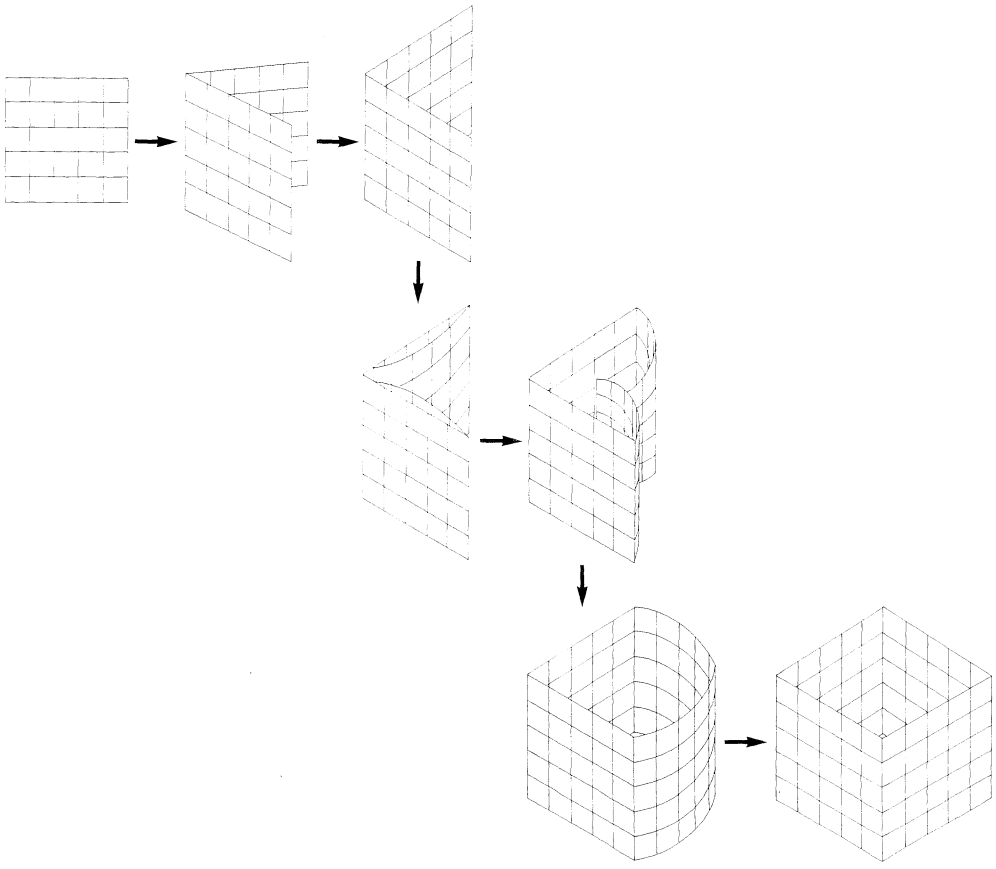
THREE-WAY METHODS FOR THE ANALYSIS OF QUALITATIVE AND QUANTITATIVE TWO-WAY DATA

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.