure for PCAMIX suggested above we have at once a quartimax and a varimax procedure for rotating the PCAMIX solution. These are given by setting γ to 0 or 1, respectively. The orthomax rotation procedure for PCAMIX has been described as the simultaneous diagonalization of a set of $\tilde{E}_j$ matrices as defined by (12). The $\tilde{E}_j$ matrices do not explicitly contain γ, but depend on γ because δ depends on γ. The $\tilde{E}_j$ matrices for the quartimax procedure are given by taking $\delta = 0$ (or $\delta = 2$, which is less convenient and therefore ignored), because then $\gamma = 0$. For the varimax procedure the $\tilde{E}_j$ matrices are given by $\delta = 1$, because then $\gamma = 1$.

So far only the QMAX and VMAX criteria have been considered. As has been shown by Kaiser (1958), for ordinary PCA component scores the procedure for optimizing the CROSSMIN, OVERMAX and KURTMAX criteria over orthogonal rotations are all equivalent to the quartimax rotation procedures. That is, these criteria are all optimized by the same orthogonal rotation matrix. For orthogonal rotation of the PCAMIX component scores again the same rotation matrix minimizes CROSSMIN and maximizes QMAX, OVERMAX and KURTMAX, as is shown below.

The CROSSMIN criterion for PCAMIX can be rewritten in terms of the squared loadings of the variables on the (rotated) components as

$$\text{CROSSMIN} = \sum_{j=1}^{m} \sum_{k<l}^{r} c_{jk}c_{jl} = \frac{1}{2} \sum_{j=1}^{m} \Big(\sum_{l=1}^{r} c_{jl} \Big)^2 - \frac{1}{2} \sum_{j=1}^{m} \sum_{l=1}^{r} c_{jl}^2. \tag{14}$$

The first term in the right-hand side contains $\sum_l c_{jl} = \sum_l \mathbf{t}_l' F' S_j F \mathbf{t}_l =$ tr $T'F'S_jFT =$ tr $F'S_jF$, which does not depend on T, because T is orthonormal. Therefore, minimizing the CROSSMIN criterion over orthogonal rotations is equivalent to maximizing the second term in the right-hand side of (14). This term is the QMAX criterion (multiplied by 1/2) expressed in squared loadings c_{jl}. Therefore, minimizing the CROSSMIN criterion is equivalent to maximizing QMAX, for PCAMIX.

The OVERMAX and the KURTMAX criteria can be rewritten in terms of squared loadings c_{jl} as

$$\text{OVERMAX} = \sum_{j=1}^{m}\sum_{l=1}^{r} c_{jl}^2 - \frac{1}{mr}\left(\sum_{j=1}^{m}\sum_{l=1}^{r} c_{jl}\right)^2 \tag{15}$$

and

$$\text{KURTMAX} = \sum_{j=1}^{m}\sum_{l=1}^{r} c_{jl}^2 \Big/ \frac{1}{mr}\left(\sum_{j=1}^{m}\sum_{l=1}^{r} c_{jl}\right)^2, \tag{16}$$

respectively. Both criteria contain the term $\sum_l c_{jl}$, which does not depend on the orthonormal rotation matrix T. As a consequence, finding the orthogonal rotation that maximizes these criteria depends only on the term $\sum_j\sum_l c_{jl}^2$, that is, again, the QMAX criterion. It can be concluded that finding the rotation matrix that optimizes the CROSSMIN, OVERMAX, and KURTMAX criteria for PCAMIX comes down to finding the orthogonal rotation that maximizes the QMAX criterion. The procedure for finding this rotation matrix has been discussed above as a special case of the procedure for orthomax rotation of the PCAMIX solution.

This concludes the discussion of simple structure rotation techniques for PCAMIX solutions. It is in no way intended to give a complete account of possible simple structure rotations for PCAMIX solutions. There are many other simple structure criteria for ordinary PCA, among which oblique simple structure rotations, as described, for instance by Clarkson and Jennrich (1988), that might be of interest for PCAMIX.

In the present section methods have been discussed for optimizing simple structure criteria for loadings of variables on PCAMIX component scores. These methods are based on rotation of the PCAMIX component scores. Such a rotation does not affect the optimality of the function that is maximized by PCAMIX, that is, $\sum_j\sum_l c_{jl}$, which can be seen as a measure for explained inertia. One might, however, want to find those components that maximize the simple structure criteria, possibly loosing the optimality of the function maximized by PCAMIX. In the case of merely quantitative variables, for instance, one might seek the components that have the maximal varimax or quartimax function value over all possible sets of orthogonal components, regardless of the variance they explain. In the next section, methods will be discussed for finding such components, that is, components that maximize the orthomax criterion for PCAMIX loadings over all possible sets of orthogonal

components.

8.5. INDOMIX and a generalization

PCAMIX is the best–known method for the analysis of mixtures of variables. In chapter 7, however, an alternative method, INDOMIX, has been developed as a compromise between PCAMIX and "PCA of quantification matrices", in that it is directed at optimally representing the variables (as PCA of quantification matrices does) while at the same time providing component scores for the objects (as PCAMIX does). In the present section it is shown that this method optimizes the quartimax criterion over all possible sets of orthogonal component scores. In addition, an alternative method is discussed, which might be used for optimizing the varimax criterion, or any other criterion that belongs to the orthomax family.

INDOMIX comes down to maximizing

$$g(X) = \sum_{j=1}^{m} \text{tr Diag}(X'S_jX)^2 = \sum_{j=1}^{m}\sum_{l=1}^{r} (\mathbf{x}_l'S_j\mathbf{x}_l)^2 = \sum_{j=1}^{m}\sum_{l=1}^{r} c_{jl}^2, \qquad (17)$$

subject to $X'X = I_r$.

Clearly, maximizing $g(X)$ is equivalent to maximizing the quartimax function. INDOMIX maximizes the quartimax function over orthogonal component scores matrices X, whereas the quartimax rotation applied to PCAMIX maximizes the quartimax function over orthogonally rotated component scores matrices FT. The class of matrices X that are only constrained by $X'X = I_r$ contains not only all rotations FT of F, but also matrices with columns outside the column–space of F. Therefore, the maximum of $g(X)$ is always at least as large as the maximum of $g(FT)$. This provides another interpretation of INDOMIX. INDOMIX maximizes the quartimax criterion over all possible orthogonal component scores matrices, in this way yielding a quartimax value that is always at least as high as the maximum possible quartimax value that can be obtained by orthogonal rotation of the PCAMIX solution. This implies that INDOMIX yields solutions that have a higher amount of simple structure than the optimally rotated PCAMIX solutions, when simple structure is defined in the quartimax sense.

The quartimax criterion being one of the first analytic simple structure

criteria, it is not the most prevalent criterion in ordinary PCA. As Kaiser (1958) pointed out, using the quartimax criterion tends to yield solutions with one general component, which is quite contrary to the purpose of achieving maximum simple structure. This tendency of yielding a general component does not seem to be present with INDOMIX, and in practice it is found that, even though INDOMIX maximizes only the quartimax function, INDOMIX tends to yield simple structure in terms of other criteria (like varimax) as well, as will be discussed later. Nevertheless, INDOMIX does not maximize these other criteria, and it seems useful to discuss methods that do maximize other simple structure criteria (in the orthomax family) over all possible sets of orthogonal component scores, in the same way as INDOMIX maximizes the quartimax criterion.

The orthomax function (8) can be rewritten as

$$\begin{aligned} f_{or}(X) &= \sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{x}_l'S_j\mathbf{x}_l)^2 - \frac{\gamma}{m}\sum_{l=1}^{r}(\mathbf{x}_l'\textstyle\sum_j S_j\mathbf{x}_l)^2 \\ &= \sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{x}_l'(S_j-\delta m^{-1}\textstyle\sum_k S_k)\mathbf{x}_l)^2, \end{aligned} \tag{18}$$

with δ again chosen such that $2\delta-\delta^2=\gamma$. The last step in the derivation of (18) is based on a similar reasoning as is made in deriving (10). Clearly, choosing $\delta=0$ (or $\delta=2$) we have $f_{or}(X)=g(X)$. However, (18) can be defined for any other γ between 0 and 1 (to which corresponds $\delta=1\pm(1-\gamma)^{1/2}$). In particular, the varimax function is maximized over all orthogonal component scores matrices X by maximizing

$$h(X)=\sum_{j=1}^{m}\sum_{l=1}^{r}(\mathbf{x}_l'(S_j-m^{-1}\textstyle\sum_k S_k)\mathbf{x}_l)^2, \tag{19}$$

over X, subject to $X'X=I_r$. That is, this method applies INDORT to the matrices $(S_j-m^{-1}\sum_k S_k)$, which are the matrices S_j "centered" with respect to their mean.

Ten Berge et al. (1988) have provided an algorithm for maximizing (17). They have shown that this algorithm converges monotonically, when the S_j are positive semi-definite. Their algorithm might be used for maximizing (18) as well, with S_j replaced by $(S_j-\delta m^{-1}\sum_k S_k)$, but then monotone convergence of the algorithm is no longer guaranteed. An algorithm for which monotone

convergence is guaranteed has been described by Kiers (in press). It is a slightly adapted version of the Ten Berge et al. algorithm.

It can be concluded that several methods are now available for finding components for mixtures of variables such that the orthomax simple structure criteria (for the loadings of the variables on the components) are maximized. One of these approaches is to perform PCAMIX first and then rotate the component scores such that they optimize the simple structure criteria over orthogonal rotations of the PCAMIX component scores. This procedure finds those sets of components that account best for the inertia as measured in PCAMIX, and, among these components, it finds that set of components that yields the highest simple structure value. The other approach, the generalization of INDOMIX that maximizes (18), is to seek those components that have the best possible simple structure, at the cost of a loss (which tends to be rather small in practice) in explained inertia.

The special case where the generalization of INDOMIX that maximizes the varimax function h(X) is applied to a set of merely quantitative variables is of special interest, because it provides an alternative to ordinary PCA. This method will yield varimax values that are always at least as high as the varimax values of rotated PCA loadings. Therefore, when one's main objective is to find components that have clear simple structure, that is, give clear clusters of variables, and when accounting for the variance is less important, then the generalization of INDOMIX might be a useful alternative to ordinary PCA. Moreover, although the explained variance $\sum_j\sum_l c_{jl}$ is no longer maximized by the generalization of INDOMIX, it cannot be very small either, because, this would contradict the maximality of h(X) which can be written as $\sum_j\sum_l(c_{jl} - m^{-1}\sum_k c_{kl})^2$, that is, the sum of column–variances of the elements of C.

In the present section, a method has been proposed for maximizing the simple structure criteria in the orthomax family, including the VMAX and QMAX criteria. In addition, maximizing the QMAX criterion over orthogonal rotations has been shown to be equivalent to optimizing the CROSSMIN, OVERMAX, and KURTMAX criteria, which are in this way also related to the orthomax family. However, when the orthomax function is maximized over all possible sets of orthogonal component scores, the equivalence between maximizing QMAX and optimizing CROSSMIN, OVERMAX, and KURTMAX can no

longer be shown to hold. Nevertheless, there are some relations between INDOMIX and optimally rotated PCAMIX solutions in terms of these criteria. These and some other results are given in the next section.

8.6. Relations between INDOMIX and simple structure rotations of PCAMIX

In the previous section, the generalization of INDOMIX that maximizes the orthomax function over all possible orthogonal component scores matrices has been discussed. Obviously, this method always yields an orthomax function value that is at least as high as the one obtained by the orthomax rotation of the PCAMIX solution. As has been mentioned above, this method is a generalization of INDOMIX (chapter 7) which maximizes the quartimax function. In the present section, it will be shown that the latter method, INDOMIX, does not only yield a quartimax function value which is at least as high as the one attained by quartimax rotation of the PCAMIX solution, but that it also yields values at other simple structure criteria that are at least as high as the ones attained by optimally rotated PCAMIX solutions. These comparisons are summarized in Results 1 to 5. In these results the component scores matrix of the INDOMIX solution is denoted as X_I, that of the unrotated PCAMIX solution as F, and that of the optimally rotated PCAMIX solution as FT_O. That is, T_O is the rotation matrix that optimizes the simple structure criterion at hand. The simple structure criteria given above as ORMAX, QMAX, VMAX, OVERMAX and KURTMAX are seen here as functions of the component scores matrices.

Result 1. $\text{QMAX}(X_I) \geq \text{QMAX}(FT_O) \geq \text{QMAX}(F)$.

Result 2. a. $\text{OVERMAX}(X_I) \geq \text{OVERMAX}(FT_O) \geq \text{OVERMAX}(F)$.

b. $\text{KURTMAX}(X_I) \geq \text{KURTMAX}(FT_O) \geq \text{KURTMAX}(F)$.

Result 3. $\text{ORMAX}(X_I) \geq \text{ORMAX}(F)$.

Result 4. $\text{VMAX}(X_I) \geq \text{VMAX}(F)$.

Result 5. $\text{ORMAX}(X_I) \geq \text{ORMAX}(FT_O) = \text{ORMAX}(F)$, if $m = 2$, and both variables are qualitative, that is, in case PCAMIX is equivalent to correspondence analysis.

The results given above are proven as follows.

Proof 1. From the fact that INDOMIX maximizes QMAX over all possible component scores matrices it follows at once that $\text{QMAX}(X_I) \geq \text{QMAX}(FT_0)$. The optimality of T_0 guarantees that $\text{QMAX}(FT_0) \geq \text{QMAX}(F)$.□

Proof 2. The OVERMAX function can be written as

$$\text{OVERMAX}(X) = \text{g}(X) - m^{-1}r^{-1}\text{f}^2(X), \tag{20}$$

where $\text{f}(X)$ is the function maximized by PCAMIX, and $\text{g}(X)$ is the function maximized by INDOMIX (17). Obviously, $\text{f}(FT_0) = \text{f}(F) \geq \text{f}(X_I)$, and $\text{g}(X_I) \geq \text{g}(FT_0)$. Because $\text{f}(X)$ is nonnegative for all X, it follows that $\text{f}^2(FT_0) \geq \text{f}^2(X_I)$. As a consequence $\left[\text{g}(X_I) - m^{-1}r^{-1}\text{f}^2(X_I)\right] \geq \left[\text{g}(FT_0) - m^{-1}r^{-1}\text{f}^2(FT_0)\right]$, that is $\text{OVERMAX}(X_I) \geq \text{OVERMAX}(FT_0)$. Obviously, the optimality of T_0 guarantees that $\text{OVERMAX}(FT_0) \geq \text{OVERMAX}(F)$. This completes the proof of Result 2a. For the proof of Result 2b it is useful to note that $\text{KURTMAX}(X) = mr\text{g}(X)/\text{f}^2(X)$, from which Results 2b follow at once, according to a similar reasoning as the one that is used in proving 2a. □

Proof 3. The ORMAX function (8) can be rewritten as

$$\text{ORMAX}(X) = \text{g}(X) - \frac{\gamma}{m}\,\text{k}(X), \tag{21}$$

where $0 \leq \gamma \leq 1$, and $\text{k}(X)$ is defined as

$$\text{k}(X) \equiv \sum_{l=1}^{r} (\mathbf{x}_l' \textstyle\sum_j S_j \mathbf{x}_l)^2. \tag{22}$$

Let an eigendecomposition of $\sum_j S_j$ be given by $\sum_j S_j = K\Lambda K'$. Then $\text{k}(X)$ can be rewritten as

$$\text{k}(X) = \sum_{l=1}^{r} (\mathbf{x}_l' K\Lambda K' \mathbf{x}_l)^2. \tag{23}$$

From the Cauchy–Schwarz theorem it follows that

$$(\mathbf{x}_l' K\Lambda K' \mathbf{x}_l)^2 = \left[(\mathbf{x}_l' K\Lambda K')(\mathbf{x}_l)\right]^2 \leq (\mathbf{x}_l' K\Lambda^2 K' \mathbf{x}_l) \tag{24}$$

and hence

$$\sum_{l=1}^{r} (\mathbf{x}_l' K \Lambda K' \mathbf{x}_l)^2 \leq \sum_{l=1}^{r} (\mathbf{x}_l' K \Lambda^2 K' \mathbf{x}_l) = \text{tr } X' K \Lambda^2 K' X. \tag{25}$$

It is readily verified (Ten Berge, 1983) that the right–hand side in (25) is smaller than or equal to the sum of the first r diagonal elements of Λ^2. Let Λ_r denote the diagonal matrix containing the first r diagonal elements of Λ, then

$$\mathrm{k}(X) = \sum_{l=1}^{r} (\mathbf{x}_l' K \Lambda K' \mathbf{x}_l)^2 \leq \text{tr } \Lambda_r^2. \tag{26}$$

Inequality (26) yields an upper bound to $\mathrm{k}(X)$. Clearly, this upper bound is attained by choosing X as K_r, the matrix with the r eigenvectors of $\sum_j S_j$ that belong to the first r eigenvalues of $\sum_j S_j$. This is precisely the unrotated PCAMIX solution F for the component scores. Therefore, $\mathrm{k}(X_I) \leq \mathrm{k}(F)$, and, because $\gamma \geq 0$, $-\gamma \mathrm{k}(X_I) \geq -\gamma \mathrm{k}(F)$. Combining this result with the fact that $\mathrm{g}(X_I) \geq \mathrm{g}(F)$ proves that $\mathrm{ORMAX}(X_I) \geq \mathrm{ORMAX}(F)$. □

Proof 4. Result 4 follows immediately from Result 3 when γ is taken equal to 1.□

Proof 5. If $m = 2$ and both variables are qualitative it can be shown that $F'S_1F = F'S_2F = \frac{1}{2}\Lambda_r$. Hence $\mathrm{f}_{or}(FT_{\mathrm{O}})$ is the maximum over T of

$$\mathrm{f}_{or}(FT) = \sum_{j=1}^{2} \sum_{l=1}^{r} (\mathbf{t}_l' (F'S_jF - \delta \tfrac{1}{2}\Lambda_r) \mathbf{t}_l)^2, \tag{27}$$

as follows from (11). Substituting $F'S_1F = F'S_2F = \frac{1}{2}\Lambda_r$ in (27) yields

$$\mathrm{f}_{or}(FT) = \sum_{j=1}^{2} \sum_{l=1}^{r} (\mathbf{t}_l' ((1-\delta) \tfrac{1}{2}\Lambda_r) \mathbf{t}_l)^2. \tag{28}$$

It is proven analogously to the proof of (26) that

$$\mathrm{f}_{or}(FT) = \sum_{j=1}^{2} \sum_{l=1}^{r} (\mathbf{t}_l' ((1-\delta) \tfrac{1}{2}\Lambda_r) \mathbf{t}_l)^2 \leq \sum_{j=1}^{2} \tfrac{1}{4}(1-\delta)^2 \text{tr } \Lambda_r^2. \tag{29}$$

The right–hand side of (29) gives an upper bound to $\mathrm{f}_{or}(FT)$, which is attained for $T = I$. That is, $\mathrm{f}_{or}(FT_{\mathrm{O}})$, the maximum over T of $\mathrm{f}_{or}(FT)$ is equal to

$f_{or}(F)$, hence ORMAX(FT_O) = ORMAX(F). With this equality ORMAX(X_I) $\geq$ ORMAX(FT_O) = ORMAX(F) follows immediately from Result 3.□

It is of interest to mention that, for $m = 2$, Result 5 implies that VMAX(F) = VMAX(FT_O) = 0. The fact that VMAX(FT_O) = 0 follows at once upon substitution of $\delta = 1$ in (29). That is, the Correspondence Analysis (MCA with $m = 2$) solution yields a varimax function which is always zero, and cannot be increased by rotating the solution.

Result 5 has been included, because it proves that VMAX(X_I) $\geq$ VMAX(FT_O) in a special case. This result does not generally hold for $m > 2$. Yet, in practice, it is often found that VMAX(X_I) $\geq$ VMAX(FT_O). This can be explained by the fact that VMAX = $\sum_j\sum_l c_{jl}^2 - m^{-1}\sum_l(\sum_j c_{jl})^2$, in which the first term is maximized by INDOMIX. The PCAMIX solution (and any rotation of this) maximizes $\sum_j\sum_l c_{jl}$. Obviously, a low value for $\sum_l(\sum_j c_{jl})^2$ for a certain PCAMIX solution would contradict the optimality of $\sum_j\sum_l c_{jl}$. Hence PCAMIX tends to find loadings for which $\sum_l(\sum_j c_{jl})^2$ is high. This explains why there is a tendency for the INDOMIX solution to yield a higher value of the VMAX criterion than any rotation of the PCAMIX solution. Along with Results 1 and 2, this result is of special interest when the INDOMIX and PCAMIX solutions are to be compared in terms of simple structure. In the next section, a comparison is made of the special cases of INDOMIX and PCAMIX where they are applied to sets of qualitative variables, that is, INDOQUAL and MCA, respectively. It will be shown that the higher simple structure values attained by INDOQUAL can be interpreted in terms of a better discriminatory capability of INDOQUAL compared to MCA, at the cost of a loss of explained inertia.

8.7. A comparison of MCA and INDOQUAL with respect to discriminatory capability

Above, it has been shown that INDOMIX attains values of several simple structure criteria which are at least as high as those attained by optimally rotated PCAMIX solutions. The present section discusses a consequence of this result for the case where only qualitative variables are involved. It has been mentioned by Van der Burg (1988, p.171 ff) that MCA can be seen as a cluster technique. It will be shown here that MCA (that is PCAMIX applied to a set

of merely qualitative variables) finds components that discriminate the objects as well as possible in terms of *all* variables, whereas in INDOQUAL each component tends to discriminate the objects mainly in terms of a subset of the variables. For different components different subsets may be involved. As a consequence, INDOQUAL yields components that discriminate the objects better than MCA does, as will be explained below.

Each qualitative variable defines a set of disjoint groups of objects that fall in different categories of the qualitative variable. Hence for each variable one can, for each component, compute the group averages on that component. Now the variance of those group averages is called the "between groups variance", or, because the groups are defined by the categories, the "between categories variance". For both MCA and INDOQUAL it is readily verified that the solution for the matrix of component scores X is centered column-wise. Therefore, for every l, we have $\mathbf{x}_l = J\mathbf{x}_l$. As a consequence, the between categories variance of $\mathbf{x}_l$ with respect to the categories of variable j can be given as

$$\sigma^2_{B(jl)} = n^{-1} \sum_{g=1}^{m_j} f_g (x_l^{gj})^2 =$$

$$n^{-1}\, \mathbf{x}_l' G_j D_j^{-1} G_j' \mathbf{x}_l = n^{-1}\, \mathbf{x}_l' J G_j D_j^{-1} G_j' J \mathbf{x}_l = c_{jl}, \qquad (30)$$

where f_g denotes the number of objects in category g of variable j, x_l^{gj} denotes the average value of $\mathbf{x}_l$ in category g of variable j, see, for instance, Tenenhaus and Young (1985, p.98). Hence the loading of variable j on component l is equal to the between categories variance for component l, with respect to (the categories defined by) variable j. This between categories variance can also be considered as the amount of discrimination, provided by component l, between the objects that fall in different categories of variable j. Hence the loading c_{jl} indicates how strongly component l discriminates the objects in terms of the categories of variable j.

The above given interpretation of the loadings provides the basis for our statement that the INDOQUAL components discriminate the objects better than the MCA components do. As will be explained below, this difference between INDOQUAL and MCA is an immediate consequence of the fact that INDOQUAL provides a solution with higher simple structure than MCA does, at least in terms of the QMAX, OVERMAX and KURTMAX criteria, and,

typically, also in terms of the VMAX criterion.

The results of the previous section on differences between INDOQUAL and MCA in terms of simple structure criteria can be interpreted as follows. The fact that the INDOQUAL loadings have simple structure values that are higher than (or equal to) those of MCA, even after optimal rotation of the MCA solution, implies that INDOQUAL finds loadings that, overall, are more diverse than those resulting from MCA. The loadings are bounded between zero and one. Therefore, a higher simple structure in the loadings implies that more loadings tend to the extreme values, zero and one. Hence INDOQUAL tends to yield more extreme loadings than MCA does. From (30) it also follows that INDOQUAL yields more extreme between categories variances than MCA does. This implies that the INDOQUAL components discriminate the objects better in terms of the categories of *certain* variables (those with large between categories variances), and, at the same time worse in terms of other variables than MCA does. In this way, it can be said that INDOQUAL finds components each of which seek to discriminate the objects, to a larger extent than MCA does, in terms of (possibly different) subsets of variables. Because INDOQUAL seeks to discriminate the objects in terms of fewer variables than MCA does, INDOQUAL will succeed better in actually discriminating the objects. MCA tries to discriminate the objects as well as possible in terms of *all* variables. When certain variables define very different groupings of objects, the MCA components will tend to make compromises by discriminating the objects a little worse both in terms of the one variable and in terms of the other variable. On the other hand, INDOQUAL will optimally discriminate the objects in terms of either one of these "opposite" variables and will possibly discriminate the objects in terms of the other variable by means of a different component.

In addition, it can be said that a subset of variables that load high on a component consists of variables that are, at least in one respect, rather strongly related to each other. That is, if all variables in a subset of variables load high on a component, this component discriminates well the categories of each of these variables. This is only possible if the partitions (groupings) defined by the different variables are highly overlapping. The latter is another way of saying that the qualitative variables involved are highly related with respect to the partition of objects into groups that are

best discriminated by the component.

It can be concluded that INDOQUAL finds components that, overall, discriminate the objects better than MCA does, and that it does so by discriminating the objects in terms of the categories of subsets of variables that have highly overlapping partitions.

Above, it has been shown that INDOQUAL finds loadings that tend more to zero and one than those in MCA. However, this does not imply that each component of INDOQUAL *always* has loadings that are greater than those of MCA. Yet, in practice INDOQUAL often yields solutions with loadings that do not only have more simple structure than those of MCA, but that are also, for each component, higher than the highest MCA loadings for a corresponding MCA component. As a consequence, INDOQUAL yields a solution in which, with respect to each component, objects fall apart more clearly than in MCA into (denser) clusters of objects that represent the categories of those highly loading variables.

It can be concluded that INDOQUAL finds loadings that tend more to zero and one than those in MCA. In addition, it has been stated that, in practice, this phenomenon often leads to INDOQUAL components with clusters of objects that are denser and more separated than those (possibly) resulting from MCA. In the next section these phenomena of better "component–wise discriminatory capability" and of "component–wise clearer clustering" are illustrated by means of an example analysis.

8.8. An example analysis of empirical data

The empirical data to be analyzed in the present section has been given by Hartigan (1975, p.228). The data consists of 24 objects like screws and nails, that are classified according to 5 qualitative variables (Whether or not they have a Thread, what type of Head they have, what Indentation they have in the heads, what kind of Bottom they have, and whether or not they are made of Brass). In addition, their Length (in half inches) is measured, which is considered here as a qualitative variable with five categories (1 through 5 half inches). Although the data is of little practical interest, it serves to illustrate the clustering phenomenon, because the objects are well–described in terms of predefined clusters (those of screws, bolts, nails and tacks),

whereas this clustering does not refer directly to a qualitative variable in the analysis.

In the present section the MCA solutions and the INDOQUAL solutions for $r = 3$ will be compared. The MCA solution will be considered both before and after varimax rotation. Table 8.1 gives the MCA loadings before varimax rotation and after varimax rotation. It can be seen that the varimax rotation changes the loadings (mainly those of the fifth variable) only slightly, and that these changes lead to increasing simple structure, as expressed by the varimax and quartimax function values. It can also be verified that the amount of explained inertia is equal in the two solutions.

Table 8.1. *MCA loadings before and after varimax rotation.*

	before varimax rotation			after varimax rotation		
	comp.1	comp.2	comp.3	comp.1	comp.2	comp.3
Thread	0.93	0.02	0.00	0.95	0.00	0.00
Head	0.95	0.64	0.74	0.96	0.64	0.73
Head Ind.	0.94	0.67	0.08	0.95	0.74	0.00
Bottom	0.55	0.02	0.00	0.50	0.05	0.01
Length	0.29	0.82	0.69	0.24	0.78	0.79
Brass	0.06	0.03	0.46	0.09	0.00	0.47
MCA–inertia ($\sum_j \sum_l c_{jl}$)	7.90			7.90		
varimax function value	0.34			0.37		
quartimax function value	0.97			1.00		

Table 8.2 gives the loadings for INDOQUAL. Clearly, the components in the INDOQUAL solution and those in the rotated MCA solution have high loadings for the same variables, but those on the INDOQUAL components are higher. This is reflected by the fact that the varimax and quartimax function values are higher for INDOQUAL than for the rotated MCA solution. It can also be seen that the higher simple structure of INDOQUAL is obtained at the cost of a (small) loss in explained inertia compared to the MCA solution.

Table 8.2. *INDOQUAL loadings.*

	comp.1	comp.2	comp.3
Thread	0.99	0.00	0.00
Head	1.00	0.87	0.81
Head Ind.	0.99	0.31	0.02
Bottom	0.40	0.00	0.02
Length	0.17	0.91	0.82
Brass	0.06	0.00	0.24
MCA–inertia $(\sum_j \sum_l c_{jl})$		7.61	
varimax function value		0.45	
quartimax function value		1.04	

In both the INDOQUAL and the MCA solutions, the first component is highly correlated with the first three variables. These are the variables that are most important in distinguishing screws and bolts on the one hand from nails and tacks on the other hand. Therefore, the object scores on the first components of both solutions are "plotted" in Figure 8.1. Each of these plots is made in the form of a stem and leaf diagram in which the component scores of the objects are divided into 30 intervals. Because it is of interest to see how well the original clustering in the data appears in the solution, the objects are indicated by the letters T (tack), N (nail), S (screw) and B (bolt).

From inspection of Figure 8.1 it follows that the objects are clustered more clearly with respect to the first INDOQUAL component than with respect to the first MCA component. In addition, the original categories appear as partly separated clusters. That is, the nails and tacks now form one cluster. The bolts and screws form different clusters, but are not separated very much. With respect to the second and third components similar plots could be made, and one would again find a clearer clustering with respect to the INDOQUAL components than with respect to the MCA components. This demonstrates the phenomenon that the INDOQUAL components have a better discriminatory capability than MCA has, as was explained in the previous section.

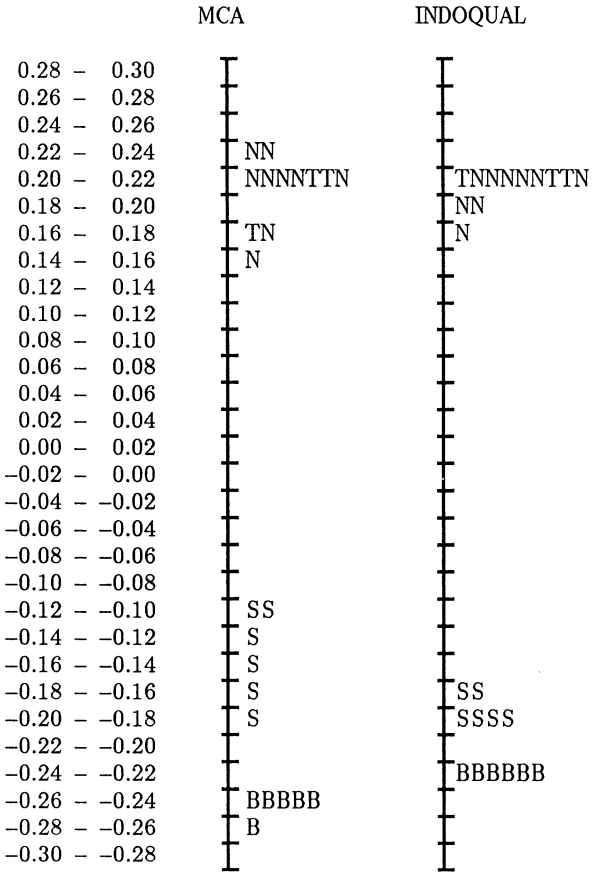


Figure 8.1. *Stem and leaf diagrams for the object scores on the first components for MCA and INDOQUAL.*

8.9. Discussion

In the present chapter methods have been described for representing mixtures of variables by component scores that optimize simple structure criteria. Two strategies have been proposed. The first one is based on first optimally accounting for the inertia in the variables, and then rotating these component scores (without loss of inertia accounted for) to optimal simple structure. The second strategy consists of finding component scores that

primarily optimize simple structure, but will still tend to explain a reasonable amount of inertia. This second strategy has a similar objective as the one that prevails in Projection Pursuit (e.g., Jones & Sibson, 1987), that is, that of revealing clustering in the data. Whether or not the methods proposed here might provide a useful alternative to Projection Pursuit methods is yet to be investigated.

Apart from providing methods to optimize simple structure by means of components for mixed variables, the present chapter describes an important difference between INDOQUAL and MCA. It has been shown that INDOQUAL has a greater discriminatory capability than MCA has. In addition, it has been explained why INDOQUAL tends to yield solutions in which the objects are more clearly clustered, per component, than in MCA solutions. This interpretation of INDOQUAL suggests comparing INDOQUAL with discriminant analysis techniques as well as cluster techniques for qualitative variables. As far as the latter is concerned, comparison with the newly developed GROUPALS technique (Van Buuren & Heiser, 1989) which combines K–means clustering procedures with optimal scaling of qualitative variables is of interest. It should be noted, however, that INDOQUAL has not been developed for purposes of discriminant or cluster analysis, but that the results presented here are merely additional properties of the method that attempts to combine the objectives of optimal representation of the objects (as MCA does) and optimal representation of the qualitative variables (as PCA of quantification matrices does).

In section 8.4 simple structure rotations have been proposed for PCAMIX solutions. Apart from simple structure rotation, an often used rotation technique in ordinary PCA is matching of the loadings to a given set of loadings. A procedure for rotating the MCA object coordinates such that the loadings optimally resemble a given set of loadings has recently been developed, and the program for it is presently being written. This offers new possibilities for comparing INDOQUAL and MCA solutions, because one may now compare the INDOQUAL solution to the rotation of the MCA solution the loadings of which optimally resemble those of the INDOQUAL solution. Furthermore, this matching procedure can be used to choose MCA solutions that optimally resemble an a priori given set of loadings, which is based on theory or on earlier results, for instance.

9. A COMPUTATIONAL SHORT-CUT FOR INDOMIX AND SOME PROPERTIES OF THE INDOMIX SOLUTION

9.1. Introduction

In the chapters 5, 6, 7, and 8 INDOQUAL and INDOMIX have been discussed. Both are based on the application of INDORT to a set of quantification matrices, $P_j = JG_jD_j^{-1}G_j'J$ for qualitative variables, and $Q_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'$ for quantitative variables, respectively, $j = 1,...,m$. In the present chapter it will be described how the solutions of these methods can be obtained. In addition, some particular properties of the solutions will be discussed.

An algorithm for INDORT, that is, for minimizing

$$\sigma(X,W_1,...,W_m) = \sum_{j=1}^{m} \| S_j - XW_jX' \|^2 \qquad (1)$$

over X, subject to $X'X = I_r$, has been given by Kroonenberg (1983, p.118). Ten Berge, Knol and Kiers (1988) have suggested an alternative algorithm for which monotone convergence is guaranteed if the matrices S_j are positive semi-definite (p.s.d.). Both algorithms are based on an iterative procedure in which an update for X is computed from the previous X and the set of S_j matrices. When these algorithms are used for INDORT on a set of quantification matrices for qualitative variables one is often faced with severe computational problems. That is, these algorithms are to be applied to a set of m matrices of order $n \times n$. Clearly, computation time increases rapidly as the sizes of the matrices increase. As a consequence, analyzing a set of qualitative variables measured on a *large* number of objects poses immense computational problems, both in terms of required memory and in terms of computation times.

Multiple Correspondence Analysis (MCA) has no such problems in handling very large numbers of objects. This is a consequence of the fact that the information that is essentially used in the computations is contained completely in the Burt-matrix, that is, the supermatrix containing the contingency tables for all pairs of variables, including the diagonal matrices of marginal frequencies. The number of objects in no way affects the size of

the Burt–matrix. Hence having a large number of objects in MCA does not give rise to computational problems.

Kiers (1989a) has shown that, as in MCA, the solution of INDOQUAL depends on the elements of the Burt–matrix only. This has been shown by means of describing the algorithm given by Ten Berge et al. (1988) completely in terms of elements of the Burt–matrix. In addition, he has shown that for determining the INDOMIX solution, again, only aggregate information is used during the computations of the solution. In this way, all computational problems for INDOQUAL or INDOMIX due to large sample sizes are resolved. The derivations and results from Kiers (1989a) are repeated here.

The new computational procedure is based on some results that are interesting in itself, because they can be used for deriving some properties of the solutions of INDOQUAL and INDOMIX. For instance, based on a procedure for weighting the objects that is developed here, the "distributional equivalence property" which is a well–known property of Correspondence Analysis, can be shown to hold for INDOQUAL as well. It will also be shown how missing data can be handled, again by applying weights to the objects. First, however, it will be shown that the Ten Berge et al. (1988) algorithm for INDORT applied to p.s.d. quantification matrices uses aggregate information only.

9.2. The Ten Berge, Knol, & Kiers algorithm for INDORT applied to quantification matrices

In the present section, the algorithm proposed by Ten Berge et al. (1988) for INDORT on p.s.d. matrices will be elaborated for the case where this algorithm is applied to quantification matrices S_j. In chapters 5 and 7, particular choices of these quantification matrices have been made. These quantification matrices (as well as most others described in chapter 3) can be decomposed as $S_j = U_j U_j'$. For instance, in the case of the quantification matrices chosen in chapters 5 and 7, $S_j = JG_jD_j^{-1}G_j'J$ for qualitative variables, and $S_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'$ for quantitative variables. Hence U_j can be chosen as $U_j = JG_jD_j^{-½}$ for qualitative variables, and $U_j = n^{-½}\mathbf{z}_j$ for quantitative variables. In the present section, an algorithm will be described for INDORT applied to any set of p.s.d. quantification matrices. In the next section, the implications of the particular choices for the quantification

matrices made in chapters 5 and 7 will be studied.

The algorithm proposed by Ten Berge et al. (1988) can be described as follows. Let X^u denote an update of X, then this update is given by

$$X^u = \sum_j S_j X W_j \left(\sum_k W_k X' S_k \sum_l S_l X W_l\right)^{-1/2}, \tag{2}$$

where $W_j = \text{Diag}\, X' S_j X$, for $j = 1,...,m$. Ten Berge et al. (1988) have shown that repeatedly updating X according to (2) monotonically decreases the function σ, or, equivalently, monotonically increases the function

$$f(X) = \sum_{j=1}^{m} \text{tr } (\text{Diag } X' S_j X)^2, \tag{3}$$

and because σ is bounded from below (and f from above), this procedure must converge to a stable function value. In (2) the term $\left(\sum_k W_k X' S_k \sum_l S_l X W_l\right)^{-1/2}$ can be computed as $K \Lambda^{-1/2} K'$ from the eigendecomposition $\left(\sum_k W_k X' S_k \sum_l S_l X W_l\right) = K \Lambda K'$.

The present elaboration of the Ten Berge et al. algorithm is based on substituting $S_j = U_j U_j'$ for S_j in the formula for the update X^u. This yields

$$X^u = \sum_j U_j U_j' X W_j \left(\sum_k W_k X' U_k U_k' \sum_l U_l U_l' X W_l\right)^{-1/2}. \tag{4}$$

Let $Y_j \equiv U_j' X$, then (4) can be rewritten as

$$X^u = \sum_j U_j Y_j W_j \left(\sum_k W_k Y_k' U_k' \sum_l U_l Y_l W_l\right)^{-1/2}. \tag{5}$$

Let $Z_j \equiv Y_j W_j$, $Z \equiv \begin{pmatrix} Z_1 \\ \vdots \\ Z_m \end{pmatrix}$, and $U \equiv (U_1 | \ldots | U_m)$, then $UZ = \sum_l U_l Z_l$, and (5) can be simplified as

$$\begin{aligned} X^u &= \sum_j U_j Z_j \left(\sum_k Z_k' U_k' \sum_l U_l Z_l\right)^{-1/2} \\ &= UZ \left(Z' U' U Z\right)^{-1/2}. \end{aligned} \tag{6}$$

It should be noted that $Z_j \equiv Y_j W_j = Y_j(\text{Diag } X' S_j X) = Y_j(\text{Diag } X' U_j U_j' X) = Y_j(\text{Diag } Y_j' Y_j)$ depends on Y_j only. As a consequence, the update X^u of X depends on the elements of U and Y only.

As has been explained above, the Ten Berge et al. (1988) algorithm for

INDORT is based on iteratively updating X by X^u described in (4). In order to compute the next update X^{uu} of X^u one uses again the expression in (4), but now with every X at the right-hand side replaced by X^u. Expression (4) can again be translated in an expression of the form of (6), when one defines $Y_j{}^u \equiv U_j'X^u$, $Y^u \equiv \begin{bmatrix} Y_1{}^u \\ \vdots \\ Y_m{}^u \end{bmatrix}$, $Z_j{}^u \equiv Y_j{}^u(\text{Diag } Y_j{}^{u\prime}Y_j{}^u)$, and $Z^u \equiv \begin{bmatrix} Z_1{}^u \\ \vdots \\ Z_m{}^u \end{bmatrix}$. Then the expression for X^{uu} is given by

$$X^{uu} = UZ^u\left(Z^{u\prime}U'UZ^u\right)^{-1/2}. \tag{7}$$

Clearly, the update X^{uu} for X^u depends on U and Z^u only. The latter itself depends on Y^u only. As a consequence, in order to compute X^{uu} one needs to have Y^u and U only. Obviously, Y^u can be computed from X^u. However, when n is large this computation is very cumbersome. Instead, Y^u is found directly from Z and U, as follows. Premultiplying (6) by U' yields

$$Y^u = U'X^u = U'UZ\left(Z'U'UZ\right)^{-1/2}. \tag{8}$$

From (8) it can be seen that Y^u can be computed by means of Z (which depends on Y only), and U. Matrix Z itself is given by the supermatrix containing $Z_j = Y_j(\text{Diag } Y_j'Y_j)$, $j = 1,\ldots,m$. In this expression Y_j depends on U_j and X only. Hence X^{uu} can be computed from X and U without, intermediately, computing X^u.

It follows from the above that the Ten Berge et al. algorithm for updating X can be modified such that it updates X *implicitly* in every iteration, while it only computes an update for Y *explicitly*. The only computations in which X is actually involved are the computation of Y from U and X at the start of the iterative procedure, and the computation of the solution for X from the final update of Y, after convergence of the iterative procedure. During the iterations computations are based on $U'U$ and Z only, while the matrix Z itself depends entirely on Y. Especially when the number of elements of $U'U$ is small compared to mn^2, the present procedure can gain a lot in computation time and memory space needed.

Ten Berge et al. (1988) have proven the monotone convergence of their algorithm by proving that $f(X^u) \geq f(X)$, $f(X^{uu}) \geq f(X^u)$, etc. In the procedure sketched above, X^u is not explicitly computed. Nevertheless, the corresponding

function value can be computed. Substituting $S_j = U_j U_j'$ for S_j in (3) we have

$$f(X^u) = \sum_{j=1}^{m} \text{tr } (\text{Diag } X^{u\prime} U_j U_j' X^u)^2$$

$$= \sum_{j=1}^{m} \text{tr } (\text{Diag } Y_j^{u\prime} Y_j^u)^2 \equiv g(Y^u). \qquad (9)$$

Because the algorithm has not been changed essentially by our procedure, it follows from $f(X^u) \geq f(X)$, $f(X^{uu}) \geq f(X^u)$ that $g(Y^u) \geq g(Y)$, $g(Y^{uu}) \geq g(Y^u)$, etc. That is, updating Y by means of (8) increases $g(Y)$ monotonically.

The algorithm proposed here can be summarized as follows.

Initialization:

Step 1. Choose a starting configuration X^0 for X.

Step 2. Compute $Y_j^0 = U_j' X^0$, $\qquad j = 1,...,m.$

Iterations:

Step 3. Compute $Z_j^i = Y_j^i (\text{Diag } Y_j^{i\prime} Y_j^i)$, $\qquad j = 1,...,m.$

Step 4. Compute $Y_j^{i+1} = U_j' U Z^i \left(Z^{i\prime} U' U Z^i\right)^{-1/2}$, $\qquad j = 1,...,m.$

Step 5. Evaluate $g(Y^{i+1}) = \sum_{j=1}^{m} \text{tr } (\text{Diag } Y_j^{i+1\prime} Y_j^{i+1})^2$.

If $g(Y^{i+1})$-$g(Y^i) > \varepsilon$, for some small value ε, then go to Step 3, else go to Step 6.

Determine the solution for the final W_j and X:

Step 6. Compute $W_j^{i+1} = (\text{Diag } Y_j^{i+1\prime} Y_j^{i+1})$ $\qquad j = 1,...,m.$

Step 7. Compute $Z_j^{i+1} = Y_j^{i+1} (\text{Diag } Y_j^{i+1\prime} Y_j^{i+1})$, $\qquad j = 1,...,m.$

Step 8. Compute $X^{i+2} = U Z^{i+1} \left(Z^{i+1\prime} U' U Z^{i+1}\right)^{-1/2}$.

The bulk of the computations is done in the Steps 3, 4, and 5. These steps only involve (parts of) the matrices $U'U$, of order $\sum_j r_j \times \sum_j r_j$, and Y and Z, both of order $\sum_j r_j \times r$, where r_j is the column-order of U_j. When n is large, the steps in which matrices of row- and/or column-orders n are involved are problematic. Such steps are Steps 1 and 2, and Step 8. In the present algorithm, these steps have to be done only once. In the original Ten Berge et al. algorithm every iteration step involves multiplication of matrices of

row– and/or column–orders n. For this reason, the procedure proposed here is much faster than the original procedure when n is large compared to $\sum_j r_j$.

Some details in the algorithm described above need further explanation. The first step of choosing a starting configuration for X can be done on several grounds. Carroll, Pruzansky and De Soete (1987) and more recently, Carroll, De Soete and Pruzansky (1989), have compared the performance of the INDSCAL algorithm when several different starting procedures are used. Their choices for starting matrices for X might be used. However, then one still has to compute Y^0 (Step 2), which might be cumbersome if n is very large. It is easier to omit Step 1, and replace Step 2 by choosing arbitrary matrices Y_j^0, but then a corresponding X^0 does not necessarily exist, and one cannot state that $g(Y^1) \geq g(Y^0)$. After the first iteration, however, a column–wise orthonormal X^i that corresponds to Y^i can be defined as $X^i = UZ^{i-1}\left(Z^{i-1\prime}U'UZ^{i-1}\right)^{-1/2}$. From $f(X^{i+1}) \geq f(X^i)$, for $i = 1,2,...$ it follows that $g(Y^{i+1}) \geq g(Y^i)$, for $i = 1,2,...$ That is, except for the first step monotone convergence is guaranteed.

In the actual iteration Steps 3 and 4 one uses Y and Z. It should be noted, however, that these need not be stored separately. That is, Y and Z can use the same memory location alternatively, because they are never needed at the same time.

This concludes the description of a modification of the Ten Berge et al. algorithm for INDORT applied to quantification matrices, in general. It can be seen that the algorithm essentially uses the elements of matrix $U'U$ only. In the next section, these sub–matrices will be described for the choices of quantification matrices made in chapters 5 and 7.

9.3. Implications for INDOMIX

In chapters 5 and 7, the quantification matrices have been chosen as $S_j = JG_jD_j^{-1}G_j'J$ for qualitative variables, and $S_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'$ for quantitative variables. Hence U_j can be chosen as $U_j = JG_jD_j^{-1/2}$ for qualitative variables, and as $U_j = n^{-1/2}\mathbf{z}_j$ for quantitative variables. In the present section, the sub–matrices $U_j'U_k$ of $U'U$ will be described for these choices of quantification matrices.

Three cases are to be distinguished. The first case is the case where both variables are qualitative. Then $U_j'U_k$ is given by

$$U_j'U_k = D_j^{-1/2}G_j'JG_kD_k^{-1/2}. \tag{10}$$

By substituting $J = (I - n^{-1}11')$ for J, and using $G_j'1 = D_j1$ we can rewrite (10) as

$$\begin{aligned} U_j'U_k &= D_j^{-1/2}G_j'(I - n^{-1}11')G_kD_k^{-1/2} \\ &= D_j^{-1/2}(G_j'G_k - n^{-1}G_j'11'G_k)D_k^{-1/2} \\ &= D_j^{-1/2}G_j'G_kD_k^{-1/2} - n^{-1}D_j^{1/2}11'D_k^{1/2}. \end{aligned} \tag{11}$$

The elements of $U_j'U_k$ can be expressed in terms of category frequencies and bivariate category frequencies, as follows. Let f_g be the frequency of category g of variable j, f_h the frequency of category h of variable k, and let f_{gh} be the number of objects that belong to both category g of variable j, and category h of variable k (called the bivariate frequency of these categories). It should be noted that when $j = k$, then $f_{gh} = f_g = f_h$ when $g = h$, and $f_{gh} = 0$, when $g \neq h$. From (11) it follows that the element (g,h) of $U_j'U_k$ is given by

$$[U_j'U_k]_{gh} = f_g^{-1/2}f_h^{-1/2}f_{gh} - n^{-1}f_g^{1/2}f_h^{1/2}. \tag{12}$$

Clearly, when $j = k$, (12) can be reduced to

$$[U_j'U_k]_{gh} = -n^{-1}f_g^{1/2}f_h^{1/2}, \qquad \text{when } g \neq h, \tag{13}$$

and

$$\begin{aligned} [U_j'U_k]_{gg} &= f_g^{-1/2}f_g^{-1/2}f_g - n^{-1}f_g^{1/2}f_g^{1/2} \\ &= 1 - n^{-1}f_g. \end{aligned} \tag{14}$$

When all variables are qualitative, all sub–matrices of $U'U$ can be computed as in (12), (13), and (14), hence, in that case, one can find the elements of the complete matrix $U'U$ in terms of the category frequencies and the bivariate frequencies. The category frequencies are given on the diagonal of the Burt–matrix (which contains all pairwise contingency tables), and the bivariate frequencies are given in the off–diagonal blocks of the Burt–matrix.

That is, the computations for the solution of INDOQUAL use the elements of the Burt–matrix only. The scores of the objects on the qualitative variables are needed only when the final matrix X is to be computed. However, when the number of objects is large, in general one is not interested in the complete matrix X. A more interesting representation of the results might in that case be the set of category centroids, as proposed by Kiers (1988), and also mentioned in section 5.5. That is, in order to summarize the solution for the objects, one might consider for each variable the means of the object coordinates in each of the categories as given by $D_j^{-1}G_j'X$. Obviously, these category centroids are computed easily from the final values in the matrices $Y_j = D_j^{-½}G_j'JX = D_j^{-½}G_j'X$ after convergence. Hence, if one is satisfied with category centroids only, one does not need to perform Step 8 in the algorithm, which is the only step in which the complete set of scores of objects on variables is needed. An important implication of this is that the present procedure makes possible the analysis of a Burt–matrix while one does not have information on the level of each observation unit. It also follows that the computational efficacy of the algorithm is in no way affected by the number of observation units on which the Burt–matrix is based.

In case some or all variables are quantitative, the sub–matrices of $U_j'U_k$ can be described as follows. If variable j is qualitative and variable k is quantitative, then

$$U_j'U_k = n^{-½}D_j^{-½}G_j'J\mathbf{z}_k = n^{-½}D_j^{-½}G_j'\mathbf{z}_k = n^{-½}D_j^{½}\mathbf{m}_{jk}, \tag{15}$$

where $\mathbf{m}_{jk}$ is the vector with the means of $\mathbf{z}_k$ in each of the categories of variable j.

If variables j and k are both quantitative, then

$$U_j'U_k = n^{-1}\mathbf{z}_j'\mathbf{z}_k = r_{jk}, \tag{16}$$

where r_{jk} is the product–moment correlation between variables j and k.

From the above it follows that the algorithm for INDOMIX involves again aggregate information only, that is, only category frequencies, bivariate frequencies, means of the quantitative variables in the categories of the qualitative variables (given in $\mathbf{m}_{jk}$), and product–moment correlations between quantitative variables. In the case where one has quantitative variables

only, the algorithm is based entirely on the product–moment correlations between the variables.

Apart from the fact that the algorithm for INDOMIX (and hence for the special case INDOQUAL) uses aggregate information only, inspection of the algorithm also yields some further information on the INDOMIX solution. That is, from (6), or, equivalently, from Step 8 of the algorithm, it follows that, after convergence of the algorithm, $X = UB$ for some matrix B of order $\sum_j r_j \times r$. That is, the columns of X form an orthonormal basis of a sub–space of the column–space of U. This implies that the maximum number of INDOMIX components is equal to the rank of U. It should be noted, however, that for this maximum number the X matrix does not necessarily provide a perfect INDORT fit. It gives the maximally attainable INDORT–fit in that case. One cannot generally increase the dimensionality of the INDOMIX solution until a perfect INDORT–fit is attained.

In the case of INDOQUAL the rank of U is equal to $\sum_j m_j - m$. Hence the maximal dimensionality for the INDOQUAL solution is $\sum_j m_j - m$. It is well–known that this is the maximal dimensionality for MCA as well. Moreover, as is readily verified, MCA also finds an orthonormal basis for a subspace of the column–space of U. Clearly, when $r = \sum_j m_j - m$, the columns of the INDOQUAL solution for X and those of the MCA solution for X both span the complete column–space of U. It follows that the solution of INDOQUAL for X is equal to the MCA solution for X, up to a possible rotation. Because the MCA solution is determined up to a rotation only, the INDOQUAL solution for $r = \sum_j m_j - m$ is also an MCA solution. It is readily verified that, when the MCA solution would be rotated such that it maximizes the quartimax criterion (see section 8.4) it would yield the INDOQUAL solution itself.

In the case of mixed variables a similar equivalence can be shown to hold for INDOMIX and PCAMIX. A more interesting case, however, seems to be the one where only quantitative variables are involved. In that case $U = n^{-1/2}Z$, hence $X = ZB$ for some $m \times r$ matrix B. That is, the components resulting from INDOMIX applied to quantitative variables, given in X, are linear combinations of Z, just as in ordinary PCA. Therefore, ordinary PCA and INDOMIX applied to quantitative variables can be seen as methods that both find linear combinations of the variables, but by optimizing different criteria. From the fact that the components from INDOMIX are linear combinations of the variables it follows that one can compute component scores

for new (or "supplementary") objects as well. This creates the possibility of cross–validation of one's results by applying the component–weights (in matrix B) resulting from an INDOMIX analysis in one sample to the variables of another sample, in order to study how the components are recovered in this second sample. Such a cross–validation can be useful for several purposes. For instance, one can use cross–validation to determine how sensitive an INDOMIX solution is to the particular sample on which the solution is based, as is done in section 10.4. Alternatively, one may want to study how strongly components that have been determined at one occasion are recovered at another occasion.

9.4. A further simplified algorithm for INDOQUAL

In section 5.4, it has been mentioned that INDOQUAL can be seen as the method that applies INDORT to the matrices $P_j = G_j D_j^{-1} G_j'$, $j = 1,...,m$, and eliminates the trivial axis, which is found consistently in this INDORT analysis. In the present section, a computational procedure for INDORT applied to the matrices $P_j = G_j D_j^{-1} G_j'$ will be given, that appears to be a little more simple than the one described for INDOQUAL above.

The INDORT analysis of the P_j matrices could be performed along similar lines as that of the S_j matrices. That is, one sets $U_j = G_j D_j^{-½}$. Again the algorithm from the previous section is used and the elements of the blocks of $U'U$ are computed as

$$[U_j'U_k]_{gh} = f_g^{-½} f_h^{-½} f_{gh}. \qquad (17)$$

Clearly, when $j = k$, $U_j'U_k$ can be written as

$$U_j'U_j = D_j^{-½} G_j' G_j D_j^{-½} = I_{m_j}. \qquad (18)$$

The algorithm of the previous section depends on the $U'U$ matrix whose elements are given as in (17) and (18). These elements are the elements of the Burt–matrix divided by the square roots of their associated marginal frequencies. It would be even simpler when the algorithm used the elements of the Burt–matrix itself. Therefore, in practice we use an algorithm which is a slight modification of the algorithm sketched above. Let $\tilde{U}_j \equiv G_j = U_j D_j^{½}$,

$\tilde{Y}_j \equiv D_j^{-1}G_j'X = D_j^{-½}Y_j$, and $\tilde{Z}_j \equiv \tilde{Y}_j(\text{Diag } \tilde{Y}_j'D_j\tilde{Y}_j) = D_j^{-½}Y_j(\text{Diag } Y_j'Y_j) = D_j^{-½}Z_j$. Then $\tilde{U}_j\tilde{Z}_j = U_jZ_j$, and $\tilde{U}\tilde{Z} = UZ$, with $\tilde{U}$ the horizontal supermatrix of the $\tilde{U}_j$ matrices, and $\tilde{Z}$ the vertical supermatrix of the $\tilde{Z}_j$ matrices. Then the iteration Steps 3, 4 and 5 are to be replaced by

Step 3': Compute $\tilde{Z}_j^{\,i} = D_j^{-½}Z_j^{\,i} = D_j^{-½}Y_j^{\,i}(\text{Diag } Y_j^{\,i\prime}Y_j^{\,i})$

$$= \tilde{Y}_j^{\,i}(\text{Diag } \tilde{Y}_j^{\,i\prime}D_j\tilde{Y}_j^{\,i}), \qquad j = 1,\ldots,m \ ;$$

Step 4': Compute $\tilde{Y}_j^{\,i+1} = D_j^{-½}Y_j^{\,i+1} = D_j^{-½}U_j'UZ^i\left(Z^{i\prime}U'UZ^i\right)^{-½}$

$$= D_j^{-1}\tilde{U}_j'\tilde{U}\tilde{Z}^i\left(\tilde{Z}^{i\prime}\tilde{U}'\tilde{U}\tilde{Z}^i\right)^{-½}, \qquad j = 1,\ldots,m \ ;$$

Step 5': Evaluate $g'(\tilde{Y}^{i+1}) = \sum_{j=1}^{m} \text{tr } (\text{Diag } \tilde{Y}_j^{\,i+1\prime}D_j\tilde{Y}_j^{\,i+1})^2$.

Steps 2, 6, 7, and 8 are to be adapted analogously. The advantage of using the present procedure over the original one is that in Step 4' one uses $\tilde{U}'\tilde{U}$ only which is exactly the Burt–matrix. It should be noted that the matrices D_j are the block–diagonal matrices of the Burt–matrix, that is, $D_j = \tilde{U}_j'\tilde{U}_j$. In this way it suffices to work with the elements of the Burt–matrix $\tilde{U}'\tilde{U}$ which are integers, instead of the elements of $U'U$, which are reals and hence require more memory space. This might enhance computer–efficiency, although the advantage is off–set by the disadvantage of a more complicated computation of the $\tilde{Z}_j$ matrices, and of the function to be evaluated. Incidentally, it should be noted that the matrix $\tilde{Y}_j$ is the matrix with category centroids. This matrix is now computed during the iterations, and it is available at once after convergence.

9.5. Applying weights to the objects by requiring distributional equivalence

In the previous sections it has been shown that the solution of INDOQUAL is based entirely on the elements of the Burt–matrix. Obviously, two objects with identical scores on all variables contribute in exactly the same way to the Burt–matrix. Hence two such objects can be seen as one "type of object" that occurs twice. Another way of putting this is that the object occurs with

weight 2 in the sample. In general, one can consider p objects that have the same scores on all variables as one object with weight p. The elements of the Burt–matrix are then computed as follows. Let object i have weight p_i, then the element of $[B_{jk}]_{gh}$, that is the bivariate frequency of category g of variable j, and category h of variable k, is given by $[B_{jk}]_{gh} = \sum_{(i \in GH)} p_i$, that is, the sum of weights of the objects that fall in category g of variable j, and category h of variable k. Clearly, it makes no difference whether all objects with the same scores on all variables are mentioned explicitly, with unit weights, or all p objects with equal scores are replaced by *one* object with weight p. This property is similar to the well–known property of "distributional equivalence" in Correspondence Analysis (CA). That is, when two rows (or columns) are proportional, CA on the matrix in which these rows (or columns) are replaced by their sum yields the same solution (e.g., Greenacre, 1984, p.95). This in turn comes down to CA on the same data with the two equal rows (or columns) replaced by one with a weight of 2. This property can readily be generalized to MCA, which, just as INDOQUAL, depends on the Burt–matrix only.

The property of distributional equivalence can be useful for INDOQUAL when a (small) number of profiles is given that are each observed many times in different frequencies. Because the program uses the Burt–matrix only and allows for differential weighting of the objects, such data can be analyzed without problem. In principle, the property of distributional equivalence also holds for INDOMIX, but it is rather unlikely that two objects have identical scores on quantitative variables. Using weights for the objects, however, opens an interesting possibility for modifying INDOQUAL and INDOMIX. That is, instead of using the weights only to summarize a number of equal profiles, one can make any choice of weights for the objects, for instance in order to down–weight the influence of certain objects that should not affect the solution unduly, or, conversely, to give them a high weight in order to let them dominate the solution to a certain extent. Another interesting application seems to be a kind of "fuzzy coding". That is, an object can be seen to belong to more than one category of a variable, to different extents. Such a fuzzy coding can be applied in INDOQUAL by replacing each object by a number of "sub–objects" (with weights summing to one) which all belong to different categories of the variable which is to be coded fuzzily (cf. Cazes, 1980, pp.391–392). The weights of the sub–objects indicate the extent to which

the object belongs to each of the categories. One particular application of such a fuzzy coding can be used in the case of missing data, as is explained in the next section. It should be noted that this type of fuzzy coding differs from the more usual definition of fuzzy coding (e.g., Van Rijckevorsel, 1987, p.104 ff). Usually, fuzzy coding comes down to replacing indicator matrices by so–called "pseudo–indicator matrices", which contain in each row a number of weights summing to one, indicating to what extent an object belongs to the category. Obviously, this second type of fuzzy coding can be incorporated in INDOMIX as well, because it simply pertains to a particular choice of quantification matrices, based on pseudo–indicator matrices.

9.6. Missing data

In the descriptions of INDOQUAL and INDOMIX no procedure has been described for handling missing data. Of course, one simple way of handling missing data is deletion of all objects for which scores on one or more of the variables are missing, often called "list–wise deletion". However, in case the number of variables is large, even with a small percentage of missing data, this procedure of "list–wise deletion" might come down to eliminating most (or even all) data. Therefore, alternative strategies are desirable.

For INDOQUAL, one particularly simple method for handling missing data is based on the procedure for fuzzy coding described in section 9.5. Suppose an object's score is missing for variable j only. Then, not knowing to which category this object belongs, one might consider this object to belong (to a certain extent) to all the categories of the variable. That is, each object is replaced by m_j "sub–objects" with certain weights adding up to 1, that belong each to a different category of variable j. The choice of weights to assign to the sub–objects might be inspired by various reasonings. A very simple procedure is to assign the weight m_j^{-1} to all sub–objects. However, if a category is rather infrequent, it is reasonable to assign a smaller weight to the sub–object falling in this category. This might be achieved by assigning the weight f_g/n to the sub–object that falls in category g, $g = 1,...,m_j$. Alternatively, one might have some information on the reason why an observation is missing. For instance, an observation on an object might be "missing" because the object in fact belongs partly to category g and partly to category h of a variable, but certainly not to any of the other categories.

Then introducing only two sub–objects (falling in categories g and h, respectively) seems to be called for, for instance with weights of 1/2 each. In all cases, the object scores for objects with missing data can be computed as the means of the object scores for the sub–objects belonging to the object concerned.

In INDOMIX the procedures described above can be used for handling missing observations on qualitative variables. In order to use such procedures for quantitative variables as well one is faced with the problem that, in principle, the number of available scores is infinite. In order yet to align the procedure for handling missing observations on quantitative variables as much as possible to the one for handling missing observations on qualitative variables, one might replace an object with a missing observation on a quantitative variable by as many sub–objects as there are scores that have been observed, with weights summing to one. It is readily verified that, when the weights are proportional to those of the observed scores on this variable, then the values of $\mathbf{z}_j'\mathbf{z}_k$ and $\mathbf{m}_{jk}$ are the same as those obtained by setting the missing observation to this variable equal to zero. Hence one can simply replace missing observations on quantitative variables by scores zero. Another way of interpreting this is by considering a missing observation to be replaced by the mean score on the variable concerned. This procedure in fact unstandardizes the variable, and it seems useful to standardize the variable again after this procedure.

Alternative procedures for handling missing data are possible as well (see Gifi, 1981, pp. 68–70, and, Meulman, 1982, for MCA). One of these is to replace the row in the indicator matrix for a missing observation on a qualitative variable by a row with zero elements only. Still another approach is to create one or more extra categories for missing observations on qualitative variables. Depending on the reason why an observation is missing one may choose for any of these options. None of them involves considerable adaptations of the computational procedures for INDOQUAL or INDOMIX.

9.7. Discussion

In the present chapter an algorithm has been described for the INDORT analysis of a set of p.s.d. quantification matrices. The algorithm described here is based on the algorithm proposed by Ten Berge et al. (1988). One of

the main problems with their algorithm, and hence with the one described here, is the fact that it is not guaranteed that the algorithm finds the global maximum of function f. Apart from using rational starts, as proposed by Carroll, Pruzansky, and De Soete (1987) and Carroll, De Soete, and Pruzansky (1989), the only (partial) remedy to this problem seems to be using several restarts.

The algorithm for INDORT applied to quantification matrices has been elaborated for the special cases of INDOQUAL and INDOMIX, but can clearly be useful for INDORT applied to many of the other quantification matrices discussed in chapter 3 as well. Furthermore, as has been discussed in chapter 4, INDORT is not the only three–way method that has been proposed to use for the analysis of a set of quantification matrices for qualitative variables. Marchetti (1988) has proposed to use Tucker's three–mode scaling (Tucker, 1972, see Kroonenberg, 1983, pp.52–53) and IDIOSCAL (Carroll & Chang, 1972) for the analysis of a set of quantification matrices. In addition, one might apply INDSCAL, that is, the unconstrained variant of INDORT, to such a set of quantification matrices. All these methods use algorithms that have computational problems when faced with m large $n \times n$ matrices. Currently, algorithms are being developed for these methods that, just as the algorithm described in the present chapter, need category frequencies and bivariate frequencies only.

An interesting implication of the fact that the INDOQUAL solution is based on the elements of the Burt–matrix only (that is, on category frequencies and bivariate frequencies) is that different data sets with the same Burt–matrix have essentially the same solution. That is, these solutions have the same W_j matrices, and the same category centroids, given by $D_j^{-1}G_j'X$. The only difference is to be found in the object coordinates. This reflects the situation in ordinary PCA, where the PCA of two sets of variables with the same correlation matrix yield the same loadings for the variables, although the component scores may differ.

PART III

ANALYSES OF EMPIRICAL DATA

10. EXPERIENCES WITH INDOQUAL AND INDOMIX

The present chapter contains several example analyses. Each of the examples have been chosen for a special purpose. The first example (section 10.2) deals with the classification of whales. As has been mentioned in section 8.7, both Multiple Correspondence Analysis (MCA) and INDOQUAL can be used as techniques for clustering objects on the basis of qualitative variables, while INDOQUAL yields a slightly better solution. This first example analysis has been treated in much detail, in order to demonstrate how one can interpret the results from an INDOQUAL analysis. In addition, a stability analysis has been performed on the MCA and INDOQUAL solutions for this data set.

In the second example (section 10.3) the results of an enquiry reported by Vegelius and Bäckström (1981) are analyzed with the purpose of finding the most important components underlying this enquiry. In this example, the emphasis is on the variables. It is demonstrated, however, that focusing on the variables *only*, as PCA of quantification matrices does, may be too limited. The results from an INDOQUAL analysis, which considers the objects and categories additionally, are discussed.

In section 10.4 an example is given of an INDOMIX analysis. The data from a survey on abortion and related issues (described in Gifi, 1981) are analyzed here. The analysis uses a special weighting option in order to control for background variables. The variables are a mixture of variables that can be considered as nominal and ordinal variables. The stability of the solution is assessed by means of cross–validation.

The fourth example (section 10.5) describes the analysis of a set of binary variables, reported by Van Zomeren and Van den Burg (1985). The purpose of this analysis is, again, to find certain subscales or clusters of variables. Some of the special possibilities of the analysis of binary variables have been used.

In order to demonstrate to what extent standardizing the variables can affect the INDOQUAL solution, the fifth example data set (section 10.6) has been analyzed both after standardizing the variables, and without standardizing the variables, the latter being the ordinary INDOQUAL procedure. The data consists of two binary and five nominal variables with

more than three categories. It appears that standardizing the variables results in a domination of the solution by the binary variables.

The sixth example (section 10.7) has been taken up in order to compare the results of an INDOQUAL solution with those of a TUCKALS–3 analysis of quantification matrices, as proposed by Marchetti (1988). The data have been analyzed by INDOQUAL with the same options as chosen by Marchetti (1988), and it is shown that the results are very similar.

Finally, the seventh data set has been analyzed to give an example of what may go wrong in blindly applying INDOQUAL to any data set of nominal variables (section 10.8). It is shown that, because the variables in the data set at hand (Sugiyama, 1975) do not form any clusters of closely related variables at all, as can be verified by means of studying the generalized correlations among them, each component of the INDOQUAL solution is determined by one and only one variable. Obviously, such a solution is not at all interesting from the point of view of data reduction.

As has been mentioned above, the stability of the solutions will be considered in two cases. In the next section, the methods will be described by means of which these aspects of the stability are assessed.

10.1. Assessing the stability of INDOQUAL and INDOMIX solutions

All methods for the multivariate analysis of qualitative and quantitative variables described here are usually applied to samples from a population. In principle, one attempts to find a description of the data that is valid for the larger population from which the sample is drawn. Observations on the complete population are rarely available. Hence it is important to know whether or not the sample is representative for the population. One way of determining this is to draw another sample. However, in practice this often is not possible. In order yet to determine how sensitive the results of an analysis are to the properties of the particular sample one may analyze the stability of the solution over deletion of certain observations, or, alternatively, one may cross–validate the results from one half of the sample by means of the other half of the sample. These two methods will be discussed briefly in sections 10.1.1, and 10.1.2, respectively.

10.1.1. Stability over deletion of certain observations (jackknifing)

The first procedure for studying how much the results of an analysis depend on the particular sample is based on alternatingly analyzing the data with one object left out. This procedure is called the "jackknife" technique (e.g., Miller, 1974), or simply "jackknifing". In INDOMIX (and hence in INDOQUAL) jackknifing will be done as follows. When the number of objects is small, the data will be analyzed n times, with each object deleted once. Each sample with one object deleted is called a jackknife. This procedure is called "complete jackknifing", because each object is eliminated once. When the number of objects is large, "random jackknives" will be used. That is, a fixed number of jackknives will be determined by leaving out one randomly determined object each time. Next, the results of the analyses of the different jackknives are compared to the results of the analysis of the original sample and to each other, in order to assess the stability of the original solution.

In both complete and random jackknifing one obtains a large number of results of INDOMIX analyses on the different jackknives. The basic elements of the jackknife solutions are the object scores and the loadings for the variables. The object scores, however, cannot all be compared over the different jackknife solutions, because in each jackknife solution partly different sets of objects are used. On the other hand, the measures called category centroids, which are based directly on the objects, can be compared across all different jackknife solutions. All jackknife solutions contain loadings and category centroids for the same variables and categories, respectively.

The purpose of the jackknifing procedure used here is to determine to what degree the separate jackknife solutions differ from each other and from the original solution. Hence one has to compare a large number of loadings and category centroids over all jackknives. As has been explained by De Leeuw and Meulman (1986), comparing different jackknife solutions in multidimensional scaling methods requires that the solutions be matched to each other, before one can compare solutions. Because the INDORT solution yields unique axes, such matchings are unnecessary for comparing results from different INDOMIX analyses.

A useful way of comparing the loadings and category centroids (both

denoted as "parameters" here) of different jackknife solutions with each other seems to be to compute the means and standard deviations of the jackknife parameters. The means over the jackknife studies can be compared with the observed scores in the original study in order to see if the jackknife results differ systematically from the original study. The standard deviations give information on the stability of the parameters over the different jackknives. The stability results provided here are by no means given for hypothesis testing. They merely serve to indicate how stable the solution is as a whole, and, specifically, how stable each of the individual parameters is.

The computation of an INDOMIX solution requires an iterative process, which is rather time consuming. In the jackknifing procedure sketched above, this procedure is to be repeated a number of times, that is, as many times as there are objects to be deleted. In addition, each INDOMIX analysis should be repeated a number of times in order to check whether or not the global optimum has been found. All this would require large computation times. However, the iterative process can be accelerated by using good start configurations. If deleting an object does not cause dramatic changes in the solution, it can be assumed that the solution of one jackknife will provide a good starting configuration for the iterative process for computing another jackknife solution. Therefore, the different jackknife solutions have been computed in this way, that is, using the solution from the previously computed jackknife in order to find the new jackknife solution.

10.1.2. Cross-validation via a split-half procedure

Apart from jackknifing, another procedure is considered for determining the dependence of a solution on particular characteristics of the sampled data. This procedure is "cross-validation" via a split-half procedure. That is, first, one randomly splits the data into two halves, and the first half of these is analyzed by means of INDOMIX. This analysis yields weights, to be explained later, for the categories of qualitative variables and for the quantitative variables. These weights are used to compute object scores for the data in the second half. These "pseudo" object scores are compared to the "original" object scores (resulting from INDOMIX on the second half) by inspecting the correlation between them. Finally, one can compute how well these object scores represent the data in the second half, by simply computing

the loadings of the object scores on the variables. Obviously, one may, in addition use the inverse procedure, that is, cross–validating the INDOMIX results of the analysis of the second half by applying the resulting weights to the variables in the first half.

The weights to be used for computing object scores from the variables are based on the following. It has been mentioned in section 9.3 that the INDOMIX object scores can be written as $X = UB$, for some matrix B of weights. These weights can be applied to any data set for which a U is given with columns referring to the same quantitative variables and/or categories of qualitative variables. In fact, these weights resemble the "component weights" in ordinary PCA, which are used to compute the component scores as linear combinations of the variables. In the present case these weights can be used to compute the object scores as linear combinations of the columns of U, that is, of the standardized quantitative variables and the columns of the transformed indicator matrices for the qualitative variables. It should be noted that the resulting object scores are not necessarily uncorrelated. Therefore, it is interesting to compute the correlation between these components as well.

10.2. The cetacea data: MCA and INDOQUAL as clustering techniques

Vescia (1985b) has collected data on 36 cetacea (whales, porpoises and dolphins) on the basis of zoological descriptions. The cetacea have been "measured" on 15 variables, describing morphological, osteological, and behavioral aspects of the animals under study. These variables have been described in detail by Vescia (1985b), as well as by Meulman (1986, pp.28–33), albeit in a different order. There were a few missing observations. Meulman (1986) has considered a missing observation on a variable as "falling in a different category". This could be justified by the fact that most of the (few) missing observations occurred systematically within one or two families of cetacea. That is, "missing" may be considered as a characteristic in its own right. In the present study missing data have been handled in the same way. The data have been analyzed essentially as they have been given (in their coded form) by Meulman (1986), except that one accidentally omitted white whale has been recovered, and the four errors that had emerged in the original data set, as pointed out by Vescia (1985b, p.13), have been corrected. One of the "missing" data was in fact based on a coding error. The data set analyzed

here is given in Table 10.1, where the corrected data are printed in bold face. As can be verified by comparing the MCA results reported by Meulman (1986) and those given here, the errors hardly affected the solution.

Table 10.1. *The cetacea data.*

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	family
1	6	1	1	1	2	3	1	4	5	2	3	1	3	4	baleen whale
1	6	1	1	1	2	3	1	4	2	2	3	1	3	4	baleen whale
1	2	1	1	4	1	3	1	4	1	2	3	1	3	4	baleen whale
1	2	1	1	1	4	4	1	4	4	1	3	1	3	4	grey whale
1	5	1	1	4	1	5	3	4	1	1	3	1	5	4	finback whale
1	5	1	1	4	4	5	3	4	4	1	3	1	5	4	finback whale
1	5	1	1	4	4	5	3	4	4	1	3	1	**5**	4	finback whale
1	1	1	1	2	1	1	1	1	1	**2**	3	4	5	1	sperm whale
1	1	1	1	1	1	1	1	1	1	**2**	1	4	2	1	sperm whale
2	3	2	3	3	1	1	2	3	1	2	2	3	3	1	beaked whale
2	3	2	3	3	1	1	2	3	1	2	2	3	5	1	beaked whale
2	2	2	3	3	1	1	2	3	1	2	2	3	5	1	beaked whale
1	4	2	3	3	1	1	2	3	1	2	2	3	3	1	beaked whale
1	2	2	3	3	1	1	2	3	4	2	2	3	5	1	beaked whale
1	2	2	2	3	3	2	1	2	5	2	2	2	3	1	dolphin
1	3	2	3	3	3	2	1	2	1	2	2	2	2	2	dolphin
1	4	2	1	3	4	2	1	2	4	2	2	2	2	1	dolphin
1	4	2	1	3	4	1	1	2	1	2	2	2	5	1	dolphin
1	3	2	3	3	3	2	1	2	1	2	2	2	3	2	dolphin
1	2	2	2	1	3	2	1	2	1	2	2	2	2	2	dolphin
2	4	2	1	3	1	2	1	2	2	2	2	2	4	2	dolphin
1	4	2	1	3	2	2	1	2	**4**	2	2	2	3	3	dolphin
1	4	2	1	3	4	2	1	2	2	2	2	2	5	1	dolphin
1	3	2	3	3	3	2	1	2	3	2	2	2	4	2	dolphin
1	3	2	3	3	3	2	1	2	1	2	2	2	4	2	dolphin
1	3	2	3	3	3	2	1	2	4	2	2	2	5	2	dolphin
1	2	2	3	3	3	2	1	2	1	2	2	2	2	2	dolphin
1	3	2	3	3	3	2	1	2	1	2	2	2	5	2	dolphin
1	2	2	1	1	1	2	1	2	4	2	2	2	5	2	porpoise
1	2	2	1	2	1	2	1	2	1	2	2	2	5	2	porpoise
2	4	2	1	1	2	2	1	2	3	1	2	3	3	2	white whale
2	**4**	**2**	**1**	**1**	**2**	**1**	**1**	**2**	**4**	**1**	**2**	**3**	**3**	**1**	white whale
2	3	2	4	2	2	2	1	3	1	1	1	5	1	2	river dolphin
2	3	2	4	2	2	2	1	1	1	1	1	5	1	2	river dolphin
2	3	2	4	2	2	2	1	3	2	1	1	2	1	2	river dolphin
2	3	2	4	2	2	2	1	3	2	1	1	5	1	2	river dolphin

The cetacea can be classified into several families. For each of the cetacea, the family name is given in Table 10.1. These families can be

grouped hierarchically into several classes. According to the theory of Grasse (1955, see also Vescia, 1985a, p.16; Meulman, 1986, p. 29) the classification given in Figure 10.1 can be made.

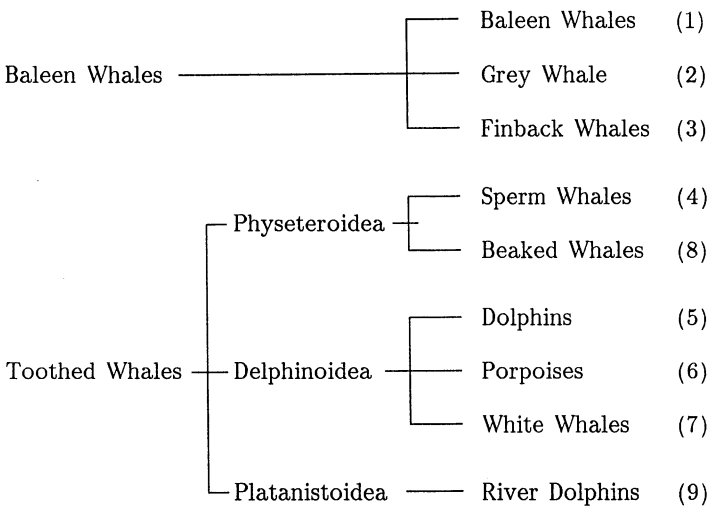


Figure 10.1. *The classification of cetacea according to Grasse.*

In the present study the data have been analyzed by means of both MCA and INDOQUAL. The former analysis has also been reported by Van der Burg (1985) and Meulman (1986, pp. 28–33). Both show that MCA (or Homogeneity analysis, as it is called there) is to a certain extent capable of distinguishing the original 9 families of cetacea, which is in accordance with Van der Burg's (1988) remark that MCA can be seen as a clustering technique. As has been described in chapter 8, INDOQUAL may be expected to yield clusters that are more compact and more clearly separated than the ones resulting from MCA. Therefore, the MCA solution will be compared to the INDOQUAL solution.

In the present analyses the two–dimensional solutions of MCA (as in the earlier analyses) and INDOQUAL are considered. This choice has been based partly on the fact that after the second component adding extra components did not increase much the amount of inertia accounted for by INDOQUAL. That is, the inertias accounted for by the one–, two–, three–, four–, and

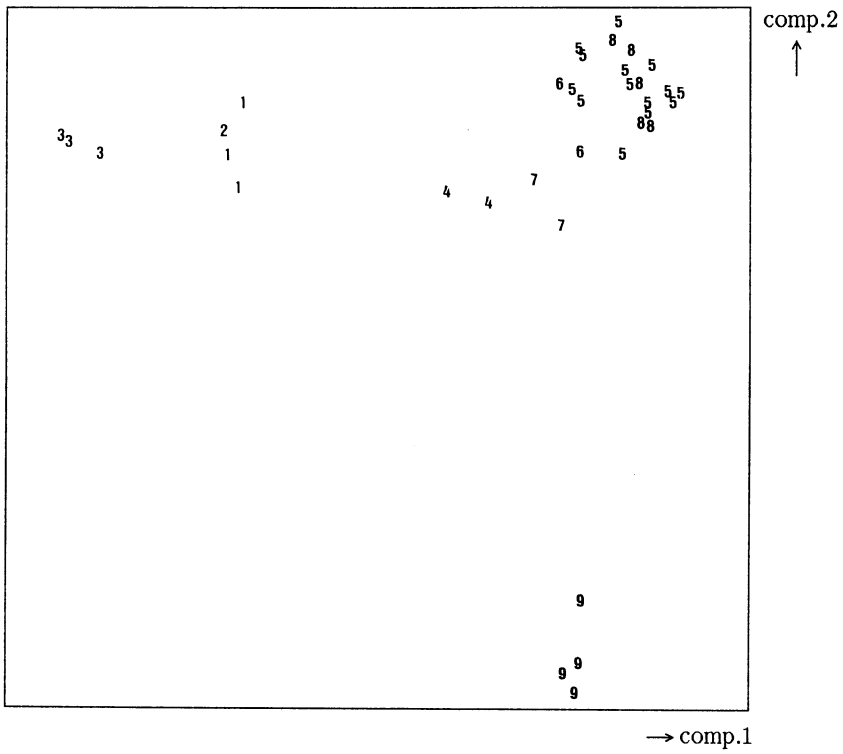


Figure 10.2. *The object coordinates resulting from MCA.*

five–dimensional INDOQUAL solutions are 6.9 (16%), 11.2 (26%), 14.1 (33%), 16.6 (38%), and 19.0 (44%), respectively. Clearly, the percentages are quite low, indicating that much of the information remains unaccounted for. One may be satisfied with a small amount of inertia accounted for, however, when one is interested in the most predominant relations between the variables. Because the two–dimensional INDOQUAL solution was well interpretable in terms of the original families of cetacea, and distinguished most of the families very well, it was decided to settle for this solution in the present study.

The two–dimensional MCA solution has been rotated to simple structure by means of a varimax rotation. The angle of rotation was 6.3°, which implies that, in this case, the principal axes solution of MCA had almost optimal simple structure.

The main purpose of the present analysis is the comparison of the object coordinates resulting from MCA and INDOQUAL. These have been plotted in Figures 10.2 and 10.3, respectively. The cetacea have been coded by their

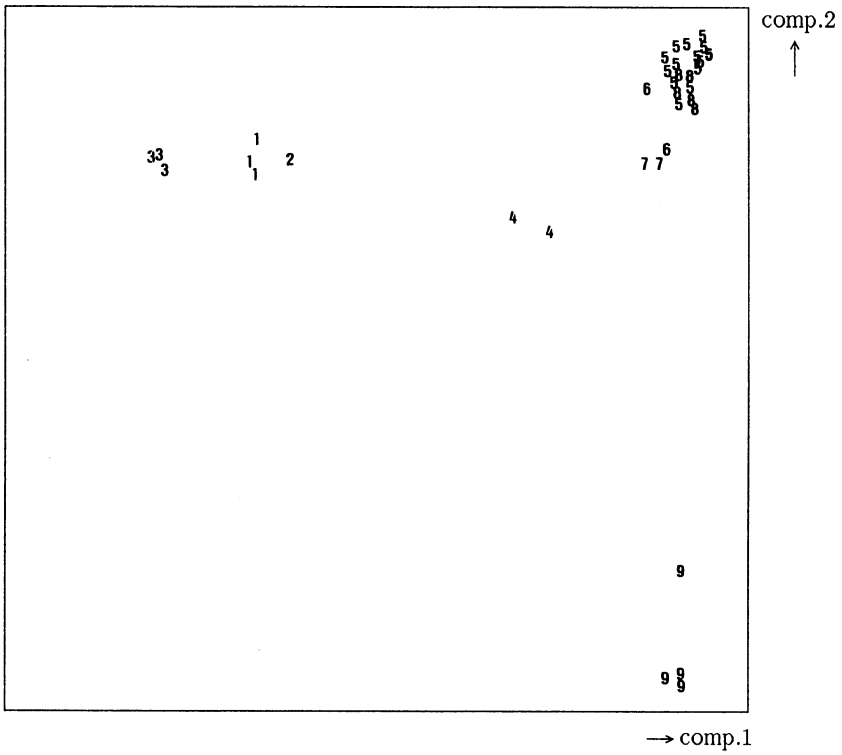


Figure 10.3. *The object coordinates resulting from INDOQUAL.*

family labels. In both plots, the families 1, 2, 3, 4, and 9, are well separated. The families 5, 6, 7, and 8 are not far apart. Moreover, the families 1, 2, and 3 are not far apart, and well separated from the other families, as would be expected from Grasse's classification. Similarly, the closeness of families 5, 6, and 7 is in agreement with Grasse's classification. Families 4 and 9 are both quite far from the other families, again in line with Grasse's classification. On the other hand, families 4 and 8 are separated much more than one would expect on the basis of Grasse's classification. The fact that family 8 is found within the cluster of families 5, 6, and 7 is unexpected as well. The results from MCA and INDOQUAL are globally equivalent. The differences between the MCA solution and the INDOQUAL solution are to be found in the details. It can be observed that in the INDOQUAL solution the families 1 and 2 are distinguished better than in the MCA solution. A similar result holds for families 5 and 6. The families 5 and 8 are completely intertwined in the MCA solution, whereas in the

INDOQUAL solution the members of family 8 are more to the bottom of the cluster with members of family 5, and "disturbed" only by a few of these members. Finally, the members of family 7 are closer to each other in the INDOQUAL solution than in the MCA solution. These slight differences all indicate that in the INDOQUAL solution the families of cetacea are distinguished better than in the MCA solution. Incidentally, it should be noted that the solution obtained by Meulman (1986, p.127) via a method that fits the distances between profiles directly yields results that are yet better interpretable in terms of the original families of cetacea. This was to be expected because her method has been designed specifically for the purpose of representing (distances between) objects. INDOQUAL on the other hand has been developed for the purpose of representing both objects and variables.

Table 10.2. *Loadings of the variables for MCA and INDOQUAL.*

	MCA + varimax		INDOQUAL	
	comp.1	comp.2	comp.1	comp.2
1. Neck	0.10	0.36	0.10	0.32
2. Form of the head	**0.80**	0.22	**0.79**	0.17
3. Size of the head	**0.82**	0.00	**0.86**	0.00
4. Beak	0.40	**0.95**	0.32	**0.96**
5. Dorsal fin	**0.77**	**0.69**	**0.76**	**0.71**
6. Flippers	0.27	0.49	0.19	0.46
7. Set of teeth	**0.94**	0.05	**0.97**	0.02
8. Longitudinal furrows	0.56	0.05	0.51	0.03
9. Blow hole	**0.93**	0.29	**0.97**	0.35
10. Color	0.13	0.17	0.09	0.13
11. Cervical vertebrae	0.17	0.45	0.13	0.42
12. Lachrymal & jugal bones	**0.87**	**0.81**	**0.91**	**0.84**
13. Head bones	**0.93**	**0.74**	**0.98**	**0.78**
14. Habitat	0.15	**0.92**	0.12	**0.93**
15. Feeding	**0.92**	0.13	**0.95**	0.08
sum of loadings	8.76	6.32	8.65	6.20
sums of squares of loadings	6.73	4.16	6.92	4.28

Next, the loadings resulting from MCA (followed by varimax rotation) are compared to those of INDOQUAL. These loadings are given in Table 10.2, with loadings $\geq$ .65 printed in bold face. The loadings from MCA and INDOQUAL are highly similar. Nevertheless, there is a tendency for loadings that are high in the MCA solution to correspond to even higher loadings in the INDOQUAL solution, and for loadings that are small in the MCA solution to correspond to even smaller loadings in the INDOQUAL solution. Table 10.2 also gives the inertia accounted for by the components in terms of the MCA–model (sums of loadings) and in terms of the INDOQUAL–model (sums of squares of the loadings).

From Table 10.2 it is clear which variables are important for the interpretation of each of the components. It is, however, not clear from these loadings how the variables are related to the components. In order to get more specific results it is useful to consider the category coordinates for all variables. These are computed as the means of the object scores of objects that belong to the category concerned. The category coordinates for the INDOQUAL solution are given in Table 10.3.

Using the results given in Tables 10.2 and 10.3, one can interpret the two components as follows. On the first component the variables 2, 3, 5, 7, 9, 12, 13 and 15 have high loadings (greater than 0.65). Therefore, this component will be interpreted in terms of the categories of these variables only. The first component can be interpreted by means of the following contrasts:

flat or convex heads	others	(variable 2)
very big heads	medium sized heads	(variable 3)
backward and falciform fin	other or no fin	(variable 5)
without teeth	with teeth	(variable 7)
double blow hole	single blow hole	(variable 9)
missing observation	other categories	(variable 12)
symmetrical headbones	asymmetrical headbones	(variable 13)
plankton eaters	others	(variable 15)

This first component is also the component that contrasts the super family of the Baleen Whales to that of the Toothed Whales, which is in agreement with the interpretation in terms of category coordinates given above.

On the second component the variables 4, 5, 12, 13, and 14 have high

loadings. In terms of the categories of these variables this component can be interpreted as follows:

narrow and long beak	other beak	(variable 4)
triangular fin	other or no fin	(variable 5)
linked lachrymal & jugal bones	independent bones	(variable 12)
missing observation	other categories	(variable 13)
river dwellers	others	(variable 14)

As is clear from Figure 10.3, this second component especially contrasts river dolphins, family 9, to the other cetacea, which corresponds well to the interpretation given above. This finishes the interpretation of the results of INDOQUAL on the cetacea data.

The stability of the INDOQUAL solution for the cetacea data has been examined by means of a jackknife study, as has been explained in section 10.1.1. That is, 34 analyses (both MCA with varimax rotation and INDOQUAL) have been performed on these data with each time one of the animals left out. Because leaving out the fourth or 22^{nd} cetacean would result in empty categories, no analyses have been performed with these animals left out. The 34 analyses resulted in sets of loadings and category coordinates that are comparable over the analyses. In order to assess the stability of all the parameters the standard deviations over the 34 jackknives have been computed. These are given, together with the means over the 34 jackknives, in Table 10.4 for the loadings, and in Table 10.5 for the category coordinates.

Comparing these results with those of the original analyses on the complete data set shows that the means of the loadings and category coordinates overthe jackknives are very close to the loadings and category coordinates in the original study. More importantly, the standard deviations over the jackknives are typically very small both in the MCA and INDOQUAL results. It can be seen that in particular the high and small loadings of INDOQUAL tend to be more stable than those in MCA, whereas the medium sized loadings tend to be more stable in MCA.

In order to indicate how one may translate the size of the standard deviation in a measure for stability, it is useful to remark that the standard deviation can be interpreted as a kind of mean deviation from the mean of the parameter value over the jackknife solutions. Instead, the true mean of

Table 10.4. *Means and standard deviations of the loadings over* 34 *jackknife studies.*

	MCA + varimax		INDOQUAL	
	comp.1	comp.2	comp.1	comp.2
var.1	0.10 (0.009)	0.36 (0.030)	0.10 (0.008)	0.32 (0.032)
var.2	0.80 (0.016)	0.22 (0.022)	0.79 (0.021)	0.17 (0.019)
var.3	0.82 (0.015)	0.00 (0.000)	0.86 (0.016)	0.00 (0.000)
var.4	0.40 (0.019)	0.95 (0.007)	0.32 (0.020)	0.96 (0.003)
var.5	0.77 (0.017)	0.68 (0.035)	0.76 (0.021)	0.71 (0.034)
var.6	0.28 (0.030)	0.48 (0.025)	0.20 (0.025)	0.46 (0.032)
var.7	0.94 (0.004)	0.05 (0.017)	0.97 (0.004)	0.02 (0.006)
var.8	0.56 (0.036)	0.05 (0.014)	0.50 (0.041)	0.03 (0.006)
var.9	0.93 (0.004)	0.29 (0.038)	0.97 (0.003)	0.36 (0.025)
var.10	0.13 (0.024)	0.18 (0.033)	0.09 (0.019)	0.14 (0.027)
var.11	0.17 (0.026)	0.44 (0.031)	0.13 (0.024)	0.42 (0.033)
var.12	0.87 (0.012)	0.81 (0.029)	0.91 (0.013)	0.85 (0.023)
var.13	0.94 (0.004)	0.74 (0.043)	0.98 (0.002)	0.77 (0.041)
var.14	0.15 (0.013)	0.92 (0.009)	0.12 (0.011)	0.93 (0.006)
var.15	0.91 (0.005)	0.13 (0.026)	0.95 (0.005)	0.08 (0.013)

Table 10.5. *Means and standard deviations of the category coordinates over* 34 *jackknife studies.*

	MCA + varimax		INDOQUAL	
var,cat	comp.1	comp.2	comp.1	comp.2
1,1	−0.19 (0.013)	0.37 (0.016)	−0.19 (0.013)	0.35 (0.019)
1,2	0.50 (0.019)	−0.97 (0.060)	0.51 (0.018)	−0.92 (0.061)
2,1	−0.27 (0.055)	−0.16 (0.120)	−0.30 (0.050)	−0.32 (0.054)
2,2	0.04 (0.047)	0.51 (0.021)	0.01 (0.050)	0.41 (0.022)
2,3	0.65 (0.021)	−0.61 (0.049)	0.58 (0.020)	−0.53 (0.045)
2,4	0.35 (0.022)	0.37 (0.033)	0.50 (0.018)	0.37 (0.022)
2,5	−2.42 (0.093)	0.05 (0.026)	−2.31 (0.087)	0.06 (0.012)
2,6	−1.57 (0.087)	−0.06 (0.051)	−1.78 (0.077)	0.04 (0.014)
3,1	−1.56 (0.056)	0.02 (0.023)	−1.60 (0.056)	−0.01 (0.011)
3,2	0.52 (0.017)	−0.01 (0.008)	0.53 (0.018)	0.00 (0.004)
4,1	−0.63 (0.027)	0.17 (0.022)	−0.57 (0.028)	0.17 (0.017)
4,2	0.55 (0.032)	0.73 (0.047)	0.50 (0.021)	0.65 (0.032)
4,3	0.70 (0.022)	0.52 (0.025)	0.61 (0.020)	0.55 (0.024)
4,4	0.47 (0.020)	−2.71 (0.140)	0.49 (0.019)	−2.73 (0.151)
5,1	−0.47 (0.066)	0.09 (0.034)	−0.47 (0.069)	0.12 (0.025)
5,2	0.31 (0.026)	−1.80 (0.097)	0.34 (0.028)	−1.84 (0.097)
5,3	0.59 (0.019)	0.53 (0.020)	0.58 (0.019)	0.54 (0.024)
5,4	−2.19 (0.090)	0.12 (0.015)	−2.17 (0.088)	0.09 (0.009)

Table 10.5. *Means and standard deviations of the category coordinates over* 34 *jackknife studies* (*continued*).

	MCA + varimax		INDOQUAL	
var,cat	comp.1	comp.2	comp.1	comp.2
6,1	−0.00 (0.055)	0.29 (0.048)	−0.03 (0.047)	0.23 (0.026)
6,2	−0.06 (0.053)	−1.19 (0.050)	−0.03 (0.055)	−1.14 (0.056)
6,3	0.70 (0.024)	0.55 (0.029)	0.60 (0.021)	0.64 (0.028)
6,4	−0.95 (0.102)	0.37 (0.031)	−0.79 (0.099)	0.30 (0.023)
7,1	0.34 (0.038)	0.34 (0.070)	0.34 (0.032)	0.23 (0.034)
7,2	0.53 (0.019)	−0.18 (0.034)	0.54 (0.018)	−0.13 (0.018)
7,3	−1.55 (0.083)	0.07 (0.043)	−1.77 (0.076)	0.10 (0.012)
7,4	−1.62 (0.058)	0.15 (0.019)	−1.59 (0.065)	0.09 (0.013)
7,5	−2.42 (0.093)	0.05 (0.026)	−2.31 (0.087)	0.06 (0.012)
8,1	0.15 (0.023)	−0.10 (0.018)	0.14 (0.022)	−0.08 (0.010)
8,2	0.63 (0.030)	0.53 (0.074)	0.57 (0.022)	0.42 (0.031)
8,3	−2.42 (0.093)	0.05 (0.026)	−2.31 (0.087)	0.06 (0.012)
9,1	−0.04 (0.062)	−1.03 (0.256)	−0.06 (0.066)	−1.17 (0.198)
9,2	0.50 (0.020)	0.44 (0.024)	0.54 (0.018)	0.49 (0.023)
9,3	0.58 (0.022)	−0.67 (0.106)	0.55 (0.019)	−0.73 (0.088)
9,4	−1.93 (0.071)	0.07 (0.013)	−1.97 (0.072)	0.08 (0.008)
10,1	0.23 (0.035)	0.02 (0.046)	0.18 (0.031)	−0.00 (0.036)
10,2	0.06 (0.080)	−0.96 (0.148)	0.07 (0.084)	−0.84 (0.135)
10,3	0.58 (0.079)	0.12 (0.107)	0.54 (0.034)	0.36 (0.077)
10,4	−0.50 (0.069)	0.36 (0.022)	−0.37 (0.063)	0.31 (0.016)
10,5	−0.54 (0.294)	0.45 (0.126)	−0.63 (0.283)	0.37 (0.080)
11,1	−0.66 (0.058)	−1.07 (0.049)	−0.58 (0.059)	−1.05 (0.052)
11,2	0.25 (0.022)	0.41 (0.021)	0.22 (0.023)	0.40 (0.021)
12,1	0.34 (0.026)	−2.20 (0.118)	0.35 (0.029)	−2.25 (0.118)
12,2	0.53 (0.018)	0.46 (0.020)	0.54 (0.018)	0.48 (0.022)
12,3	−1.74 (0.064)	0.05 (0.014)	−1.78 (0.064)	0.03 (0.009)
13,1	−1.93 (0.071)	0.07 (0.013)	−1.97 (0.072)	0.08 (0.008)
13,2	0.54 (0.020)	0.34 (0.041)	0.55 (0.019)	0.38 (0.034)
13,3	0.51 (0.030)	0.34 (0.066)	0.52 (0.021)	0.32 (0.030)
13,4	−0.27 (0.055)	−0.16 (0.120)	−0.30 (0.050)	−0.32 (0.054)
13,5	0.46 (0.021)	−2.82 (0.155)	0.48 (0.019)	−2.87 (0.163)
14,1	0.47 (0.020)	−2.71 (0.140)	0.49 (0.019)	−2.73 (0.151)
14,2	0.45 (0.039)	0.46 (0.046)	0.40 (0.031)	0.42 (0.053)
14,3	−0.27 (0.058)	0.26 (0.027)	−0.30 (0.055)	0.27 (0.017)
14,4	0.76 (0.031)	0.37 (0.044)	0.63 (0.022)	0.60 (0.034)
14,5	−0.26 (0.051)	0.35 (0.027)	−0.20 (0.047)	0.31 (0.021)
15,1	0.35 (0.029)	0.43 (0.049)	0.39 (0.026)	0.32 (0.028)
15,2	0.57 (0.021)	−0.38 (0.043)	0.54 (0.019)	−0.30 (0.028)
15,3	0.31 (0.027)	0.44 (0.031)	0.54 (0.019)	0.37 (0.025)
15,4	−1.93 (0.071)	0.07 (0.013)	−1.97 (0.072)	0.08 (0.008)

absolute deviations from the mean parameter value might have been used. The standard deviation has been chosen here because it is more sensitive to extreme deviations than the mean absolute deviation. When the standard deviation is small, this does not only imply that the deviations are small on the average, but also that extreme deviations from the mean parameter value do not occur.

It can be concluded that both MCA and INDOQUAL yield two–dimensional solutions that can be interpreted very well. The solutions differ slightly in that the INDOQUAL solution tends to give slightly more compact and more clearly separated clusters of cetacea families. In addition, most parameters in both solutions are very stable.

10.3. An enquiry about religion: Components analysis of nominal variables

Vegelius and Bäckström (1981) have analyzed the results of an enquiry among 118 theological students. The enquiry contained 24 questions on social and religious background (demographic variables), religious activities, plans for the future, and attitudes towards miscellaneous issues. All variables are considered as nominal variables. The variables are mentioned in Table 10.6, for the categories of the variables the reader is referred to Vegelius and Bäckström. Vegelius and Bäckström (1981) performed a PCA on the matrix of Tschuprow's T^2 coefficients among the variables. This method has been explained in section 4.1. Their analysis yielded 8 eigenvalues greater than one, and for this reason they reported 8 components. The loadings for the variables have been rotated to simple structure by means of a varimax rotation. This solution had a clear simple structure, and the components were well interpretable. In Table 10.6 this solution is repeated. Only the loadings greater than .40 (in the absolute sense) are reported.

As has been explained in chapter 1, PCA on generalized correlation coefficients is rather limited in that it provides only loadings for the variables, without a representation for the objects (i.e., students) and for the categories of the variables. For this reason, the data have been

Table 10.6. *Loadings from the PCA solution reported by Vegelius and Bäckström.*

Variables	Components							
	1	2	3	4	5	6	7	8
1. Sex								
2. Year of birth								.67
3. Father's social group							.75	
4. Father's education							.75	
5. Mother's social group			.79					
6. Mother's education			.78					
7. Marital status								.68
8. Future degree					.81			
9. Future job					.82			
10. Father's religious observance	.62					.41		
11. Father's denomination	.85							
12. Father's participation in church worship						.69		
13. Mother's participation in church worship						.69		
14. Mother's religious observance	.62							
15. Mother's denomination	.82							
16. Student's denomination						-.43		.40
17. Student's participation in church worship		.70						
18. Student's praying		.65						
19. Student's participation in Communion		.64						
20. Student's reading the Bible		.53						
21. Political attitude								
22. Reaction to the Jesus Movement				.77				
23. Future of the Jesus Movement				.78				
24. Place of childhood and youth								

reanalyzed* here both by means of MCA and INDOQUAL. In contrast to Vegelius and Bäckström (1981) only two components have been retained in the present analyses. This choice of dimensionality was based partly on a comparison of the one–, two–, three–, and four–dimensional INDOQUAL solutions. The inertia accounted for by these solutions was 5.2 (6%), 8.3 (10%), 9.9 (12%), and 11.4 (14%), respectively. The percentages of inertia accounted for are very small, but, again it should be noted that in order to extract the most interesting information one may be satisfied with a solution that accounts only for a small part of the inertia of the data. The INDOQUAL solutions appeared to be nested approximately. For instance, the first two components in the three– and

* The author is obliged to Anders Bäckström and Jan Vegelius who kindly made the original data available.

four–dimensional solutions were closely related to the components in the two–dimensional solution. The third and fourth components were not very interesting in that they only pertained to two variables each. Obviously, the two–dimensional solution does not describe the data at hand exhaustively. However, it does seem to capture the most interesting relations between the variables in the enquiry. For this reason, in the sequel only the two–dimensional INDOQUAL solution is reported. For reasons of comparability the same dimensionality has been chosen for the MCA solution.

The (unrotated) loadings of the MCA solution are reported in Table 10.7.

Table 10.7. *Variable loadings resulting from MCA, MCA with varimax, and INDOQUAL.*

	MCA		MCA with varimax		INDOQUAL	
	1	2	1	2	1	2
1.	0.22	0.00	0.13	0.10	0.10	0.10
2.	0.04	0.22	0.08	0.18	0.04	0.10
3.	0.04	0.08	0.05	0.07	0.03	0.04
4.	0.10	0.28	0.24	0.14	0.15	0.06
5.	0.07	0.07	0.10	0.05	0.08	0.03
6.	0.04	0.10	0.11	0.03	0.05	0.02
7.	0.10	0.12	0.06	0.15	0.04	0.08
8.	0.31	0.17	0.04	**0.44**	0.02	**0.52**
9.	**0.48**	0.27	0.14	**0.62**	0.09	**0.67**
10.	**0.69**	0.20	**0.86**	0.04	**0.93**	0.02
11.	**0.70**	0.21	**0.85**	0.05	**0.93**	0.04
12.	**0.73**	**0.40**	**0.84**	0.28	**0.90**	0.10
13.	**0.71**	**0.40**	**0.82**	0.29	**0.87**	0.11
14.	**0.68**	0.16	**0.81**	0.04	**0.90**	0.02
15.	**0.62**	0.13	**0.70**	0.05	**0.80**	0.02
16.	0.06	0.13	0.06	0.14	0.05	0.11
17.	**0.60**	0.36	0.19	**0.76**	0.11	**0.86**
18.	**0.48**	0.36	0.08	**0.76**	0.04	**0.87**
19.	**0.54**	0.35	0.16	**0.73**	0.11	**0.86**
20.	**0.49**	0.29	0.22	**0.56**	0.14	**0.60**
21.	0.24	0.10	0.08	0.26	0.08	0.18
22.	0.19	0.01	0.13	0.06	0.09	0.07
23.	0.11	0.01	0.08	0.04	0.06	0.03
24.	0.08	0.05	0.06	0.07	0.05	0.04
sum	8.32	4.47	6.89	5.91	6.66	5.55
sum of squares	4.49	1.20	4.27	2.93	4.85	3.42

The MCA solution has been rotated to simple structure by means of varimax. The axes were rotated over an angle of 38°. The MCA loadings after rotation are reported in Table 10.7 as well. Finally, Table 10.7 contains the loadings resulting from the INDOQUAL solution, with loadings $\geq .4$ printed in bold face.

From Table 10.7 it is clear that the unrotated and rotated MCA solutions lead to different interpretations of the two components. That is, in the unrotated MCA solution many variables have rather high loadings ($\geq .4$), while on the second component only the variables 12 and 13 have high loadings. The first component can be interpreted as a rather general "Religious behavior of parents and student" component. The second component, on the other hand, can easily be interpreted as "Parent's participation in church worship", but this component does not capture much of the information in the data.

After varimax rotation the picture is quite different. Now only the variables 10, 11, 12, 13, 14, and 15 have high loadings on the first component, and the variables 8, 9, 17, 18, 19, and 20 have high loadings on the second component. The first component can be interpreted as "Religious behavior of parents", and the second component as "Student's religious behavior and plans". The components resulting from the rotated MCA solution seem to be much more interesting than those of the unrotated MCA solution.

The loadings of the two-dimensional INDOQUAL solution have been reported in Table 10.7 also. The components can be interpreted in much the same way as the components of the MCA solution after varimax rotation, because loadings greater than or equal to .4 are found in both analyses for the same variables. Nevertheless, the solutions differ systematically. It can be verified that loadings above .4 in the MCA solution correspond to higher loadings in the INDOQUAL solution, and, in all but one case, loadings below .4 in the MCA solution correspond to smaller loadings in the INDOQUAL solution. This clearly demonstrates that INDOQUAL tends to find more extreme loadings than MCA, even after varimax rotation. For the interpretation this does not make a lot of difference, however. The INDOQUAL components can be said to be interpretable in the same way as the rotated MCA components, but in the INDOQUAL solution this interpretation is brought out more clearly.

On the basis of the loadings alone, one can give only a rather global interpretation of the components by means of studying the loadings of the variables on the components. Because the variables are nominal variables, it

is of interest to see how these components can be interpreted in terms of the categories of the variables as well. In Table 10.8 the category coordinates (category centroids of object coordinates) from the INDOQUAL solution are given only for the variables with loadings greater than .4.

With these category coordinates the components of the INDOQUAL solution can be interpreted as follows. Component 1 (called "Religious behavior of parents" above), contrasts, for both parents, religious observance versus no religious observance, denomination to a certain church versus no answer, and, much versus little participation in church worship. Clearly, religious behavior can be further interpreted as "amount of participation in religious

Table 10.8. *INDOQUAL category coordinates of the variables with the highest loadings.*

	comp.1	comp.2
var.8: Future Degree		
Candidate of Theology	-0.09	-0.51
Bachelor of Arts	0.31	1.39
Other degree	-0.04	-0.16
Don't know	-0.06	-0.08
var.9: Future Occupation		
Clergyman	-0.14	-0.51
Teacher	0.38	1.43
Deacon	0.52	-0.55
Other occupation	-0.45	-0.40
Don't know	-0.14	-0.35
var.10: Father's religious observance		
Yes	-1.08	-0.16
No	0.85	0.10
Don't know	0.86	0.27
var.11: Father's denomination		
Church of Sweden	-1.07	-0.29
Other denomination	-1.13	0.21
No answer	0.85	0.12
var.12: Father's participation in worship		
Never	0.91	0.46
At least once a year	0.83	0.23
At least six times a year	0.60	-0.53
At least once a month	-1.00	-0.20
At least once a week	-1.14	-0.12
Don't know	0.82	-0.45

Table 10.8. *INDOQUAL category coordinates of the variables with the highest loadings* (*continued*).

	comp.1	comp.2
var.13: Mother's participation in worship		
Never	0.89	0.63
At least once a year	0.90	0.21
At least six times a year	0.72	-0.46
At least once a month	-0.94	-0.18
At least once a week	-1.13	-0.13
Don't know	0.48	0.47
var.14: Mother's religious observance		
Yes	-1.00	-0.15
No	0.90	0.11
Don't know	0.89	0.35
var.15: Mother's denomination		
Church of Sweden	-0.85	-0.17
Other denomination	-1.02	0.01
No answer	0.89	0.13
var.17: Student's participation in worship		
Not for years	0.37	2.35
At least once a year	0.65	1.95
At least six times a year	0.52	0.70
At least once a month	0.16	-0.47
At least once a week	-0.28	-0.44
var.18: Student's praying		
Today or yesterday	-0.10	-0.39
One week ago at most	0.38	-0.42
One month ago at most	0.07	2.15
Not for years	0.46	2.19
Don't know	0.63	2.29
var.19: Student's participation in Communion		
One week ago	-0.19	-0.47
One month ago	-0.02	-0.37
Six months ago	0.72	-0.47
One year ago	-1.20	0.15
Not for years	0.42	2.03
Don't know	0.88	0.36
var.20: Student's reading the Bible		
Today or yesterday	-0.23	-0.46
One week ago	0.27	-0.04
One month ago	0.51	0.79
Six months ago	-0.89	2.71
Not for years	0.74	1.84
Don't know	1.07	-0.97

activities". On the basis of this interpretation one might suspect that the category "No answer" of variables 11, and 15, in fact means "No member of any denomination". The second component (called "Student's religious behavior and plans" above) can be interpreted as the contrast between students preparing for a clerical profession versus a profession as a layman (variables 8 and 9), as well as the "amount of the student's participation in religious activities" (variables 17–20). Again, interpretation on the basis of category coordinates complements the previous interpretation in terms of the variables in a useful way.

Finally, the results reported here are compared to those of Vegelius and Bäckström (1981). It can be seen that the components 1 and 6 in the latter solution are practically completely comprised in the first component in the INDOQUAL solution, and that the components 2 and 5 in the Vegelius and Bäckström solution are summarized well by the second INDOQUAL component. This indicates that the Vegelius and Bäckström solution provides components that could well be combined into fewer components. On the other hand, the third, fourth, seventh and eighth components in the Vegelius and Bäckström solution do not reappear in the INDOQUAL solution. This may be a consequence of the choice of a rather small dimensionality. In fact, the third component of the three–dimensional INDOQUAL solution is closely related to the seventh component in the Vegelius and Bäckström solution, and slightly related to their third component as well. These and higher components have not been considered in the INDOQUAL solution reported here.

10.4. The abortion survey: Components analysis of mixed variables

Gifi (1981, pp. 357–360) has described data from a survey among 575 respondents on attitudes with respect to abortion, capital punishment, euthanasia, and sexual freedom. In addition, the scores on several background variables were available. These data have been analyzed by Gifi by means of several options of the PRINCALS program. In the present section part of these data* is analyzed by means of INDOMIX, and the solution will be compared to that of PCAMIX.

* The author is obliged to Jacqueline Meulman who kindly made the data available.

The variables on capital punishment are hardly related to the other variables, as was found by Gifi (1981), and have therefore been excluded from the analysis, just as has been done by Gifi. In contrast to what has been done by Gifi, however, no other variables have been left out of the analysis. The data set contained many missing data. Because the present version of the INDOMIX program does not allow for missing data, all respondents with one or more missing data (22 %) have been deleted. The remaining sample contained 446 respondents. Of the 34 variables that remained after deletion of the capital punishment variables, two variables have been recoded as follows: the eleven categories of the variable "Present profession or job" (FUN) have been reduced to nine categories by merging the categories "managerial, more than ten employees" and "managerial, less than ten employees", and the categories "free profession" and "independent farmer"; the seven categories of the variable "degree of urbanization" (URB) have been reduced to five categories by merging the three categories "Amsterdam", "Rotterdam", and "The Hague". The other variables have been analyzed in their original form, as described by Gifi (1981). They are labelled here (as in Gifi) as A1 through A16 (questions concerning abortion), EU1 through EU5 (questions concerning euthanasia), SF1 through SF5 (questions concerning sexual freedom), SEX (male, female), AGE (six age categories), SOC (eight social class levels), REL (religion: four categories), POL (political preference: five categories), and EDU (four levels of education). Together with FUN and URB this leads to the total of 34 variables analyzed here.

The variables are rather different in type. The variables A1 through A8 and EU1 through EU5 are binary variables (agree or disagree with a statement). The variables A9 and A10 form six–point scales about the duration of pregnancy after which abortion is justifiable. The categories range from "until three months" (1) to "after six months" (5), while category 6 stands for "not justifiable". The latter category has been recoded as 0. The variables A11 through A14, and SF1 through SF5 are five–point Likert scale items. These variables, as well as the variables A9 and A10, have been considered as quantitative variables, because they clearly implied an ordinal scale, and, according to the theory in chapter 3, one way of dealing with ordinal variables is to treat them just as quantitative variables. The background variable SOC is also treated as a quantitative variable, because its eight categories are clearly ordered. Variables A15 and A16, just as the background

variables REL, POL, EDU, FUN, and URB, have been been considered as nominal variables.

The data described above have been analyzed by both PCAMIX and INDOMIX. The variables have not been standardized, but some of the variables have been weighted. That is, in order to let the background variables serve

Table 10.9. *Loadings from PCAMIX and INDOMIX.*

	PCAMIX		INDOMIX		INDOMIX loadings for quantitative variables	
	comp.1	comp.2	comp.1	comp.2	comp.1	comp.2
A1	0.00	0.19	0.02	0.04		
A2	**0.63**	0.00	**0.63**	0.00	0.79	
A3	**0.26**	0.10	**0.29**	0.00	0.54	
A4	**0.59**	0.00	**0.60**	0.00	0.78	
A5	0.19	0.13	**0.23**	0.02	0.48	
A6	**0.61**	0.00	**0.63**	0.00	0.80	
A7	**0.66**	0.00	**0.67**	0.00	0.82	
A8	**0.57**	0.00	**0.57**	0.00	0.75	
A9	**0.42**	0.03	**0.43**	0.01	-0.66	
A1	**0.48**	0.02	**0.49**	0.01	-0.70	
A1	**0.59**	0.01	**0.60**	0.00	0.77	
A12	**0.37**	0.18	**0.48**	0.03	-0.69	
A13	**0.49**	0.13	**0.61**	0.02	-0.78	
A14	**0.35**	**0.20**	**0.47**	0.03	-0.69	
A15	**0.33**	0.03	**0.26**	0.00		
A16	**0.21**	0.05	**0.20**	0.01		
EU1	**0.27**	0.07	**0.26**	0.00	0.51	
EU2	**0.22**	0.04	**0.20**	0.00	0.45	
EU3	**0.31**	0.05	**0.30**	0.00	0.55	
EU4	**0.21**	0.15	**0.23**	0.02	0.48	
EU5	0.10	0.00	0.07	0.00		
SF1	0.04	0.15	0.06	0.09		
SF2	0.02	**0.34**	0.07	**0.22**		0.47
SF3	0.00	**0.45**	0.02	**0.76**		0.87
SF4	0.06	**0.37**	0.12	**0.22**		0.47
SF5	0.02	**0.56**	0.06	**0.68**		0.83
SEX	0.00	0.00	0.00	0.00		
AGE	0.01	0.14	0.02	0.14		
SOC	0.00	0.01	0.00	0.01		
REL	**0.21**	0.06	**0.24**	0.04		
POL	**0.22**	0.11	**0.27**	0.02		
EDU	0.01	0.07	0.01	0.07		
FUN	0.03	0.07	0.04	0.05		
URB	0.05	0.00	0.04	0.01		

mainly as "supplementary" or "passive" variables that do not actually affect the solution, these variables have been given a very small weight (0.001) in the analyses.

In all analyses the dimensionality of the solutions was set at two, because the first two components of the three–dimensional INDOMIX solution hardly differed from those of the two–dimensional solution, and the third component essentially represented only one variable. The PCAMIX solution accounted for 12% of the inertia by means of the SUMPCA model ($IAF_S = .12$), while it accounted for 17% of the inertia in terms of the INDORT model ($IAF_I = .17$). The INDOMIX solution accounted for 19% of the inertia ($IAF_I = .19$). Clearly, the INDORT model fits the data considerably better than the more restricted SUMPCA model.

The loadings from PCAMIX (after varimax rotation) and INDOMIX are given in the first four columns of Table 10.9. Loadings $\geq .2$ have been printed in bold face. Clearly, the loadings from PCAMIX and INDOMIX differ, but when the components are interpreted on the basis of the variables that load high on them, the components are interpreted almost identically in the two analyses. That is, the variables A2 through A16 and EU1 through EU4 have high loadings on the first component, and the variables SF2 through SF5 have high loadings on the second component. In addition, the first component is related to the background variables REL and POL.

For a more detailed interpretation, it is useful to note that for quantitative variables (including binary variables) the loadings given here are the squares of the product–moment correlations. For the quantitative variables with INDOMIX loadings higher than .2 the ordinary product–moment correlations between the variables and the components concerned, are given in the last two columns of Table 10.9. The interpretation of the components in terms of the nominal variables can be facilitated by inspection of the category coordinates. For the variables A15, A16, REL and POL these are given in Table 10.10.

On the basis of Tables 10.9 and 10.10 the first INDOMIX component can be interpreted as "liberal versus conservative". This can be seen as follows. The first component is strongly positively correlated with the variables A2 through A8. These items ask whether (1) or not (2) abortion is allowed under specified circumstances. This component is negatively correlated with A9 and A10, asking after how much time abortion is still justifiable (from "not

Table 10.10. *Category coordinates on the INDOMIX dimensions for four high loading nominal variables.*

	comp.1	comp.2
A 15: Should abortion requests be handled by law ?		
law desired allowing for abortion only in special cases	0.50	0.01
law desired making abortion difficult	0.23	-0.19
no law desired, doctor decides in abortion requests	-0.57	0.03
A 16: Should abortion be permitted after twelve weeks of pregnancy ?		
after 12 weeks absolutely forbidden	0.65	-0.12
abortion after 12 weeks in special cases only	-0.11	0.02
law should not specify a time limit	-0.60	0.13
REL: Religion		
Reformed	0.24	-0.43
Calvinist	0.76	0.18
Roman Catholic	0.30	0.04
none	-0.57	0.10
POL: Political preference		
left	-0.38	-0.04
denominational	0.70	-0.02
liberal	-0.41	0.19
right	1.49	-0.68

justifiable" to after "six months"). Furthermore, the variables A11 to A14 correlate strongly with the first component. These variables indicate on five–points scales whether the respondents agree (low scores) or disagree (high scores) with certain moralistic statements on abortion. The first of these (A11) is a pro–abortion statement, which correlates positively with the first component, the others (A12 through A14) are against abortion, and correlate negatively with the first component. The variables EU1 through EU4 also correlate (positively) with the first component. These questions ask whether (1) or not (2) euthanasia is justifiable under certain specified circumstances. Summarizing, one can see that persons with low scores on the first component typically allow for abortion in many circumstances, that is, agree with many of the statements A2 through A8, find abortion justifiable even after long periods of pregnancy (A9 and A10), agree with the pro abortion statement (A11) and do not agree with the contra abortion statements (A12 through A14), find that laws regulating abortion or setting time limits regarding abortion are not needed (A15 and A16), and find that euthanasia is

justifiable under several circumstances (EU1 through EU4), while people with high scores typically take the reversed stand. This explains why this component can be labeled as "liberal versus conservative". As has been said earlier, two background variables are also related to this component. Upon inspection of the category coordinates of these variables (REL and POL), it is clear that at the lower end of this dimension one finds people without religion, and with left or liberal political preference. At the other end of the dimension one finds Reformed, Calvinist, and Roman Catholic respondents, with denominational or conservative political preferences.

The second component is positively correlated with four variables, SF2 through SF5. These variables are five–point ratings (from agree completely (1) to disagree completely (5)) of statements that are clearly contra sexual freedom. This component can hence simply be interpreted as "contra versus pro sexual freedom". This completes the description of the analysis of the abortion survey data.

As has been mentioned in section 10.1.2, one way of studying the stability of one's data is by means of cross–validation. This can be done as follows. The data set is split into two halves, which are both reanalyzed. These analyses yield weights (for the categories and the quantitative variables) which can be applied to the other half of the data (see section 10.1.2). The resulting components of object coordinates are not necessarily uncorrelated, but apart from this one can handle them as ordinary INDOMIX components. In order to see how sensitive the analysis is to particular characteristics of the sample drawn, the component scores and the loadings that can be computed after applying the weights to one half of the data are compared to the component scores and loadings resulting from INDOMIX on this half of the data. When the component scores and the loadings are very much the same, one can conclude that the analysis is hardly sensitive to the data. The component scores from the two analyses are compared by computing the correlation between them. In order to compare the loadings from one analysis to that of another, we use the coefficient proposed by Gower (1971) for the association between two numerical variables. In our case, this coefficient is equal to one minus the mean of the absolute differences between the loadings from the different solutions on a component. It has a maximum of one, which is attained when the loadings are identical.

For the present data this kind of cross–validation seems more useful than

a (random) jackknife study. Jackknifing is meant for studying changes under deletion of just one individual, which seems most appropriate in samples that intend to capture almost the whole population (as in the cetacea data). When a sample is intended to represent a much larger population, the present cross–validation procedure seems more appropriate, because it is based on comparing subsamples that should, just as the original sample, be representative for the population.

The cross–validation study undertaken here started with splitting the sample into two groups by alternatingly assigning one individual to the one and the next to the other group. First, the weights resulting from the INDOMIX analysis of the second half have been applied to the first half. The resulting components were hardly correlated (p.m.c. equal to –.13). They correlated strongly with the components resulting from INDOMIX on this half of the data, that is, the correlation between the first components was .997 and that between the second components was .978. The resulting loadings have been compared with the loadings from an INDOMIX analysis on the first half, by computing Gower's coefficients over the loadings on the (corresponding) components of the different solutions. This coefficient was .991 for the first components and .988 for the second component. These values correspond to mean absolute differences between the loadings of 0.009, and 0.012, respectively.

The inverse procedure of applying weights resulting from the first half to the data in the second half resulted in components that were nearly uncorrelated (p.m.c. equal to .10). Again, they correlated strongly with the components resulting from INDOMIX on this half of the data, that is, .997 for the first components and .980 for the second components. The coefficients for comparing the resulting loadings were 0.991 for the first component, and 0.987 for the second component, corresponding to mean absolute differences between the loadings of 0.009, and 0.013, respectively.

Clearly, the correlations between corresponding components and the values of Gower's coefficient for comparison of the loadings on the corresponding components are all very high. On the basis of this cross–validation study it can be concluded that the INDOMIX solution reported above is not very sensitive to sample fluctuations.

10.5. Residual complaints after head injury: Components analysis of binary variables

Van Zomeren and Van den Burg (1985) have observed a number of patients who had incurred a severe closed head injury. Many of them still reported some residual complaints, even after two years after the accident. These complaints have been recorded by means of 17 items, listed in Table 10.11. Apart from these (binary) variables, two measures indicating the severity of the injury have been scored as well. These latter variables are post–traumatic amnesia (PTA), and the extent to which previous work could be resumed (RTW).

The data on 50 patients have been reanalyzed* in the present study. Two of the original 52 patients have been excluded from the analysis because some observations were missing on them. The variables PTA and RTW are ordinal variables. They have been dichotomized here for simplification. PTA could very well be dichotomized by using the cut–off point proposed by Van Zomeren and Van den Burg (at 13 days). RTW was a five point–scale which could be distinguished roughly into "former work resumed" versus "former work resumed only partly or not at all". This variable has been dichotomized accordingly, yielding a set of 19 dichotomous variables.

The data, transformed as described above, have been analyzed by both MCA and INDOQUAL. Because the variables are all dichotomous, MCA comes down to PCA, and instead of INDOQUAL, INDOMIX can be used while all variables are considered numerical, which is computationally attractive (as has been discussed in chapter 9). The MCA analysis is followed by a varimax rotation.

In both analyses, two–dimensional solutions have been studied. This dimensionality has been chosen partly for reasons of comparison with the PCA solution reported by Van Zomeren and Van den Burg, and partly because it yielded an interpretable solution, in contrast to, for instance, the three–dimensional INDOQUAL solution for these data. In contrast to the other examples in the present study, the loadings that are reported here are ordinary (point–biserial) correlations between the variables and the

* The author is obliged to Pim van den Burg for kindly providing the original data

components, just as in ordinary PCA. These loadings are recorded in Table 10.11. For reasons of comparison, the last two columns of Table 10.11 contain the PCA–loadings reported by Van Zomeren and Van den Burg (1985). Loadings $\geq$ .60 are printed in bold face.

Table 10.11. *Loadings resulting from MCA and INDOQUAL, as well as loadings reported by Van Zomeren and Van den Burg* (1985).

	MCA + varimax		INDOQUAL		Original loadings	
	comp.1	comp.2	comp.1	comp.2	comp.1	comp.2
PTA	**0.76**	-0.09	**0.74**	0.00	-0.13	**0.80**
RTW	**0.71**	0.10	**0.73**	0.10	0.14	**0.70**
forgetfulness	**0.65**	0.19	**0.71**	0.07	0.19	**0.63**
irritability	0.13	0.54	0.21	0.37	0.59	-0.03
slowness	**0.69**	0.06	**0.61**	0.16	0.25	**0.66**
poor concentration	0.52	0.53	**0.63**	0.30	**0.61**	0.42
fatigue	0.39	**0.64**	0.42	0.53	**0.68**	0.31
dizziness	0.14	0.44	0.17	0.32	0.52	-0.03
incr.need of sleep	0.00	0.37	0.12	0.04	0.51	-0.24
intol. of light	0.10	**0.68**	0.25	0.30	**0.72**	-0.07
intol. of noise	0.42	0.52	0.30	**0.83**	**0.61**	0.33
loss of initiative	**0.65**	0.21	**0.72**	0.06	0.51	0.38
headache	0.13	**0.64**	0.07	0.17	0.57	-0.17
crying more readily	0.32	0.46	0.37	0.23	**0.60**	0.16
inability to do two things simultan.	**0.68**	0.38	**0.66**	0.42	0.44	**0.62**
intol. of bustle	0.22	0.55	0.09	**0.93**	**0.61**	0.12
depressed mood	**0.69**	0.04	**0.70**	0.01	0.32	0.53
more anxious	0.00	0.40	0.09	0.14	0.33	-0.05
indifference	0.04	0.48	0.09	0.18	0.24	0.13

The differences between the PCA loadings from Van Zomeren and Van den Burg (1985) and the MCA loadings reported here can be attributed to several differences in the procedures followed. Van Zomeren and Van den Burg (1985) did not dichotomize PTA and RTW, and performed a quartimax rotation instead of a varimax rotation. Especially the latter may well explain the differences,

because a quartimax rotation tends to yield a strong general component, which corresponds to the finding that Van Zomeren and Van den Burg's first component can be considered quite general. In order to check this, the solution reported by Van Zomeren and Van den Burg has been matched to the MCA solution, via an orthogonal Procrustes rotation (Green, 1952). The rotated components were related highly to the MCA components. That is, the congruence coefficient (Tucker's ϕ) measured between the loadings on the first components was .99, and that between the second components .96, indicating that the loadings yield equivalent interpretations.

The INDOQUAL solution differs from the MCA solution especially with respect to the second components (Tucker's ϕ equal to .84). The Procrustes rotation did not improve this. In MCA this component seems hard to interpret, because the variables that have high loadings on it do not seem to have much in common, except some forms of intolerance. In INDOQUAL the component can be seen as a dimension expressing more specifically intolerance of noise and bustle. In both analyses, the first component can be interpreted in the same way as Van Zomeren and Van den Burg (1985) did for their second component, that is, as a "severity"–dimension. Just as in Van Zomeren and Van den Burg (1985), this component is highly related to PTA and RTW, as well as to forgetfulness, slowness, poor concentration, inability to do two things simultaneously, and depressed mood. Unlike in Van Zomeren and Van den Burg (1985), in both analyses the first component is also related to loss of initiative, which corresponds to what one might have expected.

10.6. Characteristics of alcoholic and nonalcoholic drinks: Effects of standardizing nominal variables

For testing the INDOMIX program the author has created some fictitious data sets that are based on common sense knowledge. The data set to be considered in the present section has been constructed as follows. Some characteristics of 34 drinks, both soft–drinks and alcoholic drinks, have been given in terms of the following (seven) variables: alcoholic strength (five categories, from no alcohol to over 30 %), addition of sugar (yes or no), does drink contain carbonic acid (yes or no), kind of raw product which essentially determines the taste of the drink (fruit, grain, herbs, artificial flavors), price (four categories from cheap to expensive), taste (very sweet, sweet,

dry, bitter), and color (colorless, red, light–red, yellow, brown). The scores of the drinks on the variables are fictitious in that they are based on the author's (limited) knowledge of these drinks. As a consequence, certain data might actually be technically incorrect, but in general it can be expected that these data reflect reality rather well. The data with the names of the drinks (some of which are exclusive for certain European countries) are given in Table 10.12.

Table 10.12. *Characteristics of* 34 *drinks.*

variables	Alcohol	Sugar	Carbon	Raw prod.	Price	Taste	Color
	5 cat	2 cat	2 cat	4 cat	4 cat	4 cat	5 cat
labels							
syrop	1	1	2	4	2	1	4
cola	1	1	1	4	1	1	5
seven–up	1	1	1	4	1	1	1
orangina	1	1	1	1	1	1	4
apple juice	1	2	2	1	1	1	4
orange juice	1	2	2	1	2	2	4
red bordeaux	2	2	2	1	2	3	2
wh. bordeaux	2	2	2	1	2	3	4
red Lambrusco	2	2	1	1	2	1	2
rosé	2	2	2	1	2	2	3
Moselle wine	2	2	2	1	2	1	4
Sekt	2	2	1	1	2	2	4
Riesling	2	2	1	1	2	3	4
champagne ds	2	2	1	1	4	2	4
champagne br	2	2	1	1	4	3	4
sherry	3	2	2	1	3	2	5
port	3	2	2	1	3	1	2
Cointreau	5	1	2	1	4	1	1
jenever	5	2	2	2	3	4	1
gin	5	2	2	2	4	4	1
whisky	5	2	2	2	4	4	4
beer	2	2	1	2	1	4	4
old–br. beer	2	1	1	2	2	4	5
guinness	2	2	1	2	2	4	5
cider	2	1	1	1	2	1	4
strawberry lq	4	1	2	4	3	1	3
banana liquor	4	1	2	4	3	1	4
cherry brandy	4	1	2	4	3	1	3
bl.currant lq	4	1	2	4	3	1	2
slivovic	5	2	2	1	4	4	1
ouzo	5	1	2	3	4	1	1
Pernod	5	1	2	3	4	1	1
Jägermeister	4	2	2	3	3	4	5
rum	5	1	2	2	4	2	1

These data have been analyzed by means of INDOQUAL with and without standardization of the variables. As has been explained in chapter 3, standardization of nominal variables comes down to weighting these variables by $(m_j-1)^{-1/2}$. Clearly, standardizing the variables in the present data set implies using much larger weights for the variables Sugar and Carbon than for the other variables. In section 4.4 some grounds for choosing whether or not to normalize one's variables have been discussed. In the present section the effects of these choices on the results of an INDOQUAL analysis of the data set described just above will be compared.

The dimensionality of the solutions has been set, rather arbitrarily, to two after verifying that it gave an interpretable solution. Table 10.13 reports the loadings of the seven variables on the 2 components in both solutions. Loadings $\geq .70$ are printed in bold face.

Table 10.13. *Variable loadings resulting from INDOQUAL with and without standardization, respectively.*

	INDOQUAL on standardized variables		INDOQUAL on nonstandardized variables	
	comp.1	comp.2	comp.1	comp.2
Alcohol	0.38	0.56	**0.80**	**0.94**
Sugar	**0.96**	0.00	0.57	0.02
Carbon	0.00	**0.98**	0.06	0.10
Raw product	0.57	0.08	**0.84**	0.33
Price	0.11	0.42	0.42	**0.71**
Taste	0.53	0.02	0.54	0.12
Color	0.10	0.22	0.13	0.83

As is clear from Table 10.13, the INDOQUAL solutions on standardized and nonstandardized variables differ markedly. When the data are standardized, the solution is, in fact, dominated by the two binary variables Sugar and Carbon, the variables receiving the largest weights due to the standardization. Some of the other variables have modest loadings on these components as well, but by no means as high as those of the binary variables. In the INDOQUAL analysis of nonstandardized variables the binary variables

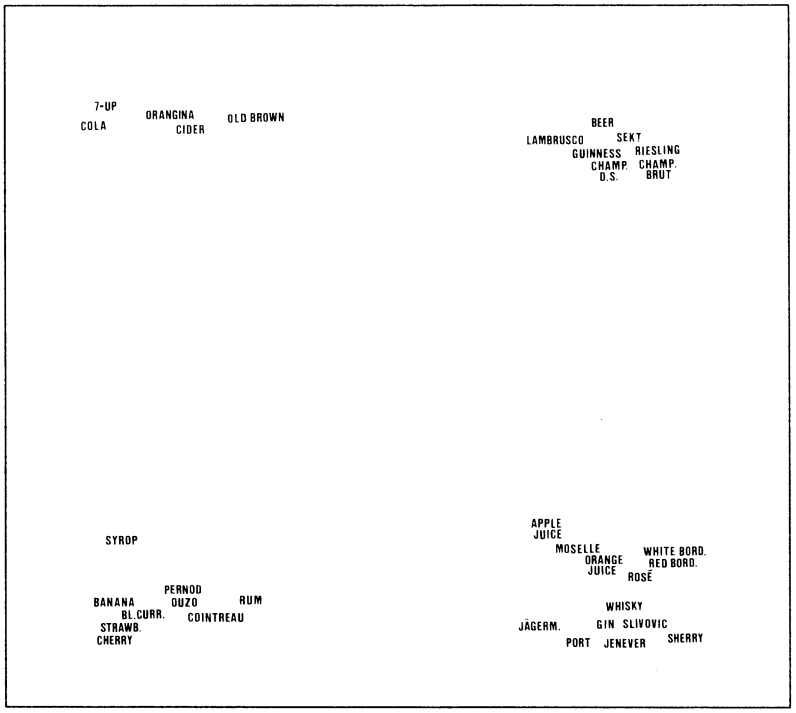


Figure 10.4. *Object coordinates resulting from INDOQUAL on standardized variables.*

only play modest roles. This solution is based more on the amount of alcohol in the drinks, the raw product determining the taste, the price, and the color of the drinks. The latter solution is more informative than the first in that it represents the data in terms of more variables.

The object coordinates of the two solutions are given in Figures 10.4 and 10.5 for the INDOQUAL solution of standardized variables and of nonstandardized variables, respectively. In the analysis of the standardized variables, the drinks are clustered in four groups with characteristics "sugar and CO_2 added", "contains CO_2 but no sugar added", "sugar added but no CO_2", "neither sugar nor CO_2 added". This results, for instance, in a cluster containing fruit juices, several wines, and strong drinks like gin and whisky at the same time, which does not seem to correspond to how one would typically classify drinks. In the solution of INDOQUAL on nonstandardized variables the drinks are not clustered as clearly as in the other analysis, but the

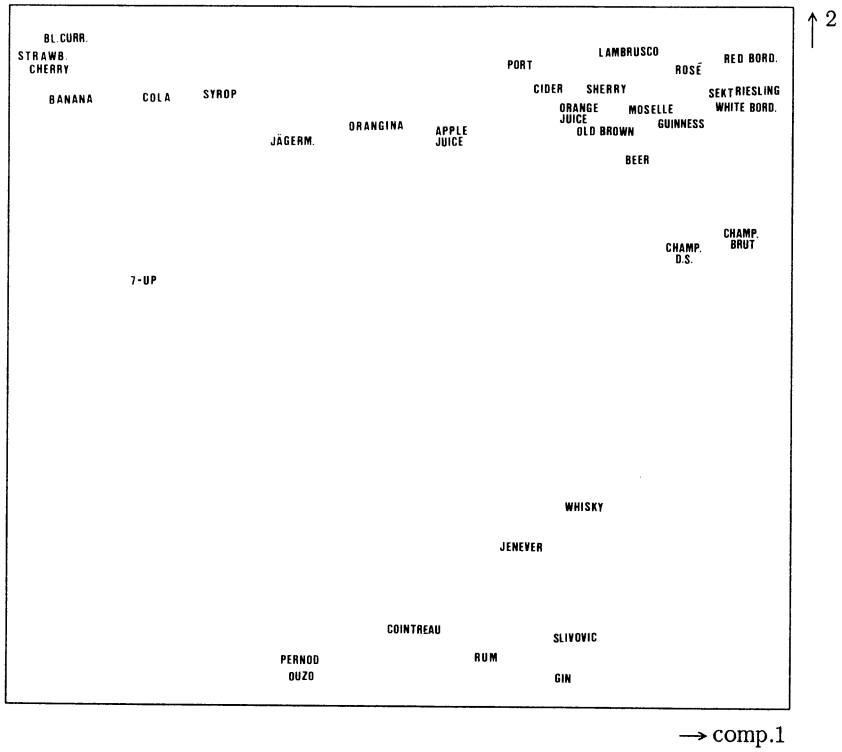


Figure 10.5. *Object coordinates resulting from INDOQUAL on nonstandardized variables.*

picture does seem to make more intuitive sense, especially separating strong drinks from soft drinks.

It can be concluded that, at least in the present case, it is not useful to standardize the qualitative variables. The original variables can be considered to yield quantification matrices in comparable units, and what is called standardizing the variables might be seen as, in fact, un–standardizing the variables, that is, letting variables with few categories have the strongest impact on the solution.

10.7. Italian freight transportation data: A comparison of INDORT and TUCKALS-3 on quantification matrices

Marchetti (1988) has described a data set collected by the Italian National Research Organization (CNR) on eight characteristics of 54 Italian

freight transportation industries. These characteristics are expressed in terms of eight variables, which are considered as nominal variables both here and by Marchetti (1988). These variables are labelled A through H, just as has been done by Marchetti (1988). They measure:

A: number of employees (three categories)

B: number of tractors (three categories)

C: use of containers (binary: yes or no)

D: number of semi–trailers (four categories)

E: number of transports in 1981 (three categories)

F: juridical status (four categories: joint–stock company, limited liability company, general or limited partnership, sole traders)

G: location (three categories: north, center, south)

H: type of firm (four categories: shipping agent, forwarding agent and carrier, lorry–conveyor, carrier).

Variables A, B, D and E are polytomized quantitative variables, but the ordering underlying the categories is not used here, following Marchetti (1988).

Marchetti (1988) has analyzed these data by means of TUCKALS–3 applied to quantification matrices. As quantification matrices he has chosen the standardized versions of the ones used in MCA and INDOQUAL. In other words, he has used the same quantification matrices as are used in MCA and INDOQUAL, but applies weights to the variables in order to standardize them. These weights are $(m_j-1)^{-1/2}$, $j = 1,...,m$, where m_j is the number of categories of variable j. As has been noted by Marchetti and has been seen in section 10.6, this may result in a solution in which variables with a small number of categories are better represented than the others.

In the present study the same data are analyzed by means of INDOQUAL, and the resulting loadings and category coordinates are compared to the ones found by Marchetti (1988). For the TUCKALS–3 analysis Marchetti (1988) has chosen both dimensionalities (i.e., for the components of variables and for the components of objects) equal to three. This corresponds best to choosing the dimensionality of the INDOQUAL solution equal to three as well.

The three–dimensional INDOQUAL solution yields a function value of 2.43, which can be seen as the inertia accounted for. Because the variables are standardized, the total inertia in the data is equal to the number of variables, that is, 8. As a result, INDOQUAL accounts for 2.43*100/8 = 30.4 %

of the total inertia. Marchetti reports that the TUCKALS–3 solution accounted for 30.8 % of the inertia. It can be concluded that INDOQUAL represents the information in the data almost as well as TUCKALS–3 does, while INDOQUAL uses a much simpler model.

The loadings for the variables found by the INDOQUAL solution are given in Table 10.14. These loadings were computed as $\mathbf{x}_l' S_j \mathbf{x}_l$, $j = 1,...,8$, $l = 1,...,3$, where S_j is the *standardized* version of the quantification matrix used in MCA and INDOQUAL. Loadings $\geq .50$ are printed in bold face.

Table 10.14. *Variable loadings resulting from INDOQUAL.*

	comp.1	comp.2	comp.3
A	**0.55**	0.02	**0.70**
B	0.13	0.10	0.02
C	0.00	**0.99**	0.00
D	0.23	0.26	0.18
E	**0.56**	0.05	0.02
F	0.24	0.05	0.03
G	0.21	0.03	0.06
H	0.21	0.12	0.00

It can be seen that only the variables A, C, and E are represented quite well, which is the same conclusion as was drawn by Marchetti (1988). This is about the only conclusion made by Marchetti as far as the variables are concerned. From his plots for the variables one might conclude that the first component mainly represents variables A and C, and to a less extent also D and E. The second component contrasts C to the other variables, and the third component contrasts A to E, while the other variables are in between these extremes. It seems difficult to give a further interpretation of these results on the basis of the loadings alone, especially as far as contrasts between nominal variables are concerned. A simple structure rotation might possibly have helped here.

The loadings from the INDOQUAL solution lead to a different interpretation. The first component of the INDOQUAL solution is highly related to variables A and E, and can be interpreted as a component expressing the "size of the company". The second component only represents variable C well, and the third component mainly represents variable A. A more detailed

comparison of the loadings found here and by Marchetti is not feasible because Marchetti (1988) only gave plots of the variables. In addition, it should be noted that the TUCKALS–3 solution can be rotated by any nonsingular matrix, hence comparing the results of the two analyses should take such rotational freedom into account. From the plots, however, it is practically impossible to see how such a rotation should be made.

Apart from plotting the loadings for the variables, Marchetti (1988) also gives a plot of the category coordinates for the first two components, and he provides a rather detailed interpretation of these results. This plot is reproduced here (with permission from Marchetti) in Figure 10.6. The capitals denote the variables, and the indices denote the categories of the variables concerned. In the same plot the category coordinates for the first two INDOQUAL components have been given after a rotation (by hand) to maximal agreement of the two configurations. The INDOQUAL category coordinates are given by lower case characters. The INDOQUAL coordinate axes are also depicted in this plot, and labelled as "DIM.1" and "DIM.2",

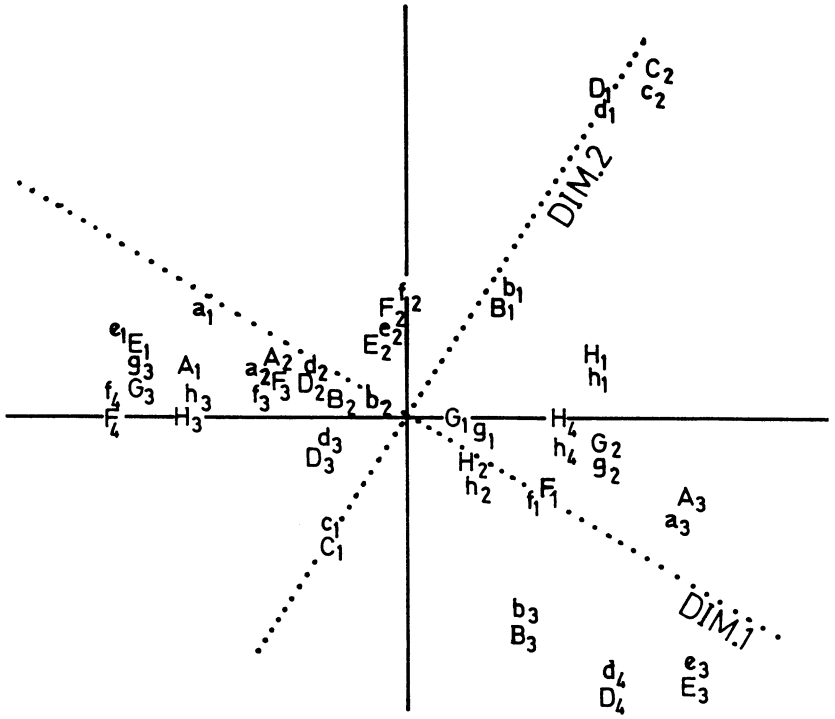


Figure 10.6. *Plot of the category coordinates on the dimensions* 1 *and* 2 *from the TUCKALS (capitalized) and INDOQUAL (lower case) solutions.*

respectively. Clearly, after rotation, the configurations of category points resulting from the two different analyses are virtually equal, and the interpretations provided by Marchetti hold for the INDOQUAL solution as well.

The components found by TUCKALS–3 for the objects and for the variables are related to each other via the elements of the core matrix. In this way TUCKALS–3 represents the data by means of a more complicated model than INDOQUAL does. In the latter method the components for the objects and those for the variables have a simple one–to–one relation with each other. Therefore, it is possible to interpret the loadings by relating them to the category coordinates and vice versa. In the present example such an interpretation can be used to interpret the first INDOQUAL component as expressing the size of a company (variables A and E mainly), and the second component as the one that contrasts the use of containers to the use of semi–trailers (variables C and to a less extent D). On the basis of the TUCKALS–3 solution a similar interpretation can be made, based on the category coordinates alone, but this interpretation of the components cannot readily be carried over to the components of the variables. Because the INDOQUAL solution accounts for practically the same amount of inertia as the TUCKALS–3 solution does, it seems that the former is to be preferred for the representation of the present data.

10.8. The Sugiyama Data: Where INDOQUAL fails

INDOQUAL has been developed as an alternative to MCA in order to find better representations for data in which subsets of variables form clusters of closely related variables, while variables of different clusters are hardly related. Such data are often characterized by rather large (generalized) correlations within subsets of variables. However, if nominal variables do not have high generalized correlations among each other, the data set may nevertheless contain interesting information. This can be the case, for instance, with preference data, e.g., binary variables indicating whether or not a stimulus is picked out of a number of stimuli. The analysis of such data has been described, for instance, by Heiser (1981). He gives several example data sets, one of which, the "Sugiyama data" (Sugiyama, 1975, see Heiser, 1981, p.142) is reanalyzed here. The data consists of six binary variables

pertaining to religious behavior. The variables can be described briefly as A: Do you make it a rule to practice religious conduct; B: Do you visit a grave once or twice a year; C: Do you occasionally read religious books; D: Do you visit shrines and temples to pray; E: Do you keep a talisman; and F: Did you draw a fortune. For the exact wordings of the questions, as well as the data themselves, the reader is referred to Heiser (1981, p.142).

Heiser has analyzed these data by MCA and found no interpretable results. A different method, better adapted to the analysis of binary proximity data (called HOMANA–BIN by Heiser, 1981), did yield good results. It turned out that the questions could be seen to form a scale ordered as F–D–E–B–A–C. In the present study the data have been analyzed by means of INDOQUAL. It turned out, however, that INDOQUAL did not produce a useful representation for this data. That is, INDOQUAL consistently yielded solutions of which each component represented one variable almost exclusively. Using different random starts tended to yield different solutions with almost the same function value, but with quite different loadings.

Clearly, these INDOQUAL results are of no value whatsoever. As has been said above, INDOQUAL is supposed to be useful when the data contains some subsets of highly correlated variables. Such subsets might be identified by inspection of the correlations between the variables, possibly complemented by a PCA of these correlations. In Table 10.15 the ϕ^2–contingency coefficients are given for this data. Clearly, no subsets of highly correlated variables are present in this data. This explains the poor quality of the INDOQUAL solutions.

Table 10.15. *ϕ^2 – contingency coefficients among the variables of the Sugiyama data.*

	A	B	C	D	E	F
A	1.00					
B	0.01	1.00				
C	0.08	0.00	1.00			
D	0.01	0.01	0.00	1.00		
E	0.01	0.03	0.00	0.08	1.00	
F	0.00	0.00	0.00	0.04	0.04	1.00

10.9. Concluding remarks

In the present chapter seven example data sets have been analyzed by some of the techniques described in the present study. There are considerable differences between these examples, as well as between the choices that have been made for analyzing them. Nevertheless, some general remarks can be made.

In most of the analyses reported here PCAMIX solutions have been compared with INDOMIX solutions. Except for the example in section 10.5, the INDOMIX solutions did not differ much from the PCAMIX solutions, at least, after the latter had been rotated by means of varimax. The modest differences that could be observed, however, did reveal a systematic tendency. High loadings in PCAMIX correspond to even higher loadings in INDOMIX, and small loadings in PCAMIX correspond to even smaller loadings in INDOMIX. This has been pointed out in particular in sections 10.2 and 10.3, but can be found in other examples as well, although not always as clear.

In two cases the stability of the INDOMIX solution has been studied. In both cases the solution appeared highly stable. Obviously, this is only partly a feature of the method. Stability is first of all determined by the homogeneity of a population or the representativeness of a sample. The special choices made in the analysis may also affect the stability. For instance, the dimensionality of the solution might be related to its stability. In the present analyses mainly small dimensionalities have been chosen. These choices might partly account for the stability of the solutions. Implicitly, this reasoning gives another criterion for determining the dimensionality of one's solution. One might check the stability of solutions of different dimensionalities and choose one's final solution only from those solutions that are sufficiently stable.

Apart from testing INDOMIX on several data sets, in passing also the varimax procedure for PCAMIX has been used consistently. The usefulness of this procedure has been discussed in one case only. It has been used in every analysis, however, and seemed very useful. Especially when one wants to interpret the components, it is useful to have a simple structure for the loadings.

It may seem rather inconsistent to use the varimax criterion for rotating the MCA solution and compare this solution with INDOMIX, which

maximizes the quartimax criterion. The reason for this apparent inconsistency is that the varimax criterion is preferred over the quartimax criterion, for the arguments given by Kaiser (1958), but the variant of INDOMIX that maximizes the varimax criterion has not yet been programmed. Moreover, INDOMIX itself has an interesting interpretation as a compromise between MCA and PCA of quantification matrices. This interpretation does not hold for the varimax based variant of INDOMIX.

The final example has not only been taken up to show the limitations of INDOQUAL (and in fact also of MCA). It also shows that analyzing one's data by just one method may hide important aspects of the data. A more useful strategy seems to be to use more than one of the methods mentioned in the hierarchy, possibly even all of these. Then one should not only consider the solutions of each of the separate analyses, but especially the larger differences in the solutions. Together with the knowledge how the methods presented in this study are related, one may determine whether or not the data can be described by these methods, and if so, which is the most useful representation of this data.

REFERENCES

Benzécri, J.–P. et al. (1973). *L'analyse des données*. Paris: Dunod.

Carroll, J.B. (1953). An analytic solution for approximating simple structure in factor analysis. *Psychometrika*, **18**, 23–38.

Carroll, J.D., & Chang, J.–J. (1970). Analysis of individual differences in multidimensional scaling via an n–way generalization of "Eckart–Young" decomposition. *Psychometrika*, **35**, 283–319.

Carroll, J.D., & Chang, J.–J. (1972). *IDIOSCAL: A generalization of INDSCAL allowing IDIOsyncratic reference systems as well as an analytic approximation to INDSCAL.* Paper presented at the Spring Meeting of the Psychometric Society, Princeton, New Jersey, March 30–31.

Carroll, J.D., De Soete, G. &, Pruzansky, S. (1989). An evaluation of five algorithms for generating an initial configuration for SINDSCAL. *Journal of Classification*, **6**, 105–119.

Carroll, J.D., Pruzansky, S., & De Soete, G. (1987). A comparison of three rational initialization methods for INDSCAL. In Diday et al. (Eds.) *Data analysis and informatics*, INRIA, Le Chesnay.

Carroll, J.D., & Wish, M. (1974). Models and methods for three–way multidimensional scaling. In D.H. Krantz et al. (Eds.) *Contemporary developments in mathematical psychology, Vol.II: Measurement, psychophysics, and neural information processing* (pp. 57–105). San Fransisco: Freeman & Co.

Cazes, P. (1980). L'analyse de certains tableaux rectangulaires décomposés en blocs: Généralisation des propriétés rencontrées dans l'analyse des correspondances multiples II. Questionaires: Variantes de codages et nouveaux calculs de contributions [ANA.BLOCS II]. *Les Cahiers de l'Analyse des Données*, **5**, 387–403.

Cazes, P., Bonnefous, S., Baumerder, A., & Pagès, J.P. (1976). Description cohérente des variables qualitatives prises globalement et de leurs modalités. *Statistique et Analyse des Données*, **1(2)**, 48–62.

Clarkson, D.B., & Jennrich, R.I. (1988). Quartic rotation criteria and algorithms. *Psychometrika*, **53**, 251–259.

Coppi, R. (1986). Analysis of three–way data matrices based on pairwise

relation measures. In *Compstat* 1986 (pp. 129–139). Heidelberg: Physica–Verlag.

Crawford, C.B., & Ferguson, G.A. (1970). A general rotation criterion and its use in orthogonal rotation. *Psychometrika*, **35**, 321–332.

Cuadras, C.M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. In Y. Dodge (Ed.) *Statistical data analysis and inference.* Amsterdam: Elsevier Science Publishers.

D'Alessio, G. (1988). Multistep principal components analysis (MPCA): A new approach for the analysis of contingency tables series. In H.H. Bock (Ed.) *Classification and related methods of data analysis.* (pp. 497–504). Amsterdam: Elsevier Science Publishers.

D'Ambra, L., & Marchetti, G.M. (1986). The analysis of three–way data matrices: a method based on relation measures between units (in Italian). In *Proceedings of the 33rd Meeting of the Italian Statistical Society, vol* 1 (pp. 171–182).

Daniels, H.E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika*, **33**, 129–135.

De Leeuw, J. (1973). *Canonical analysis of categorical data.* Doctoral Dissertation, University of Leiden.

De Leeuw, J., & Meulman, J. (1986). A special jackknife for multidimensional scaling. *Journal of Classification*, **3**, 97–112.

De Leeuw, J., & Pruzansky, S. (1978). A new computational method to fit the weighted Euclidean distance model. *Psychometrika*,**43**, 479–490.

De Leeuw, J., & Van Rijckevorsel, J.L.A. (1980). HOMALS and PRINCALS, some generalizations of principal components analysis. In E. Diday et al. (Eds.) *Data analysis and informatics II* (pp. 231–242). Amsterdam: Elsevier Science Publishers.

De Leeuw, J., & Van Rijckevorsel, J.L.A. (1988). Beyond Homogeneity Analysis. In J.L.A. Van Rijckevorsel J., & J. De Leeuw (Eds.) *Component and correspondence analysis* (pp. 55–80). New York: Wiley.

Di Ciaccio, A. (1986). Representation of a new association measure between categories using multidimensional scaling. In E. Diday et al. (Eds.) *Data analysis and informatics IV* (pp. 369–378). Amsterdam: Elsevier Science Publishers.

Domenges, D., & Volle, M. (1979). Analyse factorielle spherique: une exploration. *Annales de l'INSEE*, **35**, 3–84.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211–218.

Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Les Cahiers de l'Analyse des Données*, **4**, 137–146.

Escofier, B. (1984). Analyse factorielle en reference à un modèle. Application a l'analyse de tableaux d'échanges. *Revue de Statistique Appliquée*, **32**, 25–36.

Escofier, B., & Pagès, J. (1983). Méthode pour l'analyse de plusieurs groupes de variables – Application à la caractérisation de vins rouges du Val de Loire. *Revue de Statistique Appliquée*, **31**, 43–59.

Escofier, B., & Pagès, J. (1984). *L'analyse factorielle multiple*. Cahiers du bureau universitaire de recherche opérationnelle, no.**42**, Université Pierre et Marie Curie, Paris.

Escoufier, Y. (1970). Echantillonnage dans une population de variables aléatoires réelles. *Publ. Inst. Statist. Univ. Paris*, **19**, fax. 4, 1–47.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, **29**, 751–760.

Escoufier, Y. (1980). Exploratory data analysis when data are matrices. In K. Matusita (Ed.) *Recent developments in statistical inference and data analysis* (pp. 45–53). Amsterdam: Elsevier Science Publishers.

Ferguson, G.A. (1954). The concept of parsimony in factor analysis. *Psychometrika*, **19**, 281–290.

Fichet, B. (1986). Distances and Euclidean distances for presence–absence characters and their application to factor analysis. In J. De Leeuw, W. Heiser, J. Meulman, & F. Critchley (Eds.) *Multidimensional data analysis* (pp. 23–46). Leiden: DSWO press.

Fichet, B., & Gbegan, A. (1986). Analyse factorielle des correspondences sur signes de presence–absence. In E. Diday et al. (Eds.) *Data analysis and informatics IV* (pp. 209–219). Amsterdam: Elsevier Science Publishers.

Gifi, A. (1981). *Nonlinear multivariate analysis*. Leiden: Department of Data Theory.

Gower, J.C. (1966). Some distance properties of latent root and vector

methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–871.

Grasse, P. (Ed.) (1955). *Travaux sous la direction de P.Grasse. Traité de zoologie.* Paris: Masson.

Green, B.F. (1952). The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, **17**, 429–440.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.

Greenacre, M.J. (1988). Correspondence analysis of multivariate categorical data by weighted least squares. *Biometrika*, **75**, 457–467.

Guttman, L. (1941). The quantification of a class of attributes: A theory and method of scale construction. In P. Horst et al. (Eds.) *The prediction of personal adjustment* (pp. 319–348). New York: Social Science Research Council.

Harman, H.H. (1976). *Modern factor analysis* (3rd edition). Chicago: University of Chicago press.

Hartigan, J.A. (1975). *Clustering algorithms*. New York: Wiley.

Hayashi, C. (1950). On the quantification of qualitative data from the mathematico–statistical point of view. *Annals of the Institute of Statistical Mathematics*, **2** (1), 35–47.

Heiser, W.J. (1981). *Unfolding analysis of proximity data.* Leiden: Department of Psychology.

Heiser, W.J., & Meulman, J. (1983). Analyzing rectangular tables by joint and constrained multidimensional scaling. *Journal of Econometrics*, **22**, 139–167.

Horan, C.B. (1969). Multidimensional scaling: Combining observations when individuals have different perceptual structures. *Psychometrika*, **34**, 139–165.

Hubert, L. (1977). Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology*, **30**, 98–103.

Jaffrennou, P.A. (1978). *Sur l'analyse des familles finies de variables vectorielles.* Thèse, Université Saint–Étienne.

Janson, S., & Vegelius, J. (1978a). *Correlation coefficients for more than one scale type and symmetrization as a method of obtaining them.* (Research

Report, 78–2) University of Uppsala, Department of Statistics.

Janson, S., & Vegelius, J. (1978b). On the applicability of truncated component analysis based on correlation coefficients for nominal scales. *Applied Psychological Measurement*, **2**, 135–145.

Janson, S., & Vegelius, J. (1982). Correlation coefficients for more than one scale type. *Multivariate Behavioral Research*, **17**, 271–284.

Jennrich, R.I. (1970). Orthogonal rotation algorithms. *Psychometrika*, **35**, 229–235.

Jones, M.C., & Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society, series A*, **150**, 1–36.

Kaiser, H.F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187–200.

Kiers, H.A.L. (1988). Principal components analysis on a mixture of quantitative and qualitative data based on generalized correlation coefficients. In M.G.H. Jansen, & W.H. van Schuur (Eds.) *The many faces of multivariate analysis (Vol. I): Proceedings of the SMABS – 88 conference in Groningen* (pp. 67–81). Groningen: Rion.

Kiers, H.A.L. (1989a). A computational short–cut for INDSCAL with orthonormality constraints on positive semi–definite matrices of low rank. *Computational Statistics Quarterly*, in press.

Kiers, H.A.L. (1989b). *Hierarchical relations between three – way methods*. Manuscript submitted for publication.

Kiers, H.A.L. (1989c). INDSCAL for the analysis of categorical data. In R. Coppi, & S. Bolasco (Eds.) *Multiway data analysis* (pp. 155–167). Amsterdam: Elsevier Science Publishers.

Kiers, H.A.L. (in press). Majorization as a tool for optimizing a class of matrix functions. *Psychometrika*.

Kroonenberg, P.M. (1983). *Three – mode principal component analysis: Theory and applications*. Leiden: DSWO press.

Kroonenberg, P.M., & De Leeuw, J. (1980). Principal component analysis of three–mode data by means of alternating least squares algorithms. *Psychometrika*, **45**, 69–97.

Lauro, N., & D'Ambra, L. (1984). L'analyse non–symmetrique des correspondances. In E. Diday et al. (Eds.) *Data analysis and informatics III* (pp. 433–446). Amsterdam: Elsevier Science Publishers.

Lebart, L., Morineau A., & Tabard, N. (1977). *Techniques de la description statistique*. Paris: Dunod.

Levin, J. (1966). Simultaneous factor analysis of several gramian matrices. *Psychometrika*, **31**, 413–419.

L'Hermier des Plantes, H. (1976). *Structuration des tableaux à trois indices de la statistique*. Thèse de 3ème cycle, Université Montpellier II.

Marchetti, G.M. (1988). *Three – way analysis of two – mode matrices of qualitative data*. Research Report, Department of Statistics, University of Florence.

Marcotorchino, F. (1984). *Utilisations des comparaisons par paires en statistique des contingences* (Etudes F069, F071, F081). Centre Scientifique IBM–France, Paris.

Meulman, J.J. (1982). *Homogeneity analysis of incomplete data*. Leiden: DSWO press.

Meulman, J.J. (1986). *A distance approach to nonlinear multivariate analysis*. Leiden: DSWO press.

Miller, R.G. (1974). The jackknife: A review. *Biometrika*, **61**, 1–15.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, **49**, 115–132.

Neuhaus, J.O., & Wrigley, C. (1954). The quartimax method: An analytic approach to orthogonal simple structure. *British Journal of Mathematical and Statistical Psychology*, **7**, 81–91.

Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University Press.

Nishisato, S. (1984). Forced classification: A simple application of a quantification method. *Psychometrika*, **49**, 25–36.

Sabatier, R. (1987). *Méthodes factorielles en analyse des données: Approximations et prise en compte de variables concomitantes*. Thèse de doctorat, Université des Sciences et Techniques du Languedoc, Montpellier.

Saporta, G. (1975). *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives*. Thèse de 3ème cycle, Université de Paris IV.

Saporta, G. (1976). Quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives. *Statistique et Analyse des Données*, **1**, 38–46.

Saporta, G. (1979). Pondération optimale de variables qualitatives en analyse

des données. *Statistique et Analyse des Données*, **3**, 19–31.

Saunders, D.R. (1953). *An analytic method for rotation to orthogonal simple structure.* Research Bulletin, RB 53–10. Princeton, New Jersey: Educational Testing Service.

Sugiyama, M. (1975). *Religious behavior of the Japanese: Execution of a partial order scalogram analysis based on quantification theory.* Paper presented at the US–Japan seminar on theory, methods and applications of multidimensional scaling and related techniques, La Jolla, California.

Ten Berge, J.M.F. (1983). A generalization of Kristof's theorem on the trace of certain matrix products. *Psychometrika*, **48**, 519–523.

Ten Berge, J.M.F. (1984). A joint treatment of varimax rotation and the problem of diagonalizing symmetric matrices simultaneously in the least–squares sense. *Psychometrika*, **49**, 347–358.

Ten Berge, J.M.F., Knol, D.L., & Kiers, H.A.L. (1988). A treatment of the orthomax rotation family in terms of diagonalization, and a re–examination of a singular value approach to varimax rotation. *Computational Statistics Quarterly*, **3**, 207–217.

Tenenhaus, M. (1977). Analyse en composantes principales d'un ensemble de variables nominales ou numériques. *Revue de Statistique Appliquée*, **25**, 39–56.

Tenenhaus, M., & Young, F.W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, **50**, 91–119.

Ter Braak, C.J.F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179.

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

Tschuprow, A.A. (1939). *Principles of the mathematical theory of correlation.* New York: William Hodge.

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.

Tucker, L.R., & Messick, S. (1963). An individual differences model for

multidimensional scaling. *Psychometrika*, **28**, 333–367.

Van Buuren, S., & Heiser, W.J. (1989). Clustering n objects into k groups under optimal scaling of variables. *Psychometrika*, in press.

Van der Burg, E. (1985). HOMALS classification of whales, porpoises and dolphins. In J.–F. Marcotorchino, J.–M. Proth, & J. Jansen (Eds.) *Data analysis in real life environment: Ins and outs of solving problems* (pp. 25–36). Amsterdam: Elsevier Science Publishers.

Van der Burg, E. (1988). *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO press.

Van der Heijden, P.G.M. (1987). *Correspondence analysis of longitudinal categorical data*. Leiden: DSWO press.

Van Rijckevorsel, J. (1987). *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. Leiden: DSWO press.

Van Zomeren, A.H., & Van den Burg, W. (1985). Residual complaints of patients two years after severe head injury. *Journal of Neurology, Neurosurgery, and Psychiatry*, **48**, 21–28.

Vegelius, J. (1973). *Correlation coefficients as scalar products in Euclidean spaces*. (Research Report, 145) University of Uppsala, Department of Statistics.

Vegelius, J. (1978). On the utility of the E–correlation coefficient concept in psycholgical research. *Educational and Psychological Measurement*, **38**, 605–611.

Vegelius, J., & Bäckström, A. (1981). An enquiry about religion, analyzed by a nominal scale truncated component analysis. *Educational and Psychological Measurement*, **41**, 717–724.

Vescia, G. (1985a). Automatic classification of cetaceans by similarity aggregation. In J.–F. Marcotorchino, J.–M. Proth, & J. Jansen (Eds.) *Data analysis in real life environment: Ins and outs of solving problems* (pp. 15–24). Amsterdam: Elsevier Science Publishers.

Vescia, G. (1985b). Descriptive classification of cetacea: Whales, porpoises and dolphins. In J.–F. Marcotorchino, J.–M. Proth, & J. Jansen (Eds.) *Data analysis in real life environment: Ins and outs of solving problems* (pp. 7–13). Amsterdam: Elsevier Science Publishers.

Yanai, H. (1986). Some generalizations of correspondence analysis in terms of projection operators. In E. Diday et al. (Eds.) *Data analysis and*

informatics IV (pp. 193–207). Amsterdam: Elsevier Science Publishers.

Young, F.W., Takane, Y., & De Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, **43**, 279–281.

Zegers, F.E. (1986). *A general family of association coefficients*. Doctoral dissertation, University of Groningen.

Zegers, F.E., & Ten Berge, J.M.F. (1986). Correlation coefficients for more than one scale type: An alternative to the Janson and Vegelius approach. *Psychometrika*, **51**, 549–557.

DRIE-WEG METHODEN VOOR HET ANALYSEREN VAN KWALITATIEVE EN KWANTITATIEVE TWEE-WEG GEGEVENS

Voor de exploratieve analyse van resultaten van (sociaal–wetenschappelijk) onderzoek is Principale Componenten Analyse (PCA) een veel gebruikte en nuttige techniek. Het doel van deze methode is een aantal variabelen efficiënt samen te vatten door middel van een aantal zogenaamde komponenten.

PCA is alleen bruikbaar voor kwantitatieve variabelen. Voor het analyseren van niet–kwantitatieve variabelen, met name variabelen van nominaal meetnivo, zijn verschillende methoden voorgesteld, die elk slechts voor een deel hetzelfde doel bereiken als PCA. Enerzijds zijn er technieken ontwikkeld die gebaseerd zijn op gegeneraliseerde korrelatie–koëfficiënten. Dergelijke koëfficiënten drukken het verband uit tussen twee nominale variabelen of tussen één nominale variabele en één kwantitatieve variabele. Ze kunnen op dezelfde manier als gewone korrelatie–koëfficiënten gebruikt worden voor PCA van een stel variabelen. In tegenstelling tot PCA voor kwantitatieve variabelen, kan een dergelijke vorm van PCA voor nominale variabelen echter niet direkt een weergave van de observatie–eenheden (objekten) leveren.

Anderzijds zijn er technieken voorgesteld waarin de kategorieën van nominale variabelen worden opgevat als (binaire) variabelen en op deze variabelen een soort PCA wordt uitgevoerd. De bekendste representant van dit type methoden is Multipele Correspondentie Analyse (MCA). Deze methode heeft als voordeel boven de gegeneraliseerde korrelatie–koëfficiënten methode dat zij wel een weergave levert voor de objekten. Daar staat echter tegenover dat de techniek van het oorspronkelijke doel van PCA van de nominale variabelen afwijkt doordat het de kategorieën van de variabelen centraal stelt in plaats van de variabelen zelf.

In dit onderzoek worden methoden voor PCA van nominale variabelen ontwikkeld die een compromis vormen tussen de bovengenoemde twee typen methoden. Het gaat namelijk om methoden die een weergave leveren voor de objekten (net als MCA) en tegelijkertijd een goede weergave van de variabelen geven (net als methoden voor PCA gebaseerd op gegeneraliseerde korrelatie–koëfficiënten). Omdat nominale variabelen vaak gezamenlijk met kwantitatieve

variabelen geanalyseerd moeten worden (we spreken dan van gemengde gegevens) zijn de bovenbeschreven compromis–methoden gegeneraliseerd zodat ze toepasbaar zijn voor het analyseren van dergelijke gemengde gegevens.

Het onderzoek bestaat uit drie delen. In het eerste deel wordt een skala van methoden beschreven voor het analyseren van nominale of gemengde gegevens. Deze methoden berusten op toepassingen van drie–weg methoden op zogenaamde kwantifikatie–matrices. In hoofdstuk 2 wordt een aantal drie–weg methoden beschreven. Het gaat hierbij om een speciaal type drie–weg methoden dat bedoeld is voor het analyseren van een aantal gelijkenissen–matrices die gelijkenissen geven tussen steeds dezelfde objekten, ten aanzien van verschillende aspekten. Voor de in dit onderzoek genoemde methoden wordt aangetoond dat ze onderling een hierarchie vormen. De eerste methode is de meest algemene. De tweede methode kan opgevat worden als de eerste methode als hieraan bepaalde randvoorwaarden worden opgelegd. De derde methode is op te vatten als de tweede na toevoeging van bepaalde extra randvoorwaarden, etc.

Methoden voor het analyseren van nominale variabelen worden gekonstrueerd door drie–weg methoden toe te passen op kwantifikatie–matrices. Kwantifikatie–matrices zijn matrices die gelijkenissen tussen objekten geven zoals die kunnen worden bepaald op grond van een nominale variabele. Een heel eenvoudige vorm van zo'n kwantifikatie–matrix is die die de gelijkenis tussen twee objekten die in dezelfde kategorie van een (nominale) variabele vallen als 1 aangeeft, en de gelijkenis tussen twee objekten die in verschillende kategorieën van een variabele vallen als 0. In hoofdstuk 3 worden deze en meer gekompliceerde kwantifikatie–matrices behandeld. Naast kwantifikatie–matrices voor nominale variabelen worden kwantifikatie–matrices voor kwantitatieve variabelen besproken, en wordt aangegeven hoe kwantifikatie–matrices voor ordinale variabelen kunnen worden gekonstrueerd.

In het laatste hoofdstuk van het eerste deel, hoofdstuk 4, wordt aangetoond wat voor methoden er ontstaan als de in hoofdstuk 2 genoemde drie–weg methoden worden toegepast op de in hoofdstuk 3 genoemde kwantifikatie–matrices. Hier wordt met name aangetoond dat deze kombinatie leidt tot een aantal bestaande methoden, waaronder diverse PCA technieken die gebaseerd zijn op gegeneraliseerde korrelatie–koëfficiënten, alsmede MCA. Het levert ook een groot aantal nieuwe methoden op die elk een compromis vormen

tussen gegeneraliseerde korrelatie–koëfficiënten methoden en varianten van MCA.

Nadat in het eerste deel van het onderzoek een breed skala van voornamelijk nieuwe technieken voor het analyseren van nominale en gemengde gegevens is beschreven, wordt in het tweede deel één techniek in het bijzonder behandeld. Het gaat om de toepassing van INDSCAL met orthonormaliteits–voorwaarde (INDORT genoemd) op de kwantifikatie–matrices die ook (impliciet) in MCA gebruikt worden. Deze methode wordt INDOQUAL genoemd. In hoofdstuk 5 wordt deze methode uitvoerig besproken. Er wordt aangetoond op welke manier deze methode een compromis vormt tussen één bepaalde gegeneraliseer–de korrelatie–koëfficiënten methode en MCA. Het blijkt dat deze methode zeer goed kan worden opgevat als PCA van nominale variabelen onder de randvoorwaarde dat de komponenten voor de variabelen direkt gerelateerd zijn aan komponenten voor de objekten. Voorts wordt uitgelegd hoe de resultaten van een dergelijke analyse kunnen worden geïnterpreteerd.

In hoofdstuk 6 wordt INDOQUAL vergeleken met enkele andere methoden. Impliciet wordt hiermee een aantal nieuwe interpretaties van INDOQUAL gegeven.

INDOQUAL is alleen bruikbaar voor het analyseren van nominale varia–belen. Gemengde variabelen kunnen worden geanalyseerd met behulp van de in hoofdstuk 7 beschreven generalisatie van INDOQUAL, die INDOMIX genoemd is. Er wordt beschreven hoe deze methode, net als INDOQUAL, te zien is als compromis tussen bestaande methoden. Voorts wordt beschreven hoe de resul–taten van deze methode geïnterpreteerd kunnen worden. Tenslotte worden enkele speciale gevallen behandeld.

In hoofdstuk 8 worden relaties van INDOMIX (en dus impliciet van INDOQUAL) met simple structure rotatie–technieken beschreven. Dergelijke technieken dienen ertoe om bij gewone PCA de komponenten zodanig te roteren dat de ladingen (korrelaties van komponenten met variabelen) extreem zijn, dat wil zeggen, hetzij groot (dicht bij 1), hetzij klein (dicht bij 0). Er wordt aangetoond dat INDOMIX gezien kan worden als een methode die simple structure maximaliseert volgens het quartimax kriterium. Deze methode verschilt hierin van simple structure rotatie–technieken voor PCA dat bij INDOMIX de simple structure niet alleen gemaximaliseerd wordt over moge–lijke *rotaties* van komponenten, maar over *alle* mogelijke komponenten.

Daarnaast zijn, om een zinvolle vergelijking van INDOMIX met MCA mogelijk te maken, technieken ontwikkeld voor simple structure rotatie van MCA–oplossingen.

In hoofdstuk 9 worden algorithmen beschreven voor INDOQUAL en INDOMIX die speciaal ontworpen zijn om het mogelijk te maken gegevens–bestanden met grote aantallen objekten te analyseren. Tevens worden in dit hoofdstuk nog enige technische aspekten van de oplossing van INDOQUAL en INDOMIX besproken.

Het laatste deel van het onderzoek betreft toepassingen en ervaringen met INDOQUAL en INDOMIX. Er worden zeven toepassingen beschreven, waarin steeds andere aspekten van de methoden worden belicht. Er wordt uitvoerig ingegaan op de interpretatie van INDOQUAL en INDOMIX oplossingen voor empirische gegevens. Bovendien wordt in twee gevallen de stabiliteit van de oplossingen onderzocht. In het eerste geval wordt hierbij gebruik gemaakt van de zogenaamde jackknife methode. Daarbij wordt steeds één ander objekt uit het gegevens–bestand weggelaten, en worden de verkregen oplossingen voor deze deel–steekproeven onderling vergeleken om na te gaan hoeveel invloed het weglaten van één objekt heeft op de oplossing. In het tweede geval wordt de stabiliteit bestudeerd door middel van kruis–validatie. Hierbij wordt de totale steekproef in tweeën gesplitst, vervolgens worden de gewichten die resulteren uit de oplossing van de ene deel–steekproef toegepast op de andere, en tenslotte wordt de aldus verkregen weergave voor de objekten uit de tweede deel–steekproef vergeleken met de weergave van de objekten die gevonden wordt door een gewone INDOMIX analyse van deze deel–steekproef. Door middel van deze procedure wordt nagegaan in welke mate de oplossingen afhangen van specifieke kenmerken van elk van de deel–steekproeven. In de twee voorbeelden waar de stabiliteit is bestudeerd bleek deze zeer groot te zijn.

AUTHOR INDEX

SUBJECT INDEX

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

NOTATION

In this study it is attempted to use a uniform notation for most of the symbols. In general, lower case *italic* symbols denote integer numbers, lower case greek symbols denote real numbers, lower case **bold** symbols denote vectors, and upper case *italic* (or greek) symbols denote matrices. The elements of vectors and matrices are denoted by the same characters, but in lower case *italics*, with indices to denote their position in the vectors or matrices. Finally, some short–cuts for summation signs are described. It should be noted that chapter 9 has a partly deviating notation.

g	index for categories of variable j
h	index for categories of variable k
f_g	frequency of category g
i	index for objects, $i = 1,...,n$.
j	index for variables, $j = 1,...,m$.
k	index for variables, $k = 1,...,m$.
l	index for components (dimensions), $l = 1,...,r$
m	number of variables
m_j	number of categories of qualitative variable j
n	number of objects
p	number of object–components in TUCKALS–3
r	number of components (dimensionality)
α_j	weight for variable j
$\delta_{ll'}$	Kronecker δ; $\delta_{ll'} = 1$ if $l = l'$; $\delta_{ll'} = 0$ if $l \neq l'$
λ_j	j^{th} diagonal element of Λ, where Λ is a diagonal matrix
$\mathbf{c}$	m–vector with loadings for variables in SUMPCA
$\mathbf{h}_j$	n–vector of raw scores on variable j

$\mathbf{s}_j$	n^2–vector with elements of S_j strung–out row–wise, Vec(S_j)
$\mathbf{w}_j$	vector with diagonal elements of W_j
$\mathbf{x}_l$	n–vector of scores on component l, normalized to unit sum of squares
$\mathbf{z}_j$	n–vector of standardized scores on variable j
$\mathbf{1}$	n–vector with unit elements only
A	$m \times r$ matrix of (nonsquared) loadings of variables on components
C	$m \times r$ matrix of (squared) loadings of variables on components
D_j	$r \times r$ diagonal matrix with category frequencies of variable j
E_j	$m \times m$ symmetric matrix which is diagonalized by ordinary orthomax
$\tilde{E}_j$	$m \times m$ symmetric matrix which is diagonalized by orthomax for PCAMIX
F	matrix of component scores ($n^2 \times r$ in chapter 2, $n \times r$ in chapter 8)
F_l	$n \times n$ matrix expressing "component" l for matrices $S_1,...,S_m$
G_j	$n \times m_j$ indicator matrix for variable j
H	$p^2 \times r$ matrix containing Vec(H_1),....,Vec(H_r)
H_l	$p \times p$ matrix containing the l^{th} frontal plane of the TUCKALS–3 core
I_r	$r \times r$ identity matrix
J	$n \times n$ centering operator, given by (I_n–$n^{-1}\mathbf{1}\mathbf{1}'$)
K	$n \times n$ (orthonormal) matrix of eigenvectors of $\sum_j S_j$
K_r	$n \times r$ matrix with first r eigenvectors of $\sum_j S_j$
P_j	$n \times n$ quantification matrix $JG_jD_j^{-1}G_j'J$ for qualitative variable j
Q_j	$n \times n$ quantification matrix $n^{-1}\mathbf{z}_j\mathbf{z}_j'$ for quantitative variable j
S_j	$n \times n$ quantification matrix for variable j
S	$n^2 \times m$ matrix containing vectors $\mathbf{s}_1,...,\mathbf{s}_m$
T	$r \times r$ (orthonormal) rotation matrix
W	$r \times r$ diagonal matrix
W_j	$r \times r$ diagonal matrix with loadings for variable j on the diagonal

X	matrix of object coordinates (mostly $n \times r$, and $X'X = I_r$)
Y_j	$m_j \times r$ matrix with category means of object coordinates
Z	$n \times m$ matrix with standardized scores of objects on variables
Λ	diagonal matrix (often with eigenvalues of $\sum_j S_j$)
Λ_r	$r \times r$ diagonal matrix with first r eigenvalues of $\sum_j S_j$
$\sum_j$	summation over $j = 1,...,m$
$\sum_l$	summation over $l = 1,...,r$
Vec(.)	column–vector with the elements of a matrix strung–out row–wise
$\otimes$	Kronecker product

Apart from this list of symbols it seems useful to provide a list of abbreviations (or acronyms) of methods that are mentioned frequently in this dissertation. The names are given with, between brackets, the number of the page where they are introduced in this study and, a short description of their objectives.

INDOMIX (p. 63)

INDORT applied to a set of quantification matrices; for qualitative variables $S_j = JG_jD_j^{-1}G_j'J$, for quantitative variables $S_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'$

INDOQUAL (p. 41)

INDORT applied to a set of quantification matrices given by $S_j = JG_jD_j^{-1}G_j'J$, $j = 1,...,m$

INDORT (p. 14)

INDscal with ORThonormality constraint on object coordinates

Minimizes $\sum_j \|S_j - XW_jX'\|^2$ over X, subject to $X'X = I_r$, and over diagonal W_j

Maximizes $\sum_j \sum_l (\mathbf{x}_l'S_j\mathbf{x}_l)^2$ over X, subject to $X'X = I_r$

INDSCAL (p. 13)

INDividiual differences SCALing

Minimizes $\sum_j \| S_j - XW_jX' \|^2$ over arbitrary X $(n \times r)$ and diagonal W_j $(r \times r)$

MCA (p. 2)

Multiple Correspondence Analysis

Maximizes tr $X'J\sum_j G_j D_j^{-1} G_j' JX$ over X, subject to $X'X = I_r$

PCA

Principal Components Analysis

Maximizes tr $X'\sum_j \mathbf{z}_j \mathbf{z}_j' X$ over X, subject to $X'X = I_r$

PCAMIX (p. 62)

Generalization of MCA and PCA, for the analysis of mixed variables

Maximizes tr $X'S_jX$ over X, subject to $X'X = I_r$, where $S_j = JG_jD_j^{-1}G_j'J$ for qualitative variables and $S_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'$ for quantitative variables

PCA of quantification matrices (p. 1)

PCA of "correlation"–coefficients for qualitative variables

STATIS–1 applied to matrices S_j

Maximizes sum of squared loadings ("correlations" between variables and components)

PCA of ϕ^2–coefficients (p. 30)

STATIS–1 applied to matrices $S_j = JG_jD_j^{-1}G_j'J$

Maximizes sum of squared loadings (η^2–coefficients between variables and components)

PCA of η^2–coefficients (p. 63)

STATIS–1 applied to matrices $S_j = JG_jD_j^{-1}G_j'J$ for qualitative variables and $S_j = n^{-1}\mathbf{z}_j\mathbf{z}_j'$ for quantitative variables

Maximizes sum of squared loadings (η^2–coefficients or squared product–moment correlations between variables and components)

STATIS–1 (p. 10)

First step of STATIS method

PCA of $\text{Vec}(S_1),\ldots,\text{Vec}(S_m)$

Minimizes $\sum_j \| S_j - \sum_l c_{jl}F_l \|^2$ over $F_1,\ldots,F_r$ and C

SUMPCA (p. 15)

Minimizes $\|\sum_j S_j - X\Lambda X'\|^2$ over diagonal matrices Λ, and X, subject to $X'X = I_r$

Maximizes tr $X'\sum_j S_j X$ over X, subject to $X'X = I_r$

TUCKALS–3 (p. 11)

Method for least–squares fitting of Tucker's three–mode component analysis model

Minimizes $\sum_j \| S_j - X\sum_l c_{jl}H_lX' \|^2$ over arbitrary X $(n \times p)$, C $(m \times r)$, and H_l, $l = 1,\ldots,r$

THREE-WAY METHODS FOR THE ANALYSIS OF QUALITATIVE AND QUANTITATIVE TWO-WAY DATA

A problem often occurring in exploratory data analysis is how to summarize large numbers of variables in terms of a smaller number of dimensions. When the variables are quantitative, one may resort to Principal Components Analysis (PCA). When qualitative (categorical) variables are involved, one may choose from a variety of methods which are adaptations of PCA, designed for this purpose.

This book is about such adapted PCA methods for qualitative variables, and for mixtures of qualitative and quantitative variables.

Part I treats adapted PCA methods, like Multiple Correspondence Analysis (MCA), in retrospect. The methods are systematized such that the choice between the different methods is facilitated. In order to fill certain gaps in this system a number of new methods is developed. One of these new methods is discussed in detail in Part II, which contains various innovatory contributions. One of the main new developments is the construction of techniques for rotating the MCA solution to what is called simple structure. These techniques are generalizations of the well-known VARIMAX and QUARTIMAX techniques for rotating PCA solutions.

The former rotation techniques are implicitly constrained in the sense that the new components are rotations of the MCA solution. In order to optimize VARIMAX or QUARTIMAX from the start a different technique is offered in which this constraint is dropped. This new technique, which is based on INDSCAL, finds solutions in which clusters of variables are identified more clearly than is done in MCA, even after simple structure rotation. An important additional consequence of this is that the observation units (objects, individuals) are also represented more clearly in clusters than is the case with MCA, which implies that the method is useful as a cluster technique.

Finally, Part III shows how certain selected methods work in practice. For this purpose seven example data sets have been analyzed. The results are discussed in detail, including some stability analyses. Particular attention is paid to the clustering of variables and objects.

DSWO PRESS

ISBN 90 6695 037 4