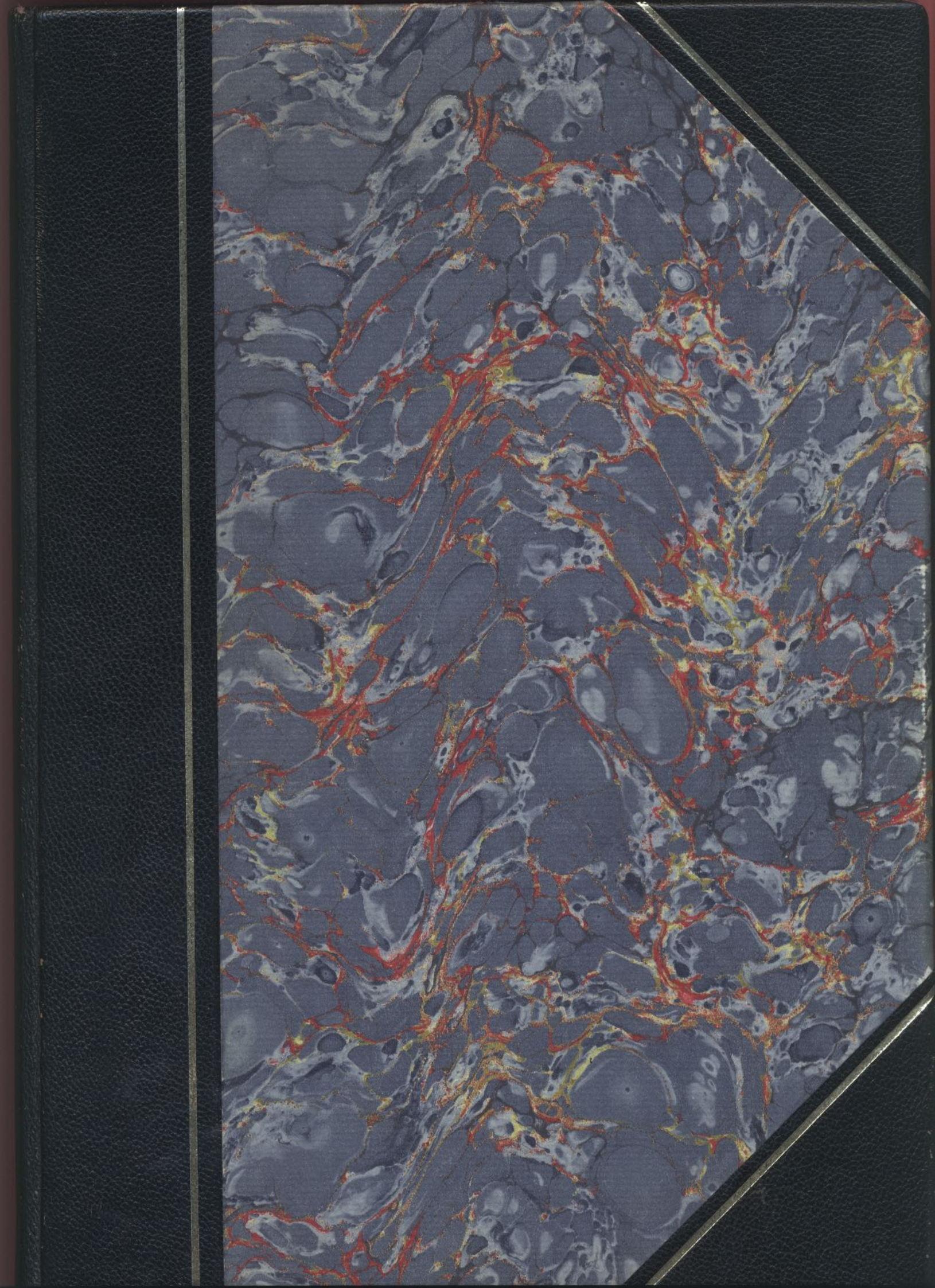


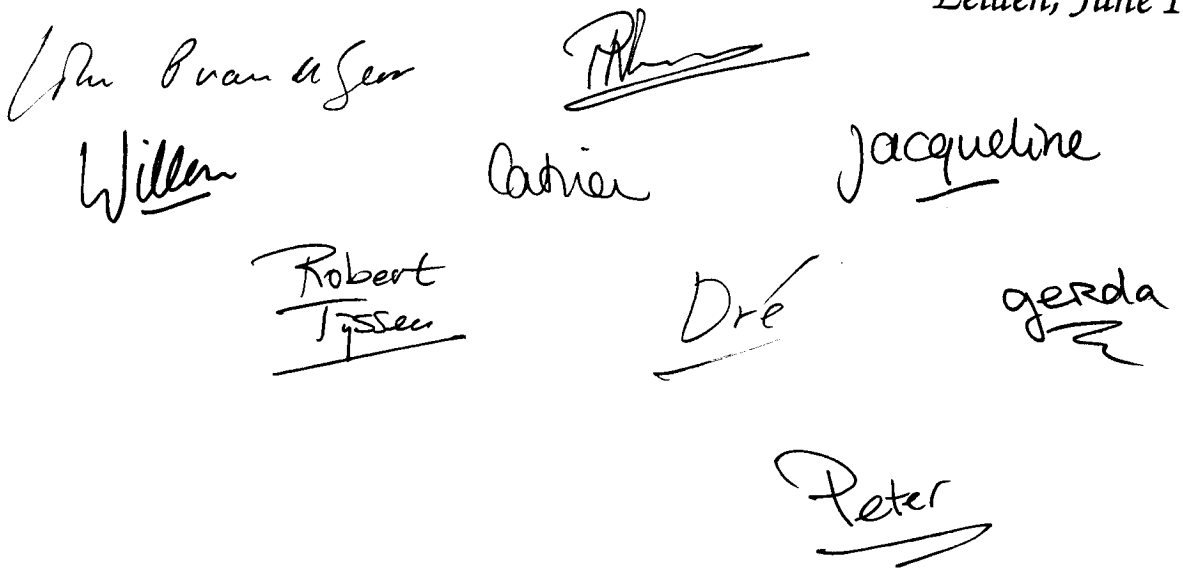
This is the only copy of the fourth, completely unrevised impression of my Nonlinear Multivariate Analysis book. This special edition has been prepared on the occasion of the departure of Jan de Leeuw from the department of Data Theory at Leiden.

"My little book is troublesome as an ague. I thought it was off my hands but it has bothered me up to this instant, when I sealed up the MS in a packet to go by post to Murray. And still there are odds and ends left and revises to come, etc. etc. But it is comparatively calm now. And it is such a small book after all. My friend F. H. Collins, who is a prince among proof correctors but cannot now leave his arm-chair, has been giving all his working time last week to putting the various contributions into better shape. The material was good but the arrangement too higgledy-piggledy."

"The new methods occupy an altogether higher plane than that in which ordinary statistics and simple averages move and have their being. Unfortunately the ideas of which they treat, and still more the many technical phrases employed in them, are as yet unfamiliar. The arithmetic they require is laborious, and the mathematical investigations on which the arithmetic rests are difficult reading even for experts; moreover they are voluminous in amount and still growing in bulk. Consequently this new departure in science makes its appearance under conditions that are unfavourable to its speedy recognition, and those who labour in it must abide for some time in patience before they can receive sympathy from the outside world."

"It's difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our own flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once. An average is but a solitary fact, whereas if a single other fact be added to it, an entire Normal Scheme, which nearly corresponds to the observed one, starts potentially into existence. Some people hate the very name of statistics, but I find them full of beauty and interest. Whenever they are not brutalised, but delicately handled by the higher methods, and are warily interpreted, their power of dealing with complicated phenomena is extraordinary. They are the only tools by which an opening can be cut through the formidable thicket of difficulties that bare the path of those who pursue the Science of man."

Albert Gifi
Leiden, June 1987



PREFACE

This book was written as the text to accompany a post-doctoral course in 'non-linear multivariate analysis', held in June 1981. An earlier Dutch version (Albert Gifi, 'Niet-lineaire multivariate analyse', Leiden 1980) served a similar purpose for a course held in March 1980. The present text is not just a translation of the earlier Dutch text. Most of the present text is newly written. Also, the present text is better organized (we hope) around the available computer programs. The text is the joint product of the members of the Department of Data Theory of the Faculty of Social Sciences, University of Leiden. 'Albert Gifi' is their 'nom de plume'. The portrait, however, of Albert Gifi shown here, is that of the real Albert Gifi to whose memory this book is dedicated, as a far too late recompense for his loyalty and devotion, during so many years, to the Cause he served.

Bert Bettonvil
Steef de Bie
Eeke van der Burg
John van de Geer
Willem Heiser
Judy Knip
Jan de Leeuw
Jacqueline Meulman
Peter Neufeglise
André Nierop
Ineke Stoop

© Department of Datatheory, Faculty of Social Sciences, University of Leiden, The Netherlands, 1981

No part of this book may be reproduced and/or published in any form, by print, photoprint, microfilm or any other means without written permission from the Department of Datatheory.

CONTENTS

1.0 Introduction

In this introductory chapter we shall try to give a definition of multivariate analysis (MVA). Of course we have to take definitions given by others into account. We do this by giving brief content analyses of the prefaces and introductory chapters of some of the more popular books on MVA. There turns out to be considerable agreement over the definition, but the main reason for this is that the definition which is most frequently used is not very specific. In our content analysis of the various books we repeatedly find a number of problems or controversies that seem interesting enough to be analyzed more in detail. This is done in the second part of the chapter.

1.1 Content analysis MVA-books

1.1.1 Roy (1957)

Roy does not pretend to give a complete coverage of MVA. In most of the chapters he analysis multinormally distributed variables, and more specifically he derives bounds for confidence intervals for certain classes of parametric functions. He mentions factor analysis, classification, and variance component analysis as his most important omissions. It is especially important for our purposes that Roy's chapter 15 is about non-parametric generalizations of MVA, by which he means techniques very similar to the modern log-linear models for multidimensional contingency tables. "Despite all the mathematical elegance and comparative simplicity of 'normal variate' analysis of variance and multivariate analysis, one cannot help feeling that the non-parametric approach (whether of this variety or of other varieties) is far more realistic and physically meaningful, and is likely, in the future, to supplant, to a large extent, the existing techniques of 'normal variate' analysis of variance and multivariate analysis, including those discussed in the first fourteen chapters of this monograph." (1c, p viii). Although Roy does not give an explicit definition, it seems that he interprets MVA as a class of techniques that can be used to test a restricted number of hypotheses about the relationships between correlated variables.

1.1.2 Kendall (1957, 1975)

Kendall does give a definition. "We may thus define multivariate analysis as the branch of statistical analysis which is concerned with the relationships of sets of dependent variables." (1957, p 6). He also gives a number of examples of typical MVA problems, and he says that such a list of problems usually works better than a simple definition. In his book he makes the important distinction between the analysis of dependence and the analysis of interdependence. In the

analysis of dependence we investigate if and how a group or set of variables depends on another group. The first set consists of the so-called dependent variables, the second group of the independent variables. There is a certain amount of asymmetry in the analysis, or, to put it differently, the direction of causal influence is from the independent to the dependent variables. In the analysis of interdependence the sets of variables are treated symmetrically, there is no distinction between dependent and independent variables. The main example of interdependence analysis is principal components analysis, the most familiar example of dependence analysis is multiple regression.

In the modernized version of Kendall's book, published in 1975, the introduction is considerably longer. The definition of MVA has not changed, and is repeated with slightly different phrasing on page 1. Kendall then proceeds to list the most important purposes of MVA techniques.

a: Structural simplification.
b: Classification.
c: Grouping of variables.
d: Analysis of dependence.
e: Analysis of interdependece.
f: Construction and testing of hypotheses.

He also mentions a number of important problems in the further development of MVA techniques.

a: In many practical situations we cannot make the usual statistical assumptions. There is often no random sample from a population, either the population is not defined or the sample is not random. "It is a mistake to try and force the treatment of such data into a classical probabilistic mould, even though some subjective judgment in treatment and interpretation may be involved in the analysis of the results." (1975, p 4).
b: Although it is practically impossible to apply MVA without the use of a computer there are certain dangers in the uncritical use of 'canned' computer programs.
c: Even if we have a random sample it is often impossible to assume multivariate normality. "Most theoretical work in multivariate statistics is based on the assumption that the parent population is multinormal, and its robustness under departures from normality is very often difficult to determine with any exactitude." (1975, p 4).
d: Pictures and graphs are very important in MVA. It is however very difficult to make satisfactory graphic representations of data in more than two dimensions
e: There is no single, natural definition of order between points in multi-dimensional space. As a consequence there is no satisfactory non-parametric MVA, comparable to univariate nonparametric statistics.

We remember from Kendall's discussion that in MVA we deal with observations on a number of interdependent variables. But the variables are not necessarily stochastic, not necessarily continuously measurable, and the observations are not necessarily independent and identically distributed. Finally it is remarkable that Kendall, like Roy, has a last chapter on the analysis of categorised or discrete multivariate data. There is no such chapter in the other books we discuss. If we compare the two editions of Kendall's book then we find, certainly in the introductory chapter, a shift from multivariate statistical analysis and multinormal analysis to a more general description of MVA.

1.1.3 Anderson (1958)

"In this book we shall concern ourselves with statistical analysis of data that consist of sets of measurements on a number of individuals or objects. ... The mathematical model on which analysis is based is a multivariate normal distribution or a combination of multivariate normal distributions." (1c, p 1). This is certainly clear enough. It is interesting to study Anderson's reasons for this apparent loss of generality.

a: The multinormal distribution often turns out to be a good description of the distribution of multivariate data arising in practice.

b: The multinormal distribution follows from the multidimensional central limit theorem, and is consequently a good approximation if the observations can be interpreted as the sum of a large number of independent small effects.

c: Distribution theory based on the assumption of multinormal parent populations is mathematically relatively simple, and consequently many interesting results have been derived in the last 75 years.

Anderson interprets MVA in most of his book as a generalisation of familiar normal-theory inferential statistics to multinormal situations. As a consequence typical multivariate techniques such as principal components analysis and canonical analysis do not get much attention in his book.

1.1.4 Cooley and Lohnes (1962, 1971)

The first version of the book demonstrates how the usual MVA techniques must be implemented on a computer. This means, of course, that it is now completely out of date. The definition of MVA is the same as Kendall's. The second version, published in 1971, has an interesting preface. Somewhere between 1962 and 1971 Cooley and Lohnes found their identity: they are now multivariate data analysts. This explains the change in the title of the book. The reasons for the change is, of course, Tukey (1962). "His gift to us was our professional identity. ... Tukey made our interest in multivariate heuristics rather than multivariate inference sound respectable." (1971, p v). This is a remarkable statement.

We must remember to investigate why social scientists studying MVA did not have a professional identity befor 1962, and we must also find out if the only effect of Tukey's paper and subsequent efforts was that some shady practices now sound respectable.

Cooley and Lohnes use Kendall's definition in their 1971 book, they also use the distinction between analysis of dependence and interdependence troughout the book. There is much popularization, application, and computation in the book.

1.1.5 Morrison (1967,1976)

The definition is more or less standard. "Multivariate statistical analysis is concerned with data collected on several dimensions of the same individual." (1967, p vii). Morrison imposes more restrictions than Kendall, however, because he assumes explicitly that the individuals are a random sample from an infinite population, and he assumes multivariate normality almost everywhere. These are possibly the reasons that Morrison, like Anderson, uses the term 'multivariate statistical analysis'. The major difference with Anderson is that Morrison has an extensive treatment of principal components analysis, canonical analysis, and factor analysis, and that principal components analysis in particular is described as a descriptive data-reduction technique. Morrison's book also has an extensive introduction to matrix algebra. The changes in the 1976 edition are not very interesting for our purposes.

1.1.6 Van de Geer (1967, 1971)

On the cover of the Dutch version of Van de Geer's book we read: "Multivariate analysis, being the art of describing the relationship between several variables by using mathematical techniques." In Cooley and Lohnes (1962) we saw emphasis on computer programs, no statistics, and only a little algebra. In Van de Geer's book there is also no statistics, but matrix algebra takes up almost half of the book. MVA only comes at the end. "In these last chapters multivariate analysis is discussed as a data-reduction technique only. In other words the book does not discuss any statistical questions in the sense of inferential statistics." (1967, p 12). The computer is mentioned as an important source of inspiration, the time we do not have to spend any more behind the adding machine can now be used to gain insight in what we are really doiing. It is clear that the insight that Van de Geer is trying to teach is mainly geometrical, wherever possible he uses figures. Bringing insight is truly bringing in sight.

In the much enlarged English edition of 1971 the starting points are stated even more clearly. "In my opinion, statistical theory is a substantially more advanced subject than is required to understand what multivariate techniques

really do with data." (1971, p ix). Van de Geer mentions the main advantages of his approach to MVA: insight in stead of a mere copying of computer output, emphasis on the similarities of the various forms of MVA, and the ensuing "cross-fertilization" between the various social sciences that use some form of MVA. Important additions in the English version are path analysis and structural equations, discussed as two new variations on the same theme.

According to Van de Geer MVA is linear analysis of data matrices, and its most important purposes are data-reduction and geometric representation. MVA is applied linear algebra, or, which amounts to the same thing, applied linear geometry. The emphasis on pictures (also pictures in the form of graphs with arrows) we also find, to a lesser degree, in Cooley and Lohnes (1971) and Kendall (1975). Van de Geer does not seem to have trouble with his professional identity, in fact his geometrical approach fits into the psychometric tradition starting with multiple factor analysis and culminating with the book of Coombs (1964).

1.1.7 Dempster (1969)

Dempster's book has a fine introductory chapter. We find the following description there. "The purpose of this book is to describe certain methods of analysis of statistical data arising from multivariate samples. A basic aim of such data analysis is to reduce large arrays of numbers to provide meaningful and reasonably complete summaries of whatever information resides in sample aggregates. Another aim is to draw inferences from sample aggregates to larger population aggregates from which the samples are drawn; that is, to understand how certain information about a sample provides uncertain information about a population." (lc, p 3). Thus data analysis and statistics are distinguished from the start. They are also treated separately in the book, starting with data analysis. "While most books on statistical theory start out with sampling theory and attempt to make methods of data analysis follow, the attitude in this book is that the methods of data analysis are carried out largely because of the intrinsic appeal of the sample quantities computed. Such, at least, were the historical origins of the methods described here. Moreover, when viewed as producing descriptive or summary statistics, the methods have value even when assumptions like randomness of samples and normality of populations are quite unwarranted." (lc, p 3). The separate treatment gives the impression that the two approaches have nothing in common. In his introduction Dempster states that statistics can be used to show that some data analysis techniques have attractive properties, or to show how data analysis techniques can be improved. In the book itself we do not find any examples of this type of interaction.

Before he starts his treatment of MVA Dempster, like Van de Geer, gives an extensive introduction to the theory of matrices and finite-dimensional linear spaces. The introduction tries to avoid the use of coordinates, and is consequently very geometrical. The computational translation of the linear operations is discussed separately. Dempster also gives a list of three important omissions.
a: Categorical data.
b: Specialized techniques such as factor analysis and structural equations.
c: Cluster analysis and related subjects.
As a consequence of these omissions Dempster's data analysis is limited to the discussion of various computational aspects of the linear model. He mentions principal components analysis and canonical analysis, there is an example of the explorative use of canonical analysis, but these typical multivariate techniques do not get much attention.

1.1.8 Tatsuoka (1971).

There are two different definitions of MVA on the first page of this book. "Multivariate statistical analysis, or multivariate analysis for short, is that branch of statistics which is devoted to the study of multivariate (or multi-dimensional) distributions and samples from those distributions." (1c, p 1). Tatsuoka says that this is the definition used by the statistician, and that it is not very useful because it sounds tautological. We may add that it also sounds rather imperialistic to define multivariate analysis as a mere abbreviation of multivariate statistical analysis. The second definition is the one used by the data analyst. "In applied contexts, particulary in educational and psychological research, multivariate analysis is concerned with a group (or several groups) of individuals, each of whom possesses values or scores on two or more variables such as tests or other measures. We are interested in studying the interrelations among these variables, in looking for possible group differences in terms of these variables, and in drawing inferences relevant to these variables concerning the populations from which the sample groups were chosen." (1c, p 1).

The book contains much matrix algebra, and a fair amount of statistics. There is considerably less geometry as in Van de Geer or Dempster, the approach of the linear model is the same as in Anderson or Morrison. Familiar univariate techniques are generalized as far as possible. "Pointing out the analogy between a given multivariate technique and the corresponding univariate method is one of the principal didactic strategies used throughout this book." (1c, p 3). The main disadvantage of this approach, as we have seen before, is that the multivariateness of the data is treated as a sort of nuisance, and that it is difficult to fit typical multivariate techniques without univariate analogues

into the framework. Principal components analysis consequently does not get much attention, canonical analysis is discussed fairly extensively, but the treatment is statistical, as in Anderson, and based on the multinormal model.

1.1.9 Harris (1975)

There is a certain pattern in our content analyses so far. In the more recent books the data analytic aspects of MVA get more attention, and the limitations of inferential multinormal analysis are stated more clearly. The book by Harris is in some respects a reaction, a sort of inferential backlash. There is an interesting first chapter, defending the inferential approach. Harris even has the courage to defend the null hypothesis and the associated significance test, while this illustrious duo is considered to be dead and buried in most data analytic circles. We do not discuss his arguments in detail here, but we shall come back to some of them further on in this chapter.

Harris' reasoning roughly goes as follows. Statistical methods are a form of quality control for scientific production. They are necessary because it has been demonstrated many times that the opinion of the investigator is not a sufficient basis for believing in the generalizability of his results. We need formal methods to study generalizability. Formal methods can be classified as descriptive or inferential. Inferential methods are designed specifically to prevent that conclusions are drawn which are not generalizable, conclusions which are only a consequence of particularities of the specific sample or the specific technique. "As will become obvious in the remaining sections of this chapter, the present Primer attempts in part to plug a 'loophole' in the current social control exercised over researchers' tendency to read too much from their data. It also attempts to add a collection of rather powerful techniques to the descriptive tools available to behavioural researchers. Van de Geer (1971) has in fact written a textbook on multivariate statistics which deliberately omits any mention of their inferential applications." (1c, p 4-5).

What are the tools that Harris uses for this plumbing job ? In the first place he chooses the now familiar starting point that MVA-techniques generalize the existing univariate techniques. They often do this by replacing groups of variables by single variables which are linear combinations of the variables in the group. The coefficients of the linear combinations are chosen to optimize some intuitively or statistically attractive criterion. Data analysts are interested primarily in the optimal coefficients, they call them 'loadings' (of the observed variables on the constructed variables), statisticians are interested primarily in the maximum value of the criterion. But it is clear that the two approaches complement each other, one way of computing the maximum

is by explicitly computing the maximizer. Harris consequently emphasizes optimality criteria with a relatively simple statistical interpretation, which lead naturally to optimum coefficients. Thus he prefers the union-intersection approach of Roy (1957) to the more usual likelihood-ratio method, because this last method usually does not give a unique set of coefficients and consequently frustrates the data analyst. Matrix algebra does not get much attention in Harris' book, because it is only a technique for efficient optimization, and obviously there is almost no geometry in the book. Multivariateness is a nuisance, even more so than in Tatsuoka's book, multivariate data must be made univariate as soon as possible. Harris' attempt to integrate multivariate statistical analysis and multivariate data analysis is interesting, but doomed to fail. We have already seen that data analysts are not primarily interested in coefficients, but in pictures. A single set of coefficients does not give very interesting pictures.

1.1.10 Dagnelie (1975)

It is nice to read the usual definition in French for a change, but if we translate it back into English it sounds a bit disappointing. "In a general sense analysis of several variables or multidimensional analysis or multivariable analysis or multivariate analysis can be interpreted as a set of statistical techniques which have the purpose of studying the relationships that exist between several dependent or interdependent variables." (lc, p 11). Dagnelie gives this definition after comparing the definitions given by Anderson, Cooley and Lohnes, Kendall, Morrison, and Press. This is somewhat disappointing, his book is oriented heavily to the anglo-american literature, it does not introduce anything new, and it does not show a typical French approach to MVA. Dagnelie discusses the usual MVA techniques, using the classification criterion that for each technique the variables are partitioned into sets containing one or more variables. There is not much algebra, not much geometry, a fair amount of cookbookery statistics, and there are some nice examples from ecology. "The user obviously must have a sufficiently precise idea what the general principles and the consitions for application of these methods are, but he must primarily concentrate his attention on the interpretation of the results he has obtained." (lc, p 18).

1.1.11 Green and Carroll (1976)

This is a different type of book. At first sight it does not even belong in our list, because its explicit purpose is to add some useful tools to the mathematical toolbox of the researcher. On second sight, however, the book wants considerably more, and it can be interpreted as fitting in the trend

we have discovered already in Van de Geer. MVA is applied linear algebra and linear geometry. There is a nice introductory chapter, in which the ideas of the authors are explained. We give some representative quotations. "Completion of this book should provide both a technical base for tackling most applications-oriented multivariate texts and, more importantly, a geometric perspective for aiding one's intuitive grasp of multivariate methods." (1c, p xii). "In function, as well in structure, multivariate techniques form a unified set of procedures that can be organized around relatively few prototypical problems." (1c, p 1). "The heart of any multivariate analysis consists of the data matrix, or in some cases, matrices. The data matrix is a rectangular array of numerical entities whose informational content is to be summarized and portrayed in some way." (1c, p 3). "To a large extent, the study of multivariate techniques is the study of linear transformations." (1c, p 14).

The message is clear. If these mathematical tools are understood, then one can recognize MVA techniques in their various disguises, and one can construct or choose the appropriate MVA technique that one needs in any practical situation. There is no statistics in the book. In the preface we get the impression that Green and Carroll see their book as an introduction to Tatsuoka, or Harris, or Morrison, but it is quite possible that some readers of the book come to the conclusion that they do not need any additional education. This is why we have interpreted the book as a book about MVA, and not only about the mathematical toolbox. Of course Green and Carroll may not agree with our point of view.

1.1.12 Caillez and Pages (1976)

In France data analysis is very popular, mainly because of the work and the influence of J.P. Benzécri. The book by Caillez and Pages is about this french form of data analysis, which differs in some important respects from the anglo-american form. French linear algebra, for example, is considerably more modern and abstract, and uses coordinates and matrices to a far lesser extent. This is in the tradition of modern french mathematicians such as Bourbaki, Dieudonne, Cartan, and Choquet. The book by Caillez and Pages has a very extensive introduction in this type of linear algebra, and has very extensive chapters on regression, principal components analysis, canonical analysis, and on Benzécri's version of metric multidimensional scaling. There is also a chapter on correspondence analysis, a technique for nonlinear principal components analysis developed by Benzécri, which is also one of the main techniques discussed in our book. There is a useful introductory chapter on sets, relations, and functions, and a useful final chapter on classification and clustering, but by far the largest part of the book is on applied linear

algebra. "We have made a definite choice in our presentation of data analytic techniques: we constantly use the notion of a projector, of an M-symmetric application, of quantification, and above all the 'duality diagram', which is an efficient and comprehensive notational device to present all aspects of the techniques of linear algebra." (lc, p vii). The book has an interesting preface by G. Morlat, which discusses the differences between data analysis and classical statistics in considerable detail. We shall come back to that discussion later in the chapter.

1.1.13 Giri (1977)

Giri's starting point is the invariance of testing problems under groups of transformations, the book can be interpreted as an application of aspects of decision theory in multivariate situations. The emphasis is on hypothesis testing, not on estimation. Everything is normally distributed. There is some material on matrix algebra and transformation groups in the book, and very little material on principal components analsysis, factor analysis, and canonical analysis.

1.1.14 Gnanadesikan (1977)

The definition of MVA is the usual one, but the approach used in the book is quite unusual. Gnanadesikan is a data analyst. "Much of the theoretical work in multivariate analysis has dealt with formal inference procedures, and with the associated statistical distribution theory, developed as extensions of and by analogy with quite specific univariate methods, such as tests of hypotheses concerning location and/or dispersion matrices. The resulting methods have often turned out to be of very limited value for multivariate data analysis." (lc, p 2). Gnanadesikan emphasizes graphical representations, but of a slightly different nature than the ones we have met earlier. The book uses probability plots and related graphics, there is less linear algebra than in Caillez and Pages and less matrix algebra than in Van de Geer. There is considerable attention for goodness-of-fit (of the multinormal distribution) and for outlier-detection. Gnanadesikan lists the most important purposes of MVA techniques.

a: Reduction of dimensionality.

b: Study of dependence.

c: Multidimensional classification.

d: Investigation of statistical models.

e: Data reduction and clarification.

He also mentions the most important problems, in a list that resembles the one given by Kendall.

a: It is difficult to find out what the client wants to know exactly.

b: If we have decided on going multivariate, it is difficult to determine the number of variables.

c: Even with the present generation of computers there are a number of MVA techniques which can only be applied on data with a relatively small number of variables and/or individuals.

d: Multidimensional pictures and graphs can easily become complicated and confusing.

e: There is no natural order in multidimensional space.

1.1.15 Kshirsagar (1978)

This is a thick book, and also a very good book, but nevertheless we can be brief. Multinormal statisticians have been very active since Anderson. A lot of computing has been done, many expansions have more terms now than in 1958. New test statistics have been thought out, and new techniques have been developed to derive complicated distributions more elegantly. Kshirsagar's book is the most complete summary of these multinormal results so far. There are no examples, there is no factor analysis, principal components analysis and canonical analysis are treated multinormally. "The theory of multivariate analysis developed so far almost invariably assumes that the joint distribution of the random variables is a multivariate normal distribution." (1c, p 1). If one defines multivariate analysis as multinormal analysis this is a trivial conclusion, if one defines multivariate analysis as multivariate statistical analysis it is incorrect because of multivariate multinomial analysis already discussed by Kendall and Roy. Kshirsagar mentions as most important omissions factor analysis, multiple time series, and categorical data. He thinks that regression analysis is the most important statistical technique, which is a typical multinormal point of view. On the other hand he does mention data reduction as an important purpose of MVA. "The aim of the statistician undertaking multivariate analysis is to reduce the number of variables by employing suitable linear transformations and to choose a very limited number of the resulting linear combinations in some optimal manner, disregarding the remaining linear combinations in the hope that they do not contain much significant information. The statistician thus reduces the dimensionality of the problem." (1c, p 2). As in the book of Harris this point of view dictates the preferred technique. Harris, a one-dimensional multivariate analyst, uses the union-intersection approach and the largest root criterion. Kshirsagar uses the Bartlett-partitioning of the multinormal likelihood ratio statistic.

1.1.16 Thorndike (1978)

Thorndike considers himself to be a data analyst, like Cooley and Lohnes. In his data analytical work he often meets the situation that his clients have a fair comprehension of analysis of variance techniques, but only a little understanding of techniques based on correlation coefficients. It is not difficult to guess that Thorndike's clients have been misformed by the ancient rituals usually known as 'statistics for psychologists'. Thorndike proposes his book as an antidote. "The approach is largely intuitive and geometric rather than mathematical..." (lc, p vi). This is a somewhat unfortunate distinction, the next quotation is less controversial. "The organization of the book reflects a conviction that understanding can be most readily developed by showing the essential unity and orderly progression of concepts in multivariate statistics. ... The geometric interpretation of the correlation coefficient as the angle between two vectors in a people space is readily generalizable to multiple and canonical correlation." (lc, p vii). There is very little mathematics in the book, as promised. "The work of other authors, notably Quinn McNemar, Harry Harman, and Bill Cooley and Paul Lohnes, is frequently cited for the reader who wants or needs the mathematical foundations of the topics discussed." (lc, pag vii). Incredible, but this is a real quotation. Thorndike's book could very well become the book on MVA for the fifties. It is remarkable that the reader who needs the mathematical foundations must go a very long way. Thorndike refers him to Cooley and Lohnes, who refer him to Tatsuoka, who refers him to Morrison, who refers him to Anderson. In the meantime the unfortunate seeker has gone back twenty years in time, and has finally arrived at a book of unquestionable quality, but completely about multinormal analysis. It would perhaps be better to refer directly to a book about linear algebra.

Thorndike also uses the Bartlett-Kendall distinction between analysis of interdependence and dependence. He calls this internal and external factor analysis, which makes multiple regression a form of external factor analysis, and which makes factor analysis a form of internal factor analysis. There is also some useful information in the book, for example about interpretation of canonical analysis results.

1.? Content analysis of tables of content

A somewhat more objective analysis of the contents of the MVA books we have discussed is also possible. In table 1 we have indicated the number of pages in the books for each of the following seven subjects.

A: MATH:
Mathematics other than statistics, i.e. linear algebra, matrices, transformation groups, sets, relations.

B: CORR:
Correlation and regression, includes path analyses, linear structural and functional equations.

C: FACT:
Factor analysis and principal components analysis.

D: CANO:
Canonical correlation analysis.

E: DISC:
Discriminant analysis, classification, cluster analysis.

F: STAT:
Statistics, includes distributional theory, hypotheses testing, and estimation. Also statistical analysis of categorical data.

G: MANO:
MANOVA, and the general multivariate linear model.

Of course it is easy to criticize this classification. There are some subjects in the same category that are only marginally related (such as discriminant analysis and cluster analysis). Moreover canonical analysis in Anderson or Kshirsagar is quite different from canonical analysis in Van de Geer or Caillez and Pages. We have chosen this classification because we wanted to be able to classify most of the material, while we do not want too many zero entries in the data matrix. Of course our selection of MVA books is also not exactly a random sample. We have used some of the books that are used most, some personal favourites, some rather extreme books, and some others which happened to be in the library at the time. The entries in table 1 are probably also not very reliable, because we have added number of pages of various chapters, and chapters are usually not completely about one single subject. Somebody else, using the same categories, would certainly arrive at different page counts. Not too different, we hope.

We applied correspondence analysis to table 1. We shall not discuss the precise nature of this technique here, this will be done in chapter 4 of this book. For now it is sufficient to know that correspondence analysis is some sort of multidimensional scaling method, which represents row- and columnpoints in a low-dimensional joint space. Two books are close together in the space (or in the picture) if they have similar contents, two subjects or topics are close if they occur in the same books in the same degree, and a book is close to a subject if the book pays relatively much attention to the subject. In correspondence

		A	B	C	D	E	F	G
1	Roy (1957)	31	0	0	0	0	164	11
2	Kendall (1957)	0	16	54	18	27	13	14
3	Kendall (1975)	0	40	32	10	42	60	0
4	Anderson (1958)	19	0	35	19	28	163	52
5	Cooley & Lohnes (1962)	14	7	35	22	17	0	56
6	Cooley & Lohnes (1971)	20	69	72	33	55	0	32
7	Morrison (1967)	74	0	86	14	0	84	48
8	Morrison (1976)	78	0	80	5	17	105	60
9	Van de Geer (1967)	74	19	33	12	26	0	0
10	Van de Geer (1971)	80	68	67	15	29	0	0
11	Dempster (1969)	108	48	4	10	46	108	0
12	Tatsuoka (1971)	109	13	5	17	39	32	46
13	Harris (1975)	16	35	69	24	0	26	41
14	Dagnelie (1975)	26	86	60	6	48	48	28
15	Green & Carroll (1976)	290	10	6	0	8	0	2
16	Caillez & Pages (1976)	184	48	82	42	134	0	0
17	Giri (1977)	29	0	0	0	41	211	32
18	Gnanadesikan (1977)	0	19	56	0	39	75	0
19	Kshirsagar (1978)	0	22	45	42	60	230	59
20	Thorndike (1978)	30	128	90	28	48	0	0

table 1.1: number of pages of MVA books devoted to several subjects.

1	ROIJ	-1.1086	-0.6144	-0.3390
2	KEN1	0.0740	0.7025	0.2526
3	KEN2	-0.2115	0.4605	-0.4923
4	ANDE	-0.7779	-0.1107	0.1556
5	COL1	0.0278	0.4065	1.0513
6	COL2	0.3578	0.6960	0.0928
7	MOR1	-0.1641	-0.1572	0.4635
8	MOR2	-0.2502	-0.1963	0.3900
9	GEE1	0.7279	-0.1945	-0.0475
10	GEE2	0.6840	0.2434	-0.1796
11	DEMP	0.0273	-0.3665	-0.4430
12	TATS	0.2680	-0.4475	0.2829
13	HARR	0.0219	0.5089	0.5172
14	DAGN	0.1205	0.4846	-0.1948
15	GREC	1.0831	-0.0321	0.0321
16	CAPA	0.6496	-0.0708	-0.1327
17	GIRI	-0.9835	-0.3927	-0.2502
18	GNAN	-0.4001	0.3292	-0.3383
19	KSHI	-0.7473	0.0810	-0.0051
20	THOR	0.5655	0.8145	-0.3526
A	MATH	1.1440	-1.5268	0.0692
B	CORR	0.8046	1.3602	-1.2504
C	FACT	0.2928	0.9827	0.6468
D	CANO	0.2605	0.8653	0.8399
E	DISC	0.2555	0.4777	-0.7914
F	STAT	-1.5635	-0.4972	-0.6071
G	MANO	-0.6745	0.2086	2.5069
LABDA		0.6038	0.5133	0.3364

table 1.2:

profections of books, profections of subjects and singular values from correspondence analysis on table 1.1.

analysis as we have used it each book-point is a weighted mean of subject points, with the weights given by the entries of table 1. Thus table 1 tells us that ROIJ = {(31×MATH) + (164×STAT) + (11×MANO)} / (31 + 164 + 11). It follows that the books are all in the convex hull of the subject points.

The projections of the books and the subjects on the three most important dimensions are given in table 2. Correspondence analysis is a singular value-singular vector technique, in which the importance of a dimension is defined as the value of the corresponding singular value. The three largest singular values are given in the last row of table 2 , the three remaining singular values are .2096, .1748 and .1007. The projections on the first two dimensions are also represented in figure 1.1. In this figure we have drawn a triangle with corners CORR, STAT and MATH. The clearest, most pure representatives of these three subjects are, respectively, THOR, ROIJ and GREC. Most of the books are on the line between CORR and STAT, these are the most classical books which vary, mainly, in degree of difficulty. Typical data analysis books are near the line between CORR and MATH, at least if we define data analyis in this context as applied linear algebra. Graphical data analysis as in GNAN fits better into the classical scale CORR-STAT. In the interior of the triangle we find four books which pay a lot of attention to both statistics and linear algebra. The book COL1 is not in the 'correct' position (i.e. not where we expect it, we expect it closer to COL2 and THOR), the third dimension in table 2 shows that this happens because COL1 gives much more attention to MANO than we expect on the basis of its 'technical level'. But of course COL1 is a manual for computing MVA solutions, and its treatment of MANO is not at all technical.

We have also repeated the analysis without the 'extremists' GREC and ROIJ. The projections on the first two dimensions are shown in figure 1.2. The position of the remaining books stays roughly the same, but the first two singular values are now considerably smaller (.5555 and .3841). The third singular value is .3333, about the same as before, and the third dimension continues to contrast COL1 with the rest. It is clear that correspondence analysis shows us the same dimensions as our content analysis, only in a more compact and comprehensive form. Of course this is due partly to our selection of the books.

1.3 A short summary and some problems

Some important points emerge from our discussion of the MVA books. In the first place there seems to be a certain difference, or even conflict, between the

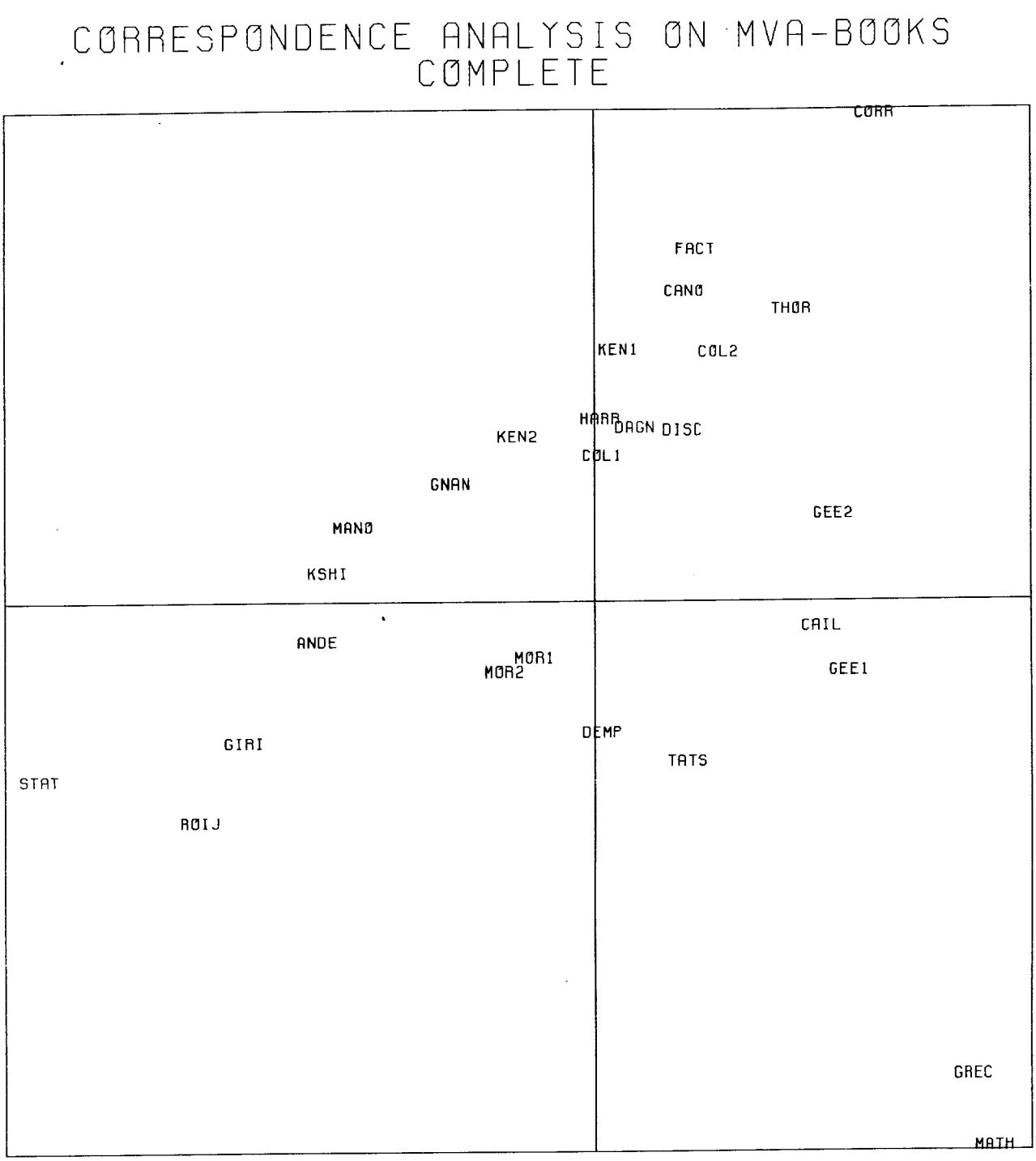
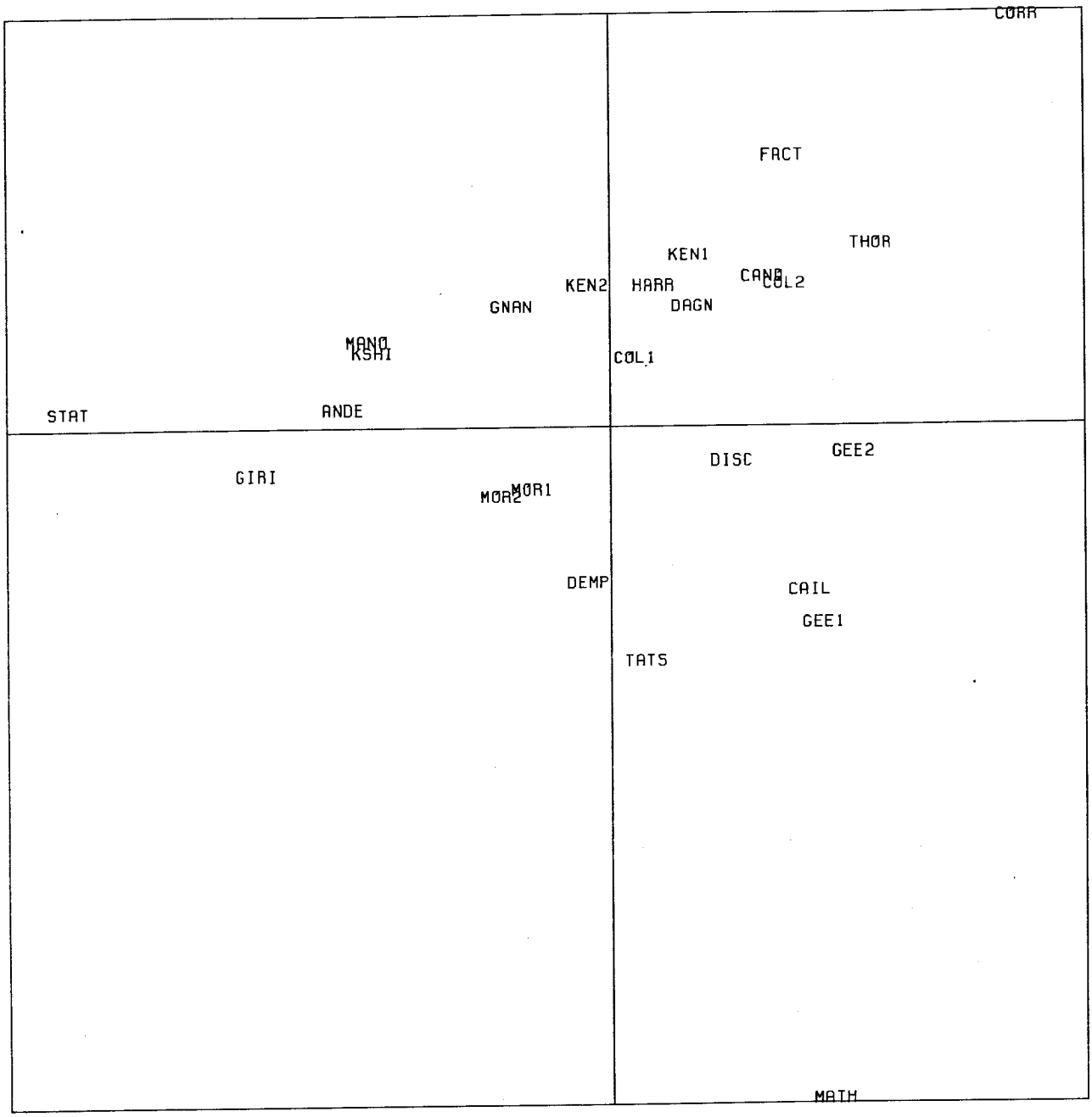

CORRESPONDENCE ANALYSIS ON MVA-BOOKS COMPLETE
CORR
FACT
CANO
THOR
KEN1
COL2
HARR
DAGN
DISC
KEN2
COL1
GNAN
GEE2
MANO
KSHI
CAIL
ANDE
MOR1
MOR2
GEE1
DEMP
GIRI
TATS
STAT
ROIJ
GREC
MATH
SINGULAR VALUES: λ1 = 0.6038 λ2 = 0.5133

figure 1.1

CORRESPONDENCE ANALYSIS ON MVA-BOOKS REDUCED



SINGULAR VALUES: $\lambda_1 = 0.5555$ $\lambda_2 = 0.3841$

figure 1.2

statistical and the data analytic approach to MVA. The statistical approach starts with a statistical model, usually based on the multinormal distribution. The model is assumed to be true, and within the model certain parametric hypotheses are constructed. The remaining free parameters are estimated, and the hypotheses are tested. The data analytic approach does not start with a model, but looks for transformations and combinations of the variables with the explicit purpose of representing the data in a simple and comprehensive, and usually graphical, way. It is certainly necessary to discuss the role and the meaning of the statistical model with the corresponding tests and estimators. More in particular we want to know more about the importance of the multinormal distribution for MVA.

A second problem is the importance of categorical data. Or, in other terms, the role of nominal and ordinal variables in MVA. The only books that pay at least some attention to categorical variables are the ones by Roy and Kendall, in the other books they only occur as codings in the context of MANOVA and discriminant analysis. In this case they are often called 'dummy variables'. In general they are interpreted as non-stochastic, they are part of the design matrix, and thus independent variables. Actually they are not even variables, they are only used to write down parametric multinormal models in compact matrix notation. A possible exception to this rule is the notion of quantification in the book by Caillez and Pages.

In our list of MVA books we have restricted ourselves to general, often introductory books, with a considerable amount of overlap. There are also a number of books that deal especially with the analysis of multivariate categorical data. The most important ones are Haberman (1974), Bishop, Fienberg, and Holland (1975), Gokhale and Kullback (1978), Fienberg (1977), and Goodman (1978). The content of these books is generally quite different from those of the books we have discussed. The models are formulated in the tradition of classical statistics, but the emphasis is on discrete distributions and on asymptotic methods. Linear algebra and geometry are not important at all. It is clear that these books constitute a completely different tradition, which is a bit strange because both forms of MVA start directly from the work of Fisher and Pearson. In classical general handbooks of statistics, such as those of Fisher, Yule and Kendall, Cramér, Kendall and Stuart, Wilks, and Rao, the two forms of MVA are both treated, but they are not or almost never related to each other. There is a gap between the discrete and the continuous approach.

Of course there also is an exception to this rule, the books by Kullback (1959) and Lancaster (1969). Kullback discusses statistical procedures based on the information-theoretic measures of divergence, which are distance measures between

parametric distributions, defined in exactly the same way for discrete and continuous distributions. To some extent the same thing is true for other work in theoretical statistics that relies heavily on general exponential models of which both the multinormal and the multinomial are special cases. Lancaster's book is partly about the decomposition of multivariate probability distributions by using orthogonal functions on the marginals, and this technique can be applied to both discrete and continuous distributions too. Another problem we shall have to discuss in this chapter is the relationship between discrete and continuous MVA.

There are also some books which are, at least in some respects, similar to our book. We mention them here, because they also do not fit into the list we used earlier in the chapter. Volume II of the treatise by Benzécri and others (1973) deals exclusively with theory and applications of correspondence analysis. And in the second place there is a very interesting recent book by Nishisato (1980) which deals with the applications of 'dual scaling' (we use the term 'optimal scaling' in our book) to various forms of categorical data. The amount of overlap with our book is considerable as far as results are concerned, mainly with our chapters 3 and 4. Our general approach to MVA, however, is quite different from Nishisato's.

1.4 Data analysis and statistics

1.4.1 Tukey's definition of data analysis

We start by discussing the ideas of John Tukey, who was the first to present data analysis as an independent discipline, distinct from statistics. "All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." (1962, p 2). "Large parts of data analysis are inferential in the sample-to-population sense, but these are only parts, not the whole. Large parts of data analysis are incisive, laying bare indications which we could not perceive by simple and direct examination of the raw data, but these too are only parts, not the whole. Some parts of data analysis, as the term is here stretched beyond its philology, are allocation, in the sense that they guide us in the distribution of effort and other valuable considerations in observation, experimentation, and analysis. Data analysis is a larger and more varied field than inference, or incisive procedures, or allocation." (1962, p 2). Tukey's reasons for not wanting

to be called a statistician any more are clear from these quotations. In the first place data analysis is more general than inferential statistics, and in the second place there are parts of (mathematical) statistics which are outside data analysis. "To the extent that pieces of mathematical statistics fail to contribute, or are not intended to contribute, even by a long and tortuous chain to the practice of data analysis, they must be judged as pieces of pure mathematics, and criticized according to its purest standards. Individual parts of mathematical statistics must look for their justification toward either data analysis or pure mathematics. Works which obey neither master, and there are thos who deny the rule of both for their own work, cannot fail to be transient, to be doomed to sink out of sight." (1962, p 3).

Tukey is obviously not very happy with the historical development of statistics. "What is needed is progress, and the unlocking of certain of rigidities (ossifications ?) which tend to characterize statistics today. Whether we look back over this century, or look into our crystal ball, there is but one natural chain of growth in dealing with a specific problem of data analysis, viz:

(a1') recognition of problem,
(a1'') one technique used,
(a2) competing techniques used,
(a3) rough comparisons of efficacy,
(a4) comparison in terms of a precise (and thereby inadequate) criterion,
(a5') optimization in terms of a precise, and similarly inadequate criterion,
(a5'') comparison in terms of several criteria.

(Number of primes does not indicate relative order).

If we are to be effective in introducing novelty, we must heed two main commandments in connection with new problems:

(A) Praise and use work which reaches stage (a3), or only stage (a2), or even stage (a1'').
(B) Urge the extension of work from each stage to the next, with special emphasis on the earlier stages.

One of the signs of the lassitude of the present cycle of data analysis is the emphasis of many statisticians upon certain of the later stages to the exclusion of the earlier ones. Some, indeed, seem to equate stage (a5') to statistics - an attitude which if widely adopted is guaranteed to produce a dried-up, encysted field with little chance of real growth." (1962, p 7).

The point of view that (a5') can be identified with statistics is discussed somewhat more in detail. "The view that 'statistics is optimization' is perhaps but a reflection of the view that 'data analysis should not appear to be a matter of judgment'. Here 'appear to' is in italics because many who hold this view would

like to suppress these words, even though, when pressed, they would agree that the optimum does depend on the assumptions and criteria, whose selection may, perhaps, even be admitted to involve judgment." (1962, p 9). "In data analysis we must look to a very heavy emphasis on judgment. At least three different sorts or sources of judgments are likely to be involved in almost every instance:

(a1) judgment based upon the experience of the particular field of subject matter from which the data come,

(a2) judgment based upon a broad experience with how particular techniques of data analysis have worked out in a variety of fields of application,

(a3) judgment based upon abstract results about the properties of particular techniques, whether obtained by mathematical proofs or empirical sampling."

(1962, p 9). "The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: ' Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.' Data analysis must progress by approximate answers, at best, since its knowledge of what the problem really is will at best be approximate." (1962, p 13-14).

Tukey's paper has been very influential, for several reasons. In the first place, as we have seen in section 1.1.4, it gave a number of people their professional identity. "It was a stroke of genius to realize that to render ' a deed without a name' respectable, you should name it (or perhaps I should say rename it), and we are all grateful for the name 'Data Analysis'. This important part of our subject can now be studied without apology or shame, and courses on it are taught and may be attended by consenting adults." (Box, 1979, p 3). In the second place Tukey's paper has pointed to some really controversial problems in the development of mathematical statistics. His conclusions have been repeated and stated even more forcefully by Wolfowitz (1969). "Except perhaps for a few of the deepest theorems, and perhaps not even these, most of the theorems of statistics would not survive in mathematics if the subject of statistics itself were to die out. In order to survive the subject must be more responsive to the needs of application." (1969, p 748). "What we must guard against is the development of a theory which, on the one hand, bears little or no relation to the actual problems of statistics, and which, on the other hand, when viewed as pure mathematics, is not interesting per se nor likely to survive." (1969, p 749). It is also clear that the same problems are the reason why the Annals of Mathematical Statistics have been replaced by two journals: the Annals of Probability, and the Annals of Statistics.

On the other hand there has always been a considerable difference between 'English' and 'American' statisticians. Wolfowitz (1969, p 745) severely

criticizes Barnard, because he has questioned the value of rigorous proofs in one of his papers. Kiefer (1964) is highly critical of the monumental treatise of Kendall and Stuart. "My main criticism of the book is that it has a very high density of errors in statements and proofs. Another negative aspect is the exclusion, in a work of this encyclopaedic nature, of much of the content and almost all of the spirit of modern mathematical statistics." (1964, p 1371). Kendall, on the other hand, views the situation quite differently "The early statisticians of the present century were competent at mathematics, but they were not great creative mathematicians. Karl Pearson was trained in mathematics, but Edgeworth was a classical scholar and Yule an engineer by training. Fisher, who was a creative mathematician, criticized his predecessors for the clumsiness of their style; but even he wrote in the tradition of English mathematics, which does not care much about extreme generalization or extreme rigor as long as it gets the right answer to its problems. The consequence was that, with a few exceptions, theoretical statistics in the forties could be understood by anybody with moderate mathematical attainment, say at the first year undergraduate level. I deeply regret to say that the situation has changed so much for the worse that the journals devoted to mathematical statistics are now completely unreadable. Most statisticians deplore the fact, but there is not very much they can do about it." (1972, p 205). As a consequence the English statisticians, who do not identify with the Annals and with the emphasis on extreme generality, are not in favor of replacing 'statistics' by 'data analysis'. "Whereas the contents of Tukey's remarks is always worth pondering, some of his terminology is hard to take. He seems to identify 'statistics' with the grotesque phenomenon generally known as 'mathematical statistics', and finds it necessary to replace 'statistical analysis' by 'data analysis'." (Anscombe, 1967, p 3). "The elevation of Data Analysis to its proper place as a subject meriting serious study makes me as happy as I would be if some neglected but important activity of the carpenter, such as the use of the saw or the chisel, had at last received proper recognition and study. But my enthusiasm for the naming of Data Analysis does not extend to the renaming of Statisticians as 'Data Analysts', any more than I should be happy to hear a carpenter described as a sawyer or a chiseler." (Box, 1979, p 3). It is clear that Box, in this quotation, identifies data analysis with what Tukey has called exploratory data analysis, which corresponds with the incisive aspects of general data analysis of his 1962 paper.

In order to prevent possible misunderstandings we also state our own position on these issues. We agree with Anscombe and others that data analysis in the broad sense of Tukey's 1962 paper is identical with statistics, and that it is consequently unnecessary to rename all statisticians. We agree with Tukey

and Wolfowitz that mathematical statistics is interesting to the extent that it has practical applications or consequences, but we also think that even very abstract and general work can have these practical consequences, although it will sometimes be difficult to trace the exact path of influence that a paper, or class of papers, has. To be more precise, it is very difficult for us to point out any one paper in the Annals which has surely not influenced the practice of data analysis. We agree with Kiefer and Wolfowitz that if a proof is given, then it should be rigorous. Many of the controversies over 'what Fisher meant', both in statistics and genetics, are simply due to the fact that Fisher skipped steps, and did not state restrictions or assumptions. We shall come back to the role of optimization in a later section.

1.4.2 Benzécri's definition of data analysis

The proper translation of 'Analyse des données' seems to be data analyses. But the principles of French data analysis, explained by Benzécri (1973), are quite different from those of Tukey's school. We translate them first, and then give some brief comments.

Principle 1: Statistics is not the same thing as probability theory. Under the name 'mathematical statistics' several writers (who, I tell you this in French, seldom write in our language ...) have built a pompous discipline, which abounds in hypotheses that are never satisfied in practice. We cannot expect a solution of our typological problems from these authors.

Principle 2: The model must follow the data, and not the other way around. This is another error in the application of mathematics to the human sciences: the abundance of models, which are built a priori and then confronted with the data by what one calls a 'test'. Often the 'test' is used to justify a model in which the number of parameters to be fitted is larger than the number of data points. And often it is used, on the contrary, to strongly reject as invalid the most judicial remarks of the experimenter. But what we need is a rigorous method to extract structure, starting from the data.

Principle 3: It is convenient to treat simultaneously information on as many dimensions as possible. As a consequence the problem of the validity of a 'test' - which, we admit, is sometimes troublesome - is not that important anymore. Nobody knows if the inequality $0.5 \neq 0.7$ must be interpreted in these practical cases as a certain empirical result, or if it is only a result of chance. But finding that in a space of two dimensions fifty points are approximately arranged on a circle is certainly a discovery (at least if the computing method does not deceive us !).

Principle 4: For the analysis of complex facts and notably for the analysis of social facts we cannot do without the computer. This principle is obviously true ...

but what would our Gaulish fathers have thought about it fifteen years ago ?
Principle 5: Using a computer means that all techniques designed before the advent of automatic computing must be abandoned. I say techniques, not science: the geometric and algebraic principles of our programmes were known to Laplace, 150 years ago. But Laplace also was the author of a treatise on celestial mechanics, that has just been republished for usage by space technicians ... And that treatis did not suffice for Napoléon to conquer the moon !
The translation is as literal as possible, the principles are on page 3,6,9,12,15 of Benzecri (1973).

The problem with these principles in general is that Benzécri seems to imply that large, unstructured multivariate data sets are the only types that are possible. Principles 3,4,5 only seem relevant in that context. Another problem, with principles 2,3, is that if they are applied consequently they lead to the type of blind empiricism that has terrorized psychometrics for a long time without any lasting results. If we continue to add additional variables, without the guidance of any theory or control whatsoever, then we end up with a lot of explained variance but with no possibilities for prediction. We do need theory, or, if you prefer, a model, because we need sensible rules on how to add variables. If the structure is there, it will become more clear if we add variables from the same domain. It will disappear, if we add enough variables from other domains. In the context Benzécri means, we agree with principles 2 and 3 and their corollaries. We shall come back both to tests and to the empiricism-rationalism problem in a later section of this chapter.

Principle 1 is quite important. The important part is stated more explicitly in Benzécri (1973, p 6): "The mathematical foundations of statistical analysis are more algebraic and geometric (and if many dimensions are involved geometric ideas merge with algebraic calculations) than probabilistic; it is better to speak of the average and the principle axes, etc...which are defined in terms of a finite number of actual data points, than to speak of the expected value, etc...defined in a potentially infinite universe. But probabilistic concepts and ideas can suggest algebraic operations and sometimes can be used to evaluate their usefulness." Compare this with the Kendall-quotation on the 'classical probabilistic mould' in 1.1.2, and also with: "A common feature of problems of this kind is that the stochastic model so central to classical statistical analysis is either absent altogether or is playing a very subordinate role. This, in my view, is inherent, and nothing could be more misguided than an attempt to force such problems into an ill-fitting classical statistical mould." (Sibson, 1972, p 311). It is also interesting that Rao's excellent 'Linear statistical inference and its applications' of 1965 has a 50-page

introductory chapter on the algebra of vectors and matrices and a 50-page introductory chapter on probability theory. Our content analysis of the books on MVA in 1.2 also shows the importance of distinguishing MATH and STAT.

1.4.3 Robust statistics

The important review papers by Huber (1972), Hampel (1973), and Bickel (1976) clearly show that the robust statisticians are trying to make a step in the direction of data analysis. "Often in statistics one is using a parametric model, such as the common model of normally distributed errors, or that of exponentially distributed observations. Classical (parametric) statistics derives results under the assumption that these models are strictly true. However, apart from some simple discrete models perhaps, such models are never exactly true." (Hampel, 1973, p 87-88). This is similar, although much more moderate, to Benzécri's Principle 1, it is also similar to: "The hallmark of good science is that it uses models and 'theory' but never believes them." (Wilk, quoted by Tukey, 1962, p 7). "We should also remember that we never know the exact distribution of ordinary data; and even if we did, or as far as we do, there remain serious questions about how to handle the excess knowledge of details. After all, a statistical model has to be simple (where 'simple', of course has a relative meaning, depending on state and standards of the subject matter field); Ockham's razor is an essential tool for the progress of science." (Hampel, 1973, p 89). Thus robust statistics defines more general 'supermodels' and finds estimates which are as good as possible for all models in the supermodel. A very important tool is the study of measures of robustness of estimators, developed mainly by Hampel. "If we look briefly to some related fields, we can recognize parallel developments there. Robustness may be viewed as a set of stability requirements, analogous to stability of ordinary differential equations, for example. Only quite recently, numerical analysts have become dissatisfied with such possibilities as obtaining a negative variance with the familiar formula $\{\sum X_i^2 - (\sum X_i)^2/n\}/(n - 1)$ and a correctly operating computer; so they started to investigate numerical stability. And in parts of prediction and control theory, engineers resorted to 'sub-optimal solutions' after finding the 'optimal' solutions offered to them inappropriate in practice." (Hampel, 1973, p 90). We must remember from the quotations of Hampel that robust statistics does not trust the usual parametric models anymore (although it trusts the supermodels), that the emphasis on optimization is replaced partly by an emphasis on stability.

Bickel remarks that the history of statistics has been marked by factional strives between various schools involved with the so-called foundations. "A conflict which I find more significant and which has serious consequences in

the field rather than just in academia is that between what I would like to call the optimists and the pessimists. The optimist confronted with a problem uses his intuition, physical knowledge, perhaps his prior opinions to construct a mathematical model for the data. Once constructed the model is unchangeable and analysis proceeds using methods optimal for that model. The pessimist, lately called a data analyst, uses models only provisionally and is always ready to change his view of the structure of the data in the light of the values that he sees. Purists of either persuasion are fortunately rare." (Bickel, 1976, p 145). "I view robustness ideas as attempts to bridge the gap between these points of view." (1976, p 145). We think that use of 'optimists' and 'pessimists' is somewhat unfortunate. In many social science situations optimism is not enough, mathematical models in these situations can only be constructed by idealists. It is also not true that purists of either persuasion are rare, there are many fanatical optimists in mathematical statistics.

There is another interesting problem we mention in this context. The optimists optimize the methods given the model. It has been pointed out by Tukey, who quotes Mallows and Barnard, that it is equally rational to optimize models given the methods (Tukey, 1962, p 10). The Mallows-quotation is: "But it seems to me to be no more reprehensible to start with an intuitively attractive design and then to search for optimality criteria which it satisfies, than to follow the approach of the present paper, starting from (if I may call it so) an intuitively attractive criterion, and then to search for designs which satisfy it." The paper Mallows mentions is on optimal designs by Kiefer, the discussion seems to indicate that English statisticians are natural pessimists.

Box (1979) discusses robustness in a wider perspective. "Some of us have had a preoccupation with optimal or best procedures. But the best, of course, is not necessarily very good. For instance, to bring in the aspect of everyday live, if I ever had to decide between cutting my throat with a razor blade or with a rusty nail, I suppose I would choose the razor blade. But although not strictly relevant to the problem as posed, one question that might cross my mind would be, 'Have I considered all my options ?' A principle that is being given more attention these days is that of 'robustification'. Here one doesn't attempt to guarantee that things will be optimal over some tractable, but perhaps very narrow set of circumstances. Instead one tries to ensure that they will be fairly good over a wide range of possibilities likely to happen in practice. Look at the human hand, for example. I doubt if there is any single thing that it does that could not be done better by some special instrument, but it is very good at doing a very large number of things that come up in

facing the world as it actually is. Another way to say this is that there is really nothing wrong with optimization per se, but that we ought to try to optimize over that distribution of circumstances which the world really presents to us. The mistake is choosing the best over too narrow a set of alternatives, suboptimization. It is sometimes argued that by doing simplified exercises we can at least optain useful pointers. However, I feel that such pointers are very likely to indicate the wrong direction, as might be true in the case of the razor blade and the rusty nail." (Box, 1979, p 1). Again we see the shifting emphasis from optimization to stability (robustness).

1.4.4 Exploration and confirmation

These two terms are becoming very popular these days. But again they are used by different people with various different meanings. We have exploratory factor analysis and exploratory multidimensional scaling, we also have confirmatory factor analysis and confirmatory multidimensional scaling. The difference between the two in this context is that the exploratory models are the most general factor analysis and scaling models, the confirmatory models impose restrictions on the parameters and representations, presumably by incorporating prior knowledge. On the other hand the point of view can also be defended that both factor analysis and multidimensional scaling are inherently exploratory, because they impose relatively little structure, deal with a large number of variables, and because constraint equations can also be used in a tentative (exploratory) way. And finally confirmatory can be used as meaning the same thing as inferential, which makes, for example, all Jöreskog's factor analysis programs and Ramsey's multidimensional scaling programs confirmatory, also the ones which are called exploratory in the first meaning we have discussed.

In order to explain in which sense we use the words, we start with some recent 'definitions' by Tukey. "If we need a short suggestion of what exploratory data analysis is, I would suggest that

1. It is an attitude, AND
2. A flexibility, AND
3. Some graph paper (or transparencies, or both).

No catalogue of techniques can convey a willingness to look for what can be seen, whether or not anticipated. Yet this is at the heart of exploratory data analysis. The graph paper -and transparencies- are there, not as a technique, but rather as a recognition that the picture-examining eye is the best finder we have of the wholly anticipated." (Tukey, 1980, p 24). Tukey disagrees, for example, with Parzen, who proposed in a recent paper to identify exploratory data analysis with confirmatory nonparametric statistical data analysis (and to identify confirmatory data analysis with confirmatory parametric statistical

data analysis). Tukey objects strongly to bringing the two under a common denominator. "Replacing chicken by tuna in chicken salad will not give fish wings or train chickens to swim." (Tukey, in the discussion of Parzen, 1979, p 122). There are other interesting contributions in the discussion of Parzen's paper, and in a sense they have much in common. They tend to emphasize the point of view that scientific discovery is a process of 'conjectures and refutations', and that exploratory data analysis is there to provide us with the conjectures, while confirmatory data analysis is there to provide us with the refutations. Statistics should cover both activities, should provide both conjecturing techniques and refuting techniques, and should also emphasize the fact that refutation of a theory or model leads to new conjectures in the form of modifications of the model. Thus the process is circular, and not linear. "It is widely recognized that the advancement of learning does not proceed by conjecture alone, nor by observation alone, but by an iteration involving both. Certainly, scientific investigation proceeds by such iteration. Examination of empirical data inspires a tentative explanation which, when further exposed to reality, may lead to its modification. This modified explanation is again put into jeopardy by further exposure to reality, and so on, in a continued alternation bewteen induction and deduction. I am continually surprised that statisticians, even good ones, still seem to ignore this iterative aspect of investigation and talk as if the movement from an initial (perhaps ill-posed) question, to design, to data collection, to analysis of the data, to the answer were a one-shot affair." (Box, 1979, p 2). It seems to us that the words exploratory and confirmatory should be used in this sense of conjectures and refutations. Tukey puts this in a historical perspective. "Once upon a time, statisticians only explored. Then they learned to confirm exactly - to confirm a few things exactly, each under very specific circumstances. As they emphasized exact confirmation, their techniques inevitably became less flexible. The connection of the most used techniques with past insights was weakened. Anything to which a confirmatory procedure was not explicitly attached was decried as 'mere descriptive statistics', no matter how much we had learned from it. Today, the flexibility of (approximate) confirmation by the jackknife makes it relatively easy to ask, for almost any clearly specified exploration, 'How far is it confirmed ?'. Today, exploratory and confirmatory can -and should- proceed side by side." (Tukey, 1977, p vii).

We also want to emphasize that this 'side by side' often has the effect that the distinction between exploratory and confirmatory gets blurred. This is perfectly allright. Many techniques have both exploratory and confirmatory aspects, although it is perhaps true that in general exploratory techniques use graph paper and confirmatory techniques use tests of hypotheses, confidence

regions, and the like. Most techniques, however, can be used both to conjecture and to refute, and perhaps all good techniques must have both aspects. Because many techniques in data analysis are not based on any explicit probabilistic assumptions it is also clear that at least in these cases refutation must be possible without using the techniques of inferential statistics (and in fact where would the natural sciences be if this were not true). The words 'conjectures' and 'refutations', of course, are taken from the work of Karl Popper. "The way in which knowledge progresses, and especially our scientific knowledge, is by unjustified (and unjustifiable) anticipations, by guesses, by tentative solutions of our problems, by conjectures. These conjectures are controlled by criticism; that is, by attempted refutations, which include severely critical tests. They may survive these tests; but they can never be positively justified: they can neither be established as certainly true nor even as 'probable' (in the sense of the probability calculus)." (Popper, 1963, p vii).

There are two other points of view with which we do not agree. The first one is explained most clearly by Morlat, in his preface to Caillez and Pages (1976). According to him data analysis is the modern form of 'descriptive statistics', and much more powerful forms of descriptions are now possible because of the computer. Tukey emphatically disagrees. "Some have suggested that 'exploratory data analysis' is just 'descriptive statistics' brought somewhat up to date. Much effort, much intelligence and understanding has been devoted in recent years to convince us that 'the map is not the region'. Perhaps an equal effort, at least among statisticians, is needed to persuade us of the equally true statement 'the usual bundle of techniques is not a field of intellectual activity' ! " (Tukey, 1980, p 24). On the other hand Morlat's interpretation of data analysis is understandable, because of the claims of Benzécri and his school, which constitute the other point of view with which we disagree. Morlat summarizes it as follows: "And the description proves to be so effectual that some do not hesitate to conclude that the essential parts of statistics, or even all of statistics, must belong to data analysis. Classical mathematical statistics, according to them, consists only of a number of somewhat arbitrary mind-games, whose only purpose is to furnish the rooms of statisticians who, at the time, did not have computers to solve more realistic problems." (Morlat, preface of Caillez and Pages, 1976, p iii). Both points of view are unnecessary extremist and polemic, and they are a serious threat to the arrival of the 'Whole Statistician', desired by Box (1979), by Tukey (1980), and also by us.

1.4.5 Inference

We have used the word 'inferential' a number of times, and is consequently desirable to specify what we mean by it. If we use the word we assume that the data are a random sample (in some sense of the word) from a (finite or infinite) population. We have certain knowledge about the sample, and we want to use this to obtain uncertain knowledge about the population. The problem is, of course, what we mean by uncertain knowledge and how we must obtain it. There has been and still is a lot of discussion on these 'foundations of statistics', much of it very polemical and much of it not very enlightening. The key issues are philosophical, and have been discussed (as 'the problem of induction') in philosophy for a very long time.

We do not want to become involved in this debate. In general we agree that scientists should behave 'rationally' or 'coherently'. It would be highly irresponsible not to agree. On the other hand we also think that 'irrational' and 'incoherent' behaviour has advanced scientific knowledge on many occasions in the past. We also take the point of view that current definitions and systems of rational statistical behaviour are 'persuasive' in the technical sense of Stevenson (1938) and Black (1949). And we remain unconvinced. Both decision theory and subjective Baysianism are very beautiful structures and heroic attempts to codify scientific behaviour in the face of uncertainty. We think that their practical relevance as prescriptive systems is limited, although they have had and will continue to have considerable influence on the practice of data analysis. Something like Carnap's principle of tolerance is clearly called for in the debate: it is not our business to set up prohibitio but to arrive at conventions. We obviously do not agree with statements such as: "In statistics, the rules are those of probability. A data analyst, not obeying these rules, will be incoherent and often do stupid things." (Lindley, in the discussion of Parzen, 1979, p 127). In most practical situations that we are aware of a decision theorist or subjective Baysian, who insists on being completely coherent, will be forced to do nothing at all. And even doing nothing at all will not necessarily be coherent according to his own criteria. Of course we do not intend to build a system which requires that people systematical do stupid things. On the other hand we find it difficult to agree with systems that try to define stupidity in a scientific context, and that, in the higher level of analysis in which choice of system is included in the process, must then be classified as stupid according to its own criteria.

1.5 Data analytic principles of this book

1.5.1 Model and technique

The procedure usually adopted in statistics is the following one. We start with a question or problem, then build a probabilistic model which we assume to have generated the data. The question is then imbedded in the model as an additional specification of the model, and the assumptions made so far also provide us with an 'optimal' technique for answering the question. Thus, given a question and an optimality criterion, the model automatically provides us with a technique. This approach tends to suggest the 'linear' model of science: question → data → → model → technique → answer. We have seen that people like Tukey, Box, Barnard, and Mallows have not been particulary happy with this linear model, and have suggested to build in a feedback loop, returning from 'answer' to 'model', i.e. which modifies the model by considering the results, or even to build in a feedback loop from 'answer' to 'data', which modifies the data by considering the results, for example by deleting offending observations. Classical statistics has very little to say about this feedback, in fact it usually advises us to start the whole process from scratch, with possibly the same question, but certainly with new data. We have also seen an important variation of this linear process, which is much less common. Given the question and the optimality criterion the technique can also be used to provide us with an optimal model. Superficially the two variations have much in common, they both try to establish a one-one correspondence between techniques and models. In practice, however, starting with the technique invites us to consider many different optimality criteria, and indeed there are many techniques which are not optimal for any model. Although the alternative approach seems equally linear, and a simple 'dual' method for approaching the same one-one correspondence, it tends to break the chain and to introduce feedback, although perhaps not explicitly.

Robust statistics can be considered as a very interesting attempt to relax the theoretical one-one correspondence between models and methods. The performance of a technique, for example computing the mean or median, is studied for a class of models. And properties of a class of models that one is interested in are used to derive a technique which has good performance for all models in the class. This two-way approach, and the study of classes of techniques and models, are important liberalizations, but the current formalizations of robust statistics tend to proceed along classical lines. The model is replaced by a supermodel, which is often a parametricized family of simple models, and the optimality criteria change accordingly. In stead of looking for a procedure which is optimal at a simple model, we look for a procedure which minimizes the worst possible performance at any simple model in the supermodel. This may be more

realistic, but it certainly is not essentially different from the classical linear approach. Consequently much practical work in robust statistics is in the spirit of data analysis, while much of the theoretical work is in the spirit of mathematical statistics.

In this book we adopt the point of view that, given some of the most common MVA questions, it is possible to start either from the model or from the technique: As we have seen in 1.1 classical multivariate statistical analysis starts from the model, generally using the multivariate normal distribution. Categorical or discrete MVA starts from a multinomial or Poisson model, which may often be more realistic, but then proceeds along the same lines. A convential optimality criterion such as asymptotic variance or asymptotic covering probability is chosen in both cases, and optimal procedures are derived or at least approximated. In many cases, however, the choice of the model is not at all obvious, choice of a conventional model is impossible, and computing optimum procedures is not feasible. In order to do something reasonable in these cases we start from the other end, with a class of techniques designed to answer the MVA questions, and postpone choice of model and of optimality criterion.

We do not think that this is the only appropriate starting point in MVA, or even that it is the best starting point in MVA, but we do think that it is the most useful starting point in exploratory multivariate situations if we want to say something about the relationship between models and techniques. Thus the technique is taken as a priori given in our theoretical work, much the same as the model in multivariate statistical analysis, and we study its properties by applying it to a number of interesting models of various kinds. The results then validate or invalidate the technique in the situation under consideration, and the results can be used in a feedback process which modifies the technique. This explains at the same time the choice of techniques in this book: they have been validated in this sense in a number of interesting situation We now discuss some of our major tools in this validating process more in detail.

1.5.2 Gauging

What do we mean by gauging of a technique ? We construct a model, with known properties, apply the technique to the model, and see if and how the technique represents or recovers the known properties. There are many types of gauges, and we mention some of the important ones.

a: Probabilistic gauges. The technique can be applied to a parametric family of probability distributions. It is interesting to see how the parameters are represented by the technique. The major probabilistic gauge in MVA is, of

course, the multinormal distribution, but other interesting examples are the Rasch model for binary multivariate data and the Markov chain for time series.

b: Statistical gauges. In stead of studying the population, as in (a), we now study the sample. It is interesting to compute theoretically what aspects of the model are 'estimated', and how well. Statistical gauges can also be compared with the corresponding probabilistic gauges to assess the effect of sampling.

c: Monte Carlo gauges. As in (b), but now we do not derive formulas, we only do computation. There are many examples in psychometrics, of which the most familiar one is perhaps computation of the null-distribution of Kruskal's stress.

d: Algebraic gauges. The data can also be generated by an algebraic model without probabilistic structure. Again we investigate what aspects of the model are represented and how well. In numerical analysis the Hilbert matrix is a familiar algebraic gauge, in psychometrics there are many scaling models formulated purely in algebraic terms such as the Guttman scale, the conjunctive and disjunctive models of Coombs (1964). Sometimes, as in the study of the Spearman-model in factor analysis, it is convenient to separate the algebraic and the probabilistic aspects when constructing gauges.

e: Empirical gauges. If we have data with well-established properties, for example measurements on a physical process which is theoretically well understood, then we can use the technique to find out if it gives results which are in accordance with theory. Interesting examples are in Wilson (1926), Wilson and Worcester (1939), and in Stigler (1978).

We emphasize that the use of statistical gauges can sometimes lead to the result that there is a model for which the technique is in some sense optimal. Consequently it is possible to study at least some of the questions that interest statisticians in this gauging-framework. We also agree with data analysts such as Benzécri that algebraic gauges are sometimes at least as important as probabilistic ones, especially in exploratory MVA. This is also in the tradition of psychometric scaling theory.

1.5.3 Stability

Even more important than gauging is the analysis of the stability of a technique. In general stability means that a small and unimportant change in data, model, or technique should lead to a small and unimportant change in the results. Both 'small' and 'unimportant' can be defined in various ways, and consequently there are many types of stability, we list the most important ones.

a: Replication stability. If we replicate the experiment under the same condition and apply the technique to the new data, then the results should not change dramatically. This requirement is, of course, fundamental for all scientific investigation, and it depends both on the properties of the technique and on the quality of the experiment (we use 'experiment' in the widest possible sense). In the social sciences independent replications are often impossible. Consequently replication stability is often not investigated directly, but the stability question is imbedded in a statistical model. If the data are a simple random sample, then the model will tell you what will happen 'on the average' if we replicate the experiment a large number of times. This is one of the reasons why statistics is needed in the social sciences. "You asked me to speak of the statistical methods of treating data. I wish you had not. It is a mean subject. Those of you who have read the biography of the great Lord Rayleigh by his son will recall his statement that he does not believe in statistical methods, that the object of repeating an experiment is to judge of the control acquired, that he even doubts the utility of averaging values to obtain a mean, though he admits that this is carrying disbelief rather far. We find very little statistical analysis in experimental physics or chemistry to-day, a smaller relative amount, I think, than was found a generation ago; and even in astronomy, for which the method of least squares was developed by Gauss and in which it was universally applied in the past, there is a strong tendency to short-cut formal statistical processes It is now to the biologist and the economist that you must go for complicated statistical analysis. Why this state of affairs ? May it perhaps lie in a contrast of the experimental and observational methods, in a difference of degree of attainable control ? Shall we say that when the control is good, when we are working in a field in which control is easy or when we are sufficiently astute or fortunate to design experiments so that those consequen in which we are interested are independent of the other variations, then we have no need of statistics and can go along with Lord Rayleigh ? Shall we admi that statistics belongs rather in the field of observation and serves to replace control when that is not attainable or is repugnant to the nature of the investigator ?" (Wilson, 1928, p 52).

b: Statistical stability. As we have seen under (a) this can be interpreted as an abstract, formalized form of replication stability, in which the model takes over the burden of replication from the investigator. A great deal of classical statistical theory can be translated as a study of the stability of data analysis techniques, the concept of standard error is a good example. If a statistician uses 'optimal' it can often be interpreted as 'optimally stable' (over independent replications). Baysian statistics, of course, is

different because it does not want to use the framework of repetitions. It is probably possible to develop and study Baysian analogues of stability, because it is also possible to talk about Baysian versions of the related concept of robustness, but we are quite happy to leave this job to somebody else.

c: Stability under data selection. This covers a multitude of procedures, all of them very objectionable from a Baysian point of view. The first one is post-experimental randomization, for example to derive permutation distributions of statistics. Another form of data selection stability is the jackknife (a good review is Miller, 1974), the bootstrap (Efron, 1979), or subsampling (Hartigan, 1969). All these approaches construct random mechanisms to perturb the given data in a probabilistic sense (the classical jackknife uses deterministic perturbations, but it is easy to construct probabilistic versions). Because the random mechanisms are all introduced conditional on the data it is not necessary, at least for some questions, to assume any probabilistic model for the data themselves. This is an extremely useful feature in many situations. The concept of the influence curve (Hampel, 1974) in robust statistics is also directly related to this form of stability. Rejection of outliers is also an interesting form of data selection which occurs very frequently. It may be true that these data selection techniques cannot be fitted into any one of the formal approaches to statistical inference. From our point of view this does not cause any inconvenience. "A sort of question that is inevitable is: 'Someone taught my student exploratory, and now (boo, hoo!) they want me to tell them how to assess significance or confidence for all these unusual functions of the data. (Oh, what can we do ?)' To this there is an easy answer: TEACH them the JACKKNIFE." (Tukey, 1980, p 25). In fact we think that the answer is a bit too easy, because there are more forms of stability than data selection stability, and there are more techniques similar to the jackknife, but we agree with the spirit of the answer.

d: Stability under model selection. A small change in the model that we fit must result in a small change in the estimates of the free parameters, and consequently also in a small change in the interpretation of the results. The study of predictor selection techniques, of multicollinearity, and of specification errors, come under this heading. Much of the work in sociology on fitting 'causal models', in psychometric genetics on fitting 'genotype-environment models', in criminology on fitting 'biosocial models', should pay more attention to this form of stability than they usually do.

e: Numerical stability. An underrated, but very important form of stability in data analysis. It studies the influence of rounding errors and of computation

with limited precision on the results given by the techniques. Study of numerical stability has profoundly influenced the field of linear least squares regression, but more complicated techniques in MVA and scaling are much more careless in this respect.

f: Analytical stability. If the possible data structures and the possible representations have enough mathematical structure, then the idea that 'a small change in the input should lead to a small change in the output' can be made precise in terms of continuity or differentiability. Statistical large sample theory, for instance, concentrates on consistency (a continuity condition) and asymptotic normality (a differentiability condition). The main difference between analysis and statistics, is that in analysis we often derive results in the form of inequalties and bounds, in statistics we derive similar results in terms of expected values. Thus analysis is often unduly pessimistic according to statistical criteria.

g: Algebraic stability. In techniques based on the procedures of linear algebra it is often feasible to derive perturbation results by algebraic means. The effect of deleting a predictor or a variable in components analysis, for example, can often be bounded by an inequality. This is closely related to (f), but often the algebra of the problem gives us simpler, more general, and more precise results.

h: Stability under selection of technique. If we apply a number of techniques, which roughly try to answer the same question, to the same data, then the results should give us roughly the same information. As the use of 'roughly' indicates this form of stability is somewhat complicated to study. But if nine out of ten techniques point to the same important characteristic of a data set, then technique number ten is disqualified if it does not show this characteristic.

It is clear that the study of stability can be made to include large parts of statistics, only changing the emphasis somewhat. An implication of our approach to data analysis is that no data analytic technique is complete without some gauging and without an investigation of its stabilities. Consequently some of these results are discussed in this book too. In the past data analytic or psychometric techniques were often considered to be justified because they 'performed well in practice', because they provided 'insight', or because the users were 'satisfied'. It is, of course, always nice if the customers do not complain, but from our point of view it is certainly not sufficient. Good advertising can sell bad products. We agree with Harris from section 1.1.9 that we need quality control.

1.6 Specific problems of MVA

1.6.1 The multinormal distribution

We have seen in the discussion of the MVA books that mathematical statisticans often assume from the start that multivariate data are multinormally distributed. This seems to result in a considerable loss of generality and applicability of the techniques. Why then is this assumption so common ? In 1.1.3 we briefly mentioned the reasons given by Anderson for using this assumption. We now discuss them in more detail.

a: Usually a good description in practice. Is it really ? Anderson's example are the classical Galton data on the distribution of length of fathers and sons, and in fact more examples of this sort can be found in antropometry. But we must not forget that Pearson discovered in studying the equally classical shrimp-data of Weldon that the normal distribution is certainly not universally valid in biometry, and that as a consequence of this discovery he constructed his famous system of skew frequency curves. Pearson is more careful than Anderson in this respect. "On the basis of a very large experience of frequency curves and surfaces we have no hesitation in saying that up to the present time no distribution has been proposed which roundly represents experience so effectively as the Gaussian frequency. One of the present writers has indicated over and over again how it fails, and he has measured the significance of its failure, but has always recognized that he must put against this the large percentage of cases in which it gives reasonable results, close enough for all practical purposes." (Pearson and Heron, 1913, p 189). Of course we must remember that Pearson's practical purposes were descriptive and not inferential, and that subsequent research on robustness has been more pessimistic. Moreover considerable efforts of Pearson's school to construct a system of bivariate or multivariate frequency surfaces comparable to the univariate system have not been successfull, which means that there are no systematic alternatives to multivariate normality. A second argument against Anderson's multinormal optimism is that 'goodness-of-fit' tests for the multinormal distribution are still fairly primitive, although it is true that they get a lot of attention in recent times. In any case normality of the marginals is not sufficient to conclude that the data are multivariate normal, and the assumption does not make sense if the variables are categorical or ordinal. □

b: The central limit theorem. The assumption that the data can be interpreted as resulting from summation of a large number of independent effects is not very natural in many situations. It may be sensible to assume something

like this in biometrical genetics or in astronomical error theory, but it seems very far-fetched in sociology or economics. We also know that convergence to the normal distribution can be very slow if the components in the summation are skew, we know that different normalizations can lead to different limiting distributions, and we know that in some cases it is more natural to think in terms of the product or the maximum of a large number of independent effects.

c: Simple formulas and many theoretical results. This seems to be the most important argument. The fact that there are many theoretical results is not only the consequence, however, of the simplicity of the formulas, but also of the large amount of interest and energy that has been invested in the multinormal because of reasons (a) and (b). Simplicity is only a relative matter, computing tetrachoric correlation for example is not simple, computing the exact distribution of the eigenvalues of a Wishart matrix is not simple either. But because the multinormal distribution is so popular, much of the complicated computing has already been done. This includes Monte Carlo work, which could in principle of course have been applied equally well in non-normal situations.

It is undoubtedly true, however, that the multinormal distribution has a number of very attractive theoretical properties, which simplify the job of the statistician considerably. We mention the ones that are most important for our purposes.

a: If a vector of random variables is multinormal, then any linear transformation of this vector is also multinormal. Because MVA often uses linear transformations of random vectors this property is extremely important.

b: If the joint distribution of two vectors of random variables is multinormal, then the conditional distribution of the first vector given the values of the second vector is again multinormal, with dispersion matrix which is independent of the value of the second vector and with a mean vector which is a linear function of the value of the second vector. The linearity of the means is called linear regression, the constancy of the dispersions is called homoscedasticity.

c: For samples from a multivariate normal distribution the sample mean and the sample dispersion matrix are independent. The sample mean and sample dispersion are also complete sufficient statistics for the multinormal parameters, and they are their maximum likelihood estimates. To put it differently: in most other distributions moments and product-moments are complicated functions of the parameters of the distribution, which implies that it is difficult to compute 'optimal' estimators, and that it is also difficult to interpret the parameters. The multinormal distribution is

exceedingly simple in this respect, and the first order moments and second order product moments contain all information in the sample.

d: Multinormally distributed variates are independent if and only if they are uncorrelated, which is true if and only if the covariance matrix is diagonal. In general zero correlation is only a necessary condition for independence, for the multinormal distribution we find again that interdependence of the variates can be described completely in terms of the second order product moments.

e: Properties of the multinormal distribution are closely connected with properties of Euclidean geometry. Points with equal probability density are on ellipsoids with the vector of mean values as the center, which implies that probability density and weighted Euclidean distance can easily be translated into each other. If, for example, two multinormal distributions have equal dispersions, then the points where the first density is larger than the second one are separated from the other points by a hyperplane.

The properties (a)-(e) are not only statistically interesting, it is clear that simple properties like these are also important from a data analytical point of view. This makes the multinormal distribution both very important and very interesting as a gauge. But in many situations the assumption of multivariate normality is not very natural, and difficult to test rigorously. This depends to a large extent on the properties of some commonly used statistical procedures. In MVA we generally test a parametric hypothesis within a more general hypothesis, and the more general hypothesis almost always contains the assumption of multinormality. The usual procedures do not test the more general hypothesis or model, they assume this to be true, and test the additional specification on the parameters. This can be quite dangerous, of course. In summary we must agree with the verdict: "Theorists of multivariate analysis clearly need to venture away from multivariate normal models." (Dempster, 1971, p 317).

In univariate statistics the normal distribution, which once was all-powerful, has already been abandoned to a much larger extent. It has been shown for several procedures (such as the t-test) that they are moderately robust, in the sense that their properties remain approximately valid under moderate deviations from normality. And a large number of non-parametric statistical procedures have been developed, which do not assume normality, or indeed any parametric model. The two have also been combined recently. In MVA both approaches have not been used systematically. Little is known about robustness of MVA procedures, and the things that are known are not very encouraging. Sample covariances, for example, are very sensitive to departures from normality in the heavy tails direction. And as we have seen, for example in our discussion

of Kendall's book, non-parametric MVA has not been developed sufficiently. There are a number of versions of some of the more simple multivariate tests, relying quite heavily on univariateness of dependent variables (compare, for example, Puri and Sen, 1971). The properties of these procedures are less satisfactory than those of univariate non-parametric statistics, and (above all) their data analytical value is limited , because the quantities that are computed do not have straightforward geometrical interpretations.

1.6.2 Tabellary analysis

In sociology, polticology, and related sciences surveys (also called observational studies) are very important. A large number of people have to respond to a large number of questions, the data are the answers to the questions, sometimes combined with background information about the respondents. Variations on this theme are questionnaires for clinical diagnostic purposes, attitude studies, panel studies, multiple choice tests, and so on. Continuously varying numerical variables are rare in investigations like these. If the variables are numerical they are usually categorized in fairly broad categories, other variables are ordinal, and background information such as religion or party affiliation can easily be nominal. Good surveys of the problems in the analysis and interpretation of observational studies are Hirshi and Selvin (1973), Cochran (1972), and McKinley (1975). We concentrate on one particular aspect.

It was conceded very soon that assuming multinormality for complex sample surveys made no sense, in stead of using Pearson's measures of association people preferred those of Yule (compare MacKenzie, 1978). According to Pearson everything in this world varied continuously on a scale, discrete variables are always discreticized continuous variables, and measures of association have the purpose tb estimate the underlying correlation between the continuous variables. Pearson made these assumptions because he was convinced that a unified conception of science was possible starting from the concept of correlation, in stead of causality. The unified conception implied the idea that biology and antropology had the same sorts of lawlike relationships as physics, i.e. functional relationships between measurable variables. The theory is outlined in the various editions of Pearson's classical "The Grammar of Science" (Pearson, 1892, 1900, 1910). In the last analysis this is, of course, a metaphysical point of view, which also explains why it was very difficult for Pearson to accept the essentially 'discrete' doctrine of inheritance of Mendel (Norton, 1975, 1978). Pearson's metaphysics has had a considerable influence in psychometrics, because many people thought that only continuous variation was 'measurable', and that only the assumption of underlying continuous variation could effect the ascent of psychology to the

level of the real sciences such as physics. Tetrachorical correlation, for example, is used almost exclusively in psychometrics, and there are still articles published which tell us how to compute this coefficient faster or better, without paying any attention to the fact that the assumption of underlying bivariate normal variation is extremely contrived for most binary data.

Yule did not subscribe to Pearson's metaphysical ideas. For him something was a measure of association if it was +1 with perfect positive relationship, -1 with perfect negative relationship, and zero in case of independence. On the basis of this axiomatic point of view he proposed a number of measures which satisfied these assumptions. They are discussed, with many modifications and an avalanche of interesting details, in the famous papers of Goodman and Kruskal (1954, 1959, 1963, 1972), now reprinted in the book Goodman and Kruskal (1979). Sociologists have never been bothered to the same extent as psychologists by the idea that their discipline should be constructed after the model and with the methods of the exact sciences. They also never had the uncomprimising empiricism and the corresponding correlationmania of the Pearson-Spearman school. The technique that became popular in sociological data analysis was making contingency tables with corresponding measures of association. This worked satisfactory at first, for obvious reasons. "Many tried and tested techniques of multivariate data analysis were invented at a time when ten was a typical number of variables in an ambitious data collection program." (Dempster, 1971, p 336). With ten variables we have 45 contingency tables, which is still a manageable amount. But in contemporary surveys 100 variables are quite common, the MMPI has approximately 700 questions, and in longitudinal studies even more variables can occur. The computer came to the aid of the sociologist, packages such as CROSSTABS were constructed, and all cross tables were printed, each table with a long list of association measures. This evidently produces enormous amounts of output, and it actually still happens that you find these mountains of paper on peoples desks, gathering dust and looking desparate. "That one still sees these long lists of little tables coming out of the printers, is because too many scientists, especially in the social sciences, have not adapted their methods to the power of these new computing tools. They are like an engineer who builds a bridge by designing blocks of concrete in the form of bricks." (Benzécri and others, 1973, p 11).

In the first place 5000 cross tables cannot be presented in a research report or paper, and consequently one must select. Usually one selects what seems interesting, which is a rather subjective criterion. It is perfectly possible that others, with different (and perhaps opposite) interests, will find other relationships in the material. In the second place a very long list of cross

tables at the very least suggest the question how these tables are related to each other. They are usually presented as independent findings, but it is clear that they certainly are not independent. It seems that reporting selected cross tables gives the impression that the relationships are stronger than they actually are, and gives moreover a very unsystematic presentation of these relationships (compare Hirshi and Selvin, 1973). We can compare it with the following procedure. Suppose we have a large number of numerical variables, and compute their large matrix of intercorrelations. The methods of tabellary analysis now suggest, that we discuss each of the individual correlations in the table that is interesting from our point of view. It seems more natural to us to look for techniques that give a compact description of the correlation matrix, and that do not assume that the observed correlations are independent statistics, which can consequently be described and discussed independently. Another habit is to discuss only significant relationships. Of course we should take into account here that with 5000 tables there will be 250 tables with 5%-significant relationships on the basis of chance alone.

The sociologists have found two different ways out of the ruins of tabellary analysis. The first one is the analysis of multidimensional contingency tables, usually by using the so-called log-linear models. Books describing these techniques are Haberman (1974), Bishop and others (1975), and Gokhale and Kullback (1978). This is the discrete MVA we have encountered earlier in our discussion of the books of Roy and Kendall. A second way out is the so-called 'causal analysis', extensively discussed in Blalock (1964) and Boudon (1967). We discuss these two recent developments in separate sections.

1.6.3 Discrete MVA

One of the main disadvantages of tabellary analysis is that is not at all clear how the various tables are related to each other. This makes it possible to find various relationships of the well-known 'spurious' sort, such as the relation between the number of imported bananas and the number of illegal births. The solution in discrete MVA is the analysis of multidimensional contingency tables, which means that we consider three or more variables at the same time, and analyse the corresponding multidimensional array of frequencies. We first mention some of the more important disadvantages of this approach.

If we have a large number of variables, then we can make an extremely large number of multidimensional tables. The problem of selection of tables thus becomes more serious. With ten variables there are 45 two-dimensional tables, 120 three-dimensional ones, 210 four-dimensional ones, and so on. There is of course only one ten-dimensional table, and it is possible in principle

to analyse only this table, which contains all information in the data. But then, unfortunately, we encounter the second disadvantage of discrete MVA. If every variable has four categories, then the ten-dimensional table has $4^{10} = 1.048.576$ cells. The number of observations will in general be much smaller than that, and consequently most of the cells will be empty. Because the inferential aspects of discrete MVA are based on the asymptotic normality of the frequencies it is necessary that the table is reasonably well-filled. According to the classical, although somewhat arbitrary, prescription of Cochran we want on the average about five observations in each cell, which means that even in this small example we need more than five million observations. If there are more variables then analysis of the complete table is not possible at all, and selection of subtables can be done in an enormous number of different ways. In this sense discrete MVA is useful if we want a fairly exhaustive analysis of the relationships between three or four variables, either because there are only three or four variables, or because there are reasons to find three or four variables extremely important. Discrete MVA can not be used for the simultaneous analysis of a large number of categorical variables.

We have seen that Roy (1957) was the first author who discussed discrete MVA in a handbook, and that he thought that this approach was more realistic than the usual multinormal one. What are the most important differences ? In MVA we are generally interested in dependence and interdependence of variables. Dependence and interdependence are properties of the probability distribution of the variables, which can be defined in various ways. For the multinormal distribution there is not much choice, all relationships can be defined in terms of the covariance matrices and derived marginal and conditional covariance matrices. For more general distributions more general definitions are needed. Roy uses, following Fisher and Bartlett, definitions in terms of conditional probabilities and in terms of the product rule for combining independent events. This is the so-called multiplicative approach, also used in the more recent work on log-linear analysis. There is also an additive aproach, used mainly by Lancaster and his school. The relationships between the two different systems is easy to describe: multiplicative analysis is the same thing as additive analysis on the logarithms of the probability measures. In general the usual probability-based definitions of independence, interdependence, and interaction can more easily be investigated in the multiplicative approach. Additive analysis has other advantages, the two techniques are compared in Darroch (1974), Lancaster (1971, 1975). In general the discreteness of the variables in discrete MVA is not essential for definitions of the interactions, it is only essential in the subsequent statistical analysis. It is possible

to define a completely general system of nonlinear multivariate analysis, valid for both discrete and continuous variables, of which discrete MVA and multinormal MVA are just special cases. In this general system we start by defining complete sets of orthogonal functions on the marginals, and we decompose the probability distribution (or its logarithm) in terms of the tensor product of the functions from the various sets. In discrete MVA the complete sets on the marginals are finite, which makes it possible to handle the analysis in practice, in multinormal MVA only linear functions contribute to the interdependence, and the complete sets consist of a single function for each variable. In the general continuous case we need bases which consist of an infinite number of functions on each variable, and consequently this case is only interesting theoretically, not in practice. In several other places in this book we shall discuss the relationships of general nonlinear MVA with the techniques we prefer.

1.6.4 Causal analysis

Causal analysis has a different historical origin than tabellary analysis. In biometrical genetics (which used to be almost identical with what we now call statistics) the dominant philosophy of science was the descriptive and empiristic system of Pearson's 'Grammar'. One of the fundaments was Pearson's doctrine that correlation is more fundamental than causation, because causality is merely the (theoretical) limit of perfect correlation. It is not necessary to look for causal relationships, you only have to compute correlation coefficients. The theory then comes automatically, because a scientific theory is merely a short and simple summary of a large number of empirical observations (for example correlations). This interpretation of science is not very popular these days, except perhaps in some isolated psychometric and biometric centres. There are at least three reasons for this. In the first place it does not work if you actually try it, as psychometric theories about intelligence or heredity or criminality have clearly shown. The number of correlations that has been computed since Pearson must run in the zillions, but no theory has as yet come out of this mountain of numbers. In the second place Yule clearly showed that correlations have their limitations. If we correlate time series, for example, we often find nonsense-correlation. And there are many, many examples that show that correlation does not imply causality (such as the income of presbyterian ministers and the import of rum from Jamaica). The third reason is more philosophical. Causality is asymmetric, implies a direction, and a temporal order. Correlation is symmetric. Nonsense correlations made some people believe that correlations can only be interpreted within an assumed causal model. Sociologists in particular, who never cared in the first

place for Pearson's empiricism, find this an attractive point of view. It is, of course, only logical that a field in which there is ten times as much theory as empirical data reacts differently to correlation coefficients than a field in which there is ten times as much data as theory.

Causal analysis was defined originally (by the geneticist Sewall Wright) for continuous multivariate data. The postulated interrelations between the variables were picture in an arrow-diagram, the arrows were interpreted as linear relations between the variables, and clearly the arrows indicated a direction in which causality operated. It is obvious that even for a small number of variables it is already possible to draw a very large number of different arrow-diagrams. This is the major problem of causal analysis. In stead of the problem of table selection we now have the problem of model selection. And again this problem is more serious if the number of variables is large. We also must not forget that the model more or less automatically implies its causal interpretation (all arrows have a direction, some possible arrows are not there), and that the estimated correlation and regression coefficients are always interpreted within the postulated model. If we had chosen another model, then the same statistics would have been interpreted differently. The problem of interpretation, which corresponds to relating the different tables in tabellary analysis, has been shifted to the a priori level, but has not been solved by this clever move. It is true that, with some additional assumptions, we can also test the goodness-of-fit of the model, but these additional assumptions are often not very realistic, and the power of these tests for complicated models with many variables is usually very low. In biometrical genetics the theory of Mendel imposes many restrictions on the choice of model, at least in relatively simple situations under direct experimental control. In studying the inheritance of intelligence, for example, genetical theory does not tell us anything useful, and consequently the choice of model is quite arbitrary, with the unpleasant effect that the same data can lead to very different interpretations.

In modern versions of causal models, inspired by biometrical genetics, by psychometry, and by econometrics, we even use 'latent' or 'unmeasurable' variables to extend the model. Genotype, for example, is almost always unmeasured, as is general intelligence. The 'latent' variables are only defined in terms of the relationships they have with each other, and with the measured variables, postulating latent variables only has consequences through the structure they impose on the covariances of the observed variables. This clearly makes the problem of model choice even more complicated than it already is, and consequently makes interpretation even

trickier. It is now even true, that if we choose another name for the unobserved variables then the same statistics are interpreted differently. The goodness-of-fit test now becomes even more overburdened, and is not of much help in the selection of an appropriate model. It is always possible that 'better' models exist, more so because most investigators only look at relatively minor variations within a model with fixed global structure. Interpretability of the results in this context (as in others) is a poor criterion of success, because the interpretation has already been largely determined at the moment of model choice.

Nevertheless we think that causal analysis is a useful attempt to incorporate prior information about the variables (for example their natural order in time) in MVA. As such it is an interesting generalization of the simple distinction between analysis of dependence and interdependence. We think that it is appropriate to incorporate some rationalistic conjectures into the untenable empiricistic optimism of the psychometricians. But there are many situations in the social sciences in which the choice of a causal model is very arbitrary, because the necessary a priori knowledge is either absent or extremely fragmentary. And this is especially the case if there is a large number of variables. In situations like these it can be misleading to interpret results in terms of the parameters of the causal model, because the choice of the model was arbitrary the same thing must be true of the interpretation. It does not make sense to use highly structured models in highly unstructured situations. The subjective choices of tabellary analysis are assumed away, and are not questioned any more. A program package such as LISREL is in many respects much more satisfactory than CROSSTABS, there is some indication of statistical stability of the results, and there is considerably less output and thus more data reduction. But the alleged statistical respectability of the approach (compar with some of the earlier alternatives) invites uncritical acceptance of the result by the uninitiated.

1.7 Definition of MVA

1.7.1 Asymmetric role of rows and columns

On the basis of the discussion in the preceding sections we can now try to give a definition of MVA. The oldest, and the most restrictive, definition was that MVA is the analysis of random samples from a multinormal distribution. The data are collected in an $n \times m$ matrix, the rows of the matrix are the n independent replications of the same multinormal m-vector. We have seen that both the assumption of independence between rows and the assumption of multinormality are too restrictive to construct a general theory of MVA.

The first, and most far-reaching, generalization (suggested by the work of Van de Geer, Caillez and Pages, and Green and Carroll) is that MVA is the analysis of an arbitrary rectangular matrix, with the explicit purpose of describing the matrix in terms of a smaller number of parameters and of making pictures of this representation. We do not make any assumptions on the origin of the matrix. If we define MVA like this, then multidimensional scaling and various forms of cluster analysis are also included in the definition. And it is true that the French followers of Benzécri include these techniques in their definition of data analysis, while the group around Krishnaiah and the Journal of Multivariate Analysis also pay attention to these 'non-statistical' forms of MVA. Nevertheless we think that this definition is somewhat too general. The name multivariate analysis implies that a number of objects are involved that we call variables or variates. In the matrix-definition of MVA the rows and columns of the matrix are treated symmetrically, variables are not mentioned. It may be better to use the term multidimensional analysis for this field.

We consequently need to preserve more elements from the classical definition, notably the asymmetric treatment of rows and columns. We also want to drop the restriction that the data are real numbers, because this is not necessarily true in discrete MVA. As the starting point we do not use an arbitrary $n \times m$ matrix, but m random variables defined on a common probability space, not necessarily real-valued. In the simplest case the space on which the variables are defined has n elements, and the probability is defined by counting the number of elements in the subset and dividing by n. The m variables can be defined by making a list of all n elements, with for each element the corresponding m values of the functions. This list can be organized in an $n \times m$ matrix. The probabilistic component in this case is essentially irrelevant, and it is consequently still possible to analyze arbitrary $n \times m$ matrices, but the approach in terms of m functions on the same space has introduced asymmetry and has made it possible to think of generalizations. In statistical terminology we study the population in this case, and the population is finite and completely observed. We can also study a finite population with a more general discrete probability distribution over the n elements, but now the probability content is no longer trivial, and we have information that is not represented in the $n \times m$ data matrix.

It is also possible that the probability space on which the variables are defined has an infinite number of elements, and that we consequently can not define our functions by giving an explicit list of values. We are still studying the population, but the population is now infinite, and can not

be observed. In stead of listing the functions explicitly we now state its properties mathematically, for example by assuming that the joint distribution of the variables is multivariate normal. It is clear that this situation is of theoretical interest only, nothing is observed, there is no data matrix. We are not doing data analysis, we are not doing statistics. The situation is of interest in the process of gauging our techniques, or in the study of various forms of stability.

In the third situation we have a data matrix again, but we now assume that the rows of the matrix are a random sample of size n, or, the rows are independent realizations of the same population model. The situation is completely different from the previous two, in which the stochastic variables were defined completely, but not necessarily observed. We can also state this by saying that we now have observations on the probability measure defined on the probability space, the probability measure itself now defines the basic random variable. We have to work with the empirical probability measure, on which we have observations, and which estimates the theoretical one. Assumptions about the theoretical measure (i.e. about the population) have consequences for the possible empirical measures we can observe, in the same way as the nature of the sample has consequences for what we observe.

On the basis of this analysis we now give the following definition: MVA studies systems of correlated random variables or random samples from such systems. We do not specify in this definition that the number of random variables is finite. This will always be the case in this book, but we do not want to exclude the analysis of continuous-time stochastic processes from the definition although an infinite number of random variables will also only occur in theory. We have also incorporated the stochastic element explicitly in our definition, but we have seen that it can be made trivial in the case of a finite population with counting measure. Thus it causes no real loss of generality. Statistics only becomes important, of course, in the special case that we actually have random samples.

1.7.2 Linear, monotone, and nonlinear MVA

We now define some specific forms of MVA which will be important in this book. MVA is linear if the results are invariant under one-to-one linear transformatio of the random variables, it is monotone if the results are invariant under one-to-one monotone transformations, and it is nonlinear if they are invariant under all one-to-one nonlinear transformations of the random variables. These definitions are somewhat vague, because we do not specify what we mean by 'the results'. The results can be formulas, which are the result of derivations,

they can be idealized numbers, wich result from substituting values for the the variables in the formulas, and they can be actual computer output, influenced by rounding errors, choice of initial configuration, or stopping criteria. It is possible that a part of the output of the actual computer programs changes and another part does not change, or even that all the results change, but in a simple way. It will become clear in the rest of the book, for all the techniques we discuss, where this distinction is important and what remains invariant. Our definitions are also idealized, because they do not take inevitable shortcomings of computerprograms into account. It is possible that transformation of the variables results in slower convergence to the desired solution, or even to convergence to an undesirable local minimum. Nevertheless the distinction is a very useful one, and using it we can now state as one of the main purposes of this book to discuss monotone and nonlinear versions of some of the more familiar linear multivariate techniques.

Up to now we have discussed the classical $n \times m$ data matrix of MVA, but we have also seen that tabellary analysis and discrete MVA use cross tabulation or contingency tables. The relationship between the two representations will be explained more formally in chapter II, in this introduction we merely observe that the relationship is that between the values of a random variable and its distribution. In the discrete case, in which the m random variables map the space into m finite sets, every possible combination of values corresponds with a cell of the m-dimensional cross tabulation. Tne data can be represented by indicating how many times each of these possible profiles occurs in the data matrix. This is true in the case of a population, the cell values are then probabilities, but it is also true in the case of a sample, in which the cell values are observed frequencies. If the variables have values in a range with an infinite number of elements, then we replace the cross tabulation with the product of these m ranges, and we replace the probabilities in the cells by a multivariate probability distribution or density. Again we see that the three special cases to which our definition of MVA applies are all covered.

It is, of course, true that a sample always has a finite number of observations, and that we can assume as a consequence of this that the range of all variables in the sample is finite. Continuous variables, if they exist at all, are always measured with finite precision, which leads to a representation with a finite number of decimals. Thus infinite ranges only occur in theoretical analysis of continuous population models. One of the basic ideas in this book is that all data are discrete (or categorical), and that continuous models are used

for gauging and to simplify calculations and approximations in some cases. This is, in fact, the way in which the normal distribution was introduced into the history of probability theory. Only much later, after Galton and especially Pearson have made continuous variation the norm, we find the point of view that discrete variables are in some sense degenerate or rounded continuous variables. This last point of view is still very important in most of the classical books on MVA, although usually implicitly.

1.7.3 Bivariate and multivariate MVA

The m stochastic variables define an m-dimensional probability distribution. This distribution has univariate marginals, bivariate marginals, and so on. There are still people whose approach to MVA is essentially univariate, by which we mean that they apply techniques which gives the same results if we apply them to another multivariate distribution with the same univariate marginals. Most people agree that such an approach can be extremely misleading (compare Rao, 1960). We have seen that multinormal MVA is typically bivariate, the techniques give the same results if we apply them to another multivariate distribution with the same bivariate marginals. In the multinormal distribution this does not lead to loss of information, because multinormal distributions are completely determined by their bivariate marginals, but in other distributio we do ignore information. Tabellary analysis is also bivariate, but tabellary analysis is nonlinear while multinormal analysis is linear. Consequently multinormal analysis gives the same results on different multivariate distributions who merely have the same variances and covariances. Moreover tabellary analysis looks at all bivariate distributions separately, and multinormal analysis looks at them jointly. Thus if two multivariate distributio have the same bivariate marginals except for one, then tabellary analysis will give the same results for all other bivariate marginals, while multinormal analysis will give different results for the complete analysis.

The techniques in this book are largely joint bivariate, although we discuss some extensions in the multivariate direction. Thus they can be considered to be somewhere in between multinormal and general multivariate analysis, we combine non-linearity and bivariateness, hoping that the bivariate marginals give sufficient information on the interdependencies in the multivariate distribution. By concentrating on bivariate marginals only we circumvent the problem of the many empty cells and the problem of difficult interpretation of higher order interactions (also familiar from the analysis of variance). By allowing for non-linear transformations we drop many of the restrictions of multinormal analysis. By analyzing all bivariate distributions jointly and by

concentrating on low-structure models we circumvent the model selection problem. And by applying as much data reduction as possible we avoid the problem of having to cope with ten pounds of output from each analysis. The emphasis in this book is consequently on large data sets with many variables, on efficient computing, on nonlinear transformations, and, whereever possible, on geometrical representations and interpretations. The idea of a 'random sample' does not play a prominent role, and the multinormal distribution is nowhere used as a starting point. Nevertheless, for gauging purposes, we are very interested in the performance of our techniques if we apply them to multinormal populations and samples. And whereever possible we use statistical stability techniques, mainly asymptotic perturbation methods and versions of the jackknife. It is difficult to indicate the position of the book in figure 1, because we are in some senses close to tabellary analysis and discrete MVA, with are not represented in the figure.

1.8 Some important ingredients

1.8.1 Join and meet problems

The techniques discussed in this book can be classified in two different groups. In the first place there are various generalizations of principal components anlysis, and in the second place similar generalizations of canonical analysis. This distinction corresponds with the already familiar distinction between internal and external MVA, or between the analysis of interdependence and the analysis of dependence. In this section we discuss a more general distinction on a verbal level. The distinction will be explained more formally in chapter 10.

Techniques such as principal components analysis try to find a subspace of the space spanned by the variables, which has minimum dimensionality and yet contains all the variables. Canonical analysis tries to find a subspace of maximum dimensionality which is contained in all groups of variables. Principal components analysis approximates from the outside, and tries to find the 'least common multiple' of all the variables, canonical analysis approximates from the inside, and tries to find the 'greatest common divisor' of all groups of variables. Although our techniques are non-linear, in the sense that the results are (often) invariant under non-linear transformations of the variables, we use computational tools from linear analysis and algebra. This is because non-linear transformations of a variable themselves define a linear space, of which the linear transformations form a subspace. We consequently work in a larger space, but the space is still a linear space in the technical sense of the word. General non-bivariate MVA works in even

larger linear spaces.

We introduce some terminology to generalize the distinction between components analysis and canonical analysis. Assume for the moment that we are dealing with an ordinary n × m data matrix. The m variables are partitioned into K sets of variables. Each set of variables defines a subset of n-dimensional space. In the linear case this is the set of all linear combinations of the variables in the set, the dimensionality is not larger than the number of variables in the set. In the nonlinear case it can be the set of all nonlinear transformations of the variables in the set, whose dimensionality is not larger than the total number of different values assumed by the variables in the set (which is equal to the number of nonempty cells in the corresponding multidimensional cross tabulation). It can also be the space of all linear combinations of separate nonlinear transformations on each of the variables, in this case the dimensionality does not exceed the sum of the numbers of values assumed by each of the variables separately. These K subsets, once they are defined, can be combined in various ways. In the first place they have an intersection, which is the largest subspace contained in all K subspaces, and they have a linear sum, which is the space of all linear combinations of K vectors, one from each of the subspaces. The linear sum is the smallest subspace that contains each of the K subspaces we started with. We borrow some terminology from lattice theory, and call the intersection the meet of the K subspaces and the linear sum their join. It is important to observe that the join of a number of subspaces generated by linear combinations of nonlinear transformations of separate variables is the same as the join of the m subspaces defined by the nonlinear transformations of each variable. Thus in this case the partitioning of variables into subsets is irrelevant in computing the join. The same thing is true in the linear case.

We translate some familiar MVA problems into this terminology. In principal components analysis we try to find p orthogonal vectors in n-space, in such a way that all variables are linear combinations of these p components. This is possible if and only if the join of the variables has a dimension not exceeding p. Thus principal components analysis is a join-problem, and we can say that it tries to compute the smallest subspace containing all variables (which is equivalent to computing the dimensionality of the join, or the join-rank). In canonical analysis there usually are only two sets of variables, and consequently only two subspaces. We try to find p orthogonal vectors in n-space that belong to both subspaces. This is possible if and only if the meet of the two subspaces has a dimension of at least p. Thus canonical analysis is a meet-problem, in which we compute the largest subspace contained in both

sets of variables (which is equivalent to computing the dimensionality of the meet, or the meet-rank). The restriction of this purely algebraic formulation to two subspaces is in no way essential, and we can extend it directly to K subspaces. The restriction of meet and join problems to finite dimensional space is also not essential, and neither is it necessary on this abstract level to consider only a finite number of variables or subspaces.

Our formulation does not use any specific coordinate system in the space. For numerical purposes, however, it is necessary that coordinates are used, and that the problem is translated into matrix algebra. Moreover we cannot expect in general that perfect solutions exist, in general the join-rank of m variables will be m, and the meet-rank of K sets will be zero. We consequently must introduce loss functions which measure departure from perfect fit, join-loss measures departure from 'join-rank = p' and meet-loss measures departure from 'meet-rank = p'. The dimensionality p is chosen by the user, the theory in this section tells us that we want to choose p as small as possible in a join-problem and as large as possible in a meet-problem. The expressions 'as large as possible' and 'as small as possible' will not be defined here, because their interpretation does not only depend on the value of meet-loss or join-loss, but also on various other properties of the data and on other data analytic considerations. The definitions and properties of the loss function will be discussed in more mathematical detail in chapter 10.

The definitions of meet and join problems above assume that all variables are linear, or at least that the possible transformations of a variable define linear subspaces. This formulation is not quite general enough for some purposes, because ordinal variables, for example, cannot be fitted into this framework. We now give a more general discussion, starting with a more precise analysis of the concept of join-rank. The partitioning into subsets is irrelevant here, so we let each variable define a subspace. In the linear case this is the subspace of linear transformations, in the nonlinear case a subspace of nonlinear transformations, but we now extend the analysis and merely assume that for each variable we can choose from a set of possible transformations (for nominal and ordinal variables transformations are also called, perhaps more appropriately, quantifications), which is not necessarily a subspace. Each choice of transformations makes it possible to compute the correlation matrix of the transformed variables, different transformations lead to different correlation matrices. If the join-rank is equal to p, then the correlation matrix will have rank less than or equal to p, no matter how we choose the transformations. Because the correlation matrix is invariant under

linear transformation of each of the variables the join rank of linear variables is p if and only if their ordinary correlation matrix has rank p.

It is now possible to use two different practical approaches to join problems, and both approaches have been used by previous authors. If we have a class of nonlinear transformations at our disposal we can suppose that we actually are in the linear situation, and that for some unfortunate reason the precise values of the numerical variables are unknown. In this case we can still speak of the correlation matrix of the variables, which is also unfortunately unknown, except for the fact that it has rank p. We choose our transformation in such a way that the correlation matrix is as close as possible to a rank p matrix. This is called the single approach to the join problem, because we only find a single transformation for each variable and only a single correlation matrix. In the multiple approach we take the nonlinear situation as our starting point, and we look for a number of different transformations which all give a correlation matrix of rank p. The definition of join-rank tells us that the number of linear independent solutions for the transformations must be equal to the dimensionality of the space of possible quantifications. Thus the single approach computes a single solution (because it believes in a true hidden transformation), the multiple approach computes more than one solution.

The two approaches to the join problem are implemented in two different computer programs, which are both discussed in this book. We mention them here, because discussing their properties shows clearly that nice theoretical distinctions do not necessarily lead to equally nice distinctions in the implementation. HOMALS and PRINCALS are discussed in chapters 3 and 5 of this book. We must emphasize in the first place that interpreting these programs in terms of solving join problems is just one possible interpretation, and not necessarily the most illuminating one, other geometrical and algebraic interpretations of especially HOMALS are possible, and will be discussed in detail in chapters 3 and 4. HOMALS implements the approach to nonlinear component analysis also known as Guttman's principal components of scale analysis, Hayashi's third method of quantification, or Benzécri's correspondence analysis of multiple disjoint tables.If interpreted as a program for solving a join problem it takes the multiple approach, and aims at solutions with a join-rank of one. HOMALS accepts only nominal variables, not numerical or ordinal ones. PRINCALS generalizes the approach used in Kruskal and Shepard's nonlinear factor analysis, in PRINCIPALS of Takane, Young, and De Leeuw, and in other similar programs by Roskam or Guttman and Lingoes. PRINCALS in its simplest versions uses the single approach, and aims at solutions with a join rank equal to any given p. The variables can be either nominal, ordinal, or numerical. Thus, briefly,

HOMALS computes p solutions of join rank one, PRINCALS computes one solution of join rank p. If perfect fit is not possible we merely have to replace 'computes' by 'approximates' in this last sentence.

Things become more complicated because PRINCALS also has the possibility of mixing the multiple and single approaches. We interpret this as HOMALS with restrictions: we use the multiple approach aiming at join rank one, but for single variables we use the restriction that the transformations in the different solutions must be proportional. It is also possible to use PRINCALS with all variables single to compute a multiple solution for general p. This amounts to computing an ordinary single PRINCALS solution, and then to compute another one, taking care in some way or another that the second solution is actually different from the first one. All this will be explained in more detail in chapters 5 and 10.

The situation is more or less the same for the meet-problem. For each transformatio of the variables and for each linear combination of the variables within the K sets we can compute a correlation matrix of order K, i.e. between sets. The meet-rank of K sets is equal to p if there are p different transformations and combinations with correlation matrix of rank one. Or, to put it differently, if there are p different transformations and combinations with generalized canonical correlation equal to one. Again we can choose for each variable if we desire a multiple or a single treatment. Now suppose K = m, i.e. all sets of variables consist of exactly one variable. Then the meet-rank is equal to p if there are p transformations that give a correlation matrix of rank one. Remember that the join rank is equal to one if <u>all</u> transformations give a correlation matrix of rank one. The multiple approach to the join rank one problem is to compute p transformations with a rank one correlation matrix, and consequently the solutions can also be used as a solution for the meet problem with rank p. Thus join problems can be identified in practice with meet-problems in which K = m. Again this will be explained in more detail in the later chapters.

Meet-problems with K < m occur in many disguises in linear MVA. Our programs implement various generalizations of these linear problems. Thus CANALS has K = 2 and generalizes canonical correlation analysis, CRIMINALS has K = 2 and generalizes multiple group discriminant analysis, PATHALS has K = 2 and generalizes path analysis, MORALS has K = 2 and generalizes multiple regression. OVERALS is a program for general K, which generalizes K-set canonical analysis. If we choose K = m in OVERALS then we are back to PRINCALS, if we choose K = m and all variables are both nominal and multiple then we are back to HOMALS. By choosing K = 2 in OVERALS we recover CANALS, and the less general programs MORALS, PATHALS, CRIMINALS, which have only been written because the special

structure of the problem leads to more efficient algorithms or to more specialized output.

1.8.2 Optimal scaling and alternating least squares

The basic computational ingredients in our computer programs are the alternating least squares method and the concept of optimal scaling. The loss functions we are minimizing have two different sets of parameters: in the first place a basis for the meet or the join (corresponding with scores for the individuals or objects), and in the second place parameters for the transformations of the variables (or for the quantifications of the categories). The loss functions are all of the least squares type, and they have the obvious property that they are zero for a particular choice of the parameters if and only if that choice defines a perfect solution for the corresponding meet or join problem. The algorithms we use are usually (but not always) of the alternating least squares type, by which we mean that in each iteration two substeps are alternated. In the first substep we compute the optimum basis for given values of the transformations, in the second substep we compute new values for the optimum transformations for the given basis computed in the first substep. Alternating these substeps obviously produces a decreasing sequence of loss function values which always converges because loss is bounded below by zero. Under some mild regularity conditions we can also prove that the basis and transformations converge to values corresponding with a stationary value of the loss function.

This particular way of computing the transformation is called optimal scaling, because the transformations are chosen in such a way that they minimize the loss function. In other forms of MVA the transformations are chosen on a priori grounds, after which ordinary linear MVA is applied. Of course optimality must not be interpreted in any wider sense, we do not pretend that our procedures always give better transformations than other procedures, they are only better in terms of the particular loss function we choose. Whether they are better in any wider sense must in principle be decided by the gauging process.

Our choice of a least squares loss function could be considered as old-fashioned. The results of robust statistics can be interpreted as showing that least squares loss functions often fail, and can almost always be replaced by more appropriate ones. Our basic join and meet philosophy is formulated in purely algebraic terms, and least squares only enters the picture if we start making use of the fact that most of the linear spaces we study in practice can be made into inner product spaces, with a corresponding (weighted) least squares distance function. But other norms for the linear spaces could in principle be used. They lead to unpleasant complications in the computational process,

and, even more importantly, they destroy most of the geometrical interpretations of our procedures which play an important role in the making and interpreting of pictures. It is quite clear, however, that in some cases the least squares loss function is a rather poor choice, for example if outliers are quite common. We explicitly study the consequences of our choice of loss function, however, in our analysis of stability and in our gauging of the technique on various theoretical and practical examples. Thus it is conceivable that in later versions of this book other norms will be studied and used, but for the moment we have to be satisfied with least squares.

We also do not wish to maintain that alternating least squares always gives the best algorithm. In fact it is clear that in some situations much more efficient computation is possible. Solving a join or meet problem is in the simplest cases (discussed mainly in chapters 3 and 4) equivalent to finding the partial or truncated singular value decomposition of a given matrix, and sometimes alternating least squares (which is equivalent to the familiar power method in these cases) is not a good algorithm for computing the singular value decomposition. Other methods are available which are faster, use less storage, and provide us with all singular values and/or singular vectors. In our ANACOR and ANAPROF programs discussed in chapter 4, for example, we use singular value algorithms not based on alternating least squares. The major advantage of alternating least squares is its generality, it can also be applied if the problem is not of the singular value type. Moreover alternating least squares can be applied to extremely large examples, in which singular value decomposition using other methods is not practical.

Alternating least squares is also useful because the transformations of the variables are often restricted in various ways. We have already seen that single variables in the multiple approach are restricted by proportionality constraints, another very common restriction is imposing monotonicity when analyzing ordinal variables. These restrictions have as a consequence that the resulting problems are not equivalent to singular value decompositions any more. The various restrictions that are possible and useful in our programs will be discussed in detail in chapters 5 and 12.

1.8.3 Dimensionality and data massage

We have been very casual so far about the choice of the dimensionality. Observe in the first place that there are two different dimensionalities involved. We have to choose the join rank or meet rank and also the number of different solutions we want to compute. These choices are usually dependent on the type of program we want to use. If we use HOMALS in a join problem, then this implies

that we choose join rank equal to one, and we only have to decide how many solutions we want to compute. If we use PRINCALS then we decide to compute only one solution, but we must decide what join rank we want to use. We have merely said so far that the dimensionality must be chosen by the user, but this is somewhat unfair because the user obviously needs some guidelines.

In almost all of the examples in this book the rank and/or dimensionality is either equal to one or equal to two. Computing just one HOMALS solution (if all variables are nominal this is the same thing as computing a PRINCALS solution with join rank one) is used quite often to compute 'optimal' transformations or quantifications of the variables. These transformed or quantified variables are then used in a subsequent linear MVA or in any other data analytic techniqu which requires numerical variables. We call this using HOMALS as a first step. It is also possible to use PRINCALS or CANALS as a first step, but this is considerably less common.

Computing two HOMALS solutions, or a PRINCALS/CANALS/OVERALS/etc with rank two is also very common. The basic emphasis in this case is on making pictures. It is clear from our experience with non-linear MVA that studying pictures in more than two dimensions is not very rewarding. The 'rotation problem' of linear MVA becomes considerably more complicated in nonlinear MVA, because we in fact often have to transform nonlinear manifolds of points to 'simple structure'. The first two dimensions often give us a fairly clear idea of the most important effects in the data. If we interpret our programs as fitting models (for example the model that the join rank is equal to two) then the routine choice of $p = 2$ is not defensible, there is no reason whatsoever why $p = 2$ should occur more 'in the real world' than any other value of p. But if we interpret our techniques as making transformations and preparing for a linear analysis, then $p = 1$ is the natural choice. And if we see them as techniques for making pictures of data then $p = 2$ is the only reasonable choice.

There are some cases in which two-dimensional solutions indicate that the results are dominated by either a single deviating object or by the properties of a single variable. These solutions are similar to 'degenerate' solutions in multidimensional scaling, although they are not quite so degenerate, because multidimensional scaling actually shifts points to infinity and collapses clusters into single points. In such degenerate cases it often happens that th interesting structure is hidden somewhere in the higher dimensions or in the remaining solutions. We could consequently study higher dimensional solutions to find the interesting structure, but we have seen that this is very problema in practice. Our solution is consequently a different one: we delete offending

objects and/or variables and compute a new two dimensional solution. In our analysis of stability in chapter 11 we will show that often this has the same effect as eliminating the offending dimension.

Some people call this massaging the data, with the implication that it is subjective, not respectable, and possibly somewhat indecent. It is true that it is difficult to formalize this massage process, because it involves what Tukey calls 'judgment' in his 1962 paper. We do not agree, however, with the point of view that the process can be used to find any conclusion or interpretation that you want to find. The main safeguard is that in reporting the analysis we should also report and motivate our deletions and other manipulations. It is also extremely naive to suppose that users of classical statistical techniques never apply massage, the major difference seems to be that exploratory data analysis encourages people to report this explicitly, while classical statistics seems to encourage the attitude that massaging should only be practiced behind closed doors (compare our discussion in section 1.4).

2 Notation and terminology

2.1 Complete indicator matrix

2.1.1

Basic in MVA is a finite set of n objects (or individuals). A variable is a function η_j that maps the set of objects into a finite set of k_j categories; this set of categories is called the range of η_j.

We shall assume that there is a finite number of m variables η_j (j=1,...m).

The Cartesian product of all categories is called the multivariate range. Its elements are the n combinations of m categories for each object; they are called profiles. Since variables are ordered (from 1 to m), each profile is an ordered m-fold.

The data matrix H is an n x m matrix with elements h_{ij} giving the category of variable η_j for object i. These elements are not necessarily numbers.

Example.

An example of a data matrix H is given in table 2.1, with n=10, m=3, k_j=3 (j=1,2,3). Elements of H are 'category labels': the first variable has categories a, b, c; the second p, q, r; the third u, v, w (with zero frequency for w).

2.1.2

The number of possible profiles equals Πk_j, the product of all k_j. It may happen that this number is much smaller than n. The data matrix then is not the most efficient way of coding. In stead one might prefer a profile frequency matrix. A complete profile frequency matrix would list all possible profiles, and indicate for each one how often it occurs. Such a matrix has Πk_j rows and (m+1) columns: the first m elements of a row give the categories of the profile and the last element shows its frequency.

Example.

Table 2.2 shows the complete profile frequency matrix derived from the data matrix of table 2.1.

Obviously, if many profiles have zero frequency, it becomes more economical to drop the corresponding row from it; we then obtain a reduced profile frequency matrix, as illustrated in table 2.3.

2.1.3

Another possible coding is as follows. Profiles correspond with the cells of an m-dimensional k_1 x k_2 x.....x k_m array. Inserting profile frequencies in the appropriate cells, a higher dimensional cross tabulation is obtained. For the example this is illustrated in table 2.4.

Table 2.1

a	p	u
b	q	v
a	r	v
a	p	u
b	p	v
c	p	v
a	p	u
a	p	v
c	p	v
a	p	v

Table 2.1.

Example of data matrix H.

Table 2.2

a	p	u	3
a	p	v	2
a	p	w	0
a	q	u	0
a	q	v	0
a	q	w	0
a	r	u	0
a	r	v	1
a	r	w	0
b	p	u	0
b	p	v	1
b	p	w	0
b	q	u	0
b	q	v	1
b	q	w	0
b	r	u	0
b	r	v	0
b	r	w	0
c	p	u	0
c	p	v	2
c	p	w	0
c	q	u	0
c	q	v	0
c	q	w	0
c	r	u	0
c	r	v	0
c	r	w	0

Table 2.2.

Profile frequency matrix.

Table 2.3

a	p	u	3
a	p	v	2
a	r	v	1
b	p	v	1
b	q	v	1
c	p	v	2

Table 2.3.

Reduced profile frequency matrix.

Table 2.4

u

	p	q	r
a	3	0	0
b	0	0	0
c	0	0	0

v

	p	q	r
a	2	0	1
b	1	1	0
c	2	0	0

w

	p	q	r
a	0	0	0
b	0	0	0
c	0	0	0

Table 2.4. Higher dimensional cross tabulation.

2.1.4

A third way of coding data will be of crucial interest for the type of analysis described in this text. For each variable η_j an n x k_j binary matrix G_j is defined, by taking

g^j_{ir} = 1 if the i^{th} object is mapped in the r^{th} category of η_j,

g^j_{ir} = 0 if the i^{th} object is not mapped in the r^{th} category of η_j.

G_j is called the indicator matrix of η_j. Such matrices can be collected in a supermatrix $G = (G_1,..,G_j,..,G_m)$ of dimension n x Σk_j, also called 'indicator matrix'. For the example of table 2.1 the indicator matrix G is shown in table 2.5.

2.2 Properties of a complete indicator matrix

The indicator matrix G_j is said to be complete if each row of G_j has only one element equal to unity and zero's elsewhere, so that row sums of G_j are equal to unity. The latter can be written as $G_j u = u$, where u is a vector of unit elements. If all G_j are complete, their combined matrix G is also said to be complete, and it then follows that Gu = um: rows of G add up to m.

Let d_j be the vector of the column totals of G_j. Its r^{th} element d^j_r corresponds to the marginal frequency of the r^{th} category of η_j. Also, the sum of the elements of d_j must be equal to $u'd_j = n$.

Write $D_j = G_j'G_j$. This matrix D_j must be diagonal (columns of G_j are orthogonal), and its diagonal elements are the same marginal frequencies as those given in d_j.

Define $C_{jl} = G_j'G_l$; it is a two dimensional cross table for variables η_j and η_l. Its elements correspond to the frequency of objects characterized by a particular combination of one category in η_j and one in η_l.

Define C = G'G. This matrix C combines all C_{jl} with along its diagonal the matrices $D_j = C_{jj}$. It is a matrix of bivariate marginals.

Define D as the superdiagonal matrix ') of C, in the sense that elements of D and C are identical in the diagonal submatrices $C_{jj} = D_j$, whereas D has zero elements in its off-diagonal submatrices. D is a matrix of univariate marginals.

For the numerical example, C and D are given in table 2.6 and 2.7. The 3 x 3 submatrices along the diagonal of D are the same as those along the diagonal of C.

') Although D is strictly a diagonal matrix for a complete indicator matrix, we prefer to think of it as a superdiagonal matrix, because of later, somewhat different, applications.

Table 2.5

a b c	p q r	u v w
1 0 0	1 0 0	1 0 0
0 1 0	0 1 0	0 1 0
1 0 0	0 0 1	0 1 0
1 0 0	1 0 0	1 0 0
0 1 0	1 0 0	0 1 0
0 0 1	1 0 0	0 1 0
1 0 0	1 0 0	1 0 0
1 0 0	1 0 0	0 1 0
0 0 1	1 0 0	0 1 0
1 0 0	1 0 0	0 1 0

Table 2.5 Indicator matrix G for data matrix H of table 2.1

Table 2.6

	a b c	p q r	u v w
a	6 0 0	5 0 1	3 3 0
b	0 2 0	1 1 0	0 2 0
c	0 0 2	2 0 0	0 2 0
p	5 1 2	8 0 0	3 5 0
q	0 1 0	0 1 0	0 1 0
r	1 0 0	0 0 1	0 1 0
u	3 0 0	3 0 0	3 0 0
v	3 2 2	5 1 1	0 7 0
w	0 0 0	0 0 0	0 0 0

Table 2.6 Matrix C of bivariate marginals

Table 2.7

	a b c	p q r	u v w
a	6 0 0	0 0 0	0 0 0
b	0 2 0	0 0 0	0 0 0
c	0 0 2	0 0 0	0 0 0
p	0 0 0	8 0 0	0 0 0
q	0 0 0	0 1 0	0 0 0
r	0 0 0	0 0 1	0 0 0
u	0 0 0	0 0 0	3 0 0
v	0 0 0	0 0 0	0 7 0
w	0 0 0	0 0 0	0 0 0

Table 2.7 Matrix D of univariate marginals

Table 2.8

	a	b	c	d	e	f
old	1	0	0	0	0	0
	1	1	0	0	0	0
	0	1	1	1	0	0
	0	0	1	1	1	0
new	0	0	0	1	1	1

Table 2.8A. Incomplete indicator matrix.

	$+^{a}_{-}$	$+^{b}_{-}$	$+^{c}_{-}$	$+^{d}_{-}$	$+^{e}_{-}$	$+^{f}_{-}$
old	1 0	0 1	0 1	0 1	0 1	0 1
	1 0	1 0	0 1	0 1	0 1	0 1
	0 1	1 0	1 0	1 0	0 1	0 1
	0 1	0 1	1 0	1 0	1 0	0 1
new	0 1	0 1	0 1	1 0	1 0	1 0

Table 2.8B. Completed indicator matrix

2.3 Quantification

Categories of variables may be numerical values, like midpoints of intervals on some continuous variable. In that case the n x m data matrix H is a "classical" multivariate data matrix and can be handled with classical techniques of linear MVA.

Most of the techniques to be discussed in this text, however, do not assume such an apriori quantification. Even in the case that prior quantification is available, such quantification could be ignored and replaced by a "nominal" categorization.

Example. Suppose we have a variable "age" that maps individuals into 15 age groups, each age group being characterized by an interval midpoint on the age scale. The data matrix would give for age a single column with 15 possible numerical values. The corresponding indicator matrix would have 15 columns, one for each age group. We might, from then on, as it were "forget" that these 15 categories were apriori ordered on an "interval scale". and interpret them as 15 nominal categories.

Quantification of categories should, of course, follow rules, with the intention to optimize some criterion, or, in other words, with the intention to minimize some loss function. For the moment we shall not discuss loss functions, however, but indicate in a global way how quantification of an indicator matrix is feasible.

Quantification of categories of variable η_j implies that these k_j categories are mapped as the k_j numerical values of a vector y_j. Then $q_j = G_j y_j$ becomes a single vector which gives the numerical result for each object with respect to η_j. Define x as the mean vector of all q_j, i.e.,

$$x = \frac{1}{m} \Sigma q_j$$

This vector x now will be the object quantification, and we will say that for some direct quantification y_j of categories, x is the induced score of objects.

On the other hand, let x be some direct quantification of the objects We then define the induced category quantification as the average of the scores of those objects which are mapped into that category. In formula:

$$y_j = D_j^{-1} G_j' x$$

(The latter assumes that D_j has an inverse, which implies that there are no categories with zero frequency. If some category has zero frequency, we may as well skip its column from the indicator matrix.)

The two procedures can be unified as follows. Let y_j be a direct quantification of the categories of the j^{th} variable. Let y be a vector that combines all y_j in a single vector with k_j elements. Induced object scores then are Gy/m. We now require that a solution for the direct quantification of objects, x, must be proportional to the induced object scores, and vice versa, that direct category quantification y_j must be proportional to the induced category quantification $D_j^{-1}G_j'x$. In chapter 3 it will be shown that this requirement not only makes solutions for x and y feasible, but that it also results in minimization of attractive loss functions.

The discussion above should not suggest that there is only one solution for x and y. In general, we might be interested in p different solutions. This implies that the category quantification corresponds to a matrix Y_j of dimension k_j x p. Induced object scores then are given in the n x p matrix GY/m. Similarly, given an n x p direct quantification X of the objects, induced category scores appear in the k_j x p matrices $D_j^{-1}G_j'X$.

2.4 Incomplete indicator matrix

2.4.1

Thus far we have described a complete indicator matrix. Its typical feature is that each row of G_j adds up to unity. This could be stated more formally by defining M_j as the diagonal matrix of row totals of G_j. For a complete indicator matrix G_j it then must be true that $M_j = I$. Also, define $M_* = \Sigma M_j$. For complete indicator matrix G this implies $M_* = mI$.

An indicator matrix G_j is incomplete if it has rows with only zero elements, and if, by adding more columns, it can be made complete.

An example is the following. Let G have n rows, one for each of n individual parliamentarians. Columns of G correspond to m different proposals. G has entry '1' if the corresponding parliamentarian voted 'in favor' of the proposal, and '0' otherwise. The matrix could be completed by adding, for each proposal, a second column in which "1" is registered if the individual voted "not in favor", and "0" otherwise. This creates for each proposal a complete indicator matrix G_j with two columns, and $M_j = I$.

Another example is that of missing data. It will be discussed more fully in section 2.5. For the moment we remark that, if an individual has missing data on the j^{th} variable, this could be coded by registering zero's only in the corresponding row of G_j. Then G_j is incomplete, but could be completed by adding a column with entries "1" for individuals with missing data.

2.4.2

An incomplete indicator matrix can be quantified according to the same principles as outlined in section 2.3 for the complete case. Again we want object scores to be proportional to the vector of average category quantification for categories that apply to the object, and, vice versa, category quantification proportional to the average scores of objects within the category. In formulae, we require

$$x = M_*^{-1} G y$$

$$y = D_j^{-1} G_j' x$$

A solution based on these requirements will be different from a solution based on the completed indicator matrix. The reason is that, in general, object scores will become more similar to the extent two objects have more categories in common. In the example with missing data, for instance, if we add to G_j a column with '1' for each object with missing data on that variable, the effect will be that such objects will be quantified closer together, as if 'having missing data' can be interpreted as a 'positive' characteristic shared by those objects.

2.4.3 Examples

(i) In the example of the parliamentarians (section 2.4.1), completion of the matrix implies that parliamentarians who vote 'not in favor' will be considered as being in the 'same category' in this respect. However, parliamentarians can vote against some proposal for opposite reasons, and it follows that analysis of the incomplete indicator matrix could be more realistic than analysis of the completed matrix.

(ii) A well-known method in archeology is 'seriation'. In the graves of an ancient grave-yard, fragments of pottery are found, with different decorative motives on them. If motives belonging to two different graves are quite similar, one might assume that the graves are about equally old. The seriation problem is solved if it appears possible to find an order for the graves, and simultaniously an order for the motives, in such a way that the 'parallelogram' structure is found which is illustrated in table 2.8A.

For the incomplete matrix the oldest and the newest grave have nothing in common and it can be expected that they will be quantified at opposite ends of a scale. In the completed matrix, however, these two graves share the characteristics b- and c-. This must lead to a solution where these two graves come closer together (the numerical solution for this example is given in section 3.1 In the context of multidimensional scaling theory the present example illustrate a Coombs scale. Typical of the representation of a Coombs scale is that it maps objects as segments of the continuum in such a way that this segment includes th categories that apply to the object. It then follows that objects which do not

Table 2.9

	a	b	c	d
1	0	0	0	0
2	1	0	0	0
3	1	1	0	0
4	1	1	1	0
5	1	1	1	1

Table 2.9A. Guttman scale

	$+^{a}_{-}$	$+^{b}_{-}$	$+^{c}_{-}$	$+^{d}_{-}$
1	0 1	0 1	0 1	0 1
2	1 0	0 1	0 1	0 1
3	1 0	1 0	0 1	0 1
4	1 0	1 0	1 0	0 1
5	1 0	1 0	1 0	1 0

Table 2.9B. Completed indicator matrix for Guttman scale

	a-	b-	c-	d-	a+	b+	c+	d+
1	1	1	1	1	0	0	0	0
2	0	1	1	1	1	0	0	0
3	0	0	1	1	1	1	0	0
4	0	0	0	1	1	1	1	0
5	0	0	0	0	1	1	1	1

Table 2.9C Ordered completed indicator matrix for Guttman scale

share some category, are not necessarily close together. (We come back to this subject in section 4.2, on unfolding theory.)

(iii) Table 2.9A gives a typical example of a Guttman scale. Objects and categories can be ordered in such a way that zero's appear only in the upper right corner. In other words, an object that has some category, must also have all categories to the left of that category. Completing the matrix results in table 2.9B (two adjacent columns, one registering 'presence' and the other 'absence' of the category).

This completed matrix can be re-ordered as in table 2.9C. This last table shows the typical Guttman 'parallelogram' structure, and the scaling solution of the completed matrix remains consistent with the ordering implied in the incomplete matrix.

2.5 Missing data

2.5.1

A special and ever recurring problem in MVA is the presence of missing data. They can occur for a variety of reasons: a subject left a blank on his response sheet, an experimental animal died, an "impossible" code has been entered in the code-book, an archeological fragment is so damaged that one cannot decide whether a certain motive was ever present or not.

In classical MVA many ways of handling missing data have been proposed, such as (i) to insert a random value selected from the range of possible values, (ii) to insert the mean of the variable, (iii) to insert the best prediction from other variables. Or, for derived data such as a correlation matrix, one might insert values that minimize the rank of this matrix, or that maximize its largest eigenvalue, etc.

Sometimes inserting values for missing data has no other purpose than making it possible to perform standard calculations. The inserted values are "stand-ins", so to speak, and should leave the scene before results are presented. But there are also situations where it is of special interest to make a sophisticated estimate about the missing value (as in the example of a damaged fragment, where the archeologist will want to make the best guess as to whether or not a motive was present).

2.5.2 One way to handle missing data is to throw away each object with missing data. This option has little to do with data analysis as such, and we shall not further discuss it.

From the point of view of data analysis we shall distinguish between the following three options:

Table 2.10

a	b	c	p	q	r	u	v
0	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1
0	0	0	0	0	1	0	1
1	0	0	1	0	0	1	0
0	1	0	1	0	0	0	1
0	0	1	1	0	0	0	1
1	0	0	1	0	0	1	0
1	0	0	1	0	0	0	1
0	0	1	0	0	0	0	1
1	0	0	1	0	0	0	1

Table 2.10A. Incomplete indicator matrix with missing data corresponding to option (i).

a	b	c	?	p	q	r	?	u	v
0	0	0	1	1	0	0	0	1	0
0	1	0	0	0	1	0	0	0	1
0	0	0	1	0	0	1	0	0	1
1	0	0	0	1	0	0	0	1	0
0	1	0	0	1	0	0	0	0	1
0	0	1	0	1	0	0	0	0	1
1	0	0	0	1	0	0	0	1	0
1	0	0	0	1	0	0	0	0	1
0	0	1	0	0	0	0	1	0	1
1	0	0	0	1	0	0	0	0	1

Table 2.10B. Completed indicator matrix, missing data single category (option (ii)).

a	b	c	?	?	p	q	r	?	u	v
0	0	0	1	0	1	0	0	0	1	0
0	1	0	0	0	0	1	0	0	0	1
0	0	0	0	1	0	0	1	0	0	1
1	0	0	0	0	1	0	0	0	1	0
0	1	0	0	0	1	0	0	0	0	1
0	0	1	0	0	1	0	0	0	0	1
1	0	0	0	0	1	0	0	0	1	0
1	0	0	0	0	1	0	0	0	0	1
0	0	1	0	0	0	0	0	1	0	1
1	0	0	0	0	1	0	0	0	0	1

Table 2.10C. Completed indicator matrix, missing data multiple categories (option (iii)).

(i) the indicator matrix is left incomplete;
(ii) the indicator matrix is completed with a single additional column for each variable with missing data;
(iii) the indicator matrix is completed with in G_j as many additional columns as there are missing data for the j^{th} variable.

Option (i) is called 'missing data deleted'. It implies that when an object has missing data for the j^{th} variable, the corresponding row of G_j is a zero row.
Option (ii) is called 'missing data single category'. It implies that one extra column is added to G_j, with entry "1" for each object with missing data on the j^{th} variable. Missing data are thus treated as if they are a category by themselves. Objects with missing data are handled as if they are in the same category in this respect.
Option (iii) is called 'missing data multiple categories'. It adds to G_j as many extra columns as there are objects with missing data on the j^{th} variable, and each such column has only one entry "1". The option handles missing data as if for each individual each single missing data forms a category of its own.
With not too many missing data, distributed randomly over objects and variables, the three options will have roughly the same results. However, when missing data cluster at some individuals (or variables), results can be rather different.

2.5.3 Example
Suppose that in the data matrix H of table 2.1 data were missing for individuals 1 and 3 on variable 1, and for individual 9 on variable 2. The three resulting indicator matrices are shown in table 2.10.

2.5.4
Note that the options for missing data also can be chosen when an indicator matrix is incomplete for different reasons, such as in the examples of section 2.4.3. E.g., if an indicator matrix is incomplete because it registers only "presence" of categories but not "absence", so that each G_j has only a single column, we might complete G_j with a second column registering "absence" as "1" and"presence" as "0", or we might complete G_j by adding as many columns as there are objects with missing data.

Table 2.11

	I	II	III
1	a	b	c
2	a	b	b
3	b	b	c
4	c	a	a
5	b	c	b

Table 2.11A. Data matrix H, 5 individuals, 3 variables.

	I			II			III		
	a	b	c	a	b	c	a	b	c
1	1	0	0	0	1	0	0	0	1
2	1	0	0	0	1	0	0	1	0
3	0	1	0	0	1	0	0	0	1
4	0	0	1	1	0	0	1	0	0
5	0	1	0	0	0	1	0	1	0

Table 2.11B. "Classical" indicator matrix.

	1	2	3	4	5
I	a	a	b	c	b
II	b	b	b	a	c
III	c	b	c	a	b

Table 2.11C. Transposed data matrix H'.

	1a	1b	1c	2a	2b	3b	3c	4a	4c	5b	5c
I	1	0	0	1	0	1	0	0	1	1	0
II	0	1	0	0	1	1	0	1	0	0	1
III	0	0	1	0	1	0	1	1	0	1	0

Table 2.11D. Reversed indicator matrix.

2.6 Reversed indicator matrix

In some cases it can be useful to 'reverse' the indicator matrix, i.e. to derive the indicator matrix from the transposed data matrix. An illustration is given in table 2.11. Table 2.11A gives the data matrix H for 5 objects and 3 variables; the 'classical' indicator matrix is shown in table 2.11B. Table 2.11C gives the transposed data matrix H'. If we now treat objects as variables (and variables I, II, III as objects), we obtain the 'reversed' indicator matrix of table 2.11D.

Analysis of this reversed indicator matrix implies that we must be prepared to accept II and III as similar on the basis of the columns 2b and 4a, showing that some individuals apply some category to II and III, but not to I.

The most logical application of a reversed indicator matrix is that of a sorting task: individuals are presented with a number of objects and are asked to sort the objects into as many categories as they like, where it is left to the individual's own fancy how he wants to define the categories. Data then consist of different groupings for each individual, on the basis of categories which are not comparable across individuals. Such an example will be given in section 3.12.

Whether the reversed indicator is also of interest in other cases depends on the question to what extent data can be interpreted as if derived from a sorting task.

Analysis of a classical indicator matrix quantifies objects and categories, but does not quantify 'variables'. Analysis of the reversed indicator matrix quantifies variables and categories per individual, but does not quantify individuals.

2.7 Indicator matrix for frequency table

An indicator matrix for a two-dimensional r x c frequency matrix is obtained as follows. The matrix will have as many rows as indicated by the total frequency (added over all cells of the frequency matrix). The matrix will have (r+c) columns: the first r columns for row categories, the last c columns for column categories. Each row will have two entries '1': one for the row category that applies to the individual, and one for the column category. Table 2.12 gives an example.

Obviously such an indicator matrix is not an efficient way of coding data, but sometimes it will be theoretically convenient to imagine data coded in

Table 2.12

	a	b	c
p	4	2	0
q	1	3	2
r	0	1	2

Table 2.12A. Frequency table for 15 objects.

	p	q	r	a	b	c
1	1	0	0	1	0	0
2	1	0	0	1	0	0
3	1	0	0	1	0	0
4	1	0	0	1	0	0
5	1	0	0	0	1	0
6	1	0	0	0	1	0
7	0	1	0	1	0	0
8	0	1	0	0	1	0
9	0	1	0	0	1	0
10	0	1	0	0	1	0
11	0	1	0	0	0	1
12	0	1	0	0	0	1
13	0	0	1	0	1	0
14	0	0	1	0	0	1
15	0	0	1	0	0	1

Table 2.12B. Indicator matrix

2.8 Grouping of categories

Sometimes it makes sense to group categories into a single category. An example is an indicator matrix G_j for one item of a multiple choice examination, with four response categories, one of which is correct. In stead of setting up G_j with four columns (one for each response category), one might take only two columns (one for the correct answer, and one for the other answers).
Results of the two types of indicator matrices will not necessarily be similar. E.g., suppose that individuals differ in response bias (some individuals systematically avoid response category a, others favour it). Such a response bias would not be revealed by the analysis based on two categories per item. Also, the analysis based on the four response categories will produce a scaling of the wrong answers as to their "degree of wrongness". This might be useful in particular when it turns o that some wrong response is chosen by the majority of individuals with man correct responses on the other items; such a result gives reason for a clo inspection of the content of the deviant item.
An illustration is given in the chapter with applications.
When categories are grouped, their quantification will not necessarily be something like the average of the quantifications they would have obtaine before grouping. When one groups categories, one makes the apriori decisi that they must have equal weight, and that their contribution to the obje score must be the same. Ungrouped categories, on the other hand, obtain d erential weights.

2.9 Grouping of variables

Here the number of categories is expanded by creating a combined variable with as many categories as the number of combinations of categories of the initial variables. For example, let one variable have categories a and b, another one categories p,g,r. The combined variable will have categories ap, ag, ar, bp, bg, br.
Grouping of variables can be useful if one suspects that the variables have "interaction" (in the sense of Analysis of Variance). Without grouping, the contribution of a combination of categories to the object score Gy/m is additive. When variables are grouped, their contribution no longer need to be additive (which is precisely the "interaction" effect).
Obviously, more than two variables can be grouped. The most extreme case is that all variables are grouped. The number of combined

categories then becomes equal to the number of posssible response patterns or profiles. Columns of the indicator matrix will add up to marginal profile frequencies. Analysis of such data will be discussed in chapter 4 (section on ANAPROF).
The following two examples illustrate special cases of grouping.

(i) A common procedure for collecting preference data for m stimuli S_i (i=1,..,m) is the method of paired comparisons. It presents all possible pairs of stimuli (S_i, S_j) $(i \neq j)$, and for each pair the individual is asked to say which of the two is preferred.
The (incomplete) indicator for n individuals would have $\frac{1}{2}m(m-1)$ columns, one for each pair, with entry "1" if the first stimulus of the pair is preferred, "0" otherwise. The indicator can be completed by interpreting each pair as a "variable", and defining indicator matrices G_{ij} with two columns, one with entry "1" if S_i is preferred, the other one with entry "1" if S_j is preferred.

(ii) The method of triads is a well known method for collecting similarity data. Given stimuli S_i (i=1,..,m) all possible m(m-1)(m-2)/6 triplets of stimuli are formed, and the individual is asked to tell, for each triplet, which two stimuli are most alike, and which two are least alike. Obviously, there are six possible responses to each triad (S_i, S_j, S_k) :

		most similar		
		ij	ik	jk
least similar	ij	-	ikj	jki
	ik	ijk	-	kij
	jk	jik	kji	-

where the entries in this table give the six possible similarity orderings (assuming that individuals are consistent and transitive).
A possible indicator would be that for each triad an indicator matrix G_{ijk} is created with six columns, one for each response pattern.

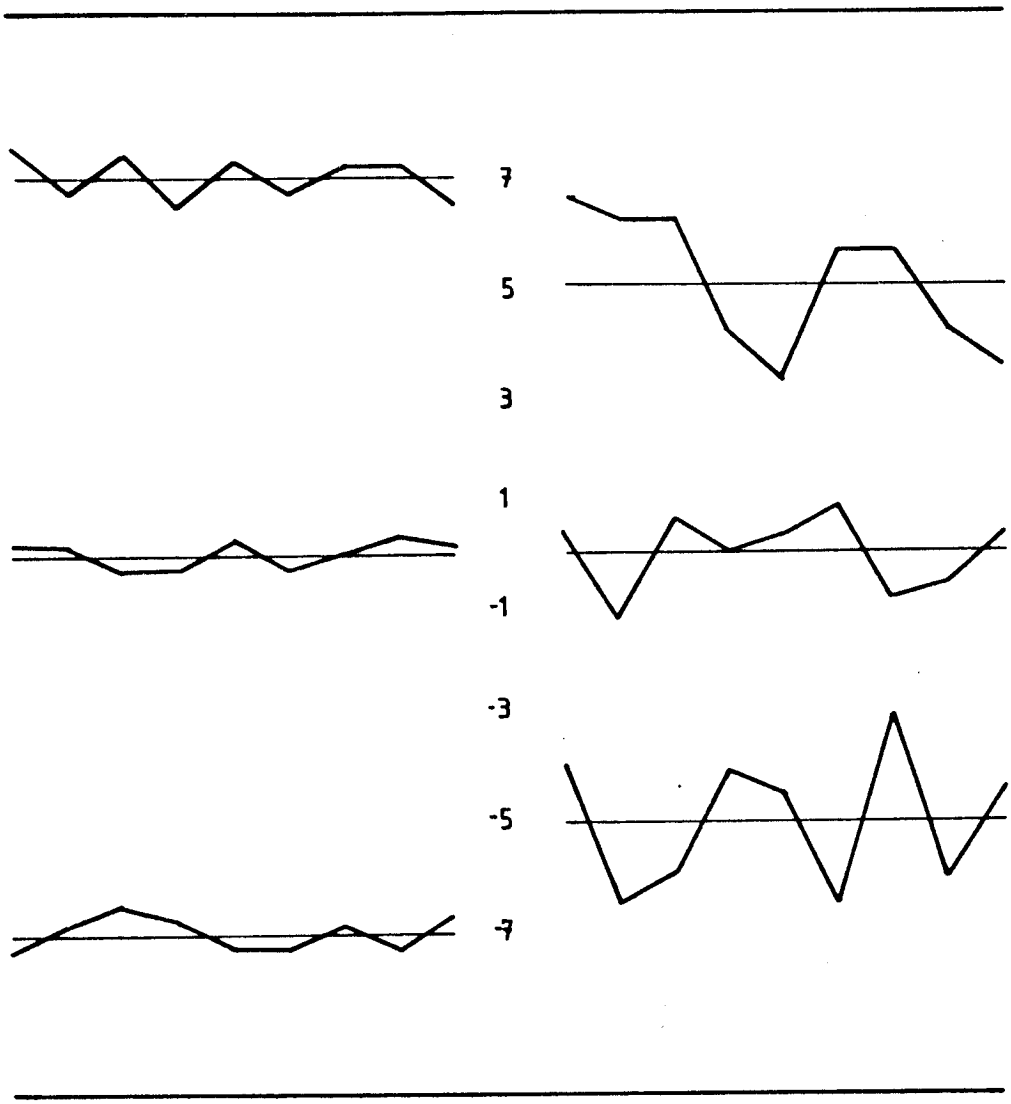


Figure 3.1. The three set of curves at the left have very different levels, but vary little within levels. They illustrate homogeneity. The three curves at the right have less variation between levels and more variation within levels. The set of curves at the right therefore is less homogeneous than the set at the left.

3.1 Homogeneity

Historically, the idea of homogeneity is closely related to the idea that different variables may measure "the same thing". If the latter were perfectly valid, the data matrix (assuming that variables are in deviations from the mean, and identically normalized) would turn out with identical values in each row. Or, if we plot observations as profiles, each profile would be a straight horizontal line.
If the idea of "measuring the same thing" would be imperfectly true (variables measure the same thing, but with random error), rows of the data matrix may have elements that vary somewhat (more to the extent that measurement error increases). A graph of profiles then would show zig-zag curves at different levels. Replacing such profiles by a straight line then implies some "loss of information". Data are homogeneous if the loss is relatively small. This is illustrated in figure 3.1.

3.2 Historical preliminaries

Since the early beginnings of quantitative social science there has been a lively interest in the problem of reduction of multivariate data to univariate scales by means of "weighted averaging". In 1888 Edgeworth wrote a paper on "The Statistics of Examinations", in which he discusses how various parts of an examination might be weighted relative to each other. In 1913 Spearman published a paper "Correlations of Sums and Differences" in which he investigates the effects of differential weighting of variables, and in which he gives a basis for multiple regression and canonical analysis. In the 1930's the problem obtained a new impetus from work on attitude scales (Thurstone, Likert).
In all these studies the basic problem was how to define the univariate scale: by simply adding scores on different variables? or by some sophisticated differential weighting method? Empirical studies, especially those from the field of mental test theory (where the problem is whether we can simply add scores on separate items of a test or should give differential weights to the items) showed that differential weighting had little effect. This literature has been reviewed by Gulliksen (1950, chapter 20), and by Burt (1950, also Burt 1948, 1951). Guilford's general conclusion was "weighting is not worth the trouble" . One should qualify this conclusion against the context of mental test theory, where items usually are selected in such a way that they correlate highly with each other. To substantiate Guilford's conclusion we give some results from classical psychometrics - without proof.

Let $\underline{h}_1, .., \underline{h}_m$ be stochastic variates, with $E(\underline{h}_j) = 0$, and $V(\underline{h}_j) = 1$. Define $r_{jk} = E(\underline{h}_j\underline{h}_k)$. Let $\underline{v} = \Sigma a_j\underline{h}_j$ and $\underline{w} = \Sigma b_j\underline{h}_j$, so that $\underline{v}$ and $\underline{w}$ are weighted sum variates with different weights a_j and b_j. Obviously r_{vw} can take any value between +1 and -1. But if we require that all weights a_j and b_j are non-negative, a lower bound for r_{vw} is the lowest correlation in the correlation matrix R of correlations between $\underline{h}_i$ and $\underline{h}_j$.

Suppose now that we simply add variables to obtain $\underline{w}$ (all weights b_j are equal to unity). In this case a better lower bound for r_{vw} is available: if all elements a_j are non negative, and all elements b_j are equal to unity, then $(r_{vw})^2$ cannot be smaller than the lowest average correlation for the rows of R. This shows that if all correlations in R are large, the correlation between any linear compound with non-negative weights and the simple sum variate necessarily must be large,too . This then is an argument to advocate the simple sum as a reasonable choice for a univariate compression of multivariate data. A recent example of such a plea can be found in Wainer (1976), from whose paper we quote: "Estimating coefficients in linear models: it don't make no nevermind." (For a critical review, cf. Rozeboom, 1979.)

Similar arguments as mentioned above can be derived from selection of random weights. This idea has been explored by Wilks (1938), Burt (1950), and Gulliksen (1950). The general conclusion is that r_{vw} approaches unity to the extent that

(i) variation in weights is small,
(ii) the average correlation is large,
(iii) the number is variables is large,

a very clean and acceptable result. Its practical significance, however, is limited. As with most probabilistic arguments, it is based on an "ideal" situation. Also, in practice one never selects random weights, but one will take weights that are in some sense "optimal".

For binary variables there are theoretical models, like those of Birnbaum or Rasch, from which optimal weights can be derived. The same is true for Spearman's "one factor model", and for Guttman scales. We come back to such models in a later chapter.

At any rate, in the context of mental test theory (how to define the overall test score as a differentially weighted sum of itemscores) the conclusion is that the weighting problem seems trivial. But this conclusion might be mainly valid because in the situation of mental tests intercorrelations between items are necessarily high, since that is the

basis upon which items are selected. But in a different context the problem remains there: to what extent can we replace a number of stochastic variates by one single variate? The problem is implied in a classical paper by Galton (1888), in which he introduced the correlation coefficient. At the end of this paper one reads (p. 144-145) :

> Neither is it necessary to give examples of a method by which the degree may be measured, in which the variables in a series each member of which is the summed effect of n variables, may be modified by their partial co-relation. After transmuting the separate variables as above, and summing them, we should find the probable error of any of them to be $\sqrt{n}$ if the variables were perfectly independent, and n if they were rigidly and perfectly co-related.

Permitting ourselves, following in Burt's steps, a liberal interpretation of this somewhat obscure quotation, it says that the average correlation is a measure for homogeneity among a number of variates. A formal proof is as follows. Let z be the candidate for replacing all h_j. Such a replacement implies loss of information, to be evaluated from the loss function

$$\sigma(z) \triangleq \frac{1}{m} \sum \text{SSQ}(z - h_j)$$

Clearly, $\sigma(z) = 0$ only if $z = h_j$ for all j, which implies that all h_j are identical. Let

$$\sigma_* = \min \{\sigma(z) \mid z\}$$

be the minimum for $\sigma(z)$. It is obtained by taking $z = h_.$ (the mean vector over all h_j) . The loss function then becomes

$$\sigma_* = 1 - \text{SSQ}(h_.) = 1 - r_{..}$$

where $r_{..}$ is the average correlation between all h_j (including $r_{jj}=1$). This result corresponds with Galton's observation.

Numerical example. Let $H = \begin{matrix} 1 & 3 & 2 \\ 3 & -1 & 1 \\ 2 & 1 & -1 \\ -6 & -3 & -2 \end{matrix}$. The question was: can we replace the columns of H by a single vector z, without re-scaling the original columns? The best solution for z is the vector of row means

$$z = \frac{1}{3} \begin{matrix} 6 \\ 3 \\ 2 \\ -11 \end{matrix}$$

The total sum of squares T for H is the trace of

$$H'H = \begin{matrix} 50 & 20 & 15 \\ 20 & 20 & 10 \\ 15 & 10 & 10 \end{matrix} \quad \text{, with } T = (50+20+10) = 80.$$

If we replace each column of H by z, the resulting sum of squares becomes $B = 3z'z = 56.67$. It is called B from Between, because it depends only on differences between rows (elements within a row are identical). A direct expression is $B = u'H'Hu/m$; a direct expression for the total sum of squares T is $T = u'Du$, where D is the diagonal matrix of H'H:

$$D = \begin{matrix} 50 & & \\ & 20 & \\ & & 10 \end{matrix}$$

Define $W = T - B$. The symbol W comes from Within since W gives the sum of squares of deviations from row means within rows. W is a measure of "absolute loss"; in the example $W = 80 - 56.67 = 23.33$. We also could define $W/T = 1 - B/T$ as a measure of "relative loss"; in the example $W/T = .292$.

Suppose now, as a further step, that columns of H are equally normalized (to unity). This can be expressed as $HD^{-\frac{1}{2}}$ (divide h_1 by $\sqrt{50}$, h_2 by $\sqrt{20}$, and h_3 by $\sqrt{10}$). The result is

$$HD^{-\frac{1}{2}} = \begin{matrix} .141 & .671 & .632 \\ .424 & -.224 & .316 \\ .282 & .224 & -.316 \\ -.849 & -.671 & -.632 \end{matrix}$$

with

$$D^{-\frac{1}{2}}H'HD^{-\frac{1}{2}} = \begin{matrix} 1.000 & .632 & .671 \\ .632 & 1.000 & .707 \\ .671 & .707 & 1.000 \end{matrix}$$

(the correlation matrix R). Row means of $HD^{-\frac{1}{2}}$ now become

$$z = \begin{matrix} .482 \\ .172 \\ .063 \\ -.717 \end{matrix}$$

and we find $B = u'D^{-\frac{1}{2}}H'HD^{-\frac{1}{2}}u/m = u'Ru/m = 2.340$, whereas $T = m = 3$, and $W = T-B = .660$. Relative loss becomes $W/T = .220$

The average correlation (averaged over all nine elements of R) becomes $r_{..} = 7.020/9 = .780$, which illustrates $W/T = 1 - r_{..}$.

Correlations between z and the columns of H are

$$r_{z,h} \quad \begin{matrix} .870 \\ .880 \\ .898 \end{matrix}$$

with average $\bar{r}_{z,h} = .883 = (.780)^{\frac{1}{2}}$, which illustrates the equality $W/T = 1 - (\bar{r}_{z,h})^2$.

3.3 Linear weights

3.3.1

Let H be a (finite) data matrix of dimension n x m with column vectors h_j. We assume that column totals are zero. Suppose now that it is allowed to apply linear transformations to h_j in an attempt to further increase homogeneity.

Let z be an arbitrary vector of dimension n and with zero mean. Let y be a vector of weights. Rescaling columns of H now comes to the same as replacing h_j by $h_j y_j$. The problem now becomes to maximize homogeneity by a suitable choice of z and y. Or, we should minimize the loss function

$$\sigma(z;y) \triangleq \frac{1}{m} \sum_j SSQ(z - h_j y_j)$$

Obviously the loss function obtains a degenerate absolute minimum of zero if we take z = 0, and y = 0. To prevent this degenerate solution it is necessary to normalize z so that $z'z \neq 0$, or to normalize y.

(Note that the argument above remains equally valid for stochastic variates $\underline{h}_j$ with $E(\underline{h}_j) = 0$.)

3.3.2

In the following we shall give algorithms for solving for z and y. These algorithms illustrate the principle of alternating least squares. This means that the algorithms proceed in alternating steps, where in one step the loss function is minimized with respect to z for fixed y, and in the other step the loss function is minimized with respect to y for fixed z. We shall describe two varieties of the algorithms, one in which z is normalized whereas the normalization of y is left free, and the other one where y is normalized while the normalization of z is left free. In order to keep the notation simple, we shall first assume that columns of the data matrix H are normalized to unity, which implies that H'H = R (the correlation matrix).

(i) First algorithm (z is normalized).

The algorithm requires an initial arbitrary choice of y ($y \neq 0$) and then proceeds with the following steps.

(1) Calculate $z = Hy/m$

(2) Redefine $z = z(z'z)^{-\frac{1}{2}}$

(3) Redefine $y = H'z$

(4) Go back to step (1) as long as results for z and y are not sufficiently stabilized (according to some selected criterion of accuracy).

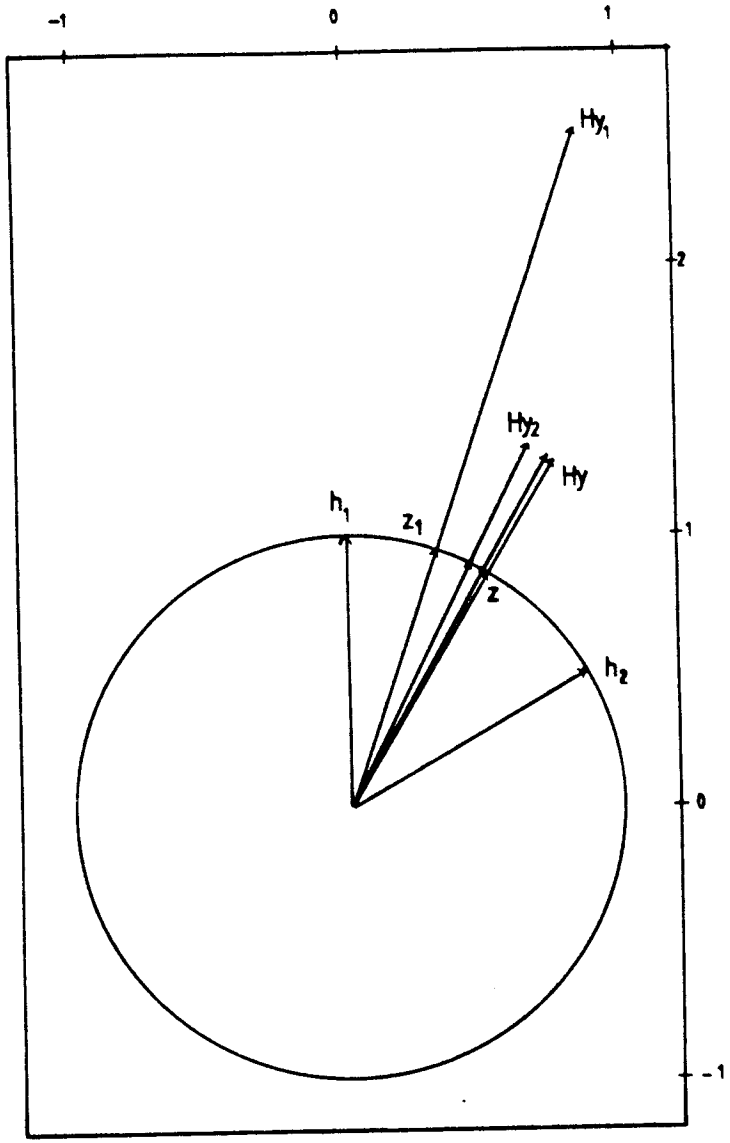


Figure 3.2. First algorithm. Vectors z_i are on the unit circle. Images Hy_i of y_i with respect to H are shown; they converge to Hy (i is the iteration index). z_i is the unit length version of Hy_i, and converges to z.

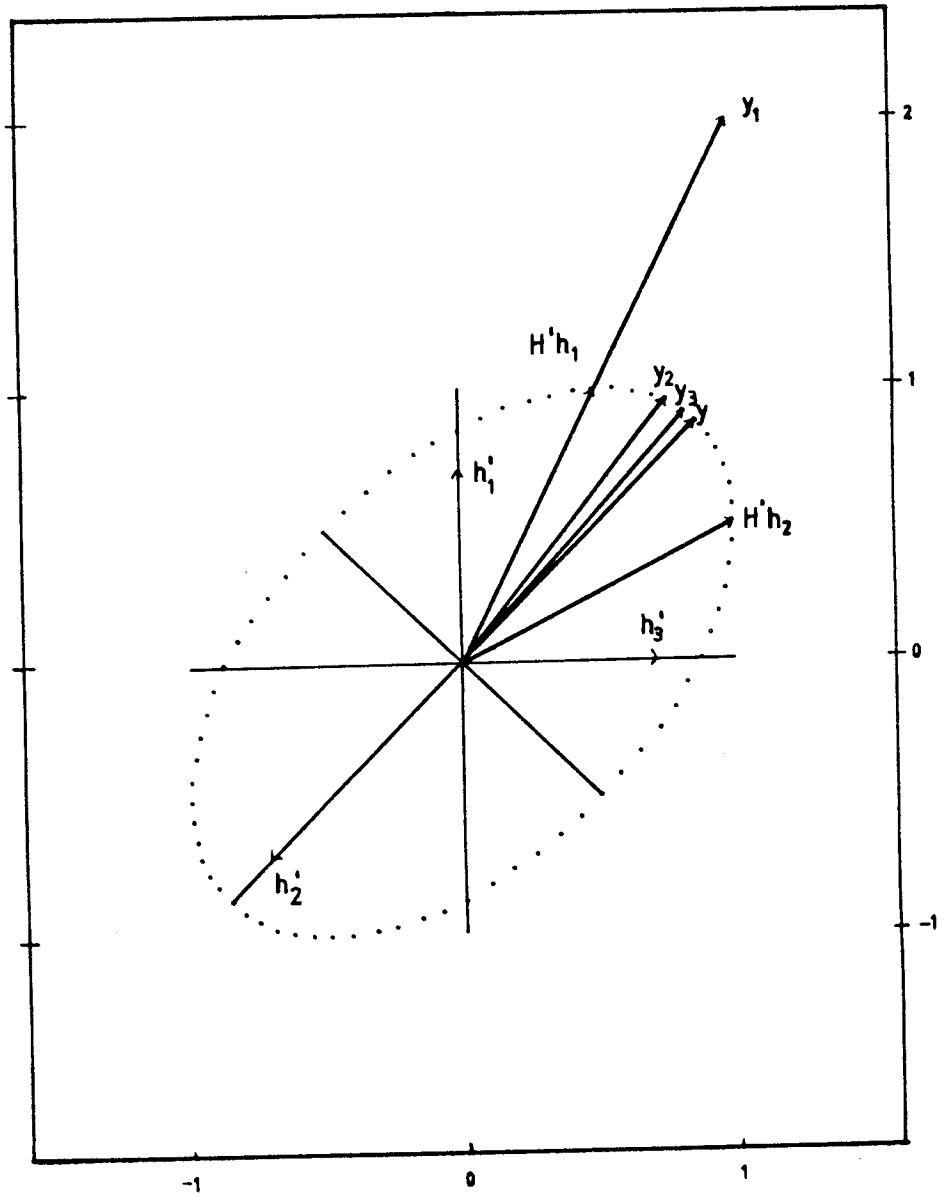


Figure 3.3. Companion to figure 3.2. y_1 is an arbitrary initial vector. The dotted ellipse is the image of the unit circle in figure 3.1. y_i (i=2,...) is the image of z_i. y_i converges to y, the long principal axis of the ellipse.

Comments.

Step (1). Note that Hy/m is a vector of row means of the re-scaled matrix with columns $h_j y_j$. The solution z = Hy/m therefore minimizes relative loss W/T for fixed weights y.

Step (2) controls the normalization of z.

Step (3) redefines y as a vector of correlations between z and the columns of H.

The appendix on Matrix Algebra discusses extensively how z = Hy/m can be interpreted as an "image" of y (and y = H'z is an image of z). The appendix also demonstrates that stationary solutions are obtained when the image H'z = H'Hy/m is proportional to y, or the image Hy = HH'z proportional to z. The algorithm above requires that z is a radius of a hypersphere with unit radius, so that H'z becomes a pseudo-radius of a hyperellipsoid. The algorithm converges to "invariant" directions, or principal axes of the latter hyperellipsoid.

In the appendix the algorithm is related to the SVD solution $H = V\Psi W'$. The algorithm converges to z as the first column vector v_1, with $y = w_1\psi_1$. This implies the equalities

$$H'z = W\Psi V'z = W\Psi V'v_1 = w_1\psi_1 = y$$

$$Hy = V\Psi W'y = V\Psi W'w_1\psi_1 = v_1\psi_1^2 = z\psi_1^2 \quad .$$

In addition, the algorithm converges to a solution with

$$B = y'H'Hy/m = \psi_1 w_1'W\Psi^2 W'w_1\psi_1/m = \psi_1^4/m$$

$$T = y'D_{H'H}y = y'y = \psi_1 w_1'w_1\psi_1 = \psi_1^2$$

with relative loss

$$W/T = 1 - B/T = 1 - \psi_1^2/m.$$

The matrix $W\Psi$ is a "factor matrix" in the sense of a PCA solution, so that $y = w_1\psi_1$ is the first column of such a factor matrix, corresponding to z as the vector of "factor scores" on the first principal component. In the comment on step (3) is was already remarked that y is a vector of correlations between z and the vectors h_j. The solution for y maximizes $y'y = \psi_1^2$ (the sum of the squared correlations).

Numerical example.

For a mini-example, let $H = \begin{matrix} .707 & .000 \\ -.707 & -.707 \\ .000 & .707 \end{matrix}$, with $R = H'H = \begin{matrix} 1.0 & .5 \\ .5 & 1.0 \end{matrix}$

where R is the correlation matrix. Table 3.1 gives results for the first two iterations and the final solution. Figures 3.2 and 3.3 give the geometry of the solution. Figure 3.2 shows the plane of the two column vectors h_j; they are radii of the unit circle. Figure 3.3 gives the image of figure 3.2 with respect to H'. The image of the unit circle now becomes an ellipse. Figure 3.3

iteration	y	Hy	z	H'z=y
1	2	1.414	.534	.944
	1	-2.121	-.801	.755
		.707	.267	
2	.944	.668	.453	.897
	.755	-1.202	-.815	.832
		.534	.362	
		---		
final	.866	.612	.408	.866
	.866	-1.225	-.816	.866
		.612	.408	

Table 3.1 Abridged history of first algorithm

iteration	y	Hy/2	H'z	y
1	.894	.316	.559	.781
	.447	-.474	.447	.625
		.158		
2	.781	.276	.547	.733
	.625	-.497	.508	.681
		.221		
3	.733	.259	.537	.717
	.681	-.500	.524	.698
		.241		
---				
final	.707	.250	.530	.707
	.707	-.500	.530	.707
		.250		

Table 3.2 Abridged history of second algorithm

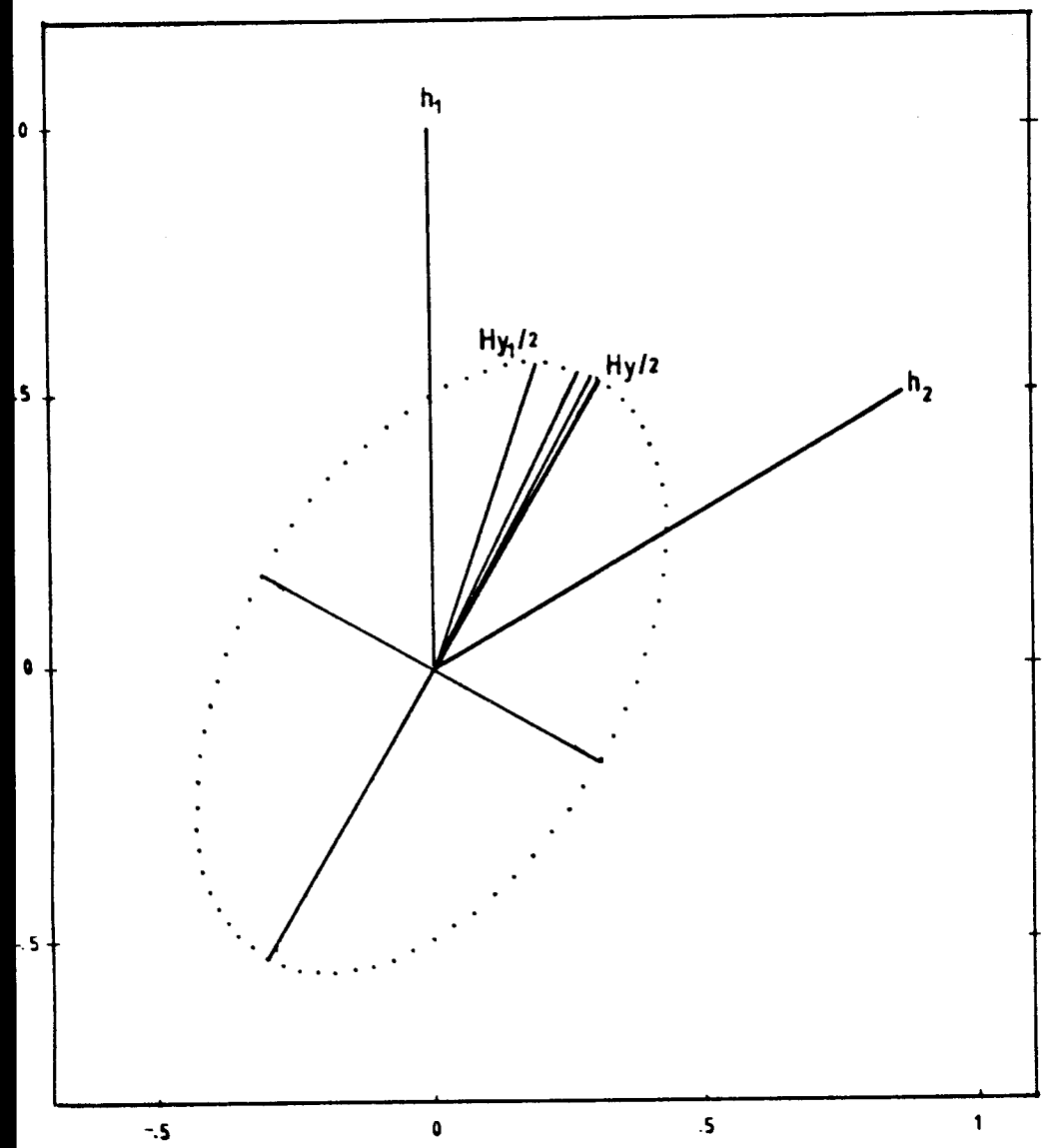


Figure 3.4. Second algorithm. $Hy_i/2$ is the image of y_i with respect to H/2. It converges to Hy/2, the long axis of the ellipse. This ellipse is the image of the unit circle in figure 3.5.

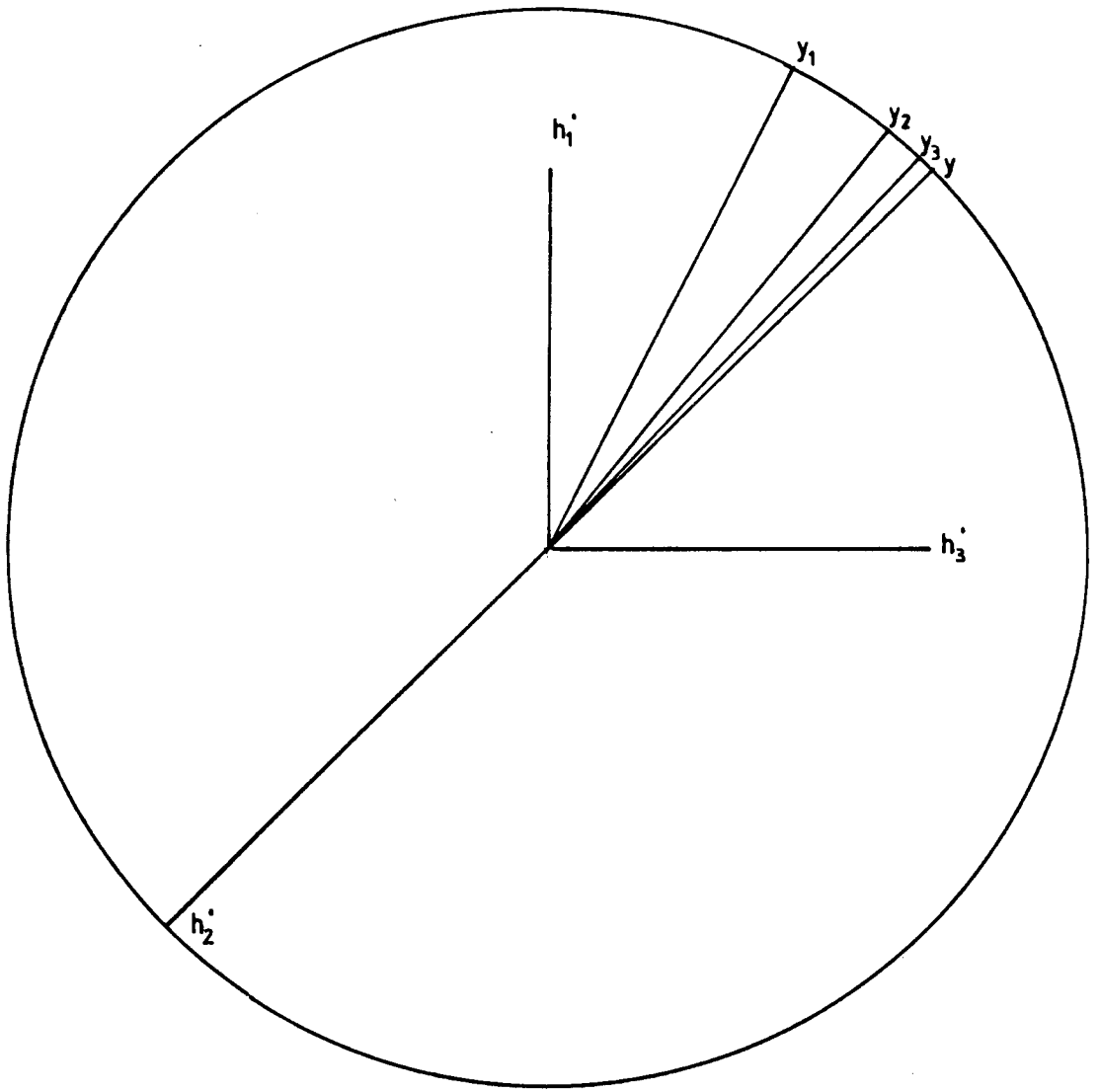


Figure 3.5. y_i is on the unit circle, and converges to y.

also shows the arbitrary vector $y_{(1)}$ (the initial choice for y); its image $Hy_{(1)}$ is in figure 3.2, where $z_{(1)}$ is obtained by giving $Hy_{(1)}$ unit length. The image of $z_{(1)}$ becomes $y_{(2)}$, where the direction of $y_{(2)}$ is closer to the principal axis of the ellipse. And so on, until finally the solution for y coincides with the principal axis.

One might verify the following properties.

(a) y gives "factor loadings" on the first principal component of R.

(b) y'y = 1.5 gives the largest eigenvalue of R, corresponding to the sum of the squared correlations between z and h_j.

(c) The ratio B/T =.75; relative loss is $1 - B/T = 1 - \psi_1^2/m = .25$.

(d) The average correlation in R is equal to B/T = .75.

(e) The average correlation between z and h_j equals $(B/T)^{\frac{1}{2}} = .866$.

(ii) Second algorithm (y is normalized)

The algorithm requires an initial choice of y (y'y=1), and then has the following steps.

(1) Calculate $z = Hy/m$

(2) Redefine $y = H'z$

(3) Redefine $y = y(y'y)^{-\frac{1}{2}}$

(4) Go back to step (1) as long as results are not stabilized.

Comments.

Step (1) minimizes the loss function W/T for fixed y.

Step (2). Since z is not unit-normalized, y now is a vector of covariances rather than correlations.

Step (3) controls the normalization.

Given the SVD solution $H = V\Psi W'$, the algorithm identifies $y = w_1$, and $z = v_1\psi_1/m$, which shows that the two algorithms are basically producing the same results apart from normalization. The result implies

$H'z = W\Psi V'v_1\psi_1/m = w_1\psi_1^2/m$

$Hy = V\Psi W'w_1 = v_1\psi_1 = zm$

$B = y'H''hy/m = w_1W\Psi^2W'w_1/m = \Psi_1^2/m$

$T = y'y = 1$

$W/T = 1 - B/T = 1 - \psi_1^2/m$

Example. For the mini-example used above table 3.2 gives the results. Figures 3.4 and 3.5 show the geometry of the algorithm for this example. Figure 3.4 shows the unit circle of which $y_{(1)}$ is a radius. Its image is $z_{(1)} = Hy_{(1)}$ in figure 3.5 and appears as the pseudo radius of an ellipse. Successive iterations move $z_{(s)}$ towards the principal axis.

3.3.4

We now drop the assumption that columns of the data matrix H are normalized to unity (but retain the assumption that they have zero means). Let D be the diagonal matrix of H'H. It then follows that $HD^{-\frac{1}{2}}$ again has unit normalized columns. The two algorithms need the following adaptations.

(i) The first algorithm starts with an arbitrary choice of y, and proceeds with

(1) Calculate $z = Hy/m$

(2) Redefine $z = z(z'z)^{-\frac{1}{2}}$

(3) Redefine $y = D^{-1}H'z$

(4) Go back to step (1) until results are stabilized.

Comments. Let $HD^{-\frac{1}{2}} = V\Psi W'$ be the SVD solution. The algorithm converges to a solution with $z = v_1$, and $y = D^{-\frac{1}{2}}w_1\psi_1$.

$B = y'H'Hy/m = \psi_1^4/m$

$T = y'Dy = \psi_1^2$

$W/T = 1 - \psi_1^2/m$

(ii) The second algorithm, if we start from an arbitrary choice of z, has steps

(1) Calculate $y = D^{-1}H'z$

(2) Redefine $y = y(y'Dy)^{-\frac{1}{2}}$

(3) Redefine $z = Hy/m$

(4) Go back to step (4) until results are stabilized.

Comment. The algorithm converges to $y = D^{-\frac{1}{2}}w_1$, and $z = v_1\psi_1/m$

Example. We take the same example as in section 3.2, with

$$H = \begin{matrix} 1 & 3 & 2 \\ 3 & -1 & 1 \\ 2 & 1 & -1 \\ -6 & -3 & -3 \end{matrix} \qquad H'H = \begin{matrix} 50 & 20 & 15 \\ 20 & 20 & 10 \\ 15 & 10 & 10 \end{matrix} \qquad D = \begin{matrix} 50 & & \\ & 20 & \\ & & 10 \end{matrix}$$

The history of the two algorithms is shown in tables 3.3 and 3.4. The optimally scaled matrix becomes Q, with columns $q_j = h_j y_j$

$$Q = \begin{matrix} .080 & .387 & .372 \\ .240 & -.129 & .186 \\ .160 & .129 & -.186 \\ -.480 & -.387 & -.372 \end{matrix}$$

of which z (second algorithm) gives the row means. Also, $D^{\frac{1}{2}}y$ (solution for y from second algorithm) is the first eigenvector of the correlation matrix for H (or Q), which has largest eigenvalue $\psi_1^2 = 2.3407$. This illustrates the close relation between optimal scaling and PCA.

iteration	y	Hy	z	$D^{-1}H'z$
1	1	6	.460	.130
	1	3	.230	.192
	1	2	.153	.268
	1	-11	-.843	
2	.130	1.242	.535	.124
	.192	.466	.202	.197
	.268	.184	.079	.283
		-1.892	-.816	
3	.124	1.281	.547	.123
	.197	.458	.196	.197
	.283	.162	.069	.284
		-1.901	-.811	
4	.123	1.282	.548	.123
	.197	.456	.194	.197
	.284	.159	.068	.284
		-1.897	-.810	

Table 3.3 History of algorithm with normalization of z

iteration	z	$D^{-1}H'z$	y	z
1	1	.48	.096	.258
	1	.60	.120	.109
	1	.80	.160	.050
	-3			-.418
2	.258	.064	.083	.277
	.109	.098	.127	.101
	.050	.141	.183	.037
	-.418			-.414
3	.277	.063	.081	.279
	.101	.100	.129	.099
	.037	.145	.185	.035
	-.414			-.413
4	.279	.063	.080	.280
	.099	.101	.129	.099
	.035	.145	.186	.035
	-.413			-.413
5	.280	.063	.080	.280
	.099	.101	.129	.099
	.035	.145	.186	.034
	-.413			-.413

Table 3.4 Iteration history of algorithm with normalization of y

For the numerical example relative loss equals W/T = .220, which is not visibly better than the result for the "unweighted" solution given in section 3.2. In fact, the elements of $D^{\frac{1}{2}}y = \begin{matrix} .567 \\ .577 \\ .588 \end{matrix}$ are almost identical, showing that Hy will be almost proportional to $HD^{-\frac{1}{2}}u$.

3.3.5

The two algorithms above were described for a one-dimensional solution. In fact, the introductory discussion in terms of homogeneity emphasized the idea of a univariate definition. If, however, variables do not "measure the same thing", but are a mixture of measurements of two or more things, the situation changes. We then want p solutions for z_s (s=1,..,p) in such a way that observed variables h_j can be described (with minimum loss) as linear compounds of the z_s.

These ideas become more tangible by looking back at the SVD solution $HD^{-\frac{1}{2}} = V\Psi W'$. A one-dimensional solution approximates $HD^{-\frac{1}{2}}$ as $v_1\psi_1 w_1'$, with relative loss $1 - \psi_1^2/m = \sum_{s=2}^{m} \psi_s^2/m$. A better approximation is obtained if we take two dimensions for the solution, so that the approximation becomes $v_1\psi_1 w_1' + v_2\psi_2 w_2'$ with loss $1 - \psi_1^2/m - \psi_2^2/m = \sum_{s=3}^{m} \psi_s^2/m$.

As to the algorithms described earlier, they converge to the solution for the first singular vectors. To adapt for a higher-dimensional solution, one should use in stead of single vectors z and y, matrices Z (of dimension n x p) and Y (of dimension m x p), and replace the normalization step by a Gram-Schmidt orthogonalization procedure. [1])

[1]) Given a matrix X, its Gram-Schmidt orthogonalization is obtained as follows: X is replaced by $\bar{X}$, with $\bar{x}_1$ identical to x_1, with $\bar{x}_2$ equal to deviations from regression of x_2 on x_1, with $\bar{x}_3$ equal to deviations from regression of x_3 on x_1 and x_2 and so on, until the final column $\bar{x}_m$, which becomes the vector of deviations from regression of x_m on all other column vectors. As a result, $\bar{X}$ will be an orthogonal matrix, with $\bar{X}'\bar{X}$ diagonal. As a final step, columns of $\bar{X}$ are unit normalized, so that after this step $\bar{X}'\bar{X} = I$.

3.4 Linear weighting for K sets of variables

Let the data matrix H be partitioned into K sets

$H = (H_1, .. H_j, .. H_K)$

where H_j has k_j columns, with $\Sigma k_j = m$. The linear optimal scaling problem now can be re-phrased as follows. Let y_j be a vector of weights for H_j, and define $z_j = H_j y_j$. Collect the K vectors z_j in the n x K matrix Z. Define $D_{Z'Z}$ as the diagonal matrix of Z'Z. Relative loss now is minimized by solving for the vectors y_j in such a way that 1 - B/T has a minimum, with

$B = u'Z'Zu/K$

$T = u'D_{Z'Z}u$

An equivalent formulation is the following. Collect all y_j in a single vector y (of dimension m). Define D as the superdiagonal matrix of H'H (i.e., D is identical to H'H in its diagonal submatrices $H_j'H_j$, but replaces off-diagonal submatrices $H_i'H_j$ ($i \neq j$) by zero submatrices). Then

$B = u'Z'Zu/K = y'H'Hy/K$

$T = u'D_{Z'Z}u = y'Dy$

so that y must be solved in such a way that y'H'Hy/y'Dy.K is maximized.

Example. We give a numerical example with K=2, $k_1=k_2=2$, m=4.

$$H = \begin{bmatrix} 1 & -1 & 3 & 0 \\ 3 & -3 & 1 & 1 \\ 2 & 1 & 2 & -1 \\ -5 & 0 & -1 & 2 \\ -1 & 3 & -5 & -2 \end{bmatrix}$$

$$H'H = \begin{bmatrix} 40 & -11 & 20 & -7 \\ -11 & 20 & -19 & -10 \\ 20 & -19 & 40 & 7 \\ -7 & -10 & 7 & 10 \end{bmatrix} \qquad D = \begin{bmatrix} 40 & -11 & 0 & 0 \\ -11 & 20 & 0 & 0 \\ 0 & 0 & 40 & 7 \\ 0 & 0 & 7 & 10 \end{bmatrix}$$

A solution can be obtained with the same algorithms as described in section 3.3 for non-normalized H; the only difference is the definition of D (a superdiagonal matrix in stead of a diagonal matrix). Using the second algorithm with normalization y'Dy = 1, the result becomes

$$y = \begin{bmatrix} .0889 \\ .1035 \\ .0097 \\ -.2297 \end{bmatrix} \qquad Hy = \begin{bmatrix} -.038 \\ -.422 \\ .584 \\ -.914 \\ .791 \end{bmatrix} \qquad Z = \begin{bmatrix} -.068 & .029 \\ -.203 & -.220 \\ .334 & .249 \\ -.445 & -.469 \\ .381 & .410 \end{bmatrix}$$

with Hy = Zu. Since y'Dy =1, we have B/T = B = 1.982/2 = .991, so that relative loss is equal to 1 - B/T = .009. The appendix on matrix algebra shows that this solution, for K = 2, is the

solution for canonical correlation, dependent on the generalized eigenvector equation $H'Hy = Dy\lambda$, with $\lambda = 1 + \rho$, where ρ is the canonical correlation coefficient (the correlation between z_1 and z_2). For the example we find $\rho = \lambda - 1 = 1.982 - 1 = .982$. (The example also shows that $z_1'z_1 = z_2'z_2 = .50$. This result is typical for the case K = 2, and has no simple equivalent when $K \geq 3$.)

3.5 More historical comments

The argument that weighting is unnecessary, implies that the largest eigenvalue λ_1 of the correlation matrix is close to the value $mr_{..}$ (where $r_{..}$ is the average of the cells of the correlation matrix). (In the example of 3.3.4 we found this eigenvalue to be equal to 2.3407. For the example $r_{..} = .7801$, with $mr_{..} = 2.3403$, very close to the eigenvalue. In fact, the example did show that differential weighting gives no visible improvement with respect to relative loss.)
It can be shown that, if R is a matrix of positive correlations,
$\min r_{j.} \leq \lambda_1/m \leq \max r_{j.}$
In words, λ_1/m must be in the range between the smallest and the largest row average of R. It follows that for variables with large and not very much different correlations (such as items of the same test) it will be true that"weighting is not worth the trouble" (cf. section 3.2), or, as Rozeboom puts it: "To put it bluntly, second digit precision in item weighting is generally a waste of effort." (1979, p. 296)
On the other hand, in cases where row means of a correlation matrix are different, weighting can be rewarding enough. The solution sketched in section 3.3, and which related differential weighting to singular value decomposition, was, according to Burt, first mentioned by MacDonell (1901, p. 209). The relevant quotation is: "Professor Pearson has pointed out to me that the ideal index characters would be given if we calculated the seven directions of uncorrelated variables, that is, the principal axes of the correlation 'ellipsoid', and, farther on the same page: "I propose to return in a later paper to this calculation". As far as we know the latter promise never has been redeemed. The quotation shows that Pearson was the actual inventor of the technique. His name is often mentioned in this connection, usually with reference to Pearson (1901). However, in this paper Pearson indicated an approach in which SSQ(Hy) is made as small as possible, which implies maximum relative loss rather than minimum.
This result is not so surprising as it may look at first sight. In somewhat modernized terminology, Pearson describes a model that assumes

Hy = 0, but where actual observations only satisfy Hy = e, with e a vector of (supposedly small) "model-disturbances". It follows that an estimate of y should minimize e'e. The solution for y then becomes the eigenvector with smallest associated eigenvalue.

Two comments. (i) If there is linear dependence among the columns of H, there will be a solution for y with Hy = 0. Such a solution demonstrates that the observations in H agree with some linear "functional relation" between the variables. In this case the smallest eigenvalue of H'H is zero.

(ii) In multiple regression theory one fits a hyperplane for given predictor variables x_j and a single criterion variable y, in such a way that the sum of the squared differences between y and some linear compound Xb is minimized. This corresponds to a minimum sum of squared distances between the observation points and the fitted hyperplane, where these distances are measured in the direction of y.
Pearson's model, on the other hand, for given data matrix X fits a hyperplane, again in such a way that the sum of the squared distances is minimized, but now the distances are measured in the direction orthogonal to the hyperplane.

The invention of linear weights also is often attributed to Hotelling (1933). This is not correct, as shown in the quotations above: Pearson was earlier. In addition, in Hotelling's paper PCA is introduced as a form of factor analysis, with the aim to minimise a loss function $\Sigma SSQ(h_j - za_j)$. Hotelling shows that PCA solves this problem. He adds: "An easily verified property of the method is that the first of our principal components has a greater mean square correlation with the tests than does any other variable" (p. 422). In other words, the first principal component has a maximum for its squared factor loadings. This does imply that the first principal component also gives the best weights (for a matrix with equally normalized columns). But Hotelling did not make this explicit. Later, the idea was explcitly mentioned by Horst (1936), Edgerton and Kolbe (1936), and Wilks (1938).

3.6 Non-linear weighting

Non-linear weighting occurs when it is allowed to transform a vector h_j into a vector $\Psi_j(h_j)$, where Ψ_j may be any function, or may be restricted to some particular class of funtion (e.g., ordinal transformations). The optimal solution should minimize the loss function

$$\sigma(z;a;\Psi) = \frac{1}{m} \sum_{j}^{m} SSQ\ (\ z - a_j\Psi_j(h_j)\ )$$

where we could impose normalization requirement $SSQ(\ \Psi_j(h_j)\ ) = n$, or

$a'a = m$. The loss function will obtain a minimum when the transformations Ψ_j are such that the resulting correlation matrix between the transformed vectors $\Psi_j(h_j)$ has maximized largest eigenvalue.

The general solution of the optimal transformation problem is far from easy (cf. chapter 5). For finite sets of observations and discrete variables, however, and when there is no restriction upon the kind of transformation, the solution becomes relatively simple. The solution then becomes basically that the optimal weighting algorithms described in section 3.3 must be applied to the indicator matrix G in stead of to the data matrix H. Since in the indicator matrix G each category is interpreted as a binary variable, it follows immediately that optimal weighting of these binary variables comes to the same as optimal scaling of categories.
This solution is called HOMALS, abbreviation of HOMogeneity analysis by Alternating Least Squares.

3.7 HOMALS for complete indicator matrix

3.7.1

In section 2.3 it was suggested that quantifications of objects and of categories for a complete indicator matrix G should obey the proportionalities

$$x \div Gy/m \qquad (3.7.1)$$
$$y \div D^{-1}G'x$$

where x is the vector of object scores and y the vector of the quantification of all Σk_j categories. The proportionalities imply

$$x \div GD^{-1}G'x \qquad (3.7.2)$$
$$y \div D^{-1}G'Gy/m = D^{-1}Cy/m$$

for which the algorithms described in section 3.3 give solutions. Also, let $GD^{-\frac{1}{2}} = V\Psi W'$ be the SVD solution, with $V'V=I$, $W'W=I$, and Ψ the diagonal matrix of singular values. It follows immediately that

$$GD^{-1}G'V = V\Psi^2$$
$$D^{-\frac{1}{2}}CD^{-\frac{1}{2}}W = W\Psi^2 \qquad (3.7.3)$$

so that, if we take x proportional to a column of V, and y proportional to the corresponding column of $D^{-\frac{1}{2}}W$, the proportionality requirements of (3.7.1) or (3.7.2) are met. The second proportionality of (3.7.2) now can be re-written as

$$Cy = Dy\psi^2 \qquad (3.7.4)$$

3.7.2

However, in section 3.3 the algorithms were applied to a data matrix H for which it was assumed that columns have zero means. This is not true for G.

We will now show that this does not matter for a complete indicator matrix. For the present section only, let S be the matrix of deviations from column means of G. The argument can perhaps be easier followed with the help of a numerical illustration. To this end we take G of table 2.5 (from which we drop the last column since category w has zero frequency). The result for S is given in table 3.5A, whereas table 3.5B shows the result for S'S. Equation (3.7.4) has a trivial solution y=u, since Cu=Dum. It follows from properties of the generalized eigenvector equation that each other solution y_i must obey $u'Dy_i=0$. This implies that y_i has zero mean. But eq.(3.7.4) also implies $D_u'Cy_i=D_u'Dy_i\psi_i^2$. Here D_u is the matrix exemplified in table 3.6; it has m columns, and in each column unit elements corresponding to the partitioning of G. For a complete indicator matrix, $D_u'C$ has all its rows equal to u'D. It follows that $D_u'Cy_i=0$ and therefore also that $D_u'Dy_i=0$. An algebraic expression for S is $S = G - uu'G/n$. Since $GD_u = uu'$, we have

$$SD_u = GD_u - uu'GD_u/n = uu' - uu' = 0$$

which shows that not only column totals of S are zero, but also row totals, and even row totals of each submatrix S_j, as one may verify in table 3.5.A. The latter result also shows that S has rank Σk_j-m, so that for y_i there also must be Σk_j-m non-trivial solutions.

An algebraic expression for S'S is

$$S'S = G'G - Duu'D/n$$

with superdiagonal matrix

$$D_{S'S} = D - DD_uD_u'D/n$$

Equation (3.7.4), re-formulated for S'S and $D_{S'S}$, becomes

$$S'Sy = D_{S'S}y\psi^2 \qquad (3.7.5)$$

or

$$(G'G - Duu'D/n)y = (D - DD_uD_u'D/n)y\psi^2$$

but since u'Dy=0, and $D_u'Dy=0$, equation (3.7.5) must have the same solutions for y and ψ^2 as equation (3.7.4).

3.7.3

For a HOMALS solution there are basically two options as to normalization (cf. the algorithms of section 3.3).

(i) y is normalized so that y'Dy is some constant. Induced object scores then are defined by x=Gy/m, which makes the object score equal to the average of its category quantifications. In a plot of such results, an object point will be the center of gravity of the points for categories that apply to the object.

(ii) x is normalized to some constant. Induced category quantifications then are defined by $y_j=D_j^{-1}G_j'x$, where each category is quantified as the average of the objects within the category. In a plot, a category point

a	b	c	p	q	r	u	v
.4	-.2	-.2	.2	-.1	-.1	.7	-.7
-.6	.8	-.2	-.8	.9	-.1	-.3	.3
.4	-.2	-.2	-.8	-.1	.9	-.3	.3
.4	-.2	-.2	.2	-.1	-.1	.7	-.7
-.6	.8	-.2	.2	-.1	-.1	-.3	.3
-.6	-.2	.8	.2	-.1	-.1	-.3	.3
.4	-.2	-.2	.2	-.1	-.1	.7	-.7
.4	-.2	-.2	.2	-.1	-.1	-.3	.3
-.6	-.2	.8	.2	-.1	-.1	-.3	.3
.4	-.2	-.2	.2	-.1	-.1	-.3	.3

Table 3.5A Indicator matrix of table 2.5 in deviations from means.

2.4	-1.2	-1.2	.2	-.6	.4	1.2	-1.2
-1.2	1.6	-.4	-.6	.8	-.2	-.6	.6
-1.2	-.4	1.6	.4	-.2	-.2	-.6	.6
.2	-.6	.4	1.6	-.8	-.8	.6	-.6
-.6	.8	-.2	-.8	.9	-.1	-.3	.3
.4	-.2	-.2	-.8	-.1	.9	-.3	.3
1.2	-.6	-.6	.6	-.3	-.3	2.1	-2.1
-1.2	.6	.6	-.6	.3	.3	-2.1	2.1

Table 3.5B Sums of squares and cross products for Table 3.5A

1	0	0
1	0	0
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
0	0	1

Table 3.6 Matrix D_u for numerical example.

will be the center of gravity of the points for objects within the category.

The standard HOMALS program takes normalization (ii), with x'x=n, so that x becomes a "standard score". There are two practical reasons for this choice. The first one is that elements of x now can be interpreted with the help of all the familiar properties of standard scores. The second is that in HOMALS applications it happens very often that n is much larger than Σk_j. For plots, normalization (ii) then gives the nicest arrangement of the picture, with object points equally spread in all directions, and category points indicating the means of subgroups of objects (objects sorted into the respective categories of a variable).

The normalization above implies, in terms of the SVD solution $GD^{-\frac{1}{2}} = V\Psi W'$, that for each of p solutions x_s and y_s (s=1,..,p) the following relations will be valid

$$x_s = v_s\sqrt{n}$$

$$y_s = D^{-\frac{1}{2}}w_s\psi_s\sqrt{n}$$

$$y_s'Dy_s = n\psi_s^2$$

However, in the HOMALS program the eigenvalue matrix is defined as

$$\Phi = \Psi^2/m$$

which implies the generalized eigenvector equation

$$Cy_s = Dy_s\phi_s/m$$

and also implies

$$y_s'Dy_s = nm\phi_s$$

The HOMALS program also defines <u>discrimination measures</u>, one for each variable, as $y_{sj}'D_jy_{sj}/n$ (where y_{sj} is the quantification for G_j in the s^{th} dimension of the solution). Discrimination measures add up to $y_s'Dy_s/n = \psi_s^2 = m\,\phi_s$, so that ϕ_s is the average of the discrimination measures in the s^{th} solution.

It can be shown that these discrimination measures are equal to the squared correlations between x_s and the optimally scaled variables $q_{sj} = G_jy_{sj}$.

Proof: in this proof we omit the index s. The squared correlation between x and q_j equals

$$(r_{x,q_j})^2 = (x'G_jy_j)^2\,(y_j'G_j'G_jy_j)^{-1}\,(x'x)^{-1}$$

Using the equality $y_j = D_j^{-1}G_j'x$, this can be re-written as

$$(r_{x,q_j})^2 = (x'G_jD_j^{-1}G_j'x)^2\,(x'G_jD_j^{-1}G_j'x)^{-1}(n)^{-1}$$

$$= (x'G_jD_j^{-1}G_j'x)/n = y_j'D_jy_j/n$$

In section 3.8 it will be shown that the discrimination measure also has an interpretation as a "squared factor loading".

3.7.4

Numerical example. For the data matrix of table 2.1, indicator matrix in table 2.5 (the example was also used in section 3.7.2) we give HOMALS results, in two versions

(i) with normalization $y'Dy = 1$, $x = Gy/m$;

(ii) with the standard HOMALS normalization $x'x = n$, $y_j = D_j^{-1}G_j'x$.

(i) Normalization $y'Dy = 1$. The generalized eigenvector equation $Cy = Dy\psi^2 = Dy\phi m$ (C and D are given in tables 2.6 and 2.7) has the three largest eigenvalues

$\psi_1^2 = 1.866$, $\phi_1 = .629$, relative loss $1 - \phi_1 = .371$

$\psi_2^2 = 1.277$, $\phi_2 = .426$, relative loss $1 - \phi_2 = .574$

$\psi_3^2 = 1.167$, $\phi_3 = .389$, relative loss $1 - \phi_3 = .611$

Category quantifications y are given in table 3.7. Using the first quantification, the optimally scaled data matrix is given in Q_1 of table 3.8, with columns $q_{1j} = G_j y_{1j}$. Objectscores based on $x = Gy/m$ are given, for the three solutions, in table 3.9.

The equality $y'Dy = 1$ implies that the sum of squares of all elements in Q_1 is unity. Sums of squares for the separate columns of Q_1 are given by $y_{1j}'D_j y_{1j}$, with result .429, .338, and .233, respectively. The larger this value, the better categories of the variable discriminate between objects.

Figure 3.6 gives a plot for the first two solution dimensions (first two columns of table 3.7 give coordinates of category points, first two columns of table 3.9 those of object points). One should verify that each object point is the center of gravity of its categories. In the figure this is illustrated for individuals 6,9 with categories c, p, and v .

Relative loss for the first two dimensions is visible in the figure as the sum of the squared distances between a category point and the object points for objects belonging to the category. In the figure, dotted lines are drawn between individual points and their categories for variable 1. The sum of the squared lengthes of these dotted lines is what variable 1 contributes to relative loss.

Since a variable discriminates better to the extent its category

y_1	y_2	y_3
-.14	-.11	.06
.39	-.19	-.09
.04	.51	-.10
-.06	.05	-.09
.55	-.34	-.09
-.05	-.06	.85
-.23	-.21	-.18
.10	.09	.08

Table 3.7 Solutions for y with y'Dy=1

variables		
1	2	3
-.14	-.06	-.23
.39	.55	.10
-.14	-.05	.10
-.14	-.06	-.23
.39	-.06	.10
.04	-.06	.10
-.14	-.06	-.23
-.14	-.06	.10
.04	-.06	.10
-.14	-.06	.10

Table 3.8 Optimally scaled data matrix Q_1, with normalization y'Dy = 1

x_1	x_2	x_3
-.15	-.09	-.07
.35	-.14	-.04
-.03	-.03	.33
-.15	-.09	-.07
.14	-.01	-.04
.03	22	-.04
-.15	-.09	-.07
-.04	.01	.02
.03	.22	-.04
-.04	.01	.02

Table 3.9 Solution for X, first three dimensions, with normalization y'Dy = 1

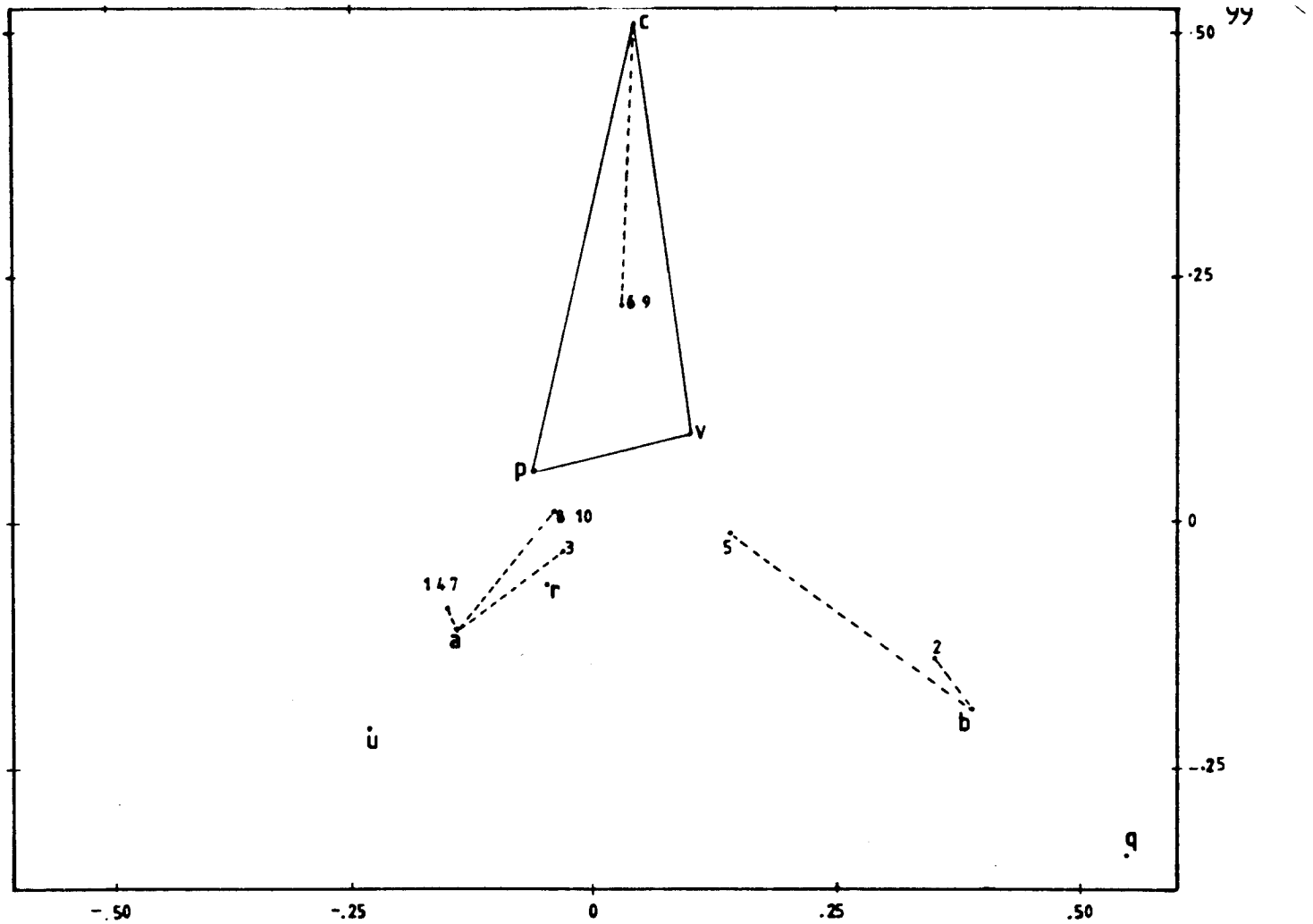


Figure 3.6. Non-standard HOMALS solution. The figure illustrates for objects 6,9 that object points are the center of gravity of category points (triangle). The figure also illustrates what variable 1 contributes to the loss function (sum of squared lengthes of dotted lines).

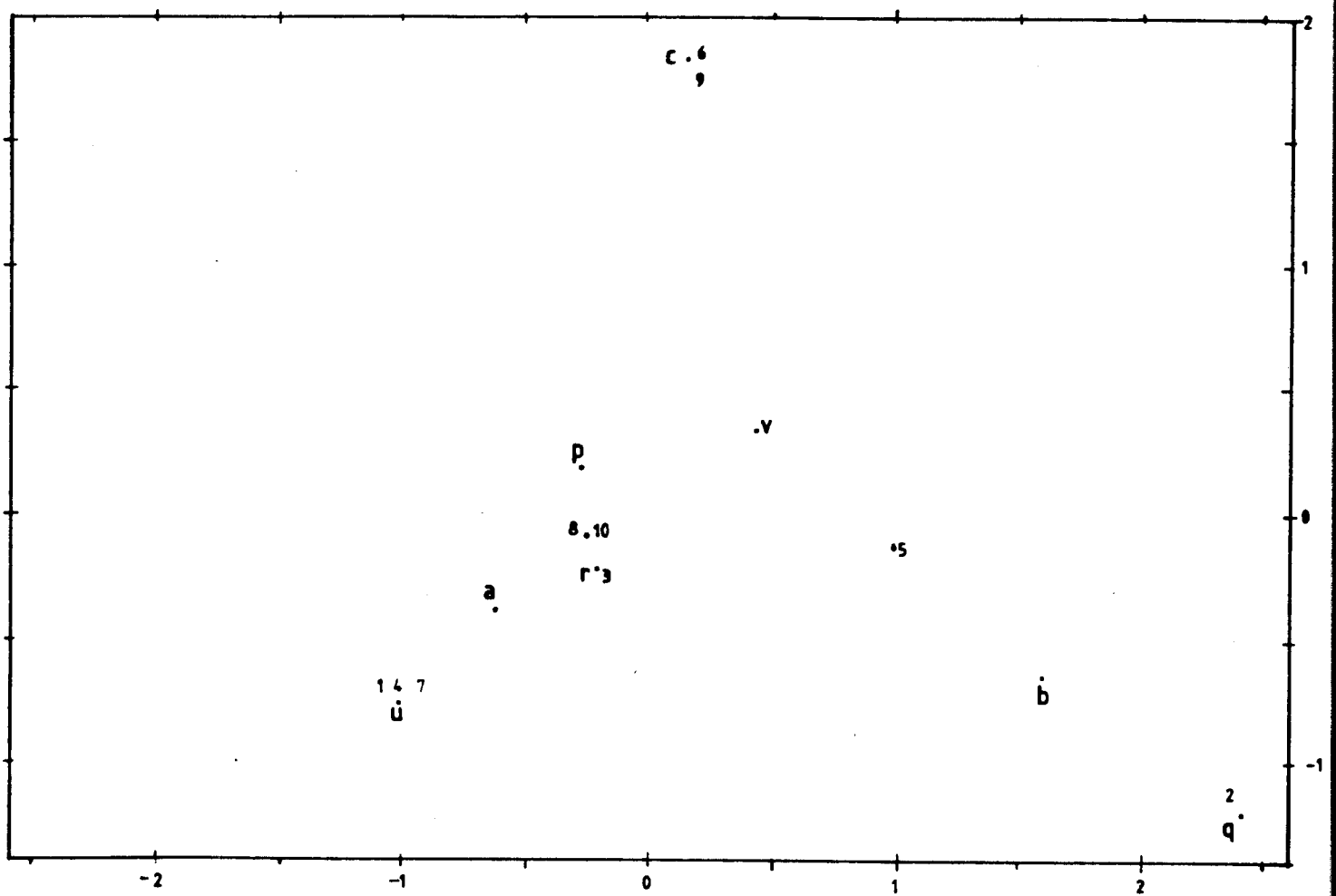


Figure 3.7. Standard HOMALS plot, with category points as the center of gravity of their objects.

x_1	x_2	x_3
-1.01	-.77	-.62
2.40	-1.20	-.32
-.21	-.22	2.89
-1.01	-.77	-.62
.98	-.12	-.32
.18	1.83	-.34
-1.01	-.77	-.62
-.25	.09	.14
.18	1.83	-.34
-.25	.09	.14

Table 3.10 Object scores X, normalization x'x=n

	y_1	y_2	y_3
a	-.62	-.39	.22
b	1.69	-.66	-.32
c	.18	1.83	-.34
p	-.27	.18	-.32
q	2.40	-1.20	-.32
r	-.21	-.22	2.89
u	-1.01	-.77	-.62
v	.43	.33	.27

Table 3.11 HOMALS category quantification.

-.62	-.27	-1.01
1.69	2.40	.43
-.62	-.21	.43
-.62	-.27	-1.01
1.69	-.27	.43
.18	-.27	.43
.62	- [illegible]	-1.01
-.62	-.27	.43
.18	-.27	.43
-.62	-.27	.43

Table 3.12 Optimally scaled data matrix, based on first HOMALS solution

points have larger spread, the sum of the squared distances between category points for some variable, and the origin of the plot, visualizes how well the variable discriminates.

(ii) Normalization $x'x = n$; standard HOMALS solution. The solution for X (first three dimensions) is given in table 3.10, the corresponding solution for $Y = D^{-1}G'X$ in table 3.11. The optimally scaled data matrix, using y_1 for the category quantification, becomes Q_1 of table 3.12. The difference with the previous solution is only in normalization. The correlation matrix for Q_1 of table 3.12 is the same as that for Q_1 in table 3.8.

Figure 3.7 gives a plot for the first two dimensions. Category points now are the center of gravity of object points, as illustrated for the first variable, where category c coincides with objects 6 and 9, category b is midway between 2 and 5, and category a is at the center of gravity of the cloud of object points 1,3,4,7,8,10. The plot also illustrates than category u coincides with objects 1,4,7, category q with object 2, and category r with object 3.

For the first dimension of the solution, discrimination measures $y_{1j}'D_j y_{1j}/n$ are .809, .638, and .439, respectively. They add up to $\psi_1^2 = 1.886$, and their mean is $\phi_1 = .629$. Discrimination measures now are squared correlations between x_1 and the columns of Q_1.

One should note that the HOMALS solution is nested, which means that the first dimension of the solution of a higher dimensional solution is identical the one-dimensional solution; the second dimension of a two-dimensional solution also is the second dimension of a higher-dimensional solution, and so on. Some of the techniques of non-linear MVA to be discussed later (e.g., PRINCALS, chapter 5) do not have this property.

We now summerize some geometrical aspects of the plot in figure 3.7, which are typical for all HOMALS plots.

(1) Categories and objects are represented as points in a joint space.
(2) Category points are the center of gravity of the object points that share the category. For each variable, categories of that variable divide the object points in "subclouds", and the category points are the means of those subclouds.
(3) A variable discriminates better to the extent the category points are further apart. In the figure, this is shown (with dotted lines) for the first variable, where category a brings objects 1,3,4,7,8,10 together, category b objects 2 and 5, and category c objects 6 and 9.

(4) The sum of the squared distances between object points in a subcloud and their mean point (the category point) is related to relative loss (in the sense that if we project distances on a dimension, and take sum of squares over all variables, this sum equals $1 - \psi_s^2/m = 1 - \phi_s$). (This implies that for an "ideal" solution, without any loss at all, all object points must coincide with their category points. This can, of course, only happen if there are at least as many categories as objects.)
(5) The spread of the category points for each variable is an indication of how much the variable contributes to relative loss, for the dimensions displayed in the plot.
(6) The distance between two object points is related to the "similarity" between their response-patterns or profiles. In particular, objects with the same response pattern will be plotted as identical points.
(However, the reverse is not necessarily true. If object points are close together in a plot of the first two HOMALS dimensions, they might be far apart in the third, fourth, etc., dimension.)
For the example, aspect (6) is illustrated by objects (8,10), or (1,4,7). Object 3 is plotted roughly midway between these two groups. However, on the third dimension (table 3.10) , we can see that object 3 is far distant from all other objects.
(7) If a category applies uniquely to only one object, the object point and this category point will coincide. (Example in the illustration: object 2 and category q). The same is true when a category applies uniquely to a grou of objects with identical response pattern (category u for objects 1,4,7).
(8) A category point with low marginal frequency will be plotted farther towards the periphery, whereas a category with high marginal frequency will be plotted nearer to the center of the plot.
(For the example this is illustrated by category q with marginal frequency of 1) versus category p (with marginal frequency of 8. However, category r, also with marginal frequency of 1, is not very far from the center. However, this category point becomes peripheral in the third dimension.)
(9) Objects with response pattern similar to the "average" response pattern will be plotted more towards the center, whereas objects with "unique" response pattern appear in the periphery. (This is mainly a corrollary of (8).) Again, this statement will be true for a plot in all revelant dimensions, and not necessarily for a plot of the first two dimensions only. (For the example, objects 8 and 10, which for each variable have the most frequently selected category, are near the center of the plot, whereas objects 2, or 6, or 9, with unique response patterns are at the periphery. However, object 3, also with unique response pattern, is not far from the cente of the plot; its uniqueness appears in the third dimension.)

3.8 Relations between HOMALS and linear MVA

3.8.1 HOMALS and PCA.

In this section we look only at the first HOMALS dimension. It is related to linear PCA in the following way. Let Q_1 be the optimally scaled data matrix, and let the correlation matrix for Q_1 be R_1. Then the first HOMALS solutions is identical to the solution for the first principal component of R_1.

To illustrate, we take the optimally scaled data matrix Q_1 of table 3.8 or 3.12. The corresponding correlation matrix R_1 is the same for the two different solutions for Q_1. It becomes

$$R_1 = \begin{matrix} 1.000 & .622 & .454 \\ .622 & 1.000 & .224 \\ .454 & .224 & 1.000 \end{matrix}$$

R_1 has eigenvalues $\lambda_{11} = 1.886$, $\lambda_{12} = .789$, and $\lambda_{13} = .325$. The eigenvectors are

$$T_1 = \begin{matrix} .655 & -.104 & .748 \\ .582 & -.563 & -.587 \\ .482 & .820 & -.308 \end{matrix}$$

and a "factor matrix" is

$$F_1 = T_1\Lambda_1^{\frac{1}{2}} = \begin{matrix} .900 & -.092 & .427 \\ .799 & -.500 & -.335 \\ .662 & .729 & -.176 \end{matrix}$$

The first column of the factor matrix gives correlations between the columns of Q_1 and the first "principal component", which is x_1. HOMALS discrimination measures for this first dimension are the squares of these "factor loadings". In addition, the algorithm implies (without proof) that the factor loadings are the positive square roots of the discrimination measures.

Figure 3.8 gives a plot for the first two principal components; the plot is based on F_1.

A proof that a HOMALS solution is a principal component of the optimally scaled data matrix Q is as follows (in this proof we omit the dimension index s). Let D_y be a Σk_j x m matrix with elements y_j in its j^{th} column and the k_j rows corresponding to the partition for the j^{th} variable, so that $D_y u = y$. The optimally scaled data matrix then can be written as $Q = GD_y$. Diagonal elements of $D_y'G'GD_y$ are $v_i = y_i'D_iy_i$. Let v be the m x 1 vector with elements v_i. Let D_v be the m x m diagonal matrix with diagonal elements v_i. Let $\{v\}^{\frac{1}{2}}$ be the m x 1 vector with elements $\sqrt{v_i}$. Then $R = D_v^{-\frac{1}{2}}D_y'G'GD_yD_v^{-\frac{1}{2}}$ is the matrix of correlations for Q. It has an

eigenvector $\{v\}^{\frac{1}{2}}$ with eigenvalue ψ^2 because

$D_v^{-\frac{1}{2}}D_y'G'GD_yD_v^{-\frac{1}{2}}\{v\}^{\frac{1}{2}} = D_v^{-\frac{1}{2}}D_y'G'GD_yu = D_v^{-\frac{1}{2}}D_y'G'Gy = D_v^{-\frac{1}{2}}D_y'G'x\psi^2 =$

$= D_v^{-\frac{1}{2}}D_y'Dy\psi^2 = D_v^{-\frac{1}{2}}v\psi^2 = \{v\}^{\frac{1}{2}}\psi^2$

Since $v_i = y'Dy = n\psi^2$, it follows that $\{v\}^{\frac{1}{2}}/\sqrt{n}$ gives the eigenvector normalized to its eigenvalue, which is the vector of "factor loadings" on the principal component. The discrimination measure v_i/n therefore gives a squared factor loading.

Notes

(1) Let Q_1 be the optimally scaled data matrix on the basis of the first HOMALS solution y_1, and let R_1 be the correlation matrix for Q_1. Then ψ_1^2 always will be the largest eigenvalue of R_1, so that the HOMALS solution corresponds to the first principal component of R_1.

(2) In addition, the HOMALS solution guarantees that Q_1 is scaled in such a way that the first eigenvalue of R_1 is maximized.

(3) For subsequent HOMALS solutions y_s, however, this solution may no longer correspond to the first principal component of R_s, but to a later one.

(4) If we take the "worst" HOMALS solution, then it will correspond to the principal component of the correlation matrix with smallest eigenvalue. In addition, the data matrix is scaled in such a way that this smallest eigenvalue is minimized (and therefore the sum of the other eigenvalues maximized).

3.8.2 Homals as a first step.

The procedure sketched in section 3.8.1 exemplifies the usage of a one-dimensional HOMALS solution only and alone in order to find an optimal quantification of the categories, whereafter the analysis of the data is continued on the basis of the optimally scaled data matrix. Such an approach not only can be useful when variables have no prior quantification (categories are nominal), but sometimes also when there is prior quantification and where one suspects that relations between variables cannot be optimally described in terms of linear relations. E.g., if one variable is 'age' of individuals, and one suspects that the other variables have a curvilinear relation with age, then the HOMALS quantification might confirm that suspicion. Also, the HOMALS quantification might in some cases suggest a well-defined transformation of the prior quantification of a variable, in the sense that the HOMALS quantification might fit with a logarithmic transformation, etc.

However, HOMALS as a first step need not necessarily be followed by PCA. For example, suppose the data matrix contains two different sets of variables, such

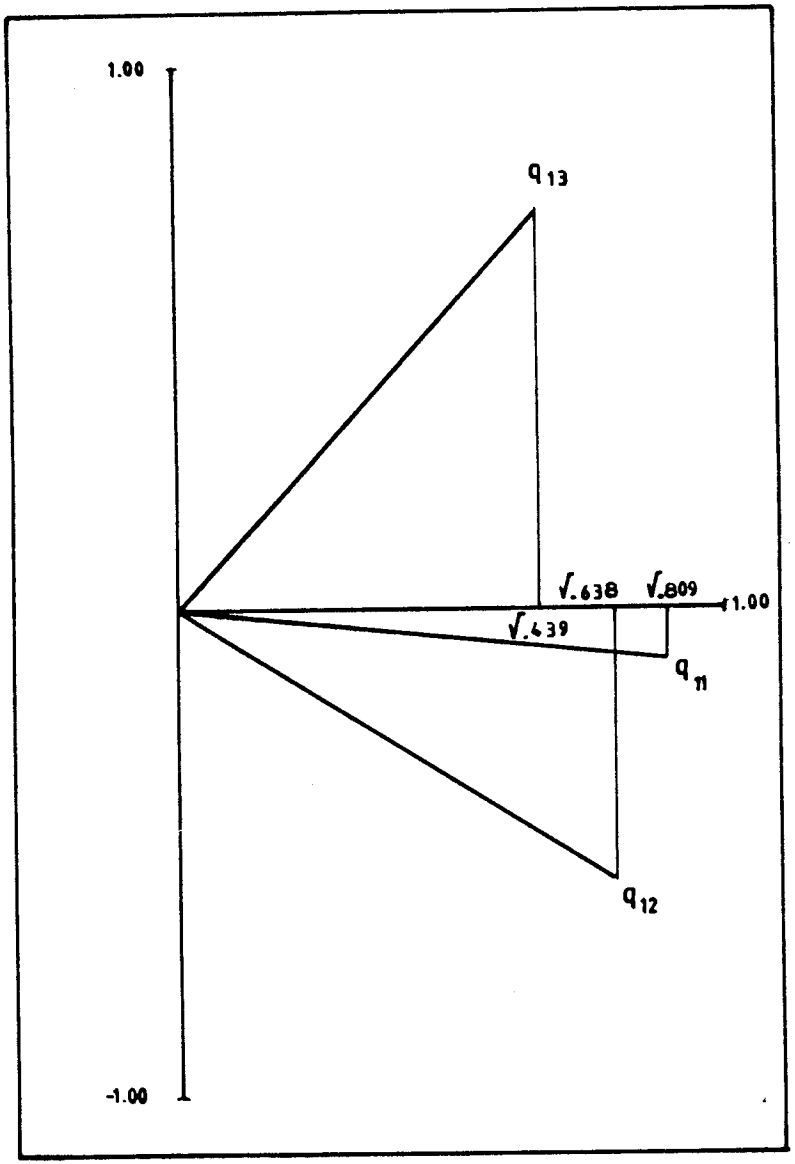


Figure 3.8. PCA solution in two dimensions for the correlation matrix of the first optimally quantified data matrix Q_1. The plot assumes that vectors q_{1j} have unit length. Their loadings on the first component are the square root of the discrimination measure.

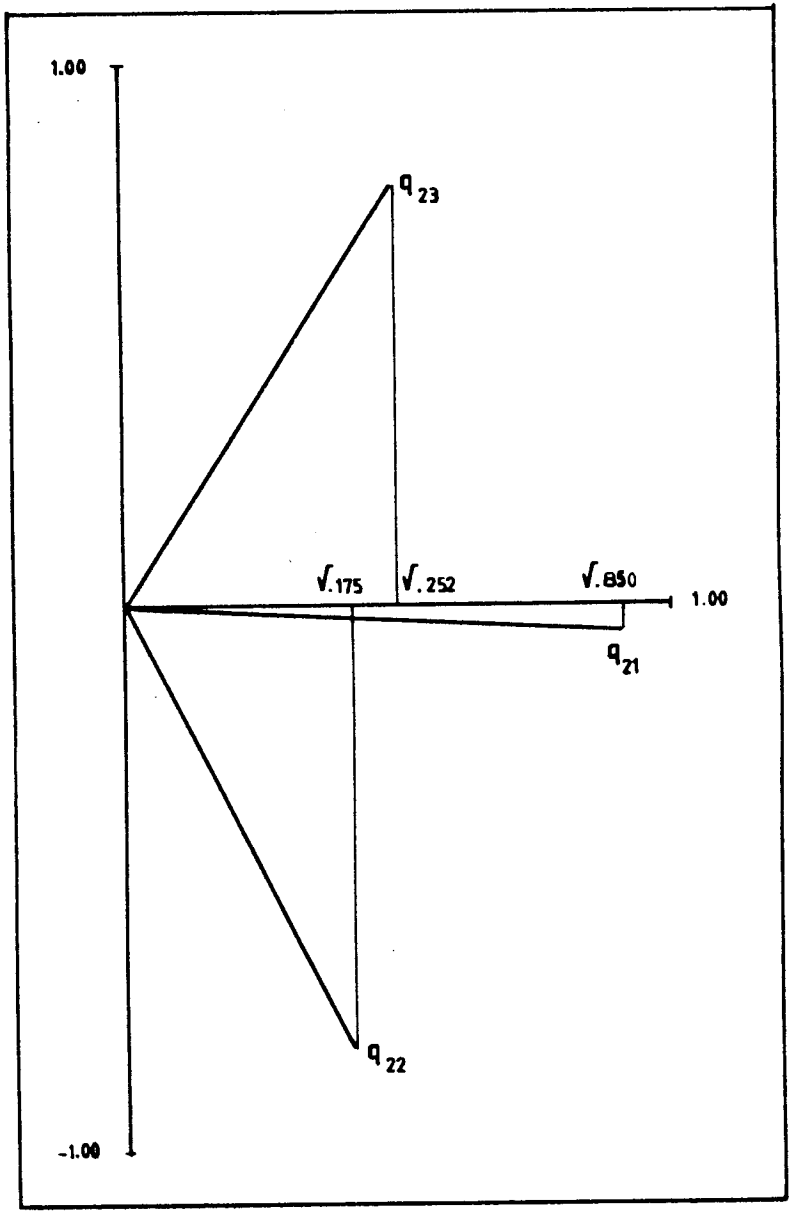


Figure 3.9. PCA solution for correlation matrix of Q_2. Compare with figure 3.8. This second solution is less good because vectors q_{2j} are not so close to their first principal component.

as "current" variables versus "background" variables. The approach of linear MVA would probably be some sort of canonical analysis. HOMALS as a first step would mean that this canonical analysis is applied to the optimally scaled data matrix. The example allows for different options: HOMALS as a first step could be applied to the data matrix as a whole, or to the two data matrices for the two sets of variables separately.

3.8.3 Multiple HOMALS and PCA.

In section 3.8.1 it was mentioned that the second HOMALS quantification creates a second bundle of vectors q_{2j} around their sumvector Gy_2. Some of the implication of PCA on the optimally scaled data matrix Q_2 are shown in the following example, based on the second HOMALS solution given in section 3.7.4 (ii). Results are given in table 3.12; this table shows the correlation matrix R_2, its eigenvector matrix T_2, and its PCA factormatrix $F_2 = T_2\Lambda_2^{\frac{1}{2}}$, where Λ_2 is the matrix of eigenvalues of R_2. For the example, these eigenvalues are $\lambda_{21} = 1.279$, $\lambda_{22} = 1.277$, and $\lambda_{23} = .445$. Note that the second eigenvalue is equal to the second HOMALS eigenvalue $\psi_2^2 = 1.277$; the second HOMALS solution now must be identified with the second PCA component of R_2. E.g., the HOMALS discrimination measures (they are .850, .175, and .252, respectively) are the squares of the second column of F_2.

The first HOMALS solution also can be described in the following way. The indicator matrix as it were "expands", or "blows up" the data matrix in the sense that a column h_j of the data matrix becomes a set of "category variables" in G_j. The HOMALS solution is a "meet" solution for these m sets of category variables, which means that HOMALS solves for m linear compounds $q_j = G_j y_j$ in such a way that these m vectors q_j are as much as possible "close" to their sumvector Gy, where Gy is proportional to x. Geometrically, this means that the vectors q_j form a sort of "bundle" in m-dimensional space around their sumvector. The HOMALS solution minimizes the sum of the squared cosines of the angles between q_j and Gy (where Gy and x coincide as to their direction). A different way of formulating the HOMALS criterion geometrically is: if we give all q_j unit length, HOMALS maximizes the sum of the squared projections of such unit-length vectos on Gy (or on x).

PCA on the optimally scaled data matrix Q solves a "join" problem. It means that in Q the indicator matrix is again "compressed" into a set of m variables q_j, and PCA investigates how these m vectors are located in their m-dimensional space, and where the PCA solution wants to describe the bundle of vectors q_j as good as possible with as few dimensions as

possible. HOMALS already gives a one-dimensional approximation of the bundle, PCA gives additional dimensions.
The second HOMALS solution "compresses" the indicator matrix in a different way, so that we obtain a different set of vectors q_{2j}. There is no simple relation between this second bundle and the first.
Figure 3.9 gives a graph for the first two dimensions of the PCA solution of R_2.
One should realize, however, that if we continue in this way for successive HOMALS solutions, things become redundant. In general, there are $\Sigma k_j - m$ (cf. section 3.7.3) possible HOMALS solutions. Each of them creates m vectors q_{sj} $(s=1,..,(\Sigma k_j - m))$, so that a "complete" HOMALS solution, with a PCA continuation for each dimension, creates a situation with $m(\Sigma k_j - m)$ dimensions (a clear example of "data production" in stead of "data reduction").

3.9 HOMALS loss functions and relations with chi-square

3.9.1 HOMALS loss functions

(i) On the basis of the SVD solution, a HOMALS loss function for a solution in p dimensions, can be written as

$$\sigma_1^{(p)} = \frac{1}{m}SSQ(GD^{-\frac{1}{2}} - \Sigma v_s \psi_s w_s')$$
$$= \frac{1}{m}(\Sigma k_j - \Sigma \psi_s^2) = \Sigma k_j/m - \Sigma \psi_2^2/m = \Sigma k_j/m - \Sigma \phi_s$$

with $s=1,..,p$ and $j=1,..,m$. If we set p equal to the number of all HOMALS solutions (of which there are $\Sigma k_j - m$; see section 3.7.3) the loss becomes

$$\sigma_1^{(\Sigma k_j - m)} = \Sigma k_j/m - (\Sigma k_j - m)/m = 1$$

That this final loss does not become equal to zero, is explained by the "trivial" non-HOMALS solution $v=u/\sqrt{n}$, $w=u/\sqrt{n}$, $\phi=1$. If we include this solution in the loss function, the "complete" solution will have zero loss.
(ii) It is perhaps more natural to define another HOMALS loss function in terms of the 'homogeneity' approach of section 3.2. This approach implies the loss function

$$\sigma_2^{(p)} = \frac{1}{mn}\Sigma\Sigma SSQ(x_s - G_j y_{sj}) = \frac{1}{mn}\Sigma\Sigma(n - y_{sj}'D_j y_{sj}) = \frac{1}{mn}\Sigma(mn - y_s'Dy_s) =$$
$$= \frac{1}{mn}\Sigma(mn - n\psi_s^2) = \frac{1}{mn}(mnp - n\Sigma\psi_s^2) = p - \Sigma\psi_s^2/m = p - \Sigma\phi_s$$

Taking all possible HOMALS solutions now gives an ultimate loss of

$$\sigma_2^{(\Sigma k_j - m)} = \Sigma k_j - m - \Sigma\phi_s = (\Sigma k_j - m) - (\Sigma k_j - m)/m =$$
$$= (\Sigma k_j - m)(1 - \frac{1}{m})$$

An alternative notation for the loss function is

$$\sigma_2^{(p)} = \frac{1}{mn}\Sigma SSQ(X - G_j Y_j)$$

where X is the n x p matrix of different solutions for object scores, and Y_j the k_j x p matrix of different quantifications of the categories in G_j. The two loss functions are related by

$$\sigma_1 - \sigma_2 = (p - \Sigma\phi_s) - (\Sigma k_j/m - \Sigma\phi_s) = p - \Sigma k_j/m$$

so that they differ only in a constant and therefore have simultaneous minimum.

σ_2 implies that loss per dimension equals $1 - \phi_s = W/T$. (It also implies that average loss, averaged over all possible Σk_j-m solutions, equals $(1 - \frac{1}{m})$. This means that loss per dimension becomes greater that average loss once $\phi_s < \frac{1}{m}$, or $\psi_s^2 < 1$.)

3.9.2 Relation with χ^2

Equation (3.7.4) has the implication that

$$D^{-\frac{1}{2}}(C - Duu'D/n)D^{-\frac{1}{2}} = W\Psi^2W' \qquad (3.9.1)$$

where the "trivial" solution $w=u/\sqrt{n}$ is not included. In section 3.7.2 it was shown that the matrix in 3.9.1 has rank Σk_j-m. The matrix at the left in equation (3.9.1) has trace Σk_j-m, and it is directly seen that the matrix at the right has trace $\Sigma\psi_s^2$. It follows that $\Sigma\psi_s^2 = \Sigma k_j$-m.

An off-diagonal submatrix of (3.9.1) has the form

$$D_i^{-\frac{1}{2}}(C_{ik} - D_i uu' D_k)D_k^{-\frac{1}{2}} \qquad (3.9.2)$$

It has the following interpretation. $D_i uu' D_k/n$ has elements equal to the "expected values" based on the univariate marginals for variables i and k. Since C_{ik} is a matrix of bivariate marginals (section 2.2) it follows that C_{ik}-$D_i uu' D_k/n$ is a matrix of differences between observed and expected values. The matrix of (3.9.2), after multiplication with $\sqrt{n}$, then has elements equal to the discrepancy between observed and expected value, divided by the square root of the expected value. The sum of the squares of those elements equals χ_{ik}^2.

For diagonal submatrices, the expression in (3.9.2) with k=i, becomes $D_i^{-\frac{1}{2}}(D_i - D_i uu' D_i/n)D_i^{-\frac{1}{2}} = I - D_i^{\frac{1}{2}}uu'D_i^{\frac{1}{2}}/n$. This is an idempotent matrix, so that its trace is equal to the sum of its squared elements. It trace equals k_i-1.

Combining results for diagonal and off-diagonal submatrices, we obtain that the sum of the squared elements of the matrix in equation (3.9.1) must be equal to

$$\sum_i (k_i - 1) + \sum_{ik}\sum \chi_{ik}^2/n$$

This sum of squares must be equal to the sum of squares of the elements of the matrix at the right in equation (3.9.1). The latter sum of squares is the trace of $W\Psi^2W'W\Psi^2W' = W\Psi^4W'$, with trace $\Sigma\psi_s^4$. It then follows that

$$\Sigma\Sigma_{ik}\chi_{ik}^2 = n\Sigma\psi_s^4 - n\Sigma k_j + mn = nm^2\Sigma\phi_s^2 - n\Sigma k_j + mn \qquad (3.9.3)$$

But we also has the result $\Sigma k_j - m = \Sigma\psi_s^2 = m\Sigma\phi_s$. Equation (3.9.3) therefore can be re-written as

$$\Sigma\Sigma_{ik}\chi_{ik}^2 = n(m^2\Sigma\phi_s^2 - m\Sigma\phi_s) = n\Sigma(m\phi_s - 1)^2$$

Equivalently,

$$\Sigma_{i<k}\chi_{ik}^2 = \tfrac{1}{2}n\Sigma(m\phi_s - 1)^2 \quad (s=1,..,(\Sigma k_j - m)) \qquad (3.9.4)$$

Under the assumption that all variables are independently distributed the last expression converges to χ^2 with $\frac{1}{2}\{(\Sigma k_j - m)^2 - \Sigma(k_j-1)^2\}$ degrees of freedom.

3.10 A numerical example

The following "real-life" example is taken from Hartigan (1975, p.228).A number of bolts, nails, screws, and tacks are classified according to a number of criteria. Table 3.13 gives the basic data matrix, and explains symbols. HOMALS was performed in p=2 dimensions. Results are given in table 3.14A for objectscores X, in table 3.14B for category quantification Y, whereas table 3.14C gives the discrimination measures. Figure 3.10 gives a plot, where we have used the labels S N B T for Screws, Nails, Bolts, and Tacks. We see from the plot that the first dimension discriminates S and B (with thread) from N and T (without thread). It also separates B (flat bottom) from S T N (sharp bottom). It does not separate N from T. The second dimension seems to separate the long SCREW1 and NAIL6 from the rest.

These results are confirmed in figure 3.11 in which discrimination measures are plotted. This plot, too, shows that the first dimension is related to variables 1 (THREAD) and 4 (BOTTOM). Closest to the second dimension is variable 5 (LENGTH) Variables 2 and 3 are in between, whereas variable 6 discriminates very poorly in the first two dimensions.[1]) In fact, it is intuitively obvious that BRASS and LENGTH

[1]) If we had plotted square roots of discrimination measures, the resulting plot would have obtained the same meaning as a plot for factor loadings in PCA. There is one difference: in a matrix of factor loadings, the sum of the squared loadings in the same row cannot be larger than 1. In HOMALS, discrimination measures for the same variable have sum not larger than k_j-1. The reason is that the discrimination measure is derived from G_j, and if we take deviations from means, G_j has rank k_j-1 (cf. section 3.3). In geometrical terms, for the HOMALS solution, variables are vectors with squared length k_j-1.
This result implies that discrimination measures are not fully comparable for variables with different number of categories. One could, of course, divide by k_j-1, but then the interpretation in terms of a squared correlation is lost.

Variables	Categories & codes
Thread	Yes = Y No = N.
Head	Flat = F Cup = U Cone = O Round = R Cylinder = Y.
Head indentation	None = N Star = T Slit = L.
Bottom	Sharp = S Flat = F.
Length	(in half inches).
Brass	Yes = Y No = N.

Object	1	2	3	4	5	6
TACK	N	F	N	S	1	N
NAIL1	N	F	N	S	4	N
NAIL2	N	F	N	S	2	N
NAIL3	N	F	N	S	2	N
NAIL4	N	F	N	S	2	N
NAIL5	N	F	N	S	2	N
NAIL6	N	U	N	S	5	N
NAIL7	N	U	N	S	3	N
NAIL8	N	U	N	S	3	N
SCREW1	Y	O	T	S	5	N
SCREW2	Y	R	L	S	4	N
SCREW3	Y	Y	L	S	4	N
SCREW4	Y	R	L	S	2	N
SCREW5	Y	Y	L	S	2	N
BOLT1	Y	R	L	F	4	N
BOLT2	Y	O	L	F	1	N
BOLT3	Y	Y	L	F	1	N
BOLT4	Y	Y	L	F	1	N
BOLT5	Y	Y	L	F	1	N
BOLT6	Y	Y	L	F	1	N
TACK1	N	F	N	S	1	Y
TACK2	N	F	N	S	1	Y
NAILB	N	F	N	S	1	Y
SCREWB	Y	O	L	S	1	Y

Table 3.13 Hartigan's hardware

object	dim1	dim2
TACK	0.75	0.46
NAIL1	0.68	0.47
NAIL2	0.96	0.52
NAIL3	0.96	0.52
NAIL4	0.96	0.52
NAIL5	0.96	0.52
NAIL6	1.00	-1.69
NAIL7	1.25	-0.74
NAIL8	1.25	-0.74
SCREW1	-0.38	-3.96
SCREW2	-0.85	0.23
SCREW3	-0.91	0.26
SCREW4	-0.57	0.28
SCREW5	-0.63	0.31
BOLT1	-1.31	0.38
BOLT2	-1.18	-0.51
BOLT3	-1.30	0.40
BOLT4	-1.30	0.40
BOLT5	-1.30	0.40
BOLT6	-1.30	0.40
TACK1	0.93	0.67
TACK2	0.93	0.67
NAILB	0.93	0.67
SCREWB	-0.54	-0.44

Table 3.14A Hartigan's hardware object scores

category	dim1	dim2
1Y	-0.96	-0.15
1N	0.96	0.15
2F	0.90	0.56
2U	1.16	-1.06
2O	-0.70	-1.64
2R	-0.91	0.30
2Y	-1.12	0.36
3N	0.96	0.15
3T	-0.38	-3.96
3L	-1.02	0.19
4S	0.43	-0.08
4F	-1.28	0.25
5/1	-0.34	0.31
5/2	0.44	0.44
5/3	1.25	-0.74
5/4	-0.60	0.24
5/5	0.31	-2.82
6Y	0.56	0.39
6N	-0.11	-0.08

Table 3.14B Hartigan's hardware category quantifications

Variables	dim1	dim2
Thread	0.930	0.024
Head	0.951	0.635
Head indentation	0.945	0.681
Bottom	0.546	0.020
Length	0.292	0.819
Brass	0.064	0.031
Eigenvalues	0.621	0.368

Table 3.14C Hartigan's hardware discrimination measures

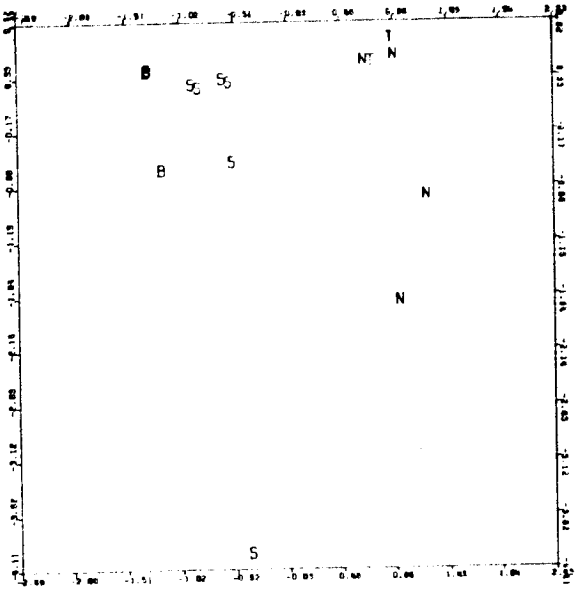


Figure 3.10. Hartigan's hardware object scores labeled by type

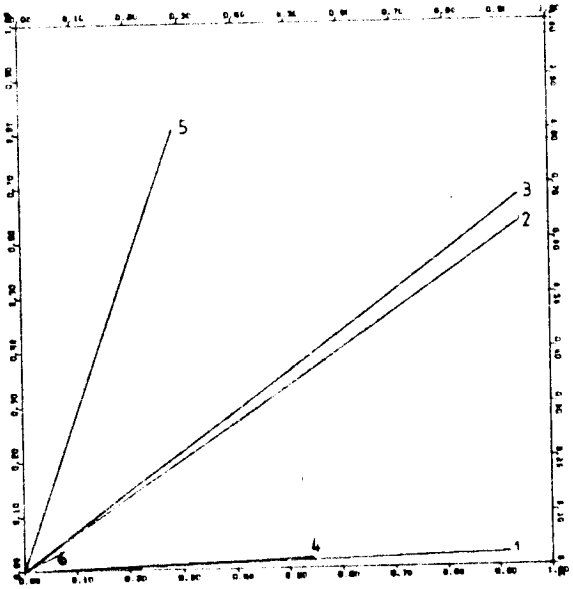


Figure 3.11. Hartigan's hardware discrimination measures

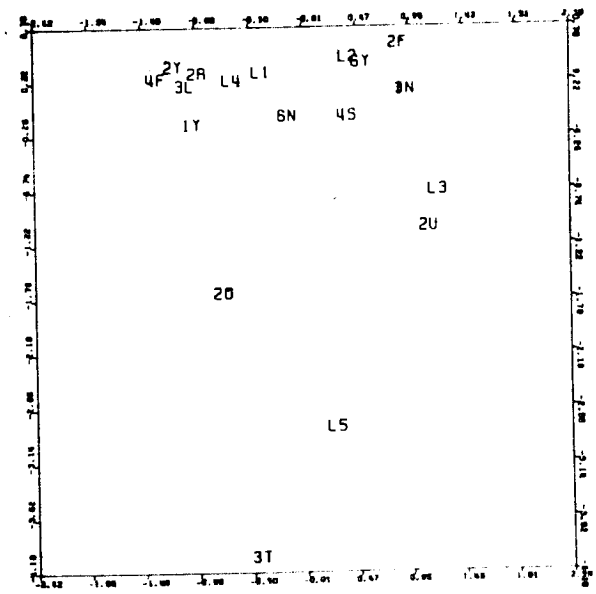


Figure 3.12. Hartigan's hardware category quantifications

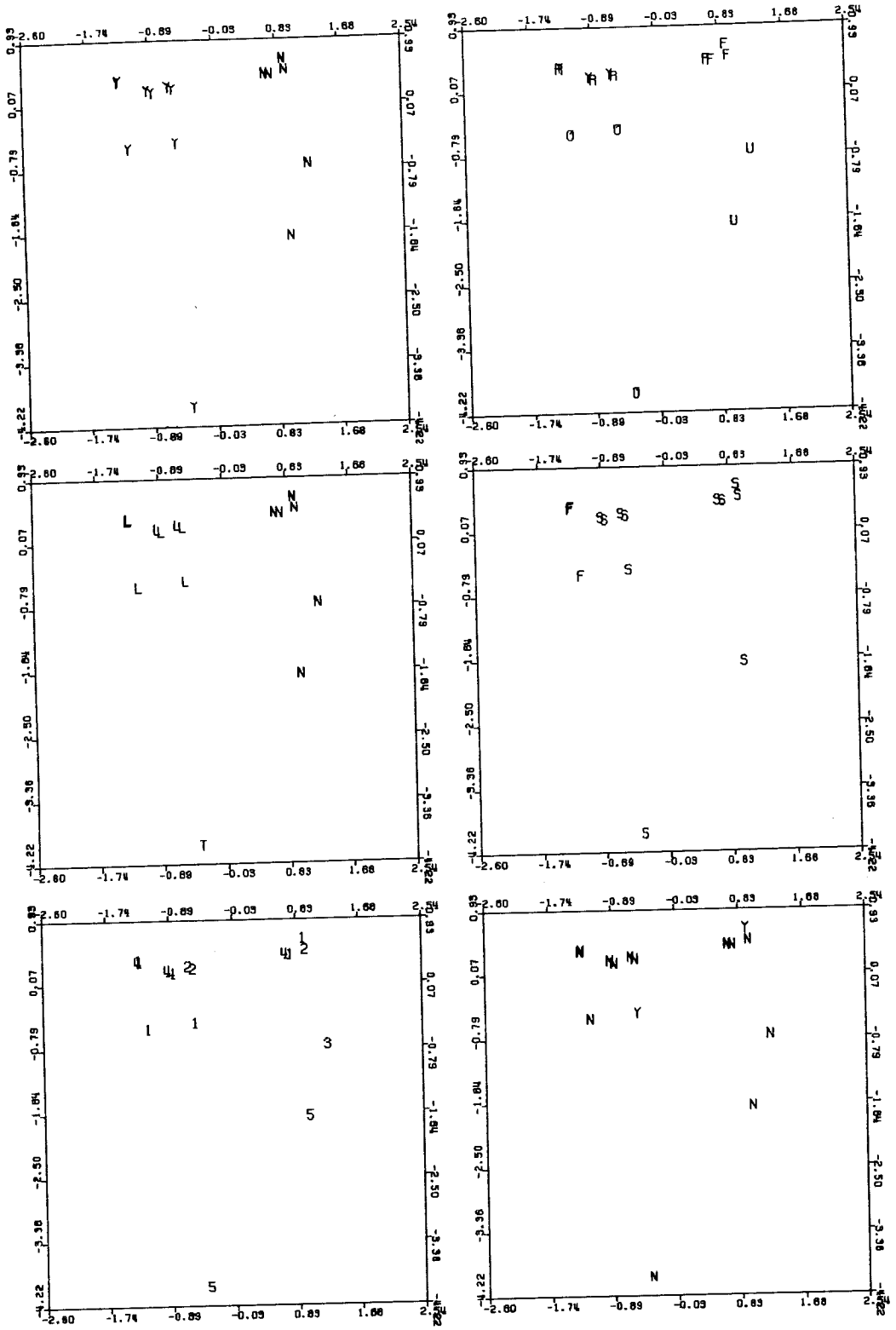


Figure 3.13. Hartigan's hardware
object scores labeled by variables 1 2
3 4
5 6

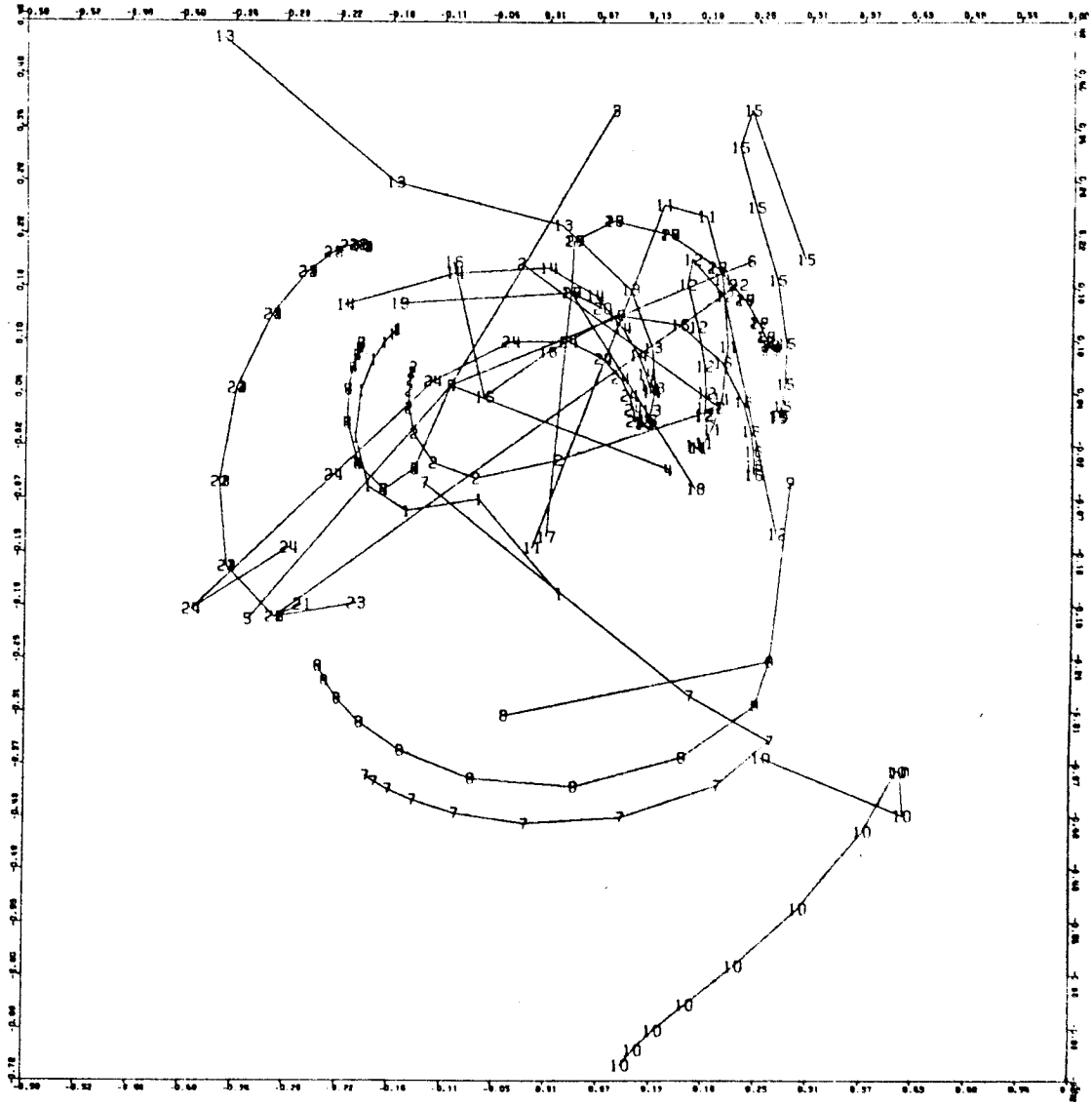


Figure 3.14. HOMALS iteration history for Hartigan's hardware example.

cannot discriminate very well, because we can make screws, nails, bolts and tacks of any length, and either in brass or not. If such variables discriminate, it must be because of peculiarities in the sample. Figure 3.12 plots category quantifications; the points in figure 3.12 are centers of gravity of selected subsets of points in figure 3.10.

A more precise and detailed analysis is possible by studying the six plots in figure 3.13. Here the object scores are plotted again, but now labeled for each variable separately, and using the labels of table 3.13. From these plots we see that variables 5 and 6 have categories that cannot be separated very well (at least in the first two dimensions). For the other variables the objects with the same label form fairly homogeneous clumps, with some exceptions (such as that for variable 2 the cylinders cannot be separated from the rounds).
There is another important point. For variable 1 the categories Y and N are not very homogeneous, in the sense that objects in the same category are not necessarily very close together. But on the other hand, it is clear that the first dimension separates Y and N perfectly. We consider this result satisfactory, although the HOMALS loss might not be small in situations like this. Or, to put it differently, in situations like this we essentially want categories to be well separated, but we do not necessarily want that objects in a category form compact clumps. The HOMALS definition of what is the "best" or "optimal" solution is stronger than the definition we really want to use. Another points illustrated by this example is that the interesting separation is almost completely along the first dimension. The second dimension capitalizes on the fact that there is one special screw in the sample: SCREW1 (too long, cone head, star identification). If we delete SCREW1 from the data, the HOMALS solution changes considerably with respect to the second dimension. The second dimension then contrasts the U-heads (variable 2) with the four brasses (variable 6). The link between these two variables is LENGTH (variable 5), because the brasses all have length 1, and the U-heads are the only objects with length 3 or 5. It thus turns out that the second dimension again capitalizes on the particular choice of objects in the sample.(Numerical results of this second HOMALS are not presented.)

All tables and plots shown so far can be produced by the standard HOMALS program. This is not true for the plot in figure 3.14. and which illustrates the history of the HOMALS algorithm. The plot maps the 24 objects in two dimensions on the basis of results for X in successive iteration steps. Lines have been drawn that connect successive positions of the same object, showing, e.g., that object 10 gradually moves towards its excentric position (SCREW1), whereas object 13 (SCREW 4) moves away from its excentric position towards the

center. Obviously, the initial positions (first iteration) very much depend on the arbitrary initial identifications of X or Y. Nevertheless, the figure shows rather nicely how HOMALS starts with big moves, which gradually become smaller, until finally they become infinitely small.

3.11 HOMALS with incomplete indicator matrix

3.11.1

The discussion in section 2.4.2 suggests for the incomplete indicator matrix a quantification that obeys the proportionalities

$$y \div D^{-1}G'x$$

$$x \div M_*^{-1}Gy$$

Such a solution is consistent with the SVD solution

$$M_*^{-\frac{1}{2}}GD^{-\frac{1}{2}} = V\Psi W'$$

with

$$x = M_*^{-\frac{1}{2}}v\sqrt{mn}$$

$$y = D^{-\frac{1}{2}}w\psi\sqrt{mn}$$

so that

$$x'M_*x = mn$$

$$y'Dy = mn\psi^2.$$

It remains true that $u'Dy = 0$ (but not that $u'D_jy_j = 0$), and it remains true that $u'M_*x = 0$ (in the complete indicator matrix case, $M_* = m.I$, so that there we have $u'x = 0$).

A consequence is that the HOMALS solution no longer can be interpreted in such a simple way as a principal components analysis on the optimally scaled data matrix. Since, with an incomplete indicator matrix G_j some rows of G_j are zero rows, it follows that the optimally scaled data matrix on its correspond places obtains zero entries. Such zero's are not the quantification of a category that was missing, although one also could reason that they are, and give missing categories the average object score, which is zero. Apart fr that, columns of the optimally scaled data matrix will not add up to zero. This among other things the consequence, already mentioned, that PCA on the optimally scaled matrix produces a different result. It no longer has the first HOMALS solution as a principal component. (Assuming, at least, that PCA is appl on the transformed data matrix after columns are re-defined as deviations fro means.)

Another consequence is that the discrimination measure $y_j'D_jy_j/n$ no longer can be intrepreted as the variance of the elements in a column of the optimally scaled data matrix, since such a column G_jy_j does not have zero mean. Also,

the discrimination measure no longer equals the squared correlation between x and $G_j y_j$, for the same reason. In addition, it now may happen that the discrimination measure becomes larger than unity

3.11.2

One special case of incomplete indicator matrix is that of missing data, as discussed in section 2.5, for the option "missing data deleted". In the present section we will compare results for the numerical example of section 2.5 for the three different options mentioned there.

(i) Missing values deleted. This is the example with incomplete indicator matrix. Results are given in table 3.15, for p=2 dimensions, with figure 3.15 as the plot.

(ii) Missing values single category. Here the indicator matrix is completed. Results are given in table 3.16, with figure 3.16 as the plot.

(iii) Missing values multiple category. Here, too, the indicator matrix is completed. Results are in table 3.17, with figure 3.17 as a plot.

Comparing these results, one may note that in solution (ii) objects 1 and 3 are brought closer together (because they "share" the single category of missing data on variable 1). In solution (iii) objects 1 and 3 are again farther apart. On the whole, solution (iii) seems to be largely affected by how missing data are handled, which, of course, in this mini-eample with 3 out of 30 data missing, is not too surprising. For actual data matrices, if the number of missing data is relatively small, and if missing data are randomly divided over objects and categories, differences between the three options will be small (and the interpretation of discrimination measures will be almost the same as for the complete case).

Options (ii) and (iii) do produce a quantification of the missing data. In option (ii) it is the average object score for objects with missing data on a variable. In option (iii) each missing data obtains the quantification of the object with missing data. In option (i) missing data are not really quantified (if we think of them as having quantification equal to 0, this becomes inconsistent with the idea that category quantification is the average of the objects within that category.

Suppose an object has missing data on all variables. Option (i) gives this object quantification 0 on all dimensions. Option (ii) will quantify this object as the average of all other objects with missing data. Option (iii) will quantify this object on a separate dimension of its own.

3.11.3

As an additional example of results for incomplete versus completed indicator matrix, figures 3.18 and 3.19 give results for the example

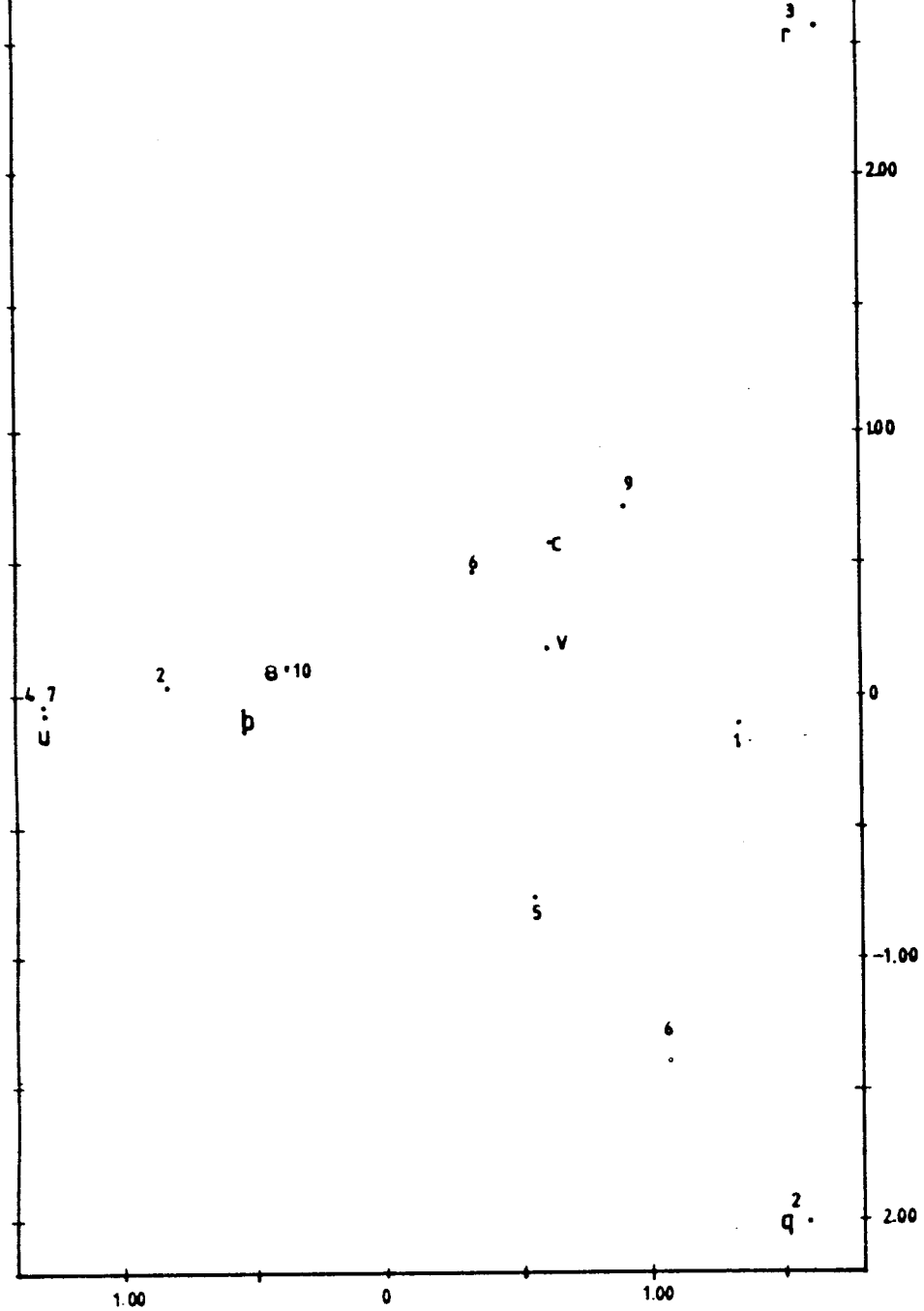


Figure 3.15. HOMALS solution with option 'missing data deleted'.

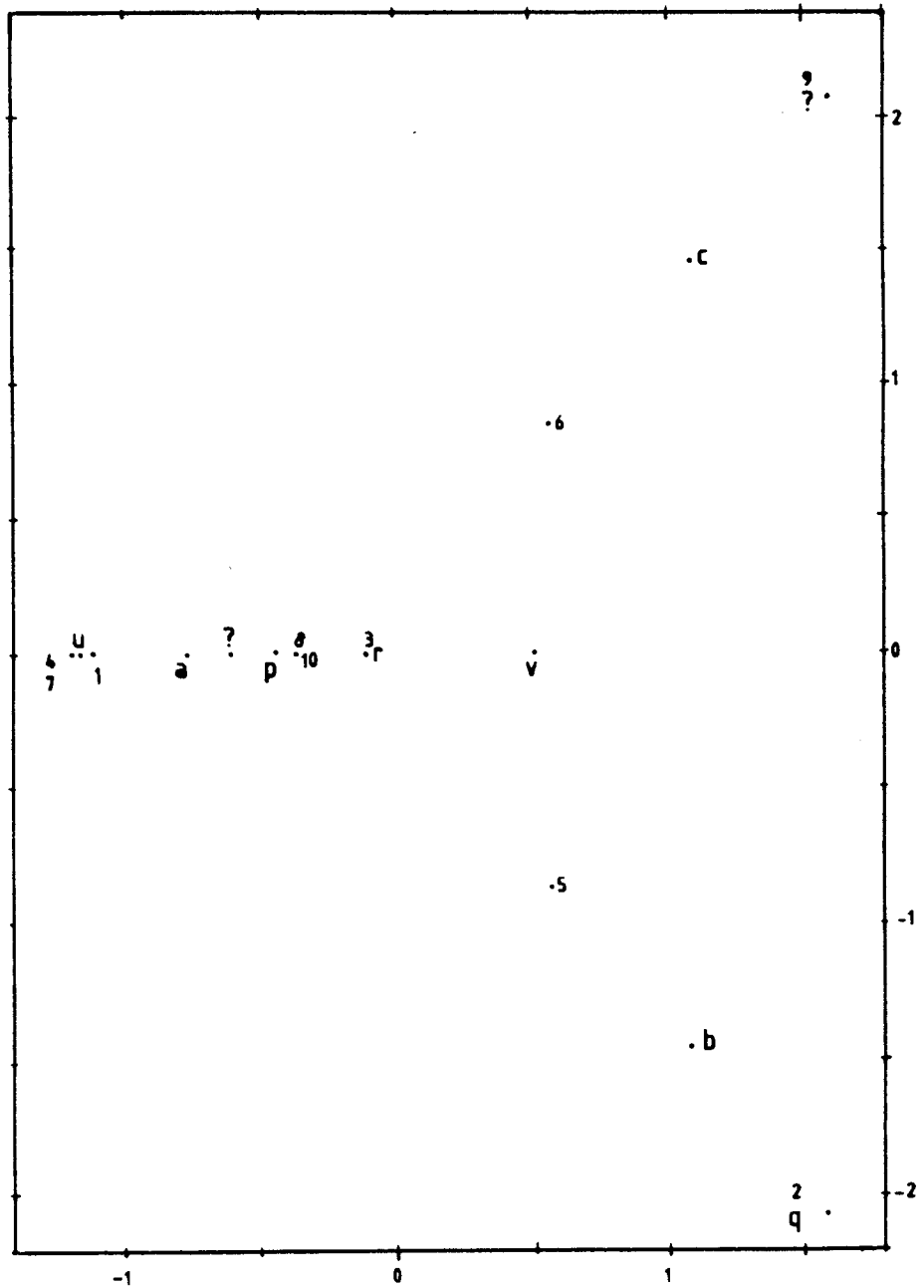


Figure 3.16. HOMALS solution with option 'missing data single category'.

Eigenvalues Φ: .6880 .5343

-1.33	-.10
1.59	-2.01
1.63	2.67
-1.29	-.04
.56	-.78
.34	.46
-1.29	-.04
-.36	.11
.90	.73
-.36	.11

Object score X

a	-.82	.03
b	1.07	-1.39
c	.62	.59
p	-.53	-.04
q	1.59	-2.01
r	1.63	2.67
u	-1.30	-.06
v	.61	.18

Category quantifications Y

Table 3.15 Results for numerical example with option (i) "missing values deleted"

Eigenvalues Φ: .6648 .5690

-1.11	.00
1.59	-2.07
-.11	.00
-1.19	-.00
.57	-.86
.57	.86
-1.19	-.00
-.36	-.00
1.58	2.07
-.36	-.00

Object scores X

a	-.77	-.00
b	1.08	-1.46
c	1.08	1.46
?	-.61	.00
p	-.44	-.00
q	1.59	-2.07
r	-.11	.00
?	1.58	2.07
u	-1.16	-.00
v	.50	.00

Category quantifications Y

Table 3.16 Results for numerical example with option (ii) "missing data single category"

Eigenvalues Φ : .7473 .6308

-1.36	-.70
.82	1.36
2.06	-2.16
-1.06	-.47
.22	.60
.22	.60
-1.06	-.47
-.32	-.06
.82	1.36
-.32	-.06

Object scores X

a	-.69	-.26
b	.52	.98
c	.52	.98
?1	-1.36	-.70
?3	2.06	-2.16
p	-.53	-.08
q	.82	1.36
r	2.06	-2.16
?9	.82	1.36
u	-1.16	-.54
v	.50	.23

Category quantifications Y

Table 3.17 Results for numerical example with option (iii) "missing data multiple categories"

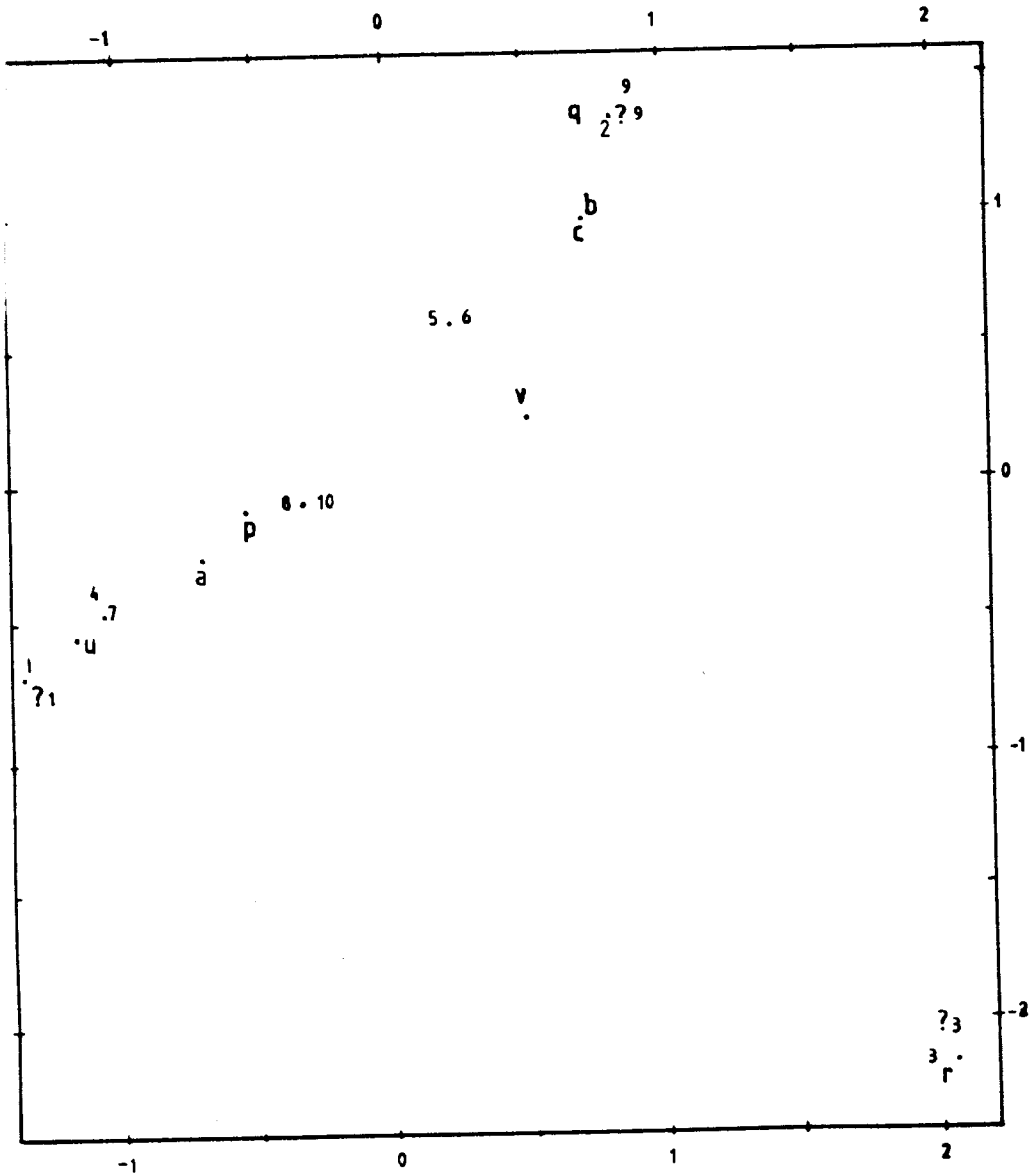


Figure 3.17. HOMALS solution with option 'missing data multiple categories'.

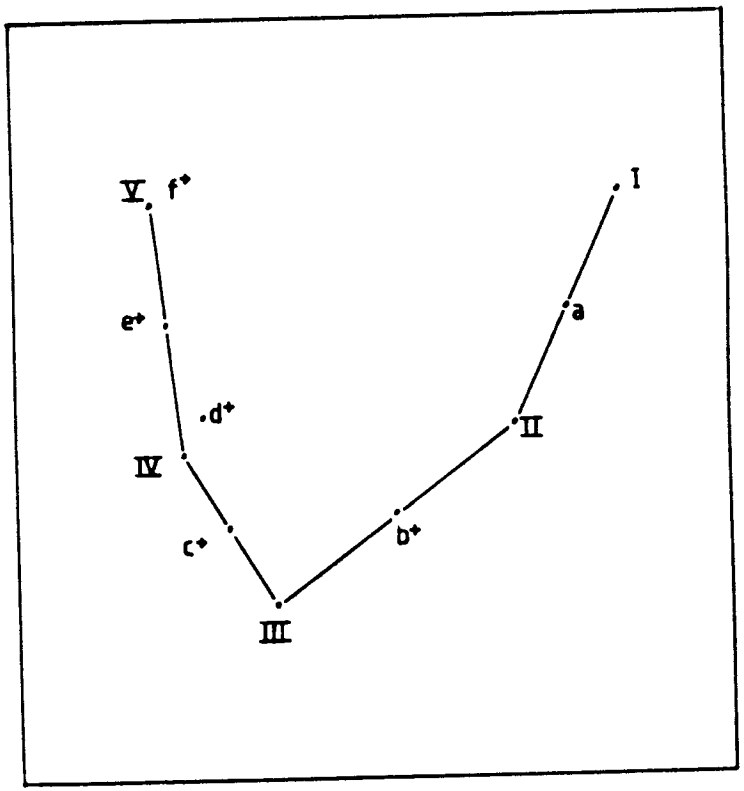


Figure 3.18. HOMALS solution for seriation example, based on incomplete indicator matrix.

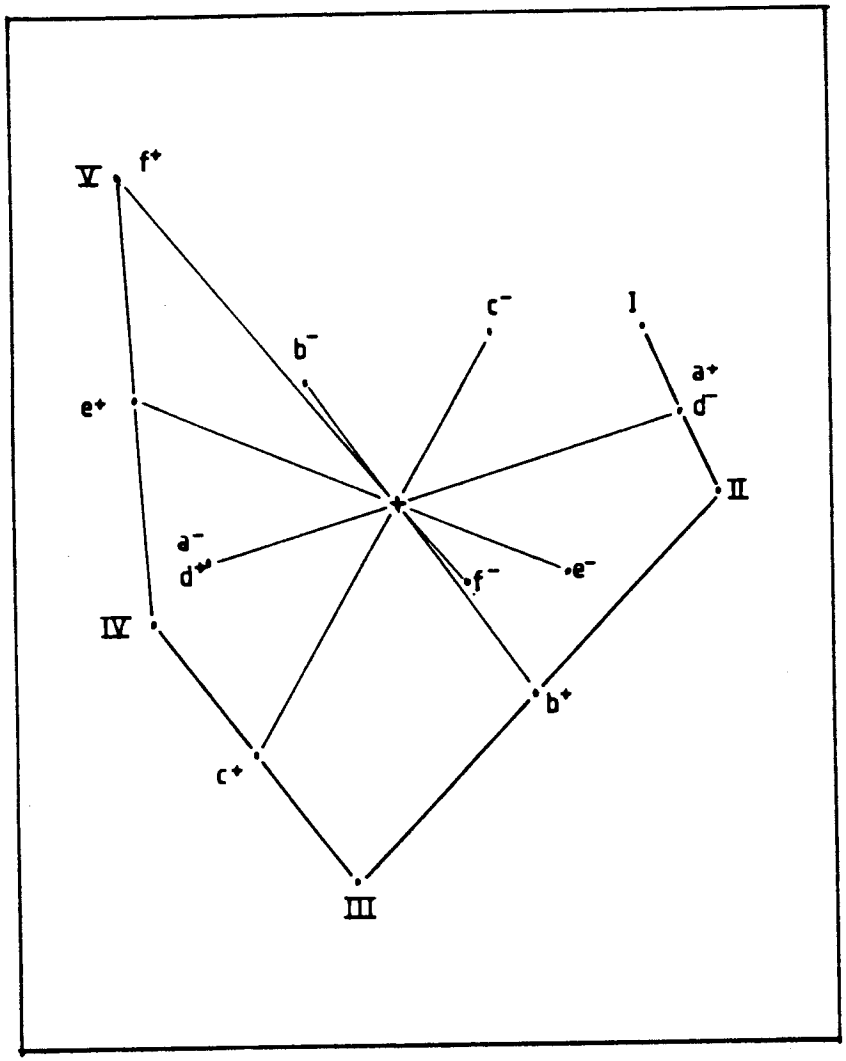


Figure 3.19. HOMALS solution for seriation example, bases on completed indicator matrix.

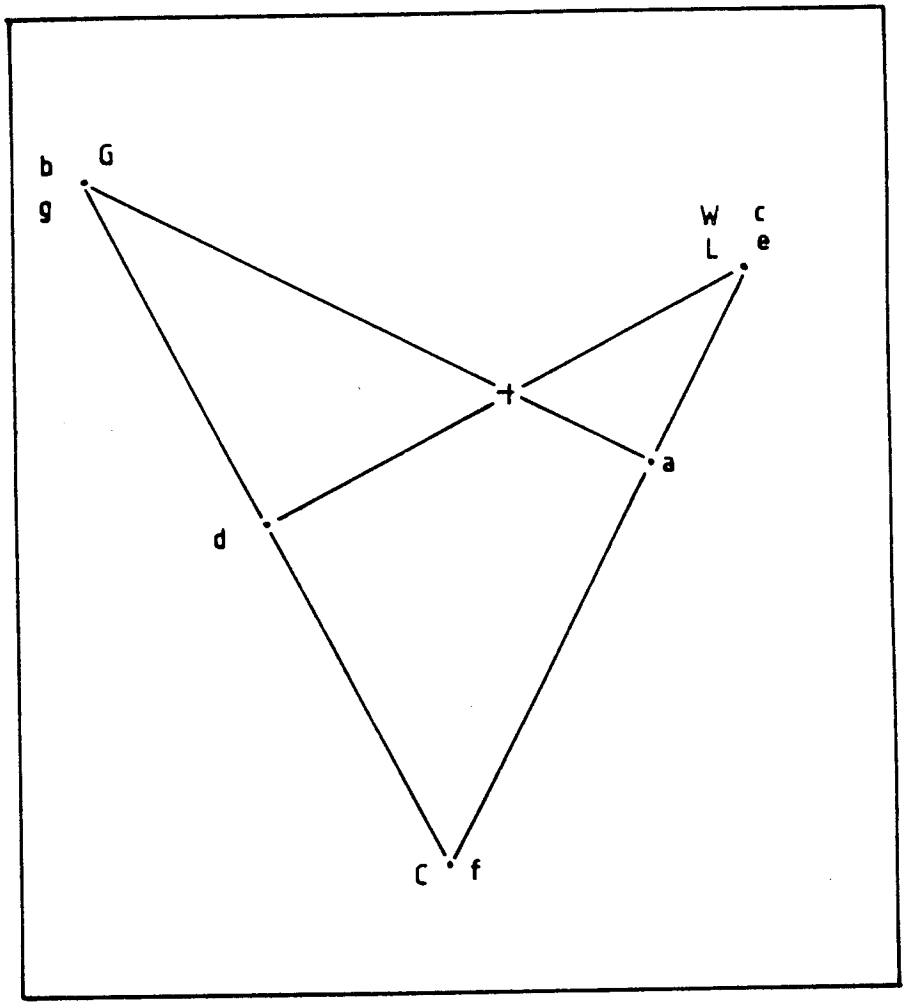


Figure 3.20. HOMALS solution for mini sorting task, illustrating solution for reversed indicator matrix.

used in section 2.4.3 (the "seriation problem"), both for the first two dimensions.

Although for both solutions the order for the objects (or for the categories) as given in table 2.8A "curves up", the first dimension of figure 3.18 still shows objects (and categories) ordered in the same way as in table 2.8A, whereas in figure 3.19 there is no longer any direction in the plot on which objects can be projected in their order of the table. Figure 3.19 therefore can be said to make objects I and V (the extremes of the scale) more similar. Note also that in figure 3.19 category point for + and - of each category are each others opposites with respect to the origin of the plot.

3.11.4

There is a technical problem related to incomplete indicator matrices. It is the following: section 3.7.2 showed that if we take deviations from the mean of a complete indicator matrix G_j, the resulting matrix (called S_j in section 3.7.2) has rank k_j-1. It follows that we might as well skip one column of S_j. the last one, say, since it is linearly dependent of other columns in S_j. One might be tempted to think that this comes to the same as the option "mssing data deleted", as if data for the last category in G_j were "missing".

This, however, is not true. The basic difference is

(i) with "missing data deleted" object scores are proportional to the average category quantification for categories that apply to the object (the number of those categories will not always be equal to m).

(ii) Deleting a column from S_j comes down to nothing but a computational short- in the analysis of the complete indicator matrix. Object scores remain calcula as if the deleted column was there.

Example. In the example of section 3.7.3, suppose we drop from G_j the last column. Let the remaining matrix be called $\bar{G}_j$, and define $\bar{G} = (\bar{G}_1, \bar{G}_2, \bar{G}_3)$. Th solution for category quantification y could as well have been derived from the generalised eigenvector equation

$$\bar{G}'\bar{G}a = D_{\bar{G}'\bar{G}}a\psi^2$$

where a_j will become a vector equal to

$$a_j = \begin{matrix} y_{j1} - y_{j,k_j} \\ \cdots\cdots\cdots \\ y_{j,k_j-1} - y_{j,k_j} \end{matrix}$$

In words, a_j is obtained if we delete the last element of y_j, and subtract this last element from each of the other elements of y_j. For the numerical examp we would obtain

$$a = \begin{matrix} -.805 \\ 1.505 \\ -.059 \\ 2.609 \\ -1.445 \end{matrix}$$

From the result we could reconstruct the last element of y_j as illustrated here for y_1. Its last element is $y_{13} = -(6 \times (-.805) + 2 \times (1.505))/10 = .182$. Then y_1 must become

$$y_1 = \begin{matrix} a_1 \\ 0 \end{matrix} + .182 = \begin{matrix} -.805 + .182 \\ 1.505 + .182 \\ 0 + .182 \end{matrix} = \begin{matrix} -.623 \\ 1.688 \\ .182 \end{matrix}$$

which is the solution we found for the complete indicator matrix.
The procedure above therefore does not imply that the column dropped from G_j to obtain $\tilde{G}_j$ is treated as if it contains missing data.

3.12 Reversed indicator matrix

In 2.6 we made the observation that the most typical application of analysis of a reversed indicator matrix would be in a sorting task. As a mini-example, suppose we have four objects. They are the names "Washington", "Lincoln", "Churchill", and "de Gaulle" (W L C G). Three individuals are asked to sort these names into categories, as many as they like, and without telling us how they define their categories. Suppose individual (1) makes the groups (W L C)(G), perhaps related to whether the objects were English speaking or not. Individual (2) sorts (W L)(C G), perhaps implicating that he sort "before 1900" versus "after 1900". Individual (3) sorts (W L)(C)(G) ,perhaps by country of birth. The data matrix and reversed indicator matrix are given in table 3.18. A HOMALS solution on this reversed indicator matrix is given in figure 3.20. Objects W and L coincide (they have the same sorting pattern), and they also coincide with categories c and e (unique for W and L). Object C coincides with category f, object G with categories b and g. Category a is in the center of gravi of W,L.C; category d in the center of gravity of C,G. Assuming that our interpretation of categories was correct, the picture shows that W and L are both born in the USA and lived before 1900, whereas G is French and lived after 190 and C,W,L are English speaking.
In section 2.6 another example was given. Its classical indicator matrix leads to the HOMALS solution given in table 3.19 for category quantification and object scores, where this solution (different from standard HOMALS) takes object scores as the averages of category quantifications. The plot is shown in figure 3.21. Clearly, object (4) takes a dimension for itself (the first dimension), with categories Ic, IIa, and IIIa coinciding with the point for this object. The second dimension contrasts objects (1) and (5), or categories

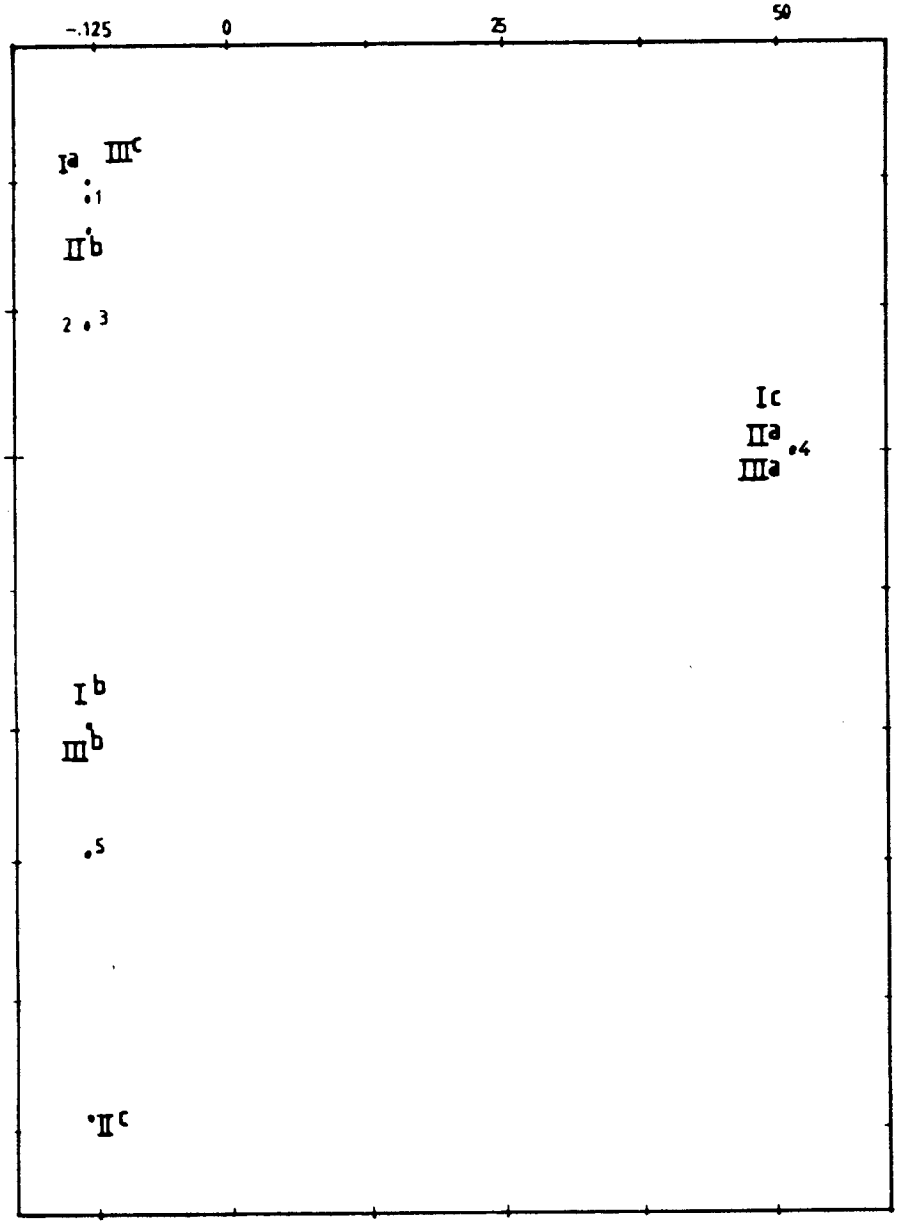


Figure 3.21. HOMALS solution for 'classical' indicator matrix.

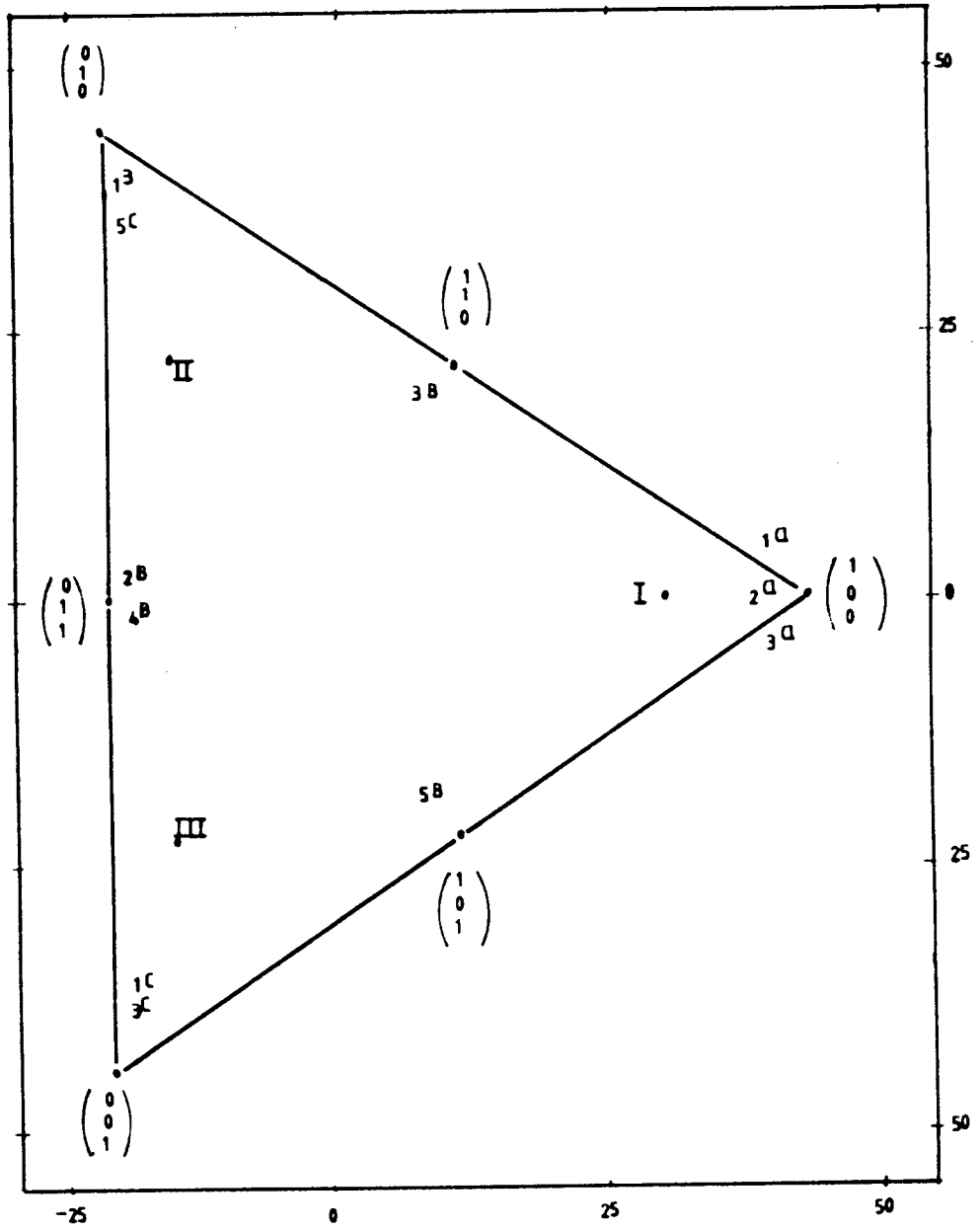


Figure 3.22. HOMALS solution of same example shown in figure 3.21, with reversed

	W	L	C	G
1	a	a	a	b
2	c	c	d	d
3	e	e	f	g

Table 3.18a

	1		2		3		
	a	b	c	d	e	f	g
W	1	0	1	0	1	0	0
L	1	0	1	0	1	0	0
C	1	0	0	1	0	1	0
G	0	1	0	1	0	0	1

Table 3.18b Example of reversed indicator matrix for sorting task.

Ia	-.129	.250
Ib	-.129	-.250
Ic	.516	.000
IIa	.516	.000
IIb	-.129	.204
IIc	-.129	-.612
IIIa	.515	.000
IIIb	-.129	-.250
IIIc	-.129	.250

1	-.129	.235
2	-.129	.068
3	-.129	.068
4	.568	.000
5	-.129	-.371

Table 3.19 Category quantification and induced object scores "classical" indicator matrix, normalization y'Dy = 1

1a	.436	.000
1b	-.218	.447
1c	-.218	-.447
2a	.436	.000
2b	-.218	.000
3b	.109	.224
3c	-.218	-.447
4a	-.218	.000
4c	.436	.000
5b	.109	-.224
5c	-.218	.447

I	.305	.000
II	-.153	.224
III	-.153	-.224

1	1.00	1.00
2	1.00	.00
3	.25	.75
4	1.00	.00
5	.25	.75

Table 3.20 Quantification of reversed indicator matrix, with discrimination measures for objects.

IIc versus (Ia, IIb, IIIc).

HOMALS for the reversed indicator matrix quantifies columns (categories for individual objects), and quantifies rows (variables) as shown in table 3.20. These results are plotted in figure 3.22. They illustrate that the first dimens contrasts I with (II,III), and also contrasts the columns of the reversed indicator matrix $\begin{matrix}1&1&1\\0&0&1\\0&1&0\end{matrix}$ with the columns $\begin{matrix}0&0&0\\0&1&1\\1&1&0\end{matrix}$. In words, it contrasts columns with upper element 1 to those with upper element 0.

Similarly, the second dimension produces a contrast between variables II and III, corresponding to columns in which the last two elements are $\begin{matrix}0\\1\end{matrix}$ versus columns with last two elements $\begin{matrix}1\\0\end{matrix}$.

HOMALS on the reversed indicator matrix quantifies variables and categories per individual, it does not quantify objects. However, for objects we now have the discrimination measure. For the example they are also given in table 3.20. Since the solution in the example is based on normalization $y'Dy = 1$, discrimination measures have been calculated as $y_j'D_jy_j\psi^2$, so that they can be interpreted again as squared correlations, now for correlations between each individual "sort" and the quantification of the variables. In the example, for the first individual, first dimension, this correlation is 1.00, because his "sort" corresponds exactly with the quantification of the variables on the first dimension.

4.1 Multidimensional scaling

Multivariate Analysis and Multidimensional Scaling (MDS) have developed along rather separate historical traditions. In MVA the basis concepts are those of 'variance', 'covariance', 'correlation', there is little emphasis on geometrical concepts, there is much concern about distributional assumptions. In MDS the interest is primarily in joint plots of objects and variables, and a basic concept is that of - Euclidean or other - 'distance', often related to the concept of 'dissimilarity' (between objects, between variables, or between an object and a variable).
In this chapter we want to show how the two traditions converge, in the sense that for certain indicator matrices the HOMALS solution also is a solution of problems of MDS.
A typical example of a data matrix in MDS is a rectangular binary matrix. It might have the following interpretation: rows refer to individuals, columns refer to objects ("stimuli"), and the cell of the matrix has 1 if the individual 'chooses' the object, 0 otherwise. What 'to choose' means, depends on the context; it micht mean agreement with a statement, preference for a kind of food, a vote in favour of some proposal,etc.

In the 1960's a variety in suggestions has been made as to how to analyze such a matrix (Coombs, 1964; Lingoes, 1968; De Leeuw, 1969) with non-metric (i.e., ordinal) MDS. Typical of such proposals is that it was stressed that assumptions about the structure of the data should be weak. Typical also was that rows and columns of the matrix were treated asymmetrically: the matrix was defined either as 'row-conditional', or as 'column-conditional'. Row-conditionality implies the general idea that in a spatial representation columns are given as points in such a way that for each row there is some separating surface. Figure 4.1. gives a very simple example, for which the columns are represented as five points, and for each row there is a line with the one's for that row on one side, the zero's on the other.
A different representation of the same data is given in figure 4.2, illustrating the 'unfolding' model. Now stimuli and individuals are both represented as points in a joint space, in such a way that an individual is closer to the stimuli he "chooses" and farther away from the stimuli he does not choose.
illustrates this with circles around the individual points: chosen stimuli are al within the circle, and not chosen stimuli are outside.
It is obvious that the representations offered in figures 4.1 and 4.2 are far from unique. One reason for this is that the example is a very small one. Another more important, reason is that, even with many individuals and stimuli, the data matrix does not impose many restrictions (Kruskal and Carroll, 1969; Heiser and De Leeuw, 1979). A consequence is that it is difficult to find algorithms which

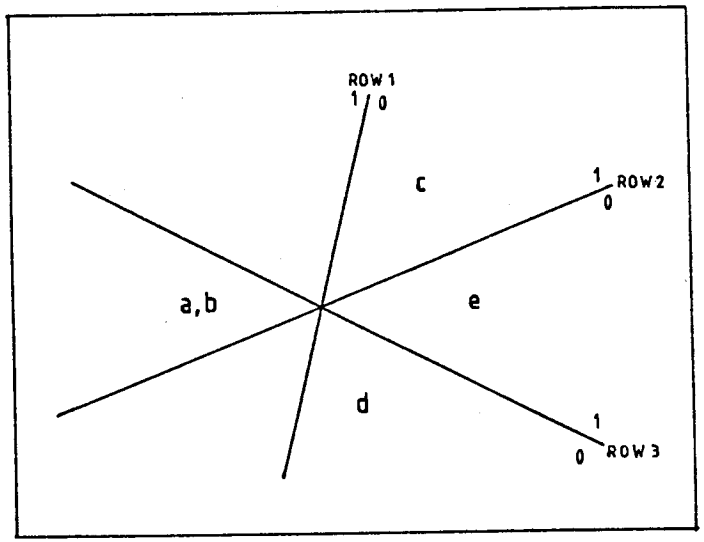


Figure 4.1. Row conditional scaling solution

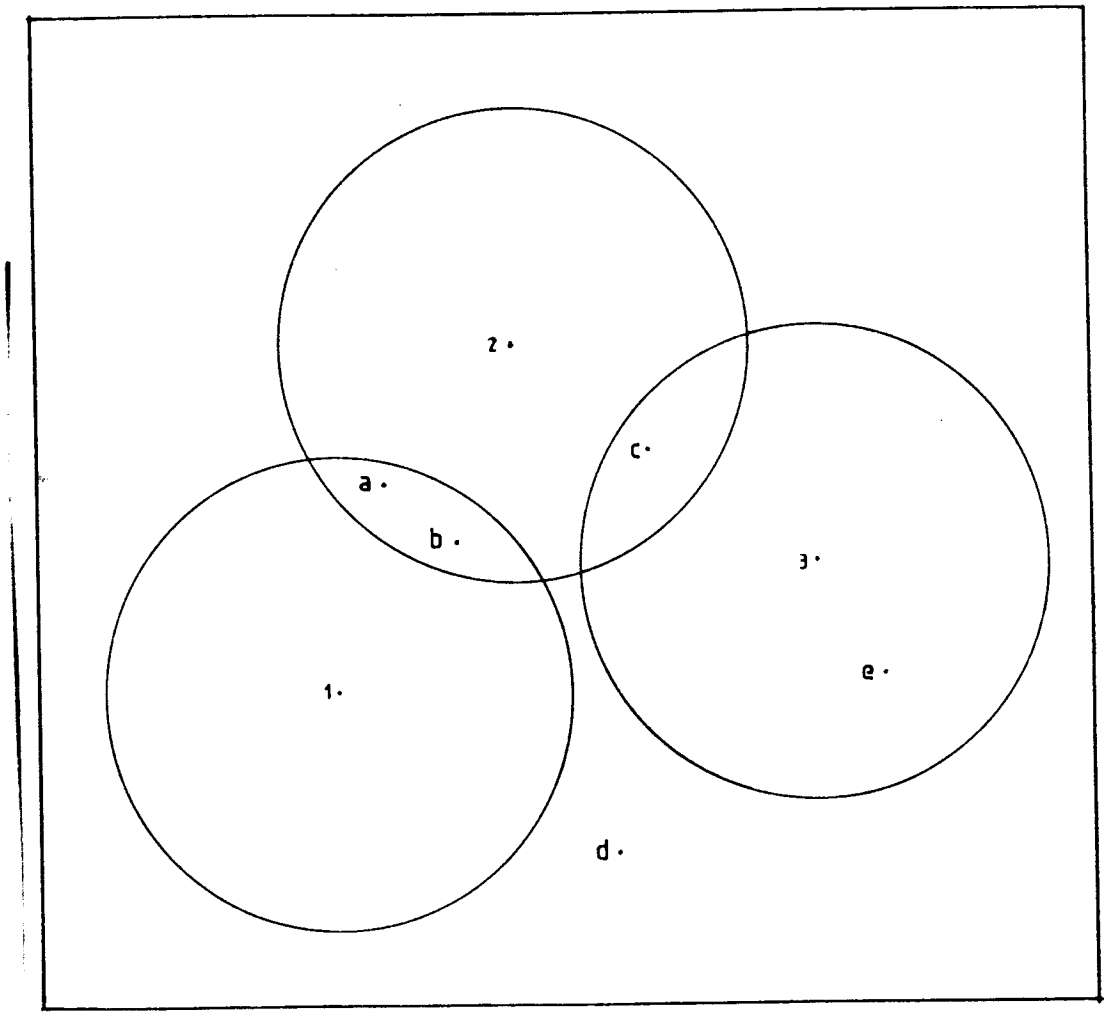


Figure 4.2. Unfolding solution for data matrix of figure 4.1.

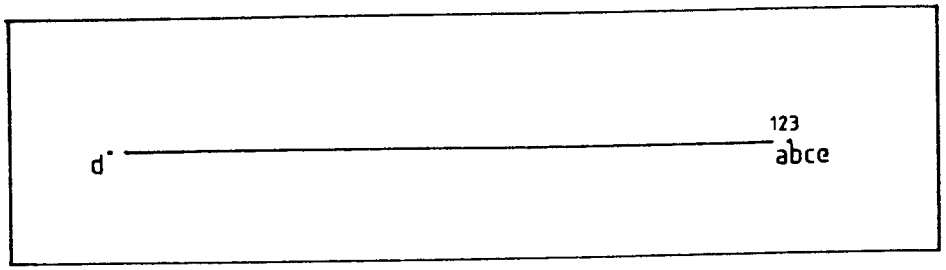


Figure 4.3. Degenerate unfolding solution for data matrix of figure 4.1.

avoid degenerate solutions. An example of such a degenerate solution is shown in figure 4.3 for the same data matrix as in figure 4.1 and 4.2. All individuals are represented as a single point, coinciding with the stimuli a b c e, and contrasting them with stimulus d, chosen by none of the individuals.

4.2 A HOMALS solution for unfolding

'Unfolding theory' was first formulated by Coombs (see Coombs, 1964). The idea is elegant and simple. Let G be a rectangular matrix (n individuals, m objects) where a row gives the preference rank order of an individual for the objects. We want a representation in which the objects are plotted as points, and where an individual also is plotted as a point in such a way that his preference rank order is the same as the order of the distances between the individual points and the object points. The individual's most preferred'object must be nearest, his least-preferred object should be farthest away. This idea implies that a preference rank order is a 'folded' version of the stimulus space, or, that the stimulus space can be obtained by 'unfolding' all preference rank orders.

In this section we shall limit the matrix G to a binary matrix, with elements $g_{ij} = 1$ if individual i chooses object j, and $g_{ij} = 0$ otherwise. The MDS solution for unfolding then requires a representation where the distance between an individual's points and an chosen object point is always smaller than the distance to a non-chosen object point.

That there is a relation with HOMALS is immediately apparent. The HOMALS solution can plot a stimulus point in the center of gravity of the individual points for those individuals who choose the object, with the consequence that, at the whole, individual points must be closer to the chosen stimuli than to the non-chosen stimuli.

For a mini example we shall use the matrix G of table 4.1, for 5 individuals and 4 binary variables, corresponding to a Coombs "pick 2" design (the individual has to tell which 2 out of 4 objects he likes best). The "pick 2" design has the consequence that rows of G add up to the same number, a property that will be used in the following derivations. It is not an essential property, but it simplifies the illustration.

(i) We first look upon G as row-conditional. I.e., we want to scale rows so that in their spatial representation row points are closer together to the extent that rows are more 'similar'. Let $=_j$ (read: "is similar with respect to column j to") be an equivalence relation for pairs of rows in the sense that

$r_i =_j r_k$ if $g_{ij} = g_{kj} = 1$

Table 4.1

	a	b	c	d
1	1	0	1	0
2	1	1	0	0
3	0	1	1	0
4	0	0	1	1
5	0	1	0	1

Table 4.1 Mini example "pick 2" design

Table 4.2

0	0	1	1	1
0	0	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	1	1

1	1	1	1	1
1	0	0	1	0
1	0	0	1	0
1	1	1	1	1
1	0	0	1	0

0	1	0	0	1
1	1	1	1	1
0	1	0	0	1
0	1	0	0	1
1	1	1	1	1

1	1	1	1	1
1	1	1	1	1
1	1	1	1	1
1	1	1	0	0
1	1	1	0	0

1/2	1/2	0	0	0
1/2	1/2	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

0	0	0	0	0
0	1/3	1/3	0	1/3
0	1/3	1/3	0	1/3
0	0	0	0	0
0	1/3	1/3	0	1/3

1/3	0	1/3	1/3	0
0	0	0	0	0
1/3	0	1/3	1/3	0
1/3	0	1/3	1/3	0
0	0	0	0	0

0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	1/2	1/2
0	0	0	1/2	1/2

5/6	3/6	2/6	2/6	0
3/6	5/6	2/6	0	2/
2/6	2/6	4/6	2/6	2/
2/6	0	2/6	5/6	3/
0	2/6	2/6	3/6	5/

Table 4.2 Dissimilarity matrices $\{\delta_{ijk}\}$ (upper row) and matrices of weights $\{w_{ijk}\}$ (lower row) for j = 1,..,4 from left to right , with their sum matrix W

The "ideal" representation would be to find a representation in p-dimensional space where the coordinates of the point for r_i are the same as those for r_k when $r_i =_j r_k$. This, of course, will in general be impossible since $r_i =_j r_k$ does not imply $r_i =_h r_k$ for $j \neq h$. We therefore have to compromise and must define a specific loss function to be minimised in the compromise. To this end we define for each of the m columns of G an n x n dissimilarity matrix with elements

$\delta_{ikj} = 0$ if $r_i =_j r_k$

$\delta_{ikj} = 1$ if $r_i \neq_j r_k$

For the example, the four dissimilarity matrices are shown in table 4.2.

In addition, we define for each column a matrix of weights, with elements

$w_{ikj} = 1/d_j$ if $r_i =_j r_k$

$w_{ijk} = 0$ if $r_i \neq_j r_k$

where d_j is the total of the j^{th} column of G. The four matrices $W_{ik.}$ are also shown in table 4.2. Note that $W_{ik.}$ gives the dissimilarities a weight reciprocal to the size of the equivalence class.

The standard MDS problem is to approximate all dissimilarity matrices with a matrix of distances between row points in a space with as small as possible number of dimensions (Torgerson, 1958; Shepard, 1962; Kruskal, 1964, 1977; Guttman, 1968; for a recent survey see Carroll and Arabie, 1980; a more theoretical discussion is De Leeuw and Heiser, 1980). The usual MDS loss function is

$$\sigma_1(X) = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{n} \sum_{k=1}^{n} w_{ikj} \left(\delta_{ikj} - \gamma_{ik}(X)\right)^2 \qquad (4.2.1)$$

where $\gamma_{ik}(X)$ gives the Euclidean distance between rows i and k of the p-dimensional matrix X of coordinates in the representation of the rows. The metric MDS problem is to find a solution for X which minimizes $\sigma_1(X)$, under appropriate normalization conditions.

The problem can be much simplified by using the special properties of the dissimilarity matrices and weight matrices. For all i,j,k we have $w_{ikj}\delta_{ikj} = 0$, and therefore

$$\sigma_1(X) = \frac{1}{2} \sum_j \sum_i \sum_k w_{ikj}\gamma_{ik}^2(X) = \frac{1}{2} \sum_i \sum_k \gamma_{ik}^2(X) \sum_j w_{ikj} \qquad (4.2.2)$$

The elements $\sum_j w_{ikj}$ can be found as the elements of the sum matrix W of the four matrices $W_{ik.}$. For the example W also is given in table 4.2. It then can be shown that $W = GD^{-1}G'$, with D the diagonal matrix of the column totals of G. The proof is simple. Obviously, $GD^{-1}G' = \sum_j g_j d_j^{-1} g_j'$ and $g_j d_j^{-1} g_j' = W_{ikj}$. It the follows that

$$\sigma_1(X) = m\ tr(X'X) - tr(X'GD^{-1}G'X) \qquad (4.2.3)$$

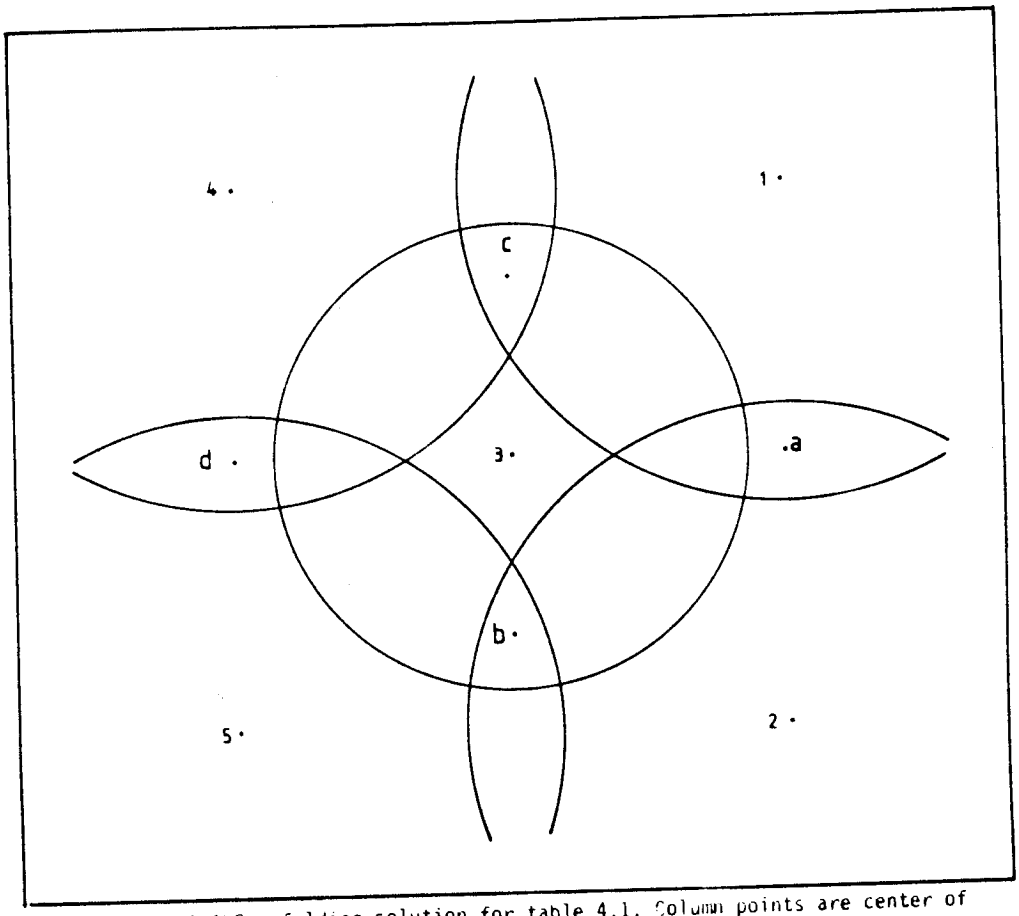


Figure 4.4. HOMALS unfolding solution for table 4.1. Column points are center of gravity of row points.

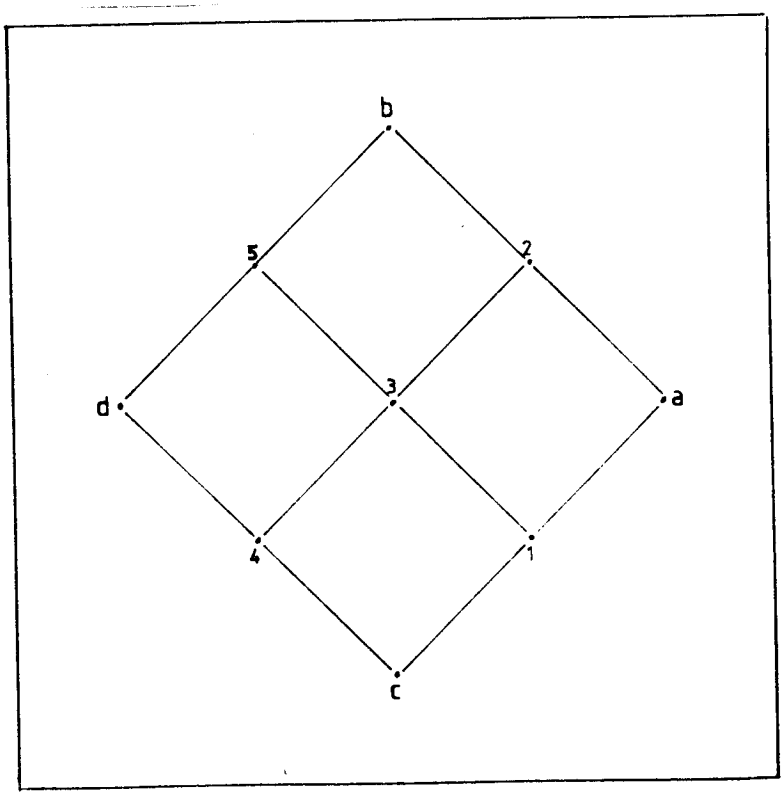


Figure 4.5. HOMALS unfolding solution of table 4.1. with row points as centroids of column points.

and we are back at the HOMALS solution for the quantification of row objects, for which we now take normalization restriction X'X = I. (In most MDS solutions the normalization restriction is less severe, it then is required only that tr(X'X) = 1.)
For the example, we find for a two-dimensional solution

$$X = \begin{bmatrix} .5 & .5 \\ .5 & -.5 \\ 0 & 0 \\ -.5 & .5 \\ -.5 & -.5 \end{bmatrix}$$

with loss $\sigma_1(X) = 1/3$. (Loss becomes 0 if we allow a third dimension, with quantification $x_3' = (1\ 1\ -4\ 1\ 1)/\sqrt{20}$). The solution is pictured in figure 4.4, where also the stimulus points are drawn as centers of gravity of individuals who have "picked" it. To illustrate 'unfolding', circle segments are drawn around each individual point, showing that stimuli which are "picked" are inside such circles, and stimuli which are not "picked" are outside.

(ii) The other approach would be to look upon G as column conditional; i.e., we primarily want to represent columns as points in p-dimensional space with coordinates Y. We shall not repeat details of the derivations. It can be shown that the loss function now becomes
$\sigma_2(Y) = \text{tr}(Y'DY) - \frac{1}{m}\text{tr}(Y'G'GY)$
which is the same as for HOMALS column quantification of the indicator matrix. For the example this solution is pictured in figure 4.5 which is essentially the same as figure 4.4 but now with individuals in the center of gravity of their chosen stimuli.

(iii) The third approach focuses on similarity relations within the pairs of one individual and one stimulus. We simply define a dissimilarity matrix $\{\delta_{ij}\}$ with

$\delta_{ij} = 0$ if $g_{ij} = 1$

$\delta_{ij} = 1$ if $g_{ij} = 0$

and a matrix of weights $\{w_{ij}\}$ with $w_{ij} = g_{ij}$. The loss function is
$\sigma_3(X) = m\ \text{tr}(X'X) + \text{tr}(Y'DY) - 2\ \text{tr}(X'GY)$
and will be minimized by the HOMALS solution for X and Y, with normalization, e.g., X'X = I, and $Y = D^{-1}G'X$, as in figure 4.4.

Summarizing: all three approaches come to the same result. Typical of the HOMALS solution is that it approaches the "ideal" $\eta_{ij}(X,Y) = 0$, if $g_{ij} = 1$. This is a stronger demand than in MDS, where in the row-conditional interpretation it is required that

if $g_{ij} = 0$ and $g_{il} = 1$, then $\gamma_{ij}(X,Y) \geq \gamma_{il}(X,Y)$

or, in the column-conditional interpretation

if $g_{ij} = 0$ and $g_{kj} = 1$, then $\gamma_{ij}(X,Y) \geq \gamma_{kj}(X,Y)$

The HOMALS solution, therefore, is obtained at the price of a metric interpretation of the data, and at the price of stronger normalization conditions, with equivalent treatment of rows and columns. On the other hand, MDS techniques with weaker assumptions not only require as to the interpretation of G as row or column conditional, they also tend to produce degenerate solutions, precisely because of their weaker assumptions.

4.3 "Analyse des correspondances"

4.3.1

Let F be an R x C frequency matrix; its cell f_{ij} gives the frequency with which row category i goes together with column category j. Let $r = Fu$ be the vector of row totals, and $c = F'u$ the vector of column totals, and let $n = u'Fu = u'c = u'r$ be the total frequency. Also, let D_r be the diagonal matrix of r, and D_c the diagonal matrix of c.

In "analyse des correspondances" as developed by Benzécri (1973) the aim is to find a representation X of the rows of F in such a way that Euclidean distances between rows of X correspond to distances between rows of F. The squared distance δ_{ik}^2 between rows i and k of F is defined as

$$\delta_{ik}^2 = n \sum_j (f_{ij}/r_i - f_{kj}/r_k)^2/c_j$$

We shall call such distances 'χ^2-distances'; this name will be explained in section 4.3.3. Obviously, δ_{ik}^2 is the same as the Euclidean distance between rows i and k of the matrix with elements

$$h_{ij} = (f_{ij}/r_i).(n/c_j)^{\frac{1}{2}}$$

This matrix also can be written as

$$H = D_r^{-1}FD_c^{-\frac{1}{2}}.n^{\frac{1}{2}}$$

4.3.2

X is a representation of F if $HH' = XX'$. Proof: $\delta_{ik}^2 = h_i'h_i + h_k'h_k - 2h_i'h_k = x_i'x_i + x_k'x_k - 2x_i'x_k$. A solution for X is found as follows. Let

$$D_r^{-\frac{1}{2}}FD_c^{-\frac{1}{2}} = K\Lambda L'$$

be the SVD solution, with $K'K = I$, $L'L = I$, and Λ the diagonal matrix of singular values. Define

$X = D_r^{-\frac{1}{2}} K\Lambda . n^{\frac{1}{2}}$

so that

$XX' = D_r^{-\frac{1}{2}} K\Lambda^2 K' D_r^{-\frac{1}{2}} . n$.

It follows from the SVD solution that

$D_r^{-\frac{1}{2}} F D_c^{-1} F' D_r^{-\frac{1}{2}} = K\Lambda^2 K'$

so that

$XX' = D_r^{-1} F D_c^{-1} F' D_r^{-1} . n = HH'$

However, the matrix $D_r^{-\frac{1}{2}} F D_c^{-\frac{1}{2}}$ has a left singular vector $D_r^{\frac{1}{2}} u$, with corresponding right singular vector $D_c^{\frac{1}{2}} u$. This is seen from the following equalities

$D_r^{-\frac{1}{2}} F D_c^{-\frac{1}{2}} D_c^{\frac{1}{2}} u = D_r^{-\frac{1}{2}} F u = D_r^{-\frac{1}{2}} r = D_r^{\frac{1}{2}} u$

$D_c^{-\frac{1}{2}} F' D_r^{-\frac{1}{2}} D_r^{\frac{1}{2}} u = D_c^{-\frac{1}{2}} F' u = D_c^{-\frac{1}{2}} c = D_c^{\frac{1}{2}} u$

which also shows that the corresponding singular value equals one. The unit normalized version of $D_r^{\frac{1}{2}} u$ is $D_r^{\frac{1}{2}} u . n^{-\frac{1}{2}}$, and it follows that X must have a column with elements 1. This column obviously does not contribute to the distance between two rows of X. We therefore drop this column from X. This comes to the same as re-defining the SVD solution as

$D_r^{-\frac{1}{2}} F D_c^{-\frac{1}{2}} - D_r^{\frac{1}{2}} u u' D_c^{\frac{1}{2}} / n = K\Lambda L'$

with the "trivial" singular vectors eliminated. This again corresponds to a replacement of H by $H - uu' D_c^{\frac{1}{2}} n^{-\frac{1}{2}}$. We come back to this matrix in section 4.3.3.

Given the solution for X, coordinates for column points are defined as $Y = D_c^{-\frac{1}{2}} L n^{\frac{1}{2}}$, with the effect that $D_r^{-1} F Y = X$. The latter equation shows that row points are the center of gravity of the column points weighted as to their frequency in the row. Benzécri calls this "le principe barycentrique." We have met the idea before as a quantification principle of HOMALS (cf. sections 2.3, 2.4.2, 3.7.3, 3.11). We come back to this principle in section 4.3.5.

Obviously, the solution of "analyse des correspondances" also could have been developed the other way round, with χ^2-distances between columns of F (in stead of rows), and column point as the center of gravity of row points. This is a matter of normalization only; the solution then becomes

$Y = D_c^{-\frac{1}{2}} L\Lambda n^{\frac{1}{2}}$

$X = D_r^{-\frac{1}{2}} K n^{\frac{1}{2}}$

and where the reversal of "le principe barycentrique" is completely comparable with the choice between the two HOMALS normalizations in section 3.7.3.

What remains invariant in both solutions (and therefore also in plots of both solutions) is the equality $XY' = D_r^{-1}FD_c^{-1}.n - uu'$. As to plots: this equality implies the following. If we draw a line through a row point, and project column points on this line, the signed lengthes of these projections will be proportional to a row of XY'. Also, if we draw a line through a column point, and project the row points on this line, the projections will have signed lengthes proportional to a column of XY'. Elements of XY' are the values $(f_{ij}-e_{ij})/e_{ij}$, where e_{ij} is the "expected value".

4.3.3

In section 4.3.2 it was found that χ^2-distances between rows of F are the same as Euclidean distances between rows of $H - uu'D_c^{\frac{1}{2}}n^{-\frac{1}{2}}$. An element is the latter matrix can be written as

$$h_{ij} - (c_j/n)^{\frac{1}{2}} = (f_{ij}/r_i).(c_j/n)^{-\frac{1}{2}} - (c_j/n)^{\frac{1}{2}} = \frac{(f_{ij}/r_i) - (c_j/n)}{(c_j/n)^{\frac{1}{2}}}$$

The numerator of the expression at the right gives the difference between a row proportion f_{ij}/r_i and the corresponding marginal proportion c_j/n. The denominator is the square root of the marginal proportion. In so far as row proportions can be interpreted as an 'estimate' of the marginal proportions, we may write

$$\Sigma_j\{h_{ij}-(c_j/n)^{\frac{1}{2}}\}^2 = \chi_i^2$$

This explains why the distances δ_{ik} in section 4.3.1 were called χ^2-distances.

Also, elements of

$$(D_r^{-\frac{1}{2}}FD_c^{-\frac{1}{2}} - D_r^{\frac{1}{2}}uu'D_c^{\frac{1}{2}}/n).n^{\frac{1}{2}} = K\Lambda L'.n^{\frac{1}{2}}$$

(where the trivial singular vectors are eliminated from the expression at the right) are equal to $(f_{ij}-e_{ij})/e_{ij}^{\frac{1}{2}}$, where e_{ij} again is the 'expected value' (compare section 3.9.2). It follows that the over-all χ^2 for F is the sum of the squared elements, equal to the trace of

$$K\Lambda L'L\Lambda K'.n = K\Lambda^2K'.n$$

Obviously, this matrix has trace equal to $n\Sigma\lambda_i^2$, so that we have the equality

$$\chi^2 = n\Sigma\lambda_i^2$$

which, on the assumption of independence between rows and columns of F, becomes the χ^2-statistic with (R-1)(C-1) degrees of freedom.

4.3.4

For a mini-example, let

$$F = \begin{matrix} 4 & 4 & 2 & 0 \\ 1 & 2 & 8 & 9 \\ 1 & 6 & 8 & 15 \end{matrix}$$

with $r' = (10 \quad 20 \quad 30)$, $c' = (6 \quad 12 \quad 18 \quad 24)$, $n=60$. Expected values are

$$E = \begin{matrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 3 & 6 & 9 & 12 \end{matrix}$$

The matrix with elements $(f_{ij}-e_{ij})/e_{ij}^{\frac{1}{2}}$ will be called $Bn^{\frac{1}{2}}$:

$$Bn^{\frac{1}{2}} = \begin{matrix} 3.00 & 1.41 & -.58 & -2.00 \\ -.71 & -1.00 & .82 & .35 \\ -1.15 & .00 & -.33 & .87 \end{matrix}$$

The sum of the squared elements of $Bn^{\frac{1}{2}}$ is $\chi^2=19.82$. B has SVD solution $K\Lambda L'$ with

$$K = \begin{matrix} .912 & .029 \\ -.282 & .766 \\ -.296 & -.642 \end{matrix}$$

and eigenvalues $\lambda_1^2=.307$, $\lambda_2^2=.023$. This confirms $\chi^2=n\Sigma\lambda^2$.

The solution for X becomes

$$X = D_r^{-\frac{1}{2}}K\Lambda.n^{\frac{1}{2}} = \begin{matrix} 1.238 & .011 \\ -.271 & .203 \\ -.232 & -.139 \end{matrix}$$

X gives the representation of the rows of F. Euclidean distances between rows of X are equal to the χ^2-distances between the rows of F. The squared distances are

$$\{\delta_{ij}^2\} = \begin{matrix} 0 & 2.32 & 2.18 \\ 2.32 & 0 & .12 \\ 2.18 & .12 & 0 \end{matrix}$$

These distances are the same as for the rows of $H-uu'D_c^{\frac{1}{2}}.n^{-\frac{1}{2}}$. This matrix could be calculated as follows. First create a matrix of row proportions:

.40	.40	.20	.00
.05	.10	.40	.45
.033	.20	.267	.50
.10	.20	.30	.40

Subtract marginal proportions from row proportions:

.30	.20	-.10	-.40
-.05	-.10	.10	.05
-.067	.00	-.033	.10

and divide by the square root of the marginals:

.949	.447	-.183	-.632
-.158	-.224	.183	.079
-.211	.000	-.060	.158

The solution for column representation becomes

$$Y = D_c^{-\frac{1}{2}} L.n^{\frac{1}{2}} = \begin{matrix} 2.416 & .766 \\ .819 & -1.368 \\ -.280 & 1.269 \\ -.804 & -.458 \end{matrix}$$

Figure 4.6 gives the plot. The distances between the row points in the plot are the χ^2-distances. Row points are the center of gravity of column points weighted as to frequencies in the row. The figure further illustrates that projections of the y-points on the line through x_2 are proportional to the second row of XY', and that projections of the x-points on the line through y_3 are proportional to the third column of XY'. Elements of XY' are $(f_{ij}-e_{ij})/e_{ij}$:

$$XY' = \begin{matrix} 3.000 & 1.000 & -.333 & -1.000 \\ -.500 & -.500 & .333 & .125 \\ -.667 & .000 & -.111 & .250 \end{matrix}$$

4.3.5

The "principe barycentrique" has been formulated independently by many authors. The oldest reference is probably Richardson (1933, quoted in Horst, 1935), who called it "the method of reciprocal averages". Horst (1935) remarked that the method "is precisely the same as that for determining the factor loadings for a group of tests" (p.373), which shows that Horst saw the relation with PCA. Horst also saw the connection with quantification of nominal categories, but he does not mention the possibility of approximating continuous non-linear transformations in this way. Neither does Guttman in his justly famous 1941 paper, in which he makes the important step to relate non-linear PCA with chi-square. Later contributions can be found in Johnson (1950), Lord (1958), Bock (1960), De Leeuw (1973), Nishisato (1980), and, of course, Benzécri (1973). The latter's influence probably works through in other publications of the "french school", such as Dauxois and Pousse (1976), and Lafaye de Michaux (1978).

A typical example where the method is re-introduced (even under its oldest name: method of reciprocal averaging) is a paper by Hill (1974). This is a paper on the 'ordination' problem in phytosociology - it is formally the same problem as the 'seriation' problem in archeology (or the Guttman 'scalogram' problem in social science). Hill's method is identical with what in this book is called HOMALS with incomplete indicator matrix. Hill compares with what he calls a PCA solution (HOMALS with complete indicator matrix, option (ii)), and also with PCA without standardization (SVD of G itself). As one could expect, he finds the latter method less satisfactory (it contaminates scale values with marginal frequencies).

4.3.6

The relation between 'analyse des correspondances' and HOMALS is very close. In section 2.7 is was shown how for a given frequency table a corresponding indicator matrix G can be created. G becomes an n x (R+C) matrix. Its category quantification for columns will be the vector y that satisfies

$Cy = Dy\psi^2$

For the indicator matrix of a frequency table, C becomes

$$C = \begin{matrix} D_r & F \\ F' & D_c \end{matrix}$$

so that

$$D^{-\frac{1}{2}}CD^{-\frac{1}{2}} = \begin{matrix} I & D_r^{-\frac{1}{2}}FD_c^{-\frac{1}{2}} \\ D_c^{-\frac{1}{2}}F'D_r^{-\frac{1}{2}} & I \end{matrix} = \begin{matrix} I & K\Lambda L' \\ L\Lambda K' & I \end{matrix}$$

It can be immediately verified that this matrix has eigenvectors $\begin{matrix} K \\ L \end{matrix}$ with eigenvalues $I+\Lambda$. It follows that the category quantification is given by $Y_r = D_r^{-\frac{1}{2}}K$ for row categories, and $Y_c = D_c^{-\frac{1}{2}}L$ for column categories (where we use the non-standard normalization $Y'DY = I$). The difference with the solution by "analyse des correspondances", therefore, is only in normalization.

4.3.7

Using the notation of section 4.3.6 it also can be easily shown that Y_r and Y_c give stationary values for the correlation between rows and columns of F. Since $u'D_rY_r = 0$, and $u'D_cY_c = 0$, columns of Y_r and Y_c are in deviations from means. If we normalize $Y_r'D_rY_r = I$ and $Y_c'D_cY_c = I$ (in stead of $Y'DY = I$, as in section 4.3.6),

$Y_r'FY_c = \{\rho\}$

becomes a matrix of correlations. But

$Y_r'FY_c = K'D_r^{-\frac{1}{2}}FD_c^{-\frac{1}{2}} = K'K\Lambda L'L = \Lambda$

so that singular values also are "canonical correlations". The relation with earlier notation for HOMALS eigenvalues is

$\psi^2 = I + \Lambda$

so that

$\Lambda = \psi^2 - I = 2\Phi - I$ (since m=2)

(where ψ^2 is restricted to eigenvalues $\psi^2 > 1$).

In terms of χ^2 we now have the relation that χ^2/n must be equal to the sum of the squared canonical correlations.

4.3.8

We may give a few comments on the history of the relations between chi-square and canonical correlation. Pearson (1904) defined a measure of "dependence", called the mean square contingency $\phi^2 = \chi^2/n$. The underlying argument is that for a continuous normal distribytion $f(\underline{x})$ there is an infinite set of orthogonal transformations $\psi^{(s)}(\underline{x})$ of degree s (s=1,...,∞) (called Hermite-Chebyshev polynomials), for which it can be shown that the correlation between $\psi^{(s)}(\underline{x})$ and $\psi^{(s)}(\underline{y})$ equals ρ^s, where $f(\underline{x},\underline{y})$ is a binormal distribution with correlation parameter ρ. It can be shown that χ^2/n is the sum of the squares of these correlations (in fact, this is consistent with the result of section 4.3.7, since for the continuous binormal stochastic distribution, the series ρ^s is a series of canonical correlations.)

This implies

$$\chi^2/n = \rho^2 + \rho^4 + \rho^6 + \ldots\ldots = \frac{\rho^2}{1 - \rho^2}$$

so that

$$\rho^2 = \frac{\chi^2}{n + \chi^2} = \frac{\phi^2}{1 + \phi^2}$$

Pearson also applied the latter coefficient to contingency tables. In so far as a contingency table can be assumed to be a "discretization" of a binormal distribution, with many classes for $\underline{x}$ and $\underline{y}$, the correlation parameter ρ can be estimated as $\phi/(1+\phi^2)^{\frac{1}{2}}$. Pearson called this the coefficient of contingency.

Gebelein (1941) further developed the theory. He showed that there are reansformations $\phi(\underline{x})$ and $\psi(\underline{y})$ for which the correlation beween $\phi(\underline{x})$ and $\psi(\underline{y})$ is maximized, and that this maximum correlation κ_{xy} is related to eigenvalues. He also made the point that, if both ϕ and ψ are linear, the "maximum" (there is not much to maximize in this case) measure is the correlation coefficient ρ, if either ϕ or ψ is linear, the maximum becomes the correlation ratio (η_{xy} or η_{yx}), and that $\rho \leq \eta \leq \kappa$.

Gebelein's work was generalised by Rényi (1959) who formulated a sort of checklist of criteria for the "ideal" measure of dependence δ_{xy} :

A: δ_{xy} is defined for all non constant $\underline{x}$ and $\underline{y}$;

B: $\delta_{xy} = \delta_{yx}$

C: $0 \leq \delta_{xy} \leq 1$

D: $\delta_{xy} = 0$ if and only if $\underline{x}$ and $\underline{y}$ are independent

E: when $\underline{x} = \phi(\underline{y})$ or $\underline{y} = \psi(\underline{x})$, then $\delta_{xy} = 1$

F: if ϕ and ψ are one-one mappings, then $\delta_{xy} = \delta_{\phi_x\psi_y}$

G: if $(\underline{x},\underline{y})$ is binormal with correlation parameter ρ, then $\delta_{xy} = |\rho|$

Applying this checklist to four measures of dependence, the correlation

coefficient ρ_{xy}, the maximum η of the correlation ratio's η_{xy} or η_{yx}, the coefficient of contingency $\phi/(1+\phi^2)^{\frac{1}{2}}$, and Gebelein's κ_{xy}, the following table can be set up:

	A	B	C	D	E	F	G
ρ_{xy}	0	1	1	0	0	0	1
η	0	1	1	0	1	0	1
$\phi/(1+\phi^2)^{\frac{1}{2}}$	0	1	1	1	0	1	1
κ_{xy}	1	1	1	1	1	1	1

Lancaster (1959, 1960a, 1960b) generalised some of such results for dependence between more than two variates.

Going back to contingency tables, Hirschfeld (1935) discussed the problem whether it is possible to quantify rows and column of such a table so that (in our notation) $u'D_r x = 0$, $u'D_c y = 0$, $x'D_r x = 1$, $y'D_c y = 1$, and $D_r^{-1}Fy = \rho x$, $D_c^{-1}F'x = \rho y$. In words: is there a quantification so that both regressions become linear? Hirschfeld showed that solutions are related to the SVD of $D_r^{-\frac{1}{2}}FD_c^{-\frac{1}{2}}$ (as shown in section 4.3.7); he also showed relations with the continuous case.

Fisher (1940) rediscovered this technique and inspired its first applications by Maung (1941a, 1941b). These studies also point out relations between (in the terminology of this book) maximum canonical correlation between indicator matrices, maximum product moment correlation for a bivariate frequency table, and maximum discrimination between categories for such tables. The same aspects are mentioned in Guttman's 1941 paper. (Further developments and applications can be found in Yates, 1948; Johnson, 1950; Williams, 1952; Lancaster, 1957; Bock, 1960). Maung described the representation

$$F = D_r\{uu' + \sum_{s}^{\min(R,C)} x_s \lambda_s y_s'\}D_c$$

(in our notation) with reference to the comparable case of continuous variates. Earlier, Mehler for this case had formulated the representation

$$f(\underline{x},\underline{y}) = f(\underline{x}).f(\underline{y})\{1 + \sum_{s=1}^{\infty} \rho^s \psi_s(\underline{x})\psi_s(\underline{y})\}$$

where ψ_s are the Hermite-Chebyshev polynomials. The Maung-Fisher equivalent is

$$f(\underline{x},\underline{y}) = f(\underline{x}).f(\underline{y})\{1 + \lambda_s \phi_s(\underline{x})\psi_s(\underline{y})\}$$

where λ_s (canonical correlations) replace ρ^s, and where ϕ_s and ψ_s are the canonical transformations. (Further developments: Lancaster, 1958; Hannan, 1961; Venter, 1966; Cambanis and Liu, 1971; Jensen, 1971; Chesson, 1976)

One specific question is: when are the canonical transformations polynomial functions? The question is of interest, because in many applications of non-linear MVA polynomial, or almost polynomial, transformations are found (examples in Barrett and Lampard, 1955; McFadden, 1966; McGraw and Wagner, 1968; Lee, 1971). Even more, in many applications the "best" transformation is almost linear, the "second best" almost quadratic, etc. Studies by Eagleson (1964), Eagleson and Lancaster (1967), and Lancaster (1975) show that for many special bivariate distributions (not only the normal, but also Poisson, gamma, hypergeometric,etc.) the canonical transformations must be classical orthogonal polynomials.
A further question then is: suppose that the canonical transformations are polynomials, what restrictions does this impose upon the canonical correlations? As indicated earlier, for a binormal distribution the canonical correlations are powers of the correlation parameter ρ. Other cases are discussed in Sarmanov and Bratoeva (1967), Eagleson (1969), Griffiths (1969,1970), Tyan and Thomas (1975). Generalizations to the multinormal case are treated in Appell and Kampé de Fériet (1926), Erdelyi (1953). Sarmanov and Zacharov (1960), Venter (1966), Naouri (1970), Dauxois and Pousse (1976).

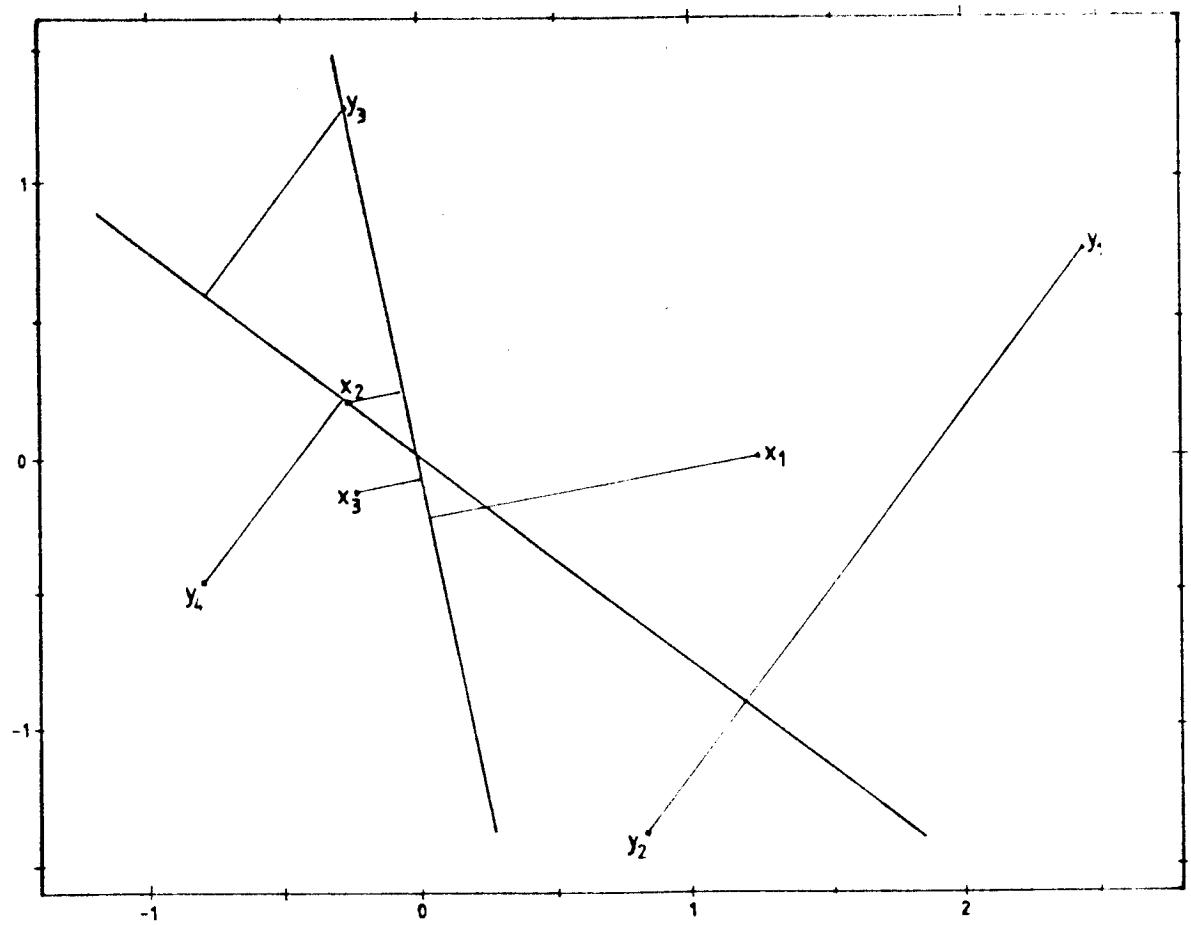


Figure 4.6. Analyse des correspondances for mini example.

4.4 The program ANACOR

4.4.1

The program ANACOR has been developed to handle data where "analyse des correspondances" is feasible. Essentially, the program solves for X and Y, as defined in section 4.3. The program has three options for normalization

(i) $X = D_r^{-\frac{1}{2}} K\Lambda . n^{\frac{1}{2}}$

$Y = D_c^{-\frac{1}{2}} L . n^{\frac{1}{2}}$ so that

$X'D_rX = \Lambda^2 . n$

$Y'D_cY = I . n$

$X = D_r^{-1} FY$

where the last equality shows that this option plots row points in the center of gravity of column points.

(ii) $X = D_r^{-\frac{1}{2}} K . n^{\frac{1}{2}}$

$Y = D_c^{-\frac{1}{2}} L\Lambda . n^{\frac{1}{2}}$ so that

$X'D_rX = I . n$

$Y'D_cY = \Lambda^2 . n$

$Y = D_c^{-1} F'X$

with column points in the center of gravity of row points.

(iii) The third option drops "le principe barycentrique" and treats rows and columns symmetrically:

$X = D_r^{-\frac{1}{2}} K\Lambda^{\frac{1}{2}} . n^{\frac{1}{2}}$

$Y = D_c^{-\frac{1}{2}} L\Lambda^{\frac{1}{2}} . n^{\frac{1}{2}}$ so that

$X'D_rX = \Lambda . n$

$Y'D_cY = \Lambda . n$

The most straightforward application of ANACOR is to two-dimensional frequency tables (as discussed in section 4.3). ANACOR then quantifies row and column categories in the same way as HOMALS would do (apart from normalizations), but ANACOR does not quantify objects. Also: ANACOR gives the complete solution for all possible dimensions (whereas HOMALS has an option for p). An illustration is given in 4.4.2.

However, ANACOR also can handle two-dimensional tables where the entries are not frequencies, but a different type of non-negative numbers. Section 4.4.3 gives an example where the entries are distances.

Finally, ANACOR can handle higher-dimensional tables, but they will be treated

Table 4.3

occupation father \ occupation son	1	2	3	4	5	6	7	
1	50	19	26	8	18	6	2	129
2	16	40	34	18	31	8	3	150
3	12	35	65	66	123	23	21	345
4	11	20	58	110	223	64	32	518
5	14	36	114	185	714	258	189	1510
6	0	6	19	40	179	143	71	458
7	0	3	14	32	141	91	106	387
	103	159	330	459	1429	593	424	3497

1. PROF: professional and high administrative
2. EXEC: managerial and executive
3. HSUP: higher supervisory
4. LSUP: lower supervisory
5. SKIL: skilled manual and routine non-manual
6. SEMI: semi-skilled manual
7. UNSK: unskilled manual

Table 4.3 Occupational mobility data

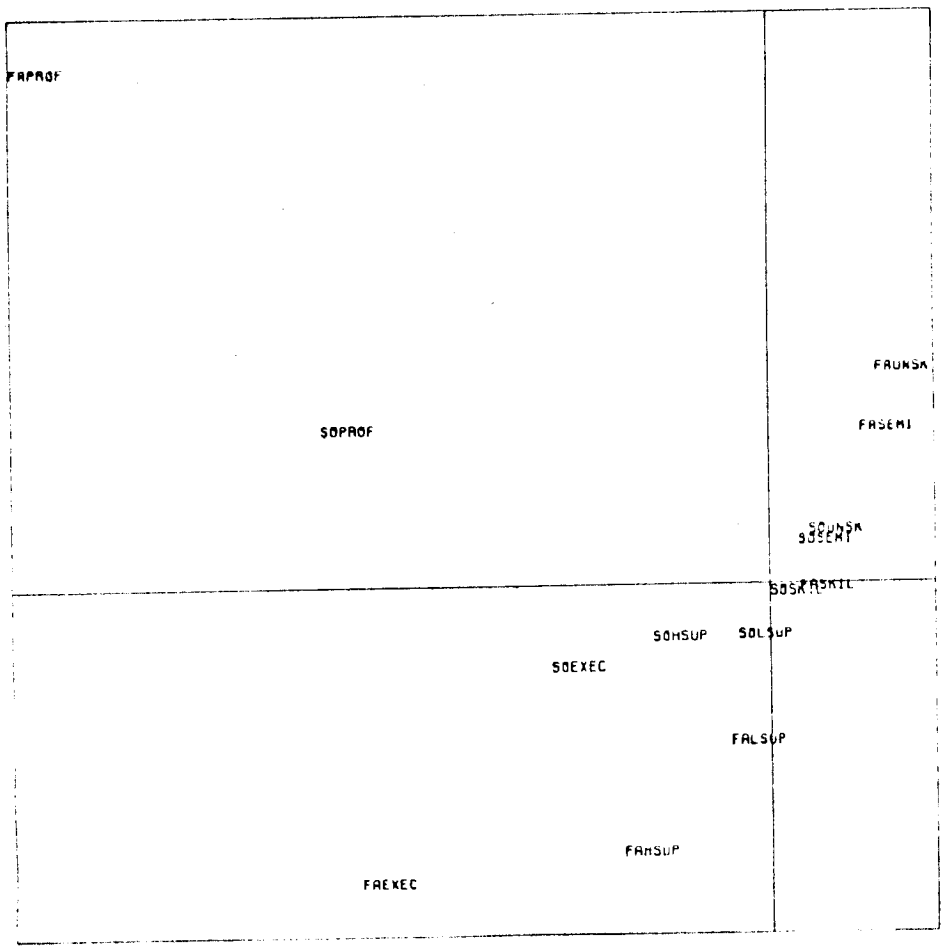


Figure 4.7 ANACOR solution for occupational mobility data

as bivariate, in the sense that the input for ANACOR is either the matrix of bivariate marginals, or a two-dimensional frequency table obtained by "grouping" dimensions. Both procedures will be illustrated in section 4.4.4.

4.4.2

Table 4.3 gives a 7 x 7 frequency table of occupational status of fathers versus occupational status of their sons for a sample of 3497 British families. These data also are discussed in Glass (1954), Goodman (1965, 1969), Haberman (1974), and Bishop e.a. (1975).
The standard ANACOR program will produce eigenvalues λ_s and quantifications of row and column categories x_s, y_s, for 6 dimensions. The program also will give plots of x_s versus y_s for each dimension, and joint plots of x_s and y_s for all pairs of dimensions. For the first two dimensions the singular values (canonical correlations) are λ_1=.526, λ_2=.267. The quantifications are given in table 4.4 (on page 151), plotted in figure 4.7 ; the solution is based on option (ii): sons are the center of gravity of fathers.

Clearly, the first dimension orders occupational status classes from low to high. The second dimension does the same with the notable exception for category PROF - one also could maintain that the second dimension is a quadratic function of the first. As a result, in figure 4.7 the sequence of labels for fathers (FA..) or for sons (So..) appears as a rather smooth curve, with category PROF much separated from the other categories.
Figure 4.8 (not a standard ANACOR plot) is a re-edit of table 4.3, with distances between categories corresponding to the solution of the first dimension, now with normalization as in option (iii) . In addition the two lines for linear regression are drawn.
As to substantive interpretation, the results above have mainly correlational significance: scaling of father and son categories cannot produce a correlation larger than .526. The analysis completely ignores any difference there might be between fathers and son as to their average score (the ANACOR solution gives the two variables a zero average), or difference as to spread. We therefore do not claim that ANACOR is the best way to analyze social mobility.

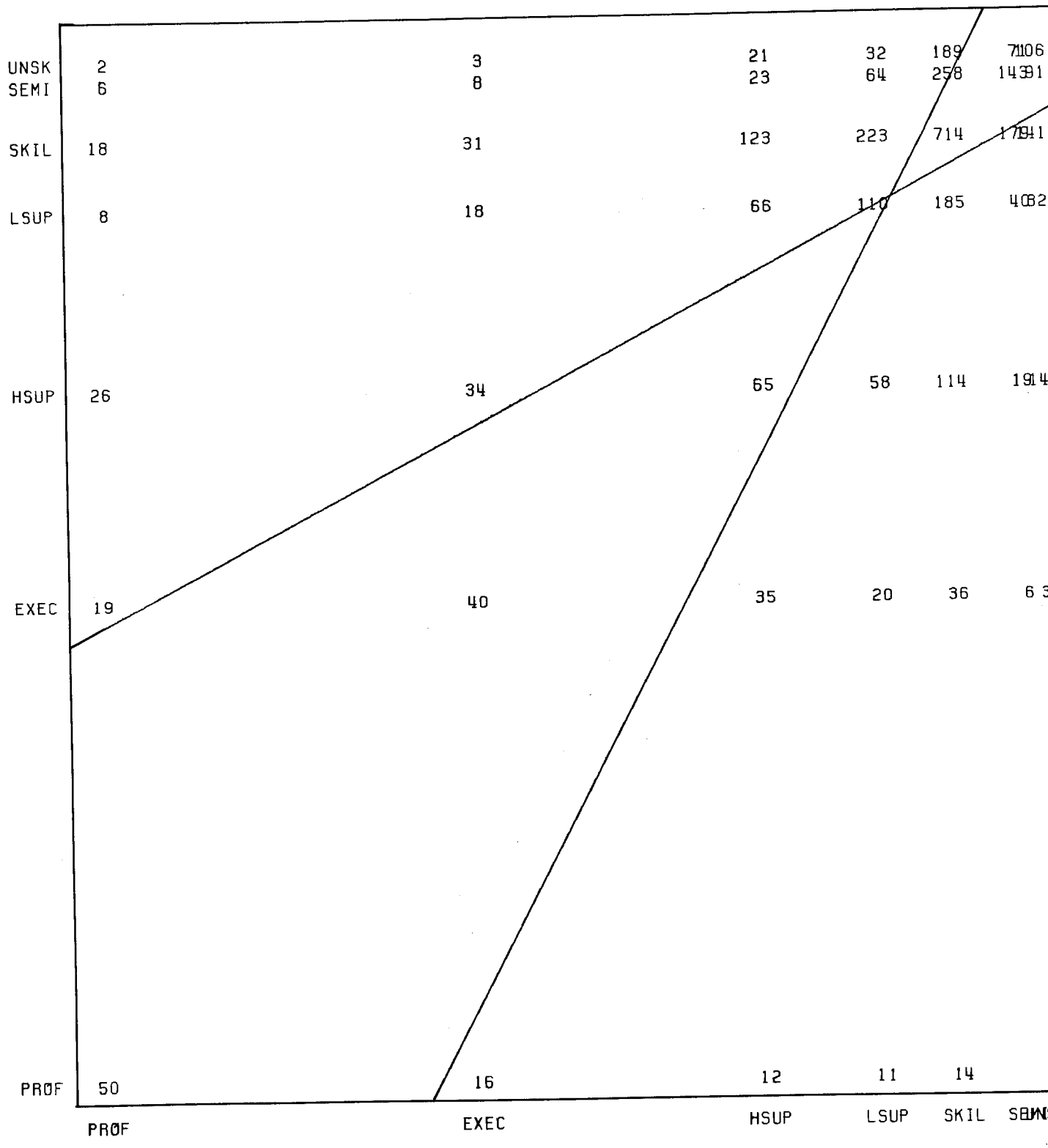
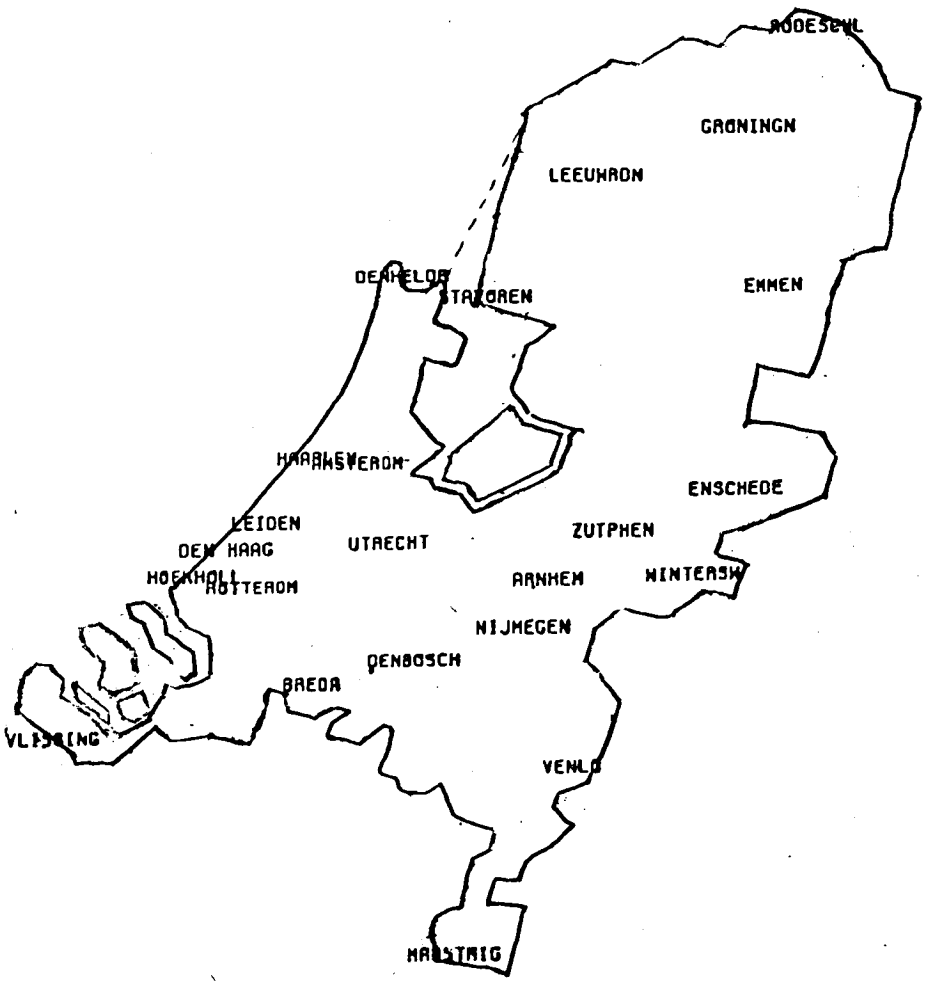
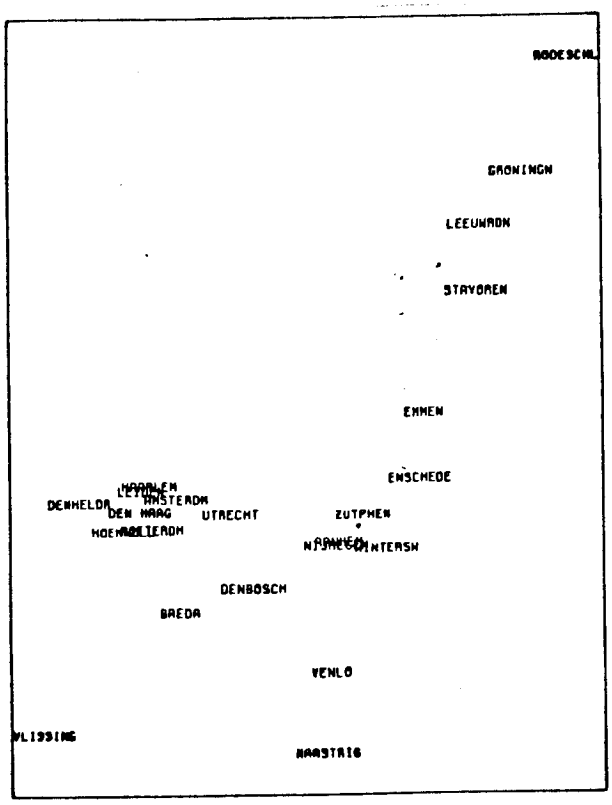


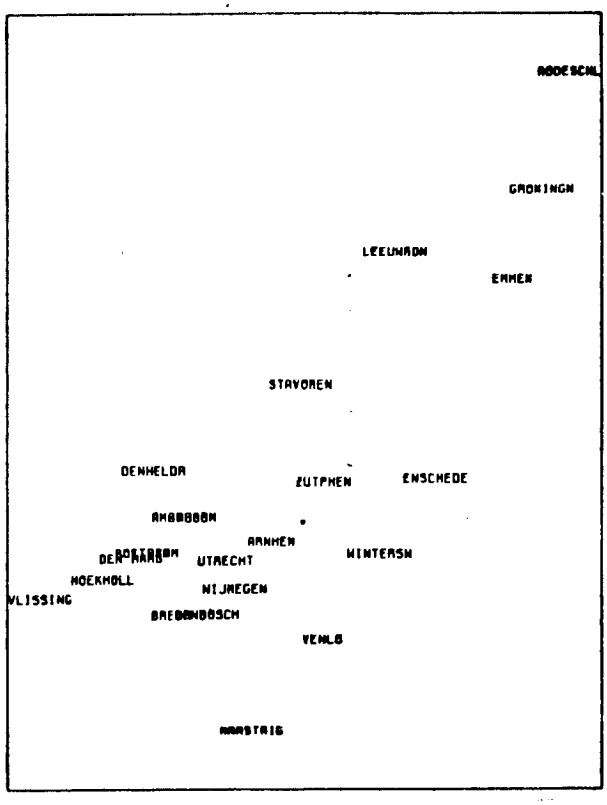
Figure 4.8 Table 4.3 graphed as a regression plot on the basis of first ANACOR solution



A: distances as the crow flies

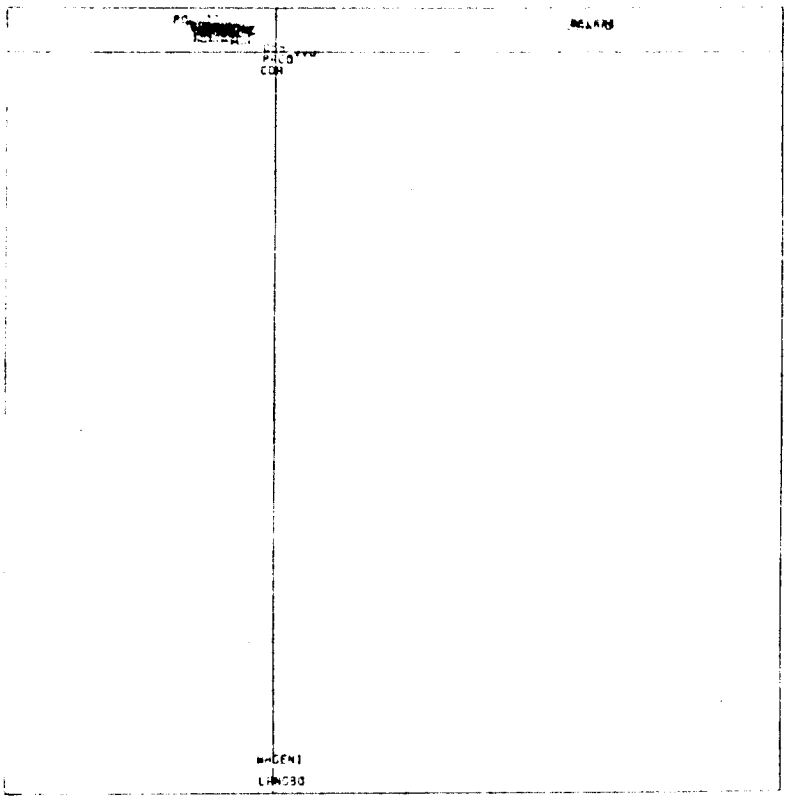


B: distances by railway

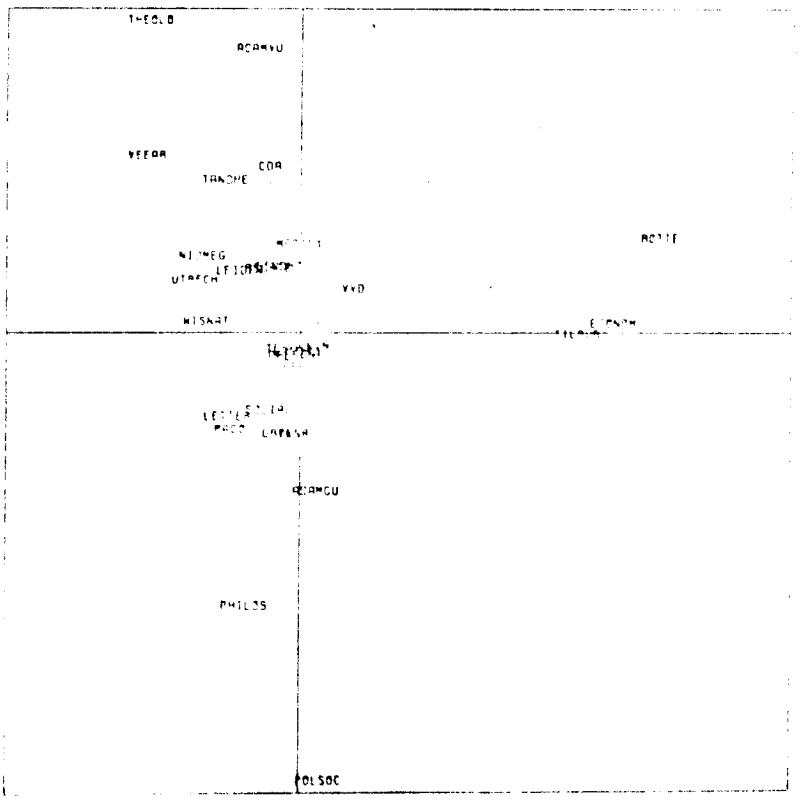


C: dichotomized distances

Figure 4.9 ANACOR solution for distances between 23 Dutch cities



A: dimensions 1 and 2



B: dimensions 3 and 4

Figure 4.10 ANACOR solution for NSR data

4.4.3

To illustrate that ANACOR is not restricted to frequency data , and to illustrate its appliction in the context of MDS, ANACOR has been applied to a matrix of distances between 23 cities in The Netherlands. Such a matrix is, by necessity, symmetric, and it follows that rows and columns will obtain identical quantification. Note that now there is no HOMALS equivalent, since it is impossible to create an indicator matrix along the lines of section 2.7.

The analysis has been done in three ways, all three with normalization option (iii). In the first analysis, F contains proximities, constructed by taking distances 'as the crow flies' and substracting them from the largest distance. In the second analysis proximities based on shortest railway connection were taken. In the third analysis F is a binary matrix with elements 0 if the distance is larger than median distance, or 1 if shorter.

Results are shown in figure 4.9. The first analysis produces a figure which is quite similar to the "real" map of the Netherlands. Figure 4.9B, for railway distances gives a typical distortion because there is no direct railway connection between Den Helder and Stavoren, whereas cities in the Western part of the country (where there are many direct connections) tend to cluster (similarly in the North-Eastern part). Figure 4.9C gives results for dichotomized distances. Given the severe simplification of the input, results are surprisingly good.

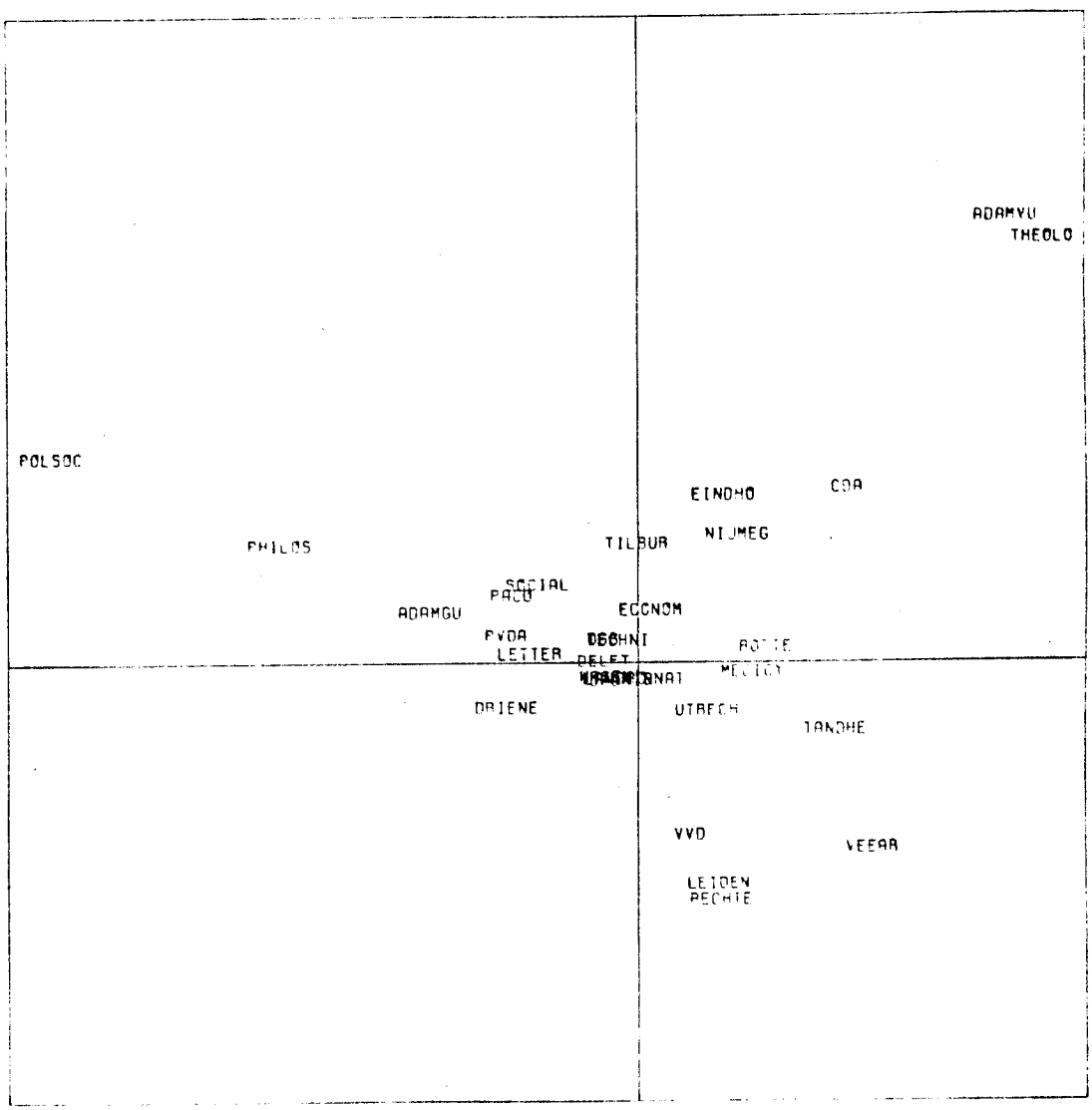
4.4.4

Table 4.9 (on page 356B) gives data collected in 1968 for a stratified sample of 1616 students from different Dutch universities (Lammers, 1969; De Leeuw, 1973). The table shows for each university a bivariate subtable of frequencies for faculties versus political preference. There are 13 different faculties, all together, but no university has all faculties, and some universities have only one faculty. As to political preference, the five categories (from left to right in the table) are

CDA : any denominational party (KVP, ARP, CHU, GPV, SGP)
VVD : conservative-liberal
PvdA: labour party
PACO: any of the smaller left-wing parties (PSP, CPN)
D'66: pragmatic-liberal

Although table 4.9 is trivariate, ANACOR handles it as bivariate. The input for ANACOR is the square matrix C (as defined for HOMALS in section 2.3. For this example C becomes a (12+13+5) x (12+13+5) matrix of univariate (diagonal) and bivariate (off diagonal) marginals.

Results are plotted in figures 4.10. The first three dimensions mainly reveal



C: dimensions 4 and

Figure 4.10 ANACOR solution for NSR data

artificial peculiarities of the data. Figure 4.10A, for the first two dimensions shows the typical difference between "technical" universities (DELFT, EINDHOVEN, DRIENENOORD) (with only one faculty that is not found at other universities), WAGENINGEN (where there is a faculty AGRICULTURE that is nowhere else), and the other universities. The third dimension (figure 4.10B) capitalizes on the fact that the ECONOMY faculties in ROTTERDAM and TILBURG are dominant. Results start to become more interesting in dimensions 4 and 5 (figure 4.10C). It shows (we mention only a few aspects) that students in THEOLOGY, at AMSTERDAM VU, favour CDA. that students in POLITICAL SCIENCE or PHILOSOPHY at AMSTERDAM GU favour PvdA or PACO; that students of the LAW faculty in LEIDEN and students in WAGENINGEN favour VVD.

The other possibility is that we group dimensions (as in section 2.9). For the present example, UNIVERSITIES and FACULTIES were grouped. Theoretically this gives 12 x 13 = 156 combinations, but many of them do not occur, so that 63 actual combinations are left. ANACOR now will produce the same quantification of categories as HOMALS with 5 columns for G_1 (political preference) and 63 for G_2 (combinations) .
Results are shown in figure 4.11, for the first two dimensions, with canonical correlations of .353 and .328. In the figure, CDA is towards the left, PvdA and PACO at the right downwards, VVD towards the top, D'66 in the center. Some other aspects: AMSTERDAM VU is favouring CDA for all faculties except PHILOSOPHY; PHILOSOPHY tends to favour PvdA and PACO, especially in LEIDEN and NIJMEGEN; THEOLOGY favours CDA except in LEIDEN and GRONINGEN; LAW faculties tend to favour VVD except at AMSTERDAM VU. The "trivial" dimensions that were dominating in the preceeding solution, now have vanished.

occupation son	PROF	-2.319	0.872	occupation father	PROF	-4.021	2.870
	EXEC	-1.051	-0.451		EXEC	-2.113	-1.639
	HSUP	-0.491	-0.291		HSUP	-0.660	-1.474
	LSUP	-0.019	-0.282		LSUP	-0.063	-0.865
	SKIL	0.160	-0.044		SKIL	0.330	-0.017
	SEMI	0.319	0.241		SEMI	0.668	0.865
	UNSK	0.376	0.299		UNSK	0.776	1.193

Table 4.4 Quantifications of occupational mobility data

4.5 The program ANAPROF

4.5.1

The program ANAPROF does the same as ANACOR or HOMALS, but is adapted to a profile frequency matrix as input. The program is quicker, and cheaper, when the number of objects n is much greater than the number of response patterns (or profiles).In general, the number of possible response patterns is equal to $\prod_m k_j$.Let P be a $\prod k_j \times \sum k_j$ matrix. A row identifies a possible profile by having entry '1' for the catefory chosen in the profile, and '0' for the non-chosen categories.

Define T as an $n \times \prod k_j$ matrix with element t_{hi} if the h^{th} object has the i^{th} profile in P. The complete indicator matrix can be written as G = TP. Also, T'T is the diagonal matrix of marginal profile frequencies. HOMALS implies the SVD solution

$$GD^{-\frac{1}{2}} = V\psi W' \qquad (4.5.1)$$

which we now can write as

$TPD^{-\frac{1}{2}} = V\psi W'$ (ignoring degenerate solutions). This implies

$T'TPD^{-\frac{1}{2}} = T'V\psi W'$ and

$(T'T)^{\frac{1}{2}}PD^{-\frac{1}{2}} = (T'T)^{-\frac{1}{2}}T'V\Psi W'$

Write $(T'T)^{-\frac{1}{2}}T'V = \bar{V}$. Then $\bar{V}'\bar{V} = I$. Proof: obviously, in the n x p matrix V rows are identical for objects with identical response pattern. In T'V such identical rows are added to their sum row. The premultiplication $(T'T)^{-\frac{1}{2}}$ divides such a sum row by the square root of its frequency. It follows that $\bar{V}'\bar{V} = V'V = I$. This implies that

$$(T'T)^{\frac{1}{2}}PD^{-\frac{1}{2}} = \bar{V}\psi W' \qquad (4.5.2)$$

also is an SVD solution, which can be solved more quickly since (4.5.1) has an $n \times \sum k_j$ matrix, while (4.5.2) has a $\prod k_j \times \sum k_j$ matrix (and where $n \geq \prod k_j$).

4.5.2

For an illustration we tale the data of Sugiyama (1975), where n = 4243 Japanese individuals responded to m = 6 binary questions about religious practice. The questions are listed in table 4.5. Theoretically, there are $2^6 = 64$ possible profiles. Table 4.6 lists them, together with their frequency of occurrence. Three response patterns have zero frequency, many have very low frequency, some have vary large frequency.

For the example, P as defined in section 4.5.1 is a 64 x 12 matrix, with G = TP the complete indicator matrix. ANAPROF gives the full SVD solution in $\sum k_j - m = 6$ dimensions. The first two eigenvalues are $\phi_1 = .269$ and $\phi_2 = .204$. Figure 4.12 shows the plot of the response patterns in the first two

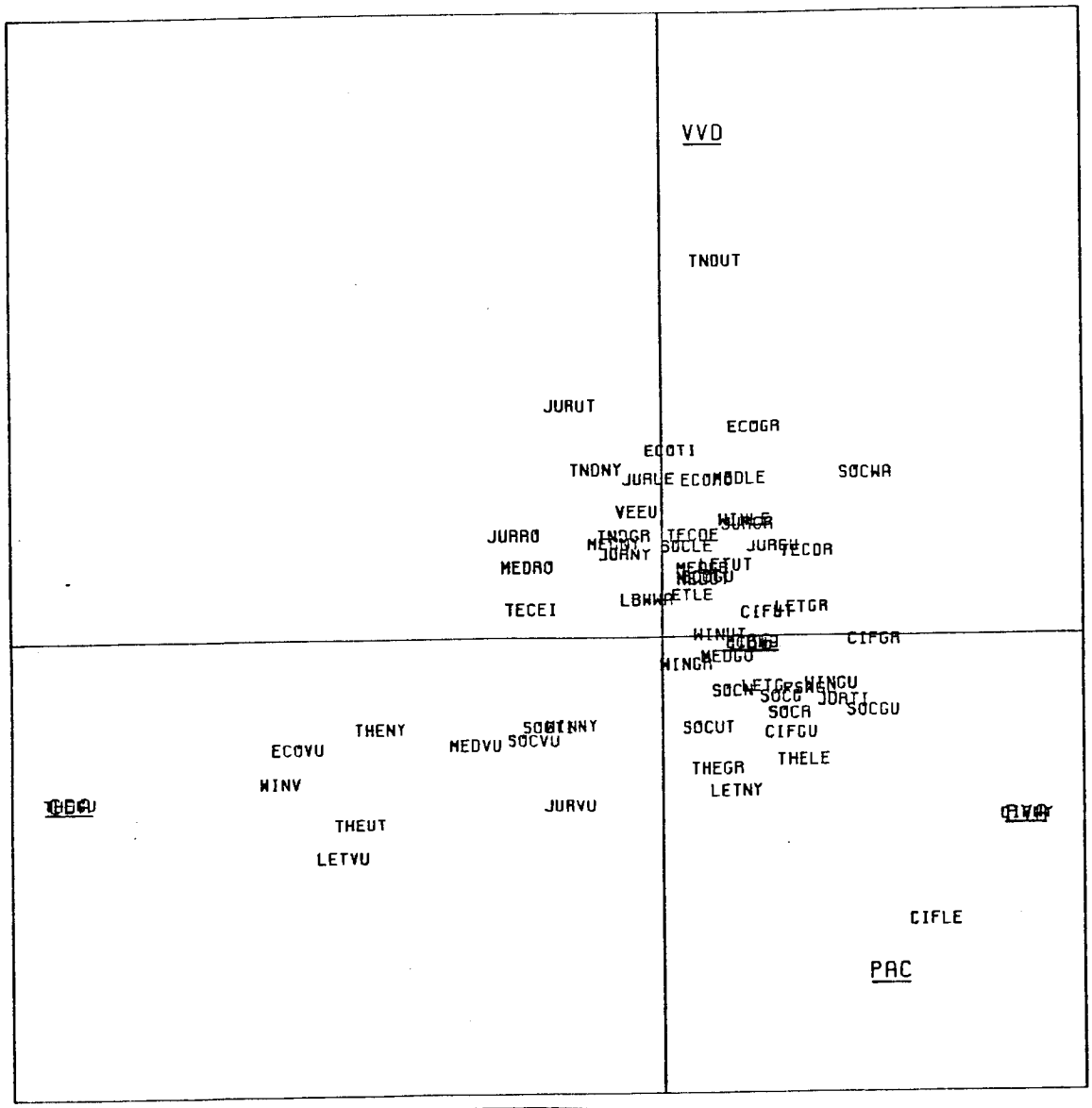


Figure 4.11 ANACOR solution for NSR data, with universities and faculties grouped as 63 categories of one variable.

A	Do you make it a rule to practice religious conduct, such as attending religious services, religious worship and missionary works and do you occasionally offer prayers or chant sutras?
B	Do you visit a grave once or twice a year?
C	Do you occasionally read religious books, such as the Bible or the Buddhist Scriptures?
D	Do you visit shrines and temples to pray for business prosperity, success in an entrance examination and so forth?
E	Do you keep a talisman, such as an amulet, charm or mascot near you?
F	Did you draw a fortune, consult a diviner or had you your fortune told within the last years?

Table 4.5 Sugiyama items.

1	1	1	1	1	1	042		0	1	1	1	1	1	011
1	1	1	1	1	0	033		0	1	1	1	1	0	007
1	1	1	1	0	1	006		0	1	1	1	0	1	002
1	1	1	1	0	0	017		0	1	1	1	0	0	005
1	1	1	0	1	1	012		0	1	1	0	1	1	004
1	1	1	0	1	0	029		0	1	1	0	1	0	008
1	1	1	0	0	1	008		0	1	1	0	0	1	004
1	1	1	1	0	0	082		0	1	1	0	0	0	044
1	1	0	1	1	1	051		0	1	0	1	1	1	072
1	1	0	1	1	0	069		0	1	0	1	1	0	126
1	1	0	1	0	1	020		0	1	0	1	0	1	045
1	1	0	1	0	0	054		0	1	0	1	0	0	142
1	1	0	0	1	1	034		0	1	0	0	1	1	080
1	1	0	0	1	0	124		0	1	0	0	1	0	258
1	1	0	0	0	1	027		0	1	0	0	0	1	137
1	1	0	0	0	0	317		0	1	0	0	0	0	760
1	0	1	1	1	1	001		0	0	1	1	1	1	000
1	0	1	1	1	0	002		0	0	1	1	1	0	002
1	0	1	1	0	1	000		0	0	1	1	0	1	000
1	0	1	1	0	0	009		0	0	1	1	0	0	004
1	0	1	0	1	1	001		0	0	1	0	1	1	004
1	0	1	0	1	0	011		0	0	1	0	1	0	003
1	0	1	0	0	1	007		0	0	1	0	0	1	006
1	0	1	0	0	0	059		0	0	1	0	0	0	030
1	0	0	1	1	1	008		0	0	0	1	1	1	033
1	0	0	1	1	0	023		0	0	0	1	1	0	048
1	0	0	1	0	1	007		0	0	0	1	0	1	038
1	0	0	1	0	0	035		0	0	0	1	0	0	064
1	0	0	0	1	1	010		0	0	0	0	1	1	042
1	0	0	0	1	0	055		0	0	0	0	1	0	096
1	0	0	0	0	1	013		0	0	0	0	0	1	090
1	0	0	0	0	0	194		0	0	0	0	0	0	718

Table 4.6 Sugiyama profile frequency matrix.

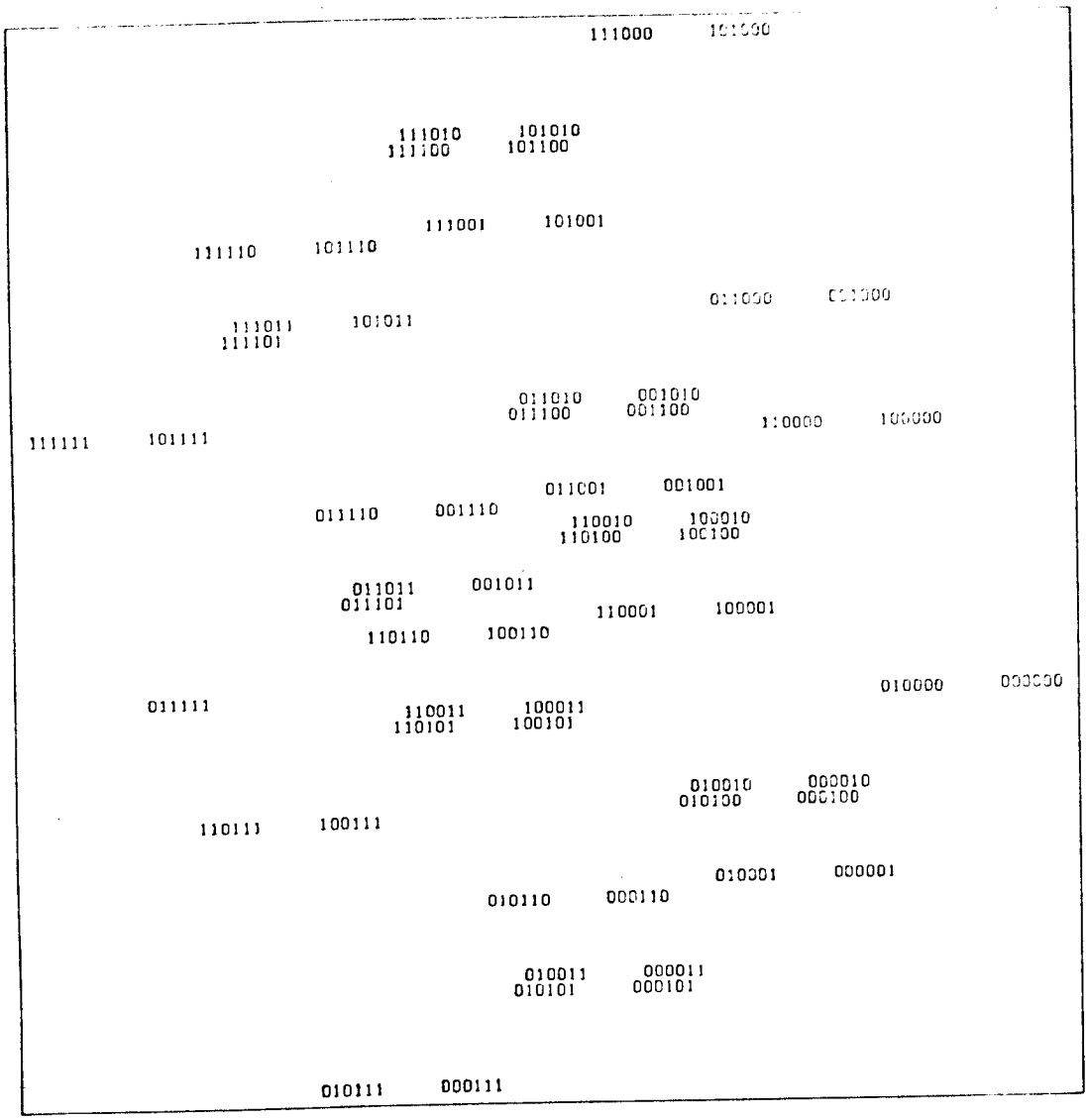


Figure 4.12 ANAPROF solution for Sugiyama data on assumption of complete indicator matrix

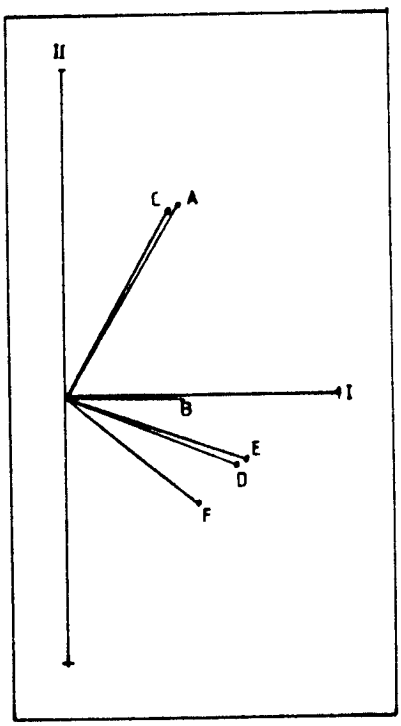


Figure 4.13 PCA solution for Sugiyama data, first two dimensions.

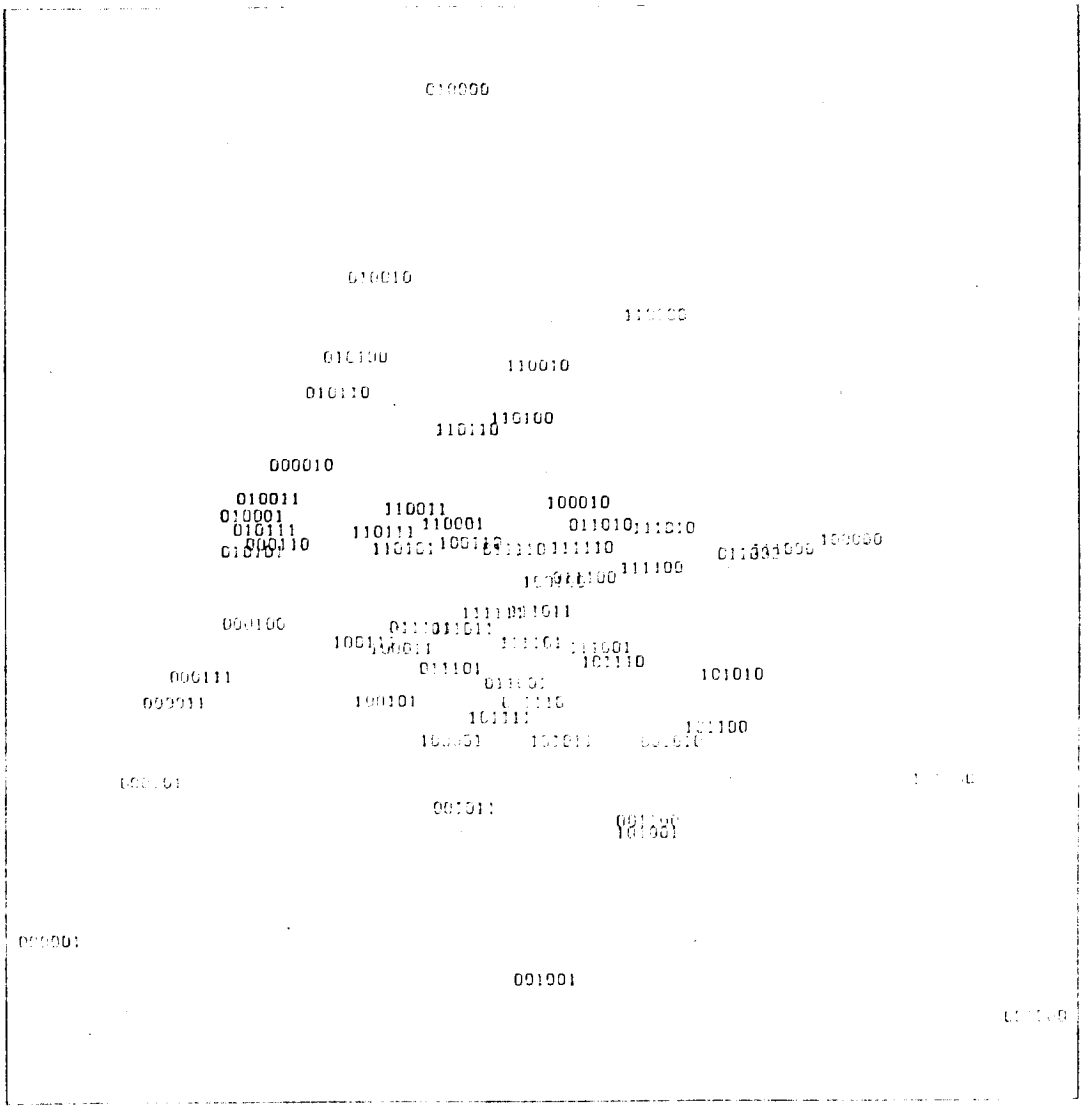


Figure 4.14. ANAPROF solution for Sugiyama data on assumption of incomplete data matrix

1.0000	.0853	.2842	.0770	.0971	-.0182
.0853	1.0000	.0522	.1119	.1631	.0614
.2842	.0522	1.0000	.0667	.0519	.0407
.0770	.1119	.0667	1.0000	.2785	.2111
.0971	.1631	.0519	.2785	1.0000	.2018
-.0182	.0614	.0407	.2111	.2018	1.0000

Table 4.7 Correlation matrix for Sugiyama items

A	B	C	D	E	F	DIA	GUT	CIR
1	1	1	1	1	1	✓	✓	
0	1	1	1	1	1	✓		
1	0	1	1	1	1			
1	1	0	1	1	1			
1	1	1	0	1	1			
1	1	1	1	0	1			
1	1	1	1	1	0	✓	✓	
0	0	1	1	1	1	✓		
1	0	0	1	1	1			
1	1	0	0	1	1			
1	1	1	0	0	1			
1	1	1	1	0	0	✓	✓	
0	1	1	1	1	0	✓		
0	0	0	1	1	1	✓		✓
1	0	0	0	1	1			✓
1	1	0	0	0	1			✓
1	1	1	0	0	0	✓	✓	✓
0	1	1	1	0	0	✓		✓
0	0	1	1	1	0	✓		✓
0	0	0	0	1	1	✓		
1	0	0	0	0	1			
1	1	0	0	0	0	✓	✓	
0	1	1	0	0	0	✓		
0	0	1	1	0	0	✓		
0	0	0	1	1	0	✓		
0	0	0	0	0	1	✓		
1	0	0	0	0	0	✓	✓	
0	1	0	0	0	0	✓		
0	0	1	0	0	0	✓		
0	0	0	1	0	0	✓		
0	0	0	0	1	0	✓		
0	0	0	0	0	0	✓	✓	

Table 4.8 Permissable response patterns for balloon scale

dimensions of the solution. The plot contrasts patterns with a 'yes' answer to items A and C (left upper) to those with a 'no' answer to these two items (right lower). The plot also contrasts patterns with 000 for the last three items (right upper) to those with 111 for these items (left lower); patterns with only one 1 and those with two '1's form bands between the two extremes. Item B is treated in a particular way: they form alternating bands.

The optimally scaled data matrix Q becomes an 4243 x 6 matrix, and its correlation matrix is geven in table 4.7, for which figure 4.13 plots the first two dimensions of the PCA factor solution. This plot confirms the similarity between A and C, and that between D,E,F, whereas item B is not very salient.

Actually, the example is a bit misleading because the correlation matrix of table 4.7 for binary variables does not depend at all on the optimal scaling: for binary variables any scaling produces the same correlations (apart from a possible change of sign). ANAPROF, for binary variables, therefore, gives the same solutions as PCA on the correlations for the columns of G.

4.5.3

It then becomes interesting to compare the solution in 5.4.2 with a solution for the incomplete data matrix G of dimension 4243 x 6 (one column only for each item, registering "yes" as '1'). Such a solution would be more in the spirit of unfolding (section 4.2). The solution, for the first two dimensions, is plotted in figure 4.14. The corresponding eigenvalues are ψ_1^2=.425, ψ_2^2=.348.

Figure 4.14 looks rather different from figure 4.12. If we take profiles with only one 1 as the positions of the items (100000 = A, etc.), we note that the items form a curved scale. An item with more '1's in it, is located at the exact centroid of the patterns with only one '1' at the same positions. E.g., 000101 is exactly midway between 000100 and 000001, pattern 111111 is the unweighted center of gravity of all six item patterns with only one 1. The unfolding structure is reflected in the fact that each pattern is closer to those elementary patterns of which it is the mean, than to other elementary patterns.

4.6 Back to MDS

4.6.1

The different results in 4.5.2 and 4.5.3 bring us back to the context of MDS. In sections 2.4.3 and 3.11.3 the distinction between Guttman scale and Coombs scale was brought in relation with the choice between a complete or an incomplete indicator matrix. There are many more types of scale. This is

illustrated in table 4.8 which shows for 6 items all 32 possible 'circular' response patterns. Imagine table 4.8 pasted on a cylinder, so that the column for the last item becomes adjacent to the column for the first. A pattern is 'circular' if on the cylinder it has no more than one run of '1's and no more than one run of '0's. In fact, the 32 patterns of table 4.8. all have just one run of '1' and just one run of '0', except for the patterns 111111 and 000000.

From the table we can select subsets of rows that fit to special cases of scaling. If we select the rows with a checkmark in the column GUT of table 4.8 we obtain items which obey a Guttman scale ordering. There are many ways to select seven items in agreement with a Guttman scale ($mx2^{m-2}=96$).

The rows checked in the column CIR form a "circumplex" of order three (all patterns that can be obtained by moving a run of three adjacent '1's around the cylinder).

The rows checked in column DIA have a "diamond" scale. Its structure is shown in figure 4.15 as a graph. The 21 patterns (pattern 000000 is left out) are ordered on the principle that at each level of the graph a pattern is the union of the two patterns directly below.

All rows of table 4.8 together form a "balloon" scale. The graph is given in figure 4.16. All previous types of scale are subsets in this graph.

4.6.2

The HOMALS solution for the balloon scale is figured in 4.17. The first dimension goes from 'North pole' (111111) to 'South pole' (000000). The graph now appears as if stretched over the surface of a sphere. Guttman scales appears as "parallels" on the surface, circumplexes appear as "meridians".

It now becomes possible to recognize in the solution for the Sugiyama data of 4.5.2 the diamond scale. Figure 4.18 shows this for the 22 response patterns that fit a graph with 'bottom' row CABEDF (rotated 90 degrees) in agreement with the ordering in figure 4.12 of the six elementary patterns from 000001 to 001000. This selection of patterns covers 3333 individuals (78% if we include the zero pattern). Figure 4.19 shows a comparable plot for the "unfolding" solution that requires 111111 to become the center of the plot and to this end folds the elementary points from 000001 to 001000 into a curve at the periphery of the plot. Note that 000000 does not appear in this plot, sice the HOMALS solution treats individuals with this pattern as "missing data deleted".

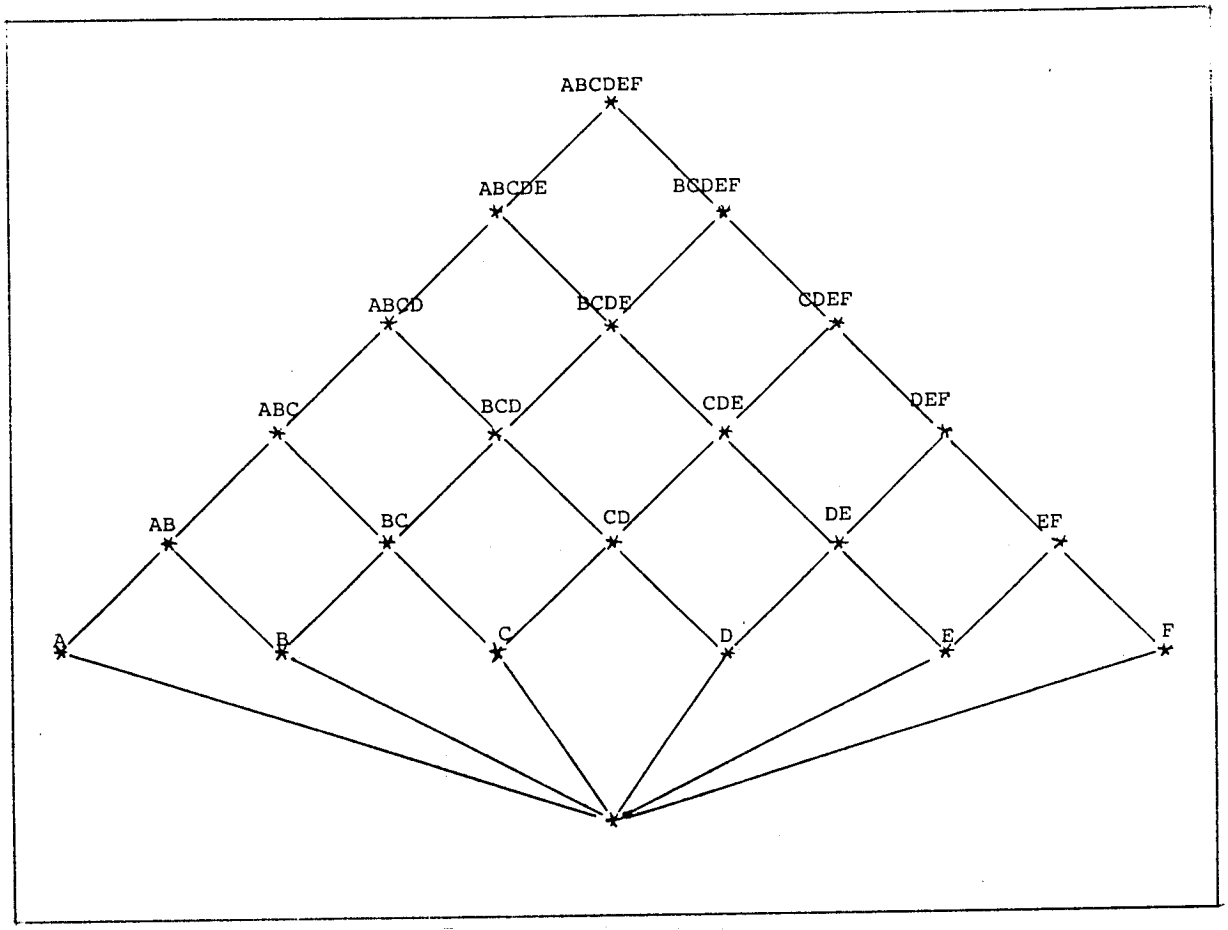


Figure 4.15 Graph of diamond scale

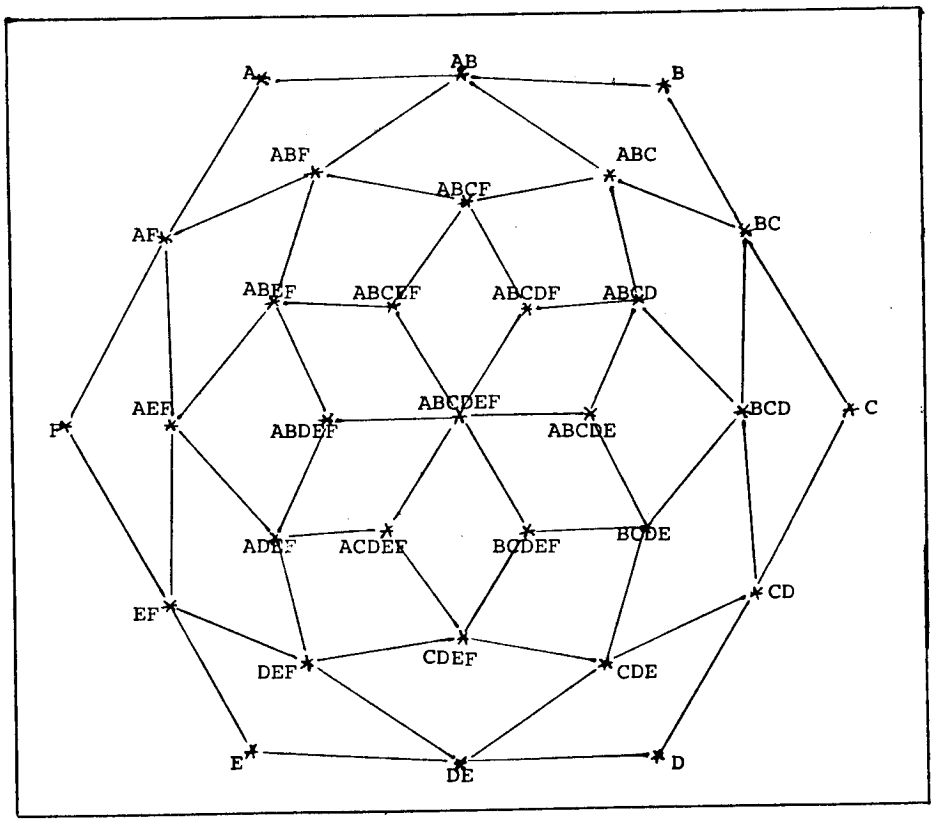


Figure 4.16 Graph of balloon scale

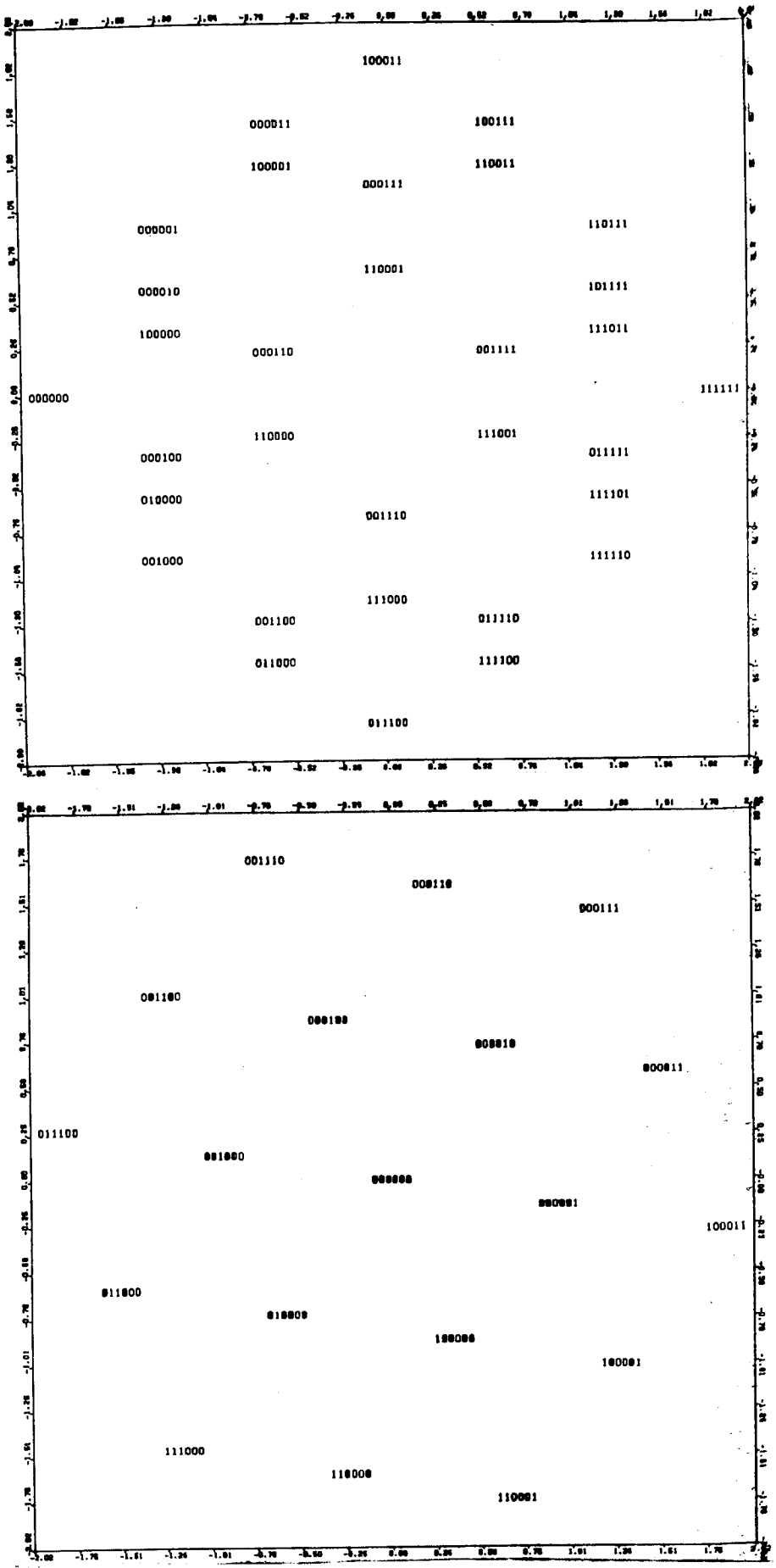


Figure 4.17 HOMALS solution for response patterns of table 4.8.
The upper figure shows dimension 1 and 2;
the lower figure shows dimension 2 and 3

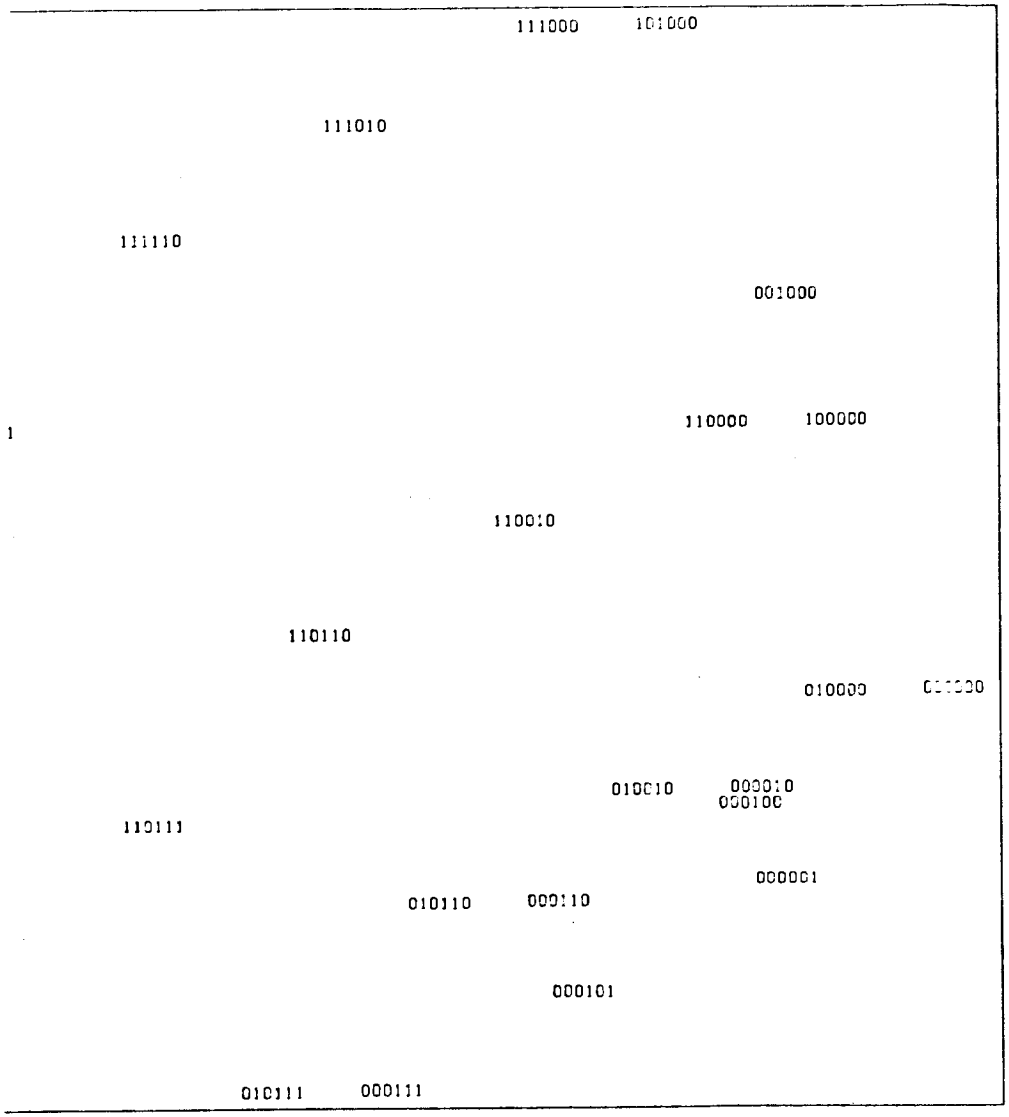


Figure 4.18 Diamond structure fitted to Sugiyama data, complete indicator matrix

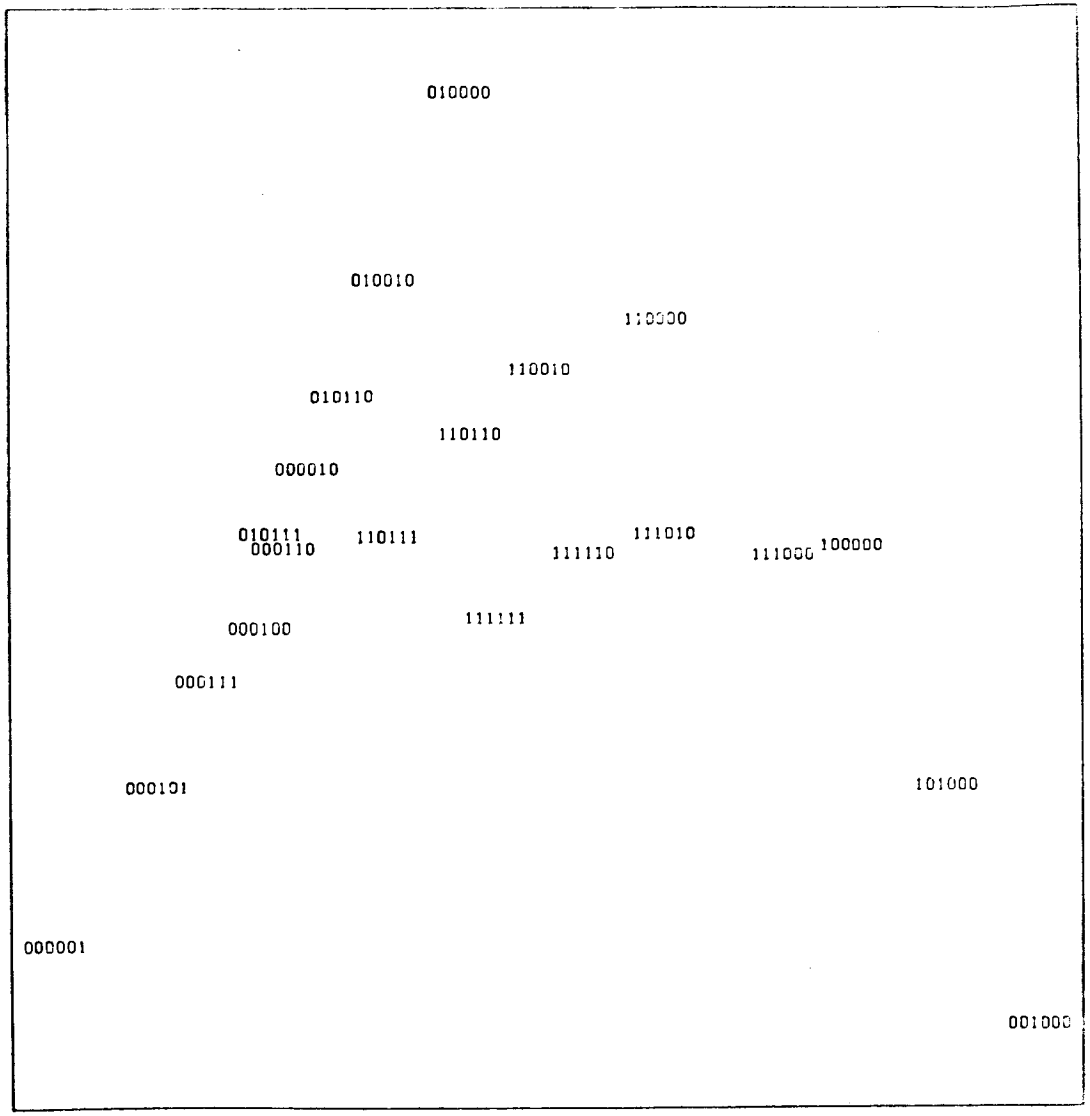


Figure 4.19 Diamond structures fitted to Sugiyama data, incomplete indicator matrix.

5 Nonmetric principal components analysis and PRINCALS

5.1 History

5.1.1 Metric principal components analysis

In chapter 3 we reviewed some of the history of homogeneity analysis, which defines one particular way to introduce principal components analysis. We have seen that Pearson introduced a 'dual' form of homogeneity analysis in 1901, and that ordinary PCA, in terms of using the principal axes of the correlation ellipsoid to compute optimal index characters, was also invented by Pearson around the same time (cf MacDonell, 1901/1902). The first systematic account of principal components analysis as a data analytic technique was Hotelling (1933), and his particular form of 'factor analysis' was introduced to statisticians by Girshick (1936). Both Hotelling and Girshick used homogeneity ideas as their starting point. They looked for the linear composite with the largest variance, or the linear composite with maximum sum of squared correlations with the variables. Additional components were introduced by looking for the second best solution under orthogonality constraints. This line of reasoning was taken up by Horst (1936), Edgerton and Kolbe (1936), Wilks (1938), and Guttman (1941). It leads directly to our form of homogeneity analysis, and to the computer program HOMALS.

At about the same time Eckart and Young (1936) introduced a natural way to define principal components analysis for a general p, i.e. to formulate optimality properties which hold for the first p components simultaneously. This is a more natural way to introduce multidimensional solutions than the earlier successive procedures, although it turns out that in the simple situations investigated by these early authors simultaneous and successive multidimensionality lead to the same solution (or, as we sometimes say, the solutions for different dimensionalities are nested). Eckart and Young used the least squares properties of the singular value decomposition (also discussed in our appendix), which were already described in a more general context by Schmidt (1906). The singular value decomposition itself was discovered, independently, by Beltrami (1873), Jordan (1874), and Sylvester (1889). Simultaneous and successive optimality conditions of principal components are reviewed by Rao (1964), Okamoto (1968), Le Roux and Rouanet (1979), Rao (1980). The Le Roux and Rouanet paper is a beautiful 'French' introduction to principal components analysis, Rao (1980) extends the approximation properties of the singular value decomposition to a more general class of matrix norms.

The Eckart-Young theorem tells us that principal components analysis of an $n \times m$ data matrix H can be formulated in terms of the loss function

$$\sigma_J(X,A) = SSQ(H - XA'),$$

where X is $n \times p$ of rank p, and where A is $m \times p$. We use the symbol σ_J because $\sigma_J(X,A) = 0$ for some X and A means, in the terminology of 1.1.8, that the join-rank of H is less than or equal to p. The appendix explains that the optimal X and A,

that minimize $\sigma_J(X,A)$, can be found by computing the singular value decomposition of H, and by retaining the largest p singular values with corresponding singular vectors. Another possible computational procedure is alternating least squares, in which we alternate the steps

$X \leftarrow HA(A'A)^+$,

$A \leftarrow H'X(X'X)^+$,

where the superscript + indicates the Moore-Penrose inverse of a matrix. This computational procedure is decribed by Daugavet (1968).

We emphasize that principal components analysis is usually not interpreted in terms of fitting a statistical model. The model $h_j = \sum x_s a_{js}$, implicit in the Eckart-Young formulation, is equivalent with the model that the rank of the covariance matrix of the h_j is less than or equal to p. If we assume a multinorma sample, then statistical theory tells us to reject this model whenever the observed sample dispersion matrix has rank larger than p. The performance characteristics of this test are very satisfactory. It is impossible to reject the model if it is true, and the probability of accepting it when it is false tends to zero exponentially fast. In practical situations, however, the test means that we shall always reject the model, which is not a very satisfactory situation. Lawley and Bartlett have studied a more realistic model, in which we suppose that the smallest m - p eigenvalues of the covariance matrix are equal (but not necessarily equal to zero). This corresponds with $h_j = \sum x_s a_{js} + e_j$ where the errors e_j are independent and have equal variances. This model, however, is better interpreted as a special case of the factor analysis model, and not as principal components analysis.

5.1.2 Non metric principal components analysis

The word 'nonmetric' in nonmetric principal components analysis could be somewhat confusing. In the history of psychological scaling we can distinguish three different uses of nonmetric. In the first place there are nonmetric data. Preference rank orders are for example usually thought of as nonmetric data, although it is of course possible to code them metrically, for example by using rank numbers or order statistics. The idea is, however, that we are only supposed to use the ordinal properties of the data in arriving at our representations. This does not mean that it is forbidden, by some mysterious central directorate for the administration of scale levels, to correlate rank numbers or paired comparisons. Of course such a forbid would not make sense, correlating rank numbe gives Spearman's ρ and correlating paired comparisons gives Kendall's τ, both of which are perfectly respectable. The idea is that if the data 'are' nonmetric, then we should first transform them to rank numbers or paired comparisons or some other set of invariant conventional numbers before we correlate them, and thus make the correlation coefficient reflect the ordinal properties of the data

only. We have used nonmetric here as almost the same thing as ordinal. This is somewhat restrictive, although it is quite common. But clearly nominal data are nonmetric too. We are not sure if missing data are nonmetric. Observe that this use of the word is consistent with our definition of nonlinear and ordinal MVA in chapter 1.

The second use of the word nonmetric is somewhat more specialized. The older scaling programs for nonmetric data produced nonmetric representations. Guttman scaling, for example, provided a rank order of the items and individuals, the unfolding theory of Coombs gave a rank order of objects (the common J-scale) only. These older algorithms are very ably reviewed in the book of Coombs (1964), who also presents similar algorithms for multidimensional unfolding and nonmetric multidimensional scaling, which rank the objects on a number of dimensions. In what is usually called the 'nonmetric breakthrough' or the 'computational breakthrough' in psychological scaling Shepard, Guttman, and Kruskal showed that it was actually possible in some circumstances to derive metric representations from nonmetric data. Or, more precisely, stable metric representations. This breakthrough produced a lot of enthusiasm at the time, but after twenty years we start to find out that it does not work very well in many situations (compare section 5.2.4 in this chapter). This makes even older approaches, due to Thurstone and Guttman, which already derived metric representations from nonmetric data by introducing stronger requirements, more interesting. As we have seen in chapter 4 this is basically how HOMALS fits into the scaling tradition.

Finally nonmetric is sometimes used for error theories and loss functions. If we have a model that is formulated in terms of a finite number of inqualities that must be true (or even a finite number of equations that must be true) then we can measure loss by counting the number of inequalities and/or equations that are not true for a particular representation, and we can find our representation in such a way that the number of violations is minimized. Such nonmetric error theories are particulary attractive in combination with nonmetric data, but they present us with considerable computational problems. The algorithms outline in Coombs (1964) were nonmetric in all three sense of the word, and could consequently be called purely nonmetric. Since the computational breakthrough the word is used almost exclusively in the first sense we discussed.

We start our brief account of the history of nonmetric principal components analysis with a quotation from Thurstone. "One of the principal assumptions underlying factorial theory is that the scores are monotonic increasing or decreasing functions of the scores on the primary factors or parameters. The fundamental observation equation of factor analysis makes the further assumption that these functions can be expressed in linear form as a first approximation. It would be possible to start with a second-degree observation equation and to develop factorial methods on that

basis. Instead of developing factorial theory more completely with observation equations of higher degree, it would probably be more profitable to develop non-me methods of factor analysis. An idea for such a development would be to determine the number of independent parameters of a score matrix by analyzing successive differences in rank order on the assumption that they are monotonic functions of a limited number of independent parameters. A score would then be regarded as mere an index of rank order, and that is essentially what we are now doing. The raw scores are transmuted into a normal distribution of unit standard deviation, and these transmuted scores are used for the correlations. Instead of dealing with the transmuted scores in this manner, one might deal with the rank orders directly or with some equivalent indices of rank order." (Thurstone, 1947, p xiii-xiv). There are no further contributions to this problem in Thurstone's book. The next contribution was Bennett (1956). This is in the purely nonmetric tradition of Coombs and his school. He determines the number of distinct ways in which n subjects can be ranked by linear functions of p factors, and uses these combinatorial results to bound the number of factors in the complete case, when all possible linear functions (tests) have been used. This is clearly not very relevant for practical data analysis, although it does tell us some interesting theoretical properties of nonmetric components analysis. Guttman (1959) is another important contribution. He adds the additional restriction that the regressions between transformed variables must be linear, and shows that there is at most one possible transformation of the variables which linearizes regressions and is monotonic. This is easy to see if there are just two variables Linearity of regression defines the stationary equations of correspondence analysis (Hirschfeld, 1935). We know, (section 4.3), that the number of solutions to these equations is equal to the number of categories of the variable with the smallest number of categories, and that the solutions are orthogonal. Orthogonality implies that if one of them is monotone, then the other certainly are not monotone. Guttman discusses one of the possible extensions of this result to more than two variables. His conclusion is interesting. "Since the metrics of observed test scores are usually arbitrary, Thurstone posed the problem of how to 'factor' them by using only rank-order considerations. One form of solution is to seek transformations that will yield new scores with a correlation matrix that is best from some point of view of factor analysis. But analysing data via their correlation matrix is justified stochastically only if the regressions are linear. Assuming only the linearity restriction on regressions, it is shown that -in general- at most one set of new scores can be found to maintain the observed rank-orders. The factor analyst has no freedom to mould the new correlation matrix by further considerations." (Guttman, 1959, summary).

In the early sixties the 'computational breakthrough' in nonmetric scaling was accomplished by Shepard and Kruskal. Shepard showed that ordinal constraints, at least in some situations, constrain the solution almost as tightly as linear constraints, and he showed that metric solutions can be obtained from ordinal data by the computer. Both ideas were revolutionary at the time, the easiest way to appreciate this is to compare them with the techniques discussed in the book by Coombs (1964). Kruskal emphasized explicit loss functions, showed that any differentiable loss function could be minimized by gradient methods, and introduced monotone regression. The introduction of monotone regression made it possible to construct nonmetric versions of all metric scaling methods, and the stability and gauging results of Shepard seemed to indicate that these new programs actually amounted to getting something (stable metric representations) from nothing (merely ordinal data).

A working program for nonmetric principal components analysis existed in 1962, in Shepard (1966) we find the first published stability results, but the basic paper was published only in 1974 by Kruskal and Shepard. The earliest published account is Roskam (1968, ch 5). Both Kruskal and Shepard use basically the same approach. They fit what they call the 'linear model' (actually 'bilinear model' is a better name) to a rectangular data matrix H. The model is defined by

$$q_{ij} = \sum_{s=1}^{p} x_{is} a_{js},$$

$$h_{ij} > h_{kj} \rightarrow q_{ij} \geq q_{kj},$$

where i,k=1,...,n and j=1,...,m. Thus each column of the data matrix (each variable) must be transformed monotonically in such a way that the model with p components fits the transformed data. The dimensionality p is important here, we can no longer expect the solutions to be nested, because choosing a different dimensionality will lead to different transformations. Another point which we must remember are that the data are interpreted as column-conditional, by which we mean that only order relations within columns are imposed. In general the elements in a column may only be partially ordered, thus there may be missing data which do not impose order constraints. The model does not specify what should be done if $h_{ij} = h_{kj}$, thus equalities in the data do not impose restrictions. This is called the primary approach to ties in Kruskal (1964a,b) or De Leeuw (1977) it is called continuous ordinal data in De Leeuw, Young, Takane (1976). If we add the restrictions

$$h_{ij} = h_{kj} \rightarrow q_{ij} = q_{kj},$$

then this defines the secondary approach or discrete ordinal data. We shall come back to this distinction later, for the moment we write the restrictions imposed by the order relations in the data in the simple form $q_j \in K_j$, with K_j a convex cone in n-space.

The loss function used by both Kruskal and Shepard and by Roskam is, except for some irrelevant details,

$$\sigma_J(Q,X,A) = \frac{1}{m} \sum_{j=1}^{m} SSQ(q_j - Xa_j) \,/\, SSQ(q_j - AVE(q_j)),$$

which is minimized over all X, A, and over all $q_j \in K_j$ (j=1,...,m). The notation in this formula deserves some attention here. The matrix A is m × p, the vector a_j has p elements, and there are m such vectors. Thus we mean that a_j is row j of A, written as a column. This sounds complicated, but no confusion is possible, and the notation is considerably simpler than something less ambiguous such as $a'_{j\rightarrow}$. We also remember (cf appendix B) that AVE(.) is the mean of a vector or the expected value of a random variable. Because of the form of σ we call it a normalized loss function, the denominator is the normalization factor. For each j the corresponding component of the loss is the variance of the residuals divided by the variance of the transformed data. The function is minimized by gradient methods, combined with monotone regression. A detailed account of the algorithms can also be found in Hartmann (1979, ch 4).

Subsequent contributions to nonlinear principal components analysis are Roskam (1977), Tenenhaus (1977), Young, Takane, and De Leeuw (1978). They all use the same loss function, but they differ in the types of data they can handle or in the algorithm. Tenenhaus assumes that variables are either numerical or nomi Young, Takane, and De Leeuw can handle any mixture of nominal, ordinal, and numerical variables. All three programs (called, respectively, MNNFAEX, PRINQUA and PRINCIPALS) use alternating least squares methods to minimize the loss function. They alternate transformation of the data and fitting of the transfor data by computing a partial or complete singular value decomposition. Another difference with the earlier programs is that PRINQUAL and PRINCIPALS use explic normalizations. They minimize

$$\sigma_J(Q,X,A) = \frac{1}{m} \sum_{j=1}^{m} SSQ(q_j - Xa_j),$$

over X and A and over q_j satisfying $q_j \in K_j$, $AVE(q_j) = 0$, $SSQ(q_j) = 1$. The homogeneity of the bilinear model proves that the problem with explicit normaliz and the problem with normalized loss functions (or: implicit normalizations) are equivalent. Using explicit normalizations leads to more compact formulas, and generally seems slightly more elegant.

It must be emphasized that all programs can also be applied to row-conditional data by simply transposing the data matrix. A very important application is to preference rank orders, in which a number of persons rank a number of objects with respect to preference (or utility, or beauty, or whatever). Tucker (1960) proposed a classical model which assumes that each person defines a direction in p-dimensional space, each object defines a point in the same space, and

preference strength for an object-person combination is the length of the projection of the object point on the person direction. In Coombs (1964) the obvious nonmetric extension of this model is discussed. If the person direction is a_j, with $SSQ(a_j) = 1$, and the object point is x_i, then

$$q_{ij} = \sum_{s=1}^{p} x_{is}a_{js}$$

is the predicted preference strength of person j for object i. If h_{ij} is the observed preference strength, then the model requires

$$h_{ij} > h_{kj} \rightarrow q_{ij} \geq q_{kj}.$$

This is clearly exactly identical to the nonmetric principal components model, but the comparisons are now within individuals. If the data are collected in an individuals × objects matrix, then they are row-conditional, and we have to transpose the matrix first before we can apply the programs discussed above. We shall come back to this important application in our examples.

5.2 Theory

5.2.1 Properties of join-loss

In this section we study some properties of $\sigma_J(Q,X,A)$, using explicit normalizatio The easiest way to do this is to define $\sigma_J(Q,*,*)$, which is the minimum of $\sigma_J(Q,X,$ over X and A for fixed Q. By the Eckart-Young theorem

$$\sigma_J(Q,*,*) = \frac{1}{m}\sum_{s=p+1}^{m} \lambda_s(R(Q)),$$

where R(Q) stands for the correlation matrix of the m vectors q_j, and the λ_s are its eigenvalues, in decreasing order. Clearly

$$0 \leq \sigma_J(Q,*,*) \leq 1 - \frac{p}{m},$$

with $\sigma_J(Q,*,*) = 0$ if and only if $\text{rank}(R(Q)) \leq p$, and $\sigma_J(Q,*,*) = 1 - \frac{p}{m}$ if and onl if R(Q) is the identity matrix.

Minimizing $\sigma_J(Q,*,*)$ over $q_j \in K_j$ with $AVE(q_j) = 0$ and $SSQ(q_j) = 1$ means that we transform or quantify our variables in such a way that the sum of the m-p smallest eigenvalues is minimized, or, equivalently, that the sum of the p largest eigenvalues of R(Q) is maximized.

5.2.2 Relations with homogeneity analysis

If we want to compare this form of component analysis with homogeneity analysis discussed in chapter 3, we must suppose in the first place that all variables are discrete and nominal. Thus the cones K_j are subspaces defined by

$$K_j = \{q_j \mid q_j = G_j y_j\},$$

with the G_j complete indicator matrices. If p = 1 then the theory of the previous section tells us that component analysis amounts to finding the y_j in such a way

that the largest eigenvalue of the correlation matrix of the $q_j = G_j y_j$ is maximized, and we have already seen in chapter 3 that this is one of the ways in which (one-dimensional) homogeneity analysis can be defined. Thus for $p = 1$ and all variables discrete and nominal the two techniques are equivalent.

If $p = m-1$, then we want to minimize the smallest eigenvalue of the correlation matrix. This is related to the 'dual' form of homogeneity analysis introduced by Pearson in 1901, and discussed briefly in section 3.6. We know from chapter 3 that the generalized eigenvalue problem corresponding with a homogeneity analysis has one trivial eigenvalue equal to one, $m - 1$ trivial eigenvalues equal to zero, and $\sum(k_j - 1)$ nontrivial solutions. In the same way as we show that the largest nontrivial eigenvalue in homogeneity analysis corresponds with the components solution for $p = 1$, we can also show that the smallest nontrivial eigenvalue in homogeneity analysis corresponds with the principal components solution for $p = m - 1$.

For intermediate values of p the situation is more complicated. One useful way of looking at the problem is as follows. We distinguish between the dimensionality p, that we have defined already in terms of the number of columns of X and A in the loss function, and the number of successive solutions r. In homogeneity analysis we have $p = 1$. We compute an optimal solution $\hat{q}$, the next step is to compute a second solution, still minimizing σ_J with $p = 1$, but imposing the additional orthogonality requirement that

$$\sum_{j=1}^{m} \hat{q}_j' q_j = 0.$$

In general we can compute additional solutions by requiring that they must be orthogonal with all previous solutions. Thus homogeneity analysis has $p = 1$ and $r \geqslant 1$, and it yields multiple quantifications of the variables. Components analysis on the other hand, has $p \geqslant 1$ and $r = 1$. We minimize σ_J for a given value of p, and then we are done. Thus components analysis gives a single quantification.

It is, of course, possible to combine the two and construct a technique with both $p \geqslant 1$ and $r \geqslant 1$. This amounts to applying components analysis r times, each time imposing an extra orthogonality constraint. We do not have any experience with this combined technique, however. It is of some interest that the definition of successive solutions implies automatically that the techniques are nested with respect to different values of r, although components analysis is usually not nested with respect to different values of p.

We can illustrate some of these points with the example of section 3.8. Not all of them, however, because in this example $m = 3$. Consequently we can only choose $p = 1$ and $p = m - 1 = 2$ here. There are 5 nontrivial solutions to the generalized eigenvalue problem of homogeneity analysis. The r successive solutions to the components analysis problem with $p = 1$ correspond with the r largest nontrivia

eigenvalues of the homogeneity problem, the r successive solutions to the components analysis problem with p = m - 1 correspond with the r smallest nontrivial eigenvalues. In table 5.1 we give the solution for Q with p = 1, with the corresponding correlation matrix, and the eigenvalues of $\frac{1}{m} R(Q)$. In table 5.2 similar results for p = 2 are given.

5.2.3 Relationships with ordinary components analysis

If the data are numerical, and we require that the transformations are all linear, then

$$K_j = \{q_j \mid q_j = \alpha_j h_j + \beta_j\}.$$

Without loss of generality we can suppose that $SSQ(h_j) = 1$ and $AVE(h_j) = 0$ for all j. Because we require that $SSQ(q_j) = 1$ and $AVE(q_j) = 0$, it follows that

$$\sigma_J(Q,X,A) = \frac{1}{m} \sum_{j=1}^{m} SSQ(h_j - Xa_j) = \frac{1}{m} SSQ(H - XA).$$

There is no freedom for choosing Q different from H, and consequently the analysis amounts to computing eigenvalues and eigenvectors of R(H).

5.2.4 Continuous variables

In De Leeuw, Young, and Takane (1976) and De Leeuw and Van Rijckevorsel (1980) a system of measurement and process levels is discussed, which can be used to define many different types of cones K_j. In this book we assume generally that data are categorical, and that the number of categories of a variable is usually much less than the number of observations. This has some interesting consequences.

Suppose, for example, that we define all K_j by the continuous ordinal option

$$h_{ij} > h_{kj} \rightarrow q_{ij} \geq q_{kj}.$$

We investigate when we can choose the $q_j \in K_j$ in such a way that rank(R(Q)) = 1, in this case we say that a perfect solution exists. Assume for convenience that the h_{ij} are integers, satisfying $1 \leq h_{ij} \leq k_j$. Suppose the n individuals are a random sample from a multivariate discrete population, in which

$$\pi_0 = \text{prob}(\underline{h}_1 = 1 \ \& \ \ldots \ \& \ \underline{h}_m = 1),$$

$$\pi_1 = \text{prob}(\underline{h}_1 = k_1 \ \& \ \ldots \ \& \ \underline{h}_m = k_m),$$

with $\pi_0 + \pi_1 > 0$. If the variables are binary items, with wrong-correct interpretatic for example, then this suppose that there are individuals in the population which will answer all items correctly or all items incorrectly. Suppose $\underline{n}_0$ is the number of individuals in the sample with $h_{ij} = 1$ for all j, and $\underline{n}_1$ is the number of individuals in the sample with $h_{ij} = k_j$ for all j. If $\underline{n}_0 + \underline{n}_1 > 0$, then a perfect solution can easily be constructed. Thus P_n, the probability that a perfect solution exists in a random sample of size n, satisfies

$$P_n \geq \text{prob}(\underline{n}_0 + \underline{n}_1 > 0) = 1 - (1 - \pi_0 - \pi_1)^n.$$

0.69	0.34	1.53
-1.88	-3.00	-0.65
0.69	0.26	-0.65
0.69	0.34	1.53
-1.88	0.34	-0.65
-0.20	0.34	-0.65
0.69	0.34	1.53
0.69	0.34	-0.65
-0.20	0.34	-0.65
0.69	0.34	-0.65

table 5.1.a:
matrix Q for p = 1,
normalized by Q'Q = nI

0.75	-0.04	-1.51
-1.73	2.35	0.65
0.75	-2.12	0.65
0.75	-0.04	-1.51
-1.73	-0.04	0.65
-0.53	-0.04	0.65
0.75	-0.04	-1.51
0.75	-0.04	0.65
-0.53	-0.04	0.65
0.75	-0.04	0.65

table 5.2.a:
matrix Q for p = 2,
normalized by Q'Q = nI

1.000		
0.622	1.000	
0.453	0.224	1.000

table 5.1.b:
R(Q) for p = 1

1.000		
-0.570	1.000	
-0.494	0.019	1.000

table 5.2.b:
R(Q) for p = 2

0.629
0.263
0.108

table 5.1.c:
eigenvalues of $\frac{1}{m}R(Q)$
for p = 1.

0.588
0.327
0.085

table 5.2.c:
eigenvalues of $\frac{1}{m}R(Q)$
for p = 2.

Note on tables 5.1 and 5.2
The eigenvalues of the corresponding homogeneity analysis are
0.629
0.426
0.389
0.139
0.085
The largest one is the largest one in 5.1.c, the smallest one is the smallest one in 5.2.c.

Thus $P_n \to 1$ if $n \to \infty$, which implies that the minimum of σ_j converges almost surely to zero if $n \to \infty$. Because all measurements are discrete in the last analysis, this result applies quite generally. It shows that the continuous ordinal option only works if the sample is small, in large samples it will almost surely find a 'trivial' perfect solution. This is not satisfactory at all, especially because P_n is generally much larger than the lower bound we have derived. Similar objections can be raised against the other continuous options mentioned in De Leeuw and Van Rijckevorsel (1981). The discrete options, which require in addition that

$$h_{ij} = h_{kj} \to q_{ij} = q_{kj}$$

are much more restrictive, and have the additional advantage that the transformation $h \to q$ is really a function in the mathematical sense.

5.2.5 Use of indicator matrices

The discrete options make it possible to use indicator matrices in the obvious way. It is then more convenient to write σ_j as a function of the category quantifications y_j. Thus

$$\sigma_j(Y,X,A) = \frac{1}{m}\sum_{j=1}^{m} SSQ(G_j y_j - Xa_j),$$

which must be minimized under the conditions

$$u'G_j y_j = 0,$$
$$y_j' D_j y_j = 1,$$
$$y_j \in K_j,$$

where K_j is now a cone in k_j-dimensional space, usually $k_j << n$.

We can now describe the algorithms of MNNFAEX, PRINQUAL, and PRINCIPALS in more detail. They are of the alternating least squares type, with two kinds of steps. Suppose we start an iteration with an estimate of the y_j satisfying the constraints. We then compute $q_j = G_j y_j$ and minimize SSQ(Q - XA') over X and A by using the Eckart-Young theorem. This is the first step. In the second step we compute new y_j for given X and A. This can be done for each j separately. Define

$$\tilde{y}_j = D_j^{-1} G_j' X a_j.$$

Then

$$SSQ(G_j y_j - Xa_j) = SSQ(G_j \tilde{y}_j - Xa_j) + (y_j - \tilde{y}_j)' D_j (y_j - \tilde{y}_j),$$

and consequently we have to minimize the second term on the right over all y_j satisfying the constraints. This is a normalized cone projection problem, for which the relevant theory is treated in an appendix. We then go back to the first step, and so on. In table 5.3 we illustrate two iterations of the algorithm, applied to the example in 3.8.

Alternatively we can also use the Daugavet-algorithm mentioned in 5.1 to construct a three-step alternating least squares method. The algorithm starts

Y	-.237	.158	.558	-.240	.178	.543	-.243	.197	.532
	-.148	.346	.839	-.146	.316	.854	-.144	.285	.868
	-.483	.207		-.483	.207		-.483	.207	
R	1.000			1.000			1.000		
	-.156	1.000		-.158	1.000		-.161	1.000	
	.491	.307	1.000	.498	.303	1.000	.504	.298	1.000
eval	.506	.379	.114	.508	.379	.113	.509	.380	.112
A	.625	-.498		.631	-.486				
	.249	.856		.237	.863				
	.739	.132		.738	.139				
X	-.542	-.073		-.543	-.076				
	.338	.245		.340	.215				
	.214	.865		.203	.882				
	-.542	-.073		-.543	-.076				
	.215	-.178		.231	-.184				
	.462	-.375		.461	-.361				
	-.542	-.073		-.543	-.076				
	-.032	.019		-.034	.020				
	.462	-.375		.461	-.361				
	-.032	.019		-.034	.020				

table 5.3: SVD-iterations to minimize σ_J.

Y	-.237	.158	.558	-.241	.179	.543	-.243	.197	.532
	-.148	.346	.839	-.146	.316	.854	-.144	.285	.868
	-.483	.207		-.483	.207		-.483	.207	
X	-.540	-.088		-.542	-.085		-.543	-.082	
	-.330	.256		.337	.220		.342	.184	
	.188	.872		.189	.885		.191	.899	
	-.540	-.088		-.542	-.085		-.543	-.082	
	.221	-.173		.234	-.180		.246	-.188	
	.473	-.360		.467	-.354		.460	-.347	
	-.540	-.088		-.542	-.085		-.543	-.082	
	-.032	.014		-.034	.020		-.035	.022	
	.473	-.360		.467	-.354		.460	-.347	
	-.032	.014		-.034	.020		-.035	.022	
A	.639	-.478		.639	-.477		.638	-.474	
	.223	.865		.223	.866		.223	.868	
	.736	.152		.736	.151		.737	.149	
σ_J	.114			.113			.112		

table 5.4: Daugavet-iterations to minimize σ_J.

each iteration with an estimate of the y_j and of X. In the first step we compute $A = Q'X(X'X)^+$. In the second step we compute a new X, by $X = QA(A'A)^+$, and the third step is the same as the second step of the previous algorithm, it computes new y_j by normalized cone regression. We also illustrate two iterations of this algorithm in table 5.4. In general, of course, this alternative algorithm has less work in each iteration, and will as a consequence often need more iterations. Many other variations are possible, because we can change the order of the three steps, and we can perform a number of iterations of the two Daugavet-steps before computing new y_j. We have not experimented seriously with these different versions to find out which is more efficient.

5.2.6 Types of cones and missing data

We have already discussed two of the more important choices of K_j in 5.3 and 5.4. A variable is treated as single nominal if y_j can be anywhere in k_j-space (only the normalization requirements are used). Adjustment for y_j in the alternating least squares algorithm consists of computing $\tilde{y}_j$, and normalizing it. A variable is single numerical if y_j is restricted to be a linear function of a given k_j-vector Thus $y_j = \alpha_j t_j + \beta_j$, where t_j is such that $h_j = G_j t_j$. The normalization requirements then imply that y_j does not change at all during the iterations, y_j remains equal to the normalized t_j. We can consequently skip the y_j-adjustment step. The third type of cone is defined by monotonicity requirements (single ordinal variables)

$$y_{j1} \leq \ldots \leq y_{jk_j}.$$

The y_j-adjustment steps apply weighted monotone regression to the vector $\tilde{y}_j$, and normalizes the result. See appendix C.

The situation becomes more complecated if there are missing data. We have already seen in 2.5.2 that there are at least three ways in which we can proceed (compare al section 3.11). Options II and III ('missing values single category' and 'missing values multiple categories') continue to work with complete indicator matrices G_j, but the cones K_j change, because we do not require monotonicity or linearity for the missing values. For single nominal variables nothing changes, for single ordinal variables monotone regression is simply performed over the nonmissing categories, for the missing categories we copy the corresponding elements of $\tilde{y}_j$, and afterwards we normalize. This is also no real complication. For single numerical variables, however, something does change. We require

$$y_j^{(1)} = \alpha_j t_j + \beta_j,$$

for the non-missing part, we do not restrict the missing part $y_j^{(2)}$. Suppose that t_j (number of elements equal to number of non-missing categories) is normalized in such a way that

$$u'D_j^{(1)}t_j = 0,$$
$$t_j'D_j^{(1)}t_j = 1,$$

with, of course, $D_j^{(1)}$ the part of D_j corresponding with the non-missing categories
We compute

$$\beta_j = - u'D_j^{(1)}t_j \;/\; u'D_j^{(1)}u,$$
$$\alpha_j = t_j'D_j^{(1)}\tilde{y}_j^{(1)},$$
$$\bar{y}_j^{(1)} = \alpha_j t_j + \beta_j,$$
$$\bar{y}_j^{(2)} = \tilde{y}_j^{(2)},$$

and we then compute y_j by normalizing $\bar{y}_j$. This adjustment process has the obvious consequence that y_j changes during the iterations, and that the principal componer problem with missing data is no longer equivalent to a simple singular value decomposition problem, as in section 5.4.

Option I ('missing values deleted') has even more far-reaching consequences (as already shown in the HOMALS-context in section 3.11. Missing values are not quantified, and we cannot interpret our results any more in terms of R(Q) and its eigenvalues. This is a major disadvantage, but as we have already seen option II often does not make sense from the interpretational point of view, while option I may lead to individuals or variables with many missing data which completely dominate the solution. Thus option I may be the only viable alternative. But we shall now show that it also leads to very unpleasant computational complications. The loss function changes to

$$\sigma_J(Y,X,A) = \frac{1}{m}\sum_{j=1}^{m}(G_jy_j - Xa_j)'M_j(G_jy_j - Xa_j),$$

where M_j is a binary diagonal matrix, indicating which observations are missing (if i is missing for j, the diagonal element i of M_j is zero, otherwise it is one). Because G_j is now an incomplete indicator matrix we have the convenient relationship $M_jG_j = G_j$, we also have $G_ju = M_ju$. The y_j-adjustment step in the alternating least squares algorithm is now simpler than in options II and III, because the vectors y_j are shorter. It is also true in the single numerical case that y_j does not change during the iterations. But, unfortunately, adjusting for X and A for fixed y_j is not a singular value problem any more. We can generali the Daugavet three-step algorithm, but it becomes more comlicated and much more expensive. Using $q_j = G_jy_j$ we find that the optimal A for fixed X and Y must be computed row-wise. Row j , written as a column a_j, is

$$a_j = (X'M_jX)^+X'q_j.$$

Thus it will take approximately m times as long to update A, compared to $A = QX(X'X)^+$. Updating X must also be done row-wise, and takes approximately

n times as long. Define N_i as a binary diagonal matrix of order m (if i is missing for j, then diagonal element j of N_i is zero, otherwise it is one). Then row i of X, written as a column x_i, becomes

$$x_i = (A'N_iA)^+A'q_i.$$

Programs MNNFAEX, PRINQUAL, and PRINCIPALS handle missing data by using option III. Of course they cannot incorporate option I, because they rely on explicit computation of the singular value decomposition.

5.2.7 Use of meet-loss

Component analysis based on σ_J has some inconvenient features. In the first place multiple quantifications must be computed successively, while in homogeneity analysis we can compute them simultaneously. And secondly missing data option I gives us computational troubles. We now present an alternative approach to components analysis, which does not have these disadvantages. Define

$$\sigma_M(X,Y) = \frac{1}{m}\sum_{j=1}^{m} SSQ(X - G_jY_j),$$

with normalization $AVE(x_s) = 0$ for $s=1,\ldots,p$ and $X'X = I$. We use the notation σ_M here, because $\sigma_M(X,Y) = 0$ for some X and Y implies that the meet-rank of the G_j is at least p. As explained in chapter 3 the alternating least squares algorithm for minimizing σ_M is

$$Y_j \leftarrow D_j^{-1}G_j'X,$$

$$Z \leftarrow \sum_{j=1}^{m} G_jY_j,$$

$$X \leftarrow GRAM(Z),$$

and it computes the first p dimensions of a homogeneity analysis simultaneously. These three steps are identical with the HOMALS algorithm (without missing data).

Now suppose that we impose in addition the rank-one restrictions

$$Y_j = y_ja_j',$$

with

$$y_j \in K_j,$$

$$u'D_jy_j = 0,$$

$$y_j'D_jy_j = 1.$$

Some simple computation, using the normalization conditions, gives

$$\sigma_M(X,Y,A) = \sigma_J(Y,X,A) + (p - 1),$$

where

$$\sigma_M(X,Y,A) = \frac{1}{m}\sum_{j=1}^{m} SSQ(X - G_jy_ja_j').$$

(The fact that we use Y_j for the multiple category quantifications, and y_j for the single category quantifications, may be confusing. On the other hand a variable is either single or multiple, not both, and has thus either a Y_j or a y_j). It follows from the last result that minimizing $\sigma_M(X,Y,A)$ and $\sigma_J(Y,X,A)$ are equivalent problems, with the same solutions.

The first advantage of using σ_M is that we can impose the conditions $Y_j = y_j a_j'$ for some variables and not for others. If we impose them for <u>all</u> variables, then we minimize σ_J, and we are doing components analysis, as before. If we do not impose them at all, then we are doing homogeneity analysis as in chapter 3. It may be interesting to mix the two options, and to give some variables single quantifications and others multiple quantifications. Especially for nominal variables single quantification often is not very natural, because it seems to suggest that the categories can be ordered on a single scale. Consequentl we might compute single quantifications for ordinal and numerical variables, and multiple quantifications for nominal variables.

Another advantage becomes clear if we analyze the implementation of option I for missing data. Now

$$\sigma_M(X,Y) = \frac{1}{m}\sum_{j=1}^{m} \text{tr}(X - G_jY_j)'M_j(X - G_jY_j),$$

with normalization

$$u'M_*X = 0,$$

$$X'M_*X = mI,$$

where

$$M_* = \sum_{j=1}^{m} M_j.$$

If $Y_j = y_j a_j'$ the simple relation between σ_M and σ_J is no longer true, and the alternating least squares algorithm for minimizing σ_M is quite simple. We start with an X satisfying the constraints. Define

$$\tilde{Y}_j = D_j^{-1}G_j'X.$$

Then

$$\text{tr}\ (X - G_jY_j)'M_j(X - G_jY_j) = \text{tr}\ (X - G_j\tilde{Y}_j)'M_j(X - G_j\tilde{Y}_j) + \text{tr}\ (Y_j - \tilde{Y}_j)'D_j(Y_j - \tilde{Y}_j)$$

Thus minimizing over Y_j can be done by minimizing the second term on the right. If the variable is <u>multiple nominal</u> we simply set $Y_j = \tilde{Y}_j$, and we are done. We can also introduce at this point a new type of variable, which fits naturally into the general framework. For a <u>multiple ordinal</u> variable we require that all columns of Y_j are either increasing or decreasing. As explained by Guttman (1959) we cannot require them all to be increasing. Computing Y_j amounts to solving two monotone regression problems for each column, the increasing and

the decreasing one, and to keep the best one. Multiple numerical variables require that all columns are linear functions of a given vector. It is easy to see, however, that this amounts to requiring that $Y_j = y_j a_j'$, with y_j known, and that consequently multiple numerical is identical with single numerical.

If a variable is single, we have to minimize

$$\text{tr } (y_j a_j' - \tilde{Y}_j)' D_j (y_j a_j' - \tilde{Y}_j)$$

over $y_j \in K_j$ and over a_j. For single nominal variables y_j is unrestricted, and we can minimize over y_j and a_j simultaneously by computing the dominant singular value and corresponding singular vectors of $D_j^{\frac{1}{2}} \tilde{Y}_j$. In general, however, we prefer alternating least squares inner iterations to solve for $y_j \in K_j$ given a_j, and for a_j given y_j. Solving for a_j, given y_j, can be done by defining

$$\tilde{a}_j = \tilde{Y}_j' D_j y_j \; / \; y_j' D_j y_j,$$

which gives

$$\text{tr } (y_j a_j' - \tilde{Y}_j)' D_j (y_j a_j' - \tilde{Y}_j) = \text{tr } (y_j \tilde{a}_j' - \tilde{Y}_j)' D_j (y_j \tilde{a}_j' - \tilde{Y}_j) + y_j' D_j y_j \, . \, \text{SSQ}(a_j - \tilde{a}_j).$$

Because a_j is unrestricted in general, this means that we can simply set $a_j = \tilde{a}_j$. It is trivial to generalize this to restricted a_j (which occurs in various forms of factor analysis). If we require $a_j \in L_j$, for some convex set L_j, then this substep of the inner iterations means that we must project a_j on L_j.

Solving for $y_j \in K_j$ for fixed a_j (the other half of one inner iteration) defines

$$\tilde{y}_j = \tilde{Y}_j a_j \; / \; a_j' a_j,$$

and uses the partitioning of the sum of squares given by

$$\text{tr } (y_j a_j' - \tilde{Y}_j)' D_j (y_j a_j' - \tilde{Y}_j) = \text{tr } (\tilde{y}_j a_j' - \tilde{Y}_j)' D_j (\tilde{y}_j a_j' - \tilde{Y}_j) + a_j' a_j \, . \, (y_j - \tilde{y}_j)' D_j (y_j - \tilde{y}_j).$$

Thus finding y_j means projecting the vector $\tilde{y}_j$ on K_j, which is a monotone regression problem in the ordinal case, a linear regression problem in the numerical case, and a question of simply setting $y_j = \tilde{y}_j$ in the nominal case.

Several comments are in order here. In the first place we can choose how many inner two-step alternating least squares iterations we want to make, before we go on to compute a new X. In the second place we do not use or assume here that $y_j' D_j y_j = 1$ or that $u' D_j y_j = 0$. In minimizing σ_J it was not necessary to normalize X and A, because Y was normalized. In minimzing σ_M it is not necessary to normalize Y and A, because X is normalized. In the third place the situation without missing data is a special case in which $M_j = I$ for all j. From the properties of the regression algorithms we find that after regression

$$u'D_j y_j = u'D_j \tilde{y}_j = u'D_j \tilde{Y}_j a_j \,/\, a_j'a_j = u'D_j D_j^{-1} G_j' X a_j \,/\, a_j'a_j = u'G_j' X a_j \,/\, a_j'a_j =$$
$$= u'M_j X a_j \,/\, a_j'a_j.$$

Thus if $M_j = I$ for all j (no missing data, or option II or III), then $u'D_j y_j = 0$ follows from $u'X = 0$, and our method can be interpreted in terms of R(Q).

The second step in the outer iteration is to compute a new X for given Y_j, on the conditions that $u'M_* X = 0$ and $X'M_* X = mI$. This consists of the following three steps. We first compute

$$Z = \sum_{j=1}^{m} M_j G_j Y_j,$$

then

$$\tilde{Z} = (M_* - \frac{M_* uu' M_*}{u'M_* u})Z,$$

and

$$X = m^{\frac{1}{2}} M_*^{-\frac{1}{2}} \text{GRAM}(M_*^{-\frac{1}{2}} \tilde{Z}).$$

If someone checks these computations, he will find out that this is <u>not</u> the least squares solution for X given the Y_j. We find the least squares solution by solving a Procrustus problem (Cliff, 1966), our Gram-Schmidt solution is a rotation of the optimal solution. It is easy to see, however, that this implies that the Y_j in the next iteration will be rotated in the same way, and that consequently σ_M will decrease as much in a major iteration if we use Gram-Schmidt. Using Gram-Schmidt gives a smaller decrease in the step which updates X, but a larger decrease in the next step which updates the Y_j. The total decrease is the same, and because Gram-Schmidt is less expensive than Procrustus we prefer it.

We remark finally that there is another way in which homogeneity analysis can be fitted into the framework of component analysis. Suppose we consider each of the Σk_j categories in a complete homogeneity analysis as a new variable, indexed $\ell = 1,\ldots,L$. For each ℓ we define M_ℓ as the diagonal matrix with on the diagonal the corresponding column g_ℓ of G. Clearly $M_* = mI$. We also define the cone K_ℓ in n-space as the set of those vectors q with the property that if objects i and k are in the category corresponding with ℓ, then $q_i = q_k$. Or: if $q \in K_\ell$ then $M_\ell q$ is proportional to g_ℓ, for normalization purposes we can simply require $M_\ell q =$ But under these conditions some computation gives

$$\sum_{\ell=1}^{L} \text{tr}(X - q_\ell a_\ell')'M_\ell(X - q_\ell a_\ell') = \sum_{j=1}^{m} \text{SSQ}(X - G_j A_j),$$

which is the usual loss function for homogeneity analysis.

5.2.8 Geometry of meet-loss

Conditions for minimal meet-loss for each variable have interesting geometrical interpretations. As in homogeneity analysis we represent object scores as points in p-dimensional space. If a variable is multiple nominal, its loss-contribution is

$$\text{tr } (X - G_jY_j)'M_j(X - G_jY_j),$$

which vanishes if and only if $M_jX_j = G_jY_j$ if and only if all objects in the same non-missing category have the same object score, which is then of course equal to the corresponding category quantification. As in homogeneity analysis we expect objects in a category to be close together. If a variable is multiple ordinal we partition the loss-contribution as

$$\text{tr } (X - G_j\tilde{Y}_j)'M_j(X - G_j\tilde{Y}_j) + \text{tr } (Y_j - \tilde{Y}_j)'D_j(Y_j - \tilde{Y}_j),$$

where

$$\tilde{Y}_j = D_j^{-1}G_j'X.$$

The loss component vanishes if and only if all object scores corresponding with a nonmissing category are the same and all category quantifications project on the dimensions in an increasing or decreasing order. This is shown in figure 5.1. There is loss because the object points do not coincide with the category points, and there is loss because the category points do not project in the correct order on the dimensions. Observe that for multiple ordinal data there generally is no complete freedom of rotation of the axes. For single variables there are three loss-components for each variable. A single additive partitioning of the loss is not possible (for fixed a we can partition with respect to y, and vice versa but not both at the same time). Geometrically, however, we want all object scores corresponding with a nonmissing category to be the same and we want category quantifications (centroids of object scores) to be on a line through the origin and we want them to be on this line in the correct way (defined by some cone K_j). This situation is pictured in figure 5.2.

Observe that for single variables the sum of all three loss components vanishes only if all objects are on the line through the origin. But this can never occur because of the normalization of X, which forces it to be of full rank. This explains that $\sigma_M = \sigma_J + (p - 1)$, and that consequently $p > 1$ implies $\sigma_M > 0$. The situation with $\sigma_M = 0$ it not very interesting, it can only occur if there are sufficiently many missing data and if we use option I. If there are no missing data (or if we use options II or III), then the contribution to meet-loss of variable j is minimized (but not zero) if the corresponding component of join-loss is zero, which means that $G_jy_j = Xa_j$. Geometrically this condition means that the object point corresponding with a category must be on parallel hyperplanes

perpendicular to the direction defined by a. Thus a more realistic representation of single loss is figure 5.3.

It is not very natural, in many situations, that the objects should be on parallel hyperplanes. It seems more satisfactory, from a geometrical point of view, to require that there exist $k_j - 1$ of these hyperplanes which separate the categories of a variable, in the sense that all objects in a category should be between two of these parallel hyperplanes. This particular way of presenting categorical data was investigated by Lingoes (1968) and by De Leeuw (1969). It corresponds to minimizing the join-loss with all variables continuous ordinal, and we have seen in 5.2.4 that such a technique will not work in many interesting situations. The other techniques discussed by Lingoes and De Leeuw, all based on the idea of separation, basically have the same problems. Guttman's MSA-I technique, for example, can be interpreted as an attempt to find representations similar to HOMALS, but with a far more realistic way of measuring loss. But unfortunately loss is defined in such a way that it is too easy to find a solution with perfect fit. MSA-I, in fact, uses HOMALS as an initial configuration, and in all of the examples we have seen the sophisticated iterative procedure either can not or does not need to get away from its starting point. The argument is the same as in section 4.2, in which we compare HOMALS with non-metric unfolding. By making the requirements on the representation unrealistically strong we prevent degeneracy and instability. This is also the reason, of course, why sometimes non-metric methods do not work and metric methods do work, even if the data are clearly ordinal (Woodward and Overall, 1976, Bentler and Weeks, 1979).

There is one special situation in which the distinction between discrete and continuous becomes irrelevant. The situation is important, because of the analysis of preference rank orders mentioned at the end of 5.1.2. If we analyze preference rank orders, then each individual defines a variable, and the objects that are compared with respect to preference are also the objects in the components analysis. This implies that each 'variable' has n categories if n objects are ranked, and thus each G_j is a permutation matrix, $D_j = I$, and $G_j\tilde{Y}_j = G_jD_j^{-1}G_j'X = X$ for all $j=1,\ldots,m$. Thus multiple loss is always equal to zero, and homogeneity analysis would give a completely arbitrary result. In fact HOMALS stops after one iteration, with X equal to the random initial configuration, and with $Y_j = G_j'X$. If the variable is single nominal, then we can minimize $SSQ(X - G_jy_ja_j')$ by setting $y_j = G_j'Xb_j$ and $a_j = b_j/b_j'b_j$, with b_j completely arbitrary, and X arbitrary except for $X'X = I$. The minimum of σ_M is $(p - 1)$ if all variables are single nominal, which is another way of saying that σ_J is trivially equal to zero.

It follows that we can throw out all variables that are either multiple or

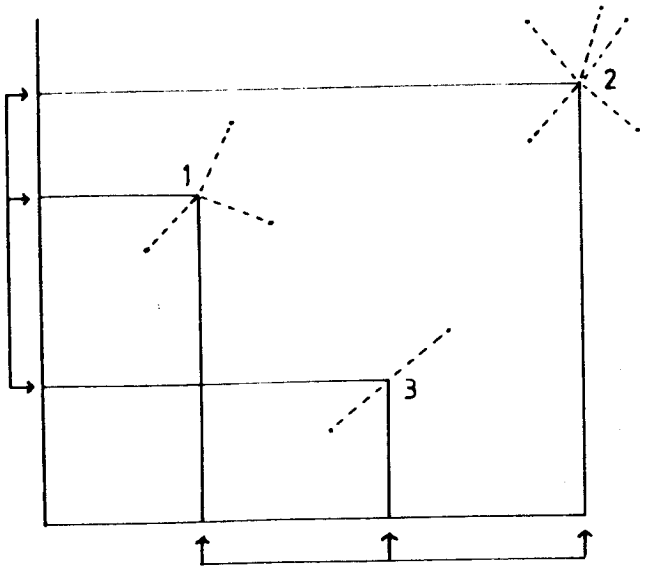


Figure 5.1 Meet-loss for a multiple ordinal variable

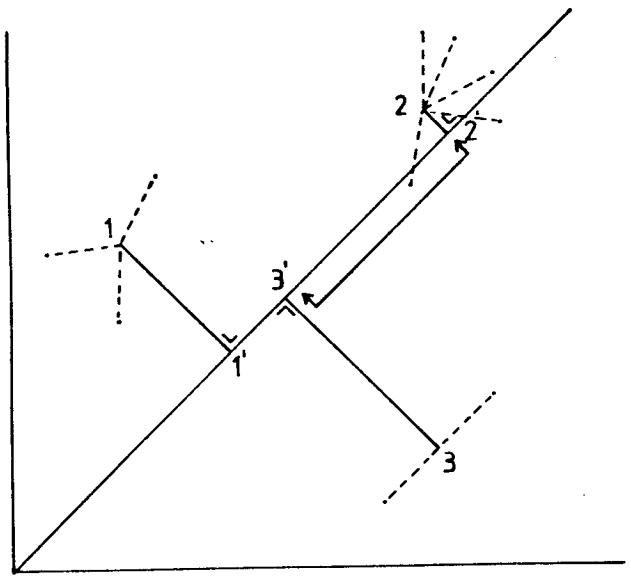
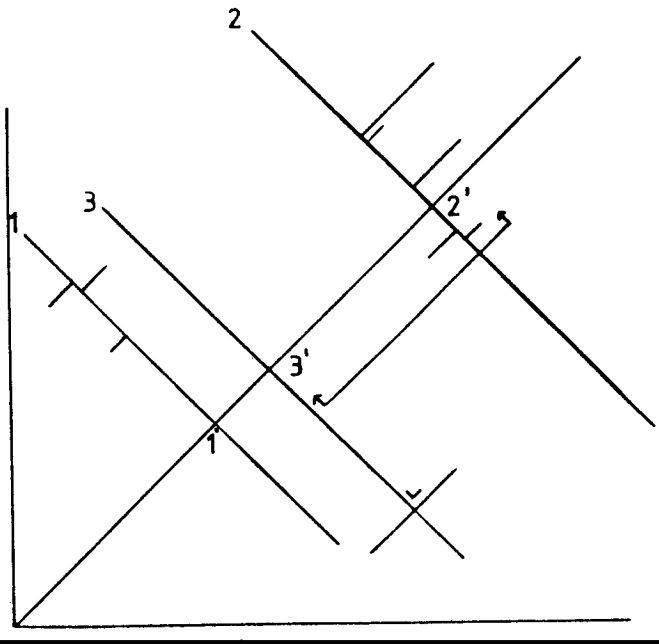


Figure 5.2 Meet-loss for a single ordinal variable



single nominal in this case, and keep only those variables that are single numerical or ordinal. In the numerical case the optimal X and A can be computed, as usual, from the matrix with columns $G_j y_j$. If the y_j are all equal to the centered rank numbers, for example, then the technique becomes identical to the one proposed by Guttman (1946), Slater (1960), Carroll and Chang (1964), and Benzécri (1967). Compare also Bechtel, Tucker, and Chang (1971), De Leeuw (1968, 1973 chapter 4), and Nishisato (1978). If all variables are ordinal, then we require that the projections of the objects on the direction corresponding with individual j (remember that the usual variables are here individuals) must be in the appropriate order. There is no multiple 'within-category' loss, all the loss is single. In fact even the rank one restrictions do not lead to nontrivial loss, only the ordinal or numerical restrictions give a contribution.

5.3 The PRINCALS program

5.3.1 Loss function

The PRINCALS program minimizes σ_M, and uses in principle option I ('missing data deleted') for missing data. This means that the loss function uses the matrices M_j, and that the interpretation in terms of R(Q) is not possible if there are missing data. Of course it is always possible to use options II and III in PRINCALS as well, but in that case the user must recode his data in the appropria way.

Because PRINCALS without missing data (or with recoded missing data under options II and III) can only perform linear regression or monotone regression over all categories this recoding will only work properly if the recoded data are treated as single nominal or multiple nominal. Otherwise we have to impose order or linearity on mssing data, which is clearly undesirable.

5.3.2 Types of variables

PRINCALS accepts multiple nominal, and single nominal, ordinal, and numerical variables. Multiple ordinal is not (yet) implemented, there are no continuous options. We can compare this with some of the other programs we mentioned. MNNFAB of Roskam and Lingoes and NMFA of Kruskal and Shepard accept single discrete ordi and single continuous ordinal. PRINQUAL of Tenenhaus accepts single discrete ordinal and single discrete numerical. PRINCIPALS of Young, Takane, and De Leeuw accepts single discrete and continuous, nominal, ordinal, and numerical, in any combination, but not multiple nominal.

5.3.3 Normalizations

As in HOMALS. Thus we normalize X by $X'M_*X = mnI$ and $u'M_*X = 0$, which implies that X is in standard scores if $M_j = I$ for all j. If the variables are single,

then we require $y_j'D_jy_j = n$. If there are no missing data (or with options II and III) this implies that the elements of a_j are correlations and can be interpreted as ordinary 'component loadings'. For multiple nominal variables the category quantifications and the discrimination measures are as in HOMALS. PRINCALS also outputs $\tilde{Y}_j = D_j^{-1}G_j'X$ for single variables, and for comparison purposes in addition y_ja_j'. In some of the programs we mentioned above other normalizations are in use. Sometimes the a_j are interpreted as direction cosines, with $a_j'a_j = 1$, which means that if we represent them as points in the joint plot they are all on a circle. This can be useful in the context of preference rank orders, but we prefer not to use this normalization in general.

5.3.4 Partitioning

PRINCALS uses the partitioning of the loss for variable j given by

$$\text{tr } (X'M_jX - \tilde{Y}_j'D_j\tilde{Y}_j) + \text{tr } (\tilde{Y}_j'D_j\tilde{Y}_j - a_ja_j').$$

The first component is called the multiple loss, the second component the single loss. For multiple variables there is no single loss. The multiple fit are the diagonal elements of $\tilde{Y}_j'D_j\tilde{Y}_j$, which are the discrimination measures of HOMALS, the single fit are the a_{js}^2, the squared component loadings. PRINCALS prints the two m × p tables with multiple and single fits.

5.3.5 Eigenvalues

If there are no missing data (or with option II or III) the eigenvalues printed by PRINCALS are those of $\frac{1}{m}R(Q)$. In that case R(Q) itself is also printed. Of course use of R(Q) supposes that all variables are single, or that p = 1, in which case single is the same as multiple. If there are missing data, and we use option I, then the eigenvalues are those of the matrix with elements $y_j'G_j'M_*^{-1}G_\ell y_\ell$, which is not necessarily a correlation matrix, although it is of course always positive semidefinite.

5.3.6 Two phases

PRINCALS computes its solution in two phases. In the first step all single variable are interpreted as numerical, and a solution is computed for multiple nominal and single numerical variables. After these iterations have converged, we use the resul as starting configuration for the second phase, which is PRINCALS with the appropriate measurement levels. The program can print both solutions. The idea behind the two phases is that the first phase will give us a good start, because PRINCALS with multiple nominal and single numerical only is a singular value problem, and consequently in the first phase the algorithm always converges to the global minimum of σ_M. This may avoid local minima in the second phase (at least some local minima). It is easy to assess from the two-step results in how far the nonmetric solution deviates from or improves upon the linear one. We

have not investigated systematically how frequently local minima occur in PRINC with single ordinal and single nominal variables. We also conjecture that if th original linear coding of the categories used in the first phase is quite diffe from the optimal coding found in the second phase, then the two-phase procedure could very well tend to introduce local minima. Because the number of iteration and the precision in the two phases can be chosen independently this is no real problem. If we only want local improvements on the linear solution we find prec solutions in both phases, if we expect dramatic changes or if we do not have th faintest idea about an optimal quantification, then we only perform a single it in the first phase. Recent numerical experience suggests that in cases with a poor fit single ordinal is much more robust with respect to local minima as sin nominal. In single nominal we impose an order on the category quantifications, without knowing which order. We want the program to find the correct order. It seems that there is a local minimum for each possible order of the category quantifications, and all but one of these local minima are avoided by imposing ordinal constraints. If $p > 1$ we advise against single nominal, either the orde of the categories is known, in which case we should use single ordinal, or the categories are unordered and unorderable, in which case we should use multiple nominal.

5.4 Some examples

5.4.1 The Guttman-Bell data

These data are discussed in Guttman (1968) and Lingoes (1968), mainly to illust the Guttman-Lingoes MSA-programs. The data were adapted by Guttman from a socio text, they characterize seven different groups of people in terms of five varia The variables, with their alternatives, are

1: Intensity of interaction.
   1: slight
   2: low
   3: moderate
   4: high

2: Frequency of interaction.
   1: slight
   2: non-recurring
   3: infrequent
   4: frequent

3: Feeling of belonging
   1: none
   2: slight
   3: variable
   4: high

4: Physical proximity.
   1: distant
   2: close

5: Formality of relationship.

1: no relationship
2: formal
3: informal

The seven objects were classified as follows:

crowd	1	1	1	2	2
audience	2	2	2	2	2
public	1	1	2	1	1
mob	4	2	4	2	3
primary group	4	4	4	2	3
secondary group	3	3	3	1	2
modern community	2	3	3	2	2

Guttman and Lingoes apply their four MSA programs to this small matrix, which is basically not such a good idea because all MSA programs impose only a small number of constraints, and can fit small examples quite perfectly. The MSA-I solution is very interesting for us, because it uses Lingoes' MAC as an initial estimate, and Lingoes MAC implements Guttman (1941), and is consequently the same as HOMALS. The HOMALS solution for the object scores is virtually the same as the MSA-I solution given by Lingoes, which is not surprising because the HOMALS solution satisfies the contiguity requirements of MSA-I perfectly. As a matter of fact it also satisfies the requirements of MSA-III perfectly as well as those of MSA-IV. These last two methods are the same as nonmetric principal components analysis with continuous single nominal and continuous single ordinal variables. We can compare our HOMALS solution (or PRINCALS with all variables multiple nominal) plotted in figure 5.4 with our PRINCALS solution, plotted in 5.5, all variables single nominal. In both figures the loss for variable 1 is illustrated. Numerical results for HOMALS and PRINCALS analysis of this example are in table 5.5. We do not think this example shows us anything deep or unexpecte about the structure of social groups, the data matrix is so small that data reduction is quite unneccessary. We merely use the example to show some of the plots that can be made, and some of the statistics that can be computed.

5.4.2 Roskam's journal preference data

In table 5.6 we have given preference rank orders of 39 psychologists for ten psychological journals (from Roskam, 1968, p 152). As usual a low element in the table indicates a high preference for the journal. The ten journals are:

1: JEXP: Journal of experimental psychology
2: JAPP: Journal of applied psychology
3: JPSP: Journal of personality and social psychology
4: MUBR: Multivariate behavioural research
5: JCLP: Journal of consulting psychology
6: JEDP: Journal of educational psychology
7: PMEK: Psychometrika
8: HURE: Human relations
9: BULL: Psychological bulletin
10: HUDE: Human development

PRINCALS, p=1, single nominal

Q	-0.812	-0.884	-0.590	0.632	-0.509
	-0.240	0.737	-0.572	0.632	-0.509
	-0.812	-0.884	-0.572	-1.581	-1.070
	1.527	0.737	1.579	0.632	1.552
	1.527	1.795	1.579	0.632	1.552
	-0,951	-0.750	-0.712	-1.581	-0.509
	-0.240	-0.750	-0.712	0.632	-0.509
R	1.000				
	0.860	1.000			
	0.964	0.810	1.000		
	0.557	0.517	0.406	1.000	
	0.964	0.817	0.976	0.499	1.000
eval	0.805	0.140	0.047	0.006	0.003

PRINCALS, p=2, single nominal

	-0.731	-0.816	-0.922	0.632	-0.675
	-0.377	0.750	-0.395	0.632	-0.675
	-0.731	-0.816	-0.395	-1.581	-0.454
	1.556	0.750	1.558	0.632	1.577
	1.556	1.783	1.558	0.632	1.577
	-0.896	-0.826	-0.702	-1.581	-0.675
	-0.377	-0.826	-0.702	0.632	-0.675
	1.000				
	0.843	1.000			
	0.979	0.840	1.000		
	0.515	0.519	0.346	1.000	
	0.978	0.789	0.990	0.357	1.000
	0.794	0.160	0.045	0.001	0.000

PRINCALS, p=3, single nominal

Q	1.011	-0.248	0.989	-0.632	0.509
	0.368	-0.818	0.763	-0.632	0.509
	1.011	-0.248	0.763	1.581	1.068
	-1.518	-0.818	-1.534	-0.632	-1.553
	-1.518	-0.926	-1.534	-0.632	-1.553
	0.278	1.530	0.276	1.581	0.509
	0.368	1.530	0.276	-0.632	0.509
R	1.000				
	0.405	1.000			
	0.984	0.352	1.000		
	0.408	0.405	0.329	1.000	
	0.975	0.494	0.964	0.499	1.000
eval	0.720	0.158	0.123	0.000	0.000

Eigenvalues HOMALS

1	0.805
2	0.625
3	0.397
4	0.334
5	0.156
6	0.084

Object scores HOMALS, first two

0.56	0.98
0.05	-0.16
1.03	1.59
-1.39	0.08
-1.62	0.14
0.94	-1.41
0.42	-1.22

Object scores PRINCALS, p = 2

0.67	-1.12
0.09	-1.06
0.80	1.38
-1.41	0.34
-1.65	0.26
0.96	1.22
0.53	-1.02

HOMALS, first two Discrimination measures		PRINCALS, p = 2. Multiple fit		Component loadings		Single fit	
0.970	0.894	0.981	0.552	-0.990	0.061	0.979	0.00
0.817	0.967	0.824	0.054	-0.907	-0.066	0.824	0.00
0.908	0.779	0.946	0.214	-0.968	0.242	0.937	0.05
0.389	0.003	0.312	0.678	-0.559	-0.823	0.312	0.67
0.938	0.480	0.940	0.439	-0.957	0.236	0.916	0.05

TABLE 5.5 Results Guttman-Bell data

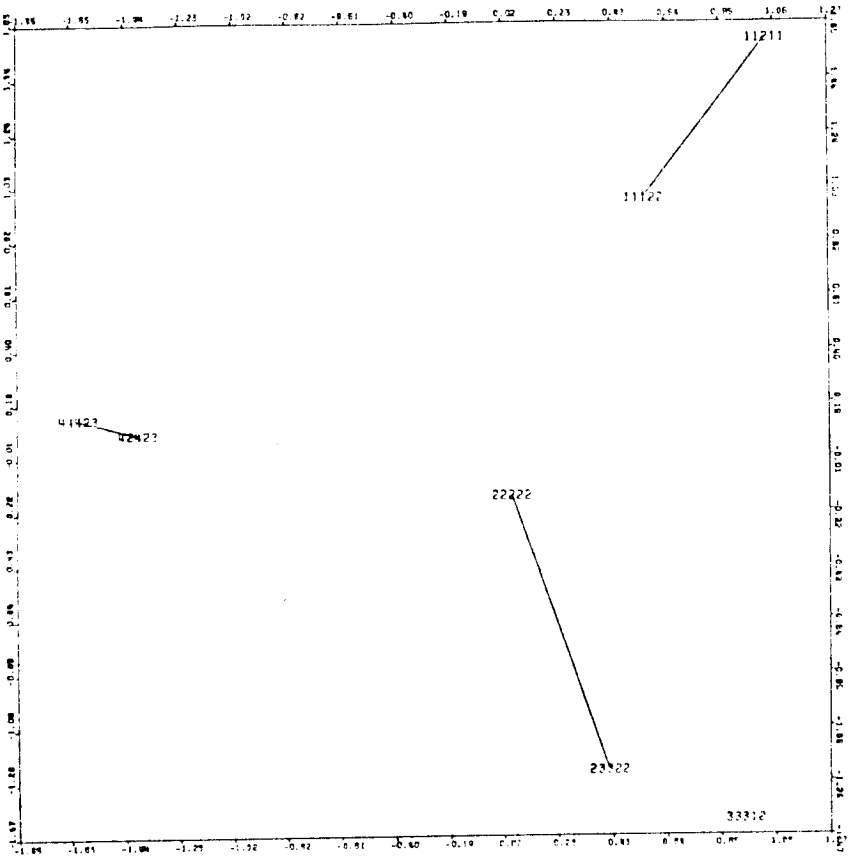


Figure 5.4 HOMALS solution of the Guttman-Bell data with the loss for variable 1

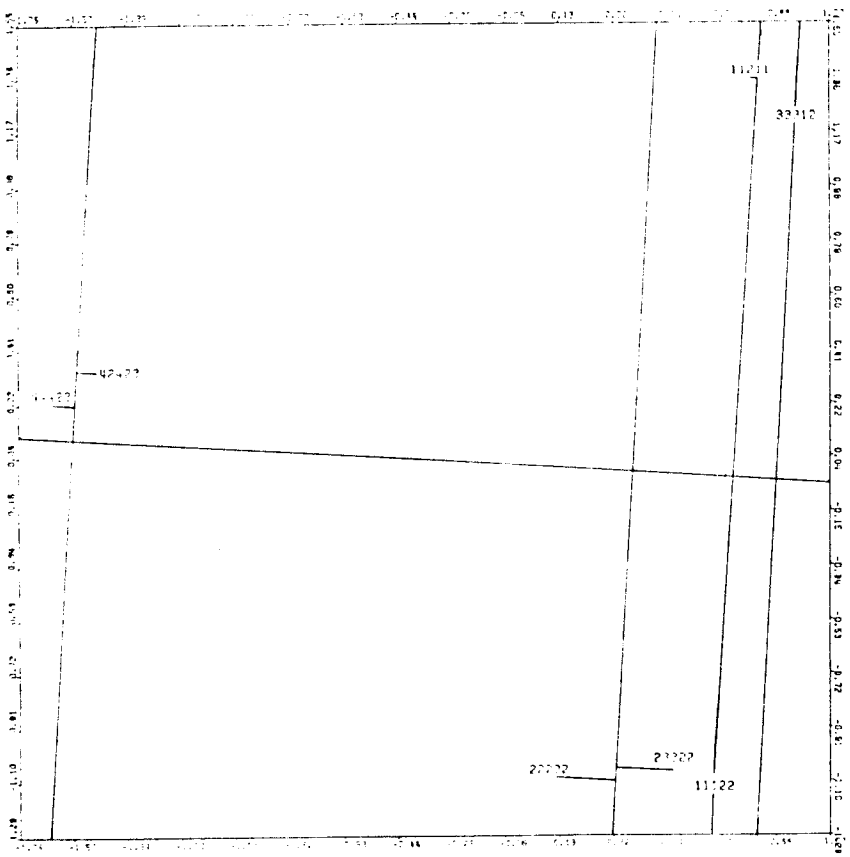


Figure 5.5 Single nominal PRINCALS solution of the Guttman-Bell data with the loss for variable 1

1:	7	4	1	8	10	9	5	2	3	6	(S)
2:	7	6	2	9	3	8	10	1	4	5	(S)
3:	10	5	1	7	4	6	8	2	3	9	(S)
4:	6	5	3	7	4	8	9	2	1	10	(S)
5:	6	3	5	10	4	2	9	7	8	1	(D)
6:	8	7	4	9	2	5	10	6	3	1	(D)
7:	5	9	4	8	6	2	10	7	3	1	(D)
8:	6	7	4	9	5	3	10	8	2	1	(D)
9:	2	3	6	4	5	8	9	7	10	1	(D)
10:	5	8	2	9	1	7	10	6	4	3	(D)
11:	7	2	6	10	5	1	9	8	4	3	(D)
12:	8	7	2	9	1	6	10	5	3	4	(C)
13:	10	7	1	9	4	6	8	2	3	5	(C)
14:	5	2	3	4	1	8	7	9	6	10	(C)
15:	6	5	2	7	1	10	9	8	4	3	(C)
16:	4	7	5	2	8	9	1	6	3	10	(M)
17:	4	7	5	3	9	8	1	6	2	10	(M)
18:	5	4	7	3	9	8	1	10	2	6	(M)
19:	1	5	6	7	10	9	3	8	2	4	(E)
20:	1	5	8	7	9	3	6	10	2	4	(E)
21:	3	7	6	2	8	4	5	9	1	10	(E)
22:	1	3	8	6	9	7	4	10	2	5	(E)
23:	1	4	6	5	9	10	2	8	3	7	(E)
24:	1	7	5	4	10	9	3	8	2	6	(E)
25:	1	8	6	5	9	4	3	10	2	7	(E)
26:	1	2	5	6	10	4	7	9	3	8	(E)
27:	1	5	6	4	8	7	2	9	3	10	(E)
28:	4	6	5	1	7	10	3	8	2	9	(E)
29:	8	7	1	2	9	10	6	3	4	5	(R)
30:	7	4	1	2	9	10	8	6	3	5	(R)
31:	9	8	2	7	1	4	10	5	6	3	(R)
32:	7	1	5	8	2	6	3	9	4	10	(T)
33:	2	3	7	8	10	9	1	6	4	5	(T)
34:	10	4	2	9	3	5	6	8	1	7	(T)
35:	3	2	10	6	8	4	7	9	1	5	(T)
36:	6	1	3	9	4	7	10	2	5	8	(T)
37:	2	1	6	4	10	9	5	7	3	8	(T)
38:	2	3	6	5	7	8	4	9	1	10	(A)
39:	2	6	7	3	10	8	4	9	1	5	(A)

table 5.6: Roskam's journal data preference rank orders

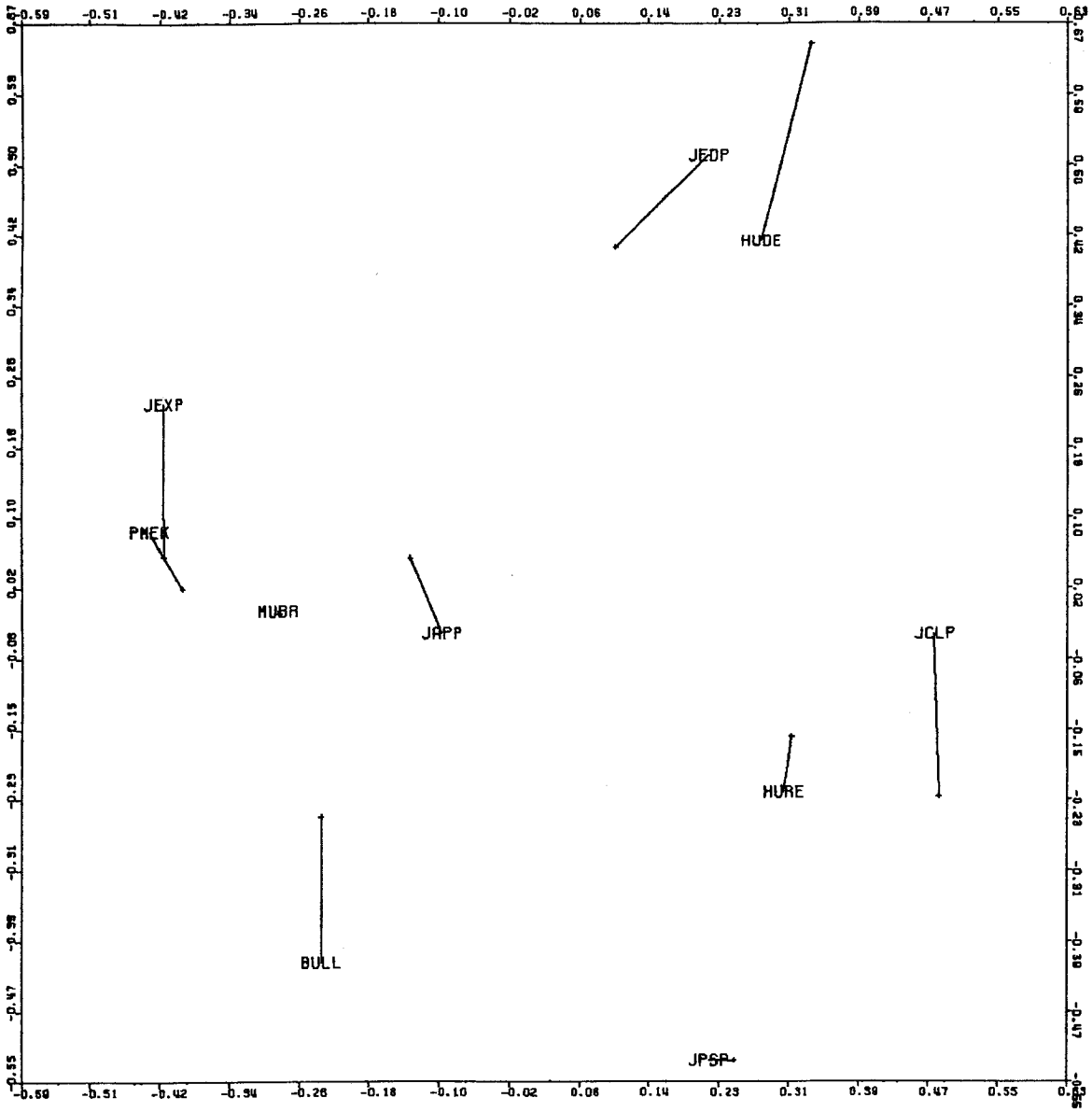


Figure 5.6 Object scores for the Roskam preference data in the single numerical solution (labeled), connected with the corresponding points in the single ordinal solution

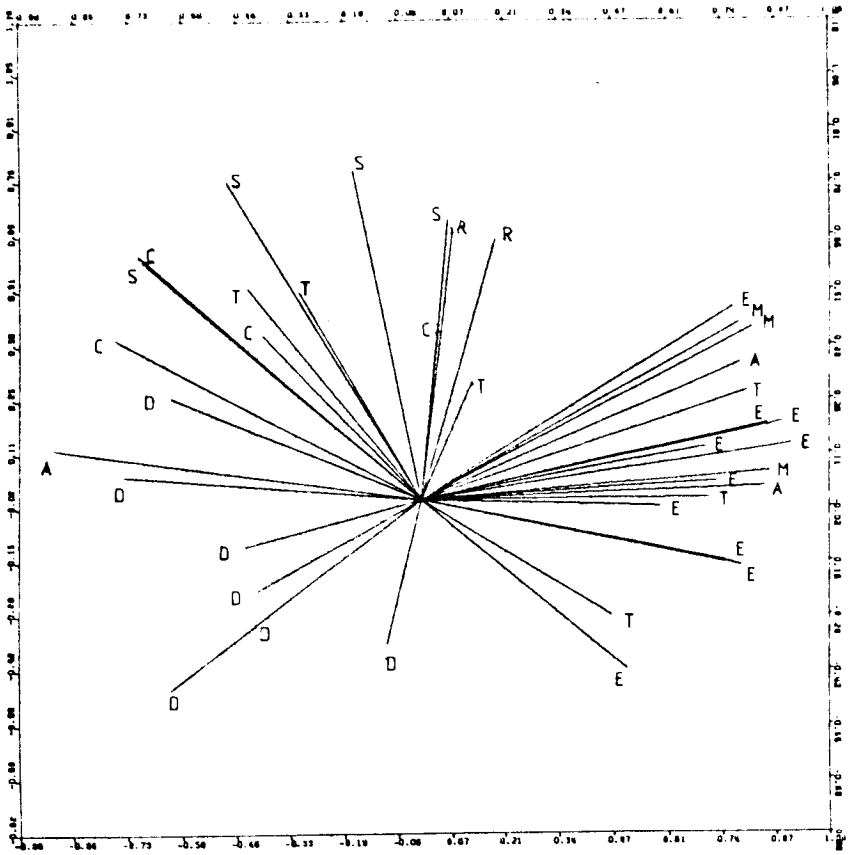


Figure 5.7 Loadings for the 39 individuals in the single numerical solution, labeled according to their speciality

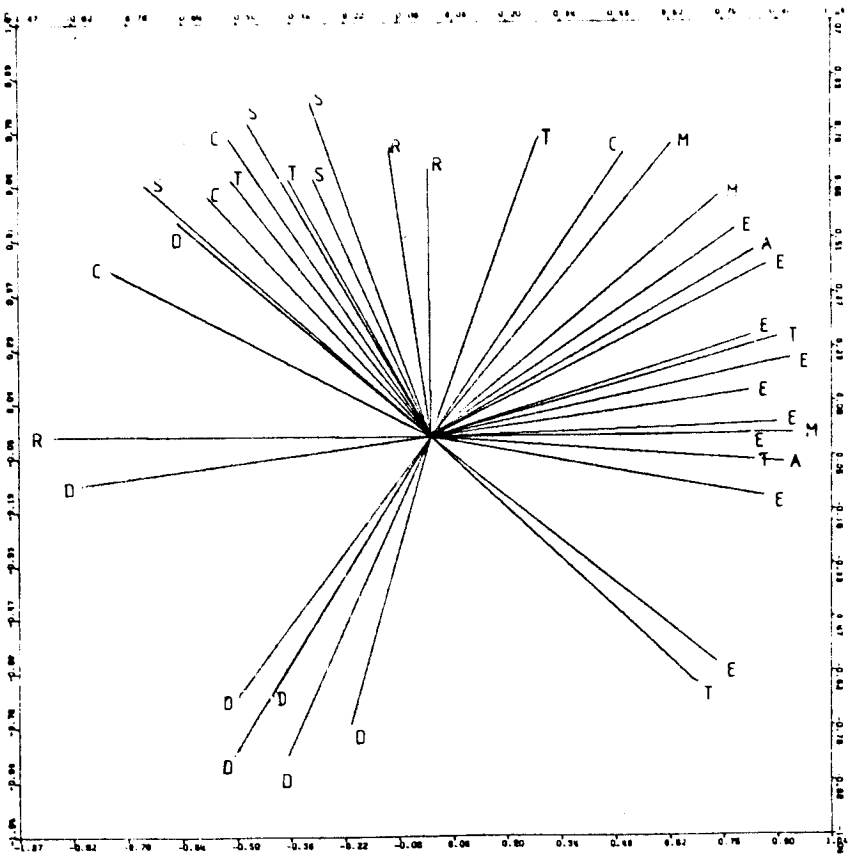


Figure 5.8 Loadings for the 39 individuals in the single ordinal solution, labeled according to their speciality

We did two PRINCALS analyses in two dimensions on these data: one with all variables single numerical, and one with all variables single ordinal. The solution with all variables numerical has loss 1.438, which means that the fit is .562, the two eigenvalues are .405 and .157. The solution with all variables ordinal has loss 1.183, fit .817, and eigenvalues .477 and .340. Figure 5.6 gives object scores for both solutions, the labeled points are the metric solution, the small crosses on the other end of the line are the nonmetric solution. The solutions are quite different from the one given by Roskam (1968, p 69) using the unfolding model nonmetrically. The journals seem to cluster into a hard group (JEXP, PMEK, MUBA, and perhaps JAPP, BULL), a developmental group (JEDP, HUDE), and a soft group (JPSP, JCLP, HURE). It is also possible to see these journals arranged on a circular structure (with JAPP in the middle). The difference between the two solutions is minimal, JEXP moves to PMEK, and BULL moves to the hard group.

The loadings for the 39 psychologists are given in figure 5.7 (numerical) and 5.8 (ordinal). Observe that the arrows point in the least preferred direction. For the labeling of the psychologists we have used Roskam's table 8.3 in Roskam (1968, p 152), which makes it possible for each subject to find out in which department he workes. The codes are

S: Social psychology
D: Educational and developmental psychology
C: Clinical psychology
M: Mathematical psychology and psychological statistics
E: Experimental psychology
R: Cultural psychology and psychology of religion
T: Industrial psychology
A: Physiological and animal psychology

The main difference between figures 5.7 and 5.8 is the vastly improved fit, which shows because it tends to make the nonmetric vectors of the same length. We also see people from the same department generally close together, with some exceptions (Industrial is not very homogeneous). Experimental, mathematical, and physiological overlap completely. Social and clinical also have considerable overlap, developmental is fairly homogeneous. The nonmetric solution seems somewhat more tightly clustered, and possibly somewhat clearer.

5.4.3 Thurstone's cylinder problem

A version of this problem has been used by Coombs and Kao, by Kruskal and Shepard, and by Young, Takane, and De Leeuw to test their approaches to principal components analysis. We made our own version with 20 objects and 10 variables. The objects varied on two dimensions, given in the first 2 columns of 5.7.a. These are uniform random variables on the unit interval. The ten variables were monotone functions of these two dimensions. If the two columns of 5.7.a are a_i and b_i, and t_{ij} is

.63	.76	3.08	1.95	.48	.06	.29	.84	1.20	6.95
.99	.37	2.14	2.12	.36	.02	.65	2.71	.37	46.64
.25	.98	3.51	.87	.24	.04	.10	.25	3.98	1.61
.72	.75	3.08	2.22	.54	.07	.33	.96	1.04	8.00
.65	.07	.96	.62	.05	.00	.96	8.96	.11	774.63
.63	.88	3.33	2.11	.56	.08	.27	.71	1.40	5.07
.27	.44	2.34	.64	.12	.01	.16	.62	1.60	9.00
.77	.48	2.45	1.88	.37	.03	.44	1.60	.62	21.10
.24	.27	1.86	.44	.07	.00	.18	.86	1.16	19.76
.36	.17	1.45	.52	.06	.00	.35	2.16	.46	81.34
.49	.90	3.36	1.63	.44	.06	.20	.54	1.84	3.79
.91	.06	.87	.79	.06	.00	1.47	15.00	.071	555.76
.90	.50	2.52	2.28	.46	.04	.51	1.79	.56	22.32
.52	.32	2.00	1.03	.16	.01	.36	1.62	.62	31.86
.99	.49	2.49	2.46	.49	.04	.56	2.00	.50	25.40
.27	.09	1.07	.28	.02	.00	.35	2.93	.34	202.87
.95	.07	.96	.91	.07	.00	1.39	12.84	.081	093.41
.50	.38	2.20	1.10	.19	.01	.32	1.30	.77	21.32
.28	.91	3.39	.94	.25	.04	.12	.30	3.30	2.08
.53	.46	2.42	1.28	.25	.02	.31	1.14	.88	15.43

table 5.7.a: cylinder problem, data

3	4	4	3	4	4	2	2	3	1
4	2	2	4	3	3	4	4	1	3
1	4	4	2	2	3	1	1	4	1
3	3	3	4	4	4	2	2	3	2
3	1	1	1	1	1	4	4	1	4
3	4	4	4	4	4	2	1	4	1
1	2	2	1	2	2	1	1	4	2
3	3	3	3	3	3	3	3	2	2
1	2	2	1	1	2	1	2	3	2
2	1	1	1	1	1	3	3	2	4
2	4	4	3	3	4	1	1	4	1
4	1	1	2	1	1	4	4	1	4
4	3	3	4	4	3	3	3	2	3
2	2	2	2	2	2	3	3	2	3
4	3	3	4	4	4	4	3	2	3
1	1	1	1	1	1	3	4	1	4
4	1	1	2	2	1	4	4	1	4
2	2	2	3	2	2	2	2	3	3
1	4	4	2	3	3	1	1	4	1
2	3	3	3	3	2	2	2	3	2

table 5.7.b: cylinder problem, data, discreticized

1.00									
-.15	1.00								
-.15	1.00	1.00							
.60	.60	.60	1.00						
.35	.79	.79	.87	1.00					
.02	.96	.96	.72	.86	1.00				
.78	-.61	-.61	.15	-.13	-.46	1.00			
.56	-.78	-.78	-.14	-.40	-.65	.91	1.00		
-.56	.78	.78	.14	.40	.65	-.91	-1.00	1.00	
.30	-.97	-.97	-.46	-.68	-.91	.73	.85	-.85	1

table 5.8: correlations PRINCALS ordinal.

.669	.280	.023	.010	.008	.005	.004	.001	.000

table 5.8: eigenvalues PRINCALS ordinal.

the data matrix, then

$t_{i1} = a_i$,

$t_{i2} = b_i$,

$t_{i3} = 2(\pi b_i)^{\frac{1}{2}}$,

$t_{i4} = 2a_i(\pi b_i)^{\frac{1}{2}}$,

$t_{i5} = a_i b_i$,

$t_{i6} = (2\pi)^{-1} a_i b_i^2$,

$t_{i7} = (2\pi)^{-\frac{1}{2}} a_i b_i^{-\frac{1}{2}}$,

$t_{i8} = a_i b_i^{-1}$,

$t_{i9} = a_i^{-1} b_i$,

$t_{i10} = 2\pi a_i b_i^{-2}$.

The cylinder problem is called the cylinder problem, because the ten variables are used in physics and engineering to describe properties of cylinders. For our purposes this is not important at all. We merely use the fact that a logarithmic transformation of all variables, followed by centering of the transformed matrix, makes the matrix exactly of rank two. Consequently nonmetric components analysis will give a perfect fit in two dimensions.

The data t_{ij} are given in 5.7.a, we have discreticized them in 5.7.b, and we fed this discreticized matrix into PRINCALS with all variables single ordinal, and with two dimensions. Because of the discretization perfect fit is no longer possible. The linear fit, which we compute in the first phase, is .9365, the nonmetric fit is .9489, which is not much higher. The correlations between the variables after the ordinal solution are given in table 5.8, together with the eigenvalues of R(Q)/m. The object scores are transformed linearly in such a way that they maximally resemble table 5.7.a. The plot in figure 5.9 has the numbered elements of 5.7.a connected to the transformed object scores of ordinal PRINCALS. The loadings are given in figure 5.10. We see that variables 1 and 2 are almost orthogonal. Variables 7, 8, and 10 make obtuse angles with variable 2 because b_i has a negative power in the formula for t_{i7}, t_{i8}, and t_{i10}. In the same way variable 9 is obtuse with variable 1, variable 2 and 3 coincide, variable 5 is between 1 and 2, 4 is closer to one than to two, and 6 is closer to two than to one. All this is easily explained by taking logarithms of t_{ij} and looking at the coefficients of log a_i and log b_j.

It is clear that PRINCALS recovers the original structure quite well, even if we have discretisized rather severely. On the other hand the fit of the metric solution already indicates that it also performs surprisingly well.

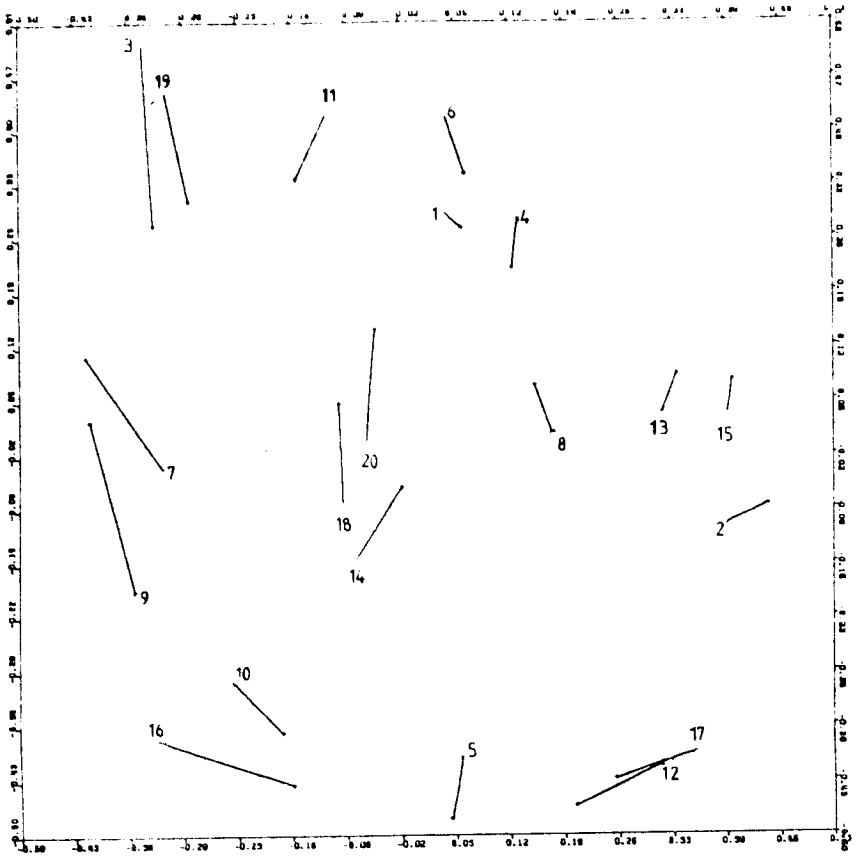


Figure 5.9 The first two columns of table 5.7.a labeled by rownumber, connected with the corresponding PRINCALS object scores

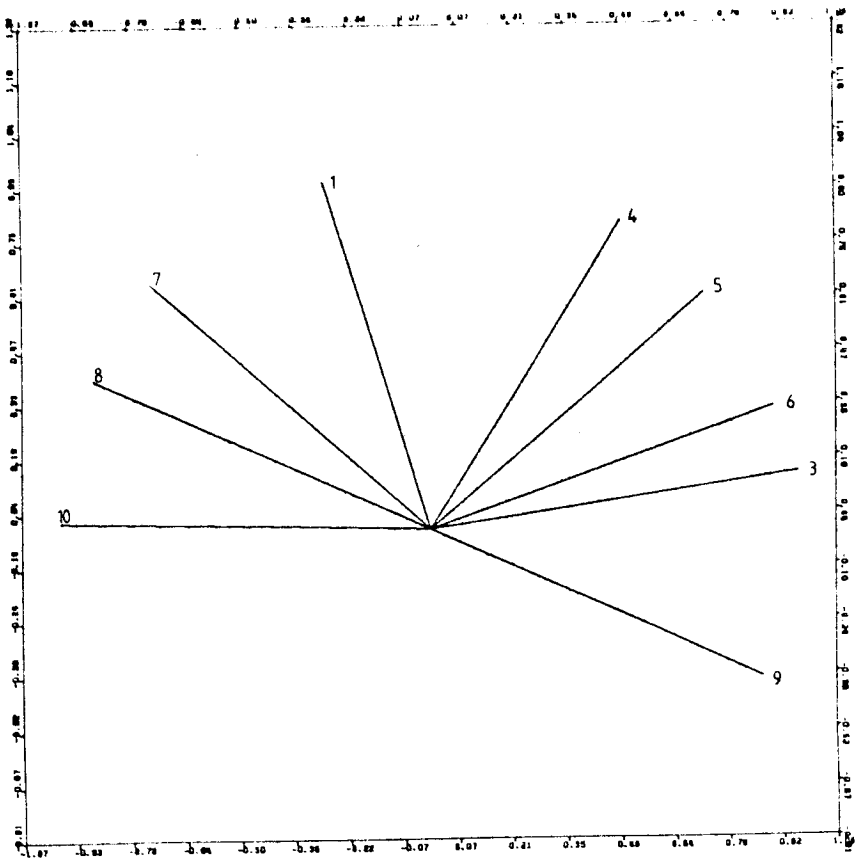


Figure 5.10 Variable loadings for the two-dimensional single ordinal PRINCALS solution of Thurstone's cylinder problem

6: K sets of variables and OVERALS

6.1 Previous work

6.1.1 General linear meet problems

We have already seen in previous chapters that homogeneity analysis can be interpreted as a technique for finding p orthogonal vectors in the meet of the m indicator matrices G_j. This is not necessarily the most natural interpretation of homogeneity analysis, but it suggests an immediate extension in which we find p orthogonal vectors in the meet of m general matrices H_j, by using the loss function

$$\sigma_M(X,Y) = \frac{1}{m} \sum_{j=1}^{m} SSQ(X - H_j Y_j),$$

with the normalization, as in homogeneity analysis, of $u'X = 0$ and $X'X = I$. We suppose, throughout this section, that the H_j are known and in deviations from the mean. We also define matrices $C_{j\ell} = H_j'H_\ell$, and $D_j = H_j'H_j$. As long as the H_j are known matrices it is not necessary to refer to their individual columns (the variables), and we can continue to use the notation we used in homogeneity analysis. Thus in this introductory section there are m sets of variables, and set j consists of k_j variables.

As usual we want to start with a historical overview of techniques which can be considered to be closely related to this general linear meet problem. In this case, however, we have some trouble to find related techniques. The literature is confused, the topic has not been systematically studied before 1960, and a great deal of the relevant papers are unpublished reasearch reports or doctoral dissertations. We shall try to bring some order into the chaos, similar systematizations have been attempted recently by Dauxois and Pousse (1976), Ten Berge (1977), and Van de Geer (1980).

Suppose the H_j are m sets of variables, and $q_j = H_j y_j$ are linear composites. Most optimality criteria are functions of the correlation matrix R(Q). This was first exploited by Kettenring (1971). In fact the most important optimality criteria are functions of the eigenvalues of R(Q). Optimizing such functions means finding optimal vectors of coefficients y_j, we consequently find a single solution to start with. We can stop after this, or proceed to solve the join problem resulting from the contraction $q_j = H_j y_j$. But we can also try to find additional solutions to arrive at a multiple solution. There are two ways in which we can find additional solutions, which are often confused in the literature. The optimal single solution is often found by differentiating the optimality criterion, and by solving the corresponding set of stationary equations (which may involve solving for undetermined multipliers). But the stationary equations may have more than one solution, and the additional solutions can be used to define a multiple solution for y_j. If this is sensible depends

on the properties of the stationary equations. We have seen in chapter 4 that the stationary equations of correspondence analysis can be interpreted in terms of the centroid principle, and they can be interpreted in terms of scoring systems that linearize the regressions. These interpretations make the stationary equations themselves interesting, and this interest is independent from the optimality criterium they were derived from. A second way to derive additional solutions is by solving the optimization problem again with added constraints that prevent reoccurrence of the same solution. In the context of m-set analysis these are often orthogonality constraints. Dauxois and Pousse (1976, p 197-198) have distinguished the weak orthogonality constraints

$$\sum_{j=1}^{m} q_j' q_j^{(s)} = \sum_{j=1}^{m} y_j' D_j y_j^{(s)} = 0,$$

for all previous solutions $y^{(s)}$, and the strong constraints

$$q_j' q_j^{(s)} = y_j' D_j y_j^{(s)} = 0,$$

which must be true for all j and for all previous solutions $y_j^{(s)}$. If we impose the strong constraints there can only be k_j solutions for a set with k_j variables, if we want more than k_j solutions than the additional solutions will have $y_j = 0$ for sets with too few variables. It is a useful exercise to write down the equations of homogeneity analysis using strong orthogonality constraints, and to consider what happens if in 'strong HOMALS' we have $k_j = 2$ for all j.

Systematizations of Kettenring (1971) and Dauxois and Pousse (1976) all introduce additional solutions by using orthogonality constraints and by computing them successively. We have already discussed the distinction between successive and simultaneous computation in chapter 5, and we have indicated that in some of the simpler cases they give the same results, in some of the more complicated cases they do not. In the context of matching configurations with orthogonal rotations, which is closely related to multiple set canonical analysis, Van de Geer (1968) en Ten Berge (1977) have also distinguished systematically between successive and simultaneous solutions. To introduce simultaneous computation in our context we merely have to define $Q_j = H_j Y_j$, with Y_j a $k_j \times p$ matrix, and to collect the $p \times p$ matrices $R_{j\ell} = Y_j' C_{j\ell} Y_\ell$ in a supermatrix R(Q) of dimension $(mp) \times (mp)$. Optimal simultaneous solutions can be computed if we have criteria which are functions of the supermatrix R(Q) or of its eigenvalues. In computing these simultaneous solutions we also impose constraints, which are strong or weak. Strong simultaneous orthogonality constraints require that $R_{jj} = I$ for all j, the weak version merely requires that $\sum R_{jj} = mI$. We have also seen in chapter 5 that it is possible in some circumstanc

stationary equations, but this rapidly becomes both wasteful and uninteresting.

6.1.2 Some optimality criteria

Steel (1951) proposes to minimize the determinant of R(Q), which is equal to the product of its eigenvalues. Clearly $0 \leq \det R(Q) \leq 1$, with $\det R(Q) = 0$ if and only if R(Q) is singular and $\det R(Q) = 1$ if and only if R(Q) is the identity. Steel derives stationary equations for minimizing the determinant, but he does not propose a specific algorithm. Chang and Bargmann apply Steel's criterion to quantification of categorical data, using a quasi-Newton algorithm. Kettenring (1971) discusses Steel's work, baptizes it with the acronym GENVAR, and proposes a relaxation algorithm closely related to alternating least squares. If we partition R(Q), using r_j for the vector of correlations of all other variables with j and R_j for the matrix of intercorrelations of all variables except j, then we can apply the Schur-formula for partitioned determinants in the form

$$\det R = (1 - r_j' R_j^{-1} r_j) \det R_j.$$

Thus minimizing the GENVAR-criterion over y_j, for fixed other y_ℓ, can be done by maximizing

$$r_j' R_j^{-1} r_j = y_j' \{ \sum_{\mu \neq j} \sum_{\nu \neq j} r^{\mu\nu} C_{j\mu} y_\mu y_\nu' C_{\nu j} \} y_j,$$

with

$$r^{\mu\nu} = (R_j^{-1})_{\mu\nu},$$

over y_j satisfying the usual $y_j' D_j y_j = 1$. This defines a generalized eigenvalue problem (cf appendix A), which is easily solved for y_j. Observe that y_j minimizes GENVAR, for fixed other y_ℓ, if it maximizes the multiple correlation of q_j with the other q_ℓ. It is also possible, with the same algorithm, to maximize the determinant, by finding the smallest generalized eigenvalue, which minimizes the multiple correlation in each step. In that case we want R(Q) to be as close as possible to the identity, for example because we want to use it for prediction purposes in the second step. Kettenring (1971) uses GENVAR in combination with strong orthogonality constraints to compute successive solutions, but we can also use essentially the same algorithm to compute p solutions simultaneously. Maximizing or minimizing the determinant of the supermatrix R(Q) can also be done by using the Schur-formula. The subproblem is maximizing or minimizing

$$\det Y_j' \{ \sum_{\mu \neq j} \sum_{\nu \neq j} C_{j\mu} Y_\mu R^{\mu\nu} Y_\nu' C_{\nu j} \} Y_j,$$

over all Y_j satisfying the strong orthogonality constraints $Y_j' D_j Y_j = I$. Again this defines a generalized eigenvalue problem, which is fairly easily solved. Both Steel and Chang and Bargmann try to defend their choice of the GENVAR

criterion by appealing to likelihood ratiotests for the multinormal distributi but the resemblance seems merely formal to us. It is better to treat GENVAR as just a possible weighted combination of the eigenvalues, and one which tend to emphasize the smallness of the smaller ones.

Horst (1961a,b) proposes four different methods, of which he clearly prefers the one we discuss now. The other three are treated as computationally conveni modifications. Kettenring uses the acronym SUMCOR, and the criterion is simply the sum of the correlations in R(Q). Observe that this is not a function of the eigenvalues, and that it depends on the sign of the elements in R(Q). The method is also discussed by Van de Geer (1968) in the matching context, by Nevels (1974), by Ten Berge (1977) also for matching, by Dauxois and Pousse (1976, p 235-244), and by Van de Geer (1980) who uses the acronym ORTHOCAN. Algorithms are generally of the 'block relaxation' type, in which the coefficients for each set are the blocks. It is clear that maximizing the SUMCOR criterion over y_j, with the other y fixed, amounts to maximizing

$$y_j' \sum_{\ell \neq j} C_{j\ell} y_\ell,$$

over $y_j' D_j y_j = 1$, which is a multiple linear regression problem, with solution proportional to

$$D_j^{-1} \sum_{\ell \neq j} C_{j\ell} y_\ell .$$

Successive solutions are computed by using strong orthogonality constraints, we can also minimize the sum of the correlations with the same algorithm (simp change all signs in the previous update of y_j), and we can compute simultaneou solutions by generalizing to the criterion

$$\sum_{j=1}^{m} \sum_{\ell=1}^{m} \operatorname{tr} Y_j' C_{j\ell} Y_\ell,$$

which must be maximized over all $Y_j' D_j Y_j = I$. The subproblems are now orthogona Procrustus problems (Cliff, 1966).

The first 'approximate' method of Horst is introduced by him as follows: "The second model specifies that the intercorrelations of the first transformed variables for the m sets shall give the best least square approximation to a ra one matrix." (Horst, 1961b, p 332). Carroll (1968) seeks, in our notation, an n vector x and composites $q_j = H_j y_j$ such that the sum of the squared correla between x and the q_j is maximized. Both formulations undoubtedly sound familia because they correspond to two of the possible ways to introduce homogeneity analysis. Kettenring uses the acronym MAXVAR, and provides the interpretation in terms of maximizing the largest eigenvalue of R(Q). He also uses this interpretation to introduce MINVAR, which minimizes the smallest eigenvalue

of R(Q). We already encountered MINVAR in chapter 5. It is clear from our chapters 3, 4, and 5 that both successive and simultaneous optimization of this criterion is possible, that it should be used prefereably in combination with the weak orthogonality constraints, and that the optimization problems are in general eigenvalue problems, which is a considerable advantage, both computationally and theoretically. This is also pointed out by Carroll (1968), and by Dauxois and Pousse, who criticize the complicatedness of the other criteria we have discussed in this section (1976, p 194, p 210-212). Van de Geer (1980) discusses MAXVAR using the acronym GENCAN. Observe that multiple solutions in this case can be computed from the stationary equations without using orthogonality. In fact orthogonality follows from the fact that two solutions (with different eigenvalues) satisfy the stationary equations.

For completeness we must mention two other methods. Kettenring (1971) also introduces SSQCOR, with criterion equal to the sum of squares of the correlations in R(Q), which is equal to the sum of squares of the eigenvalues of R(Q). This leads to subproblems of the form: maximize or minimize

$$y_j'\{\sum_{\ell\neq j} C_{j\ell}y_\ell y_\ell' C_{\ell j}\}y_j,$$

over $y_j'D_jy_j = 1$, which has an obvious simultaneous generalization. Computationally SSQCOR is similar to, although much simpler than, GENVAR. We expect SSQCOR to emphasize the largeness of the larger eigenvalues, while GENVAR emphasize the smallness of the smaller ones. Clearly a similar dual relationship exists between MAXVAR and MINVAR. The last method that we mention has not been discussed before in the literature, although it is familiar from ch.5. If we generalize PRINCALS with single variables to situations in which the indicator matrices G_j are replaced by general, but constant, H_j, then we maximize the sum of the p largest eigenvalues of R(Q), or, equivalently, we minimize the sum of the m - p smallest ones. This generalizes MAXVAR and MINVAR at the same time, and introduces some intermediate possibilities.

The different criteria have not been compared on a large scale. Horst (1961b) compares SUMCOR and MAXVAR. Kettenring uses the same example, and also computes GENVAR and SSQCOR. For this example all techniques give very similar results. Haven and Ten Berge (see Ten Berge, 1977, chapter IV) have some comparisons in a matching context. It seems to us that more research is needed, either to show that canonical analysis with m matrices H_j is stable under selection of criterion, or to indicate in which situations it can make a lot of difference which criterion we use. Dauxois and Pousse (1976) have already shown that if the techniques are extended to random vectors, then the solutions will all be the same if the random vectors are multinormal. As long as there are no results

available which tell us how to choose criterion or loss function, we think that mathematical and computational convenience indicate that MAXVAR and MINVAR are pretty good candidates.

We illustrate this point with a small example. The correlation matrix in 6.1.a is taken from De Leeuw and Stoop (1979, p 142). The basic data were from the 'From year to year' study, discussed more extensively in our later chapter with large examples. The correlation matrix is actually a correlation matrix taken from a larger matrix of order 25, in which the quantifications used to compute the correlations were derived from HOMALS. The first five variables are profession father - education father - education mother - number of children in the family - degree of urbanization. They are clearly exogeneous variables describing the situation in the family. The second set, of two variable indicates whether the child had to do one or more grades more than once and how many children there were in his class in sixth grade. The third set of four variables gives teachers advice on secondary education, score on a school achievement test, choice of secondary education, and attained level of secondary education.

The resulting 11 variables, partitioned in three sets, were analyzed with six canonical analysis techniques. We shall use new acronyms, which improve those of Kettenring, because they indicate that everything that can be maximized can also be minimized. First there is MAXMAX which maximizes the largest eigenvalue (previously known as MAXVAR or GENCAN or CARROLL), MINMIN (formerly MINVAR) minimizes the smallest eigenvalue of the correlation matrix. MAXSUM (Kettenring's SUMCOR) maximizes the sum of the correlations, MINSUM minimizes this sum. MINDET (used to be GENVAR) minimizes the determinant, and MAXSSQ (was SSQCOR) maximizes the sum of the squared correlations (or the sum of squares of the eigenvalues, or the variance of the eigenvalues). Using our system of acronyms the reader may wonder what happened to MINMAX, MAXMIN, MAXDET, and MINSSQ. The answer is simple. In this example, as in many small examples, we can choose the y_j in such a way that the correlation matrix becomes the identity. This identity matrix solves the four problems we have not mentioned.

The solutions for the y_j are given in table 6.1.b. For MAXMAX and MINMIN we have $\sum y_j'D_jy_j = 3$, for the other six techniques we have $y_j'D_jy_j = 1$ for all j. We do not give any substantial interpretations here, we merely remark that the solutions for the canonical weights are extremely similar. In table 6.1.c this also becomes obvious, we have computed all six criteria for all six solutions. The only problem here is that MAXSUM and MINSUM depend on the sign of the correlations between the three composites. For MAXSUM we consequently

choose all correlations positive, for MINSUM we get the smallest value throughout by choosing the correlation between the third composite and the other two to be negative. In table 6.1.d the numbers in 6.1.c are replaced by rank numbers. For this example MAXSUM and MINSUM do not perform very well, which is a nice result. MAXMAX and MINMIN are fair, MINDET and MAXSSQ are possibly better.

6.2 Specific theory

6.2.1 Loss function

The loss function in 6.1.1 is appropriate enough if the matrix H is constant, if the variables must be quantified or transformed, however, we need more complicated notation. Define

$$\sigma_M(X,Y) = \frac{1}{K} \sum_{k=1}^{K} SSQ(X - \sum_{j \in J_k} G_j Y_j),$$

where we assume that the index set $\{1,\ldots,m\}$ of the variables is partitioned into K sets J_k $(k=1,\ldots,K)$. This notation has the advantage that the sets J_k do not have to consist of consecutive integers, it can also be used if the J_k do not exhaust $\{1,\ldots,m\}$ or in which the J_k are not exclusive. By constructing an indicator supermatrix G^k and a supermatrix Y^k for each of the sets (k is being used as a subscript not as a power), we can also write the loss function as

$$\sigma_M(X,Y) = \frac{1}{K} \sum_{k=1}^{K} SSQ(X - G^k Y^k),$$

which makes it again very much like the loss function in 6.1.1. The reason why we usually prefer to use the individual G_j and not the G^k, is that we prefer to use the special properties of complete indicator matrices in our algorithms, and that we restrict each of the Y_j individually according to measurement level.

6.2.2 Normalization

We have seen in 6.1 that previous techniques normalize the Y_j, while we normalize X. The exception is Carroll (1968), who also normalizes X, and because at least part of the French data analytic work (Masson, 1974, Saporta, 1975, Dauxois and Pousse, 1976) is inspired by this small note of Carroll's it follows a similar approach. Kettenring (1971) also mentions Carroll's work as one of his inspirations. As we have seen in chapter 3 Carroll's work fits naturally into the tradition of homogeneity analysis started by Horst, Edgerton and Kolbe, Richardson, Wilks, and Guttman. It seems interesting to see how normalization on Y is related to our normalization $u'X = 0$ and $X'X = I$, even in the general case in which the Y_j can be restricted in various ways.

The key observation here is that we shall only be interested in restrictions on Y_j which are defined in such a way that if Y_j satisfies the restrictions, then $Y_j T$ satisfies the restrictions for all $p \times p$ matrices T. For multiple nominal variables

0.66									
0.46	0.49								
0.14	0.12	0.09							
0.16	0.12	0.15	0.17						
0.10	0.13	0.08	0.09	-0.05					
0.07	0.04	0.03	0.01	0.10	0.02				
0.37	0.40	0.30	0.15	0.07	0.33	0.08			
0.37	0.36	0.26	0.14	0.11	0.31	0.11	0.72		
0.46	0.47	0.35	0.19	0.09	0.34	0.09	0.80	0.71	
0.45	0.45	0.35	0.18	0.07	0.38	0.10	0.74	0.69	0.81

table 6.1.a: correlation matrix 11 variables, three sets.

MAXMAX	MINMIN	MAXSUM	MINSUM	MINDET	MAXSSQ
.411	.476	.398	.520	.449	.427
.483	.434	.490	.354	.443	.464
.214	.240	.208	.247	.230	.221
.272	.176	.286	.088	.209	.243
-.094	.019	.113	.131	-.024	-.063
.780	.572	.963	.992	.968	.963
.205	.104	.250	.101	.232	.252
.032	.008	.040	.029	.012	.020
.092	.106	.086	.111	.081	.080
.481	.583	.389	.387	.445	.431
.631	.670	.559	.549	.530	.539

table 6.1.b: canonical weights for six techniques

	MAXMAX	MINMIN	MAXSUM	MINSUM	MINDET	MAXSSQ
MAXMAX	1.7555	0.3914	5.1990	1.4242	0.5862	3.9628
MINMIN	1.7482	0.3868	5.1718	1.4018	0.5849	3.9540
MAXSUM	1.7553	0.3933	5.1994	1.4294	0.5878	3.9608
MINSUM	1.7282	0.3910	5.1090	1.3934	0.5952	3.9154
MINDET	1.7530	0.3875	5.1882	1.4090	0.5838	3.9619
MAXSSQ	1.7550	0.3893	5.1958	1.4170	0.5846	3.9638

table 6.1.c: 6 optimality criteria (columns) for 6 techniques (rows).

	MAXMAX	MINMIN	MAXSUM	MINSUM	MINDET	MAXSSQ
MAXMAX	1	5	2	5	4	2
MINMIN	5	1	5	2	3	4
MAXSUM	2	6	1	6	5	5
MINSUM	6	4	6	1	6	6
MINDET	4	2	4	3	1	3
MAXSSQ	3	3	3	4	2	1

table 6.1.d: as table 6.1.c, techniques ranked columnwise.

in which the Y_j are not restricted, this is trivially true. For single variables, in which we require $Y_j = y_j a_j'$, with $y_j \in K_j$, this condition is also true: if $Y_j = y_j a_j'$ then $Y_j T = y_j a_j' T$, which can be written as $Y_j T = y_j a_j'$. Observe that for multiple ordinal variables the condition is no longer true: if Y_j satisfies the constraints then $Y_j T$ will generally also satisfy the constraints for all <u>diagonal</u> matrices T. If the condition is true, then we can also minimize

$$\sigma_M(X,Y,T) = \frac{1}{K} \sum_{k=1}^{K} SSQ(X - \sum_{j \in J_k} G_j Y_j T_j),$$

over X, Y_j restricted as before, and the new variables T_j, with the same result and solution as before. We shall use a slightly different formulation, by imposing the condition that the T_j are the same within sets, and by writing them as T^k. Also define

$$Q_k = \sum_{j \in J_k} G_j Y_j,$$

and by writing $\sigma_M(*,Y)$, as usual, for the minimum of $\sigma_M(X,Y)$ over X satisfying $u'X = 0$ and $X'X = I$. We also write $\sigma_M(*,Y,*)$ for the minimum of $\sigma_M(X,Y,T)$ over X with $u'X = 0$ and $X'X = I$ and over T, where

$$\sigma_M(X,Y,T) = \frac{1}{K} \sum_{k=1}^{K} SSQ(X - Q_k T^k).$$

Our theory so far tells us that minimizing $\sigma_M(*,Y)$ over restricted Y is equivalent to minimizing $\sigma_M(*,Y,*)$ over restricted Y, where the restrictions of both problems are the same. Minimizing $\sigma_M(*,Y)$ is our original problem, and minimizing $\sigma_M(*,Y,*)$ can easily be shown to be equiavalent to maximizing the sum of the p largest eigenvalues of the matrix

$$\sum_{k=1}^{K} Q_k (Q_k'Q_k)^+ Q_k'.$$

The proof is immediate from the Eckart-Young theorem. The eigenvalues of this matrix are also equal to the generalized eigenvalues of the generalized eigenvalue problem with matrices C and D, where C has submatrices $C^{k\ell} = Q_k'Q_\ell$, and where D is 'diagonal' with submatrices $D^k = Q_k'Q_k$. The formulation and the notation have been chosen in such a way that the problem is as similar as possible to homogeneity analysis, but of course we must always remember that Q_k is not a constant matrix because it is a function of the Y_j. It is also clear from this formulation that we now can suppose without loss of generality that $\sum D^k = I$, which is more explicitly

$$\sum_{k=1}^{K} \sum_{j \in J_k} \sum_{\ell \in J_k} Y_j' G_j' G_\ell Y_\ell = I,$$

or

a	p	u
b	q	v
a	r	v
a	p	u
b	p	v
c	p	v
a	p	u
a	p	v
c	p	v
a	p	v

table 6.2.a
data matrix

a	b	c	p	q	r	u	v
1	0	0	1	0	0	1	0
0	1	0	0	1	0	0	1
1	0	0	0	0	1	0	1
1	0	0	1	0	0	1	0
0	1	0	1	0	0	0	1
0	0	1	1	0	0	0	1
1	0	0	1	0	0	1	0
1	0	0	1	0	0	0	1
0	0	1	1	0	0	0	1
1	0	0	1	0	0	0	1

table 6.2.b
indicator supermatrix

apu	apv	aqu	aqv	aru	arv	bpu	bpv	bqu	bqv	bru	brv	cpu	cpv	cqu	cqv	cru	crv
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

table 6.2.c: interactive indicator matrix.

	a	b	c	p	q	r	u	v
apu	1	0	0	1	0	0	1	0
apv	1	0	0	1	0	0	0	1
aqu	1	0	0	0	1	0	1	0
aqv	1	0	0	0	1	0	0	1
aru	1	0	0	0	0	1	1	0
arv	1	0	0	0	0	1	0	1
bpu	0	1	0	1	0	0	1	0
bpv	0	1	0	1	0	0	0	1
bqu	0	1	0	0	1	0	1	0
bqv	0	1	0	0	1	0	0	1
bru	0	1	0	0	0	1	1	0
brv	0	1	0	0	0	1	0	1
cpu	0	0	1	1	0	0	1	0
cpv	0	0	1	1	0	0	0	1
cqu	0	0	1	0	1	0	1	0
cqv	0	0	1	0	1	0	0	1
cru	0	0	1	0	0	1	1	0
crv	0	0	1	0	0	1	0	1

table 6.2.d:
transformation or
design matrix.

$$\sum_{k=1}^{K} (G^k Y^k)' G^k Y^k = I$$

which are the familiar weak simultaneous orthogonality constraints.

Some familiar special cases can be recovered easily. If J_k contains only a single variable j, then $Q_k = G_j Y_j$. If the Y_j are not further restricted (multiple nominal) we recover homogeneity analysis. If all variables are single, then $Q_k = q_j a_j'$, and

$$Q_k (Q_k' Q_k)^+ Q_k' = \frac{q_j q_j'}{q_j' q_j}$$

which gives principal components analysis.

6.2.3 Interactive variables

We have seen in chapter 2 that it is sometimes useful to combine m variables with $k_1, \ldots, k_m$ categories into a single variable with $k_1 \times \ldots \times k_m$ categories. The m indicator matrices G_j, which are $n \times k_j$, are replaced by a single indicator matrix G, which is $n \times (k_1 \times \ldots \times k_m)$. Now suppose we replace all indicator matrices in set k by the interactive indicator matrix $\underline{G}_k$ (not the same as G^k in 6.2.1, because this is the 'sum' of the G_j, which is $n \times (k_1 + \ldots + k_m)$) We rewrite the loss function as

$$\sigma_M(X,Y) = \frac{1}{K} \sum_{k=1}^{K} SSQ(X - \underline{G}_k \underline{Y}_k),$$

where we must remember that the number of rows of $\underline{Y}_k$ is now equal to the possible number of profiles for all variables in group k. The matrix $\underline{G}_k$ is a proper indicator matrix, where G^k is not, because it consists of a number of proper indicator matrices, one for each variable in the set. It is not difficult to prove that there always exists an S_k such that $\underline{G}_k S_k = G^k$. In fact it is easy to construct S_k explicitly, S_k is simply the matrix which assigns to each profile in $\underline{G}_k$ the appropriate row in G^k. Thus S_k is $(\Pi k_j) \times (\Sigma k_j)$, and is explicitly given by

$$S_k = (\underline{G}_k' \underline{G}_k)^{-1} \underline{G}_k' G^k.$$

If there are m_k variables in the set k, then $S_k u = m_k u$. By using $u' \underline{G}_k S_k = u' G^k$ we also see that S_k transforms profile frequencies in marginal frequencies.

Table 6.2 gives a small example. The data matrix for a set of variables is expanded to the indicator supermatrix G and the interactive indicator matrix $\underline{G}$. Then S is constructed. This construction so far merely capitulates material from previous chapters, but it is especially important at this point, because using $\underline{G}_k$ shows us that a group of variables can be reduced to a single variable, and that consequently all K-set problems can be reduced to homogeneity analysis

problems (while, conversely, we have seen that all homogeneity problems are K-set problems with only one variable in each set). Of course using $\underline{G}_k$ may not be very clever in some situations, because it may have far too many columns (the arguments against the discrete MVA approaches connected with the empty cell problem discussed in chapter 1 are also relevant here), on the other hand using $\underline{G}_k$ may be quite illuminating in some examples (such as the students and politics example in chapter 4). The matrix S_k is important, because $\underline{G}_k S_k = G^k$ implies, of course, that $\underline{G}_k S_k Y^k = G^k Y^k$, and thus $\underline{G}_k \underline{Y}_k = G^k Y^k$ if $\underline{Y}_k = S_k Y^k$. The restrictions $\underline{Y}_k = S_k Y^k$ can be interpreted as requiring that the category quantification of a profile must be additive, i.e. it must be the sum of quantification of the categories in the profile. Thus S_k is a sort of <u>design matrix</u>, and we can now say that K-set canonical analysis can be interpreted as homogeneity analysis with linear restrictions. Observe that for sets of single variables the restrictions are $\underline{Y}_k = S_k Y^k$ with rank-one restrictions on the submatrices of Y^k. Thus we can better write

$$\underline{Y}_k = \sum_{j\in J_k} S_{kj} Y_j = \sum_{j\in J_k} S_{kj} y_j a_j',$$

where the S_{kj} are the submatrices of S_k, which are $(\pi k_j) \times k_j$, compare table 6.2 The S_{kj} are indicator matrices, which satisfy $\underline{G}_k S_{kj} = G_j$.

6.2.4 <u>Missing data</u>

It is obvious how to generalize options II and III for handling missing data. Option I, as usual, presents some problems. The main problem is that in homogeneity analysis and in principal components analysis each variable defines a set. Consequently if an observation on a variable is missing, it is missing automatically for all variables in the set. This is no longer true if the sets contain several variables. The simplest genaralization we can think of is

$$\sigma_M(X,Y) = \frac{1}{K} \sum_{k=1}^{k} \operatorname{tr}(X - \sum_{j\in J_k} G_j Y_j)' M^k (X - \sum_{j\in J_k} G_j Y_j),$$

with M^k the minimum of the M_j in set k. This means that if an individual has a missing observation on some variable, then the set that contains this variable does not contribute to the loss for this individual.

In the case of principal components analysis $M^k = M_j$. This gives the loss funct used with option I in chapters 3 and 5.

6.2.5 <u>Algorithm</u>

As usual we treat the general case, with weighting matrices M^k for missing data, and with incomplete indicator matrices. If all M_j are equal to the identity and all G_j are complete, then we treat this as a special case of the formulas. The fitting of X for given Y_j does not present any new problems, and we refer to

chapter 5, section 8. The fitting of Y^k from the loss component

$$\mathrm{tr}(X - G^kY^k)'M^k(X - G^kY^k)$$

does present new problems, because G^k is an indicator supermatrix, and not a proper indicator matrix. This implies that in general $(G^k)'G^k$ is not diagonal, which complicates the construction of

$$\tilde{Y}^k = ((G^k)'M^kG^k)^+(G^k)'M^kX$$

Although it is possible, at least in principle, to base an algorithm on computing Y^k first, and then proceeding as in chapter 5, this will tend to become expensive. It also will only be useful for multiple variables, because for single variables we use rank one and cone restrictions defined in terms of the Y_j, which do not look very natural in terms of the Y^k. For these reasons it is probably better to fit the Y_j seperately. Define

$$\underline{X}_\ell = X - (\sum_{j\in J_k} G_jY_j - G_\ell Y_\ell).$$

then

$$X - G^kY^k = \underline{X}_\ell - G_\ell Y_\ell.$$

Thus minimizing $\sigma_M(X,Y)$ over Y, with X and the other Y_j fixed, can be done by minimizing

$$\mathrm{tr}(\underline{X}_\ell - G_\ell Y_\ell)'M^k(\underline{X}_\ell - G_\ell Y_\ell),$$

and this can be readily done by defining

$$\tilde{Y}_\ell = (G_\ell'M^kG_\ell)^+G_\ell'M^k\underline{X}_\ell,$$

and by proceeding further as in chapter 5. The matrix which must be inverted is now small (number of nonmissing categories) and diagonal. In keeping with previous notation we can define here $D_\ell = G_\ell'M^kG_\ell$. This makes it possible to use the same formulas as in chapter 5 for additional loss components for single variables. The algorithm does not have any more interesting features. The only thing that is perhaps still worth mentioning is the update formula when going from variable j in set k to variable j+1 in set k. Then

$$\underline{X}_{j+1} = \underline{X}_j + (G_jY_j - G_{j+1}Y_{j+1}).$$

6.3 The OVERALS program

Very little can be said at the moment about the program OVERALS, which is currently being written to fit the algorithms of 6.2.5. The details are more or less obvious from the theory in 6.2, and also if we consider part of our basic philosophy. The basic program in our series is HOMALS, it is the oldest program, it has been tested in many ways, and many stability and gauging results are known for this

technique. PRINCALS is constructed in such a way that it is essentially the sam as HOMALS, with an extra inner cycle if there are single variables. If all vari are multiple nominal PRINCALS performs exactly the same computations as HOMALS, also gives the same output. OVERALS is constructed in the same way. We take the PRINCALS program and add something for the case that some sets contain more tha one variable. If suitable options are chosen than OVERALS 'degenerates' to PRIN and with other options OVERALS becomes HOMALS. Thus OVERALS generalizes PRINCAL which generalizes HOMALS, and this generalization must be interpreted in a very literal sense. In the next two chapters we shall discuss some techniques which are more special, and consequently do not fit naturally into this chain.

6.4 An example

The data for this example were taken from Cailliez and Pages (1976, p 277-293). Twenty- four fish were placed in three aquariums, which were contaminated with radio-active strontium. The three aquariums were the same, but the fish stayed for a short period in the first aquarium (fish number 1-8), for a longer time in the second aquarium (fish number 9-17), and for the longest time in the third aquarium (fish number 18-24). Variables 1-9 are measures of radioactivity of various body parts of the fish after the experiment, variables 10-16 are measurements of the fish, variable 17 indicates the aquarium. We have used a discreticized version of the data matrix, also given by Caillez and Pages (lc, p 280-284). They apply metric component analysis to the first 16 variables, and find that in the plot of the object scores it is easy to separate the fish from the different aquaria. Within the aquarium clusters we find fish of the same size close together. Fish 21 and 24, which are in aquarium 3, are close to the fish in aquarium 2. There are two clusters of variables in the plot of the loadings. The first cluster consists of variables 10-14, these all measure the size of the fish rather directly. The second cluster are the variables 1,2,3,4,8, these are measurements of radio-activity of hard tissues. The two clusters are almost at right angles in the plane of the first two principal axes, which explain about 70% of the variance. The same data were also analyzed by Bouroche and Saporta (1980, p 109-121). They used discriminant analysis with 15 predictors (variable 7 was not used) to predict aquarium membership. This could be done very well, the three centroids are in the edges of an equilateral triangle, and the fish are close to their aquarium. Again the same two groups of variables can be distinguished.

The categorized data matrix is given in table 6.2 Observe that there is no fish number 17, because it died during the experiment. All variables, except the last one, have 10 categories, but many categories are empty. We first

analyze the three sets 1-9, 10-16, and 17, with the first 16 variables single numerical and the last one multiple nominal, in two dimensions. We then change the measurement level of the first 16 variables to single ordinal, and repeat the analysis. Observe that with 23 observations 10 categories for each variables is quite a lot, and the ordinal option could very well behave in the way the continuous ordinal option is expected to behave. The loss of the numerical solution is .406, which is the average of .259, .650, and .310. The loss of the ordinal solution is .010, which is the average of .003, .015, and .012. In figure 6.1 the object scores for the 23 fish are given, the numbers correspond with the numerical solution, the crosses at the end of the line with the ordinal solution. The clustering of the aquaria is clear enough in the numerical case, in the ordinal case it seems that perfect fit is not attained only because we stopped iterations too soon (computing was done in APL, which is expensive, convergence was painfully slow, in each case we stopped after approximately 55 iterations, the change in the loss function from one iteration to the next was in both cases about 20×10^{-5} at this point). The canonical weights for the two solutions are given in figure 6.2 , again the numbered endpoints correspond with the metric solution. No clear picture emerges, although it is clear that variables 10 (weight) and 11 (length) are different from the others. We have to remember, however, that the within-set correlations are very high, and this often results in instability and difficulty in the interpretation for the canonical weights. It is more informative to look at the canonical components (the correlations of the canonical variables with the original variables or the quantified variables). There are three sets of variables, consequently three pairs of canonical variables, and a fourth pair, which is their orthonormalized average X. Table 6.3.a gives the components for the numerical solution, and table 6.3.b gives them for the ordinal solution. The four subtables of 6.3.b do not differ, because of the very good fit. We have plotted table 6.3.a, first three subtables, in figure 6.3. Each variable defines three points, connected with lines to the origin, and also connected with lines to their centroid, which defines a fourth point (in general different from the corresponding row of X, which is not used). We can see from the figure that there are four groups of variables. The first one consist of the size-variables, of which especially 10-14 are important. The second group consists of variables 3,4,8 (radio-activity of gill-covers, fins, scales), the third group of variables 1 and 2 (radioactivity of eyes and gills), and the fourth group of variables 5,6,9 (radio-activity of liver, gullet, and muscles). Variable 7 (radioactivity of kidneys) does not belong with the other variables 1-9, which is probably why Bouroche and Saporta eliminated it. Clearly our classification of the variables refines the one given by Caillez and Pages, and our technique gives results which are intermediate between discriminant analysis and principal components analysis.

2	2	2	2	1	1	10	1	1	10	9	10	10	10	10	10	1
2	1	1	1	4	1	5	1	2	9	10	10	8	8	8	7	1
1	1	2	2	1	2	4	1	2	10	10	10	8	9	9	10	1
1	2	1	1	1	3	5	1	2	10	10	10	9	10	7	10	1
1	1	2	2	2	3	1	3	1	2	1	1	1	4	1	4	1
1	1	1	2	2	1	4	1	1	2	3	3	2	3	3	4	1
1	1	1	1	2	1	1	2	1	2	2	2	3	3	3	4	1
2	2	1	2	2	1	5	1	4	1	2	4	2	1	2	1	1
3	3	2	4	6	2	4	2	3	3	3	4	3	5	4	7	2
6	6	4	3	7	2	4	2	5	4	6	7	6	5	4	4	2
3	3	3	3	3	6	4	2	2	4	4	4	4	6	4	10	2
4	5	2	3	3	3	4	8	5	3	2	3	3	5	5	7	2
4	6	3	3	5	5	4	2	10	4	5	5	4	5	10	7	2
7	3	2	3	7	5	4	2	2	1	1	1	1	3	4	4	2
3	3	2	3	2	3	4	2	4	5	6	6	3	5	8	7	2
4	4	2	2	4	1	4	3	6	5	8	5	6	6	9	7	2
?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	2
10	10	10	9	5	1	7	6	3	3	3	4	3	3	5	4	3
7	7	9	9	3	9	1	7	2	2	2	3	2	3	4	4	3
10	9	8	10	10	1	1	10	10	1	1	2	2	2	4	1	3
4	5	4	5	3	1	3	1	2	7	7	8	7	9	7	10	3
7	7	10	8	1	1	6	9	3	5	5	6	4	8	6	4	3
7	7	9	9	7	10	10	7	2	5	5	5	4	5	6	7	3
6	2	6	6	2	1	4	4	1	6	5	5	5	7	5	10	3

table 6.2: Radio-active fish from an experiment by Amiard.
Data categorized by Cailliez and Pages.

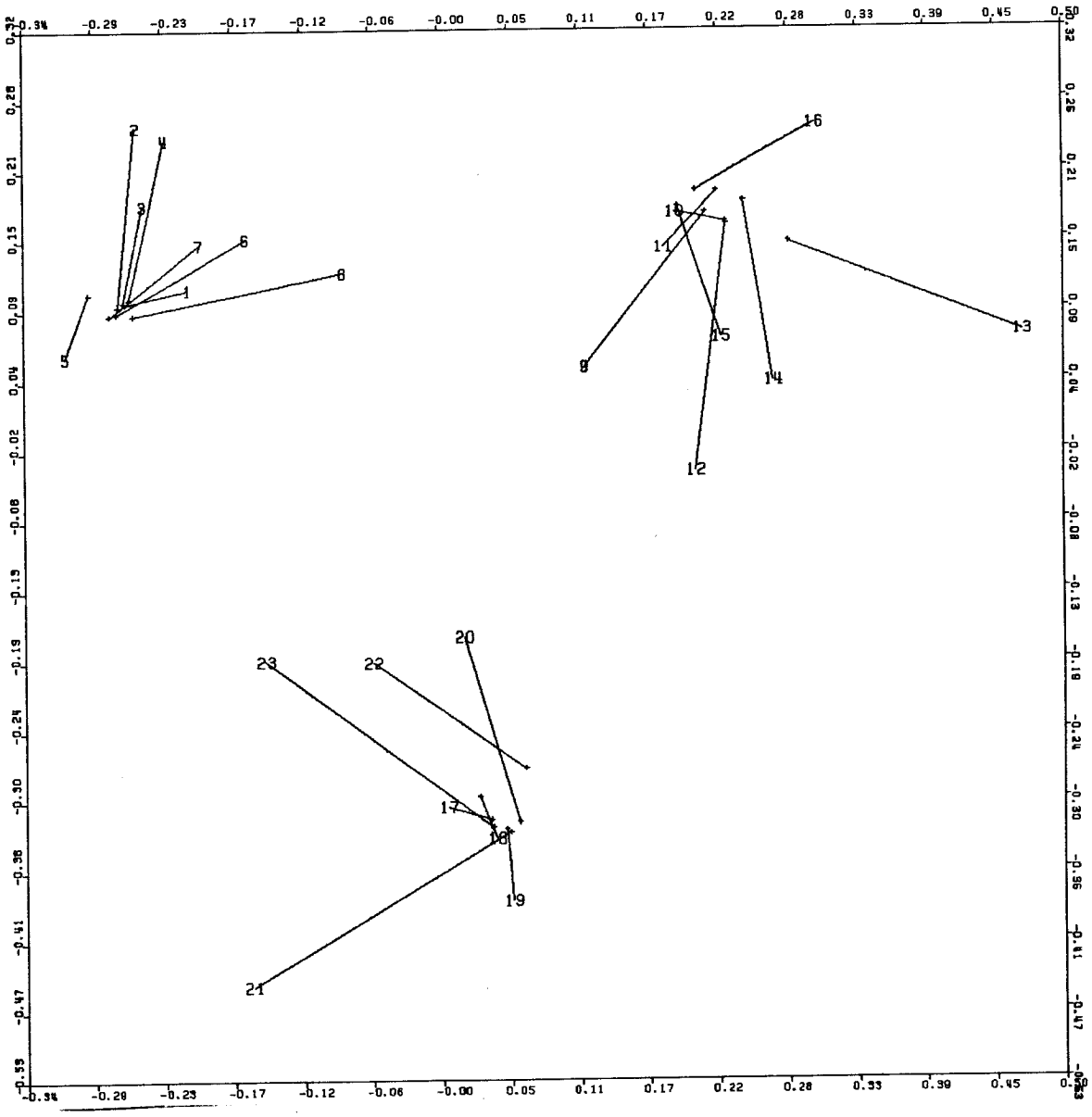


Figure 6.1 **Object scores for the fish-data in the numerical solution (labeled) connected with the corresponding points in the ordinal solution**

.38	-.83	.23	-.43	.32	-.80	.34	-.80
.44	-.79	.33	-.45	.32	-.75	.40	-.76
.00	-.94	-.01	-.44	.02	-.91	.00	-.91
.08	-.96	.06	-.49	.06	-.93	.07	-.94
.55	-.28	.36	-.14	.44	-.30	.49	-.27
.32	-.18	.20	-.06	.26	-.21	.28	-.17
-.13	.01	-.11	.11	-.08	-.06	-.12	.01
.09	-.86	.02	-.60	.12	-.70	.08	-.83
.73	-.16	.59	-.21	.47	-.09	.65	-.15
-.31	.28	-.39	.46	-.29	.14	-.36	.31
-.18	.37	-.22	.60	-.18	.21	-.21	.42
-.22	.28	-.32	.46	-.26	.14	-.29	.31
-.22	.33	-.30	.53	-.25	.17	-.28	.36
-.33	.22	-.32	.29	-.16	.06	-.29	.20
.19	.18	.21	.26	.12	.06	.19	.18
-.16	.30	-.05	.44	.07	.11	-.05	.31

table 6.3.a: canonical components for single numerical analysis (correlations of 16 original variables and four sets of two canonical variables).

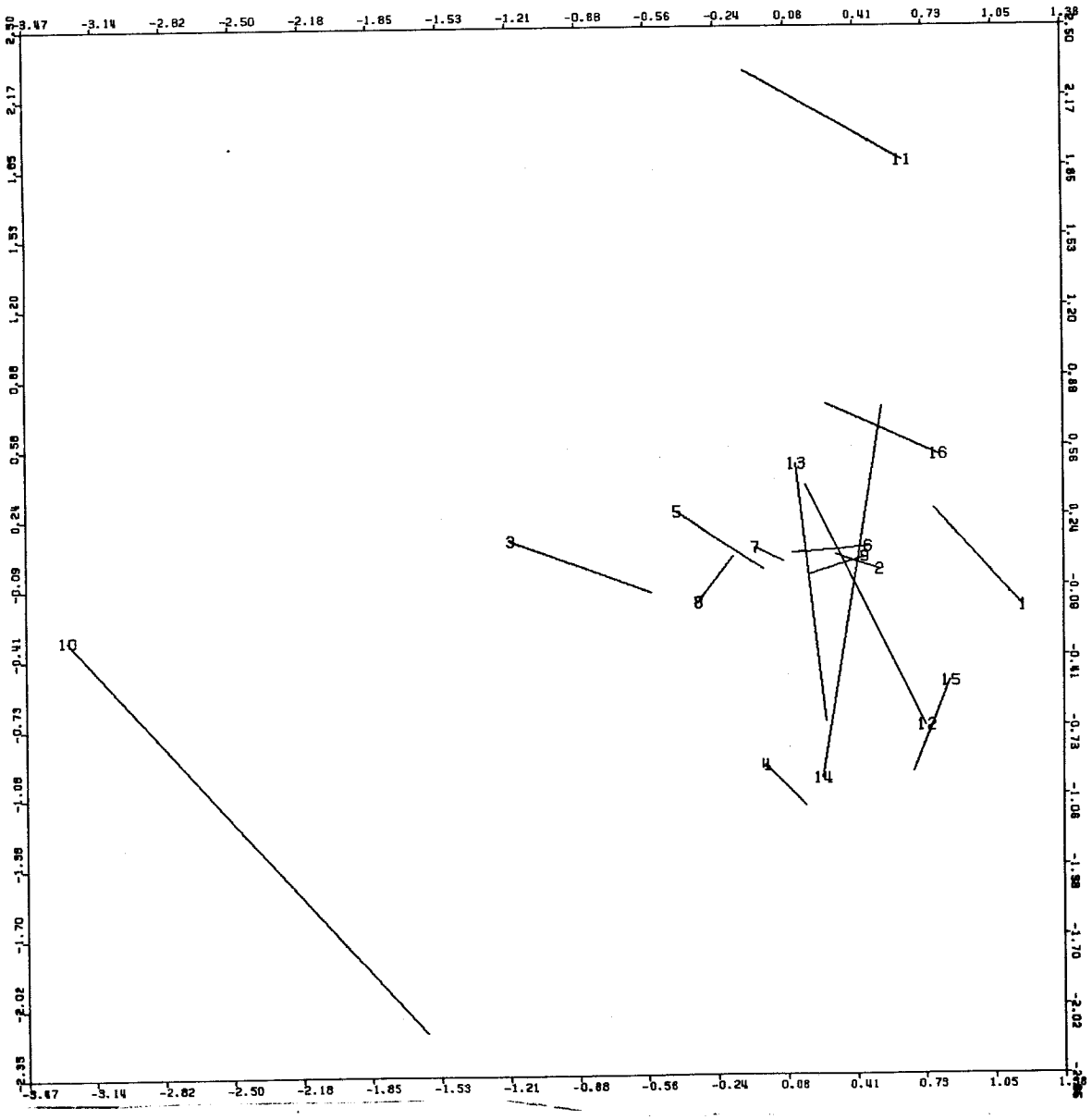


Figure 6.2 Canonical weights for the fish-data in the numerical solution (labeled) connected with the corresponding points in the ordinal solution

.86	-.46	.85	-.48	.87	-.47	.86	-.46
.90	-.20	.89	-.21	.89	-.21	.89	-.20
.22	-.87	.21	-.86	.22	-.88	.22	-.87
.24	-.97	.23	-.97	.24	-.97	.24	-.97
.32	-.14	.34	-.13	.30	-.15	.32	-.14
.34	.08	.33	.11	.34	.04	.34	.08
.05	-.18	.07	-.17	.04	-.19	.05	-.18
.14	-.57	.14	-.58	.15	-.57	.14	-.57
.55	.13	.56	.09	.54	.16	.55	.13
-.39	.01	-.40	.03	-.40	.00	-.40	.01
-.38	.24	-.40	.24	-.40	.22	-.40	.23
-.22	.04	-.23	.05	-.23	.05	-.23	.05
-.25	-.00	-.26	-.01	-.26	-.00	-.26	-.01
.21	.16	.22	.16	.22	.15	.22	.16
.40	-.16	.39	-.17	.37	-.17	.39	-.16
.20	.22	.19	.23	.18	.21	.19	.22

table 6.3.b: canonical components for single ordinal analysis (correlations of 16 quantified variables and four sets of two canonical variables).

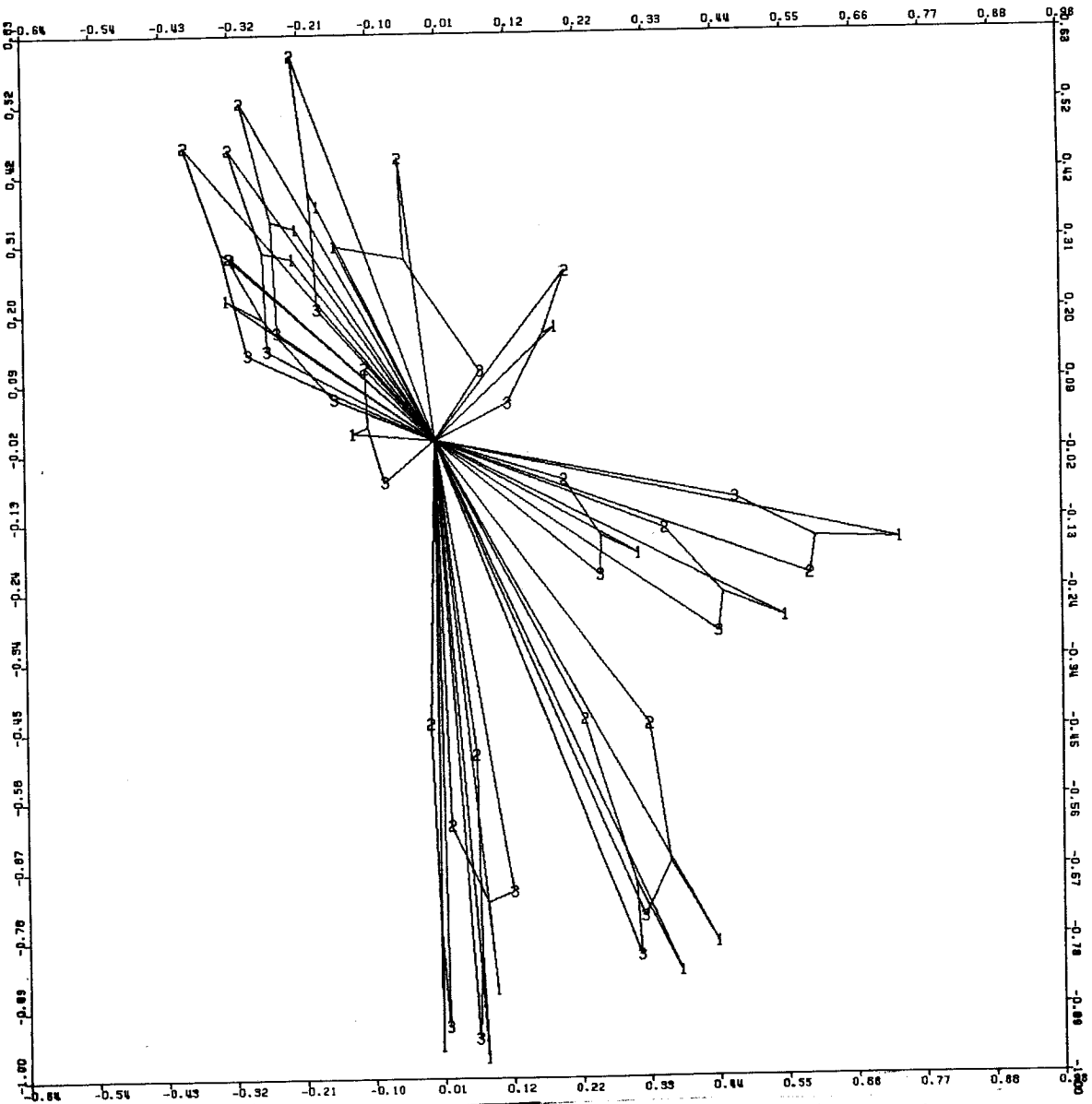


Figure 6.3 Single numerical OVERALS solution for the fish-data: canonical components of the 16 variables on 3 sets of 2 canonical variables each, labeled by

7 Canonical correlation analysis and CANALS

7.1 Previous work

We have seen in chapters 3 and 4 that if there are only two sets of variables, then the meet problem simplifies considerably, at least in the linear case. If we consider the various criteria in chapter 6.1.2, then R(Q) is 2 × 2, and the criteria are all monotonic function of the single correlation between $q_1 = H_1 y_1$ and $q_2 = H_2 y_2$. Thus all techniques amount to choosing y_1 and y_2 in such a way that this correlation is either maximized or minimized.

There is considerably more historical material on the problem of two sets. The person who formulated the problem for the first time was probably Spearman, who discussed the correlation between sums and differences in a famous paper of 1913. He observed that it was interesting to investigate how weights n_j should be computed such that the correlation between the weighted sum $n_1 a_1 + \ldots + n_p a_p$ and another variable b is as large as possible. Spearman observed that Yule had already described completely how this problem should be solved. But he adds: "By the aid of the present theorems, the still more general case of the maximum of $r_{(n_1 a_1 + \ldots + n_p a_p)(m_1 b_1 + \ldots + m_q b_q)}$ can be obtained, though sometimes through rather complicated differentiation." (Spearman, 1913, p 420, footnote). And indeed, because Spearman does not use matrix algebra and the theory of canonical forms of matrices, his expressions already become very complicated in the simple special cases that he treats. Hotelling (1935) was a big improvement in this respect. In this paper an asymmetric approach to the two sets is adopted; one set of variables consists of predictors, the other set of criteria. The problem is that there is more than one criterium, this cannot be avoided in many situations because some theoretical constructs such as college success or general price level cannot be measured by a single variable. Hotelling suggests various possible ways to combine the criteria into a single one, as soon as this is accomplished we apply multiple regression and we are done. One of the methods he proposes is to replace the criteria by their first principal component, which is quite similar to some of the 'first-step' procedures we use in our examples. Also several forms of a priori weighting are considered. "In spite of this variety of grounds for choice of a variate to be predicted, the problem of finding a linear function of the criterion variates which can most accurately be predicted from given observations, in the sense of least squares, admits a definite solution, which we shall set forth." (Hotelling, 1935, p 139). It seems that this particular formulation of the problem (the 'most predictable criterion') has confused a lot of psychometricians. In their search for 'the' criterion they thought (correctly) that there is no reason to prefer a particular criterion simply because it can be predicted most accurately. They overlooked the fact that this is only one way

to formulate canonical analysis, and that even if canonical analysis does not give us 'the' criterion, it does at least give us upper and lower bounds.

In Hotelling (1936) the more symmetrical approach is adopted. "The relations between two sets of variates with which we shall be concerned are those that remain invariant under internal linear transformations of each set separately. Such invariants are not affected by rotations of axes in the study of wind or of hits on a target, or by replacing mental test scores by an equal number of independently weighted sums of them for comparison with physical measurements." (Hotelling, 1936, p 322). This symmetric formulation seems far less objectionable, but it has been igonored by psychometricians for a fairly long time. This is despite the excellent papers by Thomson (1947), Bartlett (1948), and Burt (1948), who explained Hotelling's work to psychometricians familiar with factor analysis. Canonical analysis never became popular in the same way that multiple regression and principal components analysis became popular.

Thorndike and Weiss (1973) mention four possible reasons for this relative unpopularity. The first one is difficulty in computation, this reason was only operative until the early sixties, because during the sixties efficient computer programs for performing canonical analysis became widely available. The second reason is availability of other more familiar methods for studying the relationships between two sets of variables. Again this does not seem to be a valid reason any more. Review papers trying to convert people to using canonical analysis have been written in marketing (Green, Halbert, Robinson, 1966), in counseling psychology (Weiss, 1972), in educational research (Darlington Weinberg, Walberg, 1973). Application-oriented papers have also been published by Wood and Erskine (1976), and Thorndike (1977). We have seen in chapter 1 that canonical analysis is covered in almost all MVA textbooks, as the appropriate technique to investigate relations between two sets of numerical variables. Thus there must be other reasons. The third one mentioned by Thorndike and Weiss is the difficulty in interpreting the results of canonical analysis. The problem here seems to be that investigators often did not know exactly what to compute, what to plot, and where to look at. Again this problem has been studied quite extensively recently, and the outcome seems to be that the canonical weights should be supplemented by the correlations between the original variables and the canonical variables (we have already seen this in the example of chapter 6). The fourth reason mentioned by Thorndike and Weiss is the instability of results (the 'bouncing beta' problem of multiple regression analysis becomes doubly serious). Studies by Thorndike and Weiss (1973), Barcikowski and Stevens (1975), Thorndike (1977) indicate that indeed stability under selection of variables or individuals may be low. The studies use cross validation and Monte Carlo methods, and the results themselves are extremely difficult to

generalize, indicating that the stability of results investigating stability may also be very low. It is possible, however, to derive some analytical results which indicate that canonical analysis is generally less stable than, for example, principal components analysis. Some of these results will be reviewed in chapter 11.

A fifth reason, not mentioned by Thorndike and Weiss, is perhaps that the situation in which we have two sets of numerical variables with a symmetric role does not occur often in practice. Asymmetric versions of canonical correlations have been introduced by Stewart and Love (1968), and an asymmetric version of canonical analysis has been proposed by Van den Wollenberg (1977). We still have the impression, however, that the most interesting applications of canonical analysis are even more thoroughly asymmetric, for example because the second set consists only of a single variable (as in multiple regression), because the second set is a categorical variable (as in discriminant analysis), because the first set is a design matrix (as in analysis of variance), and so on. The fact that canonical analysis is a useful computational technique which unifies a large number of seemingly very different MVA problems was already emphasized by Bartlett (1948), and Tintner (1946). These more asymmetric versions will be discussed in the next chapter, because it is often advantageous to fit them by using specialized algorithms.

7.2 Theory

Of course it is possible to fit canonical analysis with two sets of variables with the more general algorithm for K sets outlined in chapter 6. Strictly spoken we do not need a separate chapter for two sets. We can use the special properties of K = 2, however, to simplify both theory and algorithm somewhat.

We start with a normalization result similar to that in 6.2.2. Consider

$$\sigma_M(X,Y) = \frac{1}{K}\sum_{k=1}^{K} SSQ(X - G^kY^k),$$

which must be minimized under the condition $X'X = I$ (we suppose $u'X = 0$ is automatically taken care of by the definition of the restrictions on Y^k). We assume again that the restrictions on Y^k are such that Y^kT^k is feasible whenever Y^k is feasible; no matter how we choose T^k. We have shown in 6.2.2 that minimizing $\sigma_M(X,Y)$ over $X'X = I$, means maximizing the sum of the p largest eigenvalues of

$$\sum_{k=1}^{K} Q_k(Q_k'Q_k)^+Q_k',$$

with $Q_k = G^kY^k$. And we have shown that minimizing $\sigma_M(X,Y)$ under the condition

$$\sum_{k=1}^{K} Q_k'Q_k = I,$$

is equivalent to maximizing the sum of the p largest generalized eigenvalues of the problem with matrices C and D, with submatrices $C^{k\ell} = Q_k'Q_\ell$ and $D^k = Q_k'Q_k$. Because the generalized eigenvalues of the second problems and the eigenvalues of the first problem are the same, the two problems are equivalent, and we can use either one of the normalization conditions (either on X or on Y). The condition on Y, in the case K = 2, however, shows that our problem is equivalent to minimizing

$$\sigma_M(Y^1,Y^2) = SSQ(G^1Y^1 - G^2Y^2),$$

from which X has been eliminated, and in which we require that $Q_1'Q_1 + Q_2'Q_2 = I$. Because we can choose T^1 and T^2 independently to transform Y^1 into Y^1T^1 and Y^2 into Y^2T^2, we get an equivalent solution if we require $Q_1'Q_1 = Q_2'Q_2 = I$. And finally we can also require that either $Q_1'Q_1 = I$ or $Q_2'Q_2 = I$, which again gives the same solution (except, again, possibly for a different scaling of the dimensions). It is clear that these normalization results generalize corresponding results from chapter 3 and 4 to the case in which there are restrictions on the Y_j. The important condition which makes these generalizations possible is that Y^kT^k is feasible, whenever Y^k is. This condition is satisfied by sets of single and multiple nominal variables.

The simplified algorithm for two-set canonical analysis with optimal scaling is now applied to $\sigma_M(Y^1,Y^2)$. This was done for the first time in Young, De Leeuw, Takane (1976), but the algorithm proposed in that paper does not work. The problem is in the normalization. Young, De Leeuw, and Takane require that both $Q_1'Q_1 = I$ and that $Q_2'Q_2 = I$ (actually, their CORALS algorithm is limited to compute one-dimensional solutions, but this is not essential). But this has as consequence that the constraints on Y become very complicated. If we modify one of the Y_j (by an algorithm as in 6.2.5) then we must impose both the measurement constraints on Y_j and the set-normalization constraints $(Y^kG^k)'G^kY^k = I$ which also involves the other Y_ℓ in the set. Young, De Leeuw, Takane ignore the normalization constraint while modifying the Y_j, and impose it afterwards again. This is against the general principles of alternating least squares, and it consequently does not work. CORALS diverges.

Fortunately it is fairly easy to repair the damage. We do not require that $Q_1'Q_1 =$ and $Q_2'Q_2 = I$, but that $Q_1'Q_1 = I$ or $Q_2'Q_2 = I$. Which one of these two constraints we impose at any particular time depends on which set we are modifying at the moment. If we modify the Y_j of the first set, we only impose $Q_2'Q_2 = I$. If all Y_j in the first set are modified we find a linear transformation T^1 and another linear

transformation T_2^2 in such a way that $\sigma_M(\underline{Y}^1,\underline{Y}^2) = \sigma_M(Y^1,Y^2)$ and that $Q_1'Q_1 = I$, where $\underline{Y}_j = Y_jT^1$ if Y_j is in the first set and $\underline{Y}_j = Y_jT^2$ if Y_j is in the second set. Thus we rescale both sets in such a way that the loss does not change, and that after rescaling the first set is normalized. We then proceed to modify the Y_j in the second set, without bothering about normalization. And after all these Y_j are modified we rescale again. And so on. It is clear that this procedure is convergent, because it never increases the loss. It critically depends on K = 2, because it uses the fact that we can switch between two problems which we know from theory to be equivalent.

7.3 The CANALS program

The CANALS program does not fit naturally in the series HOMALS - PRINCALS - OVERALS, because it does not use indicator matrices and cannot incorporate multiple variables. This is basically because it dates back to a period in the project in which we still wanted the possibility to incorporate continuous variables into the programs. Practical experience and some theory such as 5.2.4 have convinced us that incorporating continuous variables is not urgent, to put it mildly. CANALS is much more like the ALS-program written in North Carolina, whose descriptions can be found by looking under the references by Takane, Young, and De Leeuw (in some order). In fact it is a fairly straightforward extension of CORALS to $p \geq 1$ dimensions, in which the 'clever rescaling' mentioned in 7.2 is used to repair the error made in CORALS.

There is no need to go into technical detail about the program. The only new aspect is the rescaling. This can be explained briefly as follows. We have Q_1 and Q_2 such that $Q_2'Q_2 = I$. We want to find T^1 and T^2 such that

$$SSQ(Q_1T^1 - Q_2T^2) = SSQ(Q_1 - Q_2),$$

and that $(Q_1T^1)'Q_1T^1 = I$. The solution is simple: we find any T^2 such that $Q_1'Q_1 = T^2(T^2)'$ and we set $(T^1)' = (T^2)^{-1}$. CANALS uses Gram-Schmidt to find T^2 from Q_1 but singular value decomposition could also be used.

CANALS output will be illustrated in the examples, but we mention some of the more important peculiarities. In CANALS all variables are single, which means that we have a vector y_j of category quantifications (scaled in such a way that $u'D_jy_j = 0$ and $y_j'D_jy_j = n$) for each j. Moreover for each j we have a vector a_j of (canonical) weights. And finally we have correlations between the canonical variates $Q_1 = G^1Y^1$ and $Q_2 = G^2Y^2$ and the individual quantified variables G_jy_j. These correlations are called canonical components by Thorndike and Weiss (1973). We must distinguish R_{11}, R_{12}, R_{21}, R_{22}, which are, respectively, the correlation between the variables in the first set and the canonical variables of the first set, between the variables in the first set and the canonical variables

of the second set, between the variables of the second set and the canonical variables of the first set, and between the variables of the second set and the canonical variables of the second set. If the canonical correlations are collected in the diagonal matrix Φ, then $R_{12} = R_{11}\Phi$, and $R_{21} = R_{22}\Phi$. Using a similar notation for the correlation between the canonical variables, we also have $R_{11} = R_{22} = I$ and $R_{12} = \Phi$.

We also emphasize that CANALS does not have provisions to handle missing data according to option I, missing data are handled according to option III. Because CANALS does not use indicator matrices this simply means that the missing data approach must be incorporated in the definitions of the cones K_j.
Option II can be simulated if one codes his missing data in one extra category per variable. It would be sensible to use the nominal option in this case.

7.4 Examples

7.4.1 Prediction of a school achievement test

This first example does not illustrate the CANALS-program, but it illustrates generalization of two-sets canonical analysis and relates them to some other multivariate analysis techniques. The data are taken from CBS (1980). It is a three-way contingency table in which a sample of approximately 120000 children is classified according to sex (2 levels), fathers profession (7 levels), and school achievement test scores (5 levels). The data are in table 7.1.a. In this example we want to predict test scores from the other two variables. Thus the first set consists of the two variables S (sex) and P (profession), the second set consists of the single variable T (test).

A very common technique in situations like this is log-linear analysis. We have used the ordinary iterative proportional fitting technique to fit four different models, coded by the marginals that are fitted exactly. The results are in table 7.1.b, we give the likelihood ratio goodness-of-fit and the corresponding number of degrees of freedom. The interpretation of the models is facilitated by writing

$$\ln p_{ijk} = \alpha_{ik} + \beta_{jk} + \gamma_{ijk},$$

with p_{ijk} the proportion of individuals with father i, sex j, and test score k. The α_{ik} are the parameters for profession, the β_{jk} for sex, and the γ_{ijk} are the interactions. Thus SP = 0 (no interaction) means $\gamma_{ijk} = 0$ for all i,j,k, and SP = P = 0 means $\gamma_{ijk} = 0$ and $\alpha_{ik} = 0$ for all i,j,k. In 7.1.c we show that we can find estimates for the effect of sex and profession, and for the interactio by subtraction in various ways. The conclusion is clear: P is much more important than S, interaction SP is not very important. Because the sample is extremely large everything is significant, from a statistical point of view no simplificatio of the saturated model is possible. Inferential statistics is not very useful here

because there simply is no law generating the data, and any simplifying model simply is not true. This situation is very common in the social sciences, the common procedure in situations like this is to use large sample statistical techniques on relatively small samples, or making very strong assumptions and by testing only within these untested and often untestable assumptions. Most of the arguments in favor of high heritability of intelligence, for example, are based on sophisms like this.

From a data analysis point of view estimates of the parameters are more interesting than global significance tests. We first study additive partitioning of chi-squared, which can be based on the saturated model

$$p_{ijk} = q_{ij} r_k \left(\sum_{s=0}^{6} \sum_{t=0}^{1} \sum_{u=0}^{4} \rho_{stu} x_{is} y_{jt} z_{ku} \right),$$

where q_{ij} and r_k are the marginal proportions, and where

$$\sum_{k=1}^{5} r_k z_{ku} z_{ku'} = \delta^{uu'},$$

$$\sum_{i=1}^{7} \sum_{j=1}^{2} q_{ij} x_{is} y_{jt} x_{is'} y_{jt'} = \delta^{ss'} \delta^{tt'}.$$

Moreover $z_{k0} = x_{i0} = y_{j0} = 1$ for all i,j,k. Otherwise the x_{is}, y_{jt}, and z_{ku} are arbitrary. The saturated model can be fitted by

$$\rho_{stu} = \sum_{i=1}^{7} \sum_{j=1}^{2} \sum_{k=1}^{5} p_{ijk} x_{is} y_{jt} z_{ku}.$$

Although the choice of orthogonal functions is irrelevant from the point of view of the saturated model, it can be quite important from the data analysis point of view. In this case we have chosen orthogonal polynomials for both x_{is}, y_{jt}, and z_{ku}. Because always $\rho_{000} = 1$, $\rho_{00k} = 0$ for k=1,...,4, and $\rho_{ij0} = 0$ for $(i,j) \neq (0,0)$ there are 52 nontrivial ρ_{stu}. In 7.1.d we give all $n^{\frac{1}{2}}\rho_{stu}$. The first row is ρ_{01k}, with $k \geq 1$, it measures the effect of S. The next six rows are ρ_{i0k}, with $i \geq 1$ and $k \geq 1$, they measure the effect of P, the final six rows are ρ_{i1k}, with $i \geq 1$ and $k \geq 1$, which measures the effect of the interaction SP. Table 7.1.e has the sum of squares of the rows and columns of table 7.1.d. If in the population $\rho_{stu} = 0$ for all s,t,u (except for (s,t,u) = (0,0,0)), then the elements of table 7.1.d are independent standard normal asymptotically, and consequently the sums of squares of the rows are independent chi squared variables with four degrees of freedom, while the sums of squares of the columns are independent chi squares with 13 degrees of freedom. Again P is much more important than S, again the interaction is small. We can also see from this analysis, however, that the linear component explains most of the variance (both for P and for T approximately 93%). If the ρ_{stu} are nonzero, the sampling theory

of the statistics we have computed has been outlined recently by O'Neill (1980

The two more or less classical approaches of discrete MVA (the first one based the 'multiplicative', the second one on the 'additive' definition of interacti will now be compared with canonical analysis. In the first form of canonical a we make S and P into a composite ('interactive') variable with 14 categories, we apply multiple nominal CANALS to the resulting two sets of one single varia each. As we know in this case CANALS, HOMALS, and ANACOR are the same, and the easiest way to compute a solution is to apply singular value decomposition to the 14 × 5 table constructed from 7.1.a. This form of canonical analysis corre with constructing a general nonlinear function on S × P and a general nonlinea function on T with maximal intercorrelation. CANALS with multiple nominal variables tries to find nonlinear functions on S, P, and T separately in such a way that f(S) + g(P) correlates maximally with h(T). Again we can compu the solution in the multiple nominal case by solving a generalized eigenvalue problem. Compute the (7 + 2 + 5) × (7 + 2 + 5) table C of bivariate marginals and the matrix D which consists of the (7 + 2) × (7 + 2) and 5 × 5 principal submatrices of C, and solve the generalzied eigenproblem corresponding with C, D. The third form of canonical analysis tries to predict T from P alone, the fourth form from S alone. These last two can be computed by correspondenc analysis of the 7 × 5 and 2 × 5 tables computed from 7.1.a by summing, respe over S and P. The results of the four canonical analyses are in table 7.1.f. The first four rows are the squared canonical correlations, the fifth row is their sum, the next seven rows are the optimal scalings of P for boys, and the last seven rows are the optimal scalings of P for girls. The first three optimal scalings are also plotted in figure 7.1. In 7.1.a we see that P is much more important than S, and that there is not much interaction. In 7.1 we require that there is no interaction, and we do not loose much predictive p In figure 7.1.c we require that there is no interaction, and no effect of S, and again we do not loose much. The results of the three types of analyses we have performed so far are summarized in table 7.1.g. It is clear that the conclusions that can be drawn on the basis of the analyses are very similar, because the results are very similar. Statistical inference procedures for th canonical correlations and optimal scores have been derived by Lebart (1976), O'Neill (1978a, b).

7.4.2 Economical inequality and political stability

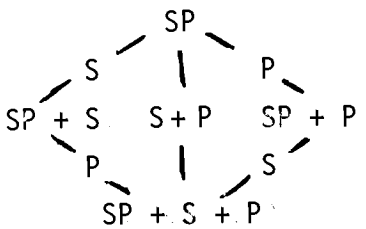
The data for this example are taken from a paper by Russett (1964), which ha reprinted in Rowney (1969). The basic hypothesis in Russett's paper is that economic inequality leads to political instability, the basic problem is how to measure these complicated constructs. "We shall be concerned with informa on the degree to which agricultural land is concentrated in the hands of a f

S1	P1	4669	6653	5543	3380	787
S1	P2	5757	12052	9526	6003	949
S1	P3	974	2188	2714	1865	276
S1	P4	468	1395	1774	1492	334
S1	P5	883	2667	3807	2971	809
S1	P6	916	3238	5963	5458	1798
S1	P7	288	1607	2906	3594	1314
S2	P1	4240	6610	5541	3368	672
S2	P2	4641	9797	9403	5448	998
S2	P3	743	2263	2459	1681	301
S2	P4	379	1185	1670	1228	301
S2	P5	766	2669	3746	2897	650
S2	P6	858	3195	5322	5028	1736
S2	P7	214	1238	3141	3571	1286

table 7.1.a: test scores as a function of Sex and fathers Profession.

Interpretation	fitted marginals	chi-squared	dfr
SP = 0	(12)(13)(23)	202.2929	24
P = SP = 0	(12)(13)	20198.82	48
S = SP = 0	(12)(23)	295.3212	28
P = S = SP = 0	(12)(3)	20294.08	52

table 7.1.b: log-linear models



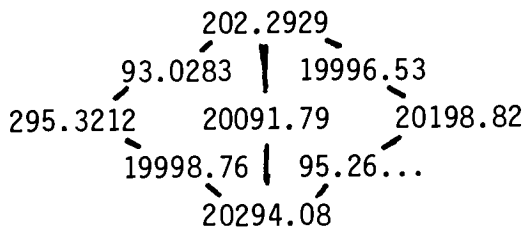


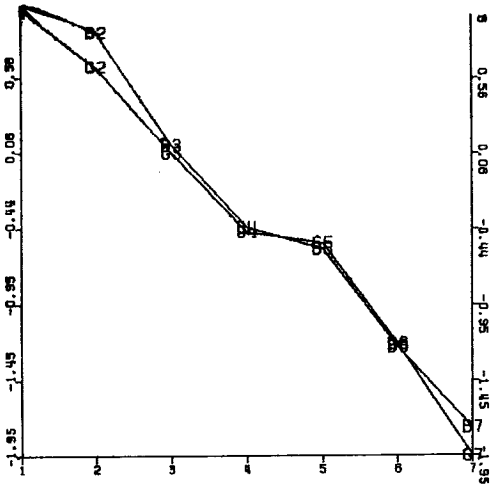
table 7.1.c: main effects and interactions by subtraction

-7.05	5.63	-0.77	-3.62	95.20
135.14	-4.72	0.30	6.51	18327.39
8.90	30.49	-3.58	1.50	1023.91
1.75	-5.69	4.46	-7.26	108.01
8.23	-1.13	-5.80	0.17	102.60
-9.00	-0.99	2.78	-2.31	95.04
-4.73	-3.81	-1.33	0.39	38.87
2.60	-2.67	0.63	0.81	14.98
-0.89	0.19	2.41	-0.33	6.73
-5.98	0.12	2.02	-6.72	85.09
1.54	2.86	-0.64	0.98	11.96
-1.47	2.72	5.35	-3.53	50.77
0.74	1.25	-1.12	3.94	18.92

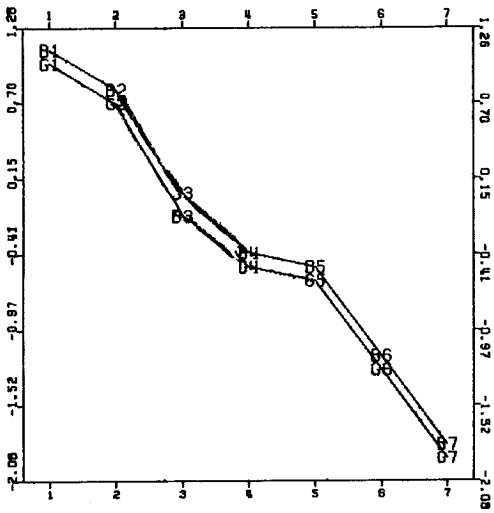
table 7.1.d: polynomial contrasts.

18614.44
1057.06
117.06
190.90

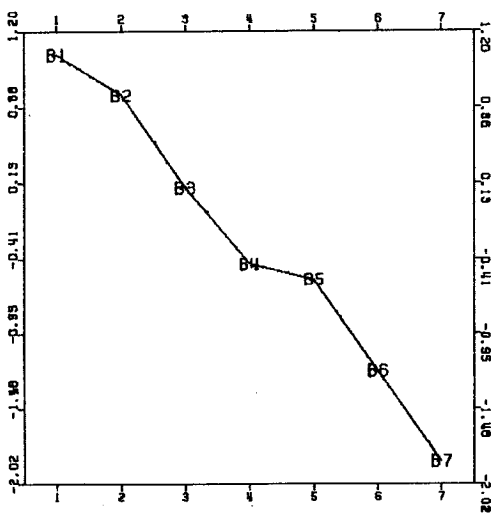
table 7.1.e: polynomial effects.
a: polynomial components S × P.
b: polynomial components A.



a S + P + SP → A



b S + P → A



c P → A

Figure 7.1 Quantifications

18675.73	18621.37	18578.18	95.20
1067.89	1045.70	1013.37	--.--
187.51	92.76	86.57	--.--
48.33	31.18	24.36	--.--
19979.47	19791.02	19702.48	95.20
1.06	1.09	1.04	-0.96
0.86	0.80	0.75	-0.96
0.11	0.13	0.09	-0.96
-0.43	-0.40	-0.45	-0.96
-0.57	-0.50	-0.56	-0.96
-1.22	-1.16	-1.21	-0.96
-1.76	-1.81	-1.86	-0.96
1.01	0.99	1.04	1.04
0.63	0.70	0.75	1.04
0.06	0.04	0.09	1.04
-0.46	-0.50	-0.45	1.04
-0.53	-0.60	-0.56	1.04
-1.20	-1.26	-1.21	1.04
-1.95	-1.91	-1.86	1.04

table 7.1.f: four canonical analyses: components, sum, and transformations.
a: on S + P + SP
b: on S + P
c: on P
d: on S

Source	Log-linear	Additive	Canonical	Dfr
S	92.97	95.20	95.20	4
P	19996.53	19695.82	19702.48	24
SP	204.58	188.45	181.79	24
Total	20294.08	19979.47	19979.47	52

table 7.1.g: summary of three analyses.

large landholders. Information on land tenure is more readily available, and is of more dependable comparability, than are data on the distribution of other economic assets like current income or total wealth." (Russett, 1964, p 444). Three variables are used to measure inequality of land distribution. We discuss them, and mention their codes in the tables and figures. GINI is the Gini index of concentration, which measures the deviation of the Lorenz curve from the line of equality. FARM is the percentage of farmers that own half of the land, starting with the smallest ones. Thus if FARM is 90%, then 10% of the farmers own half of the land. The third indicator is RENT, which is the percentage of farm households that rent all their land. Russett adds two extra economical variables, the gross national product per capita GNPR, and the percentage of the labor force employed in agriculture LABO. There are four measures of political stability. The first one is the term in office of the chief executive, on the average, in the period between 1945 and 1961.

	GINI	FARM	RENT	GNPR	LA	INST	ECK	DEAT	DEM
ARGENTINA	86.3	98.2	32.9	374	25	13.6	57	217	2
AUSTRALIA	92.9	99.6	----	1215	14	11.3	0	0	1
BELGIUM	58.7	85.8	62.3	1015	10	15.5	8	1	1
BOLIVIA	93.8	97.7	20.0	66	72	15.3	53	663	3
BRAZIL	83.7	98.5	9.1	262	61	15.5	49	1	2
CANADA	49.7	82.9	7.2	1667	12	11.3	22	0	1
CHILE	93.8	99.7	13.4	180	30	14.2	21	2	2
COLOMBIA	84.9	98.1	12.1	330	55	14.6	47	316	2
COSTA RICA	88.1	99.1	5.4	307	55	14.6	19	24	2
CUBA	79.2	97.8	53.8	361	42	13.6	100	2900	3
DENMARK	45.8	79.3	3.5	913	23	14.6	0	0	1
DOMINICAN REP.	79.5	98.5	20.8	205	56	11.3	6	31	3
ECUADOR	86.4	99.3	14.6	204	53	15.1	41	18	3
EGYPT	74.0	98.1	11.6	133	64	15.8	45	2	3
EL SALVADOR	82.8	98.8	15.1	244	63	15.1	9	2	3
PHILIPPINES	56.4	88.2	37.3	201	59	14.0	15	292	3
FINLAND	59.9	86.3	2.4	941	46	15.6	4	0	2
FRANCE	58.3	86.1	26.0	1046	26	16.3	46	1	2
GUATEMALA	86.0	99.7	17.0	179	68	14.9	45	57	3
GREECE	74.7	99.4	17.7	239	48	15.8	9	2	2
UNITED KINGDOM	71.0	93.4	44.5	998	5	13.6	12	0	1
HONDURAS	75.7	97.4	16.7	137	66	13.6	45	111	3
IRELAND	59.8	85.9	2.5	509	40	14.2	9	0	1
INDIA	52.2	86.9	53.0	72	71	3.0	83	14	1
IRAQ	88.1	99.3	75.0	195	81	16.2	24	344	3
ITALY	80.3	98.0	23.8	442	29	15.5	51	1	2
JAPAN	47.0	81.5	2.9	240	40	15.7	22	1	2
YUGOSLAVIA	43.7	79.8	0.0	297	67	0.0	9	0	3
LUXEMBOURG	63.8	87.7	18.8	1194	23	12.8	0	0	1
LIBYA	70.0	93.0	8.5	90	75	14.8	8	0	3
NETHERLANDS	60.5	86.2	53.3	708	11	13.6	2	0	1
NICARAGUA	75.7	96.4	----	254	68	12.8	16	16	3
NEW ZEALAND	77.3	95.5	22.3	1259	16	12.8	0	0	1
NORWAY	66.9	87.5	7.5	969	26	12.8	1	0	1
AUSTRIA (AST)	74.0	97.4	10.7	532	32	12.8	4	0	2
PANAMA	73.7	95.0	12.3	350	54	15.6	29	25	3
PERU	87.5	96.9	----	140	60	14.6	23	26	3
POLAND	45.0	77.7	0.0	468	57	8.5	19	5	3
SPAIN	78.0	99.5	43.7	254	50	0.0	22	1	3
TAIWAN	65.2	94.1	40.0	102	50	0.0	3	0	3
URUQUAY	81.7	96.6	34.7	569	37	14.6	1	1	1
VENEZUELA	90.9	99.3	20.6	762	42	14.9	36	111	3
UNITED STATES	70.5	95.4	20.4	2343	10	12.8	22	0	1
WEST GERMANY	67.4	93.0	5.7	762	14	3.0	4	0	2
SOUTH VIETNAM	67.1	94.6	20.0	133	65	10.0	50	1000	3
SWEDEN	57.7	87.2	18.9	1165	13	8.5	0	0	1
SWITZERLAND	49.8	81.5	18.9	1229	10	8.5	0	0	1

Table 7.2.a Original data

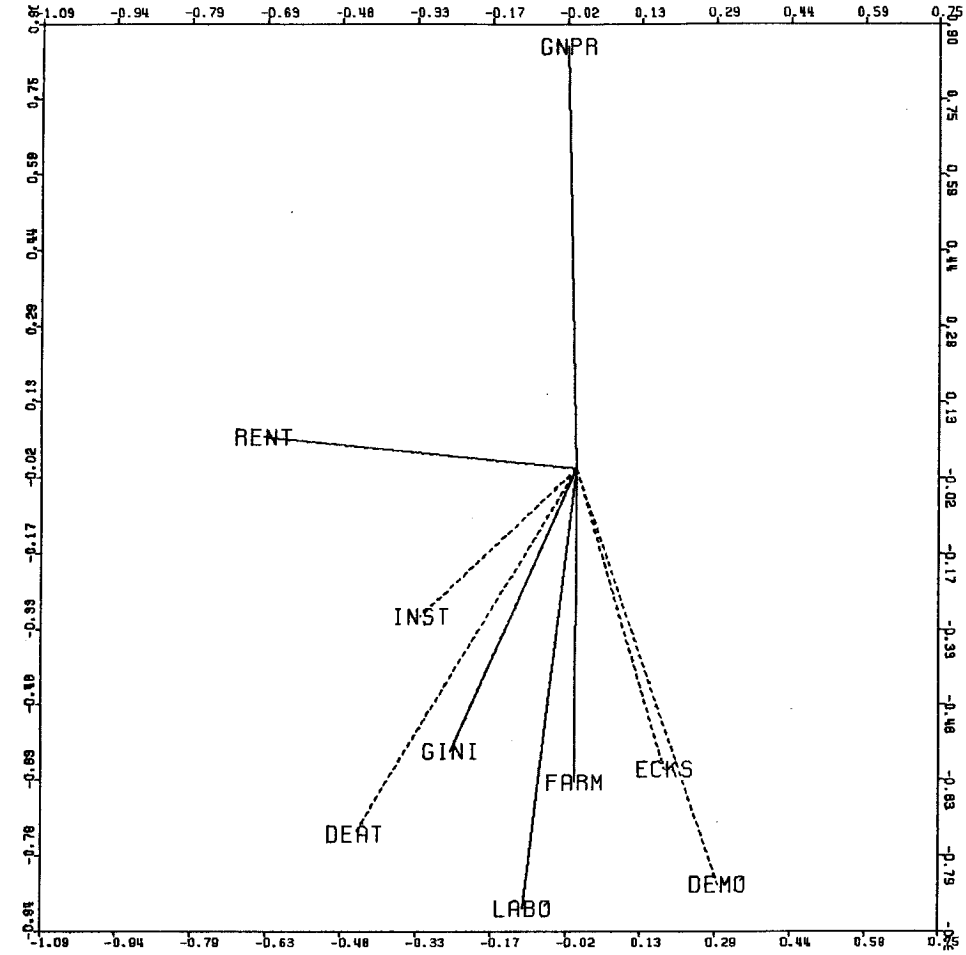
	ECON.VAR					POL.VAR			
ARGENTINA	5	5	4	4	2	5	3	4	2
AUSTRALIA	6	5	9	7	1	4	1	1	1
BELGIUM	2	3	5	6	1	6	2	2	1
BOLIVIA	6	5	3	1	4	6	3	5	3
BRAZIL	5	5	2	4	4	6	3	2	2
CANADA	1	2	2	7	1	4	2	1	1
CHILE	6	5	3	3	2	5	2	2	2
COLOMBIA	5	5	3	4	3	5	3	4	2
COSTA RICA	5	5	2	4	3	5	2	3	2
CUBA	4	5	5	4	3	5	3	5	3
DENMARK	1	1	2	6	2	5	1	1	1
DOMINICAN REP.	4	5	3	3	3	4	2	3	3
ECUADOR	5	5	3	3	3	6	3	2	3
EGYPT	4	5	3	2	4	6	3	2	3
EL SALVADOR	5	5	3	3	4	6	2	2	3
PHILIPPINES	2	3	4	3	5	5	2	4	1
FINLAND	2	3	2	6	3	6	2	1	2
FRANCE	2	3	4	6	2	6	3	2	2
GUATEMALA	5	5	3	3	4	5	3	3	3
GREECE	4	5	3	3	3	6	2	2	2
UNITED KINGDOM	4	4	5	6	1	5	2	1	1
HONDURAS	4	5	3	2	4	5	3	4	3
IRELAND	2	3	2	5	2	5	2	1	1
INDIA	2	3	5	1	4	2	3	2	1
IRAQ	5	5	5	3	5	6	2	4	3
ITALY	5	5	4	5	2	6	3	2	2
JAPAN	1	2	2	3	2	6	2	2	2
YUGOSLAVIA	1	1	1	4	4	1	2	1	3
LUXEMBOURG	3	3	3	7	2	5	1	1	1
LIBYA	3	4	2	1	4	5	2	1	3
NETHERLANDS	3	3	5	6	1	5	2	1	1
NICARAGUA	4	5	9	4	4	5	2	2	3
NEW ZEALAND	4	5	4	7	1	5	1	1	1
NORWAY	3	3	2	6	2	5	2	1	1
AUSTRIA (AST)	4	5	2	5	2	5	2	1	2
PANAMA	4	4	3	4	3	6	2	3	3
PERU	5	5	9	2	3	5	2	3	3
POLAND	1	1	1	5	3	3	2	2	3
SPAIN	4	5	5	4	3	1	2	2	3
TAIWAN	3	4	4	2	3	1	2	1	3
URUQUAY	5	5	4	5	2	5	2	2	1
VENEZUELA	6	5	3	6	3	5	2	4	3
UNITED STATES	4	5	3	8	1	5	2	1	1
WEST GERMANY	3	4	2	6	1	2	2	1	2
SOUTH VIETNAM	3	4	3	2	4	4	3	5	3
SWEDEN	2	3	3	7	1	3	1	1	1
SWITZERLAND	1	2	3	7	1	3	1	1	1

Table 7.2.b Discreticized data

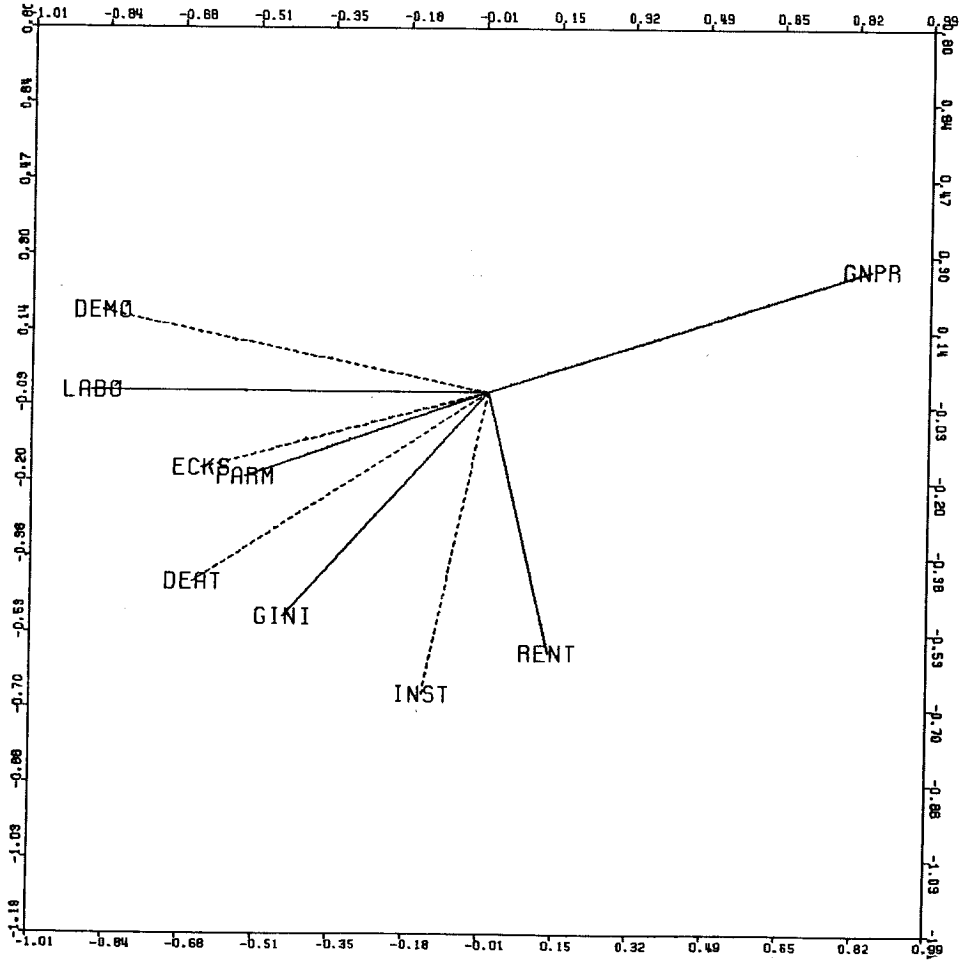
This index is called INST, it runs from zero (very stable) to seventeen (very unstable). The second index is ECKS, the total number of violent internal war incidents in 1946-1961. The third one is DEAT, the number of people killed as a result of internal group violence. The fourth one is DEMO, which is a three fold classification of countries in 'stable democracies', 'unstable democracies', and 'dictatorships'. Russett gives the values of all nine variables for 47 countries. The data are in table 7.2.a.

Russett's analysis is somewhat primitive, he simply correlates everything with everything and reports some of the higher correlations (tabellary analysis for interval scale variables, or for supposedly interval scale variables). The two economic variables GNPR and LABO were introduced to make a more refined analysis possible. Russett's discussion seems to suggest that he has a partial correlation analysis in mind, but he says that he has used multiple regression, and does not give any details. We have chosen to use canonical correlation analysis for the two sets of variables GINI, FARM, RENT, GNPR, LABO versus INST, ECKS, DEAT, DEMO. This is not necessarily the most rational choice. Both Russett's discussion and the definition of the variables suggest that it may be preferable to use three set canonical analysis (program OVERALS, explained in chapter 6) or a partial canonical correlation analysis (program PARTALS, explained in chapter 8). Because both programs are not available yet in exportable FORTRAN versions we prefer to use this example to demonstrate some of the possibilities of CANALS. All variables are treated as single ordinal, we use p = 2 dimensions. There are two different analyses, the first one is based on the rankings of table 7.2.a, the second one uses a discreticized version of this table, given in table 7.2.b, the two solutions can be compared to indicate how stable they are. Of course DEMO is already discrete, which prevents Russett from computing correlations, but which does not bother CANALS.

Figure 7.2.a and 7.2.b give the canonical components in the space of the economic variables, i.e. the correlations between the nine quantified variables and the optimal linear combinations of the first five quantified variables which define the two canonical variables of the first set. The solutions for the rankings and the discreticized data are not very different, except for the fact that the two dimensions are interchanged. For the rankings the canonical correlations are .99 and .91, for the discrete data they are .91 and .82. Figures 7.2.a and 7.2.b show that the first dimension of the discrete data, which is the second dimension of the rankings, contrasts poor dictatorships with a large percentage of agricultural labour (variables DEMO, LABO, and GNPR) with rich industrialized democracies. The other dimension is more difficult to interpret. It helps to study figures 7.3.a and 7.3.b, which are the object scores on the two canonical variables of the first set. For the dicrete data matrix we see a number of clusters. On the

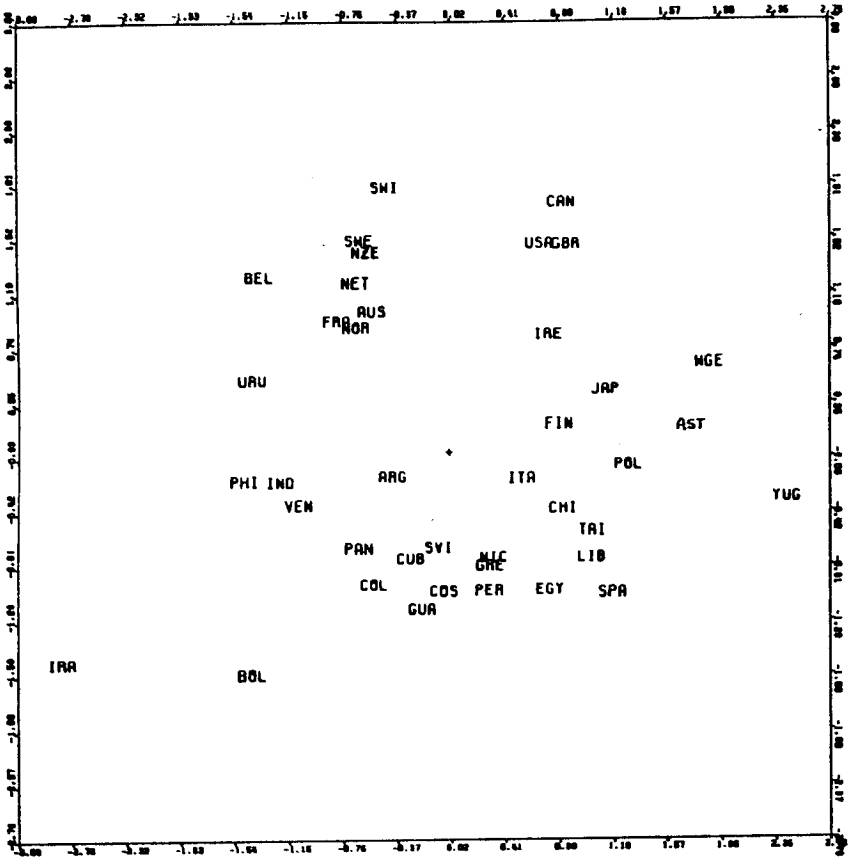


a Rankings



b Discrete

Figure 7.2 Correlations in the canonical space of economic variables



Clusters:

SWE-LUX

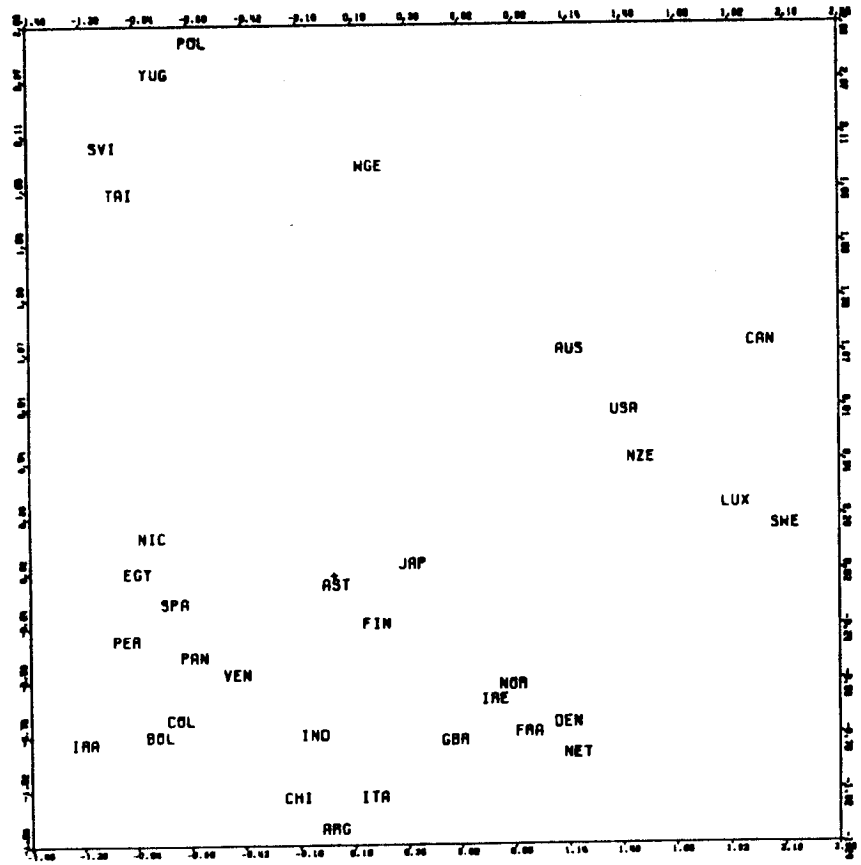
NZE-DEN

COL-ECU

COS-DOR-BRA

PER-HON-ELS

a Rankings



Clusters:

BOL-ECU-ELS-GUA

COL-COS-BRA

IND-PHI

NET-BEL

CAN-SWI

SVI-LIB

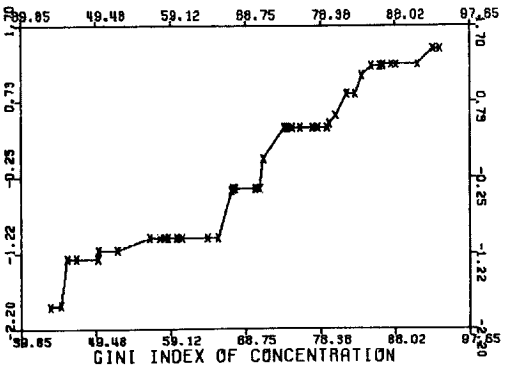
EGY-HON-DOR-GRE

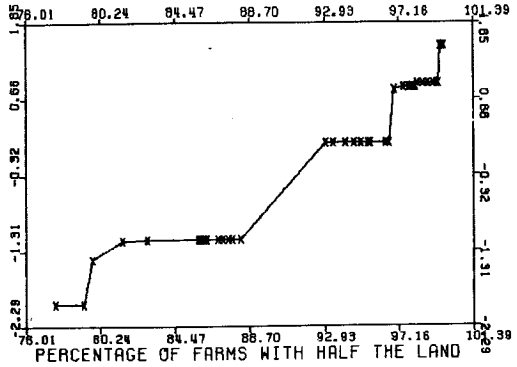
SPA-CUB

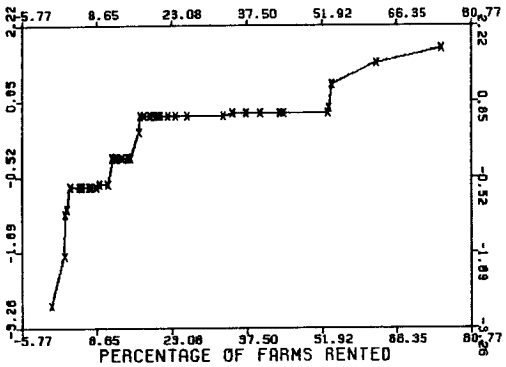
ITA-URU

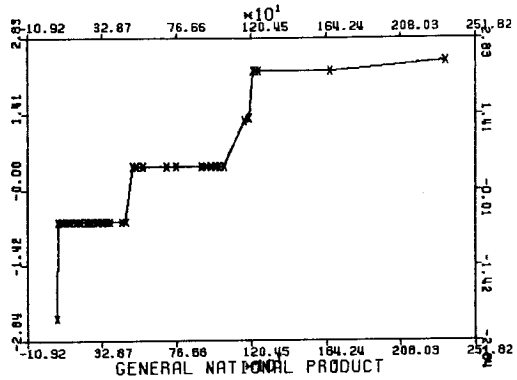
b Discrete

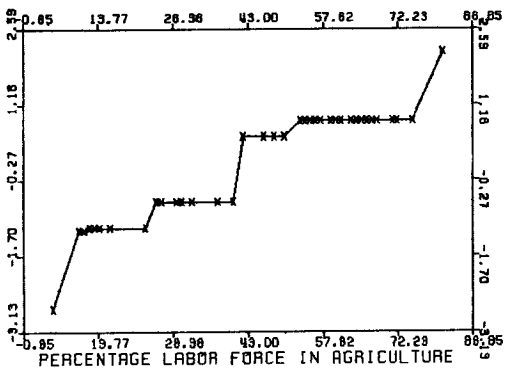
Figure 7.3 Canonical scores in the space of economic variables

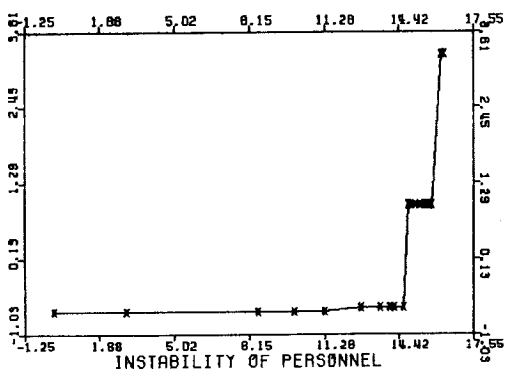


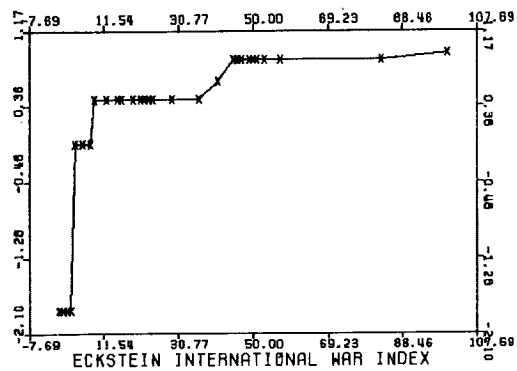


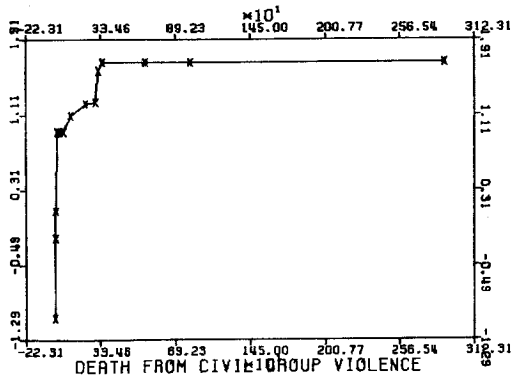












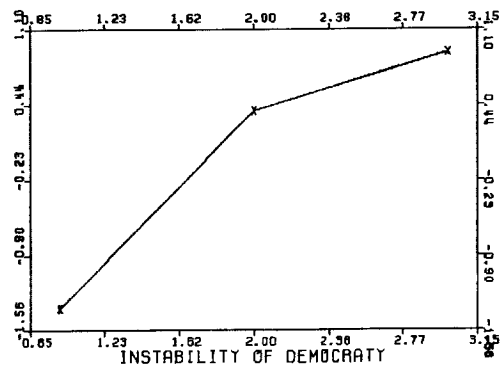
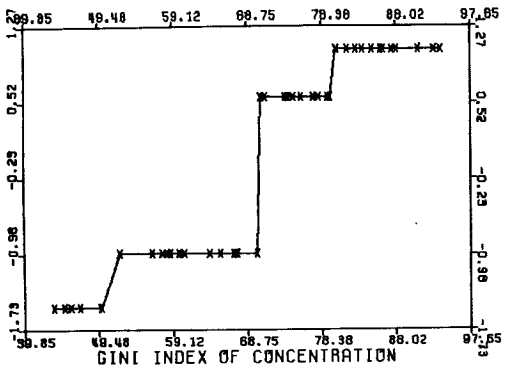
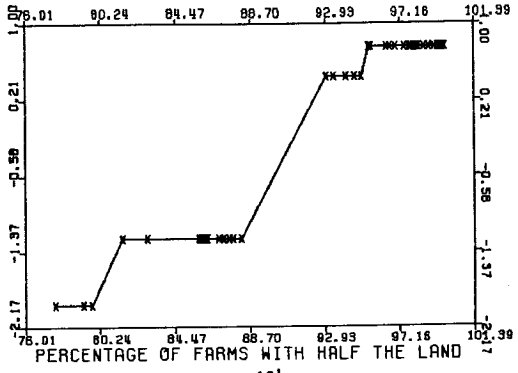
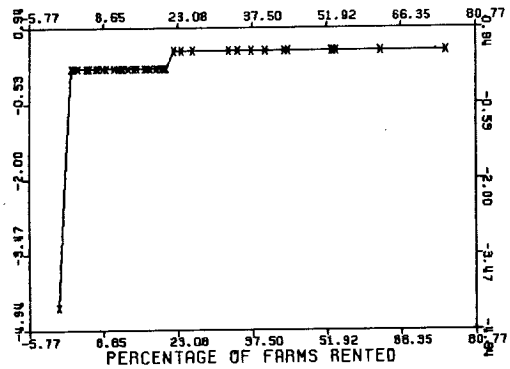
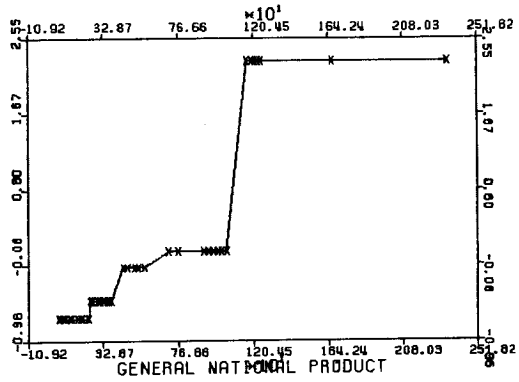


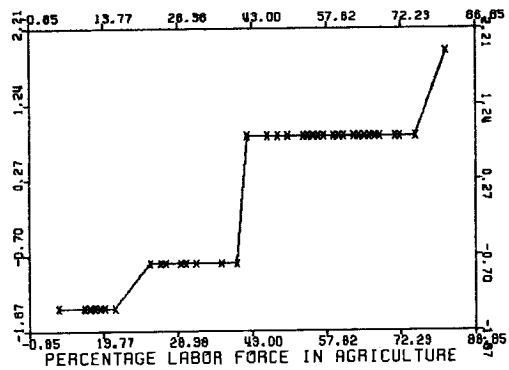
Figure 7.4 a Transformations of the rankings

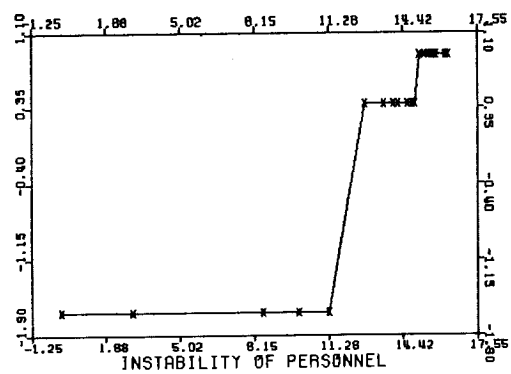


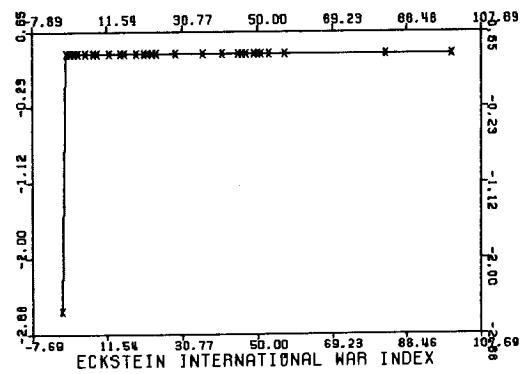


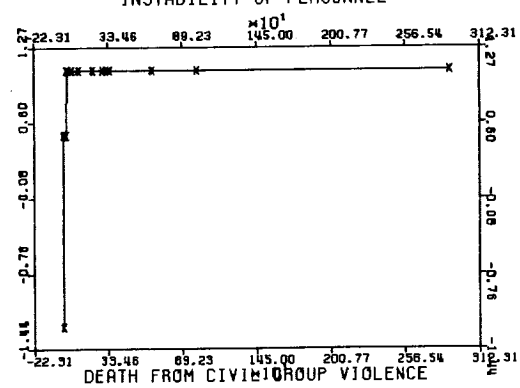












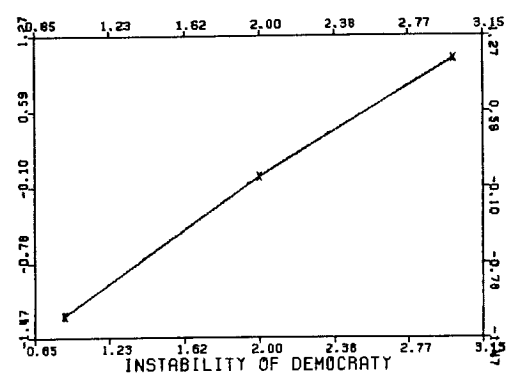


Figure 7.4.b Transformations of the discreticized data

right we see a number of countries with high GNPR, classified by Russett in his final table as stable democracies with greater than median equality. These are AUS, USA, NZE, LUX, SWE, CAN. Below that we find NOR, IRE, FRA, DEN, NET, GBR who are also democratic with a fairly high GNPR, and with greater than median equality. Only FRA is an unstable democracy. Another cluster is CHI, ITA, and ARG which are classified by Russett as unstable democracies with less than median equality. Below on the left we find dictatorships with less than median equality, in the left upper corner we find distatorships with greater than median equality. It is possible to find a refinement of Russett's classification of countries, although some countries such as WGE and SVI are not where they should be. The same clusters can be found, with much more trouble, in the figure for the rankings it is more striking however that the solutions are quite different. They are also different from the solution reported in Gifi (1980, p 227-236), which seems to indicate that the situation with stability is far from satisfactory for this example. These two new analyses are, as far as the canonical components are concerned, quite similar to each other, and also quite similar to the solution for dimension one and three of a three-dimensional CANALS solution given in Gifi (1980, figure 7.7, p 234). This seems to indicate that there are some stable effects, but they tend to occur in different dimensions in different solutions. The object scores do not seem to be stable at all. The transformations of the variables, given in figure 7.4.a and 7.4.b, are quite stable, also compared with Gifi (1980, p 236). It is possible that a solution to the instability problem is to compute more canonical variables, it is also possible that we must use more radical means in this example, for example using HOMALS or PRINCALS as a first step to quantify the variables, and perform metric canonical correlation analysis as a second step. As these possibilities have not been tried out, the same thing is true for the partial and three-set canonical analysis.

We could ask what CANALS adds to Russett's table 3, which classifies countries using DEMO and the median of the GINI-index. The clusters in figure 7.3.a show that Russett's classification is a sensible one, they also suggest some refinements, of which it is not yet known how stable they are. Figure 7.2.a and 7.2.b show the orthogonality of DEMO and INST, and the extreme importance of GNPR, showing that stable democracy and equality tend to be luxury goods, and indicating that possibly the fact that political scientists from stable Western democracies have constructed the indices has had some influence (India, Philippines, Uruguay are stable democracies; Brasil, Colombia, Argentina, Costa Rica, Chile are unstable democracies; Yugoslavia and Poland are dictatorship there is greater than median equality in Taiwan and South Vietnam). It is possible that a similar data matrix constructed in the Sovjet Union would look quite differently, and that this fact may explain some of the instability in the solutio

8: Two sets, some special cases, some future programs

8.1: Previous work

This chapter is somewhat different from the earlier chapters. As we have seen in 7.1 there are some asymmetric cases of canonical analysis which are very important in practice. In this chapter we briefly discuss multiple regression, discriminant analysis, multivariate analysis of variance, and path analysis. For each of these special cases we have planned, but not yet written, a special computer program. Clearly we could use CANALS, but the special cases make some important simplifications possible. It is also not our intention in this chapter to review the history of each of these techniques in considerable detail, we merely discuss algorithms and planned programs.

8.2: Multiple regression and MORALS

Suppose K = 2, and suppose the second set contains only one variable, which is treated as single numerical, ordinal or nominal. In this case we can suppose without loss of generality that p = 1, CANALS can be interpreted as an optimal scaling technique which maximizes the multiple correlation, and the loss function can be written quite simply as

$$\sigma_M(y_1,y_2) = SSQ(G_1y_1 - G_2y_2),$$

where G_2 is a complete indicator matrix, G_1 can be an indicator supermatrix. As far as missing data are concerned we suppose that there are no missing data in the second set. There are several ways in which missing data can occur in the first set, there are several of our options we can apply, but because we are not particulary interested in computational details of non-existing programs we ignore missing data in this chapter and refer to 6.2.4 and 6.2.5 for necessary details.

Thus $\sigma_M(y_1,y_2)$ must be minimized over the feasible y_1 and y_2. For normalization purposes we simply require $u'D_2y_2 = 0$ and $y_2'D_2y_2 = 1$, with $D_2 = G_2'G_2$. A major difference with CANALS is that we do not have to bother with switching normalizations. The program MORALS by Young, De Leeuw, and Takane (1976) already solves a similar problem (with continuous options, and without multiple nominal variables in the first set). The algorithm for adjusting y_1 is the same as in CANALS, with the obvious simplifications that result from p = 1.

As we have stated in 8.1 we do not intend to review the history of multiple regression in any detail. The linear theory is classical. It was started by Galton, and developed further by Pearson and Yule. There have been some attempts, none of them successful, to develop a theory of partial and multiple rank correlation. Box and Cox (1964) studied parametric families of transformations to improve the fit to a linear model, Kruskal (1965) was the first to use monotonic regression in this context (the first set in this case is an ANOVA-design matrix). Roskam

(1968) also fitted an additive model with Kruskal-type techniques, and related it to additive conjoint measurement. An alternating least squares program ADDALS was presented for the ANOVA application by De Leeuw, Young, and Takane (1976), this was generalized to the univariate linear model by Young, De Leeuw, and Takane (1976) with program MORALS, recently ADDALS was generalized to an individua differences additive model in the program WADDALS (Takane, Young, De Leeuw, 1980). An alternating least squares algorithm for additivity analysis, discriminant analysis, and multiple and polynomial regression was introduced earlier by De Leeuw (1969). French data analysts have also contributed a great deal. We mention Drouet d'Aubigny (1975), Tenenhaus (1977), and a very interesting recent paper by Daudin (1980).

There is, of course, much more that could be said about these generalizations of the metric linear model, but most of these additional results are in the field of gauging and stability analysis. It is also quite true that most of the questior dealing with stability and gauging have not been answered yet. In particular we can expect that the whole problem of ridge analysis, of smoothing, of cross validation, of James-Stein estimation, have analogues in this generalized multiple regression context.

8.3: Discriminant analysis and CRIMINALS

In discriminant analysis we also have a single variable in the second group, but now this variable is multiple nominal. The loss function is

$$\sigma_M(Y_1,Y_2) = SSQ(G_1Y_1 - G_2Y_2),$$

and the normalization is $u'D_2Y_2 = 0$ and $Y_2'D_2Y_2 = I$. This is convenient for the algorithm, because we do not have to switch normalizations, but it is not the most natural way to normalize the solution. Thus if we have computed the solution by the alternating least squares program CRIMINALS, we renormalize it in some convenient way, for example by requiring that $Y_1'D_1Y_1 = I$, and then by setting $Y_2 = D_2^{-1}G_2'G_1Y_1$. The canonical components of the first set G_1Y_1 are orthonormal, the Y_2 are the averages of the groups of individuals indicated by the second set. As far as history is concerned the same remarks apply as in 8.2. Discriminant analysis was invented by Fisher around 1940, it was introduced to psychologists by Rao and Slater (1948) and by Lubin (1950), and it is quite popular these days, for example in clinical psychology. A form of discriminant analysis with all variables multiple nominal was proposed by Saporta (1975), other generalizati of linear discriminant analysis are usually based on discrete MVA techniques (for example in Gilbert, 1968).

Interpretation of discriminant analysis is facilitated by defining the projector

$$P = G_2(G_2'G_2)^{-1}G_2',$$

and by observing that the minimum of $\sigma_M(Y_1,Y_2)$ over non-restricted Y_2 is equal to

$$\sigma_M(Y_1,*) = \text{tr } Y_1'G_1'G_1Y_1 - \text{tr } Y_1'G_1'PG_1Y_1.$$

The matrix $Y_1'G_1'G_1Y_1$ is the total dispersion of the canonical variables, the matrix $Y_1'G_1'PG_1Y_1$ is the between-group dispersion. Thus we maximize the trace of the between-group dispersion matrix over all quantifications with unit total dispersion matrix. If all variables are single then $Y_j = y_ja_j'$ for all j in the first group. If T is the dispersion matrix of the $q_j = G_jy_j$, and B is the between-group dispersion matrix, then

$$\sigma_M(Y,A,*) = \text{tr } A'TA - \text{tr } A'BA,$$

If we minimize this over A with restriction A'TA = I, then clearly the result is

$$\sigma_M(Y,*,*) = p - \sum_{s=1}^{p} \lambda_s(T^{-1}B),$$

or we see that our technique amounts to choosing the y_j in such a way that the sum of the p largest eigenvalues of $T^{-1}B$ is maximized. A similar interpretation is possible if not all variables in the first set are single, but the notation becomes somewhat more complicated and we omit the details. It is clearly also possible in this case to choose other criteria defined in terms of the eigenvalues of $T^{-1}B$ which must be optimized, as in 6.1.2, but we have no experience with any other choices.

8.4 Multivariate analysis of variance and MANOVALS

Suppose the indicator supermatrix G_1 in the first set is a design matrix of a complete orthogonal design. The requirement that the design is orthogonal is not as restrictive as it looks, because if we allow for missing data we can always make the design balanced in such a way that it becomes orthogonal. This is an old trick in analysis of variance literature which dates back to the iterative alternating least squares method proposed by Yates (1933). The design matrix is treated as a set of orthogonal multiple nominal variables, each of the factors can be used to define a projector P_j, and the orthogonal projectors can be used to write the loss function minimized over Y_1 (the effects) in the form

$$\sigma_M(*,Y_2) = \text{tr } Y_2'G_2'G_2Y_2 - \sum_{j=1}^{m} \text{tr } Y_2'G_2'P_jG_2Y_2,$$

more or less as in the previous section, but with Y_1 and Y_2 interchanged, and with the added refinement due to orthogonality of the design components. The technique that is most natural from our point of view maximizes the second component of the loss function with requirements $Y_2'G_2'G_2Y_2 = I$, but it is a relatively small step to see that we can actually maximize any sum of the

$Y_2'G_2'P_jG_2Y_2$, while restricting any other sum to be unity. Thus we can maximize the sum of the main effects for fixed total, we can also maximize the sum of the first order interactions for fixed main effects, and so on. This possibility was already indicated, in a metric context, by Abelson (1960). The first one to apply ANOVA to purely qualitative data, using optimal scaling, was Fisher (example 46.2 in 'Statistical methods for research workers'). Other references are in section 8.2, and in De Leeuw, Young, and Takane (1976).

8.5 Path analysis and PATHALS

Path analysis may be relatively unfamiliar to some of our readers, and we give a very short introduction. Suppose H_1 and H_2 are $n \times m_1$ and $n \times m_2$ data matrices. The columns of H_1 are n measurements on m_1 exogeneous variables, the columns of H_2 are n measurements on m_2 endogeneous variables. The exogeneous variables are assumed to influence all endogeneous variables linearly, the endogeneous variables are ordered in such a way that an endogeneous variable is influenced linearly by all previous endogeneous variables. The equation is

$$H_2 = H_1A + H_2B + E,$$

where B is upper triangular (all elements on and below the diagonal are equal to zero), and were E, the residuals, are small. We can require that the E have other properties usually associated with residuals, for example

$$E'H_1 = 0,$$

$$E'E = \text{diag}\,(E'E).$$

It is of some interest that these two structural assumptions about the errors make it possible to solve uniquely for A, B, and E. We first write the reduced form

$$H_2 = H_1A(I - B)^{-1} + E(I - B)^{-1}.$$

Now $S_{11} = H_1'H_1$, $S_{12} = H_1'H_2$, $S_{22} = H_2'H_2$, and write D for the diagonal matrix E'E. Then

$$S_{12} = S_{11}A(I - B)^{-1},$$

$$S_{22} = (I - B')^{-1}A'S_{11}A(I - B)^{-1} + (I - B')^{-1}D(I - B)^{-1},$$

or

$$S_{22} - S_{21}S_{11}^{-1}S_{12} = (I - B')^{-1}D(I - B)^{-1}.$$

This equation can be solved by Cholesky decomposition of $S_{22} - S_{21}S_{11}^{-1}S_{12}$, which is identified by diag (I - B) = I. Thus we have D and B, and we can find A from

$$A = S_{11}^{-1}S_{12}(I - B),$$

and E from

$$E = H_2(I - B) - H_1A.$$

We have merely shown in this section so far that given any partitioning of the variables into endogeneous and exogeneous ones, and given any ordering of the endogeneous variables, we can perform a complete and recursive path analysis, which is merely a transformation of the data (depending on both the partitioning and the order). A path model becomes restrictive if we suppose that some of the upper diagonal elements of B are zero, or that some of the elements of A are zero. In this case we need a loss function again.

The usual loss function is simply

$$\sigma_M(A,B) = SSQ(H_2(I - B) - H_1A),$$

which looks like two-set canonical analysis, but something remerkably simplifying is true here. The term $H_2(I - B)$ has row j equal to

$$h_{2j} - \sum_{i=1}^{j-1} h_{2i}b_{ij},$$

and because h_{2j} does not occur homogeneously (it has no coefficient) we do not have to normalize in any way. Moreover we can solve for each column separately, by using ordinary multiple regression, also if some of the elements of b_j or a_j are restricted to zero.

It is trivial to generalize path analysis to variables with single numerical, ordinal, or nominal type. We write, changing the notation somewhat,

$$\sigma_M(A,B,Y) = SSQ(Q_1A - Q_2B),$$

with $q_j = G_jy_j$ as usual, and with A and B restricted to have zeroes on specified places, moreover A must have ones on the diagonal. If desired more general restrictions are possible, in fact we can require in theory that A is in a convex set K_1, B is in a convex set K_2, and either $0 \notin K_1$ or $0 \notin K_2$. Fitting Q_1 and Q_2 for fixed A and B is already familiar. The elements of A and B can be fitted column by column, but in most cases it is quite easy to find all of A for fixed B (and Q_1 and Q_2) and all of B for fixed A (and Q_1 and Q_2). As we have seen in the linear case it is even possible if diag (A) = I to find all of A and all of B for fixed Q_1 and Q_2.

It is more difficult, but also quite interesting, to define recursive path models for multiple variables. The loss function is

$$\sigma_M(Y) = SSQ(\sum_{j\varepsilon J_1} G_jY_j - \sum_{j\varepsilon J_2} G_jY_j).$$

In stead of requiring that certain elements of A and B are zero, we remember that $Y_j = y_ja_j'$ in the single case, and require in the multiple case consequently that certain columns of the Y_j are zero. A little thought shows that this amounts

to fitting path models in the usual single-equation way, as in the linear case, but now each variable gets a different quantification in each equation that is fitted. If we predict variable j, then all exogeneous variables get a transformati which is different from their previous transformations, all preceding endogeneous variables also get a transformation which is different from their earlier ones.

8.6 Partial canonical correlation analysis and PARTALS

Partial canonical correlation analysis was introduced by Roy (1957, p 143), it was discussed more extensively by Rao (1969) and by Timm and Carlson (1976). The last authors also discuss canonical equivalents of part and bipartial correlation analysis. Of course all papers deal exclusively with the case in which all variables are numerical.

We have seen that CANALS in p dimensions transforms or quantifies the variables in such a way that the sum of the p largest canonical correlations is maximized. This is one trivial way in which we can proceed: simply define PARTALS as the technique which maximizes the sum of the first p partial, part, or bipartial canonical correlation coefficients. For completeness, and because the numerical techniques are not very familiar yet, we repeat some of the definitions given by Timm and Carlson (1976). In partial canonical correlation analysis there are three sets, the third set is partialled out of the first two, the equations are the same as for ordinary canonical correlation analysis, except for the fact that the dispersion matrix of the first two sets is replaced by the conditional dispersion matrix, with the third set partialled out. In part canonical correlation we only remove the third set from one of the two other sets, in bipartial canonical correlation analysis we have four sets, set three is removed from set one, and set four is removed from set two. 'Removing' and 'partial' are always understood in a linear sense.

We call this generalization trivial, because it is not at all clear how we want to incorporate this in an interesting algorithm. For partial canonical correlation there is an easy possibility. We add the third set to both other sets, and perform a canonical correlation analysis over the two enlarged sets. Because of the adding there will be a number of canonical correlations equal to unity (equal to the number of variables in the third set, or even more than that if some variables in the third set are multiple nominal). The remaining compounds, however, will all be orthogonal to these first 'trivial' compounds, which means that they will be orthogonal to the third set, as required. This is essentially the approach suggested by Roy (1957, p 142), and it is easy to generalize it to both single and multiple non-numerical variables.

8.7 Some examples

We use the CBS-data also used in 7.4.1 again for our examples in this chapter. In order to avoid possible misunderstandings we emphasize that all the computing in this section is done with more or less ad hoc APL-programs, and that there are no friendly FORTRAN programs for the techniques discussed in this chapter. We use three variables from the CBS-data: father's profession P, school achievement test score T, and type of secondary education chosen E.

The first analysis is of the multiple regression type. We want to predict E from P and T. The model is illustrated in figure 8.1.a. We first interprete this in a general nonlinear sense, i.e. $\omega(E) = \Psi(P,T)$, and then in the nonlinear but additive sense $\omega(E) = \psi(P) + \phi(T)$. Table 8.1 contains canonical correlations, and quantifications of E. They are extremely similar in both analyses. They illustrate the large distance between VWO/HAVO and MAVO, and the fact that LEAO/LMO is much closer to MAVO than LTS/LHNO is to MAVO. Similar results were found by De Leeuw and Stoop (1980). The quantifications of the (P,T) combinations are given in table 8.2, and plotted in figure 8.2. It is clear from the figure that children from the lower P-levels choose their secondary education too low, and that this effect is strongest for the lower levels of T and weakest for the higher levels of T. Thus not too much 'talent' is 'lost' in the high-T region, but a great deal is lost in the low-T region. We must warn the reader that our conclusions are framed in a very suggestive way, and refer only to transformations and interactions found by our technique. If the scales we construct are stable or even socially relevant remains to be seen.

There are two other path models in figures 8.1.b and 8.1.c. We fit them by using the multiple nominal approach, which means that we simply have to combine results of various regression analyses. The results (transformations standardized to unit variance, regression weights, and residual variance) are in table 8.3. The regression weights and residuals can be filled in in figure 8.1.

We can also use P and T as predictors (single nominal) in a discriminant analysis problem, where the four levels of E (multiple nominal) have to be predicted. The results are in table 8.4 We give standardized transformations of T and P, the weights A that determine the two canonical axes, the means of the four E-levels on the canonical axes, and the value of $1 - \text{tr } T^{-1}B$, which is the loss function. The solution is almost completely identical to the third column of table 8.3, which is easily explained because the group-means are on a straight line through the origin, and consequently the solution with E single nominal and $p = 1$ is almost identical to E multiple nominal and $p = 2$.

In table 8.5 the results of a partial canonical correlation are given. We predict E|T from P|T (first column) and compare this with predicting E from P (second

column). The results are quite different. It is difficult to predict E from P (compare also the analysis in chapter 7, where T is predicted from P), but if we partial out T prediction becomes worse. Observe from tables 8.3-8.5 that P explains more variance of E than of T. The difference is 0.04, which is approximately equal to the partial canonical correlation.

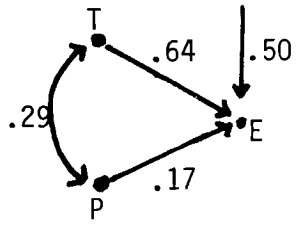


figure 8.1.a:
regression path model

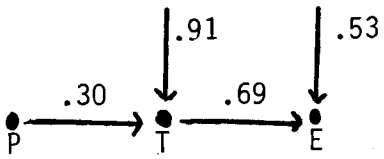


figure 8.1.b:
another path model

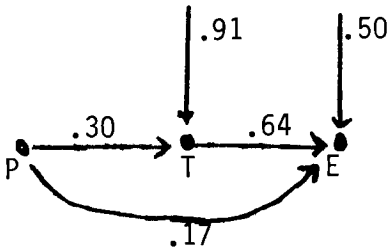


figure 8.1.c.:
a third path model

0.71	0.70	
0.37	0.35	
0.07	0.07	
1.51	1.52	VWO/HAVO
0.24	0.23	MAVO
-1.30	-1.29	LTS/LHNO
-0.83	-0.83	LEAO/LMO

table 8.1: canonical correlations and quantifications of school type for completely nonlinear and additive regression analysis.

-1.65	-0.97	0.03	0.93	1.69
-1.58	-1.01	-0.06	0.89	1.52
-1.52	-0.80	-0.09	0.83	1.79
-1.35	-0.64	0.29	1.17	1.71
-1.35	-0.54	0.34	1.11	1.70
-1.33	-0.37	0.43	1.34	1.69
-1.04	-0.06	0.72	1.41	1.79
-1.51	-0.96	-0.06	0.90	1.49
-1.54	-0.99	-0.09	0.87	1.46
-1.48	-0.93	-0.04	0.93	1.52
-1.27	-0.72	0.18	1.14	1.73
-1.23	-0.68	0.22	1.18	1.78
-1.08	-0.53	0.36	1.33	1.92
-0.89	-0.34	0.55	1.52	2.11

table 8.2: quantifications of (P,T) combinations for completely nonlinear and additive regression analysis.

	T→E	P→T	T+P→E_S
E:	-1.52	-	-1.52
	-0.23	-	-0.23
	1.28	-	1.29
	0.88	-	0.83
T:	1.59	1.66	1.59
	0.82	0.79	0.82
	-0.18	-0.23	-0.18
	-1.15	-1.01	-1.15
	-1.75	-1.96	-1.73
P:	-	1.04	0.69
	-	0.75	0.88
	-	0.09	0.63
	-	-0.45	-0.50
	-	-0.56	-0.62
	-	-1.21	-1.18
	-	-1.86	-1.95
βT	0.69	-	0.64
βP	-	0.30	0.17
$1-R^2$	0.53	0.91	0.50

table 8.3: regression results for path models

	T+P→E_M	
	-0.37	-0.66
	-0.05	-0.10
	0.31	0.56
	0.20	0.36
	1.58	
	0.83	
	-0.18	
	-1.16	
	-1.73	
	0.69	
	0.89	
	0.66	
	-0.59	
	-0.66	
	-1.17	
	-1.91	
A:	0.11	0.13
	0.30	0.56
	0.50	

table 8.4: discriminant analysis results

P|T→E|T	P→E
-2.04	-1.50
-0.31	-0.23
1.83	1.32
0.52	0.50
-	-
-	-
-	-
-	-
-	-
1.01	0.70
0.97	0.52
1.39	0.43
-0.26	-0.77
-0.27	-0.65
-0.77	-1.12
-2.07	-2.58
0.95	0.87

table 8.5: partial and ordinary nonlinear regression

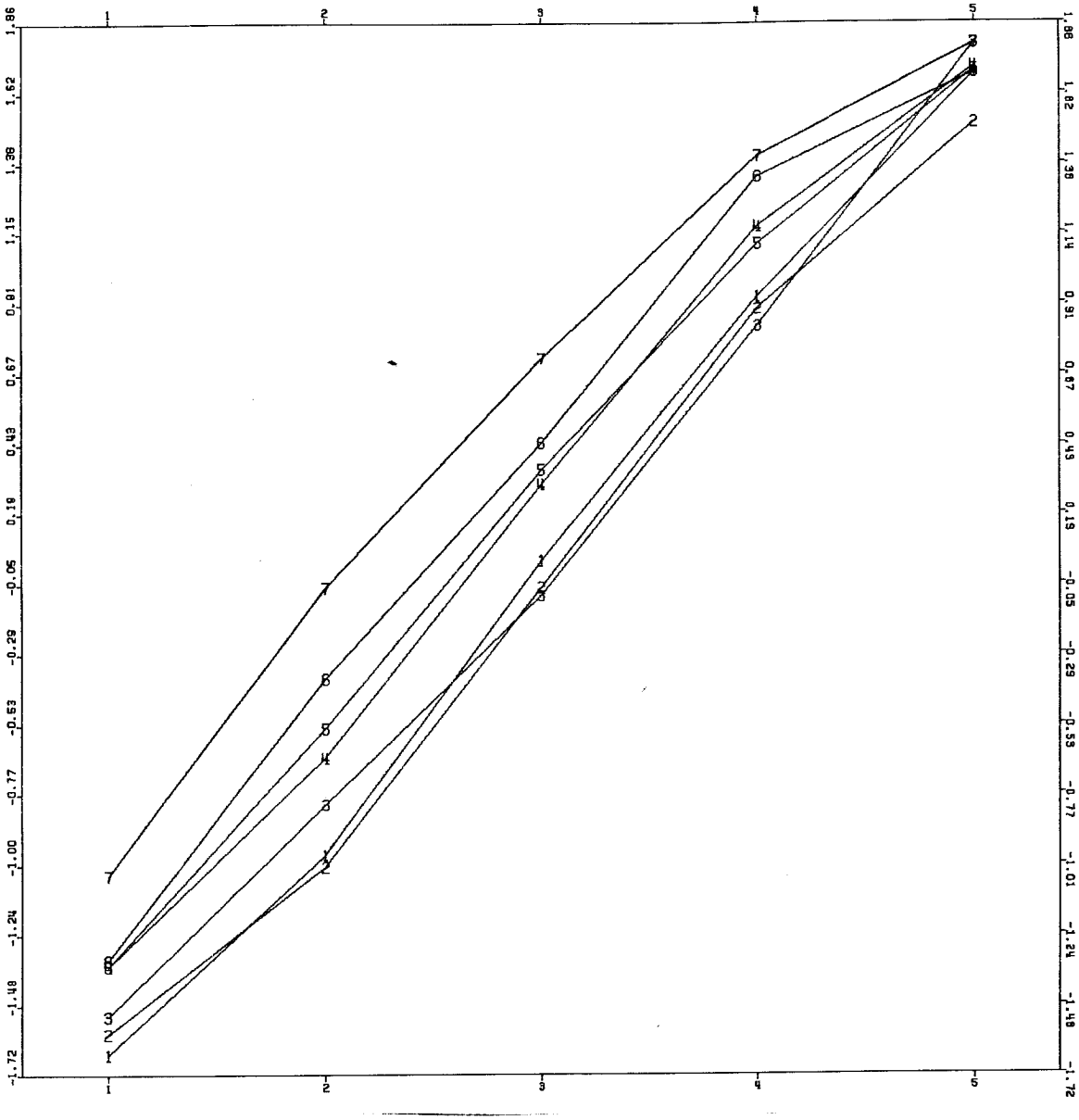


Figure 8.2.a (P,T) combinations after nonlinear regression analysis

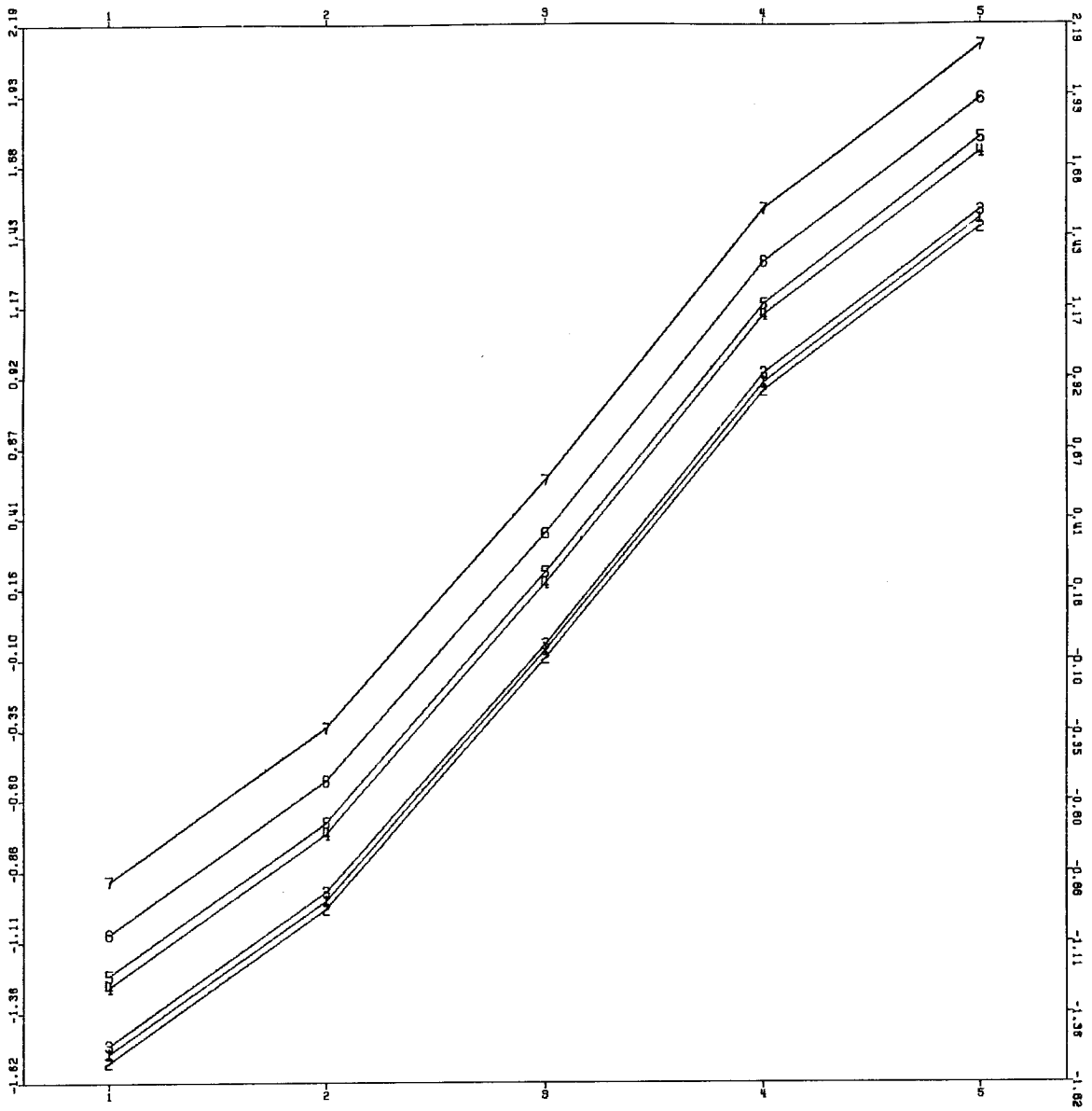


Figure 8.2.b (P,T) combinations after additive regression analysis

9: The analysis of binary variables

9.1 Introduction

In the previous chapters we have seen many times that binary variables are somewhat special. The major reason is that quantification is unnecessary, the distinction between single and multiple becomes irrelevant, the distinction between numerical, ordinal, and nominal becomes irrelevant too. We simply compute product moment correlations ('phi-coefficients') between the variables and perform linear MVA on the resulting correlation matrix. This assumes, of course, that there are no missing data and that all $n \times 2$ indicator matrices G_j are complete. In chapter 4 we have seen, however, that binary data are also interestin if they occur in incomplete indicator matrices.

Another reason to treat binary data separately is, of course, that they are very common. In the book by Coombs (1964), for example, this is amply illustrated There are many 'yes-no' data matrices, many other data matrices are actually derived from this more basic form. Most data in psychological testing, archeological seriation, ecological ordination, politicological voting analysis, are of this type. As a consequence many probabilistic and algebraic models have been developed which can be used as gauges for our techniques. This is one of the main preoccupations of this chapter. Because probabilistic gauges are quite prominent in this chapter we use the random variable notation. In chapter 1 we have already indicated that the ordinary algebraic matrix notation we have used so far is a 'degenerate' special case.

9.2 Some general formulas

Suppose $\underline{h}_j$ and $\underline{h}_\ell$ are binary variables. Define $\pi_j \triangleq \text{AVE}(\underline{h}_j)$ and $\pi_\ell \triangleq \text{AVE}(\underline{h}_\ell)$. Moreover $\pi_{j\ell} = \text{AVE}(\underline{h}_j\underline{h}_\ell)$. Then $\text{VAR}(\underline{h}_j) = \pi_j(1 - \pi_j)$ and $\text{VAR}(\underline{h}_\ell) = \pi_\ell(1 - \pi_\ell)$ and $\text{COV}(\underline{h}_j,\underline{h}_\ell) = \pi_{j\ell} - \pi_j\pi_\ell$. Because $0 \leq \pi_{j\ell} \leq \min(\pi_j,\pi_\ell)$ we find that

$$-\pi_j\pi_\ell \leq \text{COV}(\underline{h}_j,\underline{h}_\ell) \leq \min\{\pi_j(1 - \pi_\ell),\pi_\ell(1 - \pi_j)\}.$$

Thus if we define ν_j by $\nu_j = \{\pi_j/(1 - \pi_j)\}^{\frac{1}{2}}$, then

$$-\nu_j\nu_\ell \leq \text{COR}(\underline{h}_j,\underline{h}_\ell) \leq \min(\nu_j/\nu_\ell,\nu_\ell/\nu_j).$$

Thus the phi-coefficient $\phi_{j\ell} \triangleq \text{COR}(\underline{h}_j,\underline{h}_\ell)$ is bounded by functions which depend on the marginals. We can only have $\phi_{j\ell} = +1$ if $\pi_j = \pi_\ell$, we can only have $\phi_{j\ell} = -1$ if $\pi_j = \pi_\ell = \frac{1}{2}$. Because psychometricians like their correlations to be high they have suggested ϕ/ϕ_{max} as a measure of association, where ϕ_{max} is the upper bound derived above. There are other reasons to study ϕ/ϕ_{max}. They will become clear when we introduce some of the more common gauges.

9.3 Monotone latent trait models

9.3.1 General observations

Suppose $\underline{x}$ is an unobservable latent trait such that $p_j(x) \triangleq \text{AVE}(\underline{h}_j \mid \underline{x} = x)$ is

an increasing function of x. Of course $\pi_j = AVE(p_j(\underline{x}))$, where the expectation is taken with respect to the distribution of the latent trait. We also assume conditional or local independence: for all $j \neq \ell$ we have

$$AVE(\underline{h}_j, \underline{h}_\ell \mid \underline{x} = x) = p_j(x)p_\ell(x),$$

which implies

$$\pi_{j\ell} = AVE(p_j(\underline{x})p_\ell(\underline{x})).$$

Because

$$COV(\underline{h}_j, \underline{h}_\ell) = AVE\{(p_j(\underline{x}) - p_j(\underline{y}))(p_\ell(\underline{x}) - p_\ell(\underline{y}))\},$$

it is clear that $COV(\underline{h}_j, \underline{h}_\ell) \geq 0$, and thus $\phi_{j\ell} \geq 0$. In monotone latent trait models correlations are never negative. This improves the lower bound in 9.2. We remember from Perron-Frobenius theory that non-negativity of the correlation matrix implies the existence of a unique largest eigenvalue with correponding non-negative eigenvector. Assuming a monotone latent trait model thus imposes some structure on the observed correlation matrix, but not much.

Additional structure is imposed if we assume that the tracelines $p_j(x)$ do not cross each other. i.e. that $p_j(x) \lesseqgtr p_\ell(x)$ for one single x means that $p_j(x) \leq p_\ell(x)$ for all x. Such curves $p_j(x)$ are called holomorph by Mokken (1971). A measurement theory analysis of holomorphic systems, given by Levine (1970, 1972, 1975), shows that they can be written in the form $p_j(x) = p(x - \theta_j)$, with θ_j a real parameter.

Now suppose $\theta_j \geq \theta_\ell$. We can make two by two tables of items j and ℓ with another item g, and compare the entries. The tables are given below, together with their marginals. With a '+' we indicate that this element is larger than the corresponding element in the other table, which then gets a '-'. Equal elements in both tables are denoted '='.

j \ g	1	0	
1	-	-	-
0	+	+	+
	=	=	

ℓ \ g	1	0	
1	+	+	+
0	-	-	-
	=	=	

The proofs of these relations are simple. If $\theta_j \geq \theta_\ell$ then $p_j(x) \leq p_\ell(x)$ for all x. The result for the (0,0)-cell is then proved, for example, by observing that $(1 - p_j(x))(1 - p_g(x)) \geq (1 - p(x))(1 - p_g(x))$, and by taking expected values over both sides of this inequality. This ordering of the cells is an interesting structural property, which is explored for example in the book of Mokken (1971), but unfortunately it does not seem to have very clear consequences for the $\phi_{j\ell}$, unless we make additional assumptions on the $p_j(x)$.

9.3.2 The Guttman scale

Guttman (1944, 1950a,b) introduced the model with the additional assumption

$$p_j(x) = \begin{cases} 0 \text{ if } x < \theta_j, \\ 1 \text{ if } x \geq \theta_j. \end{cases}$$

The model is of a breath-taking simplicity, it has a very simple interpretation in terms of physical measurement operations, and the probabilistic aspect is essentially trivialized, which makes the model algebraic. It follows from Guttman's assumption that $\pi_{j\ell} = \min(\pi_j, \pi_\ell)$, and in the two times two bivariate distribution at least one of the off-diagonal cells is zero. If we order the variables in such a way that $\theta_1 \geq \ldots \geq \theta_m$, then also $\pi_1 \leq \ldots \leq \pi_m$ and also $\nu_1 \leq \ldots \leq \nu_m$. For $j \leq \ell$ thus $\phi_{j\ell} = \nu_j/\nu_\ell$, and for $j \geq \ell$ thus $\phi_{j\ell} = \nu_\ell/\nu_j$. It is thus necessary (and trivially also sufficient) for the existence of a Guttman-scale (or: perfect scale) that $\phi = \phi_{max}$ for all pairs of variables, or that $\omega = \phi/\phi_{max}$ is equal to one for all pairs. The use of ω in item analysis dates back to Loevinger (1947, 1948). Mokken (1971) uses $\omega_{j\ell}$ as a measure of holomorphy of his items, but this is somewhat risky. There are perfectly holomorph items with arbitrary low $\omega_{j\ell}$. Examples can be constructed by using, for example, our formula for ϕ/ϕ_{max} in the Rasch model of section 9.3.5. The programs developed by Mokken and his students must consequently be interpreted as methods to construct Guttman-scales, not as tests of general holomorphy.

It is clear that Guttman's model imposes an enormous amount of structure on the correlation matrix. It is interesting to find out if the same thing is true for the eigenvalues and eigenvectors of the correlation matrix. Guttman (1950b) gives a very satisfactory mathematical analysis, but he does not seem to know that the necessary results were derived much earlier by Gantmacher and Krein (1936 Guttman (1954) discusses the interpretation of the eigenvectors in detail, but since all eigenvectors are of course functions of the θ_j it is somewhat problematic if such interpretations are useful.

We briefly discuss the most important results for the eigenvalues and eigenvectors In the first place one could conjecture that if y is the eigenvector corresponding with the dominant eigenvalue and $\nu_1 \leq \ldots \leq \nu_m$ then also $y_1 \leq \ldots \leq y_m$. This is not true, however, in fact it is possible to prove that $y_1 \leq y_2$ and that $y_{m-1} \geq y_m$. We only show how to prove the first inequality, the second one is proved in the same way. We suppose that y is an eigenvector of a matrix R with positive elements, of order $m \geq 3$, and that all elements of y are also positive. Moreover we suppose that $1 = r_{jj} \geq r_{j,j+1} \geq \ldots \geq r_{jm}$ for all j as well as $1 = r_{jj} \geq r_{j,j-1} \geq \ldots \geq r_{j1}$ for all j. The correlation matrix of a perfect scale has all these properties. Because y is an eigenvector

$$\sum_{j=1}^{m} (r_{1j} - r_{2j})y_j = \lambda(y_1 - y_2),$$

or

$$\{\lambda - (1 - r_{12})\}(y_1 - y_2) = \sum_{j=3}^{m} (r_{1j} - r_{2j})y_j.$$

Because $r_{1j} \leq r_{2j}$ and $y_j > 0$ the term of the right is nonpositive. Moreover $\lambda > 1 - r_{12}$, because λ is the unique largest eigenvalue of R (Perron-Frobenius). Thus $y_1 \leq y_2$, and we have $y_1 = y_2$ if and only if $r_{1j} = r_{2j}$ for all j=3,...,m. This is a negative result, but there are also some positive ones. For proofs we refer to Guttman (1950b), Gantmacher and Krein (1936, 1950), Karlin (1964, 1968). In the first place the inverse of the correlation matrix of a perfect scale is tridiagonal. In the second place we can plot the elements of the eigenvectors y_s against the rank numbers of the v_j and connect successive points by straigth line segments. This produces m polygonal functions, one for each eigenvector, with interesting properties. Function s, corresponding with eigenvector y_s and s-th largest eigenvalue λ_s, has s - 1 zeroes. Moreover the zeroes of successive functions are interwoven: between each pair of successive zeroes of function s is exactly one zero of function s - 1. But this is not all: is we choose the sign of the eigenvectors in such a way that all y_{1s} are nonnegative, than the same separation properties are true for the transpose of the matrix of eigenvectors. It is clear that these properties make it easy to recognise the correlation matrix of a Guttman-scale from its eigenvalue-eigenvector properties. Another interesting property is that all eigenvalues are different if all v_j are different, which is true if and only if all π_j are different.

The elements of the eigenvector y_1 are not monotonic with the π_j, but we must remember that homogeneity analysis does not solve $Ry = \lambda y$ but $Ct = \lambda Dt$, with C the covariance matrix of the catgeories and D the diagonal matrix of the variances. The eigenvalues of the two problems are the same but the eigenvectors certainly are not, for one thing t has 2m elements, while y has only m elements. Guttman (1950) proves that the elments of t are monotone with the π_j. We only indicate the general outline of the proof, working in the general context of minimization of

$$\sigma(\underline{x},\alpha,\beta) = \frac{1}{m} \sum_{j=1}^{m} SSQ(\underline{x} - \alpha_j\underline{h}_j - \beta_j(1 - \underline{h}_j)).$$

The $\underline{h}_j$ and $\underline{x}$ are random variables defined on a common probability space, with elements that are supposed to be ordered. The $\underline{h}_j$ are known, and there are constants θ_j in the domain of the functions $h_j(\theta)$ such that

$$\underline{h}_j(\theta) = \begin{cases} 0 \text{ if } \theta < \theta_j, \\ 1 \text{ if } \theta \geq \theta_j, \end{cases}$$

which implies that the $\underline{h}_j$ are all monotonic. Because $\pi_j = AVE(\underline{h}_j)$ and we suppose that $\pi_1 \leq \ldots \leq \pi_m$ we also have $\theta_1 \geq \ldots \geq \theta_m$. We now apply our alternating least squares algorithm, with the normalization conditions $AVE(\underline{x}) = 0$ and $VAR(\underline{x}) = 1$. We know that it converges for almost all starting points to the appropriate solution. Suppose we start with a monotonic $\underline{x}$. Then the least squares estimate of α_j is $AVE(\underline{x} \mid \underline{x} \geq \theta_j)$ and that of β_j is $AVE(\underline{x} \mid \underline{x} < \theta_j)$. Thus $\alpha_1 \geq \ldots \geq \alpha_m \geq 0$ and $0 \geq \beta_1 \geq \ldots \geq \beta_m$. In the second step of the alternating least squares algorithm we compute a new $\underline{x}$ for current α and β. This new $\underline{x}$ is the standardized version of

$$\underline{z} = \sum_{j=1}^{m} \alpha_j \underline{h}_j + \sum_{j=1}^{m} \beta_j (1 - \underline{h}_j).$$

Because the elements of α are nonnegative and those of β are nonpositive $\underline{z}$ is again monotonic. From the definition of α and β moreover $AVE(\underline{z}) = 0$. After one cycle of the alternating least squares algorithm, starting with a monotone $\underline{x}$, we have found weights (or category quantifications) α and β in the appropriate order, and a new $\underline{x}$, which is still monotone. Thus subsequent iterations will not change the order-properties, and convergence is to a monotone $\underline{x}$ and to α and β in the appropriate order. The vector t which solves $Ct = \lambda Dt$ is proportional to $\alpha - \beta$, and is consequently also in the appropriate order. The oscillation theorems about sign-changes or zeroes which were true for the eigenvectors y_s of R are also true for the vectors t_s.

If we do not know the distribution of the individuals on the latent continuum, then we cannot recover the θ_j, only their order. Thus we have shown that the one-dimensional HOMALS solution recovers the interesting information without distortion, and that the remaining solutions show an interesting pattern which makes clear that we are dealing with something close to a perfect scale. There are some interesting special cases, for example $\nu_j = \nu^j$, in which the eigenvalues and eigenvectors of R can be determined explicitly (Guttman, 1950b, Goldberg, 1958, p 184-189), but we shall not go into this any further. It is also clear that computing eigenvalues and eigenvectors of R is not the simplest way to test if we are dealing with a perfect scale. Of course it is easier to use the relationship $-\ln \phi_{j\ell} = |\eta_j - \eta_\ell|$, with $\eta_j \triangleq \ln \nu_j = \frac{1}{2} \text{ logit } \pi_j$, in combination with, for example, a scaling program. Such a procedure, however, makes no sense if the data do not conform to the perfect scale pattern, while our eigenvector-eigenvalue techniques also provide interesting information in nonperfect cases.

It is sometimes said in psychological scaling literature that the Guttman scale is not realistic, and consequently not interesting. In the first place this remark seems based on the wrong interpretation of the concept of a gauge (or

model). The Guttman-scale, the Normal distribution, the Spearman hierarchy, the Republic of Plato, and the Life of Jesus, are norms. They show what happens if things are perfect. It is a completely trivial observation that the real world is not perfect. In the second place the other models for binary data are not very realistic either. People have rejected Guttman scaling partly because it became connected with nonmetric error theories and heuristic algorithms based on permutation and search. We have observed the oscillatory principal components in almost all the properly constructed attitude and aptitude scales, usually in the form of a 'horse-shoe'.

Properties of a Guttman-scale are illustrated in table 9.1. The first row in table 9.1.a is π, the second row is the variance $\pi(1 - \pi)$, the third row is ν, the next nine rows are the correlation matrix R, and the last row are the eigenvalues of $\frac{1}{m}$ R. Observe that the items in this example are rather difficult, all π_j are less than .50. As a consequence the variances are an increasing function of the number of the variable, which is not true in other examples with both difficult and easy items. Table 9.1.b contains eigenvectors of R in the first nine rows, and generalized eigenvectors of C, D in the last nine rows. The eigenvectors of R are plotted against number of the variable in figure 9.1, the first set of nine points of the first eigenvector is the first polygonal function, the horizontal axes shows the property of the increasing number of zeroes. The interlacing of zeroes is illustrated in table 9.1.c, which computes the zeroes of the polygonal functions in figure 9.1 in its first eight rows, we also have computed zeroes of the transpose of the matrix of eigenvectors and of the generalized eigenvectors. These zeroes are given, respectively, in the two other subtables of table 9.1.c. Observe that all components are 'difficulty factors' in the sense of Guilford (1941) or Ferguson (1941), because they are all functions of the item difficulty.

9.3.3 The Spearman hierarchy

We describe another gauge under the name 'Spearman hierarchy'. This name may be somewhat confusing because Spearman's one factor model, that we have in mind, was certainly not meant for binary variables. Spearman started out from the assumption of continouous variation, and constructed a simple linear latent structure model. We construct a similar linear model for binary variables, and show that it produces a similar correlation matrix as the continuous model. This is why we still use the name 'Spearman hierarchy'.

We now assume that $p_j(x) = \alpha_j x + \beta_j$. Without loss of generality we can also assume that $AVE(\underline{x}) = 0$ and $VAR(\underline{x}) = 1$. This implies immediately that $\pi_j = \beta_j$ and $COV(\underline{h}_j, \underline{h}_\ell) = \alpha_j \alpha_\ell$. Thus if $\mu_j \triangleq \alpha_j / \{\beta_j(1 - \beta_j)\}^{\frac{1}{2}}$, then $\phi_{j\ell} = \mu_j \mu_\ell$. The Spearman hierarchy, defined by a linear item characteristic function thus gives

π	.01	.04	.08	.14	.20	.26	.33	.39	.45
σ^2	.01	.04	.07	.12	.16	.19	.22	.24	.25
ν	.10	.20	.30	.40	.50	.60	.70	.80	.90

R	.50							
	.33	.67						
	.25	.50	.75					
	.20	.40	.60	.80				
	.17	.33	.50	.67	.83			
	.14	.29	.43	.57	.71	.86		
	.13	.25	.38	.50	.63	.75	.88	
	.11	.22	.33	.44	.56	.67	.78	.89
λ	.58	.18	.09	.05	.03	.02	.02	.01	.01

table 9.1.a: Guttman scale statistics

0.14	0.49	0.69	0.48	0.20	0.05	0.01	0.00	0.00
0.24	0.52	0.12	-0.51	-0.56	-0.28	-0.08	-0.01	-0.00
0.31	0.39	-0.29	-0.31	0.39	0.58	0.29	0.06	0.00
0.36	0.20	-0.40	0.17	0.36	-0.37	-0.57	-0.22	-0.03
0.38	0.00	-0.29	0.39	-0.16	-0.33	0.48	0.49	0.10
0.39	-0.16	-0.08	0.29	-0.40	0.27	0.16	-0.62	-0.29
0.38	-0.26	0.12	0.03	-0.22	0.38	-0.44	0.24	0.56
0.37	-0.32	0.26	-0.21	0.12	0.00	-0.16	0.38	-0.68
0.34	-0.32	0.31	-0.32	0.34	-0.34	0.34	-0.34	0.34
1.42	4.94	6.93	4.85	2.01	0.55	0.09	0.01	0.00
1.25	2.72	0.63	-2.65	-2.91	-1.45	-0.39	-0.05	-0.00
1.13	1.42	-1.05	-1.11	1.41	2.11	1.04	0.22	0.02
1.04	0.57	-1.16	0.49	1.06	-1.08	-1.65	-0.65	-0.08
0.96	0.01	-0.72	0.98	-0.40	-0.82	1.20	1.24	0.26
0.89	-0.35	-0.18	0.66	-0.92	0.61	0.35	-1.40	-0.66
0.82	-0.56	0.26	0.07	-0.47	0.82	-0.94	0.51	1.20
0.75	-0.65	0.54	-0.43	0.25	0.00	-0.34	0.78	-1.40
0.68	-0.65	0.63	-0.65	0.67	-0.68	0.68	-0.69	0.69

table 9.1.b: eigenvectors and generalized eigenvectors of Guttman scale.

				5.02				
			2.29		6.39			
		1.49		3.64		7.14		
	1.26		2.59		4.70		7.64	
1.16		2.32		3.61		5.55		8.00
1.11		2.21		3.34		4.54		6.26		8.32
1.08		2.14		3.22		4.31		5.44		6.72		8.53
1.05		2.10		3.15		4.20		5.26		6.34		7.45		8.67

				3.19				
			2.57		4.44			
		2.33		3.70		5.49		
	2.01		3.43		4.71		6.41	
1.72		3.22		4.42		5.60		7.20	
1.59		2.68		4.13		5.36		6.47		7.65	
1.54		2.55		3.56		4.63		6.00		7.30		8.36	
1.51		2.51		3.49		4.49		5.50		6.50		7.50		8.50

table 9.1.c: zero of polygonal Guttman eigenvectors and transposed eigenvectors

				5.02				
			2.37		6.41			
		1.65		3.69		7.14		
	1.41		2.67		4.73		7.65	
1.27		2.41		3.66		5.57		8.00
	1.19		2.27		3.39		4.58		6.27		8.33
1.14		2.19		3.26		4.34		5.47		6.73		8.53
1.10	2.13		3.18		4.23		5.28		6.36		7.46		8.67

table 9.1.c: zeroes of polygonal Guttman generalized eigenvectors

μ	.10	.20	.30	.40	.50	.60	.70	.80	.90
R	.02								
	.03	.06							
	.04	.08	.12						
	.05	.10	.15	.20					
	.06	.12	.18	.24	.30				
	.07	.14	.21	.28	.35	.42			
	.08	.16	.24	.32	.40	.48	.56		
	.09	.18	.27	.36	.45	.54	.63	.72	
λ	3.33	.99	.95	.90	.82	.72	.59	.44	.26

table 9.2.a: Spearman hierarchy statistics

0.07	0.99	0.07	0.04	0.03	0.02	0.02	0.01	0.01
0.14	-0.09	0.97	0.12	0.07	0.05	0.03	0.03	0.02
0.21	-0.05	-0.17	0.94	0.16	0.09	0.06	0.04	0.03
0.27	-0.03	-0.09	-0.25	0.90	0.18	0.10	0.07	0.04
0.33	-0.03	-0.06	-0.12	-0.34	0.84	0.20	0.11	0.07
0.38	-0.02	-0.05	-0.09	-0.16	-0.44	0.76	0.20	0.10
0.42	-0.02	-0.04	-0.07	-0.11	-0.19	-0.55	0.66	0.18
0.45	-0.02	-0.03	-0.05	-0.08	-0.13	-0.22	-0.67	0.52
0.48	-0.01	-0.03	-0.05	-0.07	-0.10	-0.14	-0.24	-0.82

table 9.2.b: eigenvectors Spearman hierarchy

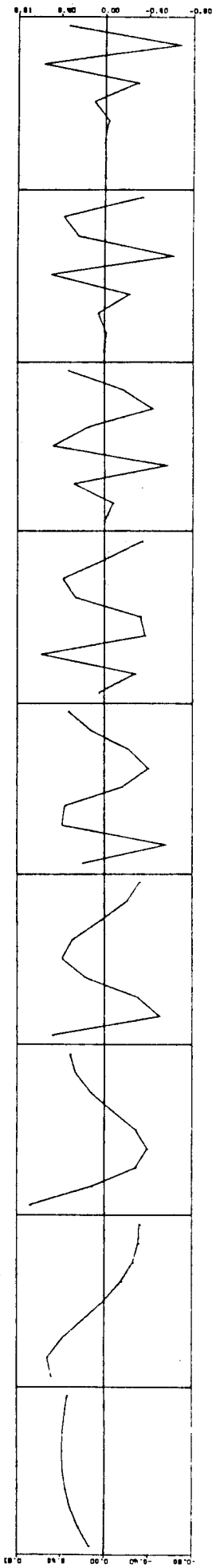


Figure 9.1 Eigenvectors of a Guttman correlation matrix

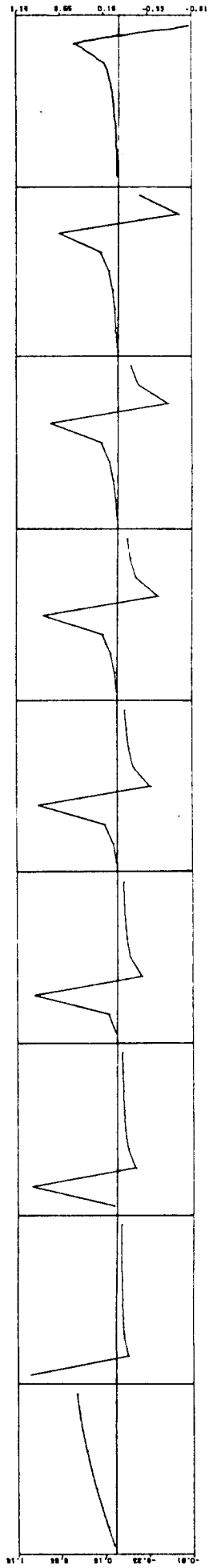


Figure 9.2 Eigenvectors of a Spearman correlation matrix

a correlation matrix of the form $R = \mu\mu' + \Delta^2$, with Δ^2 diagonal. If δ_j^2 is the j-th diagonal element of Δ^2, then $\delta_j^2 = 1 - \mu_j^2$. Observe also that if variables j and ℓ have the same parameters α and β, and thus also the same μ, then $\phi_{j\ell} = \mu^2$, which shows that $\mu_j^2 = 1 - \delta_j^2$ can be interpreted as the reliability of variable j.

At this point the reader may lose his patience, because it is somewhat illogical to model probabilities by a linear model. The $p_j(x)$ must be between zero and one, and linear functions are unbounded on the real line, and thus the variation of $\underline{x}$ is restricted by the values of all the α_j and β_j. This seems undesirable, and although the Spearman hierarchy gives a simple correlation matrix it does not seem obvious at all that it is a useful gauge or even a powerful norm such as the Guttman scale. But consider the following argument.

In a general latent trait model with item characteristic curves $p_j(x)$ we represent the latent variable $\underline{x}$ in the form $\underline{x} = \tau + \varepsilon\underline{z}$, with $AVE(\underline{z}) = 0$ and $VAR(\underline{z}) = 1$. We define $\alpha_j \triangleq p_j'(\tau)$, assuming that p_j is differentiable at τ, and $\beta_j \triangleq p_j(\tau)$. Moreover μ_j is defined in terms of α_j and β_j as in the Spearman hierarchy. What happens if ε is small, i.e. if $\underline{x}$ is distributed closely around τ. It is clear that continuity of p_j at τ already implies that $\phi_{j\ell} \to 0$ if $\varepsilon \to 0$. Assuming that p_j is twice continuously differentiable at τ, however, makes it possible to show that

$$\lim_{\varepsilon \to 0} \phi_{j\ell}/\varepsilon^2 = \mu_j\mu_\ell, \text{ or}$$

$$\phi_{j\ell} = \varepsilon^2\mu_j\mu_\ell + o(\varepsilon^2).$$

Thus for small dispersion on the latent continuum we find a correlation matrix (with small elements) which approximates the correlation matrix of a Spearman hierarchy.

We have seen that in the perfect scale the successive eigenvectors of the correlation matrix have more and more peaks and zeroes. The Spearman hierarchy is very different from this. Suppose the diagonal elements in Δ^2 are ordered as $\delta_1^2 \geq \ldots \geq \delta_m^2$. The eigenvalues of $R = \mu\mu' + \Delta^2$ then satisfy

$$\lambda_1 \leq \delta_1^2 + \mu'\mu,$$

$$\lambda_1 \geq \delta_1^2 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq \delta_m^2,$$

$$\lambda_m \leq \delta_m^2 + \mu'\mu.$$

These results are proved in many places, a convenient recent reference is Bunch, Nielsen, and Sorensen (1978). Thompson (1976) is also particulary elegant. Suppose y_j is the eigenvector corresponding with λ_j, and norm y_j in such a way that $y_j'\mu = 1$ (thus we do not consider those eigenvectors with $y_j'\mu = 0$, which

causes no real loss of generality). Then

$$y_{\ell j} = \frac{\mu_\ell}{\lambda_j - \delta_\ell^2}.$$

Plotting the elements of the eigenvectors against j gives interesting plots. For t first, largest eigenvalue the plot is monotone and non-negative (if $\mu \geqslant 0$). For th other eigenvectors there is only one sign-change, the plot jumps from high positiv down to low negative. The location of the jump shifts one place for each succesive eigenvector, in the interval in which there is no jump the polygonal functions are increasing. Table 9.2 contains Spearman hierarchy illustrations. Table 9.2.a has μ in its first row, R in its next nine rows, and λ, the eigenvalues of R, in its las row. Table 9.2.b has the eigenvectors of R, which are plotted in figure 9.2.

The pattern of eigenvector plots is again clearly very characteristic, and very different from the pattern of the Guttman scale. The patterns with the peaks has already been discussed informally in some older factor analysis literature, mainly in papers by Burt and Thomson (for example, Thomson, 1934). In the binary data context we again do not use the hierarchy as a realistic model, but as a normative gauge. Guttman shows us what happens if things are perfect, Spearman shows us what happens if things are as bad as possible. Again it is perfectly trivial to remark that real data are usually in between the two.

9.3.4 The latent distance model

There have been a number of attempts to make the Guttman scale more realistic, usually by relaxing the zero-one property which makes the Guttman model algebraic rather than probabilistic. These attempts have not been very successful, mainly because making the model more realistic means that it becomes less successful as a norm. There are many adjustable parameters, and there is a certain risk that we descend completely from the a priori level of norms, gauges, models, and theories to the empirical level of fitting 'models' with an unlimited number of degrees of freedom. This is what happened in factor analysis, when the Spearman hierarchy was replaced by Thurstone's multiple factor analysis. Although many people realized this at the time, and Guttman has emphasized it again and again in the fifties, this actually meant that all theory was banished from factor analysis, and that factor analysis was no longer psychology but statistics. Because we are primarily interested in gauges in this chapter, and not in description of reality, we do not discuss the latent polynomial models or the latent class models of Lazarsfeld or the multidimensional models of Coombs and Kao. They generalize the Guttman scale in the direction of multiple factor analysis. We refer to Coombs (1964, chapters 10, 11, 12), to Lazarsfeld and Henry (1968), and to McDonald (1967)

for some interesting discussion of these models.

In this section we discuss the simplest modification of the Guttman scale, also due to Lazarsfeld (1950). It is called the latent distance model, and it defines

$$p_j(x) = \begin{cases} \pi_j^- \text{ if } x < \theta_j, \\ \pi_j^+ \text{ if } x \geq \theta_j, \end{cases}$$

with, of course, $0 \leq \pi_j^- \leq \pi_j^+ \leq 1$. Thus $p_j(x)$ is still a step function, but it does not step from zero to one anymore. Also define

$$\kappa_j \triangleq \text{prob}(\underline{x} < \theta_j),$$
$$\varepsilon_j \triangleq \pi_j^+ - \pi_j^-.$$

After some computation it follows that

$$\pi_{j\ell} - \pi_j\pi_\ell = \kappa_j(1 - \kappa_\ell)\varepsilon_j\varepsilon_\ell, \text{ if } \theta_j \leq \theta_\ell,$$
$$\pi_{j\ell} - \pi_j\pi_\ell = \kappa_\ell(1 - \kappa_j)\varepsilon_j\varepsilon_\ell, \text{ if } \theta_j \geq \theta_\ell.$$

Thus if

$$\xi_j \triangleq \kappa_j\varepsilon_j / \{\pi_j(1 - \pi_j)\}^{\frac{1}{2}},$$
$$\zeta_j \triangleq (1 - \kappa_j)\varepsilon_j / \{\pi_j(1 - \pi_j)\}^{\frac{1}{2}},$$

then $\phi_{j\ell} = \xi_j\zeta_\ell$ if $\theta_j \leq \theta_\ell$, $\phi_{j\ell} = \zeta_j\xi_\ell$ if $\theta_j \geq \theta_\ell$, and $\phi_{jj} = 1$ for all j. This means that we can write $R = R_0 + \Delta^2$, with Δ^2 diagonal and nonnegative and with R_0 a <u>one-pair matrix</u> in the sense of Gantmacher and Krein (1936). One-pair matrices have the property that their inverse is tridiagonal, in this case symmetric tridiagonal, and that their eigenvectors have the oscillation properties we discussed in connection with the perfect scale. Thus after fitting the 'uniqueness' Δ^2 we are back in the perfect scale situation. Or, to put it differently, for the latent distance model the correlation matrix is a quasi-simplex. If we know the order of the θ_j then we can fit the model by using the logarithmic transformation of the correlations (or the covariances) again, eigenvector properties of R itself are not simple in general, they are of course simple if $\Delta^2 = \delta^2 I$.

9.3.5 <u>The Rasch model</u>

Another probabilistic generalization of the Guttman scale is the Rasch model. It has some very interesting mathematical properties which make it quite successfull as a gauge, although we shall see that the correlational properties are not very simple. In the Rasch model (Rasch, 1960, 1961, 1966, Fischer, 1974) the individuals are distributed on the positive real axes, which defines the latent trait, and

$$p_j(x) = x / (x + \theta_j).$$

This looks simple, because it is a rational function it is fairly easy to integrate, but the most important simplification is the following. It is possible to prove that

$$\pi_{j\ell} = \frac{\pi_j\theta_j - \pi_\ell\theta_\ell}{\theta_j - \theta_\ell},$$

and this relation is true no matter what the distribution of individuals on the latent continuum is. The proof is based on the algebraic identity

$$(1 - p_j(x))(1 - p_\ell(x)) = \frac{\theta_\ell(1 - p_j(x)) - \theta_j(1 - p_\ell(x))}{\theta_\ell - \theta_j},$$

which assumes, of course, that $\theta_j \neq \theta_\ell$. If we integrate both sides of the identity and collect terms we find the desired result. The most important consequence of the result is that

$$\frac{\pi_\ell - \pi_{j\ell}}{\pi_j - \pi_{j\ell}} = \frac{\theta_j}{\theta_\ell},$$

which makes it possible to compute the θ_j without any assumption about the distribution of $\underline{x}$. Rasch calls this specific objectivity, and shows that it is characteristic for his model. It certainly is a very nice property, analogous to measurement situations in the physical sciences. and quite uncommon in psychometri It is this property which makes the Rasch model a suitable gauge.

Some additional computation gives the following formula for the phi-coefficient

$$\phi_{j\ell} = \frac{1}{\theta_j - \theta_\ell}\left\{\theta_j\left[\frac{\nu_j}{\nu_\ell}\right] - \theta_\ell\left[\frac{\nu_\ell}{\nu_j}\right]\right\}.$$

This implies that the Loevinger-Mokken coefficient $\omega_{j\ell}$, in the case that $\theta_j \geq \theta_\ell$, is given by

$$\omega_{j\ell} = \frac{1}{\theta_j - \theta_\ell}\left(\theta_j - \theta_\ell\frac{\nu_\ell^2}{\nu_j^2}\right).$$

This formula illustrates our remark in 9.3.2 that perfectly holomorphic items, in this case Rasch items, can have low $\omega_{j\ell}$.

It is possible to substitute some distributions on the latent continuum into these formulas and to compute the corresponding correlation matrices. This is done, for example, in De Leeuw (1973, section 3.20) and in Gifi (1980, section 2.2.6). We shall not perform these computations here, but we mention an additional theoretical result of some interest. Gantmacher and Krein (1937, p 457) show that the Cauchy determinant formula implies that the function $1 / (x_i + y_j)$ is totally positive. According to the general composition formula of Karlin (1968, p 16-18) this means that the matrix with off-diagonal elements $\pi_{j\ell}$ and

diagonal elements $AVE(p_j^2(\underline{x}))$ is totally positive. This implies that the correlation matrix R of the Rasch model is of the form $R = R_0 + (\Delta^2 - \mu\mu')$, with R_0 totally positive, with Δ^2 diagonal, and with μ nonnegative. This representation is useful, because totally positive matrices have the same eigenvector properties as the perfect scale, and thus the Rasch correlation matrix can be thought of as the difference between a perfect scale type correlation matrix and a Spearman type correlation matrix.

9.4 Nonmonotonic latent trait models

Models with increasing $p_j(x)$ are inspired by test theory. If individual A is 'better' than individual B, then the probability that A gives a correct response is larger than the probability that B gives a correct response. For many binary data matrices, however, the notion of correctness and being better do not apply. In these cases individuals give positive responses to questions with which they agree, and the idea is that in many situations it is more realistic to suppose that the $p_j(x)$ are unimodal. Plants, for example, need a certain amount of moisture. If the soil is too moist, then certain types of plants will not occur, if the soil is too dry they will not occur either. There is a certain range which they can handle, different types of plants have different ranges. There are some plants, however, with monotone moisture behaviour. Plants who grow in water are a simple example. But it is unwise of course to use land and water plants in the same analysis, also because most models do not take into account that moisture is bounded both from above and from below. A similar situation occurs in archeology. Some objects occur in graves during a particular time interval, and not outside this interval. It is possible that there are objects which occur more and more as time progresses, but this is highly unlikely, and we never know if eventually the rate of occurrence will decrease again. Thus unimodal models are more natural in this case too, and monotone characteristic curves are best thought of as having a mode somewhere near infinity. In psychology unimodal models seem more natural in many similarity and preference contexts. To use a familiar example: people who like cold beverages usually do not want them to be completely frozen, people who like their coffee hot do not want it to be actually boiling.

The model corresponding with the perfect scale in unimodal situations is the Coombs scale. Here $p_j(x) = 1$ if $\theta_j \leq x \leq \theta_j + \varepsilon_j$, and $p_j(x) = 0$ otherwise. There is no special reason why the Coombs scale is less compelling or less natural than th Guttman scale, but it is considerable less important in practice for technical reasons. In fact if items vary in both their θ_j and their ε_j, then there is no natural way to order items completely, although there are many ways to order them partially. The restricted Coombs scale solves this problem by letting

$p_j(x) = 1$ if and only if $\theta_j \leq x \leq \theta_j + \varepsilon$, with ε the same for all j. This certainly solves the problem because now we can simply order the θ_j, and apply correspondence analysis to the incomplete indicator matrix to recover the order (an observation due to Mosteller in 1949, cf Torgerson, 1958, p 338). This result is also a consequ of general results in the next section. It indicates that we should not analyse R in the case of a Coombs scale, because analysis of R corresponds with analysis of the complete indicator matrix.

The Coombs scale is deterministic in the same sense as the Guttman scale. There ar some obvious probabilistic generalizations possible based on standard unimodal densities from the exponential family. Consider for example a latent variable distributed on the real line, with

$$p_j(x) = K_j \exp(-\tfrac{1}{2}\eta_j(x - \theta_j)^2).$$

This combines nicely with a normal distribution on the latent continuum, it also gives totally positive $\pi_{j\ell}$ in the same way as the Rasch model. It is not, however, a satisfactory gauge, because we can think of hundreds of models like this, choose convenient 'conjugate' distributions on the latent continuum, and are left with an appalling degree of arbitrariness, as in Baysian statistics. It seems quite likely, however, that a gauge similar to the Rasch model can be derived for nonmonotone items as well.

9.5 Order analysis of binary matrices

In 9.3.2 we have shown that applying correspondence analysis to a perfect scale gives category quantifications which are in the correct order. This argument appli to complete indicator matrices of order $n \times (2m)$, where in fact n can be infinite. In the previous section we have stated, but not shown, that applying correspondenc analysis to the restricted Coombs scale also produces qunatifications in the corre order, provided we code them in an incomplete indicator matrix, with order $n \times m$. In this section we provide the tools to prove this and similar statements. These tools can be used for order analysis of binary matrices, in which we order the row and columns of a binary matrix in the order of the left and right dominant eigenve from a correspondence analysis. In case of the perfect scale and the incompletely coded Coombs scale this reordering tends to produce a diagonal pattern of ones along the diagonal of the table.

Now suppose K is a pointed, closed, convex cone which is solid (has nonempty inter cf appendix C. We use the theory of K-positive matrices, a recent review is in Berman (1973 , p 51-54), who also gives the necessary references. First define the cone of $n \times n$ matrices

$$\Pi(K) = \{A \mid AK \subseteq K\},$$

where we assume that K is a cone in $\mathbf{R}^n$. Thus if $A \in \Pi(K)$, then $Ax \in K$ for all $x \in$

The following results are very useful. If $A \in \Pi(K)$ then $\rho(A)$, the spectral radius of A, is an eigenvalue of A, there is a nonzero $x \in K$ such that $Ax = \rho(A)x$ and a nonzero $y \in K^o$ (the polar cone, cf appendix C) such that $A'y = \rho(A)y$. If we assume K-positivity, i.e. $A(K - \{0\})$ is a subset of the interior of K, than it even follows that $\rho(A)$ is the unique largest eigenvalue, that x is in the interior of K, and that x is the only eigenvector in K. If P and Q are frames for K and K^o, i.e.

$K = \{x \mid x = Pt \text{ for some } t \gtrsim 0\}$,

$K^o = \{y \mid y = Qu \text{ for some } u \gtrsim 0\}$,

then $A \in \Pi(K)$ if ond only if $Q'AP \gtrsim 0$ (elementwise). Furthermore if $\rho(A)$ is an eigenvalue of A, and A is similar to a diagonal matrix, then $A \in \Pi(K)$ for some full (= pointed, solid, closed, convex) cone K. If $\rho(A)$ is the unique largest eigenvalue of A, then A maps some full cone into its interior.

We can now apply these results to the linear transformation used in correspondence analysis. If F is a nonnegative matrix, and D_c and D_r are the diagonal matrices of column totals and row totals, then we are interested in the operator $D_r^{-1}FD_c^{-1}F'$. Very similar results are of course possible for $D_c^{-1}F'D_r^{-1}F$. In the first place matrices of this form are always semi-simple (similar to a diagonal matrix) and the spectral radius is always an eigenvalue of A. The same thing is true if we remove the trivial solution and work with $D_r^{-1}(F - \frac{1}{N} D_r uu'D_c)D_c^{-1}(F - \frac{1}{N} D_r uu'D_c)' =$ $= D_r^{-1}FD_c^{-1}F' - \frac{1}{N} uu'D_r$. In this formula $N \triangleq u'D_c u = u'D_r u$. The general theory tells us that we can check whether a cone K is reproduced by this operator by using frames for K and K^o. We do this for the cone

$K = \{x \mid x_1 \lesssim \ldots \lesssim x_n\}$.

A frame for K consists of the columns of the matrix S, which has zeroes above the diagonal and ones below and on the diagonal, together with the vector -u. A frame for K^o consists of the n - 1 vectors $e_{i+1} - e_i$. Since our correspondence analysis operator A, with the trivial solution removed, satisfies $Au = 0$, we must have $(e_{i+1} - e_i)'AS \gtrsim 0$, which we can write as

$$\sum_{k=1}^{\ell} a_{i+1,k} \lesssim \sum_{k=1}^{\ell} a_{i,k},$$

for all $\ell=1,\ldots,n$. This condition is fairly easy to check in many cases. It is sometimes more convenient to work with two cones K_X and K_Y, and the operators $D_r^{-1}F$ which maps $\mathbf{R}^m$ into $\mathbf{R}^n$, and $D_c^{-1}F'$ which maps $\mathbf{R}^n$ into $\mathbf{R}^m$. We want to find out if $D_r^{-1}FK_Y$ is a subset of K_X and $D_c^{-1}F'K_X$ is a subset of K_Y. Clearly this implies that $D_r^{-1}FD_c^{-1}F' \in \Pi(K_X)$ as well as $D_c^{-1}F'D_r^{-1}F \in \Pi(K_Y)$. This is actually how we proceeded with the perfect scale. The method can also be easily adapted to infinite dimensional

spaces, as we have illustrated in section 9.3.2, where $\underline{x}$ is a random variable with finite variance, and not necessarily an n-element vector.

9.6 Dichotomized multinormal distributions

Another popular model for binary data, not based on latent trait theory, is the dichotomized multinormal. We have seen in chapter 1 that the popularity of this model is due to Pearson's philosophy that continuous variation is the norm, that correlation should replace causation, and that measures of association should be approximations of the product moment correlation of the underlying normal process. Although the fact that these techniques have been used a great deal makes them familiar and even natural, we must ask the reader to reflect a moment on the assumption that archeological findings, answers to multiple choice items, or occurrences of vegation in certain areas, result from dichotomizing a multinormal process. It seems to us that the assumption is not very natural in most situations and very silly in some. It is also not clear in what sense this model is very special, what there is that makes it into an interesting gauge. For completeness, however, we list some of the consequences of the assumptions.

Thus we suppose that $\underline{z}_1,\ldots,\underline{z}_m$ are joint multivariate normal, zero means, unit variances, with correlations $\rho_{j\ell}$, and we suppose that the observed $\underline{h}_j$ are formed by the rule $\underline{h}_j = 0$ if $\underline{z}_j < \theta_j$ and $\underline{h}_j = 1$ if $\underline{z}_j \geq \theta_j$. In connection with principal components analysis and factor analysis this model was used for the first time by Lawley (1944), although he used the classical tetrachorical theory of Pearson in his derivations. We can use the Hermite-Chebyshev polynomials ψ_s to show that we can represent the correlations in the form

$$COR(\underline{h}_j,\underline{h}_\ell) = \sum_{s=1}^{\infty} \rho_{j\ell}^s \tau_{js}\tau_{\ell s},$$

with

$$\tau_{js} \triangleq (s!)^{-\frac{1}{2}}\phi(\theta_j)\psi_{s-1}(\theta_j)\{\Phi(\theta_j)(1 - \Phi(\theta_j))\}^{-\frac{1}{2}},$$

where ϕ is the density and Φ is the distribution function of the standard normal. Observe that

$$\Phi(\theta_j)(1 - \Phi(\theta_j)) = VAR(\underline{h}_j).$$

For s=1 and s=2 we find

$$\tau_{j1} = \phi(\theta_j)VAR^{-\frac{1}{2}}(\underline{h}_j),$$

$$\tau_{j2} = \tfrac{1}{2}\sqrt{2}\ \theta_j\phi(\theta_j)VAR^{-\frac{1}{2}}(\underline{h}_j).$$

Lawley supposes in addition that the items have a Spearman structure, i.e. $\rho_{j\ell} = $ for all $j \neq \ell$. Then

$$COR(\underline{h}_j,\underline{h}_\ell) = \alpha_j\alpha_\ell\tau_{j1}\tau_{\ell 1} + \alpha_j^2\alpha_\ell^2\tau_{j2}\tau_{\ell 2} + O(a_j^3a_\ell^3).$$

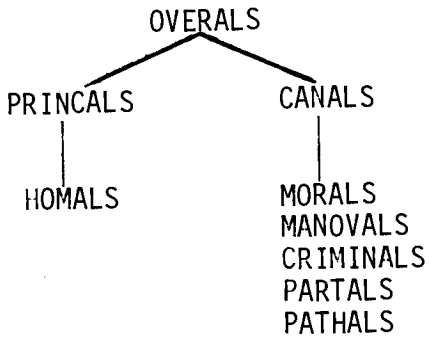
Because $\tau_{j2} > 0$ for a difficult test and $\tau_{j2} < 0$ for an easy test, the second factor in pricipal components analysis depends on the difficulty of the test. A plot of the first two factors gives, as with the continuous normal distribution and the Guttman scale, a quadratic horseshoe. We also remark that our gauges give important contributions to the problem of difficulty factors which has bothered factor analysts for a very long time. The problem is discussed very nicely in McDonald (1967), who also has the necessary references.

Alternative techniques for presumably multinormally generated binary data are usually based on tetrachoric correlations. Divgi (1979) reviews and develops efficient computational methods. Estimates of the parameters of the one-factor model are discussed by Lawley (1944), more efficients ones by Bock and Lieberman (1970). The corresponding multiple factor models have been developed by Christofferson (1975, 1977) and Muthén (1978). Compare also the very recent Bartholemew (1981). If the multinormal is categorized in more than two classes we need polychoric correlation. An interesting historical overview, and a new estimation method, are in Lancaster and Hamdan (1964). The definite proof that among the most important Swedish export products are partial derivatives of likelihood functions is given by Olsson (1979).

10:The use of restrictions

10.1: Introduction

We have seen in the previous chapters that HOMALS is a special case of PRINCALS, and PRINCALS is a special case of OVERALS. If all sets contain a single variable then OVERALS becomes PRINCALS, if all variables are multiple nominal then PRINCALS becomes HOMALS. The relationship with ANACOR is more complicated. If we are dealing with two multiple nominal variables then HOMALS becomes ANACOR, if we apply singular value decomposition to indicator matrices, complete or incomplete, or to the matrix C with bivariate cross products, then ANACOR becomes HOMALS. It is consequently better to think of ANACOR as a different technique to minimize similar loss functions, which does not fit naturally into the order HOMALS - PRINCALS - OVERALS. The techniques treated in chapter 7 and 8 are again special cases of OVERALS in which there are only two sets of variables, with often a special structure of one of the two sets. Thus the implied partial order of generality is:



This partial order is consistent with classification of MVA techniques in terms of partitioning of the variables into sets, it is also consistent with the classification into join techniques and meet techniques.

On the other hand there is the purely technical fact that both join and meet techniques can be fitted by using σ_M, meet-loss, and the fact that interactive variables can be used to introduce sets. These two points are clearly illustrated in section 5.2.7 and section 6.2.3. As a consequence we formulate the most general problem we are interested in in this chapter as minimization of

$$\sigma(X,Y) = \frac{1}{m}\sum_{j=1}^{m} \mathrm{tr}(X - G_jY_j)'M_j(X - G_jY_j).$$

Because of our use of the M_j we can suppose without loss of generality that the G_j are complete indicator matrices. If $M_j = I$ for all j then we are using option II or III for missing data. Otherwise M_j is a binary diagonal matrix, we use opt I for missing data, and the corresponding incomplete indicator matrix is M_jG_j. We can write

$$\sigma(X,Y) = \frac{1}{m} \text{tr } X'(M_* - mI)X + \frac{1}{m} \sum_{j=1}^{m} \text{SSQ}(X - M_j G_j Y_j),$$

which shows the difference between the missing data treatments very clearly. The normalization conditions in the general problem will not be defined explicitly, because we allow for general restrictions on X and/or Y of which the usual normalization conditions are just a special case. In 6.2.3 we have seen how we can use design matrices S_j to introduce the structure associated with sets of variables.

There is an interesting property of constraints which we have already discussed in 6.2.2, which simplifies the problem and makes the interpretation somewhat easier. Suppose we write the general constraints as $X \in C_X$ and $Y \in C_Y$, without necessarily implying that C_X or C_Y are cones. Suppose we assume that $Y \in C_Y$ implies that $YU \in C_Y$ for all square nonsingular U. Suppose moreover that $X \in C_X$ implies that $X'M_*X = I$. Then the same argument as in 6.2.2 shows that our problem is equivalent to maximizing SSQ(X'GY) over $X \in C_X$ and $Y \in C_Y$ that satisfy in addition $Y'DY = I$. Here G is the indicator supermatrix consisting of the m matrices $M_j G_j$ next to each other, D is the diagonal matrix of column totals of G, and M_* is the sum of the M_j, which is the diagonal matrix of row totals of G. In the same way if $X \in C_X$ implies that $XT \in C_X$ for all square nonsingular T and $Y \in C_Y$ implies that $Y'DY = I$, then our problem is equivalent to maximizing SSQ(X'GY) over $X \in C_X$ with $X'M_*X = I$ and $Y \in C_Y$. On the other hand if $X \in C_X$ implies that $X'M_*X = I$ and $Y \in C_Y$ implies that $Y'DY = I$, then the problem we are solving is equivalent to maximizing tr(X'GY) over $X \in C_X$ and $Y \in C_Y$. This is the somewhat more abstract version of the theorem familiar from chapters 3 and 4 that we can either normalize over X or normalize over Y and find the same solution (up to a nonsingular transformation). Normalizing both over X and over Y gives the same solution in simple situations such as the ones discussed in chapters 3 and 4, but may give different solutions in more complicated situations. The interesting thing is that choice of normalization remains irrelevant in general constrained situations, provided the constraint sets are closed under multiplication on the right with a nonsingular matrix. This condition is true for constraints of the form $Y_j = S_j Y_j$ we use to create sets of variables, it is also true under constraints like $Y_j = y_j a_j'$ we use for single variables. If all variables are single then the matrix X'GY transforms to X'QA, where column j of Q is given by $q_j = M_j G_j y_j$. Moreover $Y'DY = A'BA$, with $B = \text{diag}(Q'Q)$.

This summarizes most of our theory so far, and shows that it is comparatively easy to build in several types of constraints on X and Y. This does not mean that the computer programs we have discuss can routinely handle some of these constraints. It does mean, at least in some cases, that the constraints indicate that preprocessing of the data is all that is needed before we apply one of our programs.

10.2 Equality constraints

We illustrate the use of equality constraints in some simple special cases. Our analysis shows clearly how more general cases should be handled. We require $X = G_0 \underset{\sim}{X}$ for some indicator matrix G_0. This means that some of the object scores are required to be equal, for example because they are replications, or because we want all women to have one score and all men to have another. We also require that $u'M_* X = u'M_* G_0 \underset{\sim}{X} = 0$ and $X'M_* X = \underset{\sim}{X}'G_0'M_* G_0 \underset{\sim}{X} = I$. Observe that indicator matrices G have the pleasant property that G'DG is diagonal for any diagonal matrix D. Thus $G_0'M_* G_0$ is diagonal too. We also constrain the Y_j by the requirement that $Y_1 = \ldots = Y_m$ which only makes sense, of course, if all variables have an equal number of categories. This constraint seems appropriate, although certainly not strictly necessary, if the variables are Likert items, or rating scales, or equal appearing interval judgments, or t-sorts, of successive category judgments, and so on. All classical psychophysical and attitude scaling methods are based on the assumption that the category quantifications are the same for all variables, and most of the methods assume in addition that the quantifications are also known a priori.

The results of the previous section apply, and show that we compute $\underset{\sim}{X}$ and Y from a correspondence analysis of the table

$$G_0' \sum_{j=1}^{m} M_j G_j.$$

This is a nice and simple result. It easily generalizes to the situation where we only want some of the Y_j to be equal. We then must replace the indicator matrices for these variables in the indicator supermatrix by their sum, and apply ANACOR to the supermatrix consisting of sums of indicator matrices.

In the method of successive intervals (Guilford, 1954, chapter 10, Torgerson, 1958, chapter 10) we also make the assumption that all Y_j are equal, but contrary to most other methods we do not assume them to be known a priori. The method scales category boundaries and the objects simultaneously, on one-dimensional scales, so we may as well assume that p = 1 too. The idea is that stimuli are compared with boundaries and that stimulus i is less than boundary j with probability $\Phi(b_j - s_i)$, with Φ the cumulative normal. Thus in the stimuli by categories table the appropriate probabilistic model is an $n \times (m + 1)$ table, with for each stimulus a row of conditional probabilities $\Phi(b_j - s_i) - \Phi(b_{j-1} - s_i)$, with $b_0 = -\infty$, and $b_{m+1} = +\infty$. It now follows directly from the results of 9.2 that correspondence analysis recovers the correct order of the s_i and b_j if the model is exactly true, and it does that even if Φ is replaced by any other cumulative distribution function. As in the case of the Guttman scale we cannot expect to recover more than the order without making explicit assumptions about Φ. This is one of the basic messages of Coombs (1964). We can only find metric representations if we make essentially arbitrary assumptions. In our programs the arbitrariness is hidden in a special

choice of the optimality criterion.

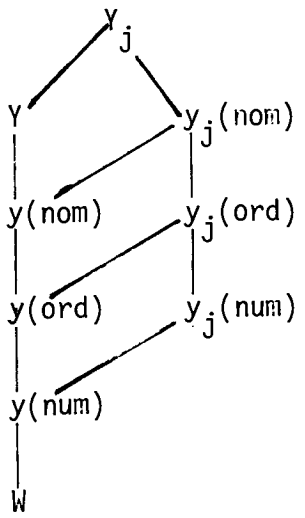
Another application of the use of inequality constraints is in the analysis of preference rankings. If n individuals rank m objects, then the data consist of n permutation matrices. Now permutation matrices are very special indicator matrices. Thus HOMALS applies, but we must remember that the ranked objects in the experiment are the individuals for the HOMALS program, and that the individuals in the experiment are the variables in the HOMALS program. We are dealing with a 'reversed indicator matrix', also discussed in 3.12. Because the G_j are permutation matrices they satisfy $G_j'G_j = G_jG_j' = I$, and this implies that all nonzero HOMALS eigenvalues are equal to one. We merely choose for x any centered m-vector, and the permutations $y_j = G_j'x$ then give a perfectly homogeneous solution. It is consequently necessary to impose some sort of constraints. We have already discussed one way to deal with the problem in chapter 5. If we suppose that all variables are single ordinal, then we can use PRINCALS and recover the familiar nonmetric approach to preference rankings using the vector model. Single nominal will not do, the same argument that showed triviality of HOMALS in this context also shows that a trivial perfect single nominal solution exists in one dimension. Single numerical is possible, of course, we shall discuss it shortly.

Another constraint is possible, without going single. We require that all Y_j are equal, the analysis now amounts to performing correspondence analysis on the sum of the permutation matrices. This matrix is known in psychophysics as the rank-order frequency matrix (Guilford, 1954, chapter 8). It can be used in combination with the normalized rank method, which uses a fixed nonlinear scale for the rank numbers and then computes scale values of the actors by averaging (a first iteration of reciprocal averaging). It can also be used in combination with Thurstonian methods, and again we can show that correspondence analysis recovers the correct orders if the simplest (case V) models are true.

Equality constraints can also be used in combination with single variables. Thus we can require $Y_j = ya_j'$, for example, in preference rankings situations, where the vector y is the same for all j. In addition we can impose ordinal or linear constraints on y. If we require that y is linear with the rank numbers, for example, then the technique becomes equivalent to the singular value decomposition of the $n \times m$ matrix of row-centered rank numbers, a technique previously invented by Slater (1960), Carroll and Chang (1964), and Benzécri (1965). This technique has the property that the first principal component for the objects is often very similar to the average rank numbers over individuals. If we require in addition, however, that weights a_j sum to zero over all j (individuals in the ranking experiment), then we must double center the matrix of rank numbers before computing the singular value decomposition. This double centering can be interpreted in terms of the squared distance model of preferential choice (Ross and Cliff, 1964). If

y is not completely known, but only ordinally constrained or not constrained at al then the situation becomes more complicated, but it is straightforward to adapt th alternating least squares algorithm of PRINCALS to this case.

If we summarize the possibilities for the analysis of preference rankings we find the following partial order.



The lower we are in the order, the more loss we incur. Methods with index j do not require equality, methods without an index require equality. Thus Y_j and y_j(nom) have trivial perfect solutions, method y(num) is Carroll and Chang's MDPREF or Benzécri's 'analyse des préférences' if we actually use the rank numbers. The W stands for Kendall's coefficient of concordance (Kendall, 1962, chapter 6), which does not only require y(num) but also $a_j = 1$ for all j. Of course more or less the same thing applies to other situations in which the number of categories for all variables is equal. In these other situtations, however, it is not necessarily true that Y_j and y_j(nom) are trivial. In a T-sort technique or for partial rankings, for example, the G_j are not square, and triviality does not apply.

The data in our first example are Guilford's spot pattern data (Guilford, 1954, p 203). He used 100 different cards with spot patterns. There were 25 groups of four cards, patterns in each group having the same number of spots. One single observer sorted the deck in nine ordered piles, with the beautiful old-fashioned instruction in mind, that he had to attempt to keep interpile distances psychologically equal. There were ten replications of this same experiment. The data can be collected into ten indicator matrices of order 100 × 9, and then be analyzed by HOMALS. These matrices are not reported by Guilford, however, he reports a single 23 × 9 matrix of groups against piles. It is not clear to us where the other two groups went, it is clear that aggregating data like this can be fit into our framework if we suppose that HOMALS must give equal scores to

cards in the same groups and equal category quantifications for all ten replications. The homogeneity approach has a natural interpretation in psychophysical contexts, in which we very commonly suppose that there 'is' a one-dimensional scale and that the different variables are merely replications.

Guilford's data matrix is reproduced in table 10.2.1. The first singular value of an ANACOR analysis on this table was .93, the remaining singular values conformed closely to $\lambda_s = (.93)^s$, which shows that we only have to pay attention to a single dimension. Table 10.2.2 contains optimal quantifications of spot patterns and of piles (intervals). Both transformations are plotted in figures 10.1 and 10.2. It is clear from the transformation of the intervals that correspondence analysis does not follow the instruction to keep interpile distances equal, the intervals in the middle are larger, near the endpoints the distances are smaller. The transformation of the stimuli is fairly linear, deviations from linearity are in the direction of concavity. In this sense correspondence analysis, which produces the 'best' scale in a specific least squares sense, confirms Fechner's law, but this is obviously a very weak confirmation. It could very well be that a similar technique with a least absolute deviation loss function disconfirms Fechner's law.

S/R	1	2	3	4	5	6	7	8	9
15:	14	18	7	1					
16:	16	19	3	2					
17:	7	18	11	4					
19:	8	18	9	3	2				
20:	3	12	14	3	6	2			
22:	1	11	14	12	2				
24:		3	12	14	9	2			
26:		2	9	18	9	2			
28:			2	20	17	1			
30:				26	11	3			
32:			2	10	16	9	3		
35:				8	17	14	1		
37:				8	18	10	4		
40:				2	14	14	10		
43:					12	19	9		
46:				2	6	18	14		
49:					2	14	23	1	
53:						10	25	5	
56:						12	22	6	
60:						5	22	11	2
64:							14	20	6
69:							7	17	16
74:							6	20	14

Table 10.2.1 Guilford's Spot Pattern data

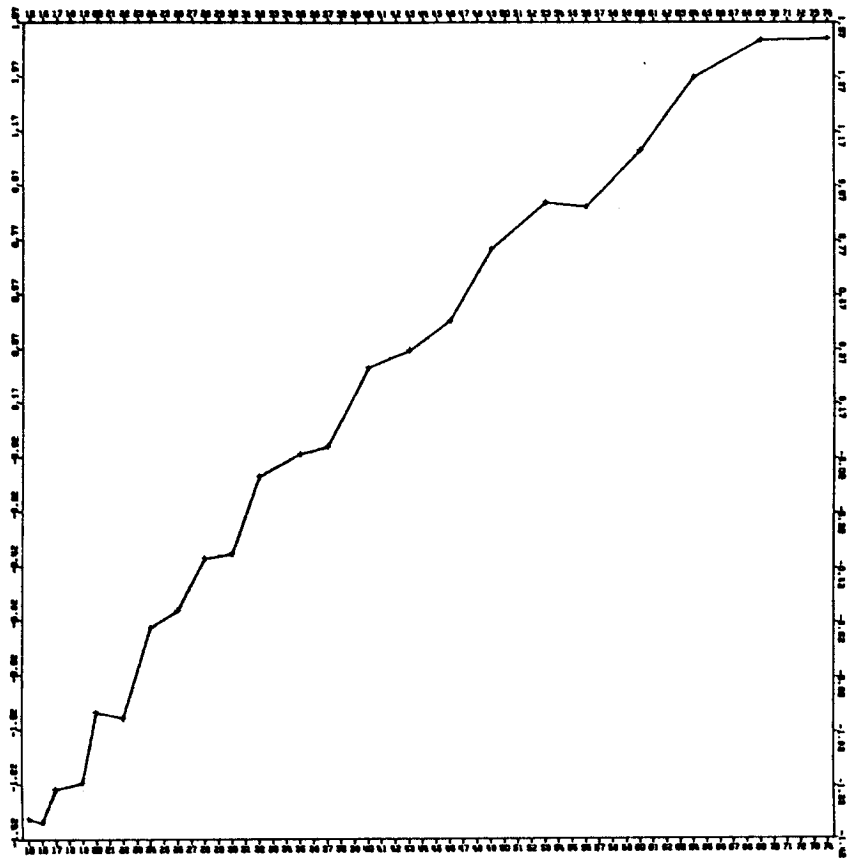


Figure 10.1 Guilford's spot patterns: transformations of the stimuli

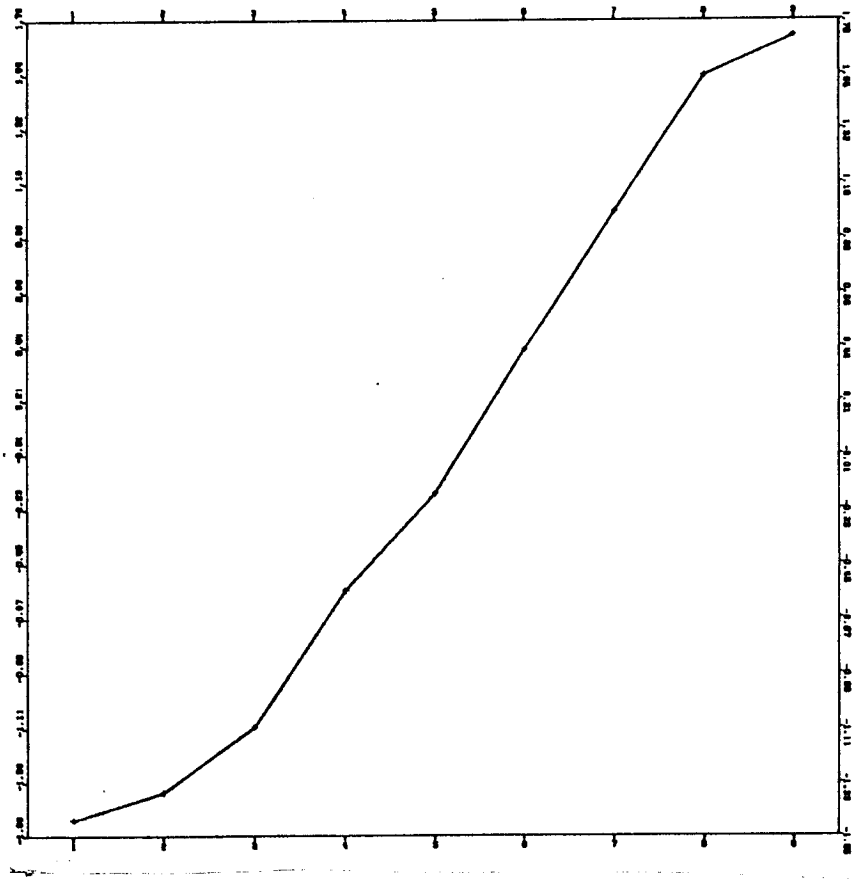


Figure 10.2 Guilford's spot patterns: transformations of the intervals

15	-1.35
16	-1.36
17	-1.24
19	-1.21
20	-0.96
22	-0.98
24	-0.65
26	-0.58
28	-0.39
30	-0.38
32	-0.09
35	-0.01
37	0.02
40	0.31
43	0.37
46	0.48
49	0.74
53	0.91
56	0.90
60	1.10
64	1.36
69	1.50
74	1.51

1	-1.49
2	-1.37
3	-1.11
4	-0.56
5	-0.17
6	0.42
7	0.98
8	1.53
9	1.69

Table 10.2.2
Guilford's spot pattern data
quantifications of cards and piles

	1	2	3	4	5	6	7	8	9	10
JEXP	8	5	2	3	4	5	5	3	1	3
JAPP	3	4	5	5	6	3	9	3	1	0
JPSP	5	6	3	3	7	9	3	2	0	1
MUBR	1	4	3	5	3	3	6	4	8	2
JCLP	5	2	2	5	3	1	2	4	8	7
JEDP	1	2	2	5	2	4	4	8	6	5
PMEK	4	2	5	3	3	3	3	3	5	8
HURE	1	5	1	0	2	6	4	8	8	4
BULL	6	9	12	7	1	2	0	1	0	1
HUDE	5	0	4	3	8	3	3	3	2	8

Table 10.2.3 Roskam's journal preference data Sum of permutation matrices

Our second example are the Roskam journal preference data already discussed in section 5.4.2. We first investigate the results with a multiple two-dimensional Y, restricted to be equal for all 39 individuals. This amounts to performing correspondence analysis on the sum of the 39 permutation matrices. This sum is given in table 10.2.3, the results of the correspondence analysis are in table 10.2.4, they are plotted in figures 10.3 and 10.4. The two-dimensional transformation of the rank numbers is horseshoe-like, with clusters (1,2,3,4), (5,6,7), and (8,9,10) along the horseshoe. These three clusters also make it possible to think of the solution in terms of a triangle. The same triangle can also be found in figure 10.4, where the Psychological Bulletin is one corner of the triangle (popular with many kinds of psychologists), the journals JEXP, JAPP, JPSP are another corner (popular with many, impopular with some), and the remaining journals are scattered around the third corner (impopular with many, popular with some). The solution can perhaps be characterized by the fact that it tries to find the one-dimensional scale and put it on a horseshoe. For these data this is not very well possible, and the reason is clear from a closer inspection of table 10.3. Most of the rows of this table are bi-modal, sometimes even U-shaped, except for BULL, and for JCLP and PMEK which are very flat and quite unpopular. In this particular case this technique is not vary satisfactory, individuals are not replications, stimuli are too 'interesting'.

A second analysis of the same example is between the nonmetric and the metric analysis already discussed in chapter 5. In the nonmetric analysis we require that $Y_j = y_j a_j'$, with the y_j in the appropriate order, in the metric analysis we require that $Y_j = ya_j'$ with y the centered rank numbers, i.e. a fixed vector. An interesting intermediate case is $Y_j = ya_j'$ with both y and a_j free, in our example we only did not require that y was in the appropriate order. A special purpose APL-program gave the transform of the rank numbers plotted in figure 10.5, the object scores in 10.6, and the loadings a_j in 10.7. The interesting feature if 10.5 is that the transformation of the rank numbers is very flat at the lower and also rather flat at the upper end. Thus the extreme opinions are flatted out in order to achieve a better fit. The dimensions in figure clearly have to do with specific-general and with hard-soft, if we compare the solution with figure 5.6 then the clinical cluster (HURE,JCLP) has merged with the developmental cluster (JEDP,HUDE). The hard cluster (PMEK,JEXP,MUBA) is still there, so is the general cluster (BULL,JAPP,JPSP), which has become somewhat more clear. In 10.7 we also see that the developmental people and the clinical people are not separated any more. On the whole this analysis seems less informative than the ones in chapter 5. Numerical results for this analysis are given in table 10.2.5.

JEXP	.17	.26
JAPP	.21	.35
JPSP	.29	.51
MUBR	-.19	.01
JCLP	-.27	-.38
JEDP	-.33	-.05
PMEK	-.12	-.30
HURE	-.39	.18
BULL	.67	-.53
HUDE	-.04	-.04
1	.28	-.06
2	.33	.04
3	.46	-.43
4	.17	-.25
5	.07	.36
6	.03	.48
7	-.09	.42
8	-.35	.07
9	-.57	-.25
10	-.33	-.37
eval	.17	.08

Table 10.2.4 Roskam's journal preference data correspondence analysis for multiple equal Y

JEXP	.38	-.02
JAPP	.11	.23
JPSP	-.19	.37
MUBR	.35	-.35
JCLP	-.49	-.05
JEDP	-.22	-.34
PMEK	.45	-.27
HURE	-.31	-.22
BULL	.18	.67
HUDE	-.26	-.04
1	.37	
2	.30	
3	.35	
4	.30	
5	.10	
6	-.11	
7	-.09	
8	-.26	
9	-.50	
10	-.46	

Table 10.2.5 Roskam's journal preference data. equal single nominal y.

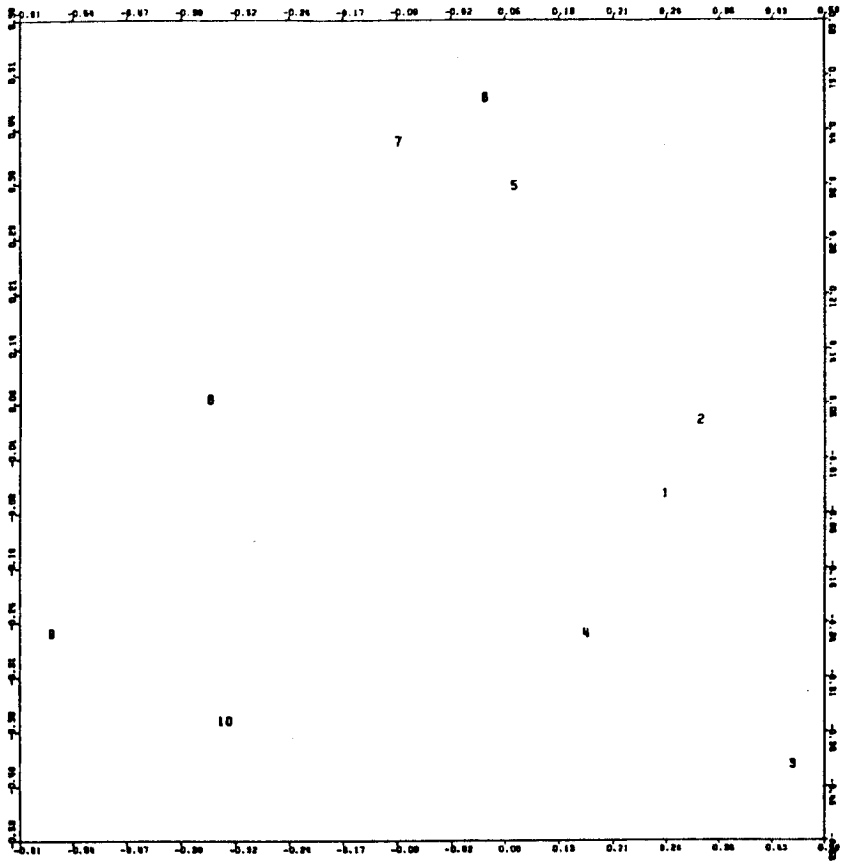


Figure 10.3 Roskam's journal preference data: category quantifications in the equal multiple analysis

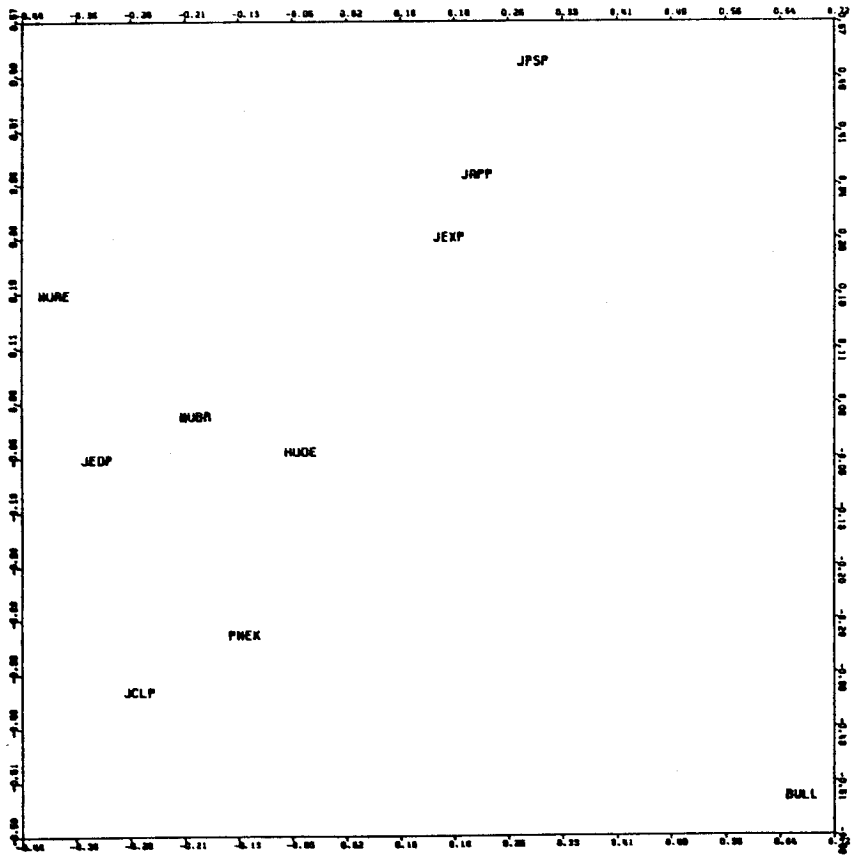


Figure 10.4 Roskam's journal preference data: object scores in the equal multiple analysis

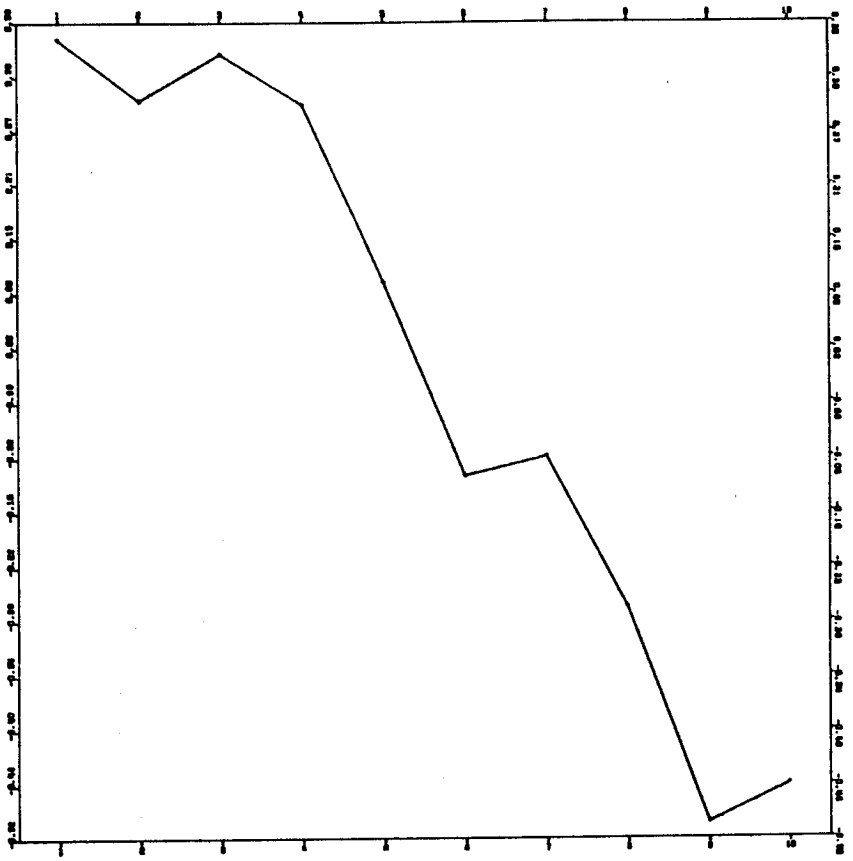


Figure 10.5 Roskam's journal preference data: transformation of the rank numbers in the single vector analysis

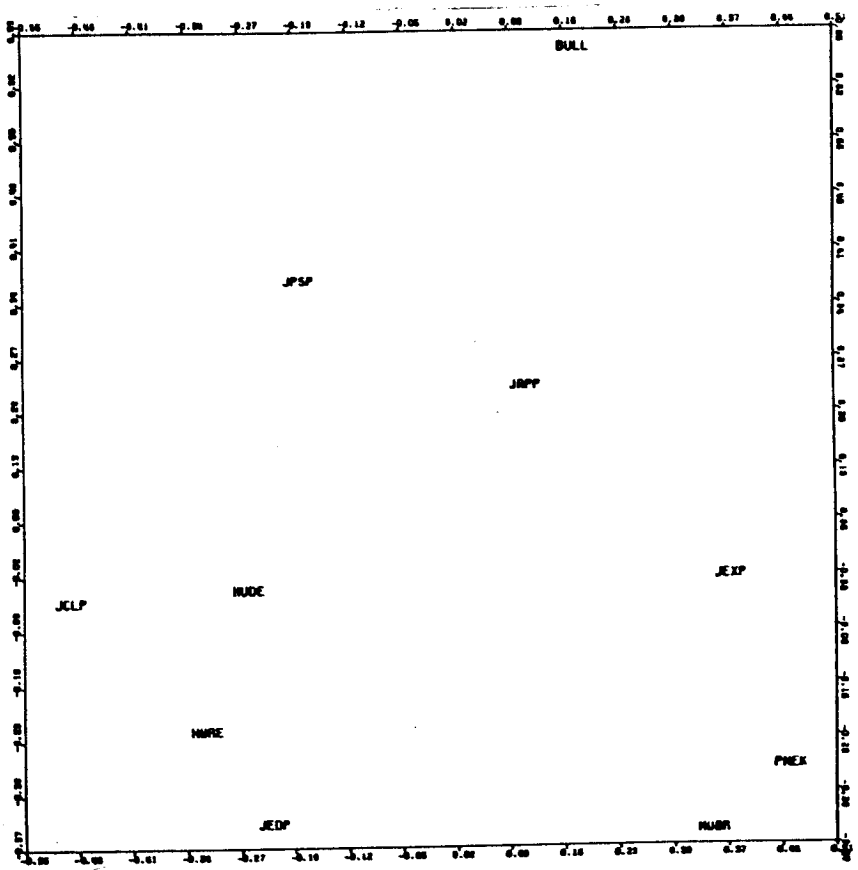


Figure 10.6 Roskam's journal preference data: object scores in the single vector analysis

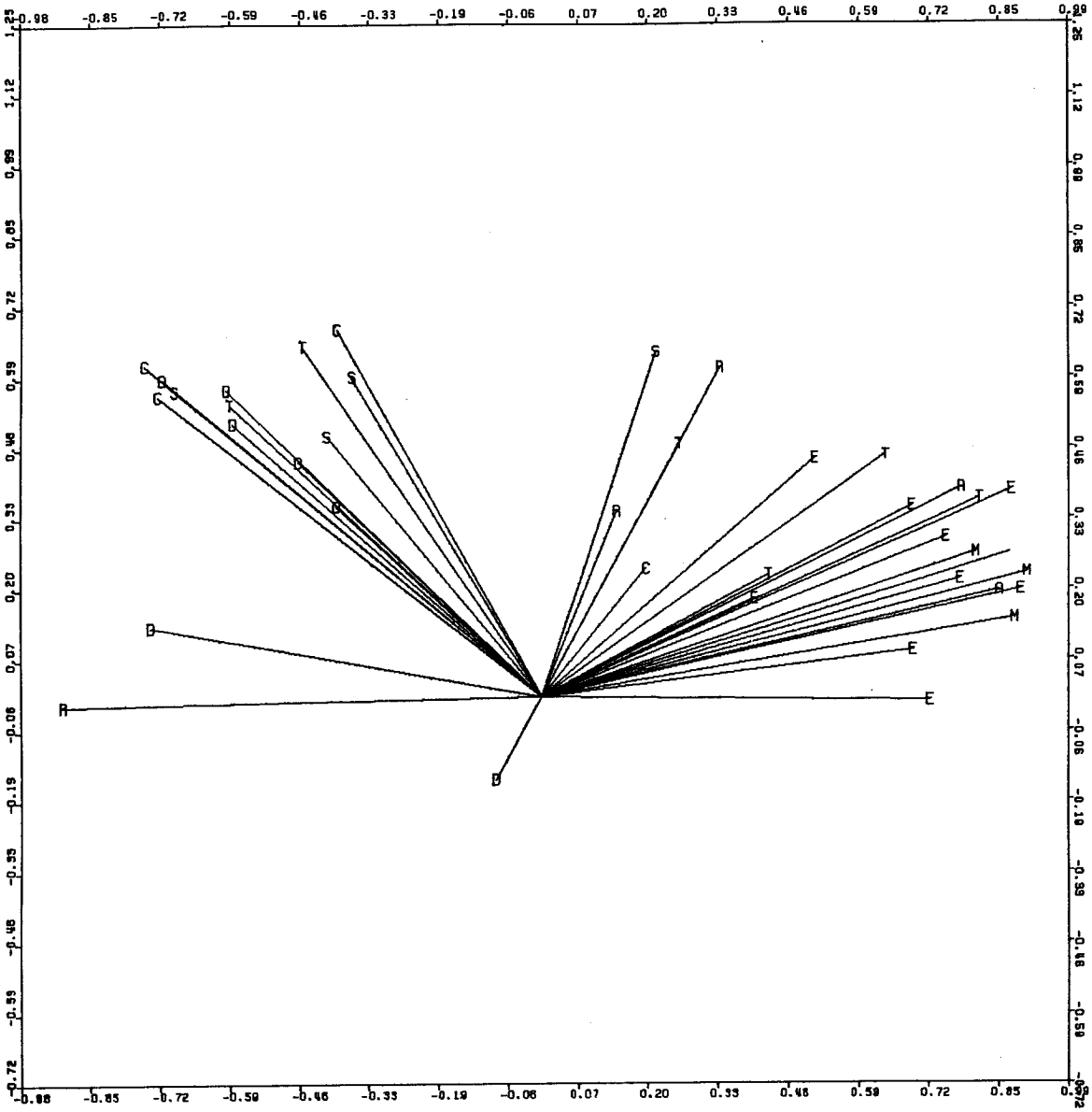


Figure 10.7 Roskam's journal preference data: loadings in the single vector analysis

10.3 Other linear constraints

Equality constraints are a simple special case of linear constraints. Additivity constraints, which we use to define sets of variables, are a special case too. In general we can require that $X = S_0 X$ and $Y_j = S_j Y_j$ for any 'design-matrices' S_j without too much complications. If we require, for example, that the category quantifications must be polynomials of degree less than or equal to s, then the S_j are polynomials on the category numbers, presumably orthogonal with respect to D_j. Additivity constraints have already been analyzed in some detail. We merely mention here that they can be used to define sets, but also in ANOVA and MANOVA situations and their categorical extensions we discussed briefly in chapters 7 and 8.

Another interesting application of linear constraints is to paired comparisons scaling. This was already discussed in Guttman (1946), it was extended to other more complicated situations by De Leeuw (1973). Nishisato (1978) rederives Guttman's method from a different starting point, his solution is also discussed in his 1980 book. We suppose that all m^2 pairs that can be formed of m objects are judged by n subjects. This defines m^2 indicator matrices, which we write as $G_{(j,\ell)}$, meaning that pair (j,ℓ) is compared. Now we use the coding in which $g_{i1}^{(j,\ell)}$ if subject i picks object j from the pair (j,ℓ), we have $g_{i2}^{(j,\ell)} = 1$ if the subject is undecided, and $g_{i3}^{(j,\ell)} = 1$ if he/she picks object ℓ. The linear constraints we use are $Y_{(j,\ell)} = S_{(j,\ell)} T$, with $S_{(j,\ell)}$ a given $3 \times m$ matrix, and with T the unknown $m \times p$ matrix of object quantifications we look for. The matrix $S_{(j,\ell)}$ is defined b $S_{1j}^{(j,\ell)} = S_{3\ell}^{(j,\ell)} = +1$ and $S_{1\ell}^{(j,\ell)} = S_{3j}^{(j,\ell)} = -1$, while all other elements of $S_{(j,\ell)}$ including all of the second row, are equal to zero. Our technique, with the normalization condition $X'X = I$, amounts to computing a singular value decompositi of the $n \times m$ matrix

$$\sum_{(j,\ell)} G_{(j,\ell)} S_{(j,\ell)}.$$

By using the definitions in the appropriate way we find that element (i,j) of this matrix is equal to

$$\sum_{\ell=1}^{m} \{(g_{i1}^{(j,\ell)} - g_{i3}^{(j,\ell)}) - (g_{i1}^{(\ell,j)} - g_{i3}^{(\ell,j)})\}.$$

If there are no ties, then this simplifies because $g_{i1}^{(j,\ell)} + g_{i3}^{(j,\ell)} = 1$, and the same thing is true for $G_{(\ell,j)}$. If we only ask the subject for $\binom{m}{2}$ judgments, and fill in the other half by mirroring then more simplification becomes possible because $g_{i1}^{(j,\ell)} = g_{i3}^{(\ell,j)}$ and $g_{i3}^{(j,\ell)} = g_{i1}^{(\ell,j)}$. If the subject is perfectly consister and his judgments represent a strict order without ties of the m objects, then the matrix we have to analyse is the row-centered matrix of rank numbers again. Thus we recover the same techniques as in the previous section, but more general

because we allow for ties and inconsistencies, and with slight modifications even for missing data. There is an interesting generalization of this technique which allows for response bias. Here T is (2m) × p, and $S_{(j,\ell)}$ is 3 × (2m) with first row equal to $e_j - e_{m+\ell}$, in stead of $e_j - e_\ell$, and with third row equal to $e_\ell - e_{m+j}$, in stead of $e_\ell - e_j$. This gives a double quantification of all objects, one as the first memeber of a pair and one as the second member. If the subjects are consistent, and thus unbiased towards choosing the first element of the pair or the second, then this new method is no generalization, because both object quantifications will be equal.

Because the coding method is a bit complicated we illustrate it in table 10.3. A single subject has judged the nine pairs that can be formed with three stimuli a, b, and c. The judgments are in table 10.3.a. All $G_{(j,\ell)}S_{(j,\ell)}$ are 1 × 3 matrices, we write them below each other in table 10.3.b, together with their sum. We also give the double coding, which allows for response bias in table 10.3.c.

a = a	0	0	0	0	0	0	0	0	0	0
a > b	+1	-1	0	+1	0	0	0	-1	0	
a > c	+1	0	-1	+1	0	0	0	0	-1	
b > a	-1	+1	0	0	+1	0	-1	0	0	
b > b	0	0	0	0	0	0	0	0	0	
b < c	0	-1	+1	0	-1	0	0	0	+1	
c < a	+1	0	-1	0	0	-1	+1	0	0	
c > b	0	-1	+1	0	0	+1	0	-1	0	
c = c	0	0	0	0	0	0	0	0	0	
	+2	-2	0	+2	0	0	0	-2	0	

table 10.3.a: judgments of a single subject
b: paired comparisons coding plus sum
c: double paired comparisons coding

It is, of course, perfectly legitimate to analyze the $G_{(j,\ell)}$ by HOMALS without restrictions. This means that we compute correlations between binary judgments, and that we quantify, either by HOMALS, or by principal components analysis of the phi-coefficients, all pairs of objects. These pair quantifications can then be used in a subsequent analysis, for example to derive a single scale of objects by analysis of variance methods such as the ones outlined by Scheffé (1952), Bechtel (1967, 1971, 1976). The scores for the individuals can be used directly, without further analysis. Observe that if the unfolding model is true, then the individuals that prefer object j to object ℓ can be separated from the individuals that prefer object ℓ to object j by a hyperplane perpendicular to the line connecting j and ℓ, and bisecting this line. From the general theory of chapter 5 this means that PRINCALS with continuous ordinal option on the binary data matrix would give a perfect solution. But we have seen that continuous ordinal does not work properly, and that consequently there is no PRINCALS with continuous ordinal

option. We have also seen that the discrete options can be interpreted as approximating the continuous ones, and that consequently HOMALS or PRINCALS on the $G_{(j,\ell)}$ can be thought of as approximating the positions the subject-points under the unfolding model.

10.4 Zeroes at specific places

Another type of linear constraint demands that some of the parameters are equal to fixed known values, while other are free to vary. In the single numerical case, for example, we fix the transformations and find optimal loadings, weights, scores In PATHALS we require that some of the regression coefficients are equal to zero and others are equal to one. We give one more interesting example of requiring zeroes at specified places. Suppose we are fitting PRINCALS with n individuals, m variables, and p dimensions. Then usually $p < m < n$, because $p > m$ is just a waste of space, we can obtain perfect fit in $p = m$ dimensions. Remember that PRINCALS with single variables and without option I missing data maximizes the sum of the p largest eigenvalues of the correlation matrix of the transformed variables. If $p = m$, then this sum is always equal to m, no matter how we choose the transformations, and consequently there is nothing to maximize.

Now suppose that $m < p < n$ and that we require zeroes in specified places in A. The loss function is as usual. Thus

$$\sigma(X,Y,A) = \frac{1}{m}\sum_{j=1}^{m} SSQ(X - G_j y_j a_j'),$$

and the normalizations are $X'X = I$, $u'G_j y_j = 0$, and $y_j'D_j y_j = 1$. We apply alternati least squares, as usual. Fitting y_j for fixed X and A is the same as in chapter 5. Fitting A for fixed X and Y is also simple. We define the $n \times m$ matrix Q with colu $q_j = G_j y_j$. We also define $\tilde{a}_j = X'q_j$, and write

$$\sigma(X,Y,A) = \frac{1}{m}\sum_{j=1}^{m} SSQ(X - q_j\tilde{a}_j') + \frac{1}{m}\sum_{j=1}^{m} SSQ(\tilde{a}_j - a_j),$$

where we use the fact that $SSQ(q_j) = 1$. It follows that the optimal a_j for given and Y is equal to $\tilde{a}_j$, with the elements of $\tilde{a}_j$ replaced by zeroes on the required places. Computing the optimal X for fixed A and Y is a bit more complicated. We must maximize tr X'QA over $X'X = I$, with both X and QA matrices of order $n \times p$, but with $\text{rank}(QA) \leq m < p$. The solution is $X = K_1L_1' + K_2TL_2'$, with K_1 the $n \times r$ matrix of left singular vectors corresponding with the nonzero singular values of QA, $r = \text{rank}(QA)$, L_1 the corresponding $p \times r$ matrix of right singular vectors, K_2 an $n \times (n - r)$ matrix of left singular vectors corresponding with the zero singular values, L_2 an $p \times (p - r)$ of corresponding right singular vectors, and T any $(n - r) \times (p - r)$ matrix that satisfies $T'T = I$. Because $r < p < n$ it follo that the solution for X is not unique, which corresponds with the by now rather

familiar 'factor score indeterminacy' problem. Clearly the choice of T determines the further course of the iterations, but the basic convergence theorems for alternating least squares still apply, basically because the mapping of QA to the set of all X of the form $X = K_1L_1' + K_2TL_2'$ is upper semi-continuous, so that Zangwill's general convergence theorems (Zangwill, 1969, chapter 4) apply. The fact that we must choose an arbitrary T in each iteration has made some people, such as Takane, Young, and De Leeuw (1976), very nervous. They consequently prefer a loss function which is defined in the space of correlation matrices $R = Q'Q$, which makes the regression problems that adapt the y_j considerably more complicated. Of course if we define $\sigma(*,Y,A)$ as the minimum of $\sigma(X,Y,A)$ over X for fixed Y and A, then

$$\sigma(*,Y,A) = \frac{1}{m}\{mp + SSQ(A) - 2\sum_{j=1}^{m}\lambda_j^{\frac{1}{2}}(A'RA)\},$$

with $\lambda_j(A'RA)$ the eigenvalues of A'RA, i.e. the squares of the singular values of QA. This is also a loss function defined in terms of R. We can now use the fact that the last term, the sum of the singular values of QA, is convex in A for given Q and convex in Q for given A to use the majorization techniques explained in a different context by De Leeuw and Heiser (1980).

The most interesting application of this algorithm is to factor analysis. This was the application studied by Takane and others (1976), who developed the algorithm FACTALS. Our point in this section is that it is possible to perform factor analysis with PRINCALS, by allowing for the possibility that $p > m$, and by making it possible to require certain patterns of zeroes in A. For unrestricted common factor analysis, for example, A is $m \times (t + m)$, and the last m columns of A must be diagonal. For restricted common factor analysis we can also require zeroes in the first t columns of A. Again we are not saying that the current version of PRINCALS can actually do all this, we are merely pointing out that the necessary modifications are easy to implement, either for us or for anyone else who is interested in categorical or nonmetric common factor analysis.

10.5 Nonlinear restrictions

Just for completeness we also discuss some intersting nonlinear restrictions. The constraints $y_j \in C_j$ we have used many times are nonlinear, and so are the constraints $Y_j = y_ja_j'$. An interesting possibility, which we merely mention, is three mode generalization of our techniques. Suppose for example that the same variables have been used on a number of different occasions or groups of individuals. We can analyze the occasions separately, for example by using HOMALS. This gives category quantifications $Y_{j\tau}$, of variable j at occasion τ. But we can also do all the HOMALSes simultaneously, with the restriction that $Y_{j\tau} = Y_jW_\tau$, with

W_τ either a full or a diagonal $p \times p$ matrix. Alternating least squares algorithms are easy to construct. Observe that if the same individuals are tested on different occasions (possibly even with different variables), then we can also require $X_\tau = XW_\tau$, and if we have both the same variables and the same individuals we can combine the two types of restrictions. Moreover it can be combined with single or additive constraints and so on. We have not the faintest idea if this is useful, but it certainly is in the spirit of modern psychometrics, and it guarantees the possibility of an unending stream of programs and publications in the appropriate journals.

In several sections of this book we have discussed the difference between HOMALS en PRINCALS in the following terms. HOMALS computes p different solutions with approximate join-rank equal to one, PRINCALS computes a single solutions with approximate join-rank equal to p. This interpretation is necessary if we want to see both programs as methods for nonlinear principal components analysis, but we have also seen that alternatively we can interprete both programs as methods of generalized K-set canonical analysis. The principal components interpretation left us with the rather mysterious possibility of computing p different solutions with approximate join rank equal to q, which generalizes both HOMALS and PRINCALS. We have been very vague about this possibility so far, mainly because it cannot be done with the existing programs, and because we have never tried it in practice. In this programmatic and speculative chapter we can suggest a fairly simple possibility. We use our ordinary loss function, with one variable in each set,

$$\sigma(X,Y) = \frac{1}{m}\sum_{j=1}^{m} SSQ(X - G_jY_j),$$

but we choose the dimensionality equal to the product of p and q. This implies that Y_j is $k_j \times pq$, we partition Y_j in p submatrices Y_{js} of order $k_j \times q$, and for each of the Y_{js} we require that $Y_{js} = y_{js}a_{js}'$. We still normalize by $X'X = I$. If we partition X in a similar way into p matrices X_s then we can write

$$\sigma(X,Y) = \frac{1}{m}\sum_{j=1}^{m}\sum_{s=1}^{p} SSQ(X_s - G_jy_{js}a_{js}'),$$

with normalization $X_s'X_s = I$ and $X_s'X_t = 0$ if $s \neq t$. This shows that we work with p PRINCALS problems at the same time, all of dimensionality q. If $p = 1$ this is PRINCALS, if $q = 1$ this is HOMALS. Again alternating least squares programs along the lines of PRINCALS are quite simple, we must require that the submatrices Y_{js} are of rank one, and thus perform rank one approximation on submatrices. This formulation also shows that it is not essential at all that q is the same for all p.

11 Nonlinear multivariate analysis: some general principles

11.1 Introduction

There are a number of unifying ideas of different levels in this book. Some of them are of a general methodological nature. We take the point of view, for example, that statistics is one of the possible methods to investigate stability of data analytic techniques, and that statistical models are possible gauges for data analytic techniques. We do not go out of our way to practice inductive inference from sample to population, we also are not trying to make coherent and rational decisions all the time. A second unifying principle is data theoretical. In the last analysis all data are categorical or discrete, measurement level is prior information which can be incorporated in the form of restrictions. Thus measurement level is not a property of the data, it is an abuse of language to say that a variable is ordinal or interval. A data analytic unifying tool are least squares loss functions in combination with geometrical interpretations. These tools are used to illustrate the essential unity of multivariate analysis and multidimensional scaling. Of course least squares is also used because of computational convenience, and because of the historical link with classical linear multinormal analysis. The major technical tools are optimal scaling, alternating least squares, and the singular value decomposition. Other ideas which occur frequently are the distinction between the multiple and the single treatment of a variable, and the 'first-step' approach which transforms variables non-linearly in such a way that subsequent linear MVA techniques give 'better' results. The classical distinction between the analysis of dependence and interdependence is generalized to the distinction between join and meet techniques. In this chapter we formalize this last distinction, starting with the finite-dimensional case, then more generally. This framework makes it possible to discuss some other related general problems such as choice of basis, discretization of continuous variables, infinite dimensional gauges, and analysis of stochastic processes.

11.2 Join and Meet

11.2.1 Finite dimension

Suppose $\mathcal{L}$ is the set of all subspaces of $\mathbf{R}^n$. We can order $\mathcal{L}$ partially by letting $L_1 \preceq L_2$ if L_1 is a subspace of L_2, we can make $\mathcal{L}$ into a complete lattice by defining meet(L_1,L_2) to be the intersection of L_1 and L_2 and join(L_1,L_2) to be the linear sum of L_1 and L_2 (i.e. the set of all $x \in \mathbf{R}^n$ of the form $x = y + z$ with $y \in L_1$ and $z \in L_2$). The meet of L_1 and L_2 is the greatest lower bound of L_1 and L_2, i.e. the largest subspace contained in both L_1 and L_2, the join of L_1 and L_2 is the least upper bound of L_1 and L_2, i.e. the smallest subspace that contains both L_1 and L_2. If $L_1,\ldots,L_m$ are subspaces of $\mathbf{R}^n$ we can also define meet$(L_1,\ldots,L_m)$ and join$(L_1,\ldots,L_m)$.

On the lattice of subspaces we can define the valuation dim( ), the dimensionality

of the subspace. Using dim( ) we define the meet-rank of $L_1,\ldots,L_m$ as

$$mrk(L_1,\ldots,L_m) \triangleq dim(meet(L_1,\ldots,L_m)),$$

and the join-rank as

$$jrk(L_1,\ldots,L_m) \triangleq dim(join(L_1,\ldots,L_m)).$$

We say that $L_1,\ldots,L_m$ have a p-meet if $mrk(L_1,\ldots,L_m) \geq p$ and $L_1,\ldots,L_m$ have a p-join if $jrk(L_1,\ldots,L_m) \leq p$.

A very useful feature of these definitions is that they are coordinate-free, i.e. they do not depend on the choice of a basis in R^n. Another useful aspect is that we can use the general results from lattice theory (Birkhoff, 1967), which give interesting result when interpreted in the data analysis framework implied by our defintions. We first give this interpretative framework. The meet-problem (qualitative version) is to find a subspace L of R^n such that dim(L) = p and $L \sim meet(L_1,..,L_m)$. Thus given $L_1,\ldots,L_m$ and given an integer p with $1 \leq p \leq n$ we try to find a subspace of dimension p which is approximately contained in all L_j. Clearly it is possible to find a subspace which satisfies these conditions exactly if the $L_1,\ldots,L_m$ have a p-meet. The meet-problem is an abstract version of the problems we solve in chapters 6, 7, 8, an important difference is that in these chapters we have introduced a least squares loss function and we can consequently define the meet-problem quantitatively. The join-problem, which is solved in pricipal components analysis, has a similar abstract version. We start with $L_1,\ldots,L_m$ and an integer p. We now try to find a subspace L such that dim(L) = p and $L \sim join(L_1,\ldots,L_m)$. Again an exact solution is possible if and onl if $L_1,\ldots,L_m$ have a p-join, and a quantitative formulation of the problem is possible by introducing loss functions.

It is clear from the definitions that if $L_1,\ldots,L_m$ have a p-meet, then they also have a q-meet for all $q \leq p$. And if they have a p-join, then they also have a q-j for all $q \geq p$. Thus we want to choose p as large as possible in meet-problems, an as small as possible in join-problems. Again 'as possible' is defined qualitative and can be translated into quantitative terminology only by introducing loss functions. Thus in the meet-problem we want to find subspaces of large dimension contained in all given subspaces, in the join problem we want to find subspaces of small dimension which contain all given subspaces.

Before we continue with the abstract discussion of our concepts it is probably useful to give some concrete examples of the subspaces we usually deal with. In metric K-set canonical analysis we try to solve a meet problem, and the K subspaces are defined as the column spaces of K given matrices. In homogeneity analysis, interpreted as a meet problem, there are m subspaces, defined as the column spaces of the indicator matrices G_j. Thus $x \in L_j$ if $x_i = x_k$ whenever

the number of categories of variable j. In metric principal components analysis we solve a join problem. Each variable defines a one-dimensional subspace, the ray of all n-vectors proportional to the observations on that variable. In nonlinear principal components analysis with discrete variables the indicator matrices are a basis for the subspace of all nonlinear transformations of the variable. Thus nonlinear principal components analysis solves to join problem for the column spaces of the G_j.

We now state some simple results on the functions mrk and jrk. This is merely classical linear algebra, translated in our notation. We start with

$$0 \lesssim \text{mrk}(L_1,\ldots,L_m) \lesssim \min_{j=1}^{m} \dim(L_j),$$

with equality on the left if and only if the intersection of the L_j is the zero vector, and with equality on the right if and only if one of the L_j contains all the others. For jrk the corresponding inequality is

$$0 \lesssim \text{jrk}(L_1,\ldots,L_m) \lesssim \sum_{j=1}^{m} \dim(L_j),$$

with equality on the left if and only if all L_j are equal to the zero vector, and equality on the right if and only if the pairwise intersection of the L_j is the zero vector (i.e. if and only if the linear sum is equal to the direct sum). We can also relate the two concepts directly

$$\text{mrk}(L_1,\ldots,L_m) \lesssim \text{jrk}(L_1,\ldots,L_m),$$

with equality if and only if $L_1 = \ldots = L_m$. Some of the general lattice theory results we mentioned earlier are the distributive inequalities

$$\text{meet}(L_1,\text{join}(L_2,L_3)) \gtrsim \text{join}(\text{meet}(L_1,L_2),\text{meet}(L_1,L_3)),$$

$$\text{join}(L_1,\text{meet}(L_2,L_3)) \lesssim \text{meet}(\text{join}(L_1,L_2),\text{join}(L_1,L_3)).$$

Because $\mathbf{R}^n$ is finite-dimensional the lattice of subspaces is modular. This means that if L_1 is a subspace of L_3 then

$$\text{join}(L_1,\text{meet}(L_2,L_3)) = \text{meet}(\text{join}(L_1,L_2),L_3).$$

An extremely important result, which shows why the case of two subspaces is very special indeed, is

$$\text{mrk}(L_1,L_2) + \text{jrk}(L_1,L_2) = \dim(L_1) + \dim(L_2).$$

Another important result can be described in words as follows. Suppose L_j is the join of k_j one-dimensional subspaces. Then $\text{join}(L_1,\ldots,L_m)$ is the join of all $\sum k_j$ one-dimensional subspaces that define the L_j. The dual result for the meet is that if L_j is the meet of k_j subspaces, then $\text{meet}(L_1,\ldots,L_m)$ is the meet of all $\sum k_j$ subspaces defining the L_j. But this result is far less interesting, because we cannot add the description 'one-dimensional' here. The result for the

A second difference is that we do not need rank restrictions on X, the subspaces have a p-join if and only if the minimum of σ_J over all Y_j and all X is equal to zero, moreover this minimum is always attained, and not always equal to zero. Again this will be shown in the sequel. A third definition of σ_J is sometimes also useful. We can write

$$\sigma_J(X;Y) = \frac{1}{m} \sum SSQ(g_\ell - Xy_\ell),$$

where the summation is over all columns of the supermatrix G, i.e. we have reduced the join problem to its one-dimensional subspaces again.

11.2.3 Further analysis of meet-loss

We now study σ_M more closely by using familiar results from least squares and eigenvalue theory. Define

$$\sigma_M(*;Y_1,\ldots,Y_m) \triangleq \min \{\sigma_M(X;Y_1,\ldots,Y_m) \mid X\},$$

thus the minimum is computed over all X, not normalized or restricted in any way. Also define

$$\overline{X} \triangleq \frac{1}{m} \sum_{j=1}^{m} G_j Y_j.$$

Then

$$\sigma_M(X;Y_1,\ldots,Y_m) = SSQ(X - \overline{X}) + \frac{1}{m} \sum_{j=1}^{m} SSQ(\overline{X} - G_j Y_j),$$

and thus

$$\sigma_M(*;Y_1,\ldots,Y_m) = \frac{1}{m} \sum_{j=1}^{m} SSQ(\overline{X} - G_j Y_j).$$

We can define the supermatrix C, of order $(\sum k_j) \times (\sum k_j)$, with submatrices $C_{j\ell} \triangleq G_j'G_\ell$, and the diagonal supermatrix D of the same order, with diagonal submatrices $D_j \triangleq C_{jj}$. Then

$$\sigma_M(*;Y_1,\ldots,Y_m) = \frac{1}{m} (\text{tr } Y'DY - \frac{1}{m} \text{tr } Y'CY).$$

These equations are familiar from the analysis of homogeneity in chapter 3, in analysis of variance terminology Y'DY is the total dispersion of the G_jY_j, $\frac{1}{m}$ Y'CY is the between-set dispersion, and thus Y'DY - $\frac{1}{m}$ Y'CY is the within-set dispersion.

Now let $r_j \triangleq \text{rank}(G_j)$, and $r \triangleq \text{rank}(D) = \sum r_j$, and $s = \min(p,r)$. Define

$$\sigma_M(*;*,\ldots*) = \min \{\sigma_M(*;Y_1,\ldots,Y_m) \mid Y'DY = I_s\}.$$

where I_s is the diagonal matrix of order p with s elements equal to one, and the remaining diagonal elements equal to zero. Then

$$\sigma_M(*;*,\ldots,*) = s - \sum_{t=1}^{s} \lambda_t(\frac{1}{m} D^{-\frac{1}{2}}CD^{-\frac{1}{2}}),$$

where $\lambda_1( ) \geq \ldots \geq \lambda_s( )$ are the ordered eigenvalues, and where $D^{-\frac{1}{2}}$ is the symmetric square root of the Moore-Penrose inverse of D.

We now reverse the roles of X and Y in this derivation. We first define

$$\sigma_M(X;*,\ldots,*) \triangleq \min \{\sigma_M(X;Y_1,\ldots,Y_m) \mid Y_1,\ldots,Y_m\},$$

where the minimum is computed over all Y_j, al unrestricted. Define, using the Moore-Penrose inverse G_j^+,

$\overline{Y}_j \triangleq G_j^+X$, then

$$\sigma_M(X;Y_1,\ldots,Y_m) = \frac{1}{m}\sum_{j=1}^{m} SSQ(X - G_j\overline{Y}_j) + \frac{1}{m}\sum_{j=1}^{m} tr\ (Y_j - \overline{Y}_j)'G_j'G_j(Y_j - \overline{Y}_j),$$

a partitioning which is already familiar from chapter 5. Thus

$$\sigma_M(X;*,\ldots,*) = \frac{1}{m}\sum_{j=1}^{m} SSQ(X - G_j\overline{Y}_j).$$

By using supermatrices this can again be written in matrix notation as

$$\sigma_M(X;*,\ldots,*) = tr\ X'X - \frac{1}{m}\ tr\ X'GD^+G'X.$$

If

$$\sigma_M(*;*,\ldots,*) \triangleq \min \{\sigma_M(X;*,\ldots,*) \mid X'X = I_s\}, \text{ then}$$

$$\sigma_M(*;*,\ldots,*) = s - \sum_{t=1}^{s} \lambda_t(\frac{1}{m} GD^+G').$$

But the nonzero eigenvalues of GD^+G' are the same as those of $D^{-\frac{1}{2}}CD^{-\frac{1}{2}} = D^{-\frac{1}{2}}G'GD^{-\frac{1}{2}}$. Consequently the fact that we have define $\sigma_M(*;*,\ldots,*)$ in two different ways does not matter, it is indeed true that

$$\min \{\sigma_M(X;Y_1,\ldots,Y_m) \mid X,\ Y'DY = I_s\} = \min \{\sigma_M(X;Y_1,\ldots,Y_m) \mid X'X = I_s,\ Y_1,\ldots,Y_m$$

We have already used this result in many places, both for theoretical purposes and for algorithm construction. We now see that it is true for meet-loss in genera even if some of the G_j are singular.

Thus the subspaces L_j (or the corresponding matrices G_j) have a p-meet if and only if the matrices $\frac{1}{m} GD^+G'$ or $\frac{1}{m} D^{-\frac{1}{2}}CD^{-\frac{1}{2}}$ have p eigenvalues equal to one. In our analysis we have included the possibility that $s < p$. In that case

$$mrk(L_1,\ldots,L_m) \leq jrk(L_1,\ldots,L_m) \leq r = s < p,$$

which implies that the L_j have no p-meet. Our results can also be used to derive some simple lower bounds for $\sigma_M(*;*,\ldots,*)$. Because

$$GD^+G' = \sum_{j=1}^{m} G_j(G_j'G_j)^+G_j',$$

we see that

$$\sum_{t=1}^{s} \lambda_t(GD^+G') \leq \sum_{j=1}^{m}\sum_{t=1}^{s} \lambda_t(G_j(G_j'G_j)^+G_j') = \sum_{j=1}^{m} \min(s,r_j).$$

with equality if and only if $mrk(L_1,\ldots,L_m) \gtrsim s$. It follows from this inequality that

$$\sigma_M(*;*,\ldots,*) \gtrsim \frac{1}{m} \sum_{j=1}^{m} \max(0, s - r_j).$$

Thus $\sigma_M(*;*,\ldots,*) = 0$ is possible only if $s \lesssim r_j$ for all j. In most situations we have $s = p$ and $r_j = k_j$, with k_j the number of columns of G_j as usual. In that case meet-loss can vanish only if $p \lesssim k_j$ for all j. If $p \gtrsim r$, then $s = r \gtrsim r_j$, and thus

$$\sigma_M(*;*,\ldots,*) \gtrsim \frac{m-1}{m} r.$$

If $p \lesssim r$ and $p \gtrsim r_j$ for all j, then

$$\sigma_M(*;*,\ldots,*) \gtrsim p - \frac{r}{m}.$$

11.2.4 Further analysis of join-loss

For join-loss the situation is much simpler. We use the supermatrix formulation $\sigma_J(X;Y)$. Define

$$\sigma_J(*;Y) \triangleq \min \{\sigma_J(X;Y) \mid X\}.$$

If

$\overline{X} \triangleq G(Y^+)'$, then

$$\sigma_J(X;Y) = \frac{1}{m} \{SSQ(G - \overline{X}Y') + tr\ (X - \overline{X})Y'Y(X - \overline{X})'\}, \text{ and consequently}$$

$$\sigma_J(*;Y) = \frac{1}{m} SSQ(G - \overline{X}Y') = \frac{1}{m} tr\ (I - YY^+)C,$$

with $C = G'G$ as usual. If

$$\sigma_J(*;*) \triangleq \min \{\sigma_J(*;Y) \mid A\}, \text{ then}$$

$$\sigma_J(*;*) = \sum_{s=p+1}^{\Sigma k_j} \lambda_s(\tfrac{1}{m}G'G).$$

If

$$\sigma_J(X;*) \triangleq \min \{\sigma_J(X;Y) \mid Y\}, \text{ then because}$$

$$\sigma_J(X;Y) = \frac{1}{m} \{SSQ(G - X\overline{Y}') + tr\ (Y - \overline{Y})X'X(Y - \overline{Y})'\}, \text{ with}$$

$\overline{Y} \triangleq X^+G$, we have

$$\sigma_J(X;*) = \frac{1}{m} SSQ(G - X\bar{Y}') = \frac{1}{m} tr\ (I - XX^+)GG'.$$

Thus

$$\sigma_J(*;*) = \min\ \{\sigma_J(X;*) \mid X\} = \sum_{s=p+1}^{n} \lambda_s(\tfrac{1}{m}GG').$$

This derivation shows, again, that there is no need to normalize if we minimize join-loss, it also shows that the partition of G into submatrices is irrelevant for join loss.

11.2.5 Meet and join problems

We have been concerned with definitions and with perfect fit so far. In the case of meet-loss we found that the subspaces have a p-meet if $\frac{1}{m} GD^+G'$ has at least p eigenvalues equal to one, they have a p-join if G has at most p nonzero singular values. It is now easy to define meet and join-solutions in an approximat sense: we simply carry out the minimizations in the previous sections to find solutions for X and Y for given p. These are the best meet and join-solutions for that particular value of p. We can use the value of $\sigma_M(*;*,\ldots,*)$ and of $\sigma_J(*;*)$ to find out if join-rank or meet-rank are really equal to p, and of course they probably are not. All in all the concepts of meet and join are gauges, it is usually not true that there is a 'true' rank we want to discover. The only reasonable model in most situations is that the meet-rank is zero and the join-rank is $\sum k_j$ (if there are no build-in singularities). We are explicitly interested in the best approximation of the p-meet and p-join for fixed p, this is the technique which should be gauged. The fact that there usually is much more information in the data than is 'explained' by our choice of p is irrelevant, if we want to 'explain' all the information we do not need a technique to analyze them, we only need a technique to copy them.

We have shown in the previous sections that join and meet are independent concept defined first without using coordinates in the lattice of subspaces, and later by using coordinates in terms of the eigenvalues of certain matrices. There are, however, some interesting relationships between the two concepts, which are most easily explained on a loss function level. We have already commented on thes relationships in previous chapters, we formulate them more generally here. Suppos $G_j'G_j = I$ for all j. Then

$$\sigma_M(X;Y_1,\ldots,Y_m) = SSQ(X) - \frac{2}{m}\sum_{j=1}^{m} tr\ Y_j'G_j'X + \frac{1}{m}\sum_{j=1}^{m} SSQ(Y_j),$$

or, using supermatrices,

$$\sigma_M(X;Y) = SSQ(X) - \frac{2}{m} tr\ X'GY + \frac{1}{m} SSQ(Y).$$

On the other hand

$$\sigma_J(X;Y) = \frac{1}{m}\sum_{i=1}^{m} k_j - \frac{2}{m} tr\ X'GY + \frac{1}{m} SSQ(XY').$$

If we assume in addition that $X'X = I$, which can be done without loss of generality, then

$$\sigma_M(X;Y) = \sigma_J(X;Y) + (p - \frac{1}{m}\sum_{j=1}^{m} k_j).$$

Thus if $G_j'G_j = I$ the problems only differ by a constant, and optimizing join-loss and meet-loss obviously has the same solution. In both cases the optimal X and Y can be computed from the eigenanalysis of GG' or $G'G$, or from the singular value decomposition of G. The relationship between the loss functions also implies

$$\sigma_M(X;Y) \geq p - \frac{1}{m}\sum_{j=1}^{m} k_j,$$

and

$$\sigma_J(X;Y) \geq \frac{1}{m}\sum_{j=1}^{m} k_j - p.$$

The condition $G_j'G_j = I$ for all j may look very special, but if we remember that the G_j were defined as any matrix for which L_j is the column space, than we see that it can actually be made without loss of generality. Or, to put it differently, if we choose the G_j as an orthonormal basis for L_j, then both the join and the meet-solution can be computed from the singular value decomposition of the super-matrix G. Moreover it is clear that the definitions of join-rank and meet-rank are certainly independent from the choice of spanning set G_j. There is another interesting special case, which seems a bit weird at first sight. Suppose $k_j = 1$ for all j. Clearly the meet-rank of $L_1,\ldots,L_m$ in this case is either one or zero, but we can still use the meet-loss as defined. We then find

$$\sigma_M(X;Y) = \sigma_J(X;Y) + (p - 1),$$

which shows that a general algorithm to minimize σ_M can also solve problems involving σ_J. We have already used this fact extensively in 5.8, and actually PRINCALS minimizes meet-loss in order to solve the join-problem.

11.2.6 Some extensions of the join and meet framework

Our formulation of meet and join so far starts with subspaces, then introduces bases or spanning sets which are essentially arbitrary, and uses these spanning sets to reduce the meet and join problems to eigenvalue-eigenvector problems. This is only possible if the variables we deal with are treated either as single numerical or as multiple nominal, we have seen in the previous chapters that in that case our problems reduce to eigen-problems for known matrices. The situation is considerably less simple if we decide to treat some variables as single nominal or single ordinal. In this case some of the columns of some of the G_j are unknown, and the definition of join and meet becomes less natural and more complicated.

We first define $\mathcal{L}_j$ which is the set of subspaces of $\mathcal{L}$ generated by all permissible choices of the columns of G_j. Thus we choose columns of G_j that are unknown or

partially unknown in their respective feasible regions (usually convex cones), for each choice of these columns we can compute the subspace spanned by these columns, the set of all such subspaces is $\mathcal{L}_j$. We then define

$$\text{mrk}(\mathcal{L}_1,\ldots,\mathcal{L}_m) \triangleq \max\{\text{mrk}(L_1,\ldots,L_m) \mid L_1 \in \mathcal{L}_1,\ldots,L_m \in \mathcal{L}_m\},$$

$$\text{jrk}(\mathcal{L}_1,\ldots,\mathcal{L}_m) \triangleq \min\{\text{jrk}(L_1,\ldots,L_m) \mid L_1 \in \mathcal{L}_1,\ldots,L_m \in \mathcal{L}_m\}.$$

These definitions make it possible to repeat most of our earlier results in this chapter, with the additional complication that in the loss functions we also have to choose the columns of the G_j in the appropriate way. Our previous theorems show directly that the meet-problem quantifies the variables in such a way that the largest eigenvalues of GD^+G' are maximized, the join-problem quantifies variables in such a way that the smallest eigenvalues of $G'G$ are minimized. More precisely we either maximize the sum of the p largest eigenvalues or we minimize the sum of the p smallest. If each G_j has only one column, then the eigenvalues of GD^+G' are those of $D^{-\frac{1}{2}}G'GD^{-\frac{1}{2}}$, which is the correlation matrix of the variables. In this general formulation there are no problems with missing data, we simply adapt the definition of the feasible columns of G_j in the appropriate way. As we have seen in previous chapters it sometimes is desirab to normalize object scores alternatively, for example by $X'M_*X = I$ in stead of X' Clearly using such an alternative normalization does not change the definition of perfect fit, i.e. of having a p-meet or a p-join, but it can change the approximate solutions considerably.

11.2.7 Extensions to infinite-dimensional space

In chapter 1 we have indicated that our basic approach does not only apply to vectors of n observations, but also to general random variables with finite variances, defined on the same probability space. The set of all such random variables is a separable Hilbert space, in which the covariance is the inner product and the variance is the square of the norm. The nonlinear transformations of a given random variable define a subspace of this same space, in the general case also a subspace of infinite dimension. Now there is no reason to think that infinite dimensional spaces are very difficult to understand, or have very complicated properties. In our case we restrict ourselves to the most simple and natural generalization of R^n, with the consequence that our coordinate-free definition of join-rank and meet-rank still applies, although some of the other results in 11.2.1 are no longer true. The biggest nuisance, however, is computati We get into trouble with infinite dimensionality as soon as we want to apply our formulas to do some actual computing, the formulas themselves do not change very much. We simply interpret G_jY_j, for example, as involving an infinite summation, G_j has possibly a countable infinite number of columns. Moreover SSQ is replaced by VAR. Later in this chapter we shall illustrate some of the

11.3 Choice of basis

11.3.1 Finite dimension

In the previous sections we have seen that choice of basis or spanning set is not important from a theoretical point of view, the minimum values of the loss functions is independent from the choice of basis. On the other hand the choice can be important for computational and/or interpretational reasons. It clearly simplifies the computations if the G_j are chosen in such a way that $G_j'G_j$ is diagonal, this happens for example if G_j is a single complete or incomplete indicator matrix. For categorical variables choice of basis is usually no problem, the indicator matrices are both general enough and have all the desirable properties. A possible exception is the case in which the variables have a very large number of categories, and in which some of the categories have very few observations. In those cases we often merge categories, which means that we choose to work in a lower-dimensional subspace of nonlinear transformations, because we expect the results to be more stable in that subspace.

An extreme case is the continuous variable. We do not mean the infinite-dimensional case, this will be treated in the next section. The number of observations is still finite and equal to n, but we assume that k_j is approximately equal to n. In this case indicator matrices for the separate categories will be approximate permutation matrices, and the situation is close to the 'm rankings' problem we discussed in chapters 5 and 10. We have seen there that one possible way out of the problem that a perfect trivial solution exists, is to use single variables and to impose ordinal or even numerical constraints. Examples such as the Roskam journal data show that this works satisfactory in these situations. On the other hand we also have to remember the result of 5.2.4, which tells us that single ordinal will usually not work if the number of categories is close to the number of observations in other situations. The difference between the two types of situations is perhaps still worth emphasizing. In the situation discussed in 5.2.4 individuals are a sample from a population, in the preference rankings situation 'individuals' in the program or technique are the objects that are ranked, 'variables' in the program are the individuals. Thus in 5.2.4 the rows of the supermatrix G are independent replications, in the rankings situation the G_j are independent $n \times n$ matrices, and the argument in 5.2.4 breaks down because now we have to let $m \to \infty$. In a situation such as 5.2.4 with as many categories as individuals any arbitrary transformation of the variables will tend to give perfect homogeneity, and consequently the only thing we can do is to restrict the admissible transformations. Requiring ordinality will usually not be enough, requiring linearity defeats our purpose, which is to generalize linear techniques. Thus we continue to work with linear subspaces, but we restrict their dimensionality.

The simplest way to do this, which also works for nominal variables, is to decreas the number of categories by merging them. In the numerical case this is the familiar problem of the choice of category boundaries, in this case it is often possible to give some rough guidelines on how to categorize in a satisfactory way. This will be treated in more detail in the next chapter. If the variables are nominal or ordinal there are no clear-cut rules, although it is probably always good practice to eliminate categories with a very small number of observati and some analysis in the next chapter shows that we can always eliminate categorie with small discrimination measures (or more generally with very bad fit). Of cours this last criterion can only be applied after some preliminary analysis.

For numerical variables there are some alternative possibilities which are also interesting. We can fit nonlinear transformations which are restricted to be polynomials of low order, this defines a low-dimensional subspace which we can parametricize in such a way that $G_j'D_jG_j = I$ (thus using D_j-orthogonal polynomials). Polynomials have the somewhat unfortunate property that they are too rigid for flexible approximation of general functions, sometimes we need a very high degree to get a good approximation, and of course choosing a polynomial of degree k_j-1 is equivalent to choosing a completely general nonlinear transformation, i.e. to using the $n \times k_j$ indicator matrix. This fact shows another advantage of considering different bases, it is sometimes nicer to interpret the Y_j in G_jY_j if G_j is a basis of D_j-orthogonal polynomials. In this case the category quantifications and their squares can be interpreted directly as the 'contributions' of the various degrees to the nonlinear transformation. Another disadvantage of polynomials is that additional constraint on them, for example that they are monotone, or nonnegative, or convex are rather difficult to incorporate. Regression problems with these constraints are of course quadratic programming problems, just as monotone regeression is a quadratic programming problem, but they are not simple quadratic programming problems.

Another possibility seems more promising than polynomials, although we have not compared the two systematically. Using indicator matrices in combination with numerical variables can be interpreted simply as approximating nonlinear transformations by step-functions. Step-functions do not approximate well in general, not only because they do not have enough parameters, but also because it is very difficult to see from the best approximate step-function what the approximated smooth function looked like. It is, of course, true that if we take sufficiently many categories (jump-points, knots) then arbitrary precise approximation is possible. Thus, as with polynomials, part of our criticism is that we need too many dimensions in the approximating subspace. Splines are intermediate between step functions and polynomials, and their current popularity

first give a very brief introduction to splines. Suppose $\ldots,a_{-1},a_0,a_1,\ldots$ is a doubly infinite increasing sequence of real numbers called knots. They play the same role as category boundaries with step functions. A spline of degree s is a function which is a polynomial of degree s in each of the intervals (a_{k-1},a_k), generally a different polynomial in each interval, with the additional property that the function is s-1 times continuously differentiable on the line. Because polynomials are infinitely many times differentiable this last condition is a restriction only at the knots, where the s-1 derivatibles from the right must be the same as the s-1 derivatives from the left. If s = 0 then our definition merely says that a spline should be constant in each interval between two consecutive knots, i.e. splines with s = 0 are just step functions. If s = 1 then a spline consists of broken line segments which are joined at the knots, which makes the resulting function continuous. Throughout our short introduction we suppose that the knots are fixed and known numbers, and that there are no multiple knots (all a_k are different).

It follows directly from the definition that splines of a given degree s on a fixed knot-sequence a_k form a linear subspace. Because there is an infinite number of knots the subspace is also of infinite dimension, with a finite number of observations, however, the knots which are outside the range of the observations do not matter, and the space becomes finite dimensional. It can be shown that in our situation the dimension is equal to the number of interior knots plus the degree of the spline plus one. We illustrate this with a small example, in which all data points are between 0 and 3, and the knot-sequence consists of the positive integers, the negative integers, and zero (if the knots are equally spaced the corresponding splines are called cardinal splines). We construct the B-spline basis for the splines of degree zero, one and two. The B-splines or basic splines were introduced by Curry and Schoenberg (1966), their properties are discussed extensively in De Boor (1978), where we can also find a proof of the theorem that the B-splines are indeed a basis of the subspace of all splines. B-splines are interesting for numerical purposes because they have local support by which we mean that they are nonzero on s + 1 consecutive knot intervals only. For s = 0 this immediately gives the step (or 'block') functions in figure 11.1.a. Clearly their normalization is immaterial, but we have normalized them in such a way that the n × 3 matrix with the values of the n observations on the three functions is an indicator matrix, with rows that add up to one. In figure 11.1.b we have drawn the four first degree splines, and in figure 11.1.c the five second degree splines. The corresponding object times function matrices are n × 4 and n × 5, we have normalized the functions in such a way that the rows add up to one again. Thus we now have straightforward generalizations of indicator matrices from using B-splines. Increasing the degree of the spline gives a basis with one more column, local support shows that at most s + 1 consecutive elements

are nonzero. The zero-degree B-splines, starting at knot k, satisfy

$g_k(x) = 1$ if $k \leq x < k + 1$,

$g_k(x) = 0$ otherwise.

The first-degree B-splines are

$g_k(x) = x - k$ if $k \leq x \leq k + 1$,

$g_k(x) = (k + 2) - x$ if $k + 1 \leq x \leq k + 2$,

$g_k(x) = 0$ otherwise.

The second-degree B-splines are

$g_k(x) = \frac{1}{2}(x - k)^2$ if $k \leq x < k + 1$,

$g_k(x) = \frac{3}{4} - (x - (k + \frac{3}{2}))^2$ if $k + 1 \leq x < k + 2$,

$g_k(x) = \frac{1}{2}(x - (k + 3))^2$ if $k + 2 \leq x \leq k + 3$,

$g_k(x) = 0$ otherwise.

Remember that these formulas are only for cardinal B-splines, for different knot-sequences we have other functions. De Boor (1978) discusses efficient and stable recursive algorithms to compute B-spline values for any value of x. In table 11.1 we illustrate how the B-splines generate matrices such as the indicator matrices, but with more nonzero entries in each row. It is clear that if G_j is defined using B-splines of degree s, then $G_j'G_j$ also has s + 1 consecutive nonzero elements in each row and column. This can be used to simplify the linear regression problem considerably. We also remark that Winsburg and Ramsay (1980, 1981) fit monotonic splines to data, as a basis for the monotonic splines they use integrals of B-splines. In fact if we normalize the B-splines such that their integral equals one, then the integrated B-splines are distribution functions and they can be used especially nicely to fit latent trace models to binary variables. B-splines and their integrals have many fascinating properties, for which we refer to the book by De Boor (1978), and its references.

We now give an example comparing step functions and (first degree) splines. This example is in the nonlinear regression area, we are currently incorporating splin routinely into a version of PRINCALS but we have no examples of that as yet. Thus there are two subspaces L_1 and L_2, as we already know the case of two subspaces i somewhat special, but we shall approach the problem as a meet problem for these two subspaces. The example is from physics, it was used by Wilson (1926) and by Wilson and Worcester (1935) to illustrate the failure of statistical data analysi techniques such as regression and components analysis to give results that confor to physical theory. Willard Gibbs has discovered a theoretical formula connecting density z, the pressure y, and the absolute temperature x of a mixture of gases w

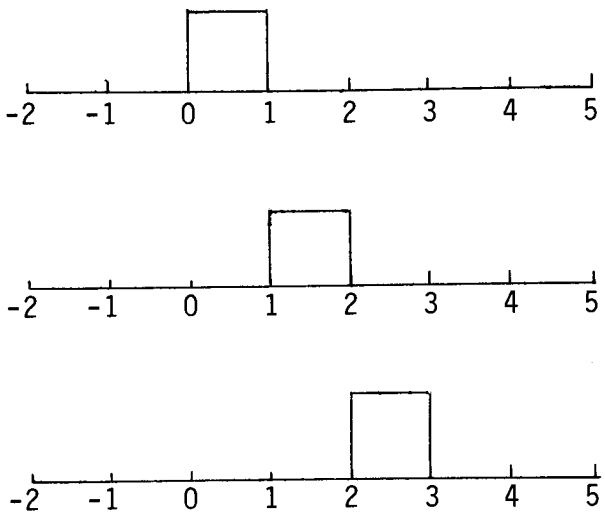


Figure 11.1.a
zero degree B-splines

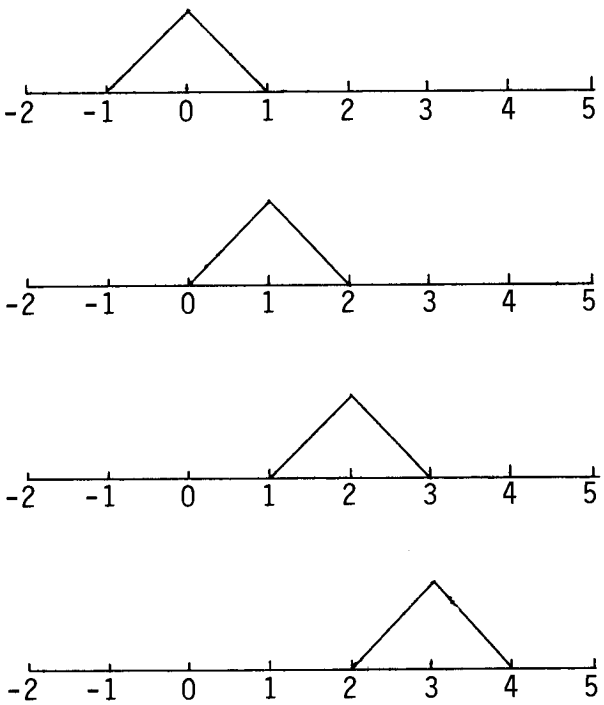


Figure 11.1.b
first degree B-splines

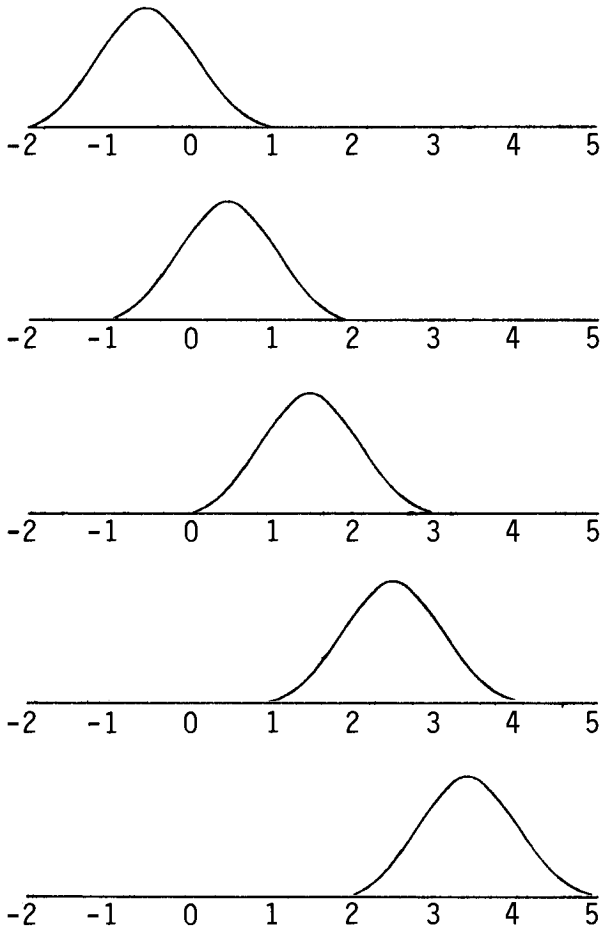


Figure 11.1.c
second degree B-splines

0.50	1 0 0	0.50	0.50	0.00	0.00	0.12500	0.75000	0.12500	0.00000	0.00000
0.75	1 0 0	0.25	0.75	0.00	0.00	0.03125	0.68750	0.28125	0.00000	0.00000
0.90	1 0 0	0.10	0.90	0.00	0.00	0.00500	0.59000	0.40500	0.00000	0.00000
2.30	0 0 1	0.00	0.00	0.70	0.30	0.00000	0.00000	0.24500	0.71000	0.04500

table 11.1: B-splines of degree zero, one, two for four values of x.

convertible components. The formula is

$$\log \frac{A(z - A)}{(2A - z)^2} = \frac{B}{x} + \log y + C.$$

The constant A is the density of the rarer component, and it can be computed from the molecular formula. The constants B and C must be estimated. Gibbs computed the from data of experiments of Cahours and Bineau. Wilson does not state, unfortunate how Gibbs computed these constants. Gibbs then applied his formula, and the estimated constants to 65 experiments of Neumann, and he discusse the systematic and accidental divergences (residuals). Wilson gives an 'empiricist' in stead of a 'rationalist' analysis of the data of Neumann, he fits three regression equations, computes partial correlations and so on, and concludes that it is very difficult to relate the results with the theory of Gibbs, and that it is impossib to deduce the rational formula of Gibbs from the regressions. Thus Wilson chooses basically, $\dim(L_1) = 2$, L_1 spanned by temperature and pressure, and $\dim(L_2) = 1$, L_2 spanned by density. The regression statistics, only one canonical correlation can be computed, are given in the first row of table 11.2. For comparison purpose we have also computed the same linear regression after applying the transformatio suggested by the formula of Gibbs, they are given in the second row of 11.2. Of course this second analysis is quite useless, we want to show that our techniques can recover functional types, and this clearly cannot be done by explicitly building them into the regression. The transformations dictated by the formula of Gibbs are plotted in figure 11.2, observe that they are monotonic, and can actually be approximated quite well by linear functions.

To recover functional types we need higher dimensional subspaces. The Gibbs formula suggests that the model is additive (i.e. linear in transformed variables What happens if we do not constrain the transformations, and only require additivity ? Then

$$L_1 = \{u \in \mathbb{R}^n \mid u_i = \psi(x_i) + \phi(y_i)\},$$

$$L_2 = \{v \in \mathbb{R}^n \mid v_i = \eta(z_i)\}.$$

In this case, however, the subspaces have too many dimensions. Temperature has only nine different values, but pressure has 65 different values, and so has density. Thus L_1 has dimension $(9 - 1) + (65 - 1) = 72$ and L_2 has dimension 64. But the number of experiments is only 65, less than 72, and thus there are 64 canonical correlations equal to unity.

The first compromise is to group the observations on pressure and density in a fairly small number of classes, which means that we have to restrict our transformations to be step-functions. For purposes of comparison we have used a crude discretization and a fine discretization. The number of intervals in the crude discretization, with their marginals frequencies, is given in table 11.3.a.

Here $\dim(L_1) = 5$ and $\dim(L_2) = 2$. The transformations corresponding with the largest canonical correlation are plotted in figure 11.3, the regression statistics are in row three of table 11.2. The results are not satisfactory at all, the multiple correlation is very low, it is impossible to get an impression of the 'true' functional types from figure 11.3. In the fine discretization $\dim(L_1) = 17$ and $\dim(L_2) = 8$, the marginals are in table 11.3.b, the regression statistics in table 11.2, row four, and the transformations in figure 11.4. Because we know the Gibbs transformations it is possible to see that the step functions are beginning to bend in the 'correct' way, the regression statistics also become quite close to the 'rational' ones, but it still seems impossible to recover functional types from these results. In other words: in situations like these step-functions do not seem to work very well. It is true, of course, that we would get a more smooth plot if we used the midpoint on the interval against the transformed value, but out purpose here is to recover the transformation over the whole range.

We now compare these results with a crude and a fine basis of first-degree B-splines, differing in the number of knots. The column sums of the crude spline basis are in table 11.3.c, the regression statistics are in row five of table 11.2, and the transformations are in figure 11.5. The transformation follow the rational curves quite closely, although $\dim(L_1)$ is only 8 and $\dim(L_2)$ is only 4. An even better recovery is obtained from the fine splines, with marginals in 11.3.d, regression statistics in row six of table 11.2, and transformations in figure 11.6. Here $\dim(L_1) = 19$ and $\dim(L_2) = 10$, still a considerable saving compared with the maximum dimensionality 64, but indicating that the crude spline approximation is not improved considerably. The transformations in figure 11.6 are quite smooth, but they show some interesting dips, which could be studied in more detail (analysis of residuals). The results of our spline-analysis shows that Wilson is too pessimistic about statistical (or data analysis) methods. It is indeed possible to recover some of the important features of the rational transformations, the statistical results make it possible to exclude a number of functional types from further rational search, and the residuals point to outlying experiments which could be studied in more detail. If there was no theory at all, then these results could probably help in reaching 'the proper induction'. We must emphasize, however, that using social science standards implies that this example is too good to be true. In the social sciences there are no comparable rational theories which can be used as gauges, and the amount of error is usually much higher. This is also illustrated if we analyze the Gibbs-Wilson data with discrete ordinal options. Then $\dim(L_1) = 72$ and $\dim(L_2) = 64$, but we are saved from triviality by the monotonicity constraints. We do not give the transformations, but the solution is very similar to the fine spline solution we discussed. If we use <u>continuous</u> ordinal we get into trouble. Because temperature is measured at only

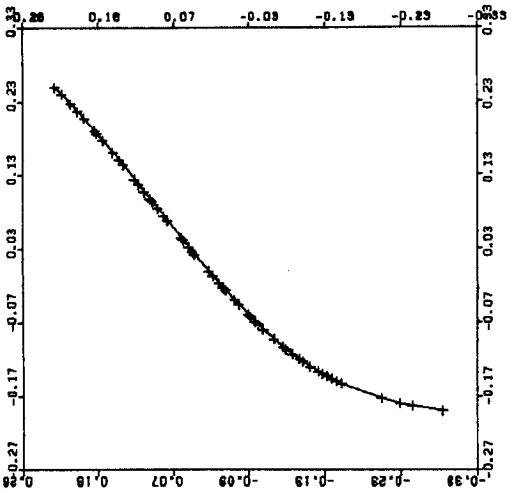
	r_{xy}	r_{xz}	r_{yz}	b_x	b_y	R^2
linear	-0.3804	0.8156	0.1534	1.0219	0.5421	0.9166
rational	-0.4008	0.8286	0.1663	1.0666	0.5938	0.9826
step 1	-0.2302	0.7248	0.2189	0.8186	0.4073	0.6825
step 2	-0.3847	0.8475	0.1406	1.0582	0.5477	0.9739
spline 1	-0.3891	0.8394	0.1663	1.0654	0.5809	0.9910
spline 2	-0.3884	0.8413	0.1655	1.0664	0.5797	0.9931

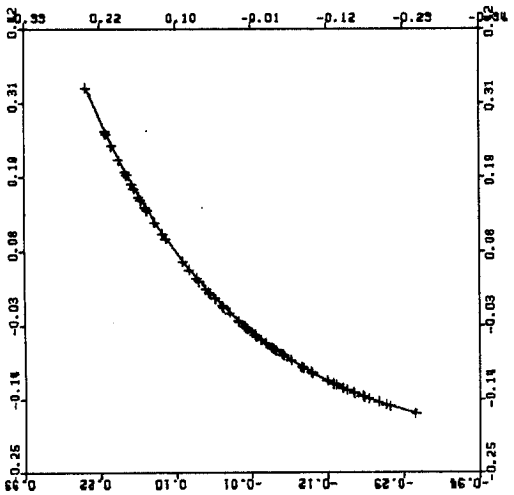
Table 11.2 Regression statistics for ten different analyses of Gibbs data.

a	temperature	16	39	10									
	pressure	34	17	13	1								
	density	28	21	16									
b	temperature	6	10	7	8	9	8	7	2	8			
	pressure	9	12	13	8	7	2	5	4	4	1		
	density	9	12	7	9	6	6	7	4	5			
c	temperature	0.2	16.2	35.0	12.8	0.8							
	pressure	2.9	29.4	19.2	11.0	2.6							
	density	3.4	24.7	21.1	14.0	1.8							
d	temperature	6.0	10.0	7.0	8.0	9.0	8.0	7.0	2.0	8.0			
	pressure	0.2	8.3	12.2	12.6	9.3	5.6	2.5	4.6	4.6	3.1	1.1	0.8
	density	1.5	7.2	11.0	8.4	8.4	6.3	6.3	6.6	4.6	4.0	0.7	

Table 11.3 Marginals for

a: step 1: step-functions crude.
b: step 2: step-functions fine.
c: spline 1: splines crude.
d: spline 2: splines fine.





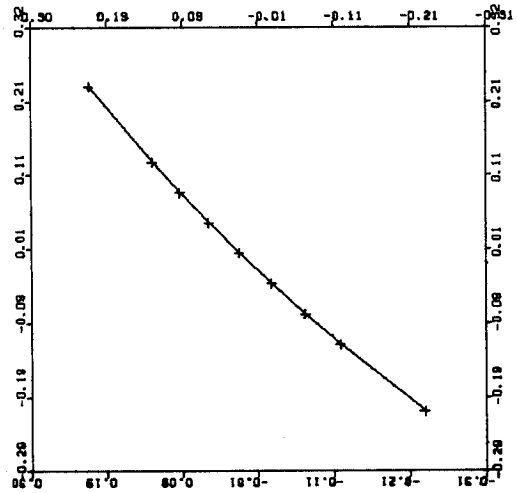
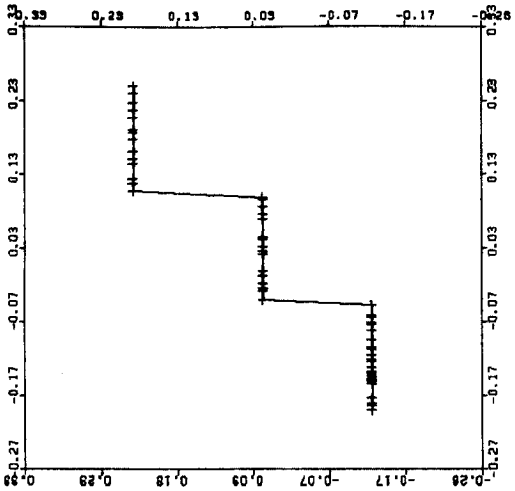
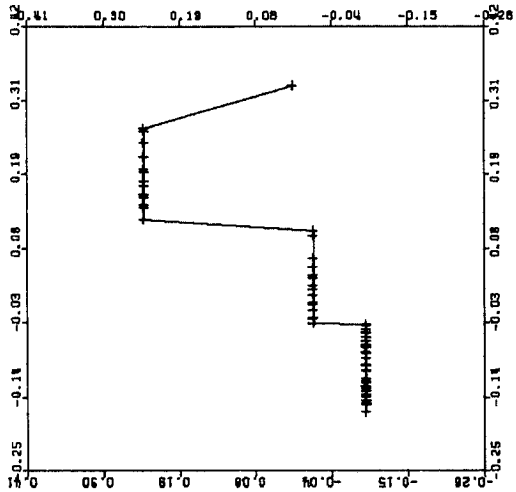


Figure 11.2 Gibbs-Wilson example rational transformations.





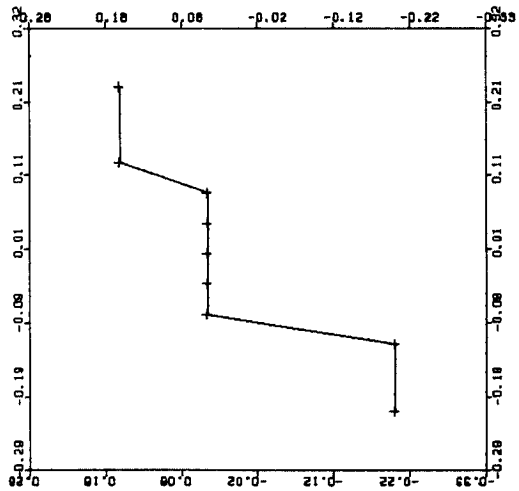
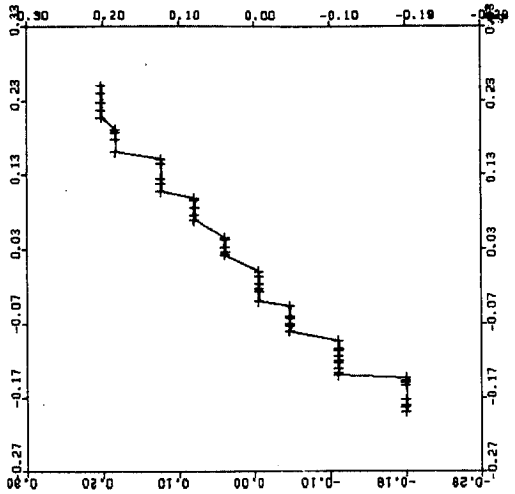
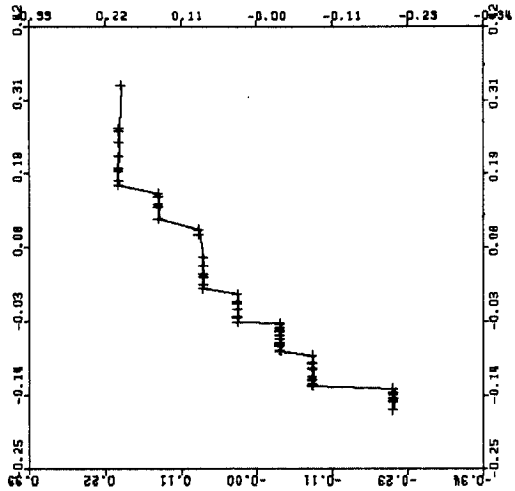


Figure 11.3 Gibbs-Wilson example crude step-functions.





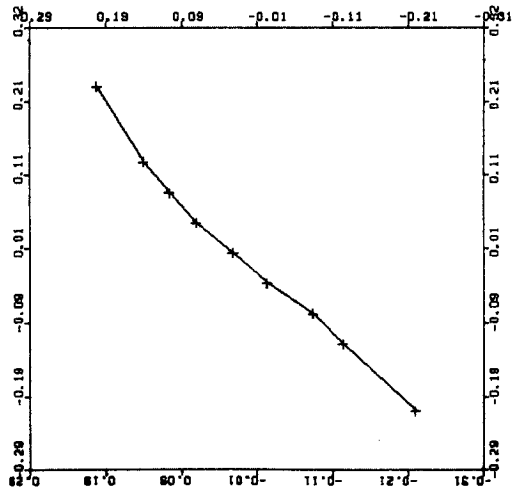
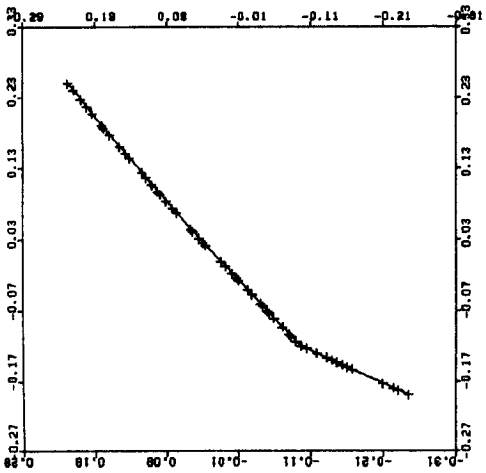
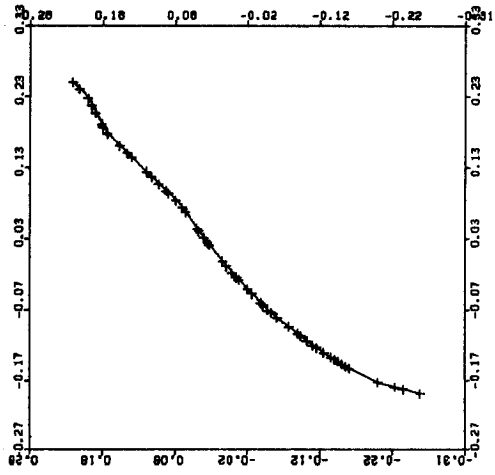
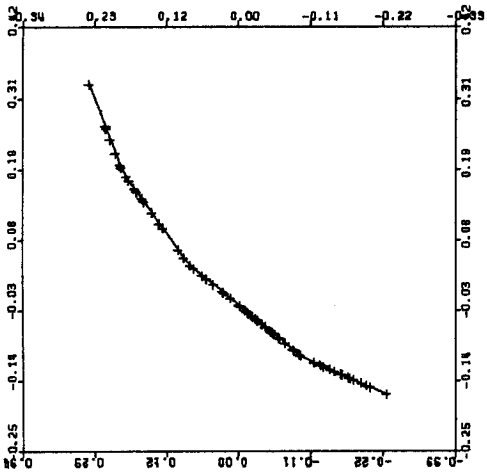
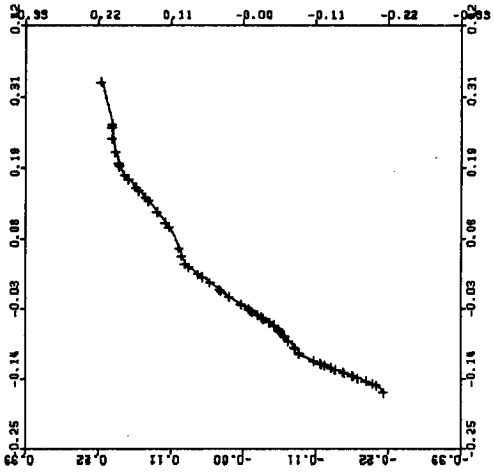


Figure 11.4 Gibbs-Wilson example fine step-functions.









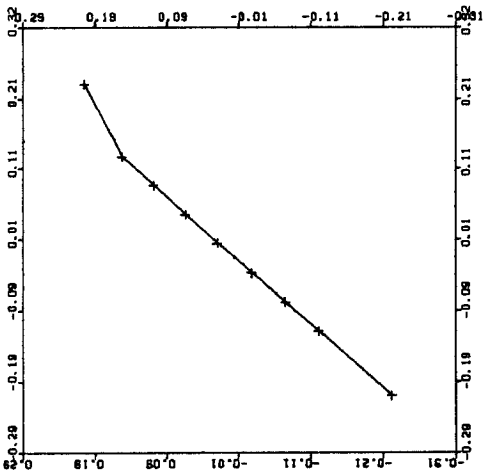


Figure 11.5 Gibbs-Wilson example crude splines.

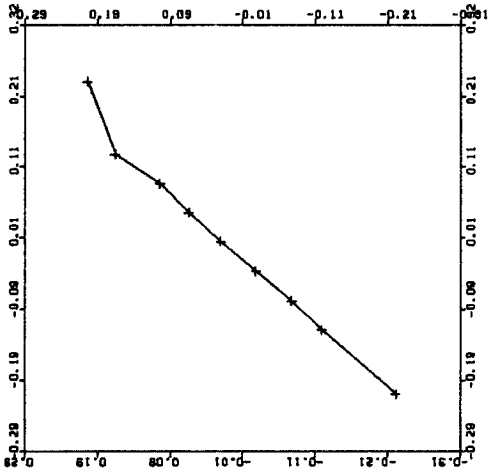


Figure 11.6 Gibbs-Wilson example fine splines.

nine levels the continuous ordinal option can use its possibility to untie ties, and produces a perfect monotone solution, and other perfect monotone solutions if the program is started from other starting points. Some of the transformations are quite like the ones found with the spline options, but in some perfect solutions are quite different from the spline transformations, and thus also from the rational ones. The fact that $\dim(L_1) = 123$ in this case puts too heavy a burden on the monotonicity constraints.

11.3.2 Interactive variables

We have already used interactive variables at a number of different places in this book. The examples in 4.4.4, 7.4.1, and 8.7. clearly show that interactive variables can be very useful for data analytic purposes. In this section we discuss a general technique to construct bases for interactive variables, given bases for the original variables. The technique is quite general, and in fact it is the correct way of interpreting most of discrete multivariate analysis in a linear algebra context. This is explained in Good (1963), Benzéri (1967), Haberman (1974). We have also used this technique in connection with the example in 7.4.1.

The technique uses the tensor product (also: direct product, Kronecker product) of linear spaces. A nice discussion of this product is given by Halmos (1948, appendix II), but many other books on linear and multilinear algebra have more extensive and more technical discussions. We first give a coordinate-free definition. Suppose L_1 and L_2 are subspaces of $\mathbb{R}^n$. Define $B(L_1,L_2)$ as the set of all bilinear functions on $L_1 \times L_2$ (here the symbol $\times$ is used for the Cartesian product, the set of all pairs). Then define $L_1 \bar{\times} L_2$, the tensor product of L_1 and L_2, as the algebraic dual of $B(L_1,L_2)$, i.e. the set of all linear functionals on $B(L_1,L_2)$.

For each $x \in L_1$ and $y \in L_2$ there is a $z = x \bar{\times} y$ in $L_1 \bar{\times} L_2$, defined by $z(u) = u(x,y)$ for all u in $B(L_1,L_2)$. Alternatively we can start with a basis $x_1,\ldots,x_r$ for L_1 and a basis $y_1,\ldots,y_s$ for L_2, and define $L_1 \bar{\times} L_2$ as the set of matrices that are linear combinations of the $n \times n$ matrices $x_i y_j'$. It is better to see this as an interpretation, the problem with matrices, as always, is that the same array of numbers can mean many things in many different contexts. From the definition, and from the interpretation, it is clear that the dimension of the direct product is the product of the dimension of the subspaces, and that the technique can be extended quite easily to more than two subspaces (the interpretation then is in terms of linear combinations of multiway arrays of rank one). Another basic results is that the $r \times s$ tensor products $x_i \bar{\times} y_j$ form a basis for $L_1 \bar{\times} L_2$. Moreover we can define an inner product in $L_1 \bar{\times} L_2$ by the rule

$$\langle x_1 \bar{\times} y_1, x_2 \bar{\times} y_2 \rangle = \langle x_1, x_2 \rangle \langle y_1, y_2 \rangle,$$

where $\langle x_1, x_2 \rangle$ and $\langle y_1, y_2 \rangle$ are the original inner products in L_1 and L_2. This shows that orthogonal bases in L_1 and L_2 give an orthogonal basis in $L_1 \bar{\times} L_2$.

The tensor product is the tool, we can now apply it in some contexts. In the simplest examples we have already replaced an $n \times k_1$ indicator matrix G_1 and another $n \times k_2$ indicator matrix G_2 by an $n \times (k_1 \times k_2)$ indicator matrix, which we can now write as $G_1 \overline{\times} G_2$. Interesting generalizations of homogeneity analysis are possible on this basis. In stead of requiring

$$x = G_j y_j$$

for all j, we can now require

$$x = G_j \overline{\times} G_\ell y_{j\ell},$$

for all j,ℓ. In stead of performing HOMALS with m variables we then perform HOMALS with $\binom{m}{2}$ product variables. If we have a basis of polynomials or splines for each j, then the direct product of these bases can be used as a basis for the direct product. We have done this, with orthogonal polynomials, in example 7.4.1. In this context we have to remember that if the polynomials are orthogonal with respect to the marginals of the variables, then the direct product of the polynomials is orthogonal with respect to the direct product of the marginals, which is generally not the same thing as the multivariate distribution of the variables.

11.3.3 Infinite dimensionality

Now suppose we replace $\mathbb{R}^n$ by an infinite-dimensional separable Hilbert space, for example the set of all random variables with finite variance on a given probability space. If the subspaces $L_1,\ldots,L_m$ are finite-dimensional, then nothing much changes. The bases for the subspaces still consist of a finite number of elements, the inner products of the elements in the bases can still be used to define finite matrices C and D, which define a finite eigenvalue-eigenvector problem. The definitions of indicator functions arising from discretization and of spline functions with a given knot sequence also do not change, of course polynomials are also easy to define.

We outline one special case, more or less as an excercise in notation. If $\underline{h}_1,\ldots,\underline{h}_m$ are the random variables on the probability space, then one-dimensional HOMALS wants to find transformations $\phi_1,\ldots,\phi_m$ and a random variable $\underline{x}$ such that

$$\sigma_M(\underline{x};\phi_1,\ldots,\phi_m) = \frac{1}{m}\sum_{j=1}^{m} \text{VAR}(\underline{x} - \phi_j(\underline{h}_j))$$

is minimized .., under the normalization condition $\text{AVE}(\underline{x}) = 0$ and $\text{VAR}(\underline{x}) = 1$. It is important to realize that this is a straightforward generalization of ordinary one-dimensional HOMALS, in which the random variables $\underline{h}_j$ are discrete and have k_j different possible values, and in which probability is counting measure. Just as in the previous chapters we can define

$$\sigma_M(\underline{x};*,\ldots,*) \triangleq \min \{\sigma_M(\underline{x};\phi_1,\ldots,\phi_m) \mid \phi_1,\ldots,\phi_m\}.$$

with $\phi_1,\ldots,\phi_m$ not constrained in any way. The formal solution to the problem of minimizing σ_M for fixed $\underline{x}$ is given by substituting the conditional expectations

$$\phi(\underline{h}_j) = \text{AVE}(\underline{x} \mid \underline{h}_j).$$

Since conditional expectation is projection on a subspace we have (appendix C) that

$$\text{VAR}(\underline{x}) = \text{VAR}(\underline{x} - \text{AVE}(\underline{x} \mid \underline{h}_j)) + \text{VAR}(\text{AVE}(\underline{x} \mid \underline{h}_j))$$

for all j, which implies that

$$\sigma_M(\underline{x};*,\ldots,*) = 1 - \frac{1}{m} \sum_{j=1}^{m} \text{VAR}(\text{AVE}(\underline{x} \mid \underline{h}_j)),$$

which shows that $\sigma_M(\underline{x};*,\ldots,*)$ is one minus the average correlation ratio of $\underline{x}$ and the $\underline{h}_j$ (remember that in the linear case principal components analsysis maximized the average correlation of $\underline{x}$ and the $\underline{h}_j$). A somewhat different notation is also convenient. If we write $\text{AVE}(\underline{x} \mid \underline{h}_j)$ as $P_j(\underline{x})$ to indicate that we project on the subspace of all functions of $\underline{h}_j$ with finite variance, then idempotency of the projector shows that

$$\sigma_M(\underline{x};*,\ldots,*) = 1 - \frac{1}{m} \sum_{j=1}^{m} \text{COV}(\underline{x},P_j(\underline{x})),$$

which indicates that maximizing over $\underline{x}$ under the condition $\text{VAR}(\underline{x}) = 1$ amounts to solving the eigen-problem for the operator

$$\frac{1}{m} \sum_{j=1}^{m} P_j.$$

Again this is a simple generalization of the finite dimensional case, in which we had $P_j = G_j(G_j'G_j)^{-1}G_j'$. This notational exercise can be generalized in the familiar directions. If we project on a cone in stead of a subspace, for example, the optimal $\phi(\underline{h}_j)$ is a generalized conditional expectation, and $P_j(\underline{x})$ is a nonlinear projector. If we compute more dimensions we have to choose again between single and multiple treatment of subspaces, if we have sets of variables we can use interactive coding and linear restrictions again.

An important practical problem is, of course, that it is impossible to carry out actual computations in infinite dimensional spaces. The matrices are too big. This is not a very serious problem from a theoretical point of view. Truly infinite dimensional problems are always gauges, because they do not only imply continuous variables but they also imply an inifinite number of observations. And if we have well-defined gauges the computations can often be carried out in the form of formulas. Suppose, for example, that the $\underline{h}_j$ are multinormal with mean zero, variance

one, and correlations $\rho_{j\ell}$. We now apply the alternating least squares algorithm for one-dimensional HOMALS theoretically. We start with a normally distributed $\underline{x}$ with $COR(\underline{x},\underline{h}_j) = \theta_j$. The first step of the ALS-algorithm computes the optimum ϕ_j for given $\underline{x}$. It gives $\phi_j(\underline{h}_j) = \theta_j\underline{h}_j$, because regressions in the multinormal case are linear. The second step computes a new optimal $\underline{x}$ for given ϕ_j, it makes the new $\underline{x}$ proportional to the average of the $\phi_j(\underline{h}_j)$. This implies that the correlation of the new $\underline{x}$ with the $\underline{h}_j$ is proportional to $R\theta$, where R is the correlation matrix of the $\underline{h}_j$. Thus the correlation between $\underline{x}$ and the $\underline{h}_j$ converges to the dominant eigenvector of R, and in the multinormal case the optimum nonlinear solution is the same as the optimum linear solution. We shall study gauges such as these in more detail in the next section, we use them here to illustrate that ALS can be carried out theoretically, and can be used to prove theorems. As we illustrated in chapter 9 it is also possible to prove ordinal properties of the eigenvector corresponding with the largest eigenvalue in this way.

For some of these theoretical computations orthogonal polynomials are quite useful. For a given probability distribution for which moments of all orders exist the orthogonal polynomials $\psi_0(\underline{x}),\psi_1(\underline{x}),\ldots$ are defined uniquely by the conditions that $\psi_s(\underline{x})$ is a polynomial of degree less than or equal to s, and $AVE(\psi_s(\underline{x})\psi_t(\underline{x})) = 0$ for $t < s$, while $AVE(\psi_s^2(\underline{x}))$ is one. The classical exponential distributions such as the normal, Poisson, gamma, binomial all have a classical set of orthgonal polynomials associated with them. We refer the reader to Tricomi (1955) for a nice treatment of orthogonal polynomials, there is also a fairly complete coverage of their statistical applications in Lancaster (1969). We illustrate how to construct $\psi_2(\underline{x})$ for the normal distribution directly. First observe that $\psi_0(\underline{x}) = 1$ and $\psi_1(\underline{x}) = \underline{x}$ if we assume that $\underline{x}$ is standard normal. Thus if $\psi_2(\underline{x}) = \alpha\underline{x}^2 + \beta\underline{x} + \gamma$, then $AVE(\psi_0(\underline{x})\psi_2(\underline{x})) = AVE(\psi_2(\underline{x})) = 0$ gives $\alpha + \gamma = 0$, while $AVE(\psi_1(\underline{x})\psi_2(\underline{x}))$ gives $\beta = 0$. Finally $AVE(\psi_2(\underline{x})) = 1$ gives $\alpha = \frac{1}{2}\sqrt{2}$. It is clear that this process of explicit orthogonalization can rapidly become tedious, but fortunately simple recursive formulas are available to compute $\psi_s(\underline{x})$ for all values of $\underline{x}$ and s. The same thing is true for the other classical orthogonal polynomials. If ϕ_1 and ϕ_2 are two functions with representation

$$\phi_1(\underline{x}) = \sum_{s=0}^{\infty} \alpha_s\psi_s(\underline{x}),$$

$$\phi_2(\underline{x}) = \sum_{s=0}^{\infty} \beta_s\psi_s(\underline{x}), \text{ then}$$

$$AVE(\phi_1(\underline{x})\phi_2(\underline{x})) = \sum_{s=0}^{\infty} \alpha_s\beta_s, \text{ and thus}$$

$$\mathrm{COV}(\phi_1(\underline{x})\phi_2(\underline{x})) = \sum_{s=1}^{\infty} \alpha_s\beta_s.$$

This result is useful because such a polynomial development is possible for all functions with finite variance. It also reduces problems with functions as arguments to problems involving infinite sums, which are easier to handle in many applications. Observe that the expansion of the covariance is not really what we are interested in, because we do not compute covariances between functions of the same variable, but covariances between functions of different variables. In such a case the polynomial expansion merely gives

$$\mathrm{COV}(\phi_1(\underline{x}_1)\phi_2(\underline{x}_2)) = \sum_{s=1}^{\infty}\sum_{t=1}^{\infty} \alpha_s\beta_t\,\mathrm{COV}(\psi_{s1}(\underline{x}_1),\psi_{s2}(\underline{x}_2)),$$

which is clearly far less simple.

It is also possible by basically the same methods to construct orthogonal polynomials on multivariate distributions. Thus we construct polynomials, say in m variables, such that

$$\mathrm{AVE}(\phi_{s_1s_2\ldots s_m}(\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_m)\phi_{t_1t_2\ldots t_m}(\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_m)) = \delta(s_1,t_1)\delta(s_2,t_2)\ldots\delta(s_m,t_m),$$

where $\delta(s,t)$ is the Kronecker delta. If such orthogonal polynomials are available it is possible to use them in the treatment of interactive variables, in which case it is not necessary to use the tensor product to construct bases. Recently Dahmen (1979,1980) has developed a system of multivariate B-splines which can be used in a similar way.

11.4 Some infinite dimensional gauges

Some of the results in this section are already discussed briefly, in a bivariate context, in section 4.3.8. There we mentioned Mehler's formula for the bivariate normal distribution, which implies in the notation of the previous section that

$$\mathrm{COV}(\psi_{s1}(\underline{x}_1),\psi_{t2}(\underline{x}_2)) = \delta(s,t)\rho_{12}^{s}.$$

Thus the covariance of two orthogonal polynomials on the standard bivariate normal distribution vanishes if the polynomials have different degree, and is equal to the ordinary correlation to the power s if they have the same degree s. Thus the covariance between arbitrary square integrable functions becomes

$$\mathrm{COV}(\phi_1(\underline{x}_1),\phi_2(\underline{x}_2)) = \sum_{s=1}^{\infty} \alpha_s\beta_s\rho_{12}^{s},$$

If we have m functions of m variables, with joint multinormal distribution, then

$$\phi_j(\underline{x}_j) = \sum_{s=0}^{\infty} \alpha_{js}\psi_s(\underline{x}_j)$$

can be used to derive an intersting result on the sum of the covariances. If $R^{(s)}$

is the matrix with correlations to the power s, and α_s is the m-vector with the α_{js} for fixed s, then

$$\sum_{j=1}^{m}\sum_{\ell=1}^{m} COV(\phi_j(\underline{x}_j),\phi_\ell(\underline{x}_\ell)) = \sum_{s=1}^{\infty} \alpha_s' R^{(s)} \alpha_s.$$

The sum of the variances is simply

$$\sum_{j=1}^{m} VAR(\phi_j(\underline{x}_j)) = \sum_{s=1}^{\infty} \alpha_s' \alpha_s.$$

This shows what the eigenvalueproblem that must be solved in homogeneity analysis amounts to in the case of the multinormal distribution. The eigenvalues are the m eigenvalues of $R^{(1)}$, the m eigenvalues of $R^{(2)}$, and so on. It follows from general results on Hadamard products (Styan, 1973) that the largest eigenvalue of all these eigenvalues is always the largest eigenvalue of $R^{(1)}$, which is the ordinary correlation matrix. Thus the largest eigenvalue has all transformations $\phi_j(\underline{x}_j)$ linear, a result proved by other techniques in the previous section. The second largest eigenvalue is more of a problem, however. It can be the second largest eigenvalue of $R^{(1)}$ or the largest eigenvalue of $R^{(2)}$. In the first case the transformations on the second dimensions are all linear too, in the second case they are all quadratic. The second place typically occurs if the largest eigenvalue of $R^{(1)}$ is much larger than the second one, because in that case the largest eigenvalue of $R^{(2)}$ is often larger than the second eigenvalue of $R^{(1)}$. In this case the plot of the first two dimensions shows a horse-shoe. Thus multinormal data in the inifinite dimensional generalization of HOMALS tend to give a horse-shoe. The implication is that approximately multinormal data, discreticized, and from a finite sample, will give an approximate horse-shoe.

We now illustrate this with a small example. Suppose we have four normal variates, with correlation matrix.

```
1 A B C
A 1 C B
B C 1 A
C B A 1
```

The eigenevectors of R are, columnwise,

```
+1 +1 +1 +1
+1 +1 -1 -1
+1 -1 +1 -1
+1 -1 -1 +1
```

and the corresponding eigenvalues are

$\lambda_1 = 1 + A + B + C,$

$\lambda_2 = 1 + A - B - C,$

$\lambda_3 = 1 - A + B - C,$

$\lambda_4 = 1 - A - B + C.$

This example is convenient because $R^{(s)}$ has the same form as R, and consequently also the same eigenvalues. The order of the eigenvalues depends, of course, on

the size of A, B, C. If we also look at the eigenvalues of $R^{(s)}$ then the order problem becomes quite complicated, even for this small example. We want to use this example to show that multinormal data do not necessarily give horse shoes. Thus we want the second eigenvalue of $R^{(1)}$ to be larger than the dominant eigen of $R^{(2)}$. We choose A, B, C in such a way that their difference is as large as possible. After some problem solving we find that this occurs for A = $\frac{1}{2}$ and B = for which the first two eigenvalues of $R^{(s)}$ are $1 + (\frac{1}{2})^s$ and the second two are $1 - (\frac{1}{2})^s$. Thus the second eigenvalue of $R^{(1)}$ is 1.5, the first eigenvalue of $R^{(2}$ is 1.25. We now draw a sample of 1000 individuals from this multinormal distribu and discreticize the continuous variables in four categories. If ψ_1 and ψ_2 are the normal orthogonal polynomials (the Hermite-Chebyshev polynomials), then we expect the first four eigenvectors of HOMALS to look row-wise like

$(+\psi_1 \ +\psi_1 \ +\psi_1 \ +\psi_1)$,
$(+\psi_1 \ +\psi_1 \ -\psi_1 \ -\psi_1)$,
$(+\psi_2 \ +\psi_2 \ +\psi_2 \ +\psi_2)$,
$(+\psi_2 \ +\psi_2 \ -\psi_2 \ -\psi_2)$,

with the first two eigenvalues equal to 1.50 and the second two to 1.25. The actual eigenvalues in the discreticized sample are 1.5311, 1.4306, 1.1807, 1.15 The eigenvectors are plotted against category number in figure 11.7, each row i an eigenvector,each column a variable. It is clear that the predicted pattern i there.

The statement that we find orthogonal polynomials if the data are approximately multinormal is much too weak. If the multivariate distrbution is a mixture of multivariate normals, then the eigenvectors are again the Hermite-Chebyshev polynomials, only the eigenvalues are more complicated functions of the various correlation matrices. If the data are multivariate log-normal (or more generall one-to-one nonlinear transformations of normal variates), then our technique finds the inverse transformation to normality and then finds the Hermite-Chebyshev polynomials again. The marginals of the transformed variables can be anything, even rectangular. The work of Lancaster (1969), Eagleson (1969), Griffiths (1969, 1970), Tyan and Thomas (1975) shows that there are many other classical bivariate distributions with orthogonal polynomials as eigenvectors. They often have the same solution structure as the multinormal, although the relation with the matrices $R^{(s)}$ is not necessarily true. The work of Karlin (1964, 1968) shows that oscillatory eigenvectors,which cannot be distinguished from orthogonal polynomials in discrete situations, often occur because of the general condition of total positivity. As a consequence we can make the somewha stronger statement that HOMALS tends to find orthogonal polynomial transformati whenever it is used as a variable transformation technique on data which intend to measure a single scale. Thus the horse shoe is the HOMALS equivalent of the

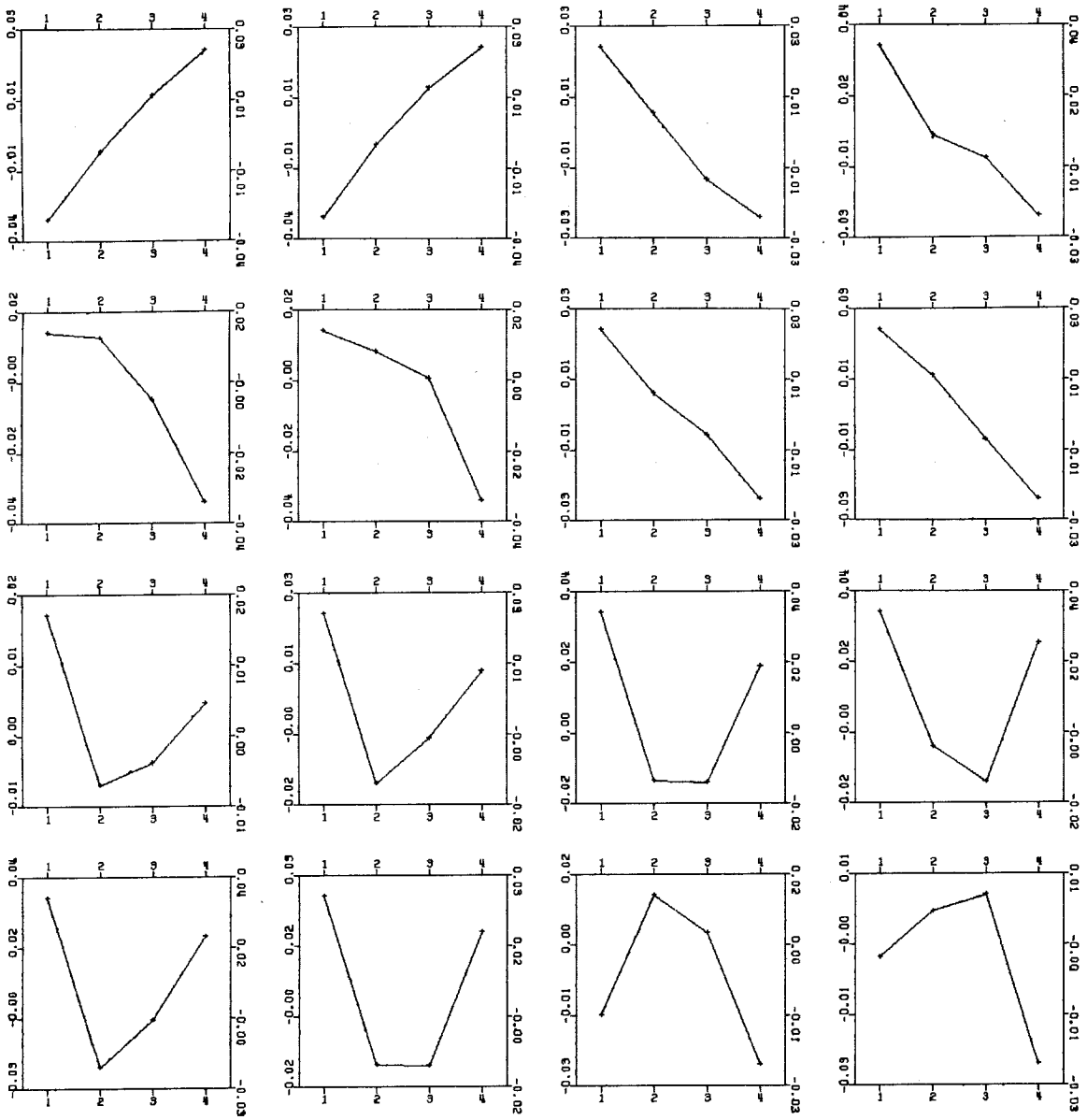


Figure 11.7 Four eigenvectors (rows), four variables (columns).

general factor. In many cases, however, HOMALS is used as a scaling technique on data which are not a random sample on a set of m variables, on data in which some of the variables are purely nominal, or on data in which there are positive and negative and high and low correlations. In that case we do not find the familiar quadratic pattern.

The multinormal distribution can gauge HOMALS, but can it also be used to gauge CANALS and PRINCALS ? We start with OVERALS and use the multivariate version of the Hermite-Chebyshev polynomials discussed by Appell and Kampé de Fériet (1926), Erdelyi (1953), and used in this context by Venter (1966) and Dauxois and Pousse (1976). For these polynomials it can be proved that

$$COV(\psi_{s_1\ldots s_m}(\underline{x}_1,\ldots,\underline{x}_m),\psi_{t_1\ldots t_m}(\underline{y}_1,\ldots,\underline{y}_m)) = \delta(s_1,t_1)\ldots\delta(s_m,t_m)\rho_1^{s_1}\ldots\rho_m^{s_m},$$

where $\rho_1 \geq \ldots \geq \rho_m$ are the canonical correlations (we assume for the moment that both sets contain an equal number of variables, merely to simplify notation). This implies

$$COV(\phi_j(\underline{x}_j),\phi_\ell(\underline{x}_\ell)) = \sum_s \alpha(s)'R(s)\alpha(s),$$

In this equation the $\underline{x}_j$ are vectors of random variables, the index s is a vector $(s_1,\ldots,s_m)$, and for each value of this index there is a 2×2 matrix R(s) with element off the diagonal

$$(R(s))_{12} = \rho_1^{s_1}\ldots\rho_m^{s_m}.$$

where the ρ_s are the canonical correlations. This shows in the same way that CANALS with all variables multiple nominal amounts to solving eigenvalue problems for the matrices R(s) separately, the same way as HOMALS. Ordering the eigenvalues of the various matrices is even more complicated in this case, although it is easy to see that the largest eigenvalue is the largest eigenvalue of R(1,0,...,0) the matrix with the off diagonal element equal to the largest canonical correlation, which again corresponds to linear transformations of all variables. If there are more than two sets of variables the situation becomes much more complicated again.

For single variables additional complications are introduced too. We have seen that HOMALS on the multivariate normal can select its successive solutions from the various $R^{(s)}$, depending on the relative size of the eigenvalues of these matrices. PRINCALS computes a single solution (transformation), and the single restriction forces PRINCALS to select its p dimensions from a single $R^{(s)}$. It may be possible to construct a correlation matrix for which the sum of the p largest eigenvalues of $R^{(2)}$ is larger than the sum of the p largest eigenvalues of $R = R^{(1)}$, but we have not found one yet. Generally PRINCALS selects its p dimensions from $R^{(1)}$, which means that using PRINCALS in the multinormal (and

similar) cases should give the same solution as using HOMALS 'as a first step', followed by linear principal components analysis 'as a second step'. A similar result is true for CANALS with single variables. We emphasize, however, that we have not proved that the linear solution gives the maximum, we merely have proved that the linear solution (and all other solutions based on the orthogonal polynomials) give stationary values of the loss function. The proof is simply by substitution in the stationary equations and is not given here.

11.5 Analysis of stochastic processes

Our components analysis techniques can also be used for the analysis of stochastic process data. This does not add anything new for discrete time, the time-points t_1, t_2, ... define variables, in practical data analysis problems there is only a finite number of time points, and consequently a finite number of variables. For continuous time the situation is considerably more interesting. For details we refer the reader to the admirable paper by Deville and Saporta (1980). We merely emphasize the formal similarity with HOMALS. The eigen equation we have to solve is

$$\int_T D_t^{-1} C_{ts} y_s ds = \lambda y_t,$$

with $D_t = G_t'G_t$ and $C_{ts} = G_t'G_s$, and with G_t the indicator matrix (or the basis of indicator functions, polynomials, or splines) at time t. Thus we now have an infinite number of variables, for which our basic join and meet philosophy does not necessarily apply without further regularity conditions on the subspaces and operators. In full generality these conditions are treated by Dauxois and Pousse (1976), for the particular application we discuss here they simplify considerably and are discussed by Deville and Saporta (1980). The simplifications are due to the fact that we deal with a discrete and finite state space here.

Deville and Saporta also discuss data analysis aspects of continuous time. Again we have to discreticize time in some way or another to get finite matrices. The first idea is to select a point in each of a finite number of time intervals, and use the indicator matrix of this time point, weighted with the length of the time interval, in a correspondence analysis. A better alternative is to use generalized indicator matrices, in which element $g_{i\ell}^j$ indicates the percentage of time spend by individual i in state ℓ in time interval j. Thus the G_j are not binary, but their rows still add up to one. Again the generalized indicators are weighted by the length of the time interval, and subjected to correspondence analysis. The technique seems very interesting, but we do not have practical experience with it and we cannot show an example.

Deville and Saporta do not mention any gauges for their techniques, but quite a number of interesting results are available. In the same way as in the previous section gauges can be derived from the canonical decomposition of bivariate

distributions, in this case from the transition probabilities of symmetric stationary Markov processes with a continuous time parameter. For continuous state space this was done by Wong and Thomas (1962) and Sarmanov (1963), for birth and death processes with a discrete state space by Eagleson (1969). These authors investigate in detail the conditions which guarantee that the canonical variables are orthogonal polynomials, with results similar to those obtained in the static case. In a very interesting recent paper of Cooper, Hoare, and Rahman (1977) the transition probabilities of a class of Markov chains are analyzed. The canonical variables in this case are discrete orthogonal polynomials, by letting converge various paparmeters to limiting values we obtain continuous state processes with continuous orthogonal polynomials as canonical components.

11.6 Alternative criteria

11.6.1 Correlation matrices and their eigenvalues

In chapter 6 we have discussed a number of different criteria that could be optimized in the case of metric K-set canonical analysis. We can now formulate the same problem more generally. Suppose $\underline{x}_1,\ldots,\underline{x}_m$ are m vector valued random variables, which can be discrete or continuous and defined on a finite or infinite probability space. We study the transforms $\underline{q}_j = \phi_j(\underline{x}_j)$, where all kinds of restrictions can be imposed on the ϕ_j (they could be additive, additive univariate splines, additive and monotonic, and so on). We agree to choose the ϕ_j from feasible sets Φ_j in such a way that a function of the correlations $COR(\underline{q}_j,\underline{q}_\ell)$ is maximized or minimized. As explained in previous chapt this formulation covers HOMALS (maximize the largest eigenvalue of the correlatio matrix, compute more than one solution), it covers PRINCALS (maximize the sum of the p largest eigenvalues), it covers CANALS (maximize, for a given partition into two sets, the sum of the first p canonical correlations), and it covers OVERALS (maximize, for a given partition into K sets, the sum of the first p generalized canonical correlations). It also covers the alternative criteria mentioned in chapter 6.

It is interesting to find a more general class of criteria for which some results can be shown to be true without specifying to some particular criterium in the class. This is possible if we specify that the function of the correlation matrix R we are trying to maximize is a unitary invariant norm, that is a matrix norm with the additional property that $K'RK$ has the same norm as R for all orthogonal K, and not necessarily with the property that the norm of a product is less than or equal to the product of the norms of the matrices in the product. Thus if $\omega(.)$ is our criterion, then

$R \neq 0$ implies $\omega(R) > 0$,

$\omega(aR) = |\alpha|\omega(R)$,

$\omega(R_1 + R_2) \leq \omega(R_1) + \omega(R_2)$,

$\omega(K'RK) = \omega(R)$.

The sum of the p largest eigenvalues of R is a particular case of a unitarily invariant norm, it fact it is quite a special case. A theorem by Ky Fan (1951) says that $\omega(R_1) \gtrsim \omega(R_2)$ for all unitarily invariant norms if and only if $\lambda_1(R_1) \gtrsim \lambda_1(R_2)$ and $\lambda_1(R_1) + \lambda_2(R_1) \gtrsim \lambda_1(R_2) + \lambda_2(R_2)$ and ... and $\lambda_1(R_1) + \ldots + \lambda_{m-1}(R_1) \gtrsim \lambda_1(R_2) + \ldots + \lambda_{m-1}(R_2)$. Thus a set of transformations is optimal for all unitarily invariant norms if and only if it is optimal for PRINCALS in 1,2,...,m-1 dimensions.

Now suppose that R has a representation

$$r_{j\ell} = \sum_{s=1}^{\infty} \alpha_{js}\alpha_{\ell s}c_{j\ell s},$$

or, in matrix notation,

$$R = \sum_{s=1}^{\infty} A_s C_s A_s,$$

with A_s diagonal, and

$$\sum_{s=1}^{\infty} A_s^2 = I.$$

For unitarily invariant norms it is possible to prove that

$$\omega(R) \lesssim \max_{s=1}^{\infty} \omega(C_s),$$

and equality is possible by choosing $A_s = I$ for one s and $A_s = 0$ for all other s. The same type of inequality can be proved for other matrix norms which do not necessarily satisfy $\omega(K'RK) = \omega(R)$, but which satisfy $\omega(R_1R_2) \lesssim \omega(R_1)\omega(R_2)$. The inequality implies that we can investigate all C_s separately, and keep the one with the largest $\omega(C_s)$. In the special case of the multinormal distribution $C_s = R^{(s)}$. We know that if we select $\omega(C_s) = \lambda_1(C_s)$, then the maximum of $\omega(R)$ is $\lambda_1(R^{(1)})$. In PRINCALS we have assumed that 'usually' the sum of the p largest eigenvalues of R is also maximized for $R^{(1)}$, but we have not proved this. Oppenheim's inequality (Styan, 1973, Beckenbach and Bellman, 1965, p 71) can be used to show that for $\omega(R) = (\det(R))^{1/m}$ we also only have to look at the $\det(C_s)$ individually, and that if $C_s = R^{(s)}$, then $\det(R^{(1)})$ is the smallest of the determinants $\det(R^{(s)})$. Thus the determinant criterion in chapter 6, which had to be minimized, gives linear transformations in the multinormal case too.

11.6.2 Least squares or other

We have already given our reasons to use least squares in many places, either explicitly or implicitly by emphasizing pictures, inner products, or multinormal gauges. Currently popular alternatives to least squares are motivated statistically, often by robustness considerations. As explained earlier we do not think that statistical gauges are necessarily the best way to select data analysis techniques, the situation is often the other way around. On the other hand if we have definite

reasons to take some gauge seriously, then other methods than least squares certainly have to be taken into account. We do not discuss them in detail, we merely indicate how some of our techniques can be generalized. Suppose, for example, that we want to minimize

$$\sigma(x;y_1,\ldots,y_m) = \frac{1}{m}\sum_{j=1}^{m}\phi(x - G_j y_j),$$

where x must satisfy the normalization condition $\psi(x) = 1$. In this formulation the functions $\phi(\ )$ and $\psi(\ )$ are now assumed to be positively homogeneous and convex. Algorithms to minimize functions of this sort are derived easily from the work of Robert (1967) or Boyd (1974), they have been used in a somewhat different context by De Leeuw (1977). The extension to random variables is easy enough, the extension to multidimensional solutions is somewhat more complicated because a natural definition of orthogonality is not necessarily available. We do believe, however, that it is essentially trivial to translate most of our theory into normed linear space terminology. Certainly the concept of join and meet are independent of any norm, and they can be translated into loss function terminology by using any norm.

12 Stability of multivariate analysis methods

12.1 Introduction

In this chapter we give an overview of most of the techniques that can be used to analyze stability, we give some applications to nonlinear multivariate analysis, and we discuss some possible research projects in connection with stability. It is a somewhat unfortunate situation that all results are collected in one chapter, we have made it quite clear in chapter one that data analysis techniques are incomplete without an extensive analysis of stability. Consequently it would be more appropriate to discuss the stability results in the chapters in which the various techniques are introduced. Our results, however, are still rather fragmentary, and more easily decribed as applications of general techniques for the investigation of stability than as substantial contributions to the analysis of stability of particular data analysis techniques.

It is true, however, that we have already used some general techniques to assess stability in an informal way in earlier chapters. Some examples have been analyzed with various different methods, for example the Roskam journal preference data in chapters 5 and 10. More results about 'stability under method selection' will be discussed in the next chapter. Choice of measurement level of the variables, both between single and multiple, and between numerical, ordinal, and nominal, has also been investigated in various places in the previous chapters. It is a particular case of 'stability under model selection', which has been investigated more closely in chapters 7 and 8. The general idea here is that if we analyze data by various methods, 'assuming' various models, and using different techniques, then it will be possible to separate properties of the data from those of the models and techniques. This is an old idea, which also underlies the idea of experimental design. If we manipulate the factors in the design (data, model, technique, and possibly others such as algorithm, loss function, gauge), and we do this systematically and with as little bias as possible, then we can isolate the effects of all factors and their possible interactions. This is obviously an idealized statement, the selection from possible data, techniques, models is largely subjective without any clear cut rules. The approach is, however, considerably more realistic than the one which assumes that one particular model is 'true', that one particular principle to construct loss functions is always optimal or very close to optimal, and that consequently there also is a unique technique which should be used. It seems to us that this approach ignores the data. Data analysis is not a mechnical process, although processes such as LISREL, ALSCAL, MULTISCAL and HOMALS are often used in such a way. Almost by definition this is not the fault of the users, at least sometimes the people to construct, produce, and sell these processes promote or do not

discourage these practices.

12.2 Analytical stability

12.2.1 Eigenvalue problems

We first study the problem of solving

$$AX = BX\Lambda,$$

$$X'BX = I.$$

The matrices A and B are given real symmetric matrices of order m, we assume that B is positive definite. The system must be solved for the matrix X, also of dimension $m \times m$, and for the $m \times m$ diagonal matrix Λ. The problem solved by OVERALS with all variables either numerical or multiple nominal is of this type, and consequently (for the same types of variables) CANALS, PRINCALS, HOMALS, and ANACOR are special cases.

Analytical stability in this context studies the following problem. Suppose the matrices A and B depend on a number of parameters, thus they are not fixed matrices, but matrix valued functions defined on a parameter space. Since the solution to the generalized eigenvalue problem is usually unique (up to permutation of dimensions) these problems also define X and Λ as matrix valued and diagonal-matrix valued functions of the parameters. We now want to know what the effect of a small change in the parameters is on the value of X and Λ. We proceed formally in this section, assuming that A and B are differentiable functions of the parameters, and by differentiating our functions in the neighborhood of a parameter-point where all elements of Λ are different. The justification of our formal procedure is due to Rellich, who published a series of five papers on the subject in the Mathematische Annalen between 1937 and 1942. All the necessary results, and much more than that, are reviewed in the book of Kato (1976).

For the moment we assume that there is just one parameter, and we write the partial derivatives of a matrix with respect to that parameter by using the symbol for the matrix with a dot above it. Thus differentiating the two equations defining X and Λ gives

$$\mathring{A}X + A\mathring{X} = BX\mathring{\Lambda} + B\mathring{X}\Lambda + \mathring{B}X\Lambda,$$

$$X'B\mathring{X} + \mathring{X}'BX + X'\mathring{B}X = 0.$$

If we premultiply the first of these by X', and simplify, we find

$$X'\mathring{A}X - X'\mathring{B}X\Lambda = \mathring{\Lambda} + (X'B\mathring{X}\Lambda - \Lambda X'B\mathring{X}).$$

The matrix in parenthesis on the right has a zero diagonal. If we consider diagonal terms only we find

$$\dot{\Lambda} = \text{diag}(X'\dot{A}X - X'\dot{B}X\Lambda).$$

If we define S implicitly by $\dot{X} = XS$, then it follows that

$$S + S' = -X'\dot{B}X,$$

or

$$\text{diag}(S) = -\tfrac{1}{2}\,\text{diag}(X'\dot{B}X).$$

Also

$$X'\dot{A}X - X'\dot{B}X\Lambda = \dot{\Lambda} + (S\Lambda - \Lambda S),$$

or

$$\text{nondiag}(X'\dot{A}X - X'\dot{B}X\Lambda) = S\Lambda - \Lambda S,$$

where the nondiag(.) of a matrix is the matrix minus its diagonal. In more familiar notation the result for $\dot{\Lambda}$ shows that

$$\frac{\partial \lambda_s}{\partial \theta_k} = \sum_{i=1}^{m}\sum_{j=1}^{m} x_{is}x_{js}\left(\frac{\partial a_{ij}}{\partial \theta_k} - \lambda_s \frac{\partial b_{ij}}{\partial \theta_k}\right).$$

The corresponding results for $\dot{X}$ are considerably more complicated. We first define

$$q_{st}^k \triangleq \sum_{i=1}^{m}\sum_{j=1}^{m} x_{is}x_{jt}\frac{\partial a_{ij}}{\partial \theta_k},$$

$$r_{st}^k \triangleq \sum_{i=1}^{m}\sum_{j=1}^{m} x_{is}x_{jt}\frac{\partial b_{ij}}{\partial \theta_k}.$$

With this notation we find for the partials

$$\frac{\partial x_{js}}{\partial \theta_k} = \sum_{t\neq s}^{m} x_{jt}\frac{q_{ts}^k - \lambda_t r_{ts}^k}{\lambda_s - \lambda_t} - \tfrac{1}{2}x_{js}r_{ss}^k \qquad \text{and} \qquad \frac{\partial \lambda_s}{\partial \theta_k} = q_{ss}^k - \lambda_s r_{ss}^k.$$

These formulas are not very interesting in themselves, but it is nice to have them around. By differentiating them again we can find formulas for the second partials, but these formulas look downright horrible, and we do not reproduce them here. In some interesting special cases the formulas simplify a lot. A familiar special case, for example, arises when B is a constant matrix. Then $\dot{B} = 0$, and all formulas simplify. Another interesting special case is the one in which both A and B are linear combinations of matrices A_k and B_k. We shall use some of the resulting simplifications in our examples. Computationally it is interesting that the formula for the partials of eigenvalue s only involve eigenvalue and eigenvector s, while the partials for eigenvector s involve all eigenvalues and eigenvectors. This last results seems inconvenient in data analysis techniques which compute only a few (often one or two) of the largest eigenvalues and their corresponding eigenvalues. It is, however,

possible to repair this by using the formulas

$$\dot{x}_s = -(A - \lambda_s B)^+(\dot{A} - \lambda_s \dot{B} - \dot{\lambda}_s B)x_s - \eta_s x_s,$$

with

$$\dot{\lambda}_s = x_s'(\dot{A} - \lambda_s \dot{B})x_s,$$

$$\eta_s \triangleq \tfrac{1}{2} x_s' \dot{B} x_s.$$

It is possible to use these formulas in combination with various iterative schemes In general it seems true, however, that the best approach to computing the partials depends on the particular application we have in mind.

12.2.2 Some simple applications

12.2.2.1 Eliminating a variable in HOMALS

One way to formulate the problem of homogeneity analysis is solving the eigenproblem

$$P_* X = M_* X \Lambda,$$

$$X' M_* X = I,$$

with

$$P_* = \frac{1}{m} \sum_{j=1}^{m} P_j,$$

$$M_* = \frac{1}{m} \sum_{j=1}^{m} M_j,$$

$$P_j = M_j G_j (G_j' M_j G_j)^+ G_j' M_j.$$

Now suppose we do another homogeneity analysis, on the same data but with the first variable eliminated. The matrix P_* changes to

$$P_* + \frac{1}{m-1}(P_* - P_1),$$

and M_* changes to

$$M_* + \frac{1}{m-1}(M_* - M_1).$$

For general perturbations of the form $P_* + \varepsilon(P_* - P_1)$ and $M_* + \varepsilon(M_* - M_1)$ we find, differentiating with respect to ε,

$$\lambda_s(\varepsilon) = \lambda_s(0) - \varepsilon\, x_s'(P_1 - \lambda_s M_1)x_s + o(\varepsilon).$$

Thus, for large m, the eigenvalue s of the HOMALS solution without variable one will be approximately equal to

$$\lambda_s - \frac{1}{m-1} x_s'(P_1 - \lambda_s M_1)x_s.$$

This suggest that the importance of a variable for a dimension can be defined (or approcimated) by the quantity $x_s'(P_j - \lambda_s M_j)x_s$. Observe that the average

over j of these quantities is zero. If $M_j = I$ for all j, then they are equal to $x_s'P_jx_s - \lambda_s$, which is the discrimination measure of variable j on dimension s minus the average of the discrimination measures of all variables on dimension s (which is equal to the eigenvalue λ_s). If we eliminate a variable with a small discrimination measure the eigenvalues will not change much.

Results for eigenvectors are far less satisfactory in many ways. It the first place they are necessarily more complicated, in the second place the terms $(\lambda_s - \lambda_t)^{-1}$ suggest that if eigenvalues are close, then small perturbations can have large effects on eigenvectors. Another important component in eigenvector perturbation are the off-diagonal elements q_{st}^k and r_{st}^k. If these are approximately zero, then small perturbations of the matrices will have little effect on the eigenvectors. In our example with a variable deleted from HOMALS, with in addition $M_j = I$ for all j, eigenvector s does not change much if the off-diagonal elements in row and column s of $X'P_1X$ are small. This indicates that in homogeneity analysis the discrimination measures, on the diagonal of $X'P_jX$, give interesting information on the influence of the variables on the eigenvalues, the off-diagonal elements give information about the influence on the variables on the eigenvectors.

It is not difficult to see that if we add a variable, with corresponding projector P_o, then P_* changes to

$$P_* - \frac{1}{m+1}(P_* - P_o),$$

and we can apply essentially the same results in the same way.

12.2.2.2 Merging categories in HOMALS

Merging categories can be formalized algebraically as using indicator matrices H_j, of dimension $k_j \times \ell_j$, with $\ell_j \leq k_j$, to replace G_j by G_jH_j. If $k_j = \ell_j$, then it follows that $H_j = I$, and nothing changes for this variable. We assume that $M_j = I$ for all j to avoid unnecessary and inessential complications. P_j is replaced by

$$\tilde{P}_j \triangleq G_jH_j(H_j'D_jH_j)^+H_j'G_j',$$

and the first order approximation to the new eigenvalues is

$$\tilde{\Lambda} = \Lambda + \text{diag}(X'(\tilde{P}_* - P_*)X) + o(||\tilde{P}_* - P_*||).$$

Again many possible applications can be studied, the simplest one is merging the first two categories of the first variable. Suppose the category frequencies of these categories are d_1 and d_2, and suppose the category quantifications are y_{1s} and y_{2s}. Then

$$\tilde{\lambda}_s = \lambda_s - \frac{1}{m}\frac{d_1d_2}{d_1+d_2}(y_{1s} - y_{2s})^2 + o(\tfrac{1}{m}).$$

Merging two categories with approximately the same quantifications hardly changes the eigenvalues.

12.2.2.3 Eliminating a subject in HOMALS

In this case it is better to start from the generalized eigenvalue problem

$$CY = mDY\Lambda,$$

where C contains bivariate marginals and D contains univariate marginals. C is of the form

$$C = \frac{1}{n}\sum_{i=1}^{n} t_i t_i',$$

where the t_i are vectors with $\sum k_j$ elements, composed of the rows of m indicators matrices. Moreover D = diag(C). If we eliminate the first subject, we find a new C, related to the old C by

$$\tilde{C} = C + \frac{1}{n-1}(C - t_1 t_1').$$

If we let $T_1 = \text{diag}(t_1 t_1')$, then

$$\tilde{D} = D + \frac{1}{n-1}(D - T_1).$$

It follows from our general results that, with normalization $y_s' D y_t = \delta^{st}$,

$$\tilde{m\lambda}_s = m\lambda_s - \frac{1}{n-1}((y_s' t_1)^2 - m\lambda_s y_s' T_1 y_s) + o((n-1)^{-1}).$$

If we use the stationary equation $Gy_s = m\lambda_s x_s$, which is satisfied in HOMALS with the usual normalization, then $y_s' t_1 = m\lambda_s x_{1s}$, and thus

$$\tilde{\lambda}_s/\lambda_s = 1 - \frac{1}{n-1}(m\lambda_s x_{1s}^2 - y_s' T_1 y_s) + o((n-1)^{-1}).$$

The quantities $m\lambda_s x_{is}^2 - y_s' T_i y_s$ can be interpreted as the importance of a subject for a dimension, analogous to the discrimination measures. For each s these quantities sum to zero. Our general theory applies directly if we use the normalizations $y_s' D y_s = 1$, which implies together with $Gy_s = m\lambda_s x_s$ and $G'x_s = Dy_s$ that $x_s' x_s = (m\lambda_s)^{-1}$. But of course we can renormalize x_s and y_s in other ways, and adjust the equations correspondingly. Again the same approach can be used if we add a subject. More generally we can define C as

$$C = \sum_{i=1}^{n} p_i t_i t_i',$$

in which each 'profile' t_i has a relative frequency p_i. We then study

$$C(\varepsilon) = \sum_{i=1}^{n}((1-\varepsilon)p_i + \varepsilon\delta^{i1})t_i t_i' = C - \varepsilon(C - t_1 t_1')$$

with the same methods as before. Perturbations of this form are studied in classical asymptotic statistics, and in jackknife-type methods. We shall return to them later in the chapter.

12.2.3 Problems which are not eigenvalue problems

The use of restrictions, notably the use of cone restrictions and the use of single variables, often leads to problems which are not eigenvalue problems for a given fixed matrix, but eigenvalueproblems for a matrix which depends on the quantifications. For applications of this sort we need more general approaches to compute partial derivatives. We do not study problems of this kind in any detail, we simply indicate some possible approaches and we list some of the possible mathematical tools.

Suppose a loss function $\sigma(\gamma;\theta)$ is a function of two sets of parameters. The first set are the parameters we are interested in, the second set are the perturbation parameters. Define

$$\sigma(*;\theta) \triangleq \min \{\sigma(\gamma;\theta) \mid \gamma \in \Gamma\},$$

where we assume that the problem is defined in such a way that the minimum is attained for all $\theta \in \Theta$, the set of perturbations we are interested in. We also define

$$\Omega(\theta) \triangleq \{\gamma \in \Gamma \mid \sigma(\gamma;\theta) = \sigma(*;\theta)\}.$$

Thus $\Omega(\theta)$ is the set of minimizers for perturbation θ, this is a point-to-set map because the minimizer is not necessarily unique. In recent years there has been an enormous amount of research in the mathematical programming and convex analysis literature on theorems which guarantee that the functions $\sigma(*;\theta)$ and $\Omega(\theta)$ are continuous or even differentiable. We do not intend to review this literature here, we merely want to point out that it is extremely useful in many problems of psychometrics, data analysis, and statistics, and that some of the key references are Demjanov and Malozemov (1974), Hogan (1973a, 1973b), Hiriart-Urruty (1978), Springarn (1980). The simplest and in many respects most useful result is that if $g(\gamma,\theta)$ is the vector of partial derivatives of σ with respect to the perturbation parameters, evaluated at $\gamma \in \Gamma$ and $\theta \in \Theta$ then

$$\sigma(*,\theta + \varepsilon\delta) = \sigma(*;\theta) + \varepsilon \min \{ \delta' g(\gamma,\theta) \mid \gamma \in \Omega(\theta)\} + o(\varepsilon), (\varepsilon > 0),$$

uniformly in the directions δ. This can be used in most of our problems. A simple application, for example, generalizes some of the results of 12.2.2. If $P_*(\varepsilon) = P_* + \varepsilon(P_* - P_i)$ and is P_* has a largest eigenvalue with multiplicity r, with a corresponding $n \times r$ matrix of eigenvectors X, then

$$\lambda_{max}(P_*(\varepsilon)) = \lambda_{max}(P_*) + \varepsilon \lambda_{max}(X'(P_* - P_i)X) + o(\varepsilon), (\varepsilon > 0),$$

which generalizes 12.2.2.1 where we have to assume that $r = 1$. On the other hand the result for $r > 1$ is definitely less useful, because the perturbation of the eigenvalue is no longer asymptotically linear with the perturbation, i.e. not differentiable in the usual sense.

12.3 Algebraic stability

12.3.1 Eigenvalue problems with restrictions

Consider the problem of computing

$$\sigma_1 \triangleq \sup \{\text{tr}(X'BX)^+X'AX \mid X \in S\},$$

where S is a subset of $\mathbf{R}^{n\times p}$, the set of all $n \times p$ matrices. We also define

$$\sigma_0 \triangleq \sup \{\text{tr}(X'BX)^+X'AX \mid X \in \mathbf{R}^{n\times p}\}.$$

We assume that B is positive semidefinite, that A is positive semidefinite, that $\text{rank}(B) \geq p$, and that $A \lesssim B$ in the sense that $x'Ax \leq x'Bx$ for all $x \in \mathbf{R}^n$. These assumptions are true in most of the problems studied in the previous chapters. They guarantee that

$$\sigma_0 = \sum_{s=1}^{p} \lambda_s,$$

and that the sup is attained for $X = Z$, where $AZ = BZ\Lambda$, $Z'BZ = I$, and $\lambda_1 \geq \ldots \geq \lambda_p$ are the p largest generalized eigenvalues of this problem. Clearly $0 \leq \sigma_1 \leq \sigma_0$. In this section we first try to improve the lower bound.

For this purpose we suppose that $X \in S$ implies that $XQ \in S$ for all $p \times p$ matrices Q. We have already seen in the previous chapters that this assumption is usually true in our nonlinear multivariate analysis problems, even if they involve cone restrictions and single variables. Simple special cases are the restriction that each column of X must be in a subspace L of $\mathbf{R}^n$, or the restriction that X must have rank less than or equal to some given q. We also define

$$\varepsilon^2 \triangleq \inf \{\text{tr}(Z - Y)'B(Z - Y) \mid Y \in S\},$$

and we suppose the inf is attained in some $U \in S$. Remember that Z is the solution of the unrestricted generalized eigenvalue problem corresponding with A and B. By the definition of U and the fact that S is closed under multiplication on the right we see that

$$\text{tr}(Z - UQ)'B(Z - UQ)$$

is minimized over Q by $Q = I$. This means that $U'BZ = U'BU$. Now define $W \triangleq U'BU$ and $r = \text{rank}(W)$. Clearly

$$(Z - U)'B(Z - U) = I - U'BU = I - W,$$

which means that all eigenvalues of W are less than or equal to one, and

$$\varepsilon^2 = \text{tr}(I - W),$$

which shows that W is positive definite, and thus $r = p$, if $\varepsilon^2 < 1$. We now define $V \triangleq UW^+$. Thus $V \in S$ and $(V'BV)^+ = W$. Moreover

$$Z'B(V - Z) = Z'BUW^+ - I = WW^+ - I.$$

This implies

$$\text{tr}(V'BV)^{+}V'AV = \text{tr } W\{Z + (V - Z)\}'A\{Z + (V - Z)\} =$$

$$= \text{tr } W\Lambda + \text{tr } W(V - Z)'A(V - Z) + 2 \text{ tr } WZ'A(V - Z).$$

The last term on the right vanishes, because

$$\text{tr } WZ'A(V - Z) = \text{tr } W\Lambda Z'B(V - Z) = \text{tr } W\Lambda(WW^{+} - I) = 0.$$

Thus

$$\text{tr}(V'BV)^{+}V'AV = \text{tr } W\Lambda + \text{tr } W(V - Z)'A(V - Z).$$

Now suppose τ is such that $A - \tau B \gtrsim 0$, i.e. $x'Ax \gtrsim \tau x'Bx$ for all $x \in R^n$. Clearly $\tau = 0$ satisfies this condition, but a better choice is to take τ equal to the smallest generalized eigenvalue of A and B. Now

$$\text{tr}(V'BV)^{+}V'AV \gtrsim \text{tr } W\Lambda + \tau\text{tr } W(V - Z)'B(V - Z).$$

The second term on the right can be simplified by using

$$(V - Z)'B(V - Z) = V'BV - Z'BV - V'BZ + Z'BZ = W^{+} - 2WW^{+} + I,$$

so that

$$\text{tr}(V'BV)^{+}V'AV \gtrsim \text{tr } W\Lambda + \tau \text{ tr}(WW^{+} - W).$$

Making the necessary substitutions proves our final result, which is

$$\sigma_1 \gtrsim \sigma_0 - \sum_{s=1}^{p} (\lambda_s - \tau)\varepsilon_s^2 - \tau(p - r),$$

where ε_s^2 is diagonal element s of I - W. This lower bound is true for all τ not larger than the smallest generalized eigenvalue of A and B. This result generalizes a theorem by Weinberger (1974, p 68) who supposes that S is defined by a subspace of R^n, and who assumes that r = p, in fact even that $\varepsilon^2 < 1$. An easy generalization of our result can be obtained if we replace R^n by a separable Hilbert space, A and B by bounded self-adjoint positive semidefinite linear operators, and S by a set in a finite dimensional subspace of the Hilbert space.

We have indicated how we can bound σ_1 by computing the solution to the generalized eigenproblem without restrictions, and by projecting this solution on S in the metric defined by B. If Z is close to S, then σ_1 is close to σ_0, and our result gives precise quantitative bounds which tell us how close. A similar result for the solutions of the two problemsis considerably more complicated. We derive a result which is at least easy to apply.

Suppose Z solves the unrestricted problem and Y solves the restricted problem, both Z'BZ = I and Y'BY = I. We write Y in the form Y = ZT + Q, with Q'BZ = 0 and T = Z'BY. Then

$$Y'AY = T'\Lambda T + Q'AQ,$$

$$Y'BY = T'T + Q'BQ.$$

Multiply the second equation by λ_{p+1}, and subtract it from the first. This gives

$$Y'AY - \lambda_{p+1}I = T'(\Lambda - \lambda_{p+1}I)T + Q'(A - \lambda_{p+1}B)Q.$$

We now use the result that

$$\lambda_{p+1} = \max \{\frac{x'Ax}{x'Bx} \mid Z'Bx = 0\}.$$

This shows that $Q'(A - \lambda_{p+1}B)Q \lesssim 0$. From this point we can proceed in many different ways, we choose what seems to be the simplest one. By taking traces we find

$$\sigma_1 - p\lambda_{p+1} \lesssim \sum_{s=1}^{p} (\lambda_s - \lambda_{p+1})\kappa_s^2,$$

where $\kappa_1 \gtrsim \ldots \gtrsim \kappa_p$ are the singular values of T. This implies

$$\kappa_1^2 \gtrsim \frac{\sigma_1 - p\lambda_{p+1}}{\sigma_0 - p\lambda_{p+1}},$$

which is a convenient, although possibly not very sharp estimate of the size of T. If we combine this with our upper bound on σ_1, then we find

$$\kappa_1^2 \gtrsim 1 - \frac{\sum_{s=1}^{p} (\lambda_s - \tau)\epsilon_s^2}{\sum_{s=1}^{p} (\lambda_s - \lambda_{p+1})},$$

where we have assumed that $p = r$. Thus we see that Y and Z are close if ϵ^2 is small, but also if λ_{p+1} is much smaller than the larger λ_s.

12.3.2 Applications of the previous results

The applications in this section are different from those in 12.2.2. There we dealt with asymptotic approximations, here with inequalities which are valid without any assumption on the size of the perturbation, although the results will only be nontrivial if the perturbations are relatively small. Both in 12.2.2 and here we assume that an eigenvalue problem has been solved, and we want to say something about the solutions of a related problem without actually computing them. In 12.2.2 the related problems are eigenvalue problems of the same order, here we have considerably more freedom. The results of 12.3.1 can be applied as soon as we can solve the problem of minimizing

$$\text{tr } (Z - Y)'B(Z - Y)$$

over Y in S fairly easily. This gives us the ϵ_s^2, and the required upper bounds.

Suppose, for example, that we eliminate a variable from HOMALS. We use the eigenvalue problem $CY = mDY\Lambda$, and we require that $Y_j = 0$. Then ϵ_s^2 is the discrimination measure $z_{js}'D_jz_{js}$ of variable j on dimension s. Merging categories is equally simple. In the situation of 12.2.2.2 we require that $y_{1s} = y_{2s}$

for all s. The projection problem is very easy to solve, and we find that

$$\varepsilon_s^2 = \frac{d_1 d_2}{d_1 + d_2} (z_{1s} - z_{2s})^2,$$

as expected. If we want to eliminate a subject we start with $P_* X = M_* X\Lambda$, and require that $x_{is} = 0$ for all s. Clearly $\varepsilon_s^2 = m_{*i} x_{is}^2$. Thus the examples treated in 12.2.2 are fairly trivial in this context, the only remarkable thing is that we always start with the 'dual' eigenproblem in this new approach. An additive perturbation in the approach of 12.2 is a subspace restriction in the approach of 12.3, and a subspace restriction is an additive perturbation in the dual eigenproblem, and vice versa.

We have already indicated that it is also possible to treat nonlinear restrictions. If we change all HOMALS (i.e. multiple nominal) variables to PRINCALS single variables, then

$$\varepsilon^2 = \sum_{j=1}^{m} \text{tr } (Z_j - y_j a_j')' D_j (Z_j - y_j a_j'),$$

where y_j and a_j are chosen to minimize this quantity, i.e. y_j and a_j are the optimal rank-one approximation to Z_j in the metric D_j. Our experience indicates that in many examples of this particular sort ε^2 will <u>not</u> be small if $p > 1$, i.e. often ε_1^2 will be small, but the other ε_s^2 will be large. We have explained this behaviour in the previous chapter, in many common gauges HOMALS creates horseshoes and PRINCALS does not.

The procedure in the previous example suggests an alternative proof of the inequalities relating σ_0 and σ_1. It turns out that they can be proved, at least as quickly but somewhat less generally, by starting from meet-loss σ_M, and applying the same partitioning of meet-loss as in chapters 5, 6, 7, 8. In these chapters we also used the closedness of the constraint sets under matrix multiplication on the right to show equivalence with certain singular value and eigenvalue problems. Both in this section and in section 12.2 it is possible to generalize the treatment to, for example, dropping both subjects and variables at the same time. It is clear that in this case we have to study perturbations of the singular value problem directly, not of one of the derived dual eigenproblems. But the singular value problem is, of course, equivalent to an augmented eigenproblem, which means that our general theory still covers it.

12.3.3 <u>Discretization</u>

We have already pointed out that the results of 12.3.1 also apply to approximation of an infinite dimensional eigenproblem by using finite dimensional subspaces. In particular we can approximate the general homogeneity analysis problem outlined in the previous chapter by using step functions. Suppose $\phi_j(\underline{h}_j)$ is

the optimal transformation without restrictions, and suppose $-\infty = a_0 < \ldots < a_{k_j} = +\infty$ are the discretization points. Moreover we suppose that $\underline{h}_j$ has marginal probability density $p_j(x)$. Then the contribution to ε^2 of variable j is

$$\sum_{\ell=1}^{k_j} \int_{a_{\ell-1}}^{a_\ell} (y_\ell - \phi_j(x))^2 p_j(x) dx,$$

where the y_ℓ are chosen to minimize this quantity, i.e.

$$y_\ell = \text{AVE}(\phi_j(\underline{h}_j) \mid a_{\ell-1} < \underline{h}_j < a_\ell).$$

In many important gauges the optimal function we are approximating is $\phi_j(x) = x$. In this case it also makes sense, and it is computationally feasible, to define optimal discretization points a_ℓ. These points are <u>not</u> optimal in the sense that they maximize the largest eigenvalue over all step functions and over all sets of discretization points. They are optimal in the 'marginal' sense that they minimize ε^2, and consequently minimize the upper bound to the largest eigenvalue. These marginal optimal discretization points, together with the optimal step function, can be found by using an alternating least squares algorithm to minimize ε^2. We have already indicated how to minimize the function over the y_ℓ for fixed a_ℓ, if we are approximating $\phi_j(x) = x$ then the optimum a_ℓ for fixed y_ℓ is given by

$$a_\ell = \tfrac{1}{2}(y_\ell + y_{\ell+1}).$$

The problem of discretization has been treated in statistical literature by Sheppard (1898) and Cox (1957), but much more completely in the communications analysis literature under the name 'quantization'. We do not discuss this very extensive literature in detail, some of the key references are Max (1960), Roe (1964), Wood (1969), Elias (1970), Gish and Pierce (1968), Sharma (1978), Gersho (1979). This literature is largely unknown to statisticians and data analysts, but it contains some very interesting results. There are some interestir asymptotic expressions for ε^2 if the number of discretization points is large, and there is considerable literature which suggests that these asymptotic approximations are already very good for small k_j. A general result of this type is that if $k_j \to \infty$ then $k_j^2 \varepsilon^2 \to C$, where C depends on p(x). This suggests that discretization error in, for example, HOMALS will always be much smaller than sampling error. Results such as these are investigated by Monte Carlo methods in Van Rijckevorsel, Bettonvil, and De Leeuw (1980) for the normal distribution in HOMALS with various descretization strategies, and by Van der Burg and De Leeuw (reported in Gifi, 1980, p 266-277) for CANALS with normal distributions, various numbers of discretization points, and various types of intercorrelations. The conclusion is that discretization type or fineness is of small influence in the cases studied, sample size is much more important, and in the case of CANALS the intercorrelation structure (degree of multicollinearity) is very

important too. The discretization results in this section are much more generally valid than for normal distributions and for step functions, we can also use them to locate the knots of spline functions for any distribution. This creates the problem of a 'robust' choice of knots, which is good for most distributions in a particular family of gauges. The empirical results so far suggest that both uniform spacing of knots or discretization points (which is optimal for the rectangular distribution) and choice of knots so that the marginals become rectangular is fairly robust.

12.3.4 Interactive variables

We have seen in chapters 6 and 11 that the general OVERALS problem can be interpreted as HOMALS with restrictions on the category quantifications. These restrictions are often additivity, which is used to create groups of variables. Additivity is a linear restriction, and it is easy to estimate ε^2 for additivity. In fact computing ε^2 amounts to performing an analysis of variance on the unrestricted quantications, although the design is often not orthogonal because we have to weight by the marginals in D. As the examples in chapters 7 and 11 using the CBS-data show it is often possible to predict quite precisely what CANALS does from what HOMALS on the interactive variables does, because there typically seem to be only small interactions in examples like these.

12.3.5 Other techniques for algebraic stability

The method discussed in 12.3.1 can be extended in various directions. A very common one is to extend it to other linear spaces of infinite dimensions. This work is often of limited practical value for the data analysis problems we are interested in, but on the other hand it does show clearly in which direction some of our results could possibly be improved. Reviews of work in this area are the book by Vainikko (1976), and the paper by Chatelin (1979). More direct generalizations of our results to not necessarily bounded and not necessarily self-adjoint generalized eigenproblems in not necessarily separable Hilbert spaces are contained in Kolata (1978), our bounds for the approximation of eigenvectors are improved by Weinberger (1960). Our approach is based on investigating how well a solution of the eigenproblem satisfies the restrictions. It is also possible to proceed the other way around. We can start with a solution of the restricted problem and check how well it satisfies the eigenequation. Bounds can also be derived from this. Some pertinent references are Wilkinson (1961), Yamamoto (1980), Symm and Wilkinson (1980), and in an infinite dimensional context Werner (1974).

Many additional results can be derived from the max-min characterization of eigenvalues. These results have two forms again, one for additive perturbations and one for eigenvalues of submatrices. The two forms are 'dual' and can often

be translated into each other by using the singular value decomposition. Some of the sharpest results have been proved by Robert Thompson and his collaborators. We mention Thompson and Freede (1970, 1971), Thompson and Theranios (1972a,b), Thompson (1975, 1976), but this list is not even approximately complete. The min-max results, which only refer to eigenvalues and not to eigenvectors, have been used in connection with principal components analysis and correspondence analysis by Escofier and Le Roux (1972). In that paper the results are used in their simplest form, the results of Thompson and others make it possible to improve them a great deal.

The study of eigenvectors is more complicated. There are two basic papers, one by Davis and Kahan (1970), and another one by Stewart (1973). Both papers review previous work, they are essentially concerned with converting first order analytic perturbation results into rigorous bounds in the form of inequalities (using algebraic methods). Stewart has carried on his research since 1973, additional results are reported in Stewart (1975, 1979) and in the review paper (1978). The results of Davis and Kahan (1970) have been used in components analysis by Escofier and Le Roux (1977), who study the effect of adding and dropping a variable. The results of Davis and Kahan are limited to ordinary eigenvalue problems, results of Stewart make it possible to extend them to generalized eigenvalue problems and singular value problems.

It is clear that this section is merely a list of references, together with some hints on how these results could be applied in nonlinear multivariate analysis. It is clearly a considerably research project to apply them to various techniques, using various types of perturbations.

12.4 Replication stability

12.4.1 The delta method

The delta method is one of the classical methods of asymptotic statistics. In its simplest form it assumes that a sequence $\underline{x}_n$ of random variables is given, that $n^{\frac{1}{2}}(\underline{x}_n - \mu)$ converges in distribution to a multinormal variable with mean zero and dispersion Σ, and that $\underline{y}_n = \Phi(\underline{x}_n)$ defines a new sequence of random variables, where the mapping Φ is assumed to be differentiable at μ. The basic result of the delta method then tells us that the asymptotic distribution of $n^{\frac{1}{2}}(\underline{y}_n - \Phi(\mu))$ is the same as the asymptotic distribution of $n^{\frac{1}{2}}G(\mu)(\underline{x}_n - \mu)$, with $G(\mu)$ the derivative of Φ evaluated at μ. And this last asymptotic distribution is multinormal with mean zero and dispersion $G(\mu)\Sigma G(\mu)'$.

This basic result can be generalized in several directions, all these generalizati are useful for our purposes. In the first place we can relax the assumption that $n^{\frac{1}{2}}(\underline{x}_n - \mu)$ is asymptotically normal, it is sufficient for some purposes to assume that $K_n(\underline{x}_n - a_n)$ has some asymptotic distribution for some sequence a_n, and for

some sequence of norming constants $K_n \to \infty$. It remains true that $K_n(\Phi(\underline{x}_n) - \Phi(a_n))$ has the same asymptotic distribution as $K_n G(a_n)(\underline{x}_n - a_n)$, which may be considerably easier to compute. In the second place for some purposes it suffices to assume differentiability in all directions. In that case the asymptotic distribution of $n^{\frac{1}{2}}(\Phi(\underline{x}_n) - \Phi(\mu))$ is the same as that of $n^{\frac{1}{2}}G(\mu,\underline{x}_n - \mu)$, with $G(\mu,z)$ the directional derivative. The last asymptotic distribution is the distribution of $G(\mu,\underline{z})$, where $\underline{z}$ is multinormal with mean zero and dispersion Σ, provided $G(\mu,z)$ is continuous in its second argument. Another generalization is that we do not insist that Φ is the same for all n, but we study $\underline{y}_n = \Phi_n(\underline{x}_n)$. And finally we can be interested in the asymptotic distribution of statistics such as $n(\Phi(\underline{x}_n) - \Phi(\mu) - G(\mu)(\underline{x}_n - \mu))$, which give as it were higher order information.

How do we apply the delta method to our nonlinear multivariate analysis. Our techniques compute statistics which are all functions of the number of observations of the cells in the multidimensional contingency table, or equivalently of the frequencies of the profiles. We can apply the delta method if we show that these functions are differentiable, and if we imbed our observations in a sequence of random variables in a reasonable way. In our case it is obvious how a statistician would perform this imbedding. We suppose that the individuals are a simple random sample from an infinite population, this makes the cell frequencies multinomially distributed and consequently asymptotically normal. Some methodological comments, which repeat some of the things said in chapter one, are in order here. If we want to assess the variability (the replication stability) of a statistic, then the appropriate thing to do is to repeat the experiment a couple of times under the same conditions, and to recompute the statistic after each experiment. This is what is done in the experimental sciences to find out what the 'measurement error' is, to find out in how far we do indeed control the relevant conditions. In the social sciences it is often impossible to repeat experiments, if we do repeat them we find different results, and we do not know if these results are due to measurement error or to uncontrolled variables, because we do not know what the relevant variables are. Consequently we replace actual replication by the statistical model of independent identically distributed random variables, and we replace actual control by pre-experimental randomization. We think that randomization is a very useful device, which simply cannot be dispensed with. But the statistical model, which actually tells us that replications are unnecessary because probability theory and statistics shows us what replications would give if we performed them, carries too heavy a burden in many situations. It is nice to be able to predict what replication would give, but it is not nice that we have no conceivable way to falsify these predictions. And, if we try to falsify them, we usually succeed, only we never know if this happens because the statistical model is not true or because our attempt to falsify the predictions was faulty. We can only verify the predictions of statistical models in Monte Carlo experiments,

but of course this is more or less tautological, or it is a method to find out if we made an error in our proof of a theorem, or it is a method to find out if a particular approximation holds 'sufficiently well'. Thus pre-experimental randomization does not replace control, it is merely a passive way to dodge the criticism that some choices are arbitrary. The solution is to make all choices completely arbitrary by definition, except one or two for which the criticism does not apply because they are the experimental conditions. This is the only possible way to proceed in many situations. On the other hand assuming a simple statistical model does not replace actual replication, in fact it does not even make replication less desirable, and there is a simple alternative way to proceed. The alternative way is not to assume a model, and not to proceed as if the model was true, and to face all kinds of instabilities. All this has very little to do with inference from sample to population (which is simply a misleading way to formulate the problem of replication stability in many situations), it also has very little to do with prescriptions on how to behave rationally, which belong to ethics or normative psychology.

We first give a simple example of the delta method to show how it works. Suppose A_k are K fixed matrices of order m, and A_k has probability π_k of occurring in the population from which we are sampling. A multinomial simple random sample of size n gives relative frequencies $\underline{p}_k$, our technique is to compute eigenvalues and eigenvectors of $\underline{A}_* = \sum \underline{p}_k A_k$. In the preference rankings techniques of chapter 10, for example, the A_k can be permutation matrices P_k or rank-one matrices of the form $P_k yy'P_k$, with y the vector of centered rank numbers. From the results of 12.2 we obtain

$$\frac{\partial \Lambda}{\partial p_k} = \text{diag}(X'A_k X),$$

where X is the matrix of normalized eigenvectors of A_*, evaluated at π. We suppose that all elements of Λ, evaluated at π, are different. Clearly

$$\sum_{k=1}^{K} \pi_k \frac{\partial \Lambda}{\partial p_k} = \text{diag}(X'A_* X) = \Lambda.$$

The $\underline{p}_k$ are jointly asymptotically normal, more precisely the vector $n^{\frac{1}{2}}(\underline{p} - \pi)$ is asymptotically multinormal with mean zero and dispersion $\Pi - \pi\pi'$, where Π is the diagonal matrix with on the diagonal the elements of π. Thus $n^{\frac{1}{2}}(\Lambda(\underline{p}) - \Lambda(\pi))$ is asymptotically normal with mean zero, and

$$n \cdot \text{COV}(\lambda_s(\underline{p}), \lambda_t(\underline{p})) \to \sum_{k=1}^{K} \pi_k \Big(\sum_{i=1}^{m} \sum_{j=1}^{m} x_{is} x_{js} a_{ijk} \Big) \Big(\sum_{i=1}^{m} \sum_{j=1}^{m} x_{it} x_{jt} a_{ijk} \Big) - \lambda_s \lambda_t.$$

By substituting the sample $\underline{x}_{is}$, $\underline{\lambda}_s$, and $\underline{p}_k$ on the right hand side we find a consistent estimate of the asymptotic variances and covariances. In MDPREF we have $A_k = P_k yy'P_k$. If we define $\rho_{ks} = x_s'P_k y$, the correlation between eigenvecto

and vector of centered rank numbers, then

$$n\ \mathrm{COV}(\lambda_s(\underline{p}),\lambda_t(\underline{p})) \rightarrow \sum_{k=1}^{K} \pi_k \rho_{ks}^2 \rho_{kt}^2 - \lambda_s\lambda_t.$$

Because

$$\sum_{k=1}^{K} \pi_k \rho_{ks}^2 = \lambda_s,$$

it follows that the asymptotic covariance of the eigenvalues is equal to the covariance of the ρ_{ks}^2. These formulas still look fairly simple and they can be interpreted to some extent. If in generalized eigenproblems the matrices A and B both depend on the parameters, possibly nonlinearly, then the situation becomes more unpleasant. If we compute derivatives of eigenvectors or higher derivatives of eigenvalues then the formulas becomes even more forbidding. It does not help very much to give these formulas here for various specific techniques, using the results of 12.2.1 they can be derived quite easily.

The delta method can also be used for bias correction, in the previous development we used it mainly for variance estimation. For the eigenvalues, for example, we can use the development

$$\lambda_s(\underline{p}) = \lambda_s(\pi) + \sum_{k=1}^{K} \frac{\partial \lambda_s}{\partial p_k}(\underline{p}_k - \pi_k) + \tfrac{1}{2}\sum_{k=1}^{K}\sum_{\ell=1}^{K} \frac{\partial^2 \lambda_s}{\partial p_k \partial p_\ell}(\underline{p}_k - \pi_k)(\underline{p}_\ell - \pi_\ell) + o_p(n^{-1}),$$

where the notation $o_p(n^{-1})$ means that n times the residual converges to zero in probability. If we take expectations on both sides we find

$$\mathrm{AVE}(\lambda_s(\underline{p})) = \lambda_s(\pi) + \tfrac{1}{2}\sum_{k=1}^{K}\sum_{\ell=1}^{K} \frac{\partial^2 \lambda_s}{\partial p_k \partial p_\ell}(\delta^{k\ell}\pi_k - \pi_k\pi_\ell) + o(n^{-1}).$$

By using the appropriate formula for second derivatives, and by substituting sample quantities, we can use this formula for bias correction. Of course this practice more or less implies that we are interested in the population eigenvalues, and it certainly implies that we take our multinomial problem imbedding seriously. The same thing is true if we use asymptotic normality and consistent estimates of dispersions to test the hypothesis that the smallest eigenvalues are zero. It may be of some comfort to some that such a test is fairly easy to compute, consistent, although not asymptotically most powerful against close alternatives. But from our point of view the statistical model which assumes that some population eigenvalues are zero is useless. The model has no rational interpretation in terms of the familiar gauges, indeed most gauges predict that the eigenvalues are all nonzero. The only model which does make some sense, for example in HOMALS or ANACOR, is that all eigenvalues are zero, i.e. that all variables are independent in pairs. This model will of course be rejected in all interesting data sets.

If we assume that all eigenvalues are zero the theory of 12.2.1 does not apply any more, because there we assumed that all eigenvalues were different. The results

of 12.2.3 can still be used, however. In Lebart (1976) this is done for correspondence analysis, we use our simpler example of preference rank orders. Suppose that the model is that $A_* = \sum \pi_k A_k$ is of rank one. For the first eigenvalue we still have asymptotic normality, but for the remaining eigenvalues the theory of 12.2.3 tells us that

$$\lambda_s(\underline{p}) = \lambda_s\left(\sum_{k=1}^{K} (\underline{p}_k - \pi_k)(I - xx')A_k(I - xx')) \right) + o_p(n^{-\frac{1}{2}}).$$

with x the eigenvector corresponding with the largest eigenvalue. Thus directional derivatives are eigenvalues of a matrix with asymptotically normal elements. Under some additional symmetry conditions the elements of this making are asymptotically indepedent, and thus the eigenvalues are the square roots of the eigenvalues of a standard Wishart matrix. The distribution of the eigenvalues of a standard Wishart matrix has been tabulated extensively, and has been used for example by Lebart (1976).

We end this section with some references and some comments. The delta method is classical, it is discussed in all important textbooks, particulary neatly for example in Rao (1965, section 6a). A proper treatment of the delta method is possible if we realize that there are three types of convergence involved. In the first place weak convergence, or convergence in distribution. The key theorem here is the Mann-Wald-Rubin theorem on the preservation of weak convergence by mappings, treated for example by Billingsley (1968, section 1.5) and more generally by Topsoe (1967). The second type of convergence is convergence of moments (including expected values and variances), which is guaranteed by weak convergence together with uniform integrability (Billingsley, 1968, p 31-33). The third type of convergence is convergence of probability measures, which is studied for delta method expansions in Bhattacharya and Ghosh (1978). It seems to us that convergence of moments is most useful for our problems. Hurt (1976) gives some useful general theorems, first order asymptotic distribution theory for principal components analysis is given in the normal case by Anderson (1963), and in the nonnormal case by Davis (1977) and Waternaux (1978). First order asymptotic theory for correspondence analysis has been given by O'Neill (1978a,b), O'Neill (1980) applies the delta method in the related problem of decomposition of a discrete multivariate distribution by using tensor products of orthogonal functions on the marginals.

We have incorporated the delta method in our programs ANACOR and ANAPROF. This means that these programs do not only give eigenvalues and eigenvectors, but also an estimate of the dispersion matrix of the eigenvalues, and for each row and column point an estimate of the dispersion matrix of the p coordinates. This dispersion matrix estimate can be used to draw 95% confidence ellipsoids

around each of the points. This is illustrated for the social mobility data of table 4.3. Figure 12.1 shows that the ANACOR structure is fairly stable, that the ellipses of row and column points have considerable overlaps and that categories 6 and 7 of both sons and fathers are virtually indistinguishable. The ANAPROF solution for the Sugiyama data on religious practise from table 4.6 is plotted with 95% ellipses in figure 12.2. Because of the large sample size the ellipses are much smaller in this example and in fact they are, with one tiny exception, all disjoint. This indicates considerable replication stability under the multinomial model. In general the ellipsoids are quite useful, even if the multinomial assumptions are not very intuitive. This is because the critical quantities in stability analysis of this type are the partial derivatives. Analytical stability gives us first order approximations, which can in some cases be expressed as inequalities. Now inequalities are nice and rigorous, but they tend to be rather pessimistic. The delta method provides us with another method to assess the size of the derivatives, because basically the estimates of the sampling variances of the statistics are the sampling variances of the partial derivatives of the statistics. Thus we do not have to buy the statistical model in order to compute ellipsoids. An alternative way to compute ellipsoids, for example, is simply to use the second derivatives of the loss function evaluated at the solution. For some loss functions this provides asymptotically the same ellipsoids, for others they will perhaps differ from the delta method elllipsoids. The 95% confidence interpretation may be natural in some situations, but it does refer to imaginary replications.

A disadvantage of the delta method in some situations is that it is sometimes difficult and/or expensive to compute partial derivatives. We have already seen that the partials of the eigenvectors are quite complicated, if we want to correct eigenvectors for bias we need the second partials which are very complicated. The methods described in the next section have been developed partly to avoid these very complicated computations.

12.4.2 Randomization methods

12.4.2.1 General theory

We describe this class of methods in the discrete case only. Suppose there are K possible profiles, the data consist of the frequencies $n_1,\ldots,n_k$ with which these profiles occur. Let $n \triangleq \sum n_k$ and $p_k \triangleq n_k/n$. Observe that up to now we have not introduced any probabilistic structure. Our technique consists of the computation of a function $\phi(p)$, which we shall assume to be real valued for the time being. Now suppose we have a rule which associates with each pair (p,n) vectors $p_1(p,n),\ldots,p_m(p,n)$. The vectors $p_j(p,n)$ are nonnegative, and their elements add up to one. With each $p_j(p,n)$ our rule also associates a

probability $\pi_j(p,n)$. Now compute

$$\mu_n(p) \triangleq \sum_{j=1}^{m} \pi_j(p,n)\phi(p_j(p,n)),$$

$$\sigma_n^2(p) \triangleq \sum_{j=1}^{m} \pi_j(p,n)\{\phi(p_j(p,n) - \mu_n(p))\}^2.$$

The vectors $p_j(p,n)$ are perturbation vectors, the $\pi_j(p,n)$ are perturbation probabilities, the number of perturbation vectors m may depend on n. The basic idea behind the method is that we have observed p, but we might as well have observed the $p_j(p,n)$.

This 'might as well have' sounds intuitive and difficult to define precisely. Before we discuss it in more detail we first discuss some general techniques to construct the $p_j(p,n)$. The first one, which is the most familiar one, is the jack-knife. Miller (1974) reviews theory and applications of the jackknife. It is easiest to explain the construction of the perturbations if we work with the $n \times K$ indicator matrix of the profiles, thus $p = \frac{1}{n} G'u$. In the jackknife we suppose that we might as well have observed any one of the $(n - 1) \times K$ submatrices of G obtained by deleting a single row. There are obviously n of these submatrices, but only K of them are different, depending on which profile we delete. The corresponding perturbation vectors are

$$p_k(p,n) = p + \frac{1}{n-1}(p - e_k),$$

with probability

$$\pi_k(p,n) = p_k.$$

A second method to construct the perturbations has been suggested by Hartigan (1969, 1970). It is called subsampling. The basic idea is that we might as well have observed any of the $2^n - 1$ subsamples, i.e. submatrices of G, excluding only the empty sample, all with equal probability. In other versions of the method we might as well have observed a particular set of submatrices, which has a particular balanced group structure, again all with equal probability (cf Hartigan, 1969, or Gordon 1974a,b). It is clear that the marginals of the submatrices can have all possible values $\nu_1,\ldots,\nu_K$, provided that $\nu_k \leq n_k$ for all k. The probability of a perturbation vector found by norming ν is equal to

$$\binom{n_1}{\nu_1} \cdot \binom{n_2}{\nu_2} \cdots \binom{n_K}{\nu_K} \Big/ (2^n - 1).$$

The third and last method we discuss is due to Efron (1979). It is called the bootstrap. The basic idea is that we might as well have observed any matrix G of dimension $n \times K$ consisting of the same rows, but in different frequencies. The probability of another G is n^{-n} times the multinomial coefficient of

$\upsilon_1,\ldots,\upsilon_n$, where υ_i indicates the number of times row i occurs in the new G. Many of these new G have the same marginals, the probability of a set of marginals $\nu_1,\ldots,\nu_K$ is the probability that we observe $\nu_1,\ldots,\nu_K$ if we take a sample of size n from the multinomial with parameters $p_1,\ldots,p_K$. Jackknifing, subsampling, and bootstrapping are the main randomization methods. In discrete situations it seems to use that the jackknife is the simplest method, and the bootstrap is the most natural one. Efron emphasizes that if the data are a sample from a multinomial then the perturbation vector is related to the sample proportions in the same way as the sample proportions are related to the population parameters. Subsampling has never been used by us, although interesting theoretical and practical work has been done by Hartigan (1975), Forsythe and Hartigan (1970), Gordon (1974a,b). In the next sections we consequently concentrate on bootstrap and jackknife.

The idea that we might as well have observed something else is related to the idea of a random sample. If the data are a random sample from a population, then we might indeed have observed any of the perturbation vectors as well, with the indicated probabilities, which estimate corresponding population probabilities. All three approaches assume that the n individuals are equally important and interchangeable, conversely if we assume this then the randomization methods make sense. On the other hand it is not strictly necessary to commit oneself to any probability model concerning the original observations. We shall see that the randomization methods can be used as Monte Carlo methods or as discretization methods to approximate first and/or second derivatives of ϕ, and we already know that first and second derivatives of loss functions and solutions give interesting information on stability in general.

12.4.2.2 Approximation properties of the jackknife

In case of the jackknife we have seen that

$$\mu_n(\mathbf{p}) = \sum_{k=1}^{K} p_k \phi(p + \frac{1}{n-1}(p - e_k)).$$

If we assume that ϕ has a bounded third derivative on the unit simplex in $\mathbf{R}^K$, then

$$\mu_n(p) = \phi(p) + \tfrac{1}{2} \frac{1}{(n-1)^2} \operatorname{tr} H(p)(P - pp') + O((n-1)^{-3}),$$

with H(p) the matrix of second derivatives of ϕ, evaluated at p, and with P the diagonal matrix with the p_k. We can use this relationship directly for bias correction. It follows that

$$n\phi(p) - (n-1)\mu_n(p) = \phi(p) - \tfrac{1}{2} \frac{1}{n-1} \operatorname{tr} H(p)(P - pp') + O((n-1)^{-2}),$$

If p is based on a random sample with multinomial probability π, then we know

that the delta method gives

$$\text{AVE}(\phi(\underline{p})) = \phi(\pi) + \tfrac{1}{2}\frac{1}{n}\,\text{tr}\, H(\pi)(\Pi - \pi\pi') + O(n^{-2}).$$

Combining the last two results gives

$$\text{AVE}(n\phi(\underline{p}) - (n-1)\mu_n(\underline{p})) = \phi(\pi) + O(n^{-2}).$$

Thus computing $n\phi(\underline{p}) - (n-1)\mu_n(\underline{p})$ corrects for bias, without making it necessary to compute second, or even first, derivatives.

In the same way it is possible to prove that

$$\sigma_n^2(p) = (n-1)^{-2}\, g(p)'(P - pp')g(p) + O((n-1)^{-3}),$$

which means that the variance of the pseudo-values

$$n\phi(p) - (n-1)\phi(p + \frac{1}{n-1}(p - e_k)),$$

each with probability p_k is $g(p)'(P - pp')g(p) + O((n-1)^{-1})$, where $g(p)$ is the vector of partials at p. It follows that if $\underline{p}$ is multinomial, then the variance of the pseudovalues can be used as a consistent estimate of the asymptotic variance of $\phi(\underline{p})$, again computable without formulas for derivatives.

We have just explained the theoretical jackknife, which computes all K different pseudovalues, their average and variance exactly. There are some familiar variations of the jackknife, in one variation we omit s observations at the same time, in another variation we add one observation, a third variation omits observations continuously, and so on. For these variations we refer to the literature. There is another variation which interests us more. Suppose $\underline{e}_1,\ldots,\underline{e}_R$ are independent and identically distributed random variables, which assume the values $e_1,\ldots,e_K$ with probabilities $p_1,\ldots,p_K$. Define

$$\underline{\mu}_n(p) = \frac{1}{R}\sum_{r=1}^{R}\phi(p + \frac{1}{n-1}(p - \underline{e}_r)).$$

Now $\underline{\mu}_n(p)$ is a random variable which converges in probability if $R \to \infty$ to the constant $\mu_n(p)$. In fact $R^{\frac{1}{2}}(\underline{\mu}_n(p) - \mu_n(p))$ converges to a normal distribution with mean zero and variance $(n-1)^{-2}g(p)'(P - pp')g(p) + O((n-1)^{-3})$. These results do not depend on any probabilistic assumptions on the distribution of p, in fact p is assumed to be a fixed set of constants. We can also compute the R pseudo-values

$$n\phi(p) - (n-1)\phi(p + \frac{1}{n-1}(p - \underline{e}_r)).$$

The expected value of their sample mean is $\phi(p) - \tfrac{1}{2}\frac{1}{n-1}\,\text{tr}\, H(p)(P - pp') +$ $+ O((n-1)^{-2})$, the expected value of their sample variance is

$$\frac{R-1}{R}\, g(p)'(P - pp')g(p) + O((n-1)^{-1}).$$

If both $\underline{p}$ and $\underline{e}_r$ are random, the $\underline{e}_r$ are defined conditionally on $\underline{p}$, and $\underline{p}$

is marginally multinomial with parameter π, then the expected value of the sample mean of the R pseudovalues

$$n\phi(\underline{p}) - (n-1)\,\phi(\underline{p} + \frac{1}{n-1}(\underline{p} - \underline{e}_r))$$

is $\phi(\pi) + O(n^{-2})$, and the expected value of the sample variance is

$$\frac{R-1}{R}\, g(\pi)'(\Pi - \pi\pi')g(\pi) + O((n-1)^{-1}).$$

This could be called the Monte Carlo version of the jackknife. It is easy to apply if K is very large, in which case the theoretical jackknife can easily become computationally rather demanding.

12.4.2.3 Approximation properties of the bootstrap

For the bootstrap we can write

$$\mu_n(p) = \sum_{\nu} \pi(\nu|p)\phi(\nu/n),$$

where $\pi(\nu|p)$ is the probability of integer vector ν if we draw a random sample of size n from a multinomial with parameters p. A first remark which is in order here is that $\mu_n(p)$ is the n-th Bernstein polynomial of ϕ, evaluated in p. There is an enormous literature dealing with Bernstein polynomials, the older literature is reviewed in Lorenz (1953), but many more results have been derived since this book. The literature is valuable, because any result about Bernstein polynomials is a result about the bootstrap. An analysis of the bootstrap in terms of the Bernstein polynomials will be published elsewhere. In this book we proceed in the same way as with the jackknife.

In the first place

$$\mu_n(p) = \phi(p) + \tfrac{1}{2}\, n^{-1}\, \text{tr}\, H(p)(P - pp') + O(n^{-2}).$$

Thus

$$2\phi(p) - \mu_n(p) = \phi(p) - \tfrac{1}{2}\, n^{-1}\, \text{tr}\, H(p)(P - pp') + O(n^{-2}),$$ and in the multinomial case

$$\text{AVE}(2\phi(\underline{p}) - \mu_n(\underline{p})) = \phi(\pi) + O(n^{-2}).$$

Thus the theoretical bootstrap can also be used to correct for bias, in the same way as the jackknife, although generally with more computational effort. For the bootstrap we also find

$$\sigma_n^2(p) = n^{-1}\, g(p)'(P - pp')g(p) + O(n^{-2}).$$

Bootstrap pseudovalues are defined by

$2\phi(p) - \phi(\nu/n)$, which has probability $\pi(\nu|p)$. Their variance is $\sigma_n^2(p)$.

In the theoretical bootstrap summation is over all possible ν. This is expensive, even in small examples, it is simply impossible in larger ones. Consequently we need a Monte Carlo version of the bootstrap too. We define it directly in terms

of pseudovalues as $2\phi(p) - \phi(\underline{\nu}_r/n)$, where the $\underline{\nu}_r$ are independent samples from the multinomial with probabilities p. If the original observations or indicator matrices are still available we can sample the $\underline{\nu}_r$ by drawing simple random samples with replacement from our n individuals or from the n rows of the indicator matrices, until we have a sample of size n too. We repeat this procedure R times, and use these R bootstrap samples to compute pseudovalues. In the same way as before the mean of the R pseudovalues can be used to estimate $\phi(\pi)$ in the multinomial case, the variance of the pseudovalues can be used to estimate the standard error of $\phi(\underline{p})$. The formulas are very much like those for the jackknife, and we do not give them here.

Figures 12.3 and 12.4 give ANACOR/ANAPROF solutions for ten bootstrap samples from the social mobility and religious practice data analyzed in chapter 4. The ellipses are computed with the delta-method, they are the same as in figures 12.1 and 12.2. It is clear that the bootstrap gives roughly the same information as the delta-method in these examples, although in some cases the bootstrap vectors emanating from the solution point tend to point in one direction only. This implies that the average of the bootstrap pseudo-values will give an appreciable 'bias-correction'.

12.4.2.4 Comparison of jackknife and bootstrap

We have seen that both methods lead to roughly the same formulas, and any choice between them should be based on comparison of second order terms in the expansions Such comparisons will be published elsewhere. The results that we have suggest that it is difficult to recommend one of the two techniques on asymptotic grounds. Moreover we must also compare the techniques with subsampling and with the original delta method. Recent theoretical work in mathematical statistics seems to suggest that the delta method corrected for bias will give the smallest mean square error, but we do not only compare the methods in terms of loss functio In fact it is clear from the examples in this chapter and the applications in the next chapter that jackknife and bootstrap are very powerful data analysis techniques. We prefer the Monte Carlo version of the bootstrap for the time being, because it is very easy to explain, very easy to compute, and very natural in multinomial situations. If the delta method can be applied the two techniques give very similar information on stability. The bootstrap values $\phi(\underline{\nu}_r/n)$ have a probabilistic interpretation and an asymptotic distribution even if p is nonrandom the jackknife fails in some situations in which the bootstrap does not fail. This is because the bootstrap smoothes more, in spline-terminology the bootstrap is variation diminishing, and can approximate even discontinuous functions. The theoretical jackknife is easier to compute than the theoretical bootstrap, because there are fewer interpolation points. For the bootstrap $\mu_n(p)$ is a polynomial in p which is also convenient in theoretical investigations.

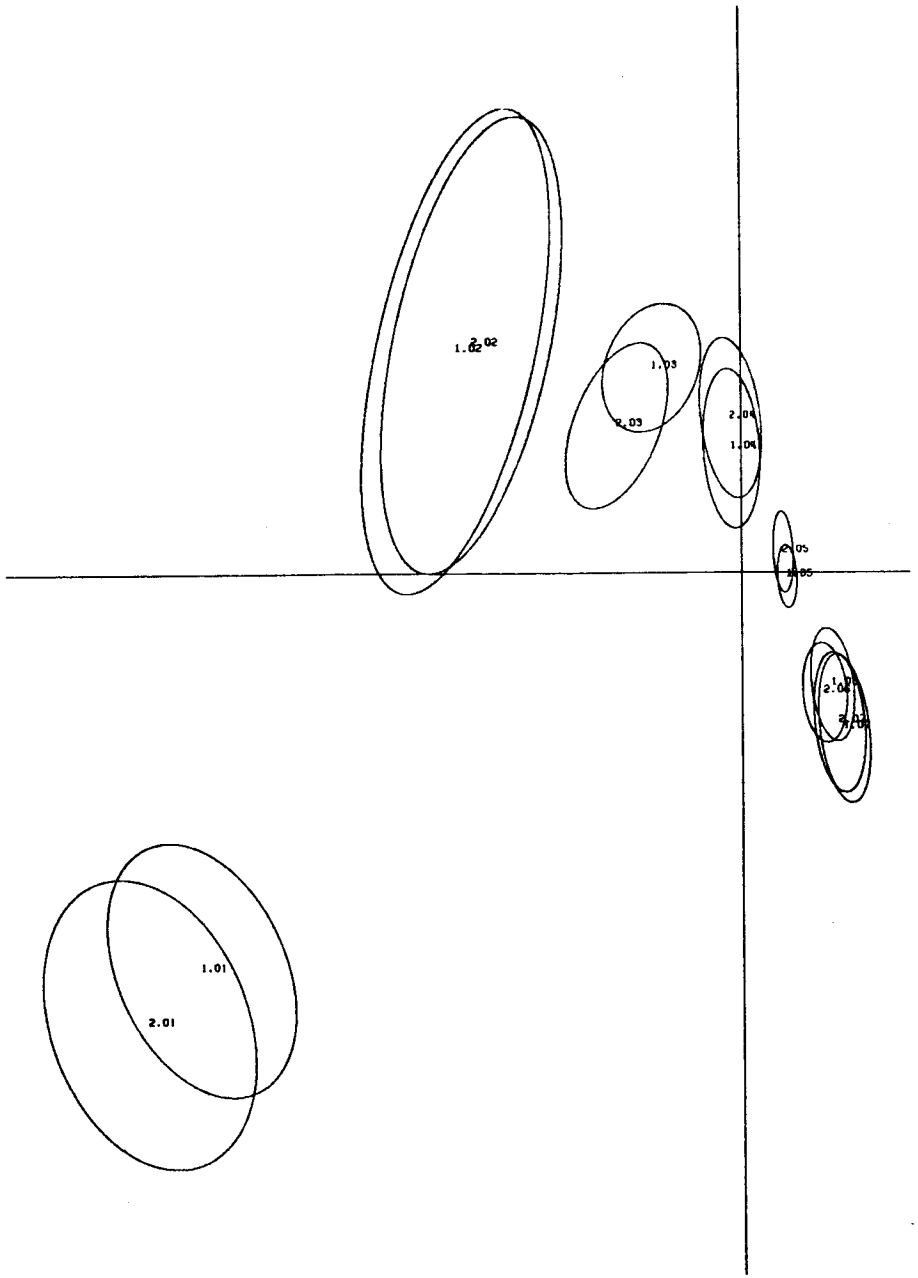


Figure 12.1 ANACOR solution of the social mobility data with 95% confidence ellipses

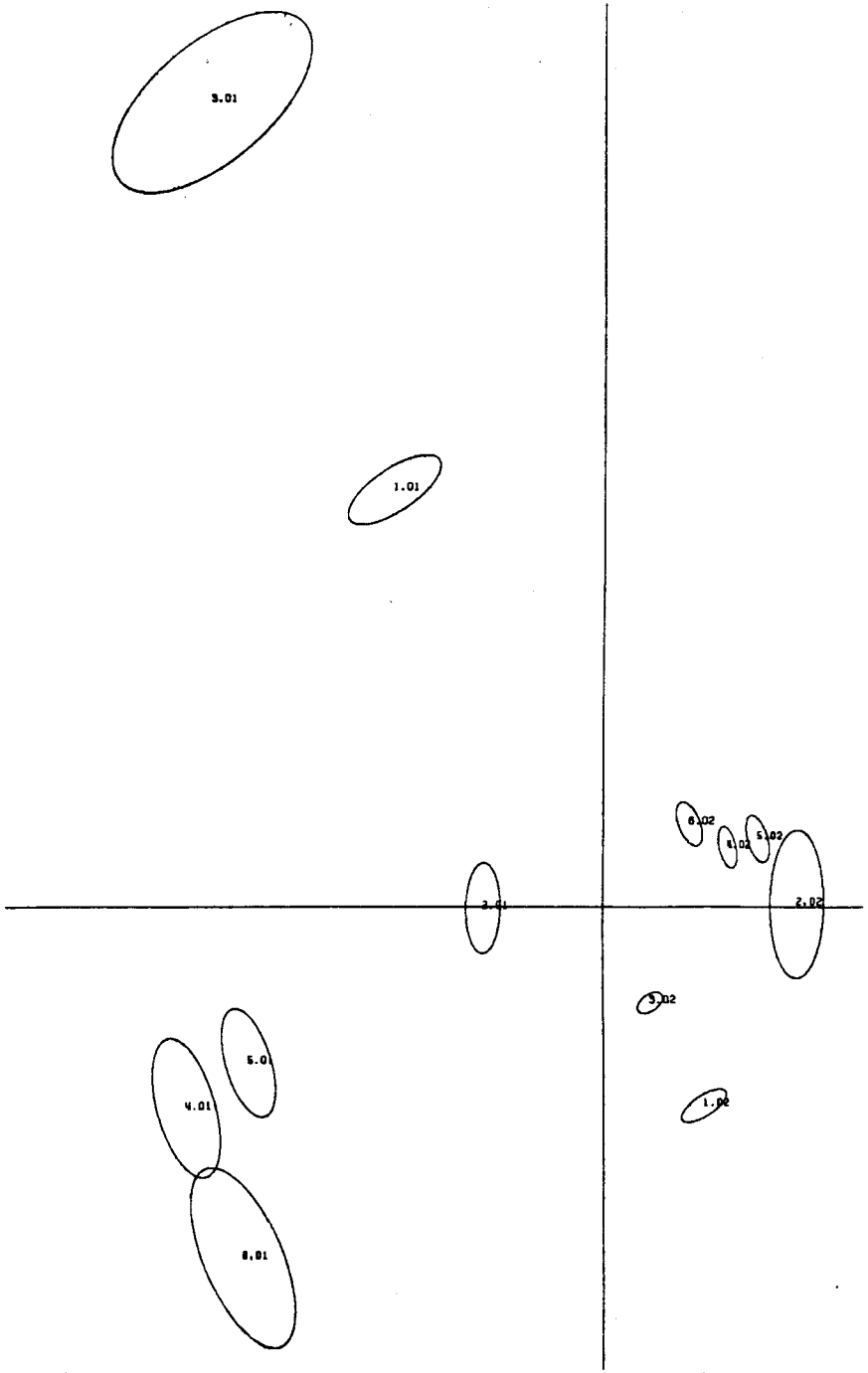


Figure 12.2 ANAPROF solution of the Sugiyama data with 95% confidence ellipses

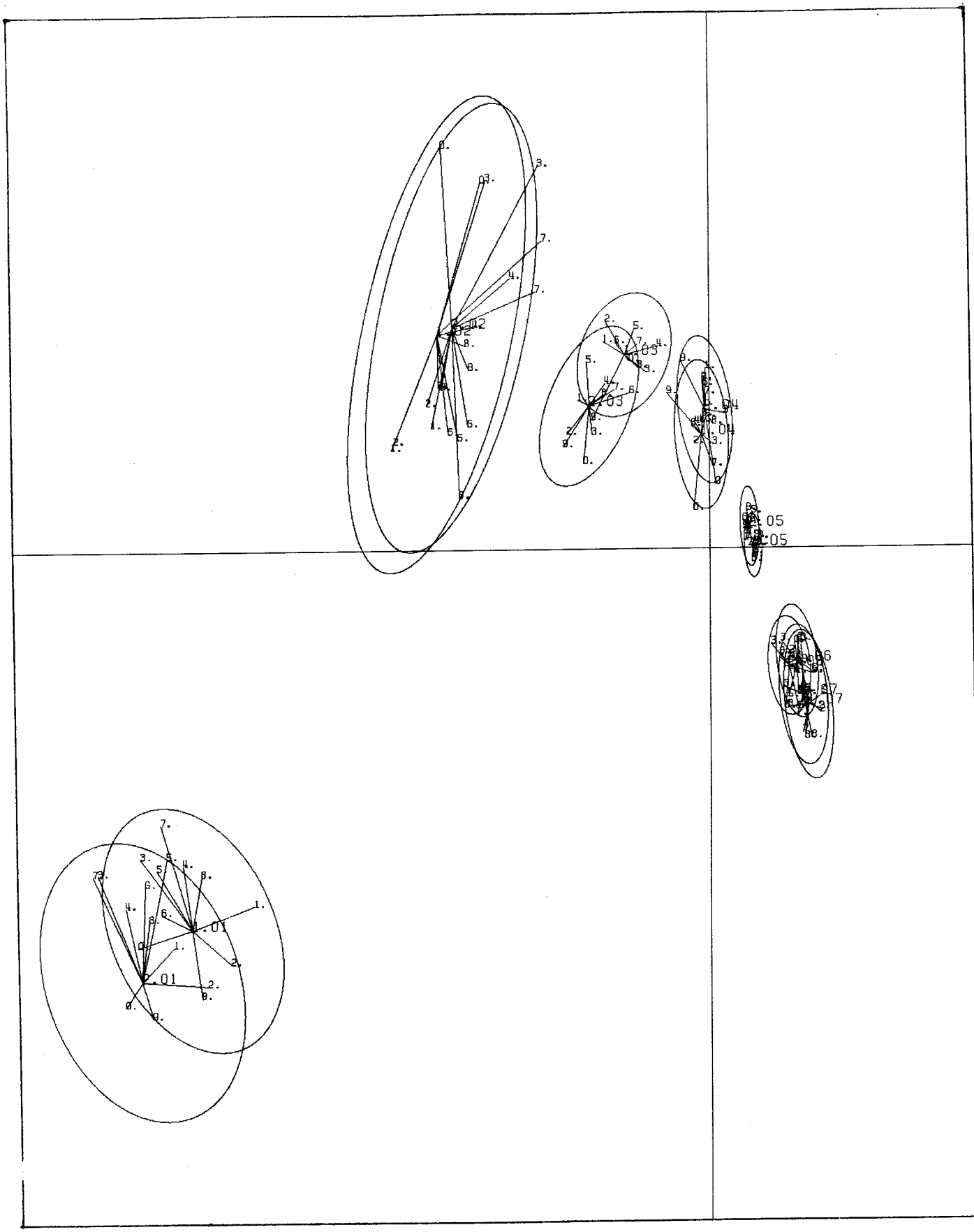


Figure 12.3 Stability of the ANACOR solution of the social mobility data: 95%-ellipses and 10 bootstraps

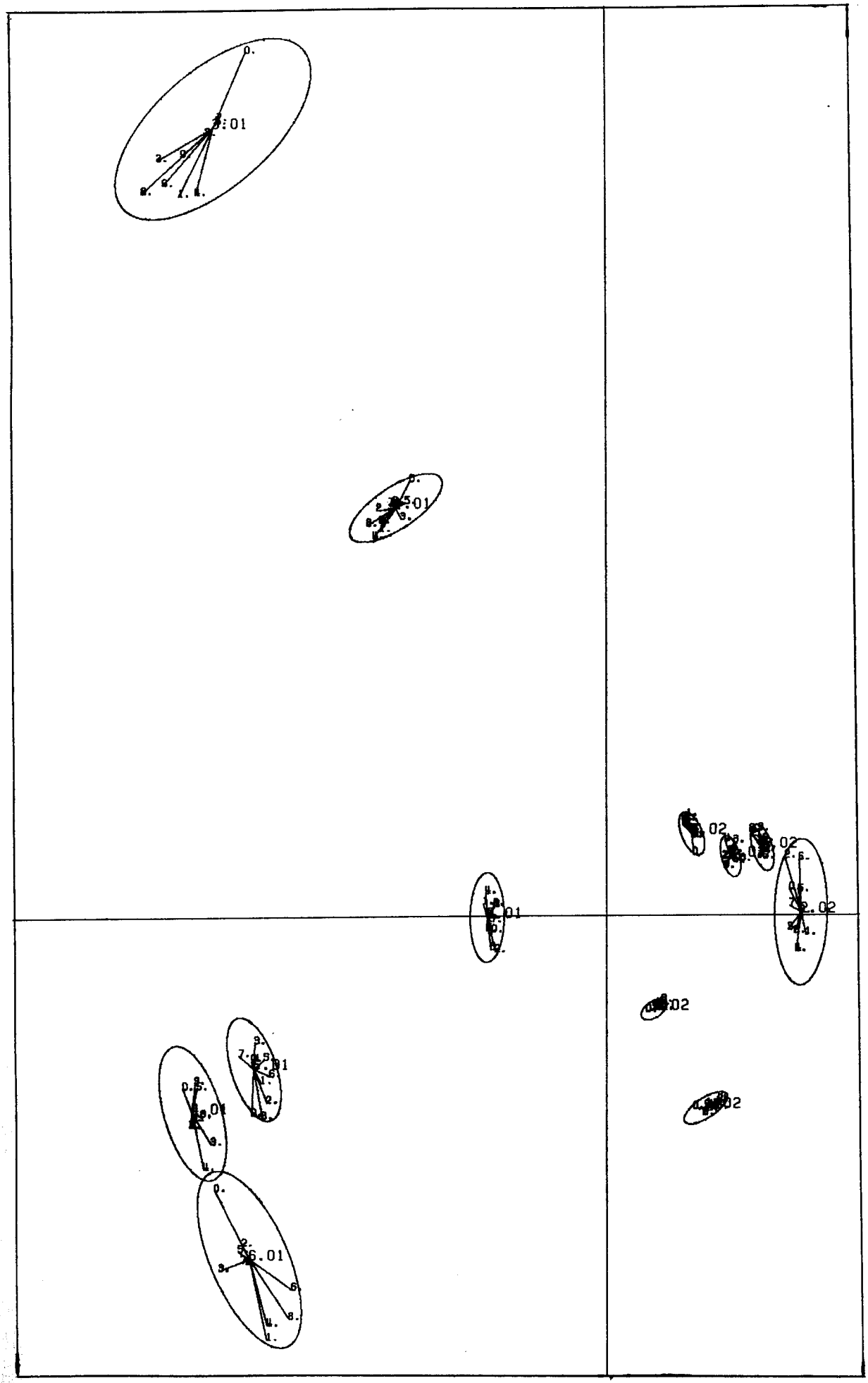


Figure 12.4 Stability of the ANAPROF solution of the Sugiyama data: 95%-ellipses and 10 bootstraps

13.1 Introduction

This chapter gives five examples of application of non-linear data analysis to real-life data. Purpose of the chapter is to show what kind of results the various techniques do produce, what kind of plots can be made, how such plots should be interpreted, etc. Purpose is also, to discuss some of the typical problems often met with in practice. For all examples the discussion remains focused on aspects of the data analysis, and not on details of substantive interpretation.

It might be emphasized right here that we do not claim that the analysis presented in the examples is better than any other analysis that could be applied to the same data.

The five examples are:

13.2 Multiple choice examination,

13.3 Abortion survey,

13.4 "From Year to Year",

13.5 Parliament survey,

13.6 Crime and fear.

13.2 Multiple choice examination

13.2.1 Introduction.

A multiple choice examination consists of items, or questions, where for each item the individual can make a choice from a number of alternative answers, one of which is "best". Every examiner knows that it is difficult to construct "wrong" answers that nevertheless have some plausibility (so that the student really has to know the subject-matter in order to reject the "wrong" alternative). Every examiner also knows that it often happens that an alternative supposed to be wrong, nevertheless is chosen by many of the "best" students. This indicates a sort of conflict between "apriori" and "aposteriori" evaluation of items.

The present example refers to a multiple choice examination on "Introduction to Psychology", taken by 190 pregraduate students of Psychology in Leiden, January 1980. There were 30 items, each of them with four response alternatives. The indicator matrix therefore becomes a 190 x 120 matrix, with 30 matrices G_j, each of them having 4 categories.

13.2.2 Data.

Table 13.2.1 gives the basic data. This table lists for each question the marginal frequencies of the four response alternatives, with an additional column for "missing data". The correct alternative is marked with the symbol "=". A "+" before the number of the question means that the item had been used in earlier examinations (such items ought to be "better" than new items that have not been validated earlier). The right half of the table summarizes results in terms of correct/incorrect

answers, with missing counted as incorrect. The table also gives columns for discrimination measures; they will be discussed in the next section.

13.2.3 HOMALS, one dimension.

The results of a one-dimensional HOMALS solution are shown in figure 13.2.1. This figure, for each item, plots the quantification of the four alternatives; the figure also indicates which alternative is "correct" (using the label C for the correct alternative and the lable W for the three wrong ones). This first HOMALS dimension is scaled in such a way that low (negative) quantification goes with "correct". One should expect, therefore, that in figure 13.2.1 for each question the category labelled G should be lower than the other three categories labelled F. This is, in fact, the case for most items, but not for all. Exceptions are items 5 and 19. This results agrees with the discrimination measures shown in table 13.2.1 (left half): items 5 and 19

4 categories							2 categories (correct/wrong)				
nr	marginal frequencies					discr.	nr	marginal frequencies			discr.
	M	1	2	3	4			M	1-correct	2-wrong	
+ 1	0	16	24	10	140=	.103	+ 1	0	140	50	.105
+ 2	0	11	147=	25	7	.285	+ 2	0	147	43	.264
3	0	17	34	32	107=	.067	3	0	107	83	.088
+ 4	0	22	9	17	142=	.068	+ 4	0	142	48	.041
5	0	102=	0	81	7	.001	5	0	102	88	.004
6	0	1	29	68	92=	.284	6	0	92	98	.155
7	0	4	44	52	90=	.264	7	0	90	100	.182
8	0	49	11	100=	30	.182	8	0	100	90	.214
9	0	1	153=	31	5	.062	9	0	153	37	.102
+ 10	0	109=	26	2	53	.136	+ 10	0	109	81	.113
11	0	26	121=	39	4	.260	11	0	121	69	.195
+ 12	0	11	36	5	138=	.201	+ 12	0	138	52	.183
13	0	90	4	92=	4	.081	13	0	92	98	.018
14	0	19	9	36	126=	.026	14	0	126	64	.021
15	0	40	14	28	108=	.134	15	0	108	82	.128
+ 16	0	5	5	144=	36	.302	+ 16	0	144	46	.211
17	0	106=	48	27	9	.074	17	0	106	84	.110
18	0	5	17	15	153=	.262	18	0	153	37	.144
+ 19	0	66	1	102=	21	.058	+ 19	0	102	88	.000
20	1	3	35	138=	13	.252	20	0	138	52	.228
+ 21	0	12	47	103=	28	.093	+ 21	0	103	87	.102
22	0	22	38	10	120=	.102	22	0	120	70	.105
23	0	8	11	148=	23	.150	23	0	148	42	.142
24	0	44	126=	6	14	.154	24	0	126	64	.116
25	0	154=	11	10	15	.380	25	0	154	36	.213
+ 26	1	103=	29	35	22	.058	+ 26	0	103	87	.040
27	0	13	135=	37	5	.131	27	0	135	55	.102
28	0	117=	28	39	6	.163	28	0	117	73	.198
29	0	5	167=	9	9	.377	29	0	167	23	.290
+ 30	0	27	149=	11	3	.309	+ 30	0	149	41	.293
hom.						.1673	hom.				.1369

Table 13.2.1 Marginal missing entries frequencies, discrimination measures and homogeneity for the multiple choice examination

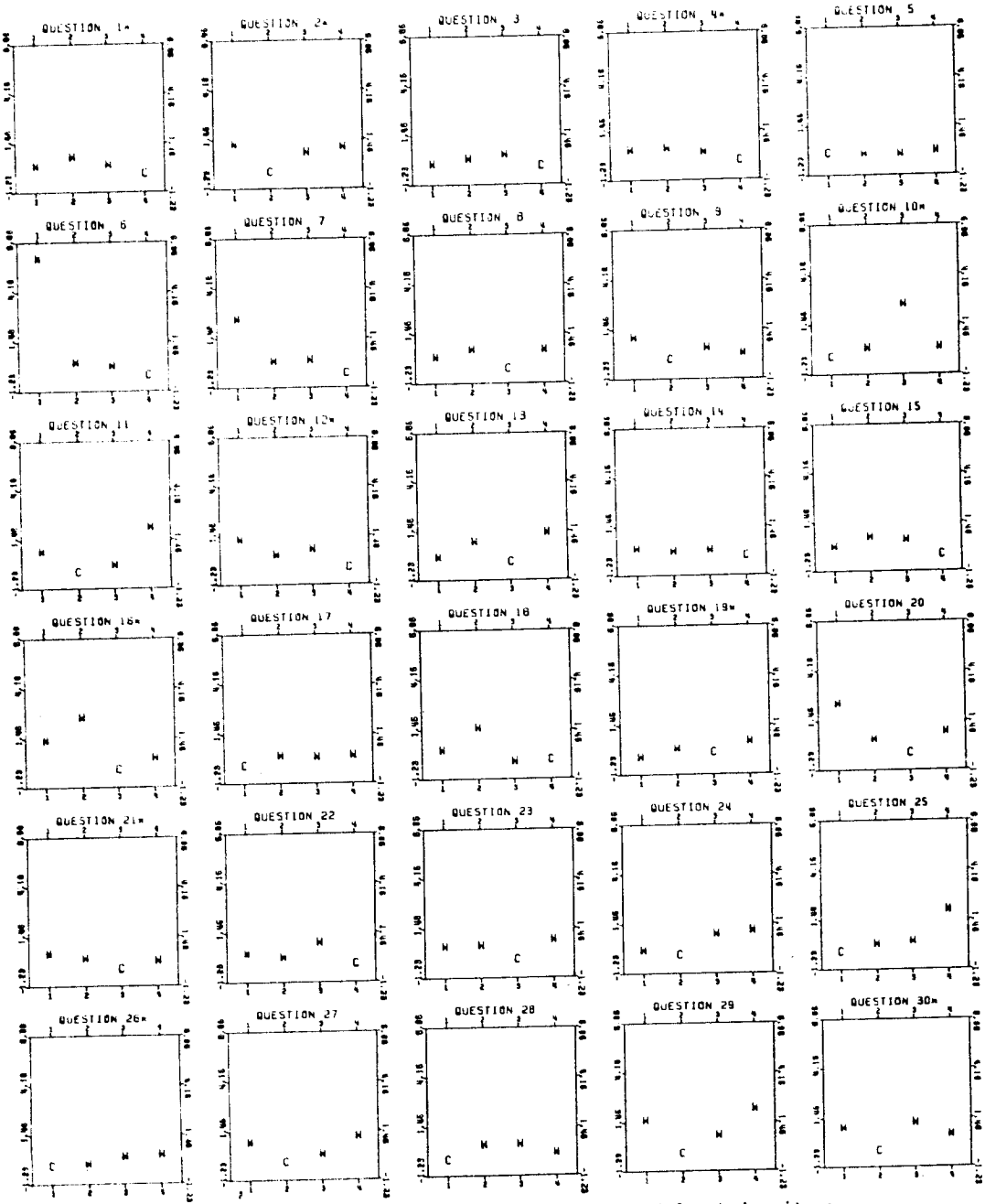


Figure 13.2.1 HOMALS optimal scaling of categories of multiple choice items.
C=correct answer, W=wrong answer

have low discrimination measures. Items 5 and 19 obviously do not discriminate between individuals with many correct answers on other items, versus individuals with few correct answers; these two items could as well have been skipped from the examination.
Figure 13.2.2 plots HOMALS individual scores against total number of correct answers. The plot shows a high correlation. Nevertheless, there are some discrepancies. An example is that the individual with lowest HOMALS score (the best HOMALS result) has 27 correct responses, whereas there are three individuals with higher number of correct responses. The explanation is simple: the individual with best HOMALS score made three "errors", on items 5, 13, and 19. These are items with low discrimination measure, and the individual's wrong answers are precisely those answers selected by most individuals with high scores.

13.2.4 HOMALS, grouped categories.

An alternative analysis is that incorrect answers are grouped into one category, as in the right part of table 13.2.1. This means that the HOMALS indicator matrix becomes a 190 x 60 matrix, with two columns for each item. The result for the first HOMALS solution, is shown in figure 13.2.3, again in the form of a plot against total number of correct answers. Discrimination measures, listed in the right half of table 13.2.1, again show that items 5 and 19 are among the poor ones.
On the whole, discrimination measures for the analysis based on two categories are smaller than those for the analysis with four categories. This, however, is partly an artificial result (see footnote to section 3.10). Clearly, HOMALS with responses grouped into two categories, gives less information than HOMALS with four categories. Once 'incorrect' answers are grouped, we shall never be able to discriminate between incorrect answers that are blatantly incorrect, and incorrect answers that are only slightly incorrect (in the sense that they are often chosen by individuals with many correct answers on other items). In the example, item 13 has correct answer in category c, but this answer competes with the incorrect alternative a which is far "less incorrect" than the other two categories b and d. Such a result indicates, in fact, that the examiner might better have a close look at item 13.

13.2.5 Reversed indicator matrix.

HOMALS also has been applied to the reversed indicator matrix for these data, as if individuals were sorting items into four categories. This reversed indicator matrix is a 30 x 760 matrix.

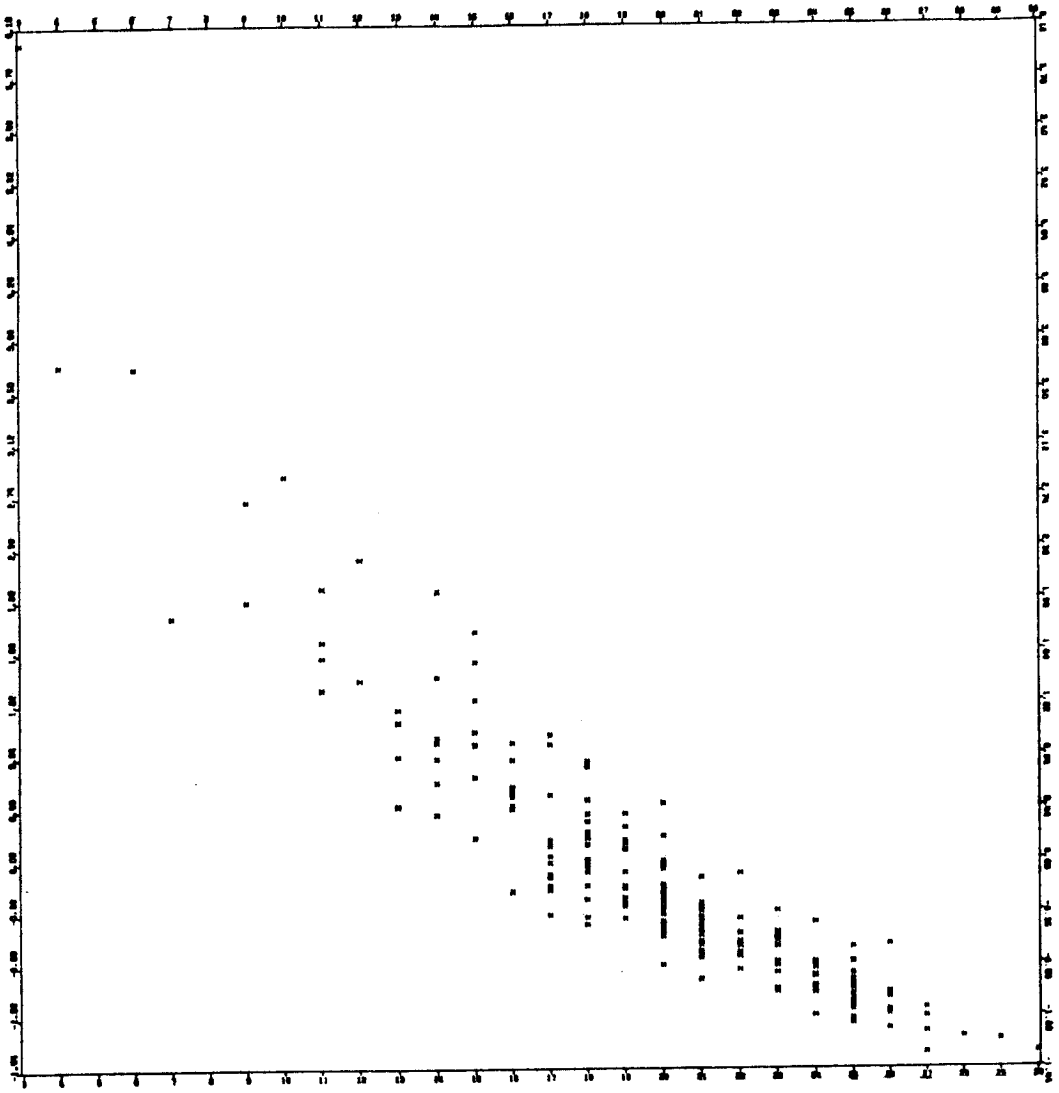


Figure 13.2.2 Individual HOMALS scores (4 categories) plotted against number of correct answers

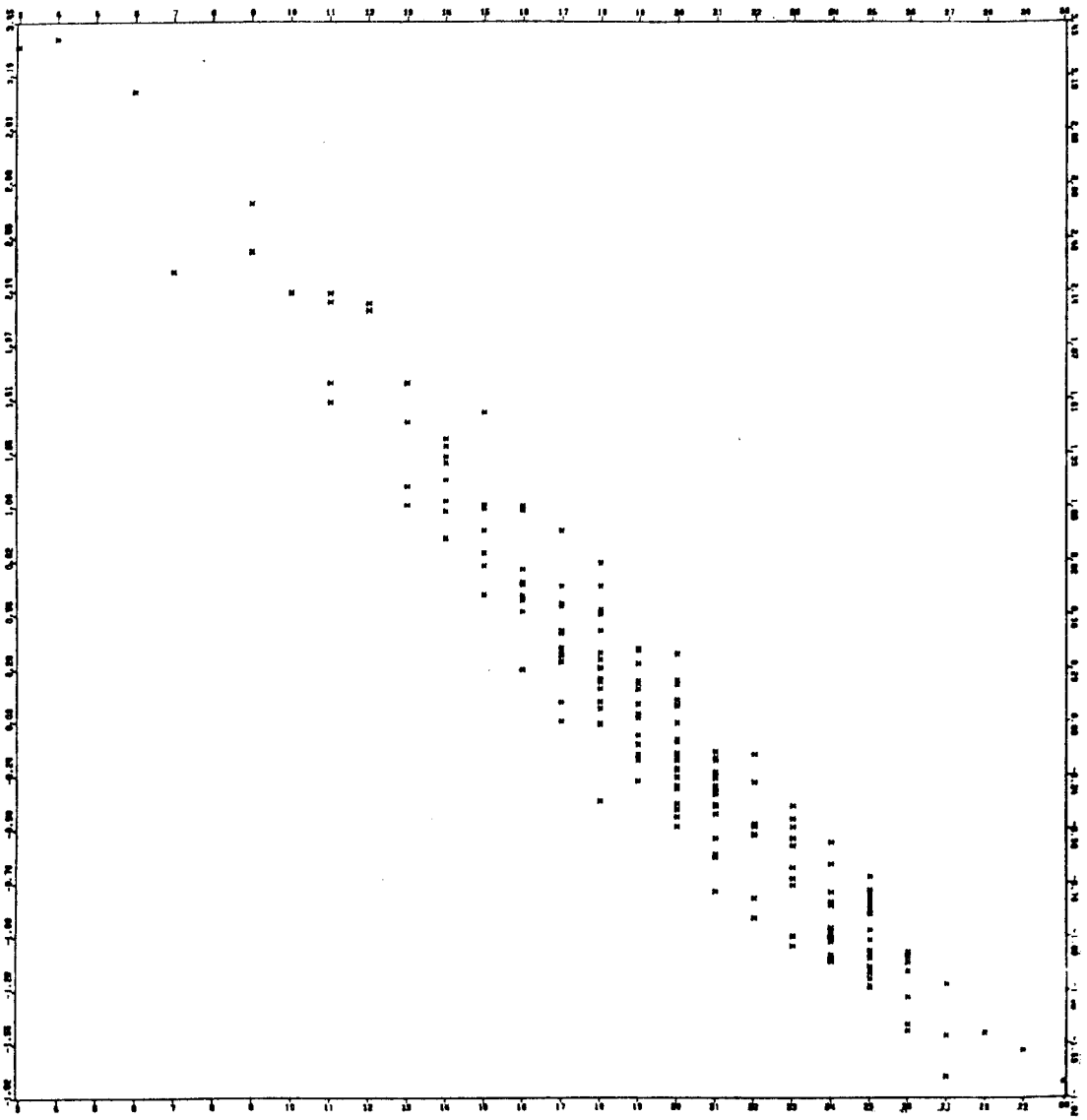


Figure 13.2.3 Individual HOMALS scores (2 categories) plotted against number of correct answers

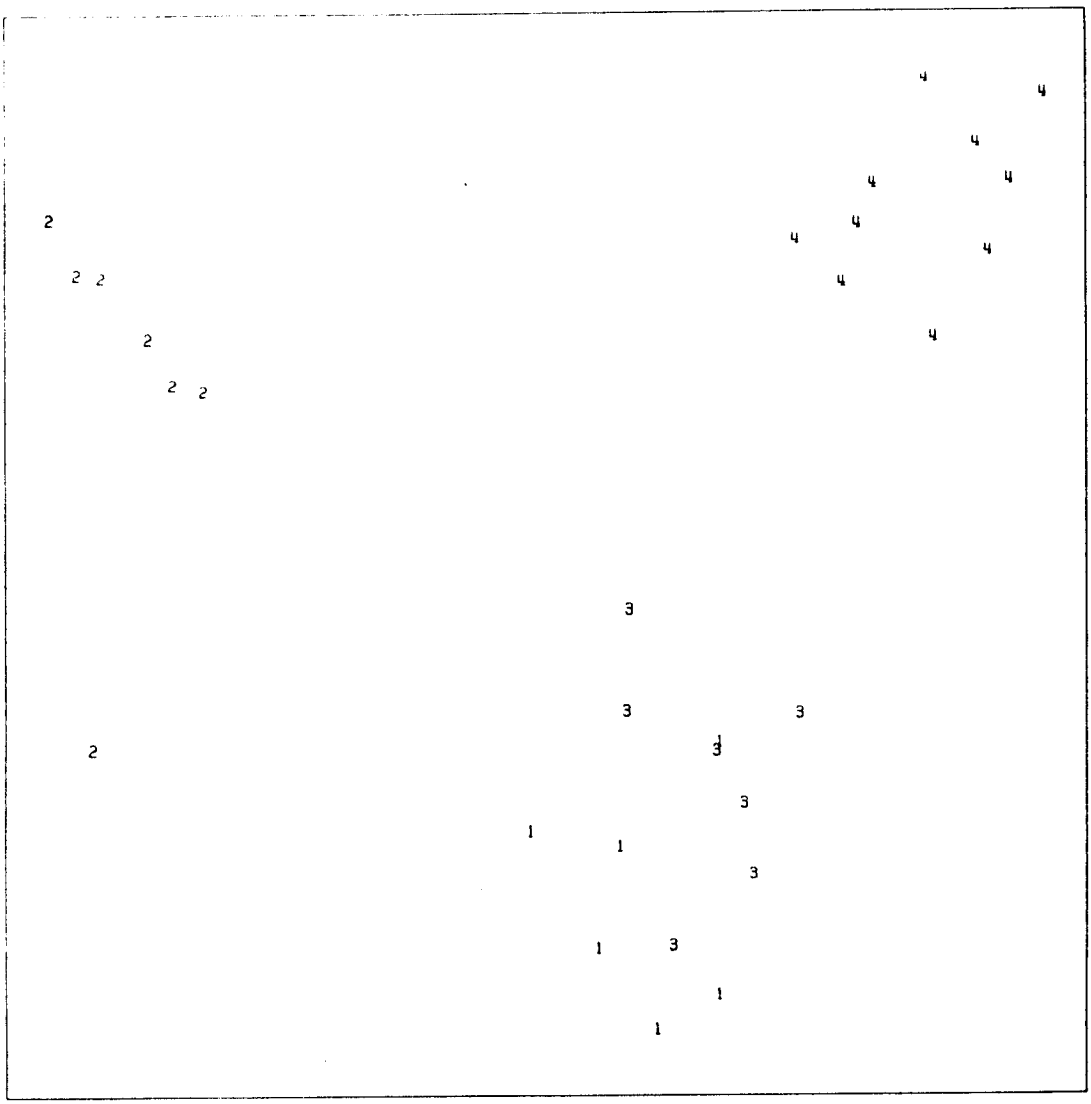


Figure 13.2.4 HOMALS item scores: First two dimensions of analysis of reversed indicator matrix

Results are trivial. Since individuals tend to select the correct answer in the first place, they sort items into four groups: one group where category a is the correct answer, a second group with category b as the correct answer, a third group wehere c is correct, a fourth group where d is correct. Figure 13.2.4 plots results for the first two HOMALS dimensions on this reversed indicator matrix. The plot shows that items with b or d as the correct answer form clusters (in fact, category d is more often correct than any other category) - one should expect that category a and category c will form similar clusters in subsequent HOMALS dimensions. Analysis of the reversed indicator matrix gives discrimination measures for individuals. Such discrimination measures, for the first two HOMALS dimensions, will be correlated with individual total number of correct answers, to the extent that the latter depends on items with category b or d as the correct answer.
In fact, the third HOMALS dimension for the reversed indicator matrix (numerical results are not shown) reveals a discrimination between items with a or c as the correct category, with items 5 and 19 (the items that discriminate poorly) excepted.
In conclusion, HOMALS on the reversed indicator matrix for this example turns out to be a rather naive idea. Although the results of the analysis tell us something about individuals (in terms of discrimination measures), they tell us more about how the examiner allocated correct answers over the categories a to d.

13.3 Abortion survey

13.3.1 Introduction

The abortion data are the results of a survey among 575 respondents over 37 variables, in 1974. A description of the variables is given in section 13.3.7. That section also gives a table of marginal frequencies.

13.3.2 HOMALS one dimensional.

Table 13.3.5 gives the discrimination measures for the first HOMALS solution. Results show that variables CP 1,2 and 3 do not discriminate, and that of the eight background variables only REL and POL are related to the first HOMALS solution. Figure 13.3.1 gives a plot of the quantification of the remaining 28 variables. The plot shows that the HOMALS solution is related to such basic variables as A11-14 and that low values are in the direction of "in favour of legal abortion" whereas high values are in the "anti-abortion" direction. The quantification of POL shows that 'left' and 'liberal' are on the 'pro-abortion' side, CDA (the combination of denominational parties) is 'anti'. Other results of figure 13.3.1 speak for themselves.

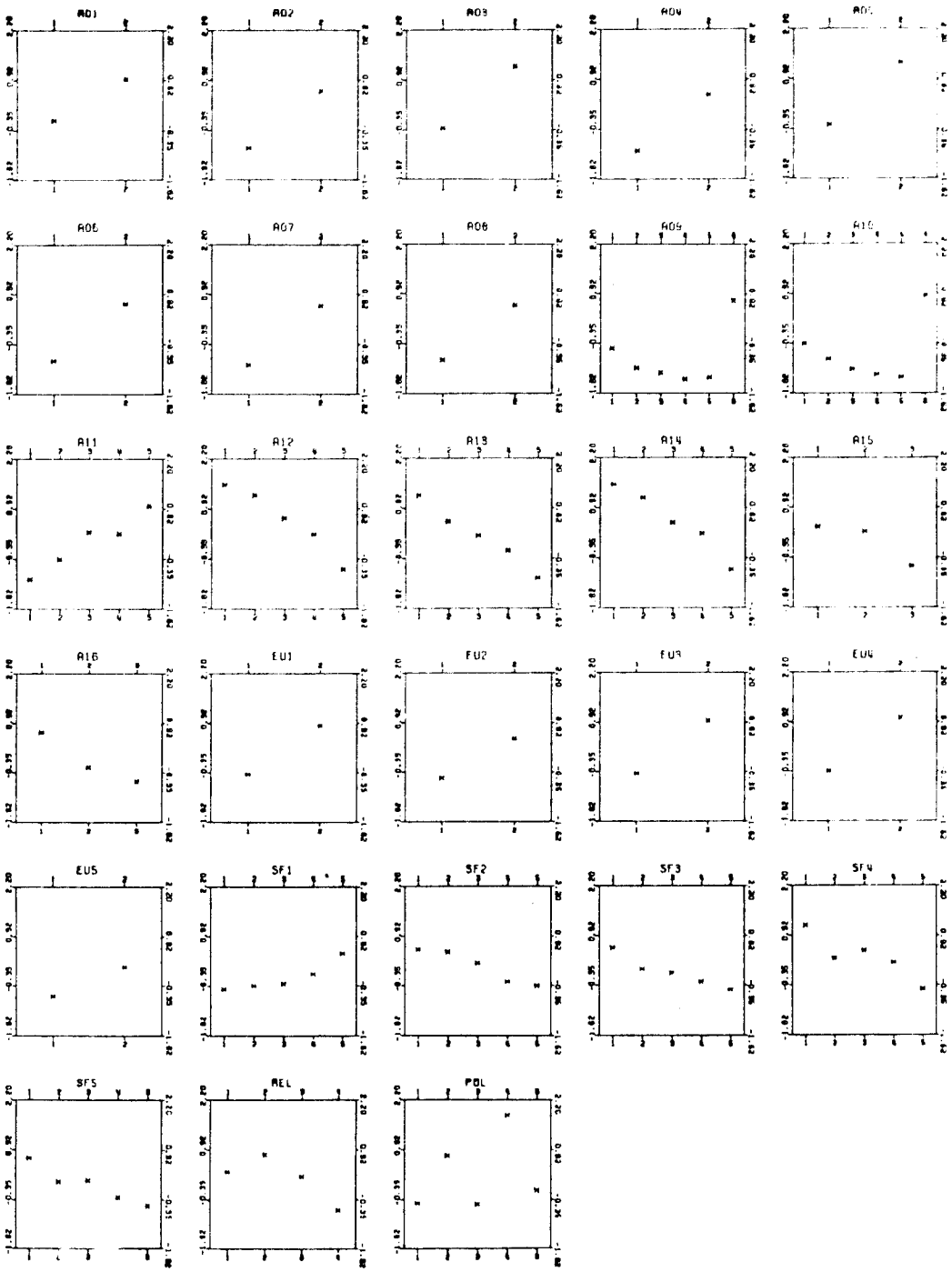


Figure 13.3.1. HOMALS optimal scaling of categories Abortion study

13.3.3 Likert scales.

A Likert scale consists of a number of statements to each of which the respondent can answer with a choice out of a number of response categories (usually five) ranging from 'strongly approve' to 'strongly disapprove'. It has become current practice to evaluate such items on the basis of item-total correlations, and to calculate a respondent's score on the scale only on the basis of the items that reach the criterion of sufficiently high item-total correlation, whereas items that fail to reach the criterion are ignored.

In the abortion data 11 items are candidate for a Likert scale: A9-14, and SF1-5. Table 13.3.1A shows their intercorrelations and item-total correlations, after reversal of items A11 and SF1. Also, variables A9 and A10 needed apriori re-coding: response category 6 rejects abortion under all circumstances, but categories 1 to 5 in this order become more "favourable" towards abortion; their order must be reversed in order to become consistent with category 6. The average squared item-total correlation in table 13.3.1A is .4109.

	A09	A10	A11	A12	A13	A14	SF1	SF2	SF3	SF4	SF5	I-T
A09		.55	.39	.35	.41	.37	.27	.24	.25	.26	.20	.61
A10			.46	.37	.43	.41	.23	.19	.22	.27	.20	.64
A11				.50	.57	.49	.24	.14	.18	.26	.21	.65
A12					.71	.68	.26	.27	.28	.36	.30	.71
A13						.70	.28	.27	.28	.39	.34	.77
A14							.23	.26	.25	.38	.38	.72
SF1								.18	.30	.25	.30	.52
SF2									.37	.39	.39	.52
SF3										.37	.54	.58
SF4											.50	.63
SF5												.63
											homogeneity	.4109

Table 13.3.1A Intercorrelations and item-total correlations original scores (after corrections).

One dimensional HOMALS over the same set of 11 variables gives the quantification shown in figure 13.3,2, with correlations shown in table 13.3.1B. Note, first of all, that the HOMALS quantification by itself takes care of the reversal of items A11 and SF1. Also, HOMALS very nicely takes care of the recoding in variables A9 and A10, with the first five categories going downwards but category 6 obtaining the highest quantification. Secondly, for the HOMALS solution the average squared item-total correlation is .4541, slightly better than for the Likert solution. In fact, HOMALS maximizes this index (it is the HOMALS eigenvalue ϕ).

	A09	A10	A11	A12	A13	A14	SF1	SF2	SF3	SF4	SF5	I-T
A09		.65	.52	.49	.56	.51	.32	.31	.27	.35	.29	.73
A10			.58	.52	.59	.57	.27	.23	.25	.36	.29	.75
A11				.55	.58	.51	.26	.15	.19	.28	.24	.69
A12					.73	.68	.29	.28	.28	.40	.32	.78
A13						.71	.30	.29	.30	.41	.36	.83
A14							.27	.27	.25	.40	.40	.79
SF1								.15	.30	.28	.29	.47
SF2									.36	.43	.40	.49
SF3										.38	.54	.52
SF4											.51	.63
SF5												.60
										homogeneity		.4541

Table 13.3.1B Intercorrelations and item-total correlations after HOMALS quantification.

Thirdly, the Likert solution would calculate individual scores as the sum of the item scores. In HOMALS, however, the individual score becomes a weighted sum of item scores: items with better discrimination contribute more than items with poor discrimination.

Fourthly, the Likert procedure assumes that variables are measured on interval scale level. The HOMALS solution can be interpreted as a test of that assumption: if the assumption is correct, the plots in figure 13.3.2 (after re-ordering, when necessary, such as in A9 and A10) should be linear. We find that this is roughly true for a number of items, but not for all of them. In SF1, e.g., the difference between categories 2 and 3 is negligible (they might as well have been grouped).

Fifthly, in the Likert solution the A items have larger item-total correlations than the SF items. This tendency is even stronger in the HOMALS solution, where the item-total correlations for SF become smaller than in the Likert solution. This indicates that the 11 items cannot be one-dimensional: there must be a second dimension for which the SF items are better indicators. To explore this further, PCA was applied to both correlation matrices of table 13.3.1, with results shown in figure 13.3.3. The two solutions are very similar, apart from a slight "rotation" (the first component in HOMALS is somewhat closer to the A-items, the first component of the Likert solution somewhat closer to the SF-items).

In conclusion: the HOMALS approach has a number of advantages over the Likert approach. It takes care of reversals and possible re-ordering of item categories, it does not assume interval scale level of measurement, it produces optimal scale values for individuals. On the other hand, the results here also demonstrate that the net gain of HOMALS over the Likert solution is small.

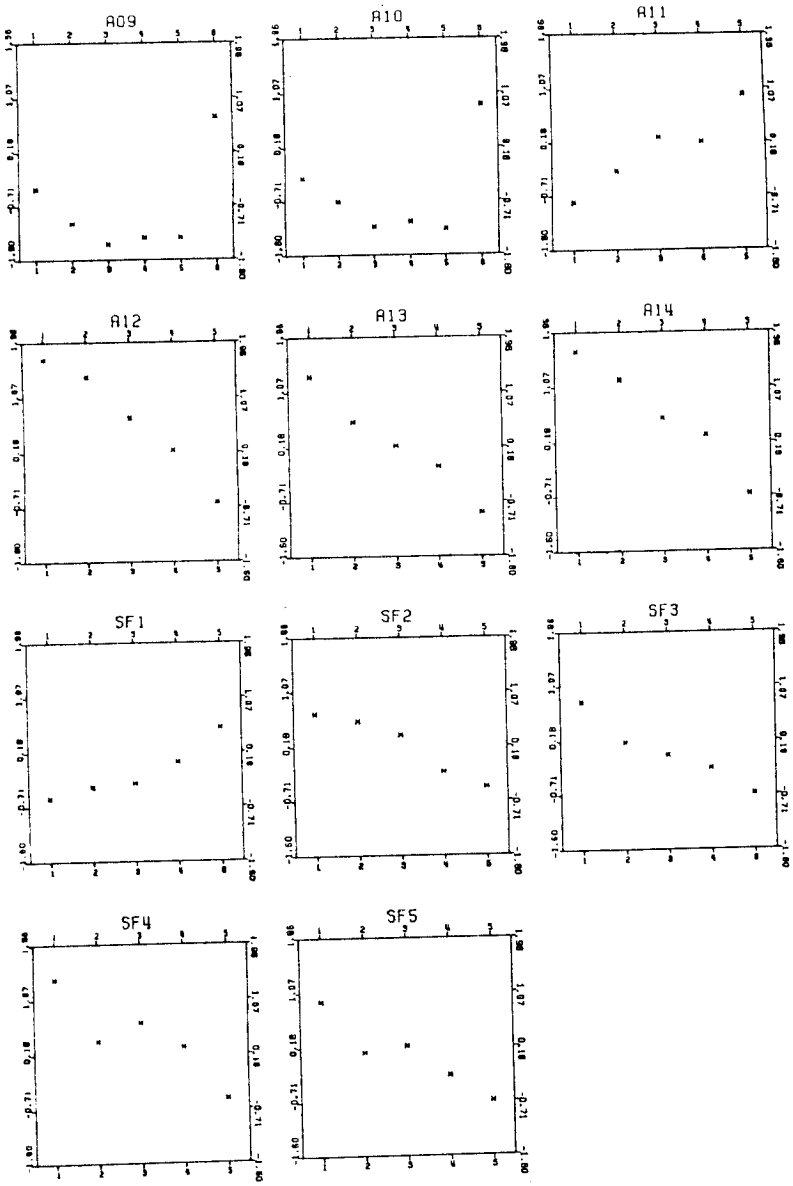


Figure 13.3.2 Optimal scaling of 'Likert' items.

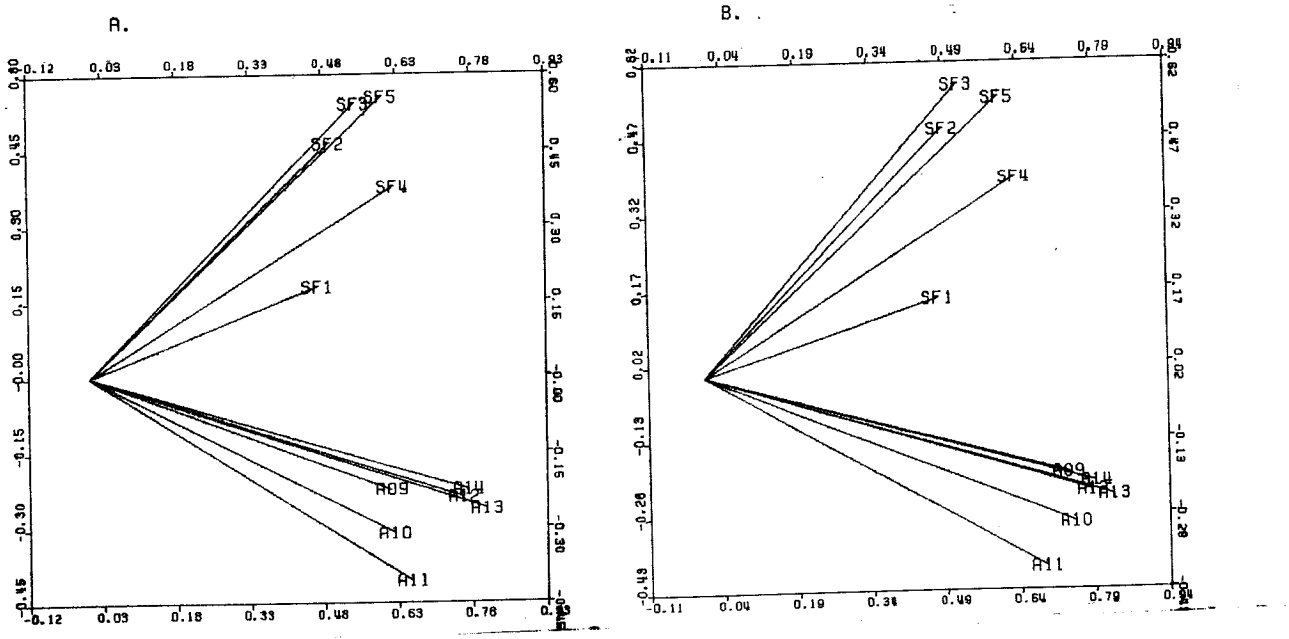


Figure 13.3.3 PCA results, first two dimensions for initial data matrix (A) and optimally scaled data matrix (B)

13.3.4 Guttman scales.

A Guttman scale consists of binary items and assumes that only a limited number of response patterns can occur (see sections 2.4.3, 4.6). Very often items will not fit a Guttman scale perfectly, in the sense that there are response patterns that "should not have been there". Guttman (1946) suggested the 'coefficient of reproducibility' (REP) an as index of how well empirical data approximate a perfect Guttman scale. The index is defined

$$REP = 1 - \frac{\text{number of errors}}{\text{total number of responses}}$$

where an 'error' is defined as a reversal one has to make in order to change an anomalous response pattern into an acceptable one. If REP is large (larger than .85, is the rule of thumb), one may assume that anomalous patterns are a matter of error and that the hypothesis of a one-dimensional underlying structure can be maintained.

An objection against REP is that the Guttman scale will order the items in such a way that the number of errors is minimized, so that REP always will be larger than zero. The minimum of REP often will be larger than .5 (Mokken, 1971, p. 50-54). Define MMR (coefficient of minimum marginal reproducibility) as

$$MMR = \frac{\text{total number of responses in modal categories}}{\text{total number of responses}}$$

where 'a response in a modal category' is a response in the category most often used. This leads to other rules of thumb, such as Edwards (1957) rule that not only REP should be larger than .85, but also substantially larger than MMR.

Starting from the other end, one might define a coefficient of scalability

$$S = 1 - E/E_o$$

where E is the total number of errors, and E_o the expected number of errors if responses were completely at random (Niemöller, 1976). This approach implies a probabilistic rather than a deterministic model. The idea has been further developed by Mokken (1971). The'Mokken scale'orders items according to proportion p_i of positive answers to item i. But the scale accepts items only if they obey "double monotony": if p_i is increasing with i, then for each fixed item h the probability of a positive answer to both h and i must be increasing with i, and the probability of a negative response to both h and i must be decreasing with i. An index for scalability is H_i, essentially the ratio between the sum of the covariances between i and all other items, and the sum of the maximum covariances allowed by the marginal proportions.

In the abortion data, variables A1-8 qualify as binary items. Their Guttman scale solution is given in table 13.3.2. Biserial item-total correlations order the items as

7 - 4 - 6 - 2 - 8 - 5 - 3 - 1

The HOMALS solution is given in table 13.3.3; the discrimination measure (squared item-total correlation) orders the items as

7 - 4 - 6 - 2 - 8 - 3 - 5 - 1

so that in this respect results are quite comparable (the more so since in both solutions items 3 and 5 are close together).

Results for the Mokken scale are given in table 13.3.4. As to "degree of difficulty" results are block-wise comparable to the Guttman and Homals ordering. But as to scalability results are very much different.

13.3.5 HOMALS, two dimensions.

For the 28 variables of figure 13.3.1

a two-dimensional HOMALS solution for categories is plotted in figure 13.3.4. Results show the rather typical 'horse shoe', where the second dimension is a quadratic function of the first, and as such has no other substantive interpretation than that it contrasts categories in the middle with those at the extremes. Going from upper right downwards to the middle, and from there to the upper left, one finds categories ordered from 'anti-abortion' and political 'right', towards 'pro legalization of abortion' and political 'left' or 'liberal'. Categories for variables SF (sexual freedom) follow roughly the same pattern, but because of their lower discrimination measures are located more towards the center. Variable EU5 (very low discrimination measure) has all its categories near the center of the plot.

Figure 13.3.5 plots the corresponding points for individuals, with a similar, but more blurred, horse shoe pattern.

13.3.6 PRINCALS.

PRINCALS has been applied to 26 variables (those of 13.3.5, with A1 and EU5 dropped because of their low discrimination). In this analysis the variables REL and POL obtained multiple nominal scaling (each category as a single nominal variable). The typical Likert items (A11-16, SF1-5) were treated as single ordinal. For binary variables (A2-8, EU2-4) the choice of measurement level is irrelevant. For the special variables A9 and A10 single nominal scaling was required.

Figure 13.3.6 gives a plot of the weights (correlations) of the variables in the two-dimensional PRINCALS solution. The plot also shows the boundary hyperplanes (lines) for the categories of A9 and SF5. Note that

ITEM	A01		A02		A03		A04		A05		A06		A07		A08		TOTAL
RESP	A	D	A	D	A	D	A	D	A	D	A	D	A	D	A	D	
	ERR		ERR		ERR		ERR		ERR		ERR		ERR		ERR		
DISAGREE 8	0	17	0	17	0	17	0	17	0	17	0	17	0	17	0	17	17
	ERR																
7	40	4	3	41	1	43	0	44	0	44	0	44	0	44	0	44	44
			ERR														
6	38	6	36	8	14	30	0	44	0	44	0	44	0	44	0	44	44
					ERR												
5	97	4	91	10	98	3	4	97	6	95	4	97	1	100	2	99	101
							ERR										
4	56	4	57	3	57	3	17	43	21	39	20	40	8	52	4	56	60
									ERR								
3	47	1	47	1	45	3	31	17	23	25	21	27	16	32	10	38	48
											ERR						
2	45	5	49	1	46	4	36	14	38	12	30	20	34	16	22	28	50
													ERR				
1	33	3	36	0	34	2	32	4	23	13	29	7	35	1	30	6	36
															ERR		
0	143	0	143	0	143	0	143	0	143	0	143	0	143	0	143	0	143
SUMS	499	44	462	81	438	105	263	280	254	289	247	296	237	306	211	332	543
PCTS	92	8	85	15	81	19	48	52	47	53	45	55	44	56	39	61	
ERROR	0	27	3	23	15	15	4	78	27	50	45	27	59	1	68	1	442
B.C.	.3556		.7143		.7083		.9049		.8438		.8668		.9568		.9284		

A - agree
D - disagree

coefficient of reproducibility = .8
minimum marginal reproducibility = .6
percent improvement = .2
coefficient of scalability = .6

Table 13.3.2 Guttman scale solution for Abortion data

	agree	disagree	discr.
A01	-0.07	0.82	0.058
A02	-0.88	0.72	0.623
A03	-0.29	1.22	0.342
A04	-1.05	0.65	0.673
A05	-0.24	1.33	0.308
A06	-0.84	0.78	0.657
A07	-0.98	0.74	0.718
A08	-0.84	0.72	0.598
hom.			0.497

Table 13.3.3 HOMALS solution

var	difficulty	H(I)
A01	.9183	.3928
A02	.4452	.6450
A03	.8000	.6943
A04	.3774	.7434
A05	.8435	.7203
A06	.4783	.6824
A07	.4243	.7152
A08	.4504	.6312
	scale	.6730

Table 13.3.4 Mokken scale

The resulting Mokken scale depends on the order in which the items are selected. This order is : (A05,A07), A04,A03,A06,A02,A08,A01

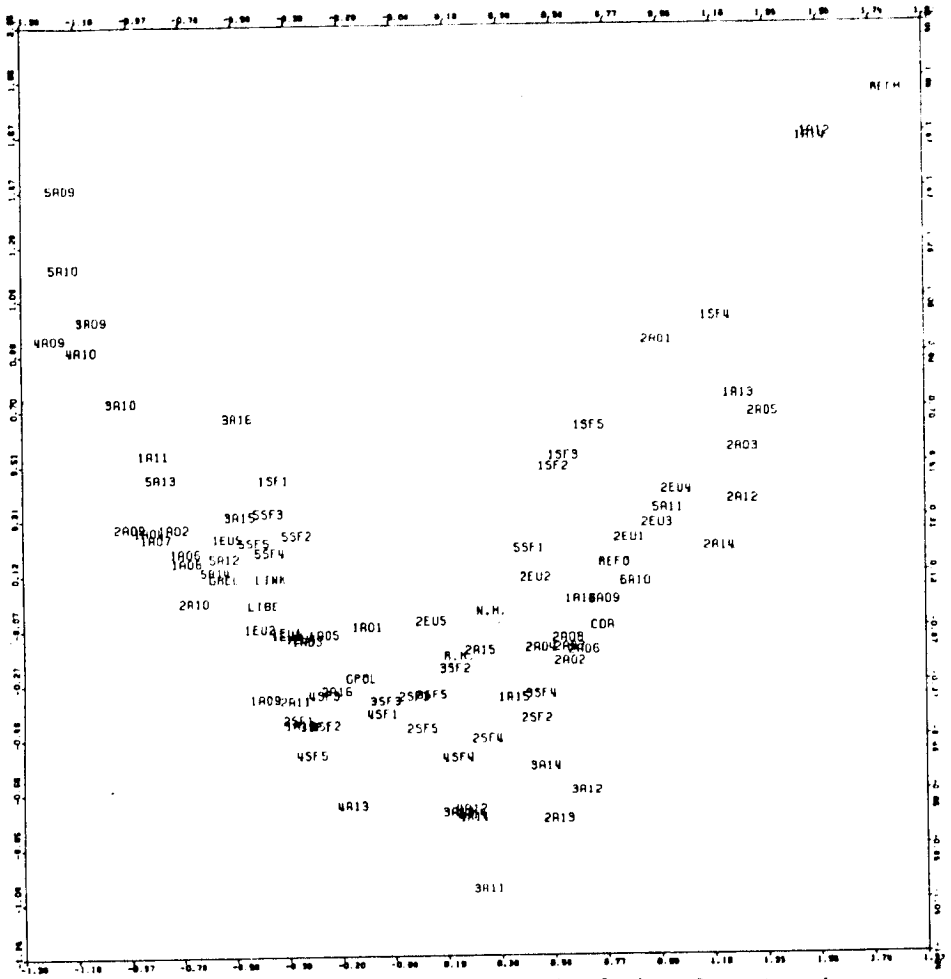


Figure 13.3.4 Abortion data. First two HOMALS solutions for categories

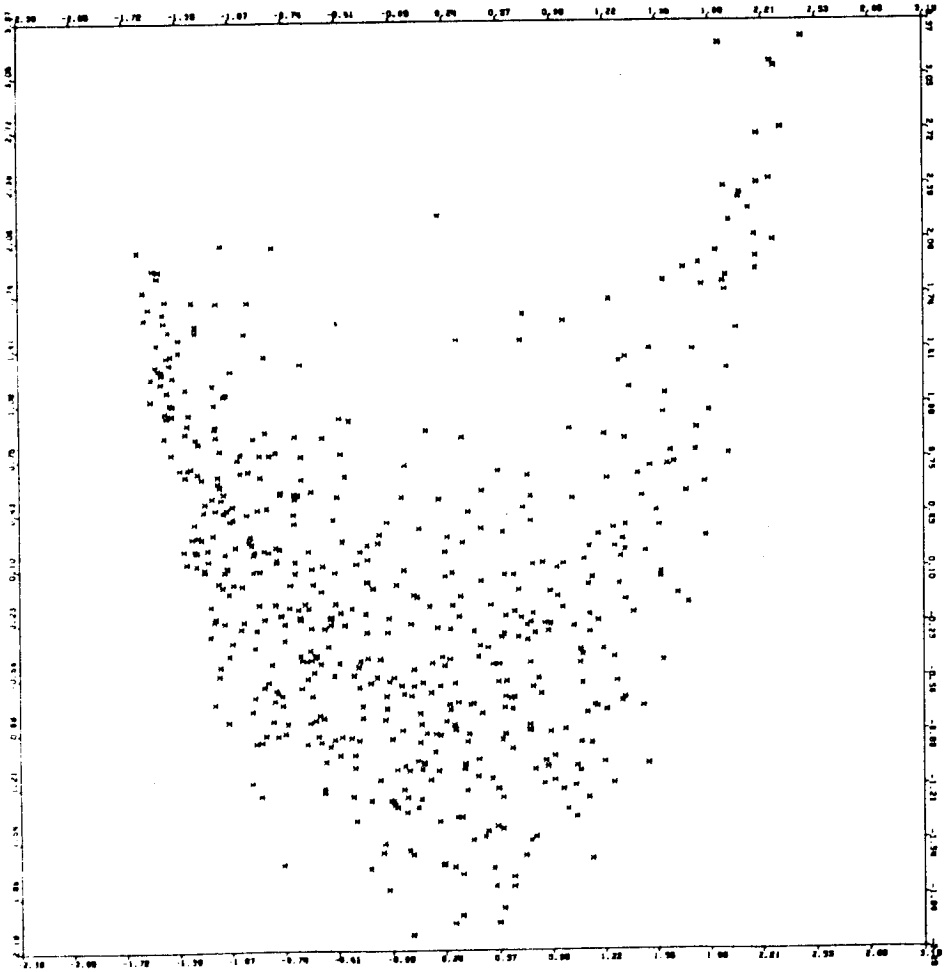


Figure 13.3.5 Abortion data. First two HOMALS solutions for individual scores

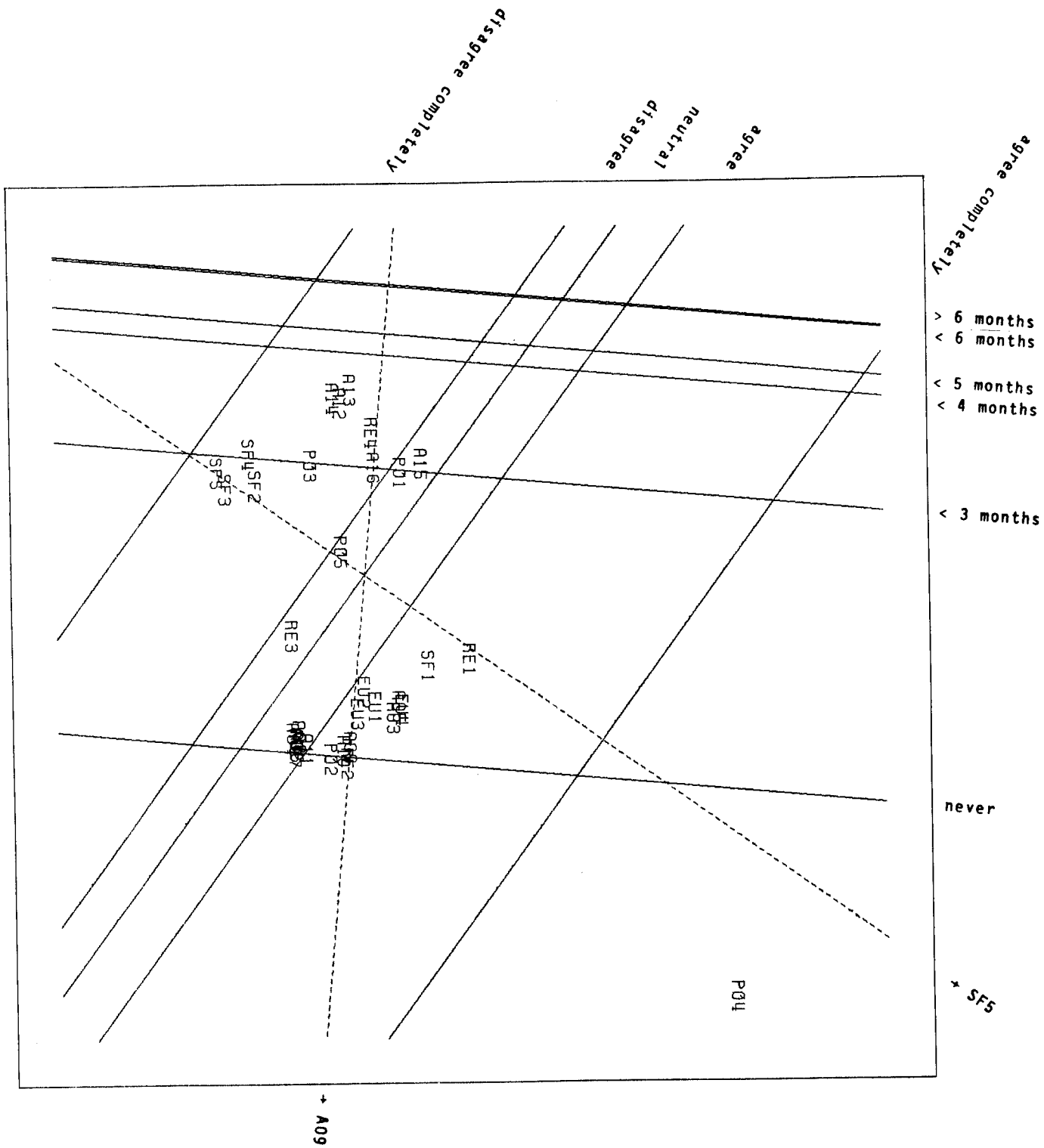


Figure 13.3.6 Two-dimensional PRINCALS solution Abortion data. Drawn lines separate between categories of variables A09 and SF5.

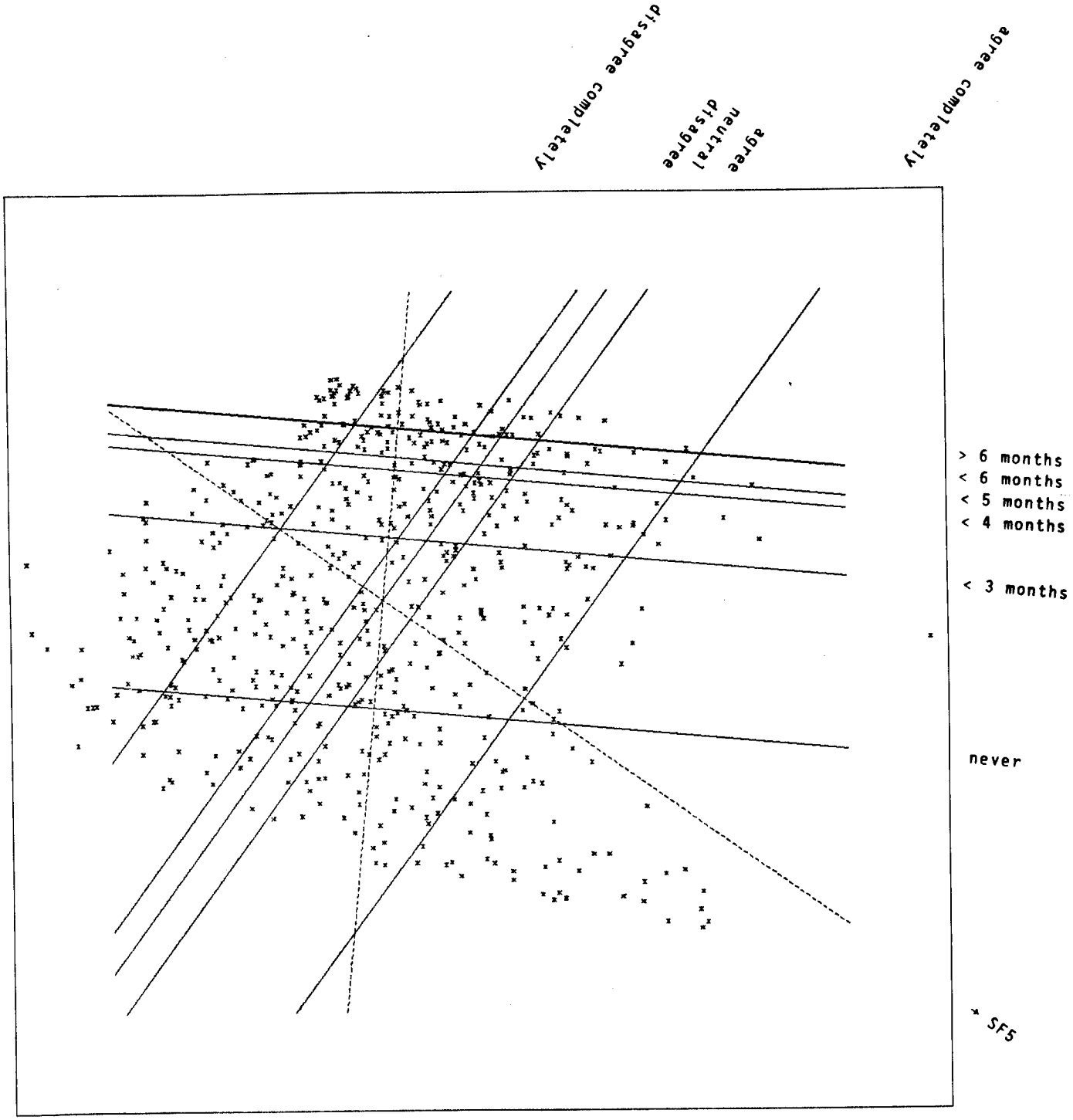


Figure 13.3.7 Two-dimensional PRINCALS, individual scores. Drawn lines give optimal separation between

	LEIDEN					A'DAM VU					A'DAM GU					UTRECHT				
JUR	13	32	9	2	18	1	0	0	1	1	3	14	12	2	10	7	16	1	0	5
MED	3	20	7	4	10	7	1	3	1	2	4	6	4	7	18	6	11	9	2	11
W&N	2	15	4	5	15	6	0	1	0	1	1	4	13	4	19	10	14	16	7	21
SOC	5	12	4	3	14	2	0	1	0	2	1	8	28	11	19	6	4	5	10	11
LET	5	8	6	3	5	2	0	0	1	0	7	8	16	8	16	3	10	4	4	13
TEC																				
PSW											1	1	5	0	4					
VEE																3	6	1	1	5
TND																0	3	0	0	1
THE	1	1	3	2	0	5	0	0	0	0	1	0	0	0	0	5	0	0	2	1
LBW																				
CIF	0	0	1	2	0	0	0	0	0	1	2	2	6	5	7	0	1	0	1	4
ECO						4	0	0	0	2	7	16	8	6	20					

	WAGENINGEN					GRONINGEN					DRIENENOORD					NIJMEGEN				
JUR						2	8	6	0	7						5	7	3	1	6
MED						3	7	3	2	11						6	9	3	2	5
W&N						7	5	9	2	12						7	2	2	4	6
SOC	0	1	1	0	0	4	5	11	6	10						4	2	5	4	22
LET						1	5	6	2	6						4	0	8	2	4
TEC											0	7	3	4	5					
PSW																				
VEE																				
TND						1	2	0	1	1						1	2	0	0	2
THE						1	0	2	0	1						6	0	0	0	5
LBW	11	14	7	6	14															
CIF						0	1	1	1	0						0	0	1	0	0
ECO	0	0	0	0	1	1	13	5	0	7										

	TILBURG					ROTTERDAM					EINDHOVEN					DELFT				
JUR	0	0	1	0	2	2	2	1	0	0										
MED						8	7	1	1	7										
W&N																				
SOC	5	1	0	3	6	1	1	3	1	1										
LET																				
TEC											12	7	3	0	13	24	66	22	20	50
PSW																				
VEE																				
TND																				
THE																				
LBW																				
CIF																				
ECO	3	11	2	0	9	9	33	15	2	16										

Table 4.9 Political preference (discussed on page 149)

Order: CDA,VVD,PvdA,PACO,D'66

A9 is mainly related to the first PRINCALS dimension, whereas SF5 also is related to the second. This confirms results of the Likert scale analysis of 13.3.3 above. Note also that PRINCALS, too, re-orders the categories of item A9. Figure 13.3.7 plots the individual scores in the same way. Although, with a bit of good will, one can still discern the horse shoe pattern in it, this pattern is much less outspoken than in the two-dimensional HOMALS solution. The HOMALS solution, as it were, absorbs the SF items in the curve of the horse shoe, whereas PRINCALS tends to give these items a dimension of their own.

As a consequence, in figure 13.3.7, one finds at the upper right individuals who are anti-abortion, politically at the right, religious, versus in the upper left individuals who "favour" abortion, are politically left or liberal, and non-religious. In the lower middle one finds individuals who favour sexual freedom to some extent, but oppose abortion.

13.3.7 Description of the abortion survey.

Data were collected in 1974.

Reference: Veenhoven and Hentenaar (1975). Total number of respondents: 575. Number of variables: 37. Marginal frequencies, number of missing data, discrimination measures first HOMALS dimension, are given in table 13.3.5.

Description of the variables:

3 variables with respect to capital punishment (CP1-CP3)

16 questions about abortion (A01-A16)

5 questions about euthanasia (EU1-EU5)

5 questions about sexual freedom (SF1-SF5)

8 background variables.

CP: three statements on capital punishment. Response from (1)=agree completely to (5)=disagree completely.
CP1: Taking hostages should be punishable by death
CP2: Murder should be punished by death
CP3: Killing people in time of war is justifiable

A01-A08. Eight statements. Response (1)=agree, (2)=disagree. All eight statements start with "I find abortion justifiable if"; they continue with:
A01: prolonged pregnancy is a danger to the mother's life or health
A02: the woman wants it for whichever reason, and if there are no medical counterindications
A03: there is a high chance that the child will be deformed or handicapped
A04: the expectant mother is unmarried and does not want to marry the father of the child
A05: the pregnancy is the result of rape or indecent attack
A06: the woman has a large family already and it is undesirable to have more children
A07: the expectant mother is unmarried and not able to marry the father of the child
A08: there are chances that the child will have an unhappy childhood because its parents do not really love it.

NAME	NR	M	1	2	3	4	5	6	7	8	9	10	11	DIM.
CP1	1	1	188	129	61	113	83							.026
CP2	2	1	167	112	77	108	110							.019
CP3	3	3	86	131	89	108	158							.014
A01	4	1	528	46										.082
A02	5	7	256	312										.505
A03	6	8	460	107										.372
A04	7	6	217	352										.467
A05	8	6	485	84										.307
A06	9	4	275	296										.515
A07	10	8	244	323										.539
A08	11	10	259	306										.461
A09	12	4	217	48	18	8	21	259						.544
A10	13	8	205	62	31	12	33	224						.562
A11	14	0	178	115	36	93	153							.537
A12	15	0	41	32	77	111	314							.558
A13	16	0	114	60	69	117	215							.638
A14	17	1	43	54	62	110	305							.584
A15	18	26	249	50	250									.225
A16	19	33	158	266	118									.219
EU1	20	6	396	173										.316
EU2	21	4	299	272										.247
EU3	22	14	405	156										.346
EU4	23	4	435	136										.333
EU5	24	7	104	464										.083
SF1	25	1	130	85	56	98	205							.168
SF2	26	2	84	67	85	115	222							.161
SF3	27	1	124	109	100	114	127							.156
SF4	28	1	49	42	56	126	301							.288
SF5	29	4	124	97	88	88	174							.221
SEX	30	1	248	326										.002
AGE	31	1	39	100	130	95	98	112						.046
SOC	32	0	17	23	65	120	201	75	70	4				.012
REL	33	19	103	62	168	223								.259
POL	34	0	176	128	92	20	159							.350
EDU	35	20	311	134	43	67								.041
FUN	36	3	3	27	8	13	12	70	84	56	33	21	245	.076
URB	37	1	44	44	40	166	65	64	151					.062
														hom .2734

Table 13.3.5 Missing entries, marginal frequencies and discrimination measures for Abortion data

A09-A10. Two statements. Responses: (1)=justifiable until 3 months, (2)=until 4, (3)=until 5, (4)=until 6 months, (5)=after six months, (6)=not justifiable.

A09: A woman, 45 of age, when menstruation fails to come, thinks menopause has started, and does not worry. Later she appears to be pregnant. She has a family with grown-up children. Until which month of pregnancy do you feel that abortion in this special case is still justified? Or is in your judgment abortion in this case not justified?

A10: A girl of 15 -unmarried- suspects she is pregnant. She is scared to talk about it with the family doctor or with her parents. As a result it takes much longer for her than necessary to enlist for medical aid. Until which month (etc., as in A09)

A11-A14. Four statements. Response from (1)=agree completely, to (5)= disagree completely.

A11: It is the woman's right to have abortion when she wants it

A12: Medical practitioners who perform abortion are not better than murderers

A13: People who agree with abortion have little respect for life

A14: Abortion is justifiable under no circumstances

A15.One statement with three response categories: (1)=abortion law, only special cases, (2)=law should make abortion difficult, (3)=no law, doctor decides.

A15: During the last years, up to now, politicians of different parties have been working on proposals with respect to abortion. In your judgment, should there be a law that allows for abortion in special cases only, or should there be a law that makes it difficult to have an abortion, or do you think there is no need for a law and that it is up to the doctor to decide whether or not he will help the woman?

A16.A statement with three response categories: (1)=after 12 weeks absolutely forbidden, (2)=after 12 weeks in special cases only, (3)=no time limit.

A16: Discussions on abortion during the last half of this year focused on whether abortion should be permitted or not after the 12th week of pregnancy. In your opinion, should there be a law that rules out abortion absolutely after 12 weeks, or a law that limits abortion after 12 weeks to special cases only, or do you say that the law should not specify a time limit for abortion.

EU1-EU5. Five statements on euthanasia. Responses: (1)=justifiable, (2)=unjustifiable. All statements begin with "I find euthanasia justifiable if"; they continue with

EU1: the ill person asks for it because he or she knows that the illness is terminal

EU2: close relatives ask for it, and the ill person is unconscious, whereas there is no hope for recovery

EU3: at the birth of a child it becomes evident that the child can be kept alive in a strictly technical medical sense but never will be able to have human contact

EU4: dying persons suffering from unbelievable pain in this way can be relieved from their misery

EU5: elderly people no longer are able to take care of themselves and express the wish they prefer to die.

SF1-SF5.Five statements on sexual freedom. Response from (1)=agree completelely, to (5)=disagree completely.

SF1: I don't object against children below the age of ten to walk around on the beach naked

SF2: If sexual intercourse would be separated from procreation it would soon become pure egoism

SF3: Parents should forbid children to have sexual play
SF4: Young people who have sexual intercourse before marriage do not have respect for each other
SF5: Parents should impress upon their children that it is better to have control over yourself and not to indulge in masturbation

Eight background questions.
SEX: (1)=male, (2)=female
AGE: (1)=below 20, (2)=between 20 and 30, etc., until (6)=above 60
SOC: social class, ranging from (1)=high to (8)=low
REL: religion. (1)=protestant, (2)=reformed, (3)=RC, (4)=none
POL: political preference. (1)=left, (2)=denominational, (3)=liberal, (4)=right, (5)=none
EDU: education level. (1)=LO,VGLO, (2)=ULO, (3)=VHMO, (4)=professional training or university
FUN: present profession or job. (1)=managerial, more than 10 employees, (2)=ibid., less than 10 employees, (3)=free profession, (4)=independent farmer, (5)=higher employees and civil servants, (6)=ibid.,middle level, (7)=ibid.,lower level, (8)=schooled workers, (9)=unskilled labour, (10)=students, (11)=house-wives.
URB: degree of urbanization. (1)=Amsterdam, (2)=Rotterdam, (3)=The Hague, (4)=medium size cities, (5)=small cities, (6)=industrialized rural, (7)=agricultural rural

13.4 From Year to Year

13.4.1 Introduction.

The study called "From Year to Year" is a longitudinal survey, meant to find out the determinants of the education level finally attained by children after they left elementary school. A description of the variables is given in section 13.4.7.The crucial variable is EIN. A number of the variables (BVA BIL INT URB BIM DLO ADV LL6 KGS PRE SEX) stem from a nation-wide survey held in 1965 over more than 11,000 children. Later, it was decided to do a follow-up study for a subgroup of some 2,000 children. This study was done in 1970, and collected data for the variables OPV OPM AKG ASO ASL OOA DWO BMB INS KLS TON AOS LLS EXT and also EIN for those children who by that time had left secondary school. For the other children EIN was measured in 1974. Ultimately, the complete data set contained data for 1845 children. It should be added, perhaps, that the categories of EIN (final level) exclusively refer to the level of secondary full-time education : part time education and tertiary education (MBO) have not been taken into account.

3.4.2 HOMALS, all variables.

Results of one-dimensional HOMALS over all variables are plotted in figure 13.4.1. Corresponding discrimination measures can be found in table 13.4.1. Looking at variables with high discrimination measure, it becomes evident that the first HOMALS dimension is related to school success (EIN TON PRE ADV).

We shall give a few comments in the quantification shown in figure 13.4.1. Firstly, HOMALS will, on the whole, quantify categories in such a way that their distribution becomes more similar to a normal distribution. This implies that for variables which already have a normal distribution, the HOMALS quantification will tend to be linear. Such is the case for the stanine variables BIL BIM PRE. On the other hand, HOMALS will correct for skewness (an example is ASO). Variable URB had an almost rectangular distribution before quantification; HOMALS corrects for this by making the difference between categories (1) and (2), and that between (3) and (4), larger than that between (2) and (3).

A second comment is that many of the nominal variables (BVA OPV OPM ADV TON EIN) are quantified as a monotonous function of the category numbers, sometimes even almost linear. This would justify an analysis in which interval scale level of measurement is assumed, as was the case in some of the earlier analyses (Dronkers, 1978; Dronkers and Jungbluth, 1979). But there are typical exceptions. One example is BVA (where categories 3 and 5 have to be interchanged, which, in hindsight, looks plausible enough). OPV and OPM also have some (interpretable) interchanges. For EIN we find some irregularities which cannot be understood so easily. Probably they are related (among other things) to the large differences in marginal frequency for the categories of EIN (for a more detailed discussion, see Stoop, 1980). Finally, the quantification of EXT nicely puts category 6 (no extracurricular activities) below category 1 (only one extracurricular activity), which demonstrates that HOMALS often automatically corrects for a somewhat awkward apriori numbering of categories.

13.4.3 HOMALS separated for subgroups of individuals.

Separate HOMALS analyses were done for boys and girls. Table 13.4.1 gives the discrimination measures, and figure 13.4.2 gives a plot of discrimination measures for boys versus girls. The plot indicates that AOS and BIM have higher discrimination for boys, whereas LLS and EXT have higher discrimination for girls; the differences, however, are far from dramatic.

Separate HOMALS solutions also were calculated for each of the categories of BVA (occupational level of father). Table 13.4.2 gives discrimination measures. On the whole, we note that variables characterizing secondary education (AOS LLS EXT) have increasing discrimination measure with increasing level of BVA. Rather large differences are found also in OPV ASO INT BMB.

In order to capture table 13.4.2 in a plot, ANACOR was applied to the table, with results shown in figure 13.4.3. The figure shows that

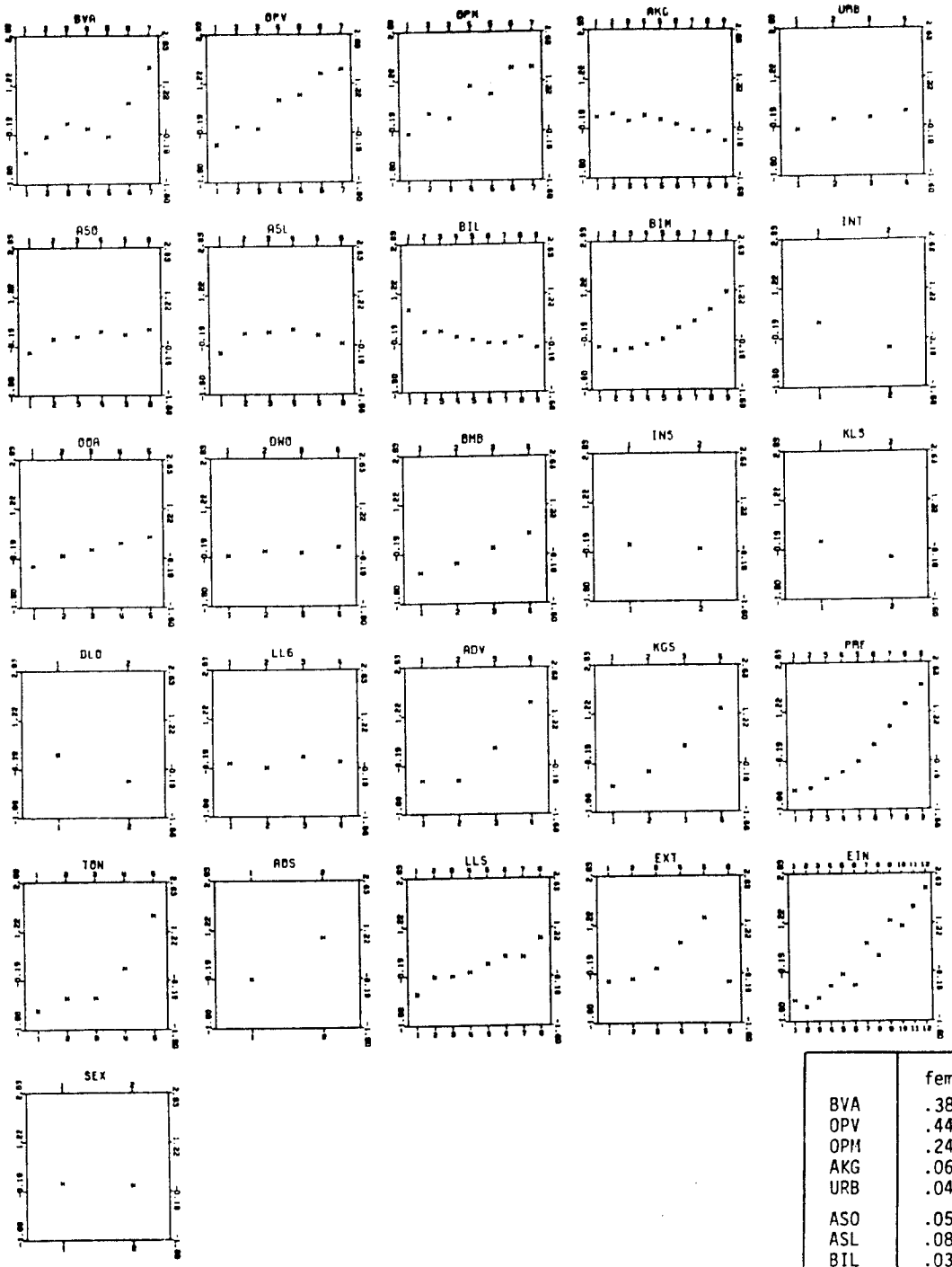


Figure 13.4.1.

From Year to Year. Optimal scaling of categories.

	female	male	total
BVA	.388	.376	.384
OPV	.442	.384	.411
OPM	.244	.274	.259
AKG	.065	.055	.055
URB	.044	.026	.034
ASO	.058	.041	.049
ASL	.088	.052	.067
BIL	.038	.112	.059
BIM	.134	.273	.186
INT	.125	.087	.109
OOA	.063	.055	.053
DWO	.011	.004	.008
BMB	.184	.153	.165
INS	.000	.009	.003
KLS	.028	.014	.021
DLO	.123	.129	.125
LL6	.016	.021	.015
ADV	.652	.670	.651
KGS	.386	.434	.412
PRE	.606	.645	.627
TON	.783	.776	.774
AOS	.137	.364	.218
LLS	.245	.107	.145
EXT	.440	.301	.359
EIN	.737	.709	.711
SEX	----	----	.001
hom.	.2414	.2428	.2269

Table 13.4.1 HOMALS discrimination

variables with high discrimination for all professional categories (EIN TON PRE ADV) are located in the center of the plot, whereas variables for which the discrimination measure varies a lot, have excentric location. Note that the professional categories in the plot move away from the center towards the variable with largest discrimination for that category: BV7 has moved towards ASO and LL6, BV3 towards BIM and INS, BV1 towards ASL, etc.

Figure 13.4.3 illustrates how information in a large table (table 13.4.2) can be summarized in a picture. One should be cautious, however, about the interpretation in this illustration. Discrimination measures are squared correlations between variables and the first HOMALS dimension, and changes in correlation may turn up for a variety of reasons. One of them is 'restriction of range'. Division of data into subgroups always entails the danger that restriction of range makes correlations lower.

13.4.4 MORALS.

As an illustration of MORALS, table 13.4.3 shows results of eight applications, all of them with EIN as the variable to be predicted, but with increasing number of variables in the predictor set. The first analysis makes use only of BVA OPV OPM AKG URB as predictors, in the second analysis the variables ASO ASL BIL BIM are added to the predictor set, etc. The idea is that the selection of variables follows the "time factor". Variables BVA OPV OPM AKG URB can be assessed, as it were, even before the child is born, whereas variables AOS LLS EXT are available only a short time before EIN itself can be assessed.

Table 13.4.3 shows that the multiple correlation increases to the extent more predictor variables are added. This, of course, is a trivial result; it could not be otherwise. More interesting is that the table also shows that for each application of MORALS the correlation between EIN and the predictor variables changes (this, of course, is not the case in linear multiple regression analysis). E.g., in the first MORALS the correlation between EIN and BVA equals .468; in the last analysis this correlation goes down to .195. That such correlations are subject to change stems from the fact that each MORALS gives a different quantification to all variables included in the analysis.

The table also shows that,especially in the last three analyses, predictor variables are added that correlate more highly with EIN than earlier variables. For instance, once TON enters the predictor set (and TON is highly correlated with EIN), all earlier variables become more or less irrelevant. In the table of regression weights TON then becomes the only

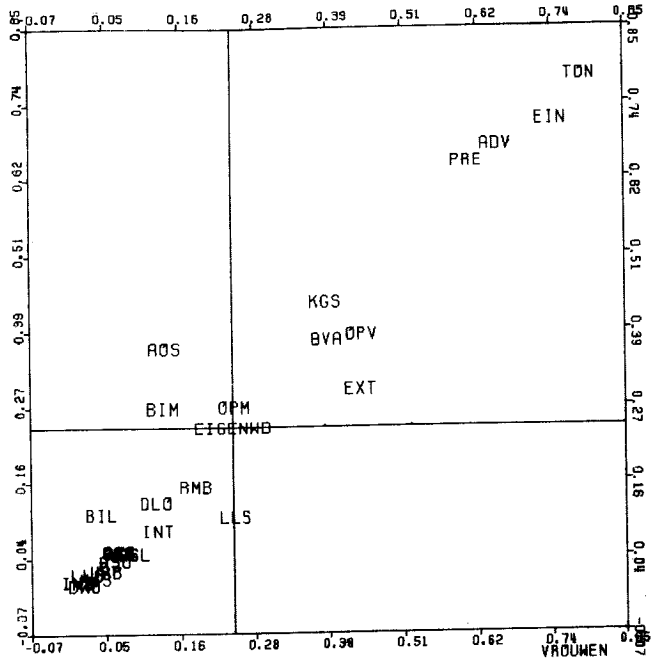


Figure 13.4.2 HOMALS discrimination measures for males (vert.) versus females (hor.).

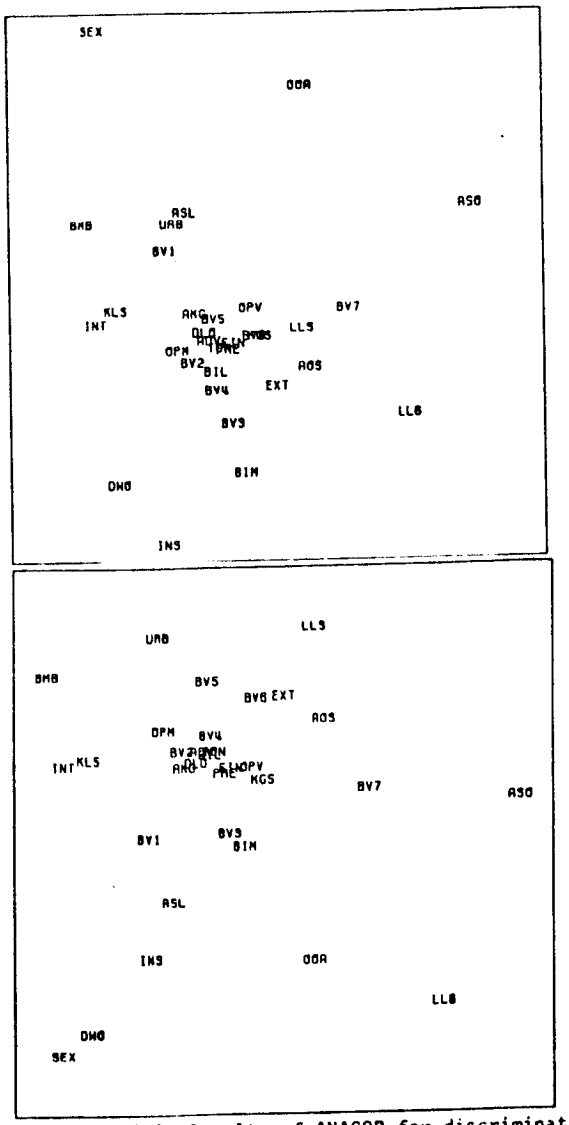


Figure 13.4.3 Results of ANACOR for discrimination measures of variables in seven professional groups A. Dimensions 1 and 2; B. Dimensions 1 and 3

	agricultural + unskilled labour	skilled labour	clerical	small business	farmers	manegerial	high manegerial + free professions
OPV	.174	.064	.150	.185	.188	.165	.202
OPM	.098	.105	.127	.080	.151	.116	.056
AKG	.078	.142	.045	.071	.083	.036	.088
URB	.033	.028	.000	.030	.068	.006	.028
ASO	.064	.011	.043	.030	.095	.087	.295
ASL	.137	.027	.072	.020	.058	.069	.063
BIL	.077	.074	.127	.106	.133	.081	.086
BIM	.090	.171	.339	.288	.095	.116	.194
INT	.130	.056	.091	.113	.129	.056	.004
OOA	.110	.027	.009	.006	.015	.040	.110
DWO	.035	.031	.078	.018	.011	.006	.004
BMB	.163	.099	.015	.093	.156	.100	.007
INS	.008	.013	.033	.016	.006	.001	.005
KLS	.028	.028	.001	.035	.000	.023	.006
DLO	.173	.090	.171	.162	.167	.191	.117
LL6	.019	.027	.088	.025	.006	.010	.108
ADV	.613	.683	.646	.633	.711	.684	.580
KGS	.308	.293	.389	.292	.317	.417	.447
PRE	.576	.635	.723	.632	.684	.611	.691
TON	.666	.758	.783	.768	.824	.799	.727
AOS	.093	.084	.185	.253	.149	.295	.278
LLS	.062	.165	.127	.189	.328	.299	.309
EXT	.118	.340	.392	.375	.412	.479	.445
EIN	.617	.668	.779	.616	.766	.713	.762
SEX	026	.011	.000	.001	.002	.001	.009
hom.	.1799	.1852	.2165	.2015	.2222	.2160	.2249

FINAL LEVEL

	Correlations								Regression weights							
BVA	.468	.455	.450	.450	.436	.411	.216	.195	.285	.236	.228	.229	.222	.122	.107	.069
OPV	.451	.439	.416	.416	.376	.376	.201	.088	.235	.218	.163	.161	.135	.082	.038	.015
OPM	.329	.326	.307	.304	.297	.306	.168	.070	.152	.133	.112	.110	.102	.049	.033	.019
AKG	.177	.178	.175	.175	.174	.165	.014	.064	.127	.123	.099	.100	.087	.072	.038	.013
URB	.065	.056	.009	.018	.001	.054	.004	.001	.021	.028	.027	.031	.028	.020	.009	.009
ASO		.187	.064	.073	.159	.145	.060	.194		.178	.036	.035	.035	.030	.041	.048
ASL		.172	.170	.168	.151	.149	.121	.062		.092	.079	.078	.076	.034	.028	.017
BIL		.083	.088	.092	.091	.091	.098	.226		.053	.052	.056	.056	.061	.043	.063
BIM		.345	.330	.330	.316	.276	.216	.222		.219	.190	.190	.160	.066	.031	.077
INT			.335	.334	.327	.234	.163	.183			.210	.210	.201	.113	.105	.092
OOA			.210	.185	.248	.195	.034	.168			.049	.048	.055	.033	.041	.034
DWO			.178	.105	.212	.123	.023	.109			.067	.038	.115	.042	.031	.076
BMB			.350	.338	.347	.267	.023	.016			.097	.090	.084	.024	.013	.004
INS			.238	.220	.232	.161	.163	.237			.233	.203	.205	.166	.134	.220
KLS				.083	.080	.087	.090	.070				.044	.043	.015	.026	.003
DLO					.431	.398	.395	.272					.323	.132	.071	.021
LL6					.072	.094	.062	.033					.073	.023	.027	.020
ADV						.744	.485	.456						.419	.105	.371
KGS						.366	.207	.145						.086	.033	.026
PRE						.674	.395	.170						.189	.069	.018
TON							.880	.709							.726	.066
AOS								.750								.254
LLS								.806								.291
EXT								.850								.208
MC	.557	.630	.680	.682	.739	.853	.917	.983	.557	.630	.680	.682	.739	.853	.917	.983

Table 13.4.3 Eight MORALS applications with EIN as the criterion variable

SEX

	Correlations									Regression weights								
BVA	.143	.134	.134	.134	.133	.133	.131	.114	.114	.141	.129	.130	.130	.132	.123	.118	.115	.126
OPV	.106	.108	.106	.106	.105	.104	.102	.102	.099	.108	.103	.104	.103	.102	.102	.103	.101	.102
OPM	.090	.085	.085	.085	.085	.081	.082	.080	.081	.089	.073	.070	.072	.065	.052	.055	.060	.060
AKG	.061	.055	.052	.052	.050	.050	.051	.054	.052	.061	.061	.062	.062	.067	.071	.066	.057	.058
URB	.042	.038	.031	.028	.037	.041	.041	.042	.042	.038	.042	.031	.031	.037	.042	.041	.054	.058
ASO		.256	.264	.264	.263	.263	.263	.263	.262		.258	.282	.283	.278	.242	.241	.226	.209
ASL		.056	.061	.060	.061	.060	.061	.062	.061		.085	.087	.089	.091	.084	.075	.087	.080
BIL		.218	.222	.222	.223	.223	.219	.200	.199		.174	.168	.169	.169	.161	.160	.145	.143
BIM		.211	.212	.212	.211	.212	.213	.210	.209		.160	.155	.157	.150	.148	.152	.129	.120
INT			.192	.193	.192	.190	.191	.183	.182			.193	.194	.191	.184	.182	.170	.168
OOA			.135	.134	.100	.123	.067	.080	.150			.083	.083	.063	.046	.056	.049	.054
DWO			.190	.190	.190	.186	.190	.189	.184			.275	.279	.272	.244	.284	.306	.282
BMB			.220	.220	.220	.215	.220	.216	.213			.242	.242	.226	.196	.252	.262	.231
INS			.161	.163	.151	.134	.150	.086	.072			.110	.112	.106	.088	.105	.065	.066
KLS				.008	.008	.008	.008	.008	.008				.021	.020	.031	.030	.043	.042
DLO					.088	.089	.089	.088	.088					.069	.079	.073	.084	.086
LL6					.075	.072	.074	.072	.070					.069	.046	.047	.039	.037
ADV						.144	.140	.114	.111						.131	.156	.109	.114
KGS						.135	.134	.134	.134						.138	.126	.118	.112
PRE						.035	.031	.010	.001						.052	.026	.008	.012
TON							.099	.056	.055							.137	.054	.042
AOS								.148	.151								.153	.165
LLS								.376	.375								.331	.326
EXT								.200	.202								.096	.086
EIN									.176									.189
m.c.	.213	.426	.483	.484	.492	.521	.532	.634	.650	.213	.426	.483	.484	.492	.521	.532	.634	.650

Table 13.4.4 Eight MORALS applications with SEX as the criterion variable

variable with substantive weight; for most other variables the regression weight goes down. Such a result also can be expected in linear regression analysis (contrary to Gresham's law we could say: a good predictor drives out bad predictors). Typical for MORALS is that also the correlations between other variables and EIN go down once TON is added to the predictor set (in linear analysis such correlations do not change, of course). But the table for regression weights shows also that TON's dominant position is immediately undermined when the even better predictors AOS LLS EXT are added to the predictor set: TON's regression weight goes down from .726 to .066 (TON's correlation does not go down that drastically). Again, such a result is not typical for MORALS. We might find the same thing in linear regression analysis, where regression weights also will become very unstable if we add new predictors correlated with old ones.

As a further example, the same type of successive "step up" MORALS analysis has been applied to SEX as the criterion variable (and EIN as a ninth addition to the predictor set). Results are shown in table 13.4.4. Now results are much more stable. The reason is that the new variables added in each successive analysis do not have typically larger correlations with SEX.

Finally, MORALS has been applied to girls and boys separately. Results are summarized in table 13.4.5 (with EIN as criterion variable, and all other variables in the predictor set). Clearly, for girls, variables BVA OPV INT OOA DLO KGS are more important than for boys. But again, interpretation of such results is ambiguous, mainly because again changes in correlations can depend upon many different things (such as restriction of range). Table 13.4.5, therefore, invites us to go back to the data with new hypotheses as to what the data might reveal. Why are BVA OPV INT OOA DLO KGS more "important" for girls? Table 13.4.5 does not give an answer to this question; it sends us back to scrutinizing the data. And, in fact, that is what usually happens in non-linear MVA: we do not get ready made answers to questions we never posed, but we are made aware of peculiarities of the data that need further investigation.

13.4.5 PRINCALS.

A one-dimensional PRINCALS solution on 25 variables (SEX excluded) has been calculated under both ordinal and linear restrictions. Table 13.4.6 gives results in terms of squared correlations between re-scaled variables and the first dimension of PRINCALS. The linear restriction implies that results are the same as in classical PCA. The corresponding

		Correlations		weights	
1	BVA	.14	.02	.08	.08
2	OPV	.34	.24	.10	.03
3	OPM	.20	.26	.04	.04
4	AKG	.16	.10	.05	.05
5	URB	.01	.00	.03	.01
6	INT	.26	.08	.10	.07
7	ASO	.17	.06	.03	.04
8	OOA	.14	.02	.03	.02
9	DWO	.13	.21	.05	.06
10	BMB	.09	.10	.00	.01
11	KLS	.10	.09	.00	.01
12	DLO	.46	.37	.08	.04
13	LL6	.08	.02	.02	.03
14	ADV	.65	.62	.19	.25
15	KGS	.30	.13	.05	.05
16	BIL	.14	.17	.07	.05
17	BIM	.21	.28	.03	.08
18	PRE	.48	.40	.09	.09
19	TON	.89	.89	.38	.28
20	AOS	.57	.77	.02	.15
21	LLS	.68	.74	.25	.20
22	EXT	.77	.82	.10	.14
23	ASL	.09	.12	.05	.04
24	INS	.17	.25	.11	.16
M.C.		.96	.96		
		F	M	F	M

Table 13.4.5 MORALS on boys and girls seperately, with EIN as the criterion variable

Variable code	PRINCALS numerical	PRINCALS ordinal	HOMALS multiple
BVA	.56	.63	.63
OPV	.63	.65	.65
OPM	.49	.51	.51
AKG	.23	.24	.24
URB	.17	.18	.18
ASO	.19	.22	.22
ASL	.14	.25	.27
BIL	.20	.25	.24
BIM	.40	.44	.43
INT	.39	.35	.35
OOA	.22	.22	.22
DWO	.09	.09	.09
BMB	.44	.42	.41
INS	.07	.05	.05
KLS	.14	.15	.15
DLO	.42	.39	.38
LL6	.05	.10	.12
ADV	.77	.81	.81
KGS	.65	.65	.65
PRE	.80	.81	.80
TON	.86	.89	.89
AOS	.42	.45	.46
LLS	.32	.35	.37
EXT	*.30*	*.42*	*.59*
EIN	.84	.86	.86

Table 13.4.6 PRINCALS squared component loadings (single fit) and HOMALS discrimination mea

correlations then are classical "factor loadings". Ordinal PRINCALS allows for an ordinal re-scaling of categories, and under this restriction the first eigenvalue of the correlation matrix is maximized. The HOMALS solution in table 13.4.6 allows for any optimal scaling of categories, and maximizes the first eigenvalue of the optimally scaled data matrix. Obviously, this first eigenvalue will increase to the extent less restrictions are imposed, and we find

PRINCALS linear (same as PCA without re-scaling): 5.36
PRINCALS ordinal 5.85
HOMALS 6.05

Figures 13.4.4 and 13.4.5 give plots of the category transformations for linear transformation (13.4.4) and ordinal transformation (13.4.5). The nominal HOMALS transformation was already shown in figure 13.4.1. We shall not give detailed comments on the comparison between figures

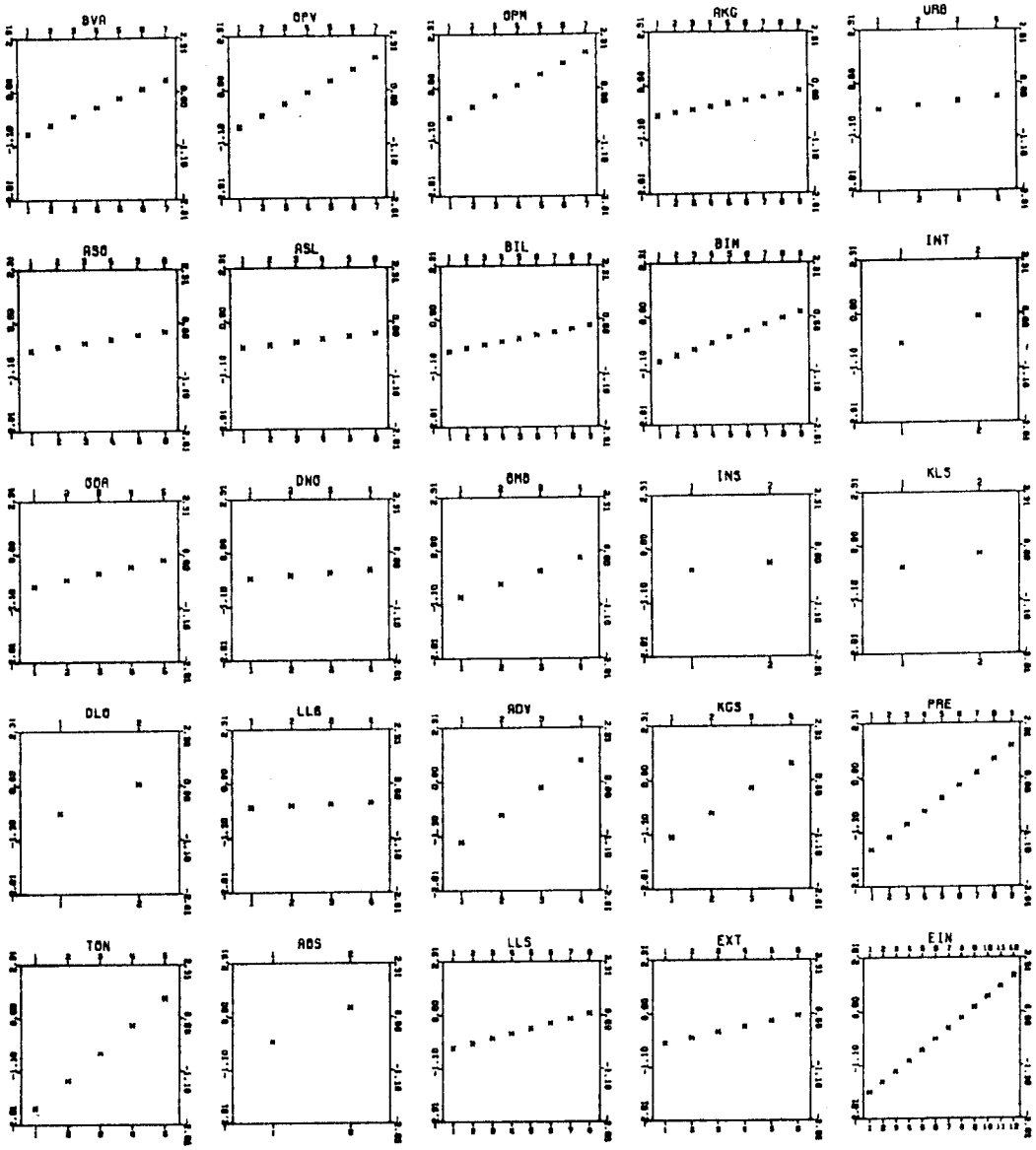


Figure 13.4.4 One-dimensional PRINCALS numerical solution

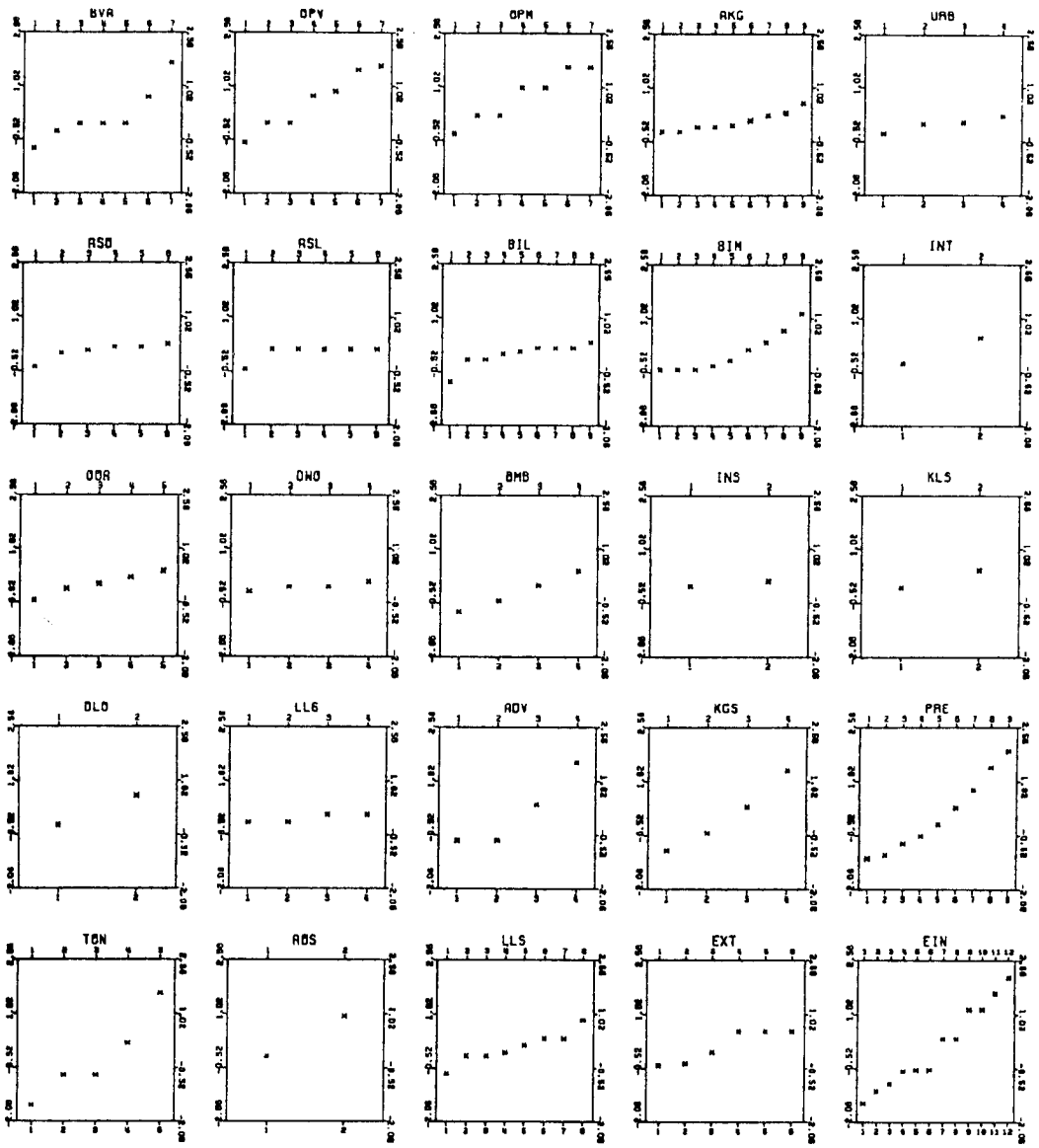


Figure 13.4.5 One-dimensional PRINCALS ordinal solution

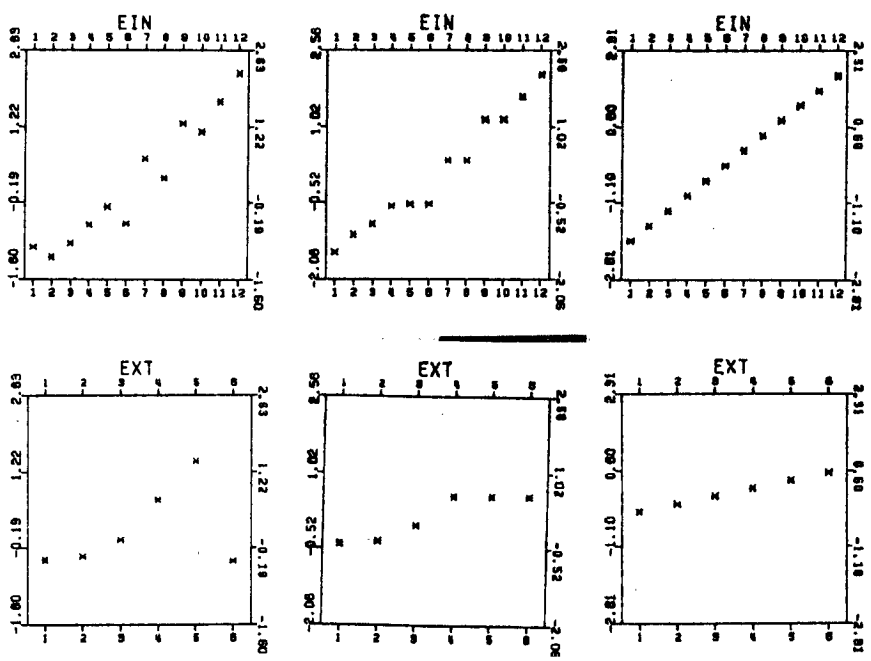


Figure 13.4.6 Three different transformations illustrated for variables EIN and EXT.
A. nominal B. ordinal C. numerical

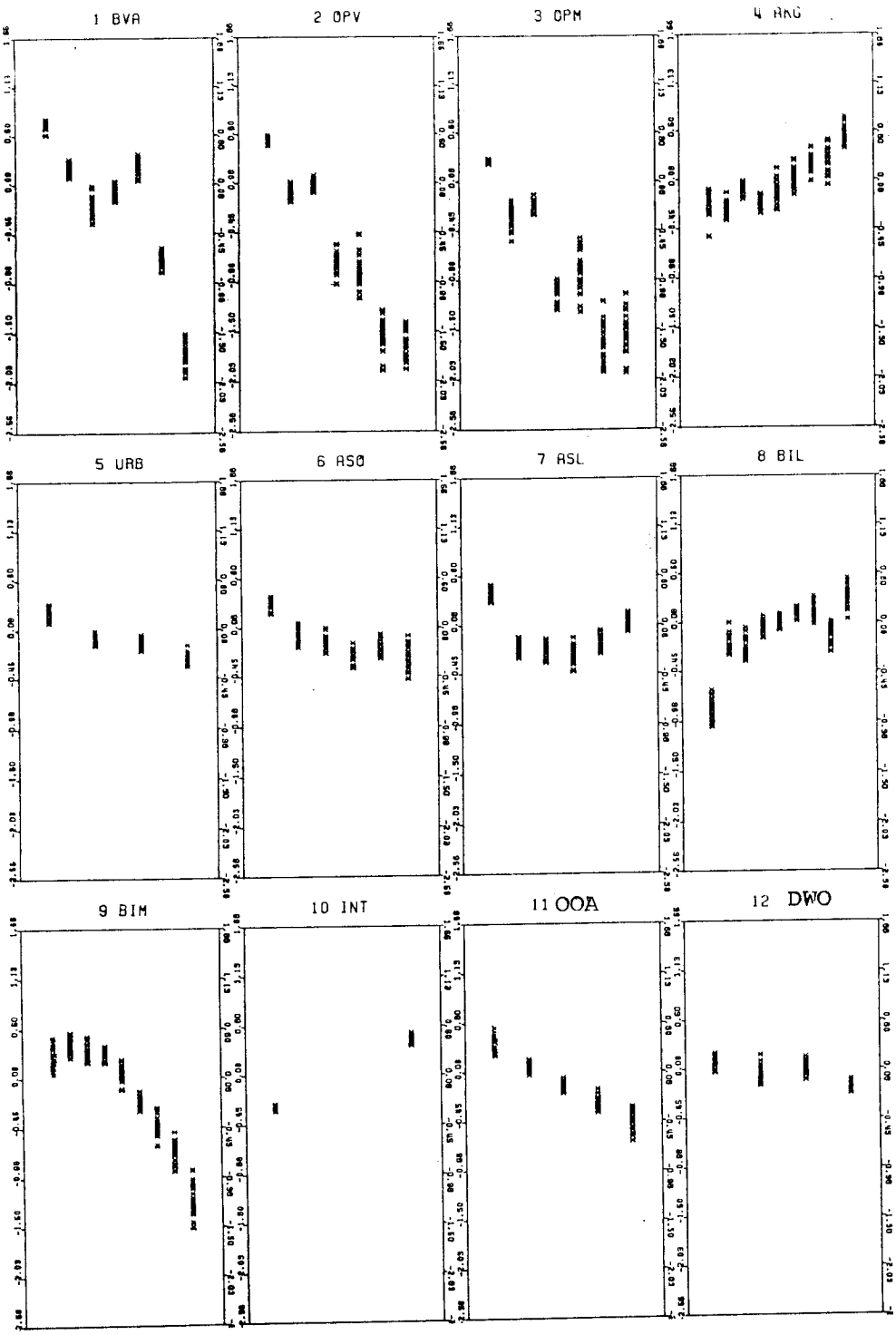
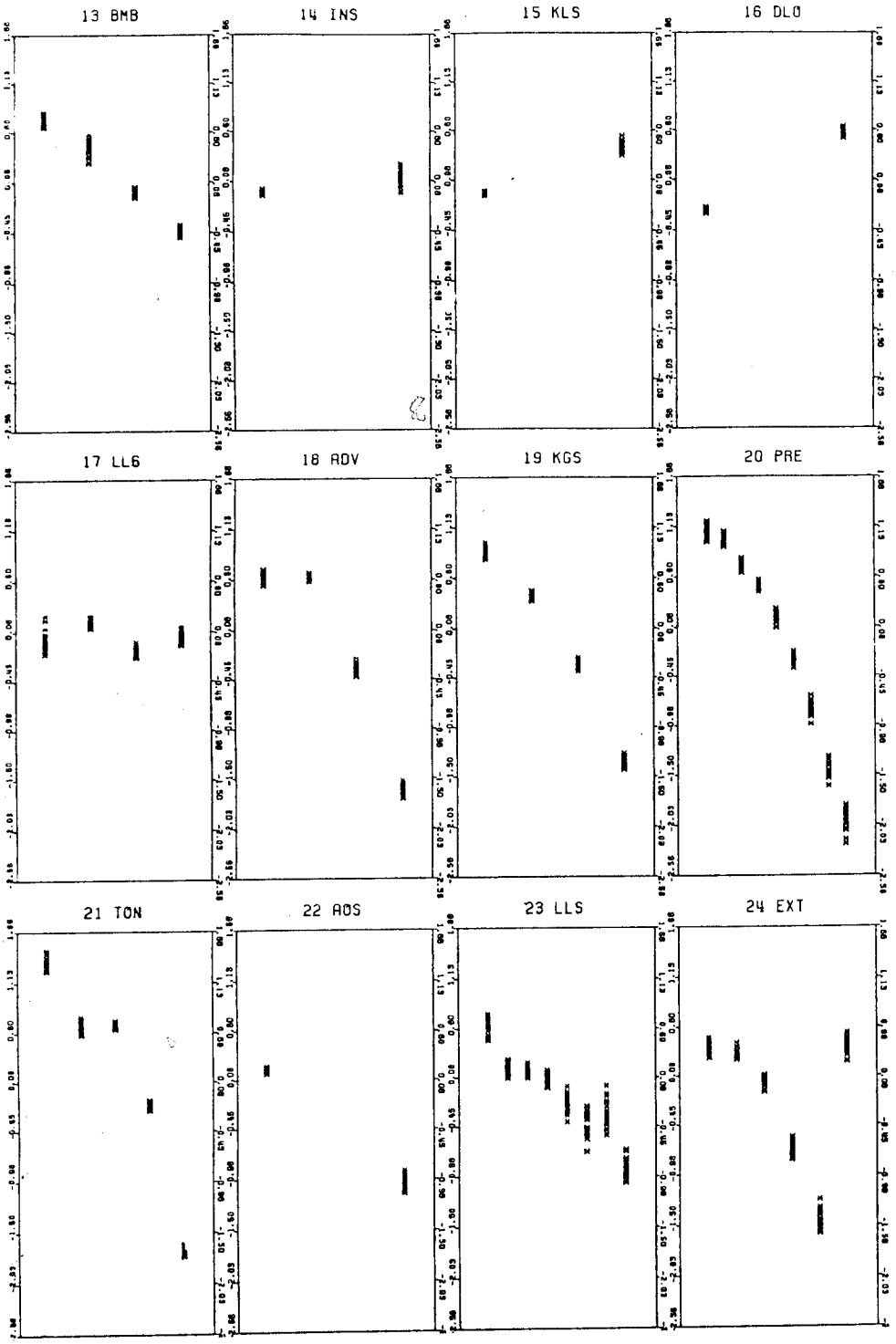


Figure 13.4.7 HOMALS bootstrap for categories


13 BMB
14 INS
15 KLS
16 DLO
17 LL6
18 ADV
19 KGS
20 PRE
21 TON
22 AOS
23 LLS
24 EXT

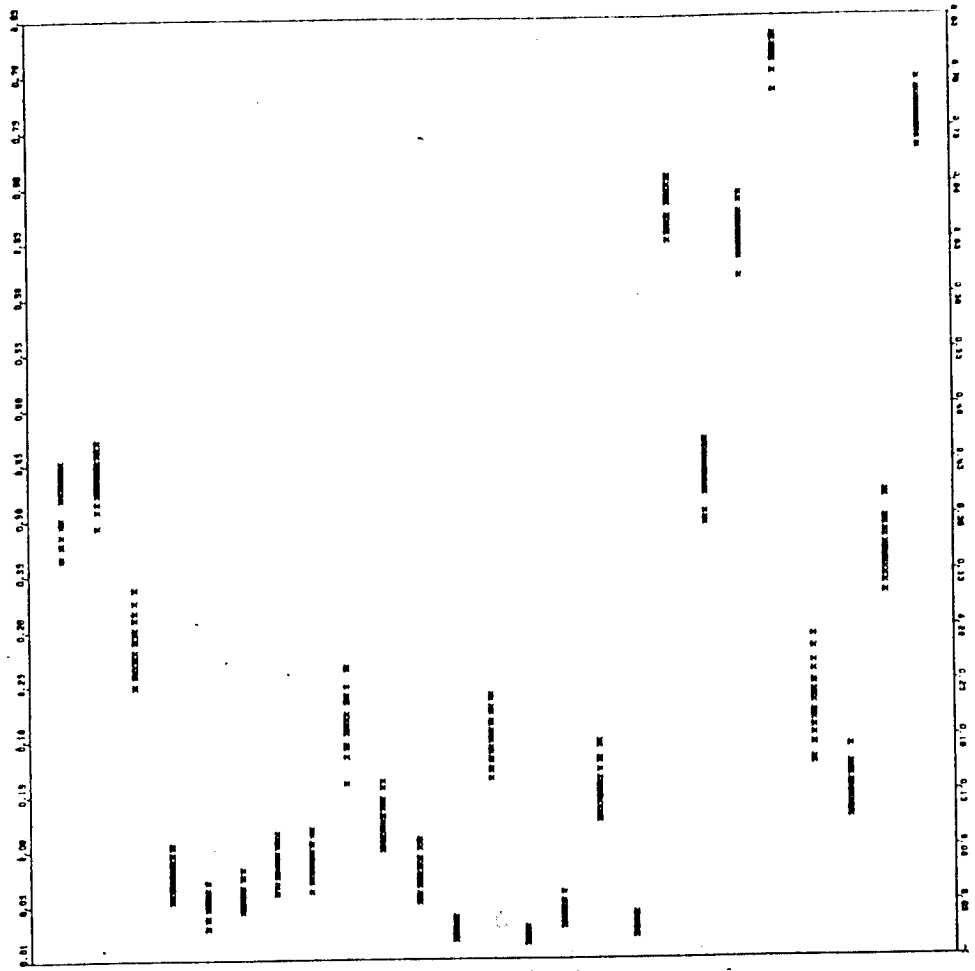


Figure 13.4.8 HOMALS bootstrap for discrimination measures

13.4.1, 13.4.4, and 13.4.5. It is clear that the plots in 13.4.4 always are linear, that the plots in 13.4.5 always are monotonous, and that the plots in 13.4.1 have more freedom; to what extent this affects the quantification of each separate variable can be checked by the reader. Figure 13.4.6 summarizes results for two selected variables, EIN and EXT. Obviously, the HOMALS solution recognizes the special position of category (6) in EXT; the other two solutions compromize on that. As to EIN, HOMALS shows the typical reversals, PRINCALS ordinal flattens them, PRINCALS numerical ignores them. Of course, for EXT the HOMALS solution is "better"; on the other hand, the categories of EXT were a bit muddled up to start with, so that PRINCALS results for EXT are unduly at a disadvantage. As to EIN, we remarked in section 13.4.2 that the HOMALS quantification may be affected by the unequal marginal frequencies; PRINCALS ordinal corrects for that.

13.4.6 HOMALS bootstrap.

Figure 13.4.7 gives results for a bootstrap on the first HOMALS dimension, based on 25 samples with replacement out of the 1845 individuals in the complete survey. Figure 13.4.7 should be compared with figure 13.4.1. Results show, on the whole, that the bootstrap has smaller spread for variables with large discrimination measure (e.g., TON). But also, there is a relation with marginal frequency: OPV-1 with marginal frequency of 761 has small bootstrap spread, whereas OPV-6 with marginal frequency of 56 has large bootstrap spread.

Bootstrap results confirm the reversal of categories (3) and (5) in BVA. They also show that in OPV the categories (2) and (3), or (4) and (5), or (6) and (7) might as well be grouped together.

Figure 13.4.8 shows bootstrap results for the discrimination measures. The plot suggests that when a discrimination measure is large, it also is relatively stable. Low discrimination measures also tend to be stable: they remain consistently low. Larger bootstrap spread is found for variables with intermediate discrimination measure, such as OPV, or AOS.

13.4.7 Description of "From Year to Year" variables.

Details about the survey 'From Year to Year' can be found in the original survey proposal (ITS, 1968), and in reports by Kropman and Collaris (1974), Collaris and Kropman (1978), Dronkers (1978).

BVA: occupational level of father. (1)=agricultural or unskilled labor, (2)=skilled labour, (3)=clerical, (4)=small business, (5)=farmers, (6)=managerial, (7)=higher managerial and free professions

OPV: education level of father
OPM: education level of mother
For both variables categories ranging from (1)=primary school only, to (7)=university training

AKG: number of children in family. (1)=1, (2)=2, etc., until (9)=more than 8.

URB: degree of urbanization of residence. (1)=rural, (2)=villages, (3)=small cities, (4)=large cities

ASO: aspiration level of parents (Reisman,1953). Parents were asked whether they find their child later should accept a "very good job" even if this would imply a specific disadvantage such as "no time for hobbies". Eleven such disadvantages were mentioned, for each of which parents could respond with (1)=unacceptable to (4)=acceptable. The score is the sum of the category numbers chosen, later condensed into 6 classes.

ASL: aspiration level of child. Same as ASO, answered by child.

BIL,BIM: two scales for 'interest' in various professional activities, at lower level (BIL) and at higher level (BIM)

INT: parents' interest in child's school performance, as rated by teacher in last form of primary school. (1)=much interest, (2)=not much interest

DWO: a four point scale, indicating agreement with the statement that parents decide about educational and professional career of their child irrespective of what child wants. (1)=agree, to (4)=disagree

OOA: a five point scale indicating agreement with the statement that a child's career should be based on advice from teachers and results of mental tests. (1)=agree, to (5)=disagree

BMB: a four point scale indicating agreement with the statement that professional training is unimportant for girls. (1)=agree, to (4)= disagree

INS: do parents agree with present educational choice? (1)=yes,(2)=no

KLS: has child been at nursery? (1)=yes,(2)=no

DLO: did child ever repeat a form in primary school? (1)=no,(2)=yes

LL6: number of children in last form of primary school. (1)=less than 10, (2)=10 to 19, (3)=20 to 30, (4)=more than 30

ADV: teacher's advice as to education after primary school, on four levels increasing from (1) to (4)

KGS: average score on achievement test in last form of primary school, from (1)=low, to (4)=high

PRE: result on test supposed to predict achievement on secondary school. Stanine scoring.

TON: level of secondary school first selected immediately after primary school. Categories (1) to (5) indicate increasing level.

AOS: whether first selected secondary school had a differentiated curriculum. (1)=no, (2)=yes

LLS: number of pupils at first secondary school. (1)=less than 100, (2)=100 to 200, etc., until (8)=more than 700

EXT: number of extra-curricular activities at secondary school: library, school-paper, excursions, clubs, school council. Categories (1) to (5) are a count of number of activities mentioned. (6)=child cannot mention any such activity

EIN: final level of secondary education
(1) - LO only
(2) - VGLO without certificate
(3) - VGLO with certificate, or LBO 1 year, or "brugklas"
(4) - LBO or ULO/MAVO/VHMO unfinished after 2 years
(5) - LBO or ULO/MAVO/HAVO/MMS unfinished after 3 years
(6) - LBO finished
(7) - ULO/MAVO/HAVO/MMS unfinished after 4 years, or VHMO unfinished after 3 years
(8) - ULO/MAVO finished
(9) - HAVO/MMS unfinished after 5 years, or VHMO unfinished after more than 4 years
(10) - HAVO/MMS finished
(11) - HBS finished
(12) - ATHENEUM/GYMNASIUM finished

SEX: sex. (1)=male, (2)=female

13.5 Parliament survey

13.5.1 Introduction.

In 1968 and in 1972 members of the Second Chamber of Dutch Parliament were asked to co-operate in an extensive questionnaire study, some data of which will be used for the present example. The data refer to two items of the questionnaire.

A. Preference order. The respondent was asked to give a preference rank order for the parties represented in parliament (the actual question was to order parties as to "the degree respondent feels congenial to them").

B. Issue statements. Respondent is asked to indicate his or her position with respect to seven political issues. The issues are listed in table 13.5.1. (This item was included only in the 1972 questionnaire.)

	1 DEVELOPMENT AID	
the government should spend more money on aid to developing countries	(1).....(9)	the government should spend less money on aid to developing countries
	2 ABORTION	
the government should prohibit abortion completely	(1).....(9)	a woman has the right to decide for herself about abortion
	3 LAW & ORDER	
the government takes too strong action against public disturbances	(1).....(9)	the government should take stronger action against public disturbances
	4 INCOME DIFFERENCES	
income differences should remain as they are	(1).....(9)	income differences should become much less
	5 PARTICIPATION	
only management should decide important matters in industry	(1).....(9)	workers too must have participation in decisions important for industry
	6 TAXATION	
taxes should be increased for general welfare	(1).....(9)	taxes should be decreased so that people can decide for themselves how to spend their money
	7 DEFENSE	
the government should insist on shrinking the Western armies	(1).....(9)	the government should insist on maintaining strong Western armies

Table 13.5.1 The seven issues and the meaning of lowest and highest category.

13.5.2 Preference rank orders, 1968.

Preference rankings for political parties were available from 141 Members of Parliament, 1968. Their party allegiance is given in table 13.5.2. The data have also been analyzed by Daalder and Rusk (1972), De Leeuw (1973), Daalder and Van de Geer (1977).

The data matrix in this case would be a 141 x 14 matrix. It has 14 columns since apart from the 12 parties mentioned in table 13.5.2 there were in 1968 two more parties in Parliament, each with only one representative. A row of the data matrix consists of the rank numbers(the actual numbers were from 2 to 15) in some permutation. If we take each party as a variable, and each rank number as a category, the indicator matrix for a HOMALS analysis would become a 141 x 196 matrix. Such an analysis would completely ignore the ordinal nature of the rankings.

As an alternative one might think of analyzing the reversed indicator matrix, treating respondents as variables, and parties as objects. This would result in a 14 x (141 x 14) matrix, in which for each respondent the indicator matrix G_j would be some permutation of the identity matrix. As a consequence the HOMALS solution will degenerate: all eigenvalues ϕ_i will become equal to one.

For the present illustration it was decided to analyze the data with PRINCALS, both with numerical and with ordinal option. Also, PRINCALS was applied to the transposed data matrix, with 12 rows (the two small parties mentioned above were dropped, only the parties listed in table 13.5.2 were retained), and 141 columns.

The numerical option comes to the same as a linear analysis (as if the rank numbers in a column of the transposed data matrix are at interval scale level). The solution is plotted in figure 13.5.1. The figure shows 12 points for the parties, and 141 for respondents. It is a typical 'joint plot' (section A10.1), in the sense that if we draw a vector though the point of a respondent, and project party points on this vector, such projections will be approximately proportional to the entries of a column of the

PARTY	NUMBER OF RESP.	DESCRIPTION OF PARTY	LABEL IN FIGURES
CPN	0	communist	-
PSP	4	pacifist-socialist	P
PvdA	37	labor	L
D'66	7	pragmatic-liberal	6
PPR	0	radical	-
KVP	42	catholic	K
ARP	15	protestant	A
CHU	12	protestant	U
VVD	17	conservative-liberal	V
BP	4	farmers	B
SGP	2	reformed	S
GPV	1	reformed	G

Table 13.5.2 Party allegiance of respondents 1968

transposed data matrix (after subtraction of the column mean). Conversely, if we draw a vector through a party point, and project all 141 respondent points on it, such projections will be approximately proportional to a row of the transposed data matrix.

Note that figure 13.5.1 shows the familiar 'horse shoe' for respondents, whereas for parties the points almost are located on a closed circle. If we would give parties an apriori ordering from left to right (CPN, PSP, etc., until GPV, SGP, BOP) , the extremes on this scale are in the plot close together.

Figure 13.5.2 gives the plot for the PRINCALS solution with ordinal option. Since the ordinal option is less restrictive than the numerical, eigenvalues for the ordinal option cannot be smaller than those for the numerical solution. For the numerical solution, the eigenvalues are ϕ_1=.56, ϕ_2=.25; corresponding eigenvalues for the ordinal solution are ϕ_1=.64, and ϕ_2=.28. The improvement is not impressive. In fact, figure 13.5.2 is very similar to figure 13.5.1 (the rotation over 180^0 is trivial).

Figure 13.5.3 gives results of a bootstrap for the ordinal solution, based on 10 samples. In the plot, samples are labelled from A to K (symbol I not used). The plot is a mere juxtaposition of results, without any attempt to make the 10 configurations more similar by means of rotations or scale corrections. Results show some overlap between KVP and CHU, between SGP and GPV, and to some extent also between PvdA, PPR, and D'66. The two least preferred parties, CPN and BP, tend to come close together.

13.5.3 Political preference and issue positions, 1972.

For the data from the 1972 questionnaire, CANALS has been applied to the preference orderings as one set of variables, and the responses to issue statements as the second. Preference rankings were reduced to rankings for the four largest parties only (in terms of representation in the Second Chamber: PvdA (39 seats), KVP (35), VVD (16), and ARP (13)). In the first instance, data were available from 141 respondents. (3 respondents had too many missing data and were excluded from the analysis).

In this application of CANALS, the preference data, although they are collected as row-conditional, are treated as column-conditional. I.e., each of the four parties becomes a 'variable', with as many categories as the number of different rankings given to the party.

A two-dimensional CANALS with ordinal option produced a canonical correlation of 1.00 for the first dimension. Such a result (not infrequent in CANALS) usually means that some respondent has an atypical response pattern in both

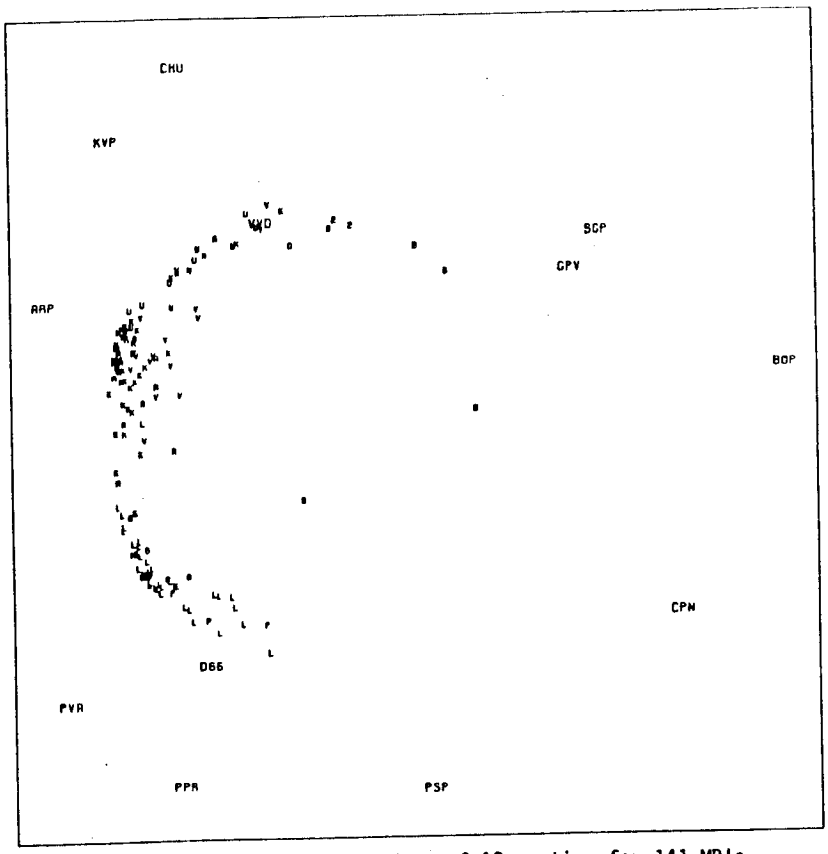


Figure 13.5.1 Preference rank orders of 12 parties for 141 MP's. Initial metric configuration.

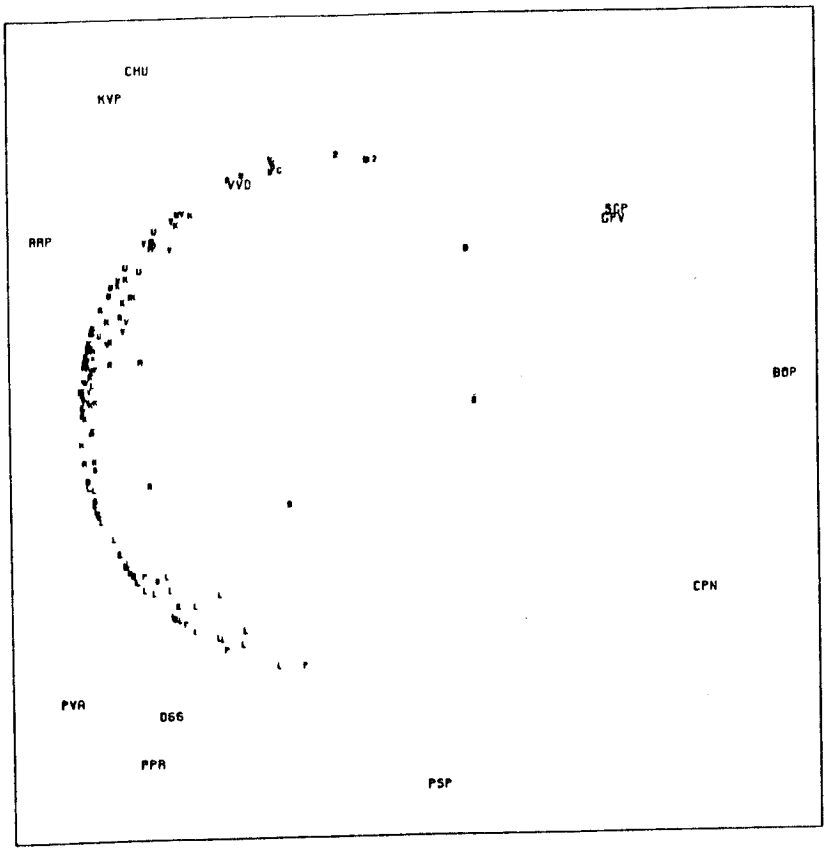


Figure 13.5.2 Preference rank orders of 12 parties. Final ordinal solution

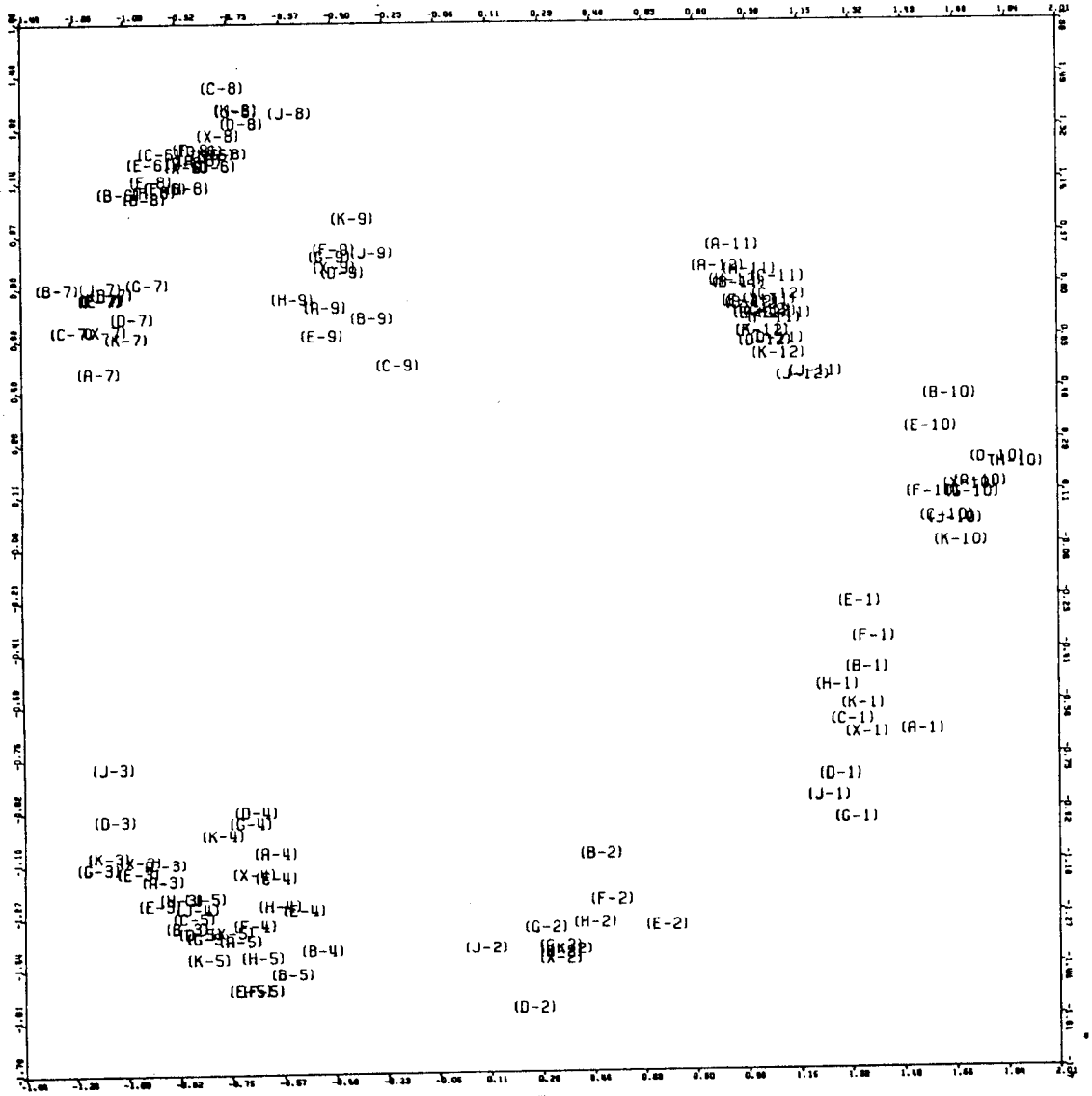


Figure 13.5.3 PRINCALS bootstrap on party preferences.

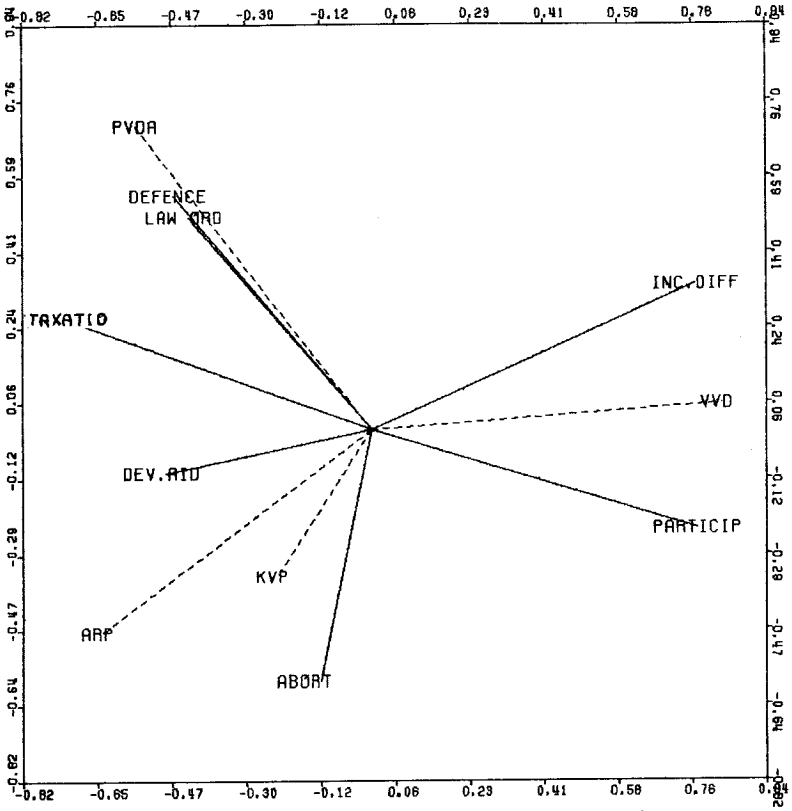
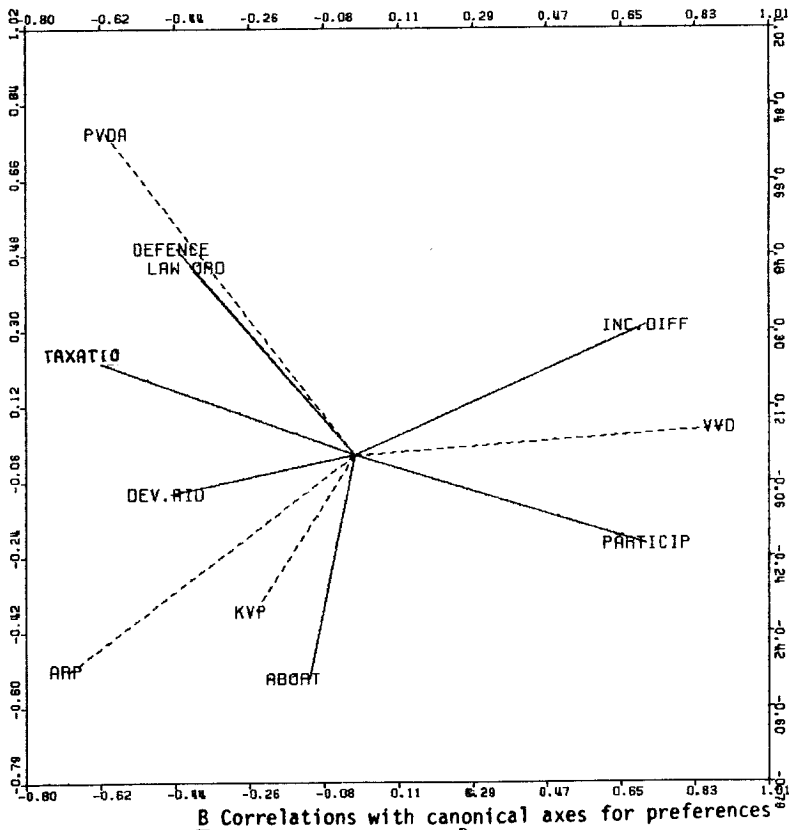


Figure 13.5.4 A Correlations with canonical axes for issues.



B Correlations with canonical axes for preferences

sets, which will create a separate dimension with perfect canonical correlation. This was also found to be the case in the present example: one respondent had the unique response pattern of combining very low sympathy for KVP with a missing score on LAW&ORDER. It was therefore decided to repeat CANALS with this respondent left out. The two-dimensional solution with ordinal option, now produced canonical correlations of .92 and .87, respectively. Figure 13.5.4 gives one possible plot of the results, with the variables as vectors, projected on the plane of the canonical variates for issues in figure A, and on the plane of the canonical variates for preferences in figure B. Since the canonical correlations are high, the two figures become very similar. As to substantive interpretation of the figures, it is useful to remember that vectors may be mirrored. E.g., TAXATION and PARTICIPATION are almost opposite vectors. Both vectors point in the direction of a high score, which for TAXATION means that taxes should decreased, and for PARTICIPATION that workers must have more participation. If we reverse the direction of the vector TAXATION,it will point in the direction "taxation must be increased", and this then goes along with "higher worker's participation".

Figure 13.5.5 gives two joint plots, for respondents and variables together, in the canonical plane for issues (A), and in the canonical plane for preferences (B). Variables are indicated as directions in the plane only, with both the positive and the negative pole. Again, if we project the points for respondents on such a direction, these projections will be approximately proportional to the quantification of the responses to that variable. The figure shows, e.g., that TAXATION contrasts respondents from VVD and DS70 to respondents from PvdA and many from KVP. ABORTION separates respondents from denominational parties from the others. One should, however, remain careful with the interpretation of the figures, and realize that they always remain an approximate summary of the data. E.g., the figures strongly suggest that TAXATION (decrease) and PARTICIPATION (less worker's participation) are very closely related, and indicate a characteristic position of VVD and DS70. From an inspection of the raw data, however, it could be seen that it is true that VVD and DS70 have rather similar outspoken views on TAXATION, whereas for PARTICIPATION VVD is on the middle of the scale (the average is close to initial category (5)), whereas DS70 has a much higher average (to the direction of increased worker's participation).

For completeness it must be said that in figure 13.5.5A two points for respondents have been omitted, and in figure B four. These are points for VVD respondents; their scores are so extreme (in SW direction) that they

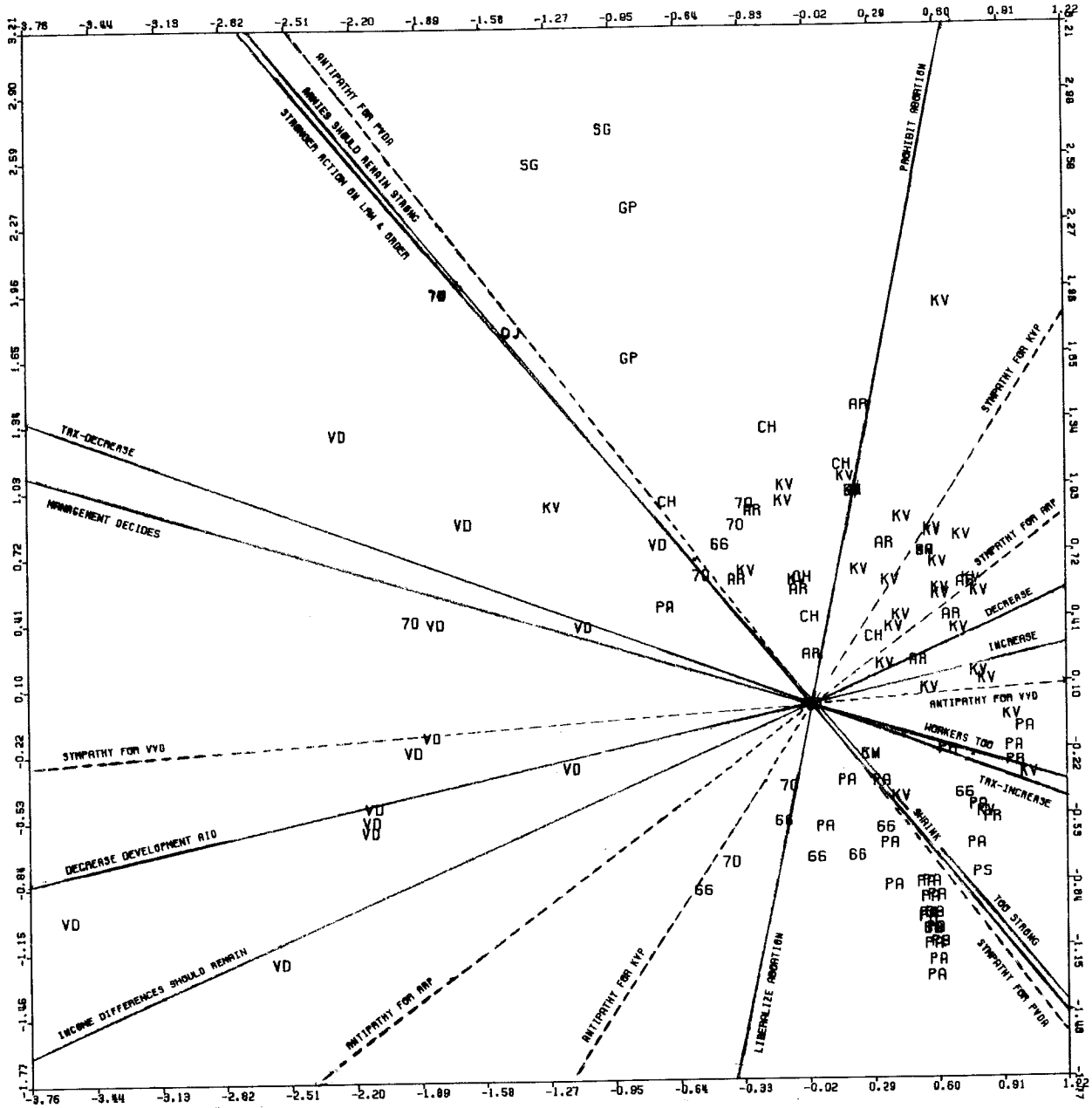
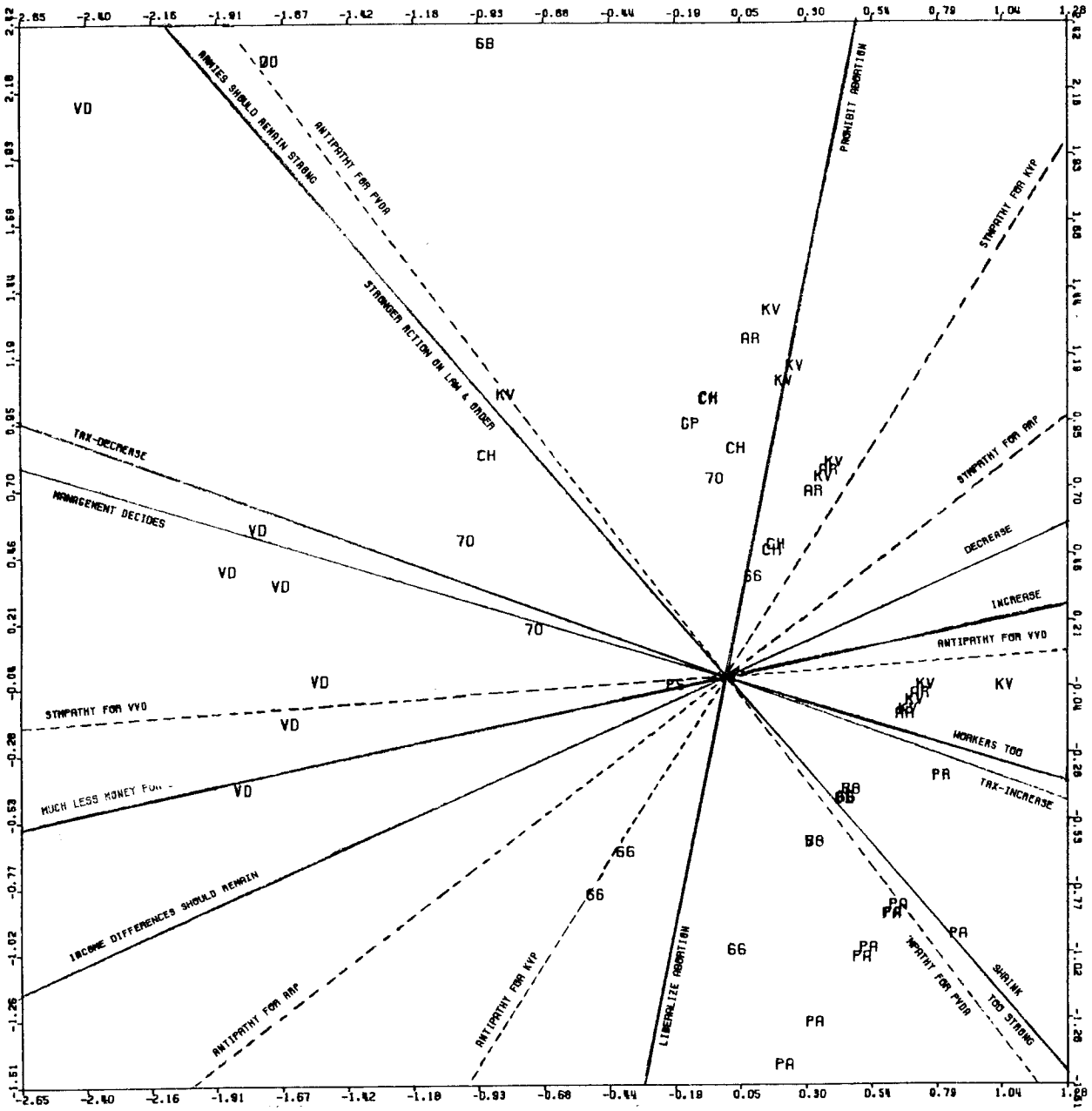


Figure 13.5.5 A Joint plot of respondents and issue-variables in the issue space

PA=PvdA	VD=VVD	GP=GPV
KV=KVP	66=D'66	SG=SGP
AR=ARP	70=DS70	DJ=one man party
CH=CHU	PS=PSP	



B Joint plot of respondents and issue-variables in the preference space

should be plotted far outside the figures as shown (if the picture would be re-scaled in order to accommodate for these points, the other points would have become so crowded together in the center of the plot that labels would be blurred altogether).

Figure 13.5.6 gives a plot of the ordinal transformation of the category numbers. It shows that the first five categories of DEVELOPMENT AID might as well be grouped together, as well as categories 4 to 9 of INCOME DIFFERENCES, 2 to 5 of PARTICIPATION. As to the ordinal transformation of preferences, the plot for ARP shows that this becomes an almost binary variable. The plots for KVP and VVD reveal gradations in sympathy more than in antipathy.

The figure also shows the average quantification for bootstrap samples.These dotted curves do not diverge much from the drawn curve, except for the lower category number of PARTICIPATION. These categories are infrequently chosen, always by VVD respondents, and it follows that the bootstrap samples will be sensitive to whether or not such response categories are in the sample.

Figure 13.5.7 gives results for the same bootstrap analysis with respect to category quantifications. It is striking, of course, that spread is most often large at one extreme of the variables, such as category (8) of DEVELOPMENTAL AID, or category (1) of ABORTION, or category 10 of ARP. In all these cases marginal frequencies of the categories are low, so that the bootstrap result depends on whether or not such a category is sampled. Figure 13.5.8 illustrates this further: the figure plots bootstrap variance as a function of marginal frequency, and it shows that all cases of extreme bootstrap variance are related to categories with small marginal frequency.

13.5.4 Relation between position on issues and party allegiance, 1972.

To illustrate CRIMINALS, an analysis was performed on 119 respondents representing the seven largest parties in Parliament, who had no missing entries. These parties are PvdA, KVP, VVD, ARP, D'66, CHU and DS70. CRIMINALS then comes to the same as CANALS with, on the one hand, a set of seven binary variables that code whether or not a respondent is a member of a party, and on the other hand the seven variables for the issues, each with 9 categories. CRIMINALS was solved for the first two dimensions of the solution, using four different options:

(1) Multiple nominal. All categories are treated as nominal. In addition, the quantification of categories for the second dimension is different from that used in the first dimension.

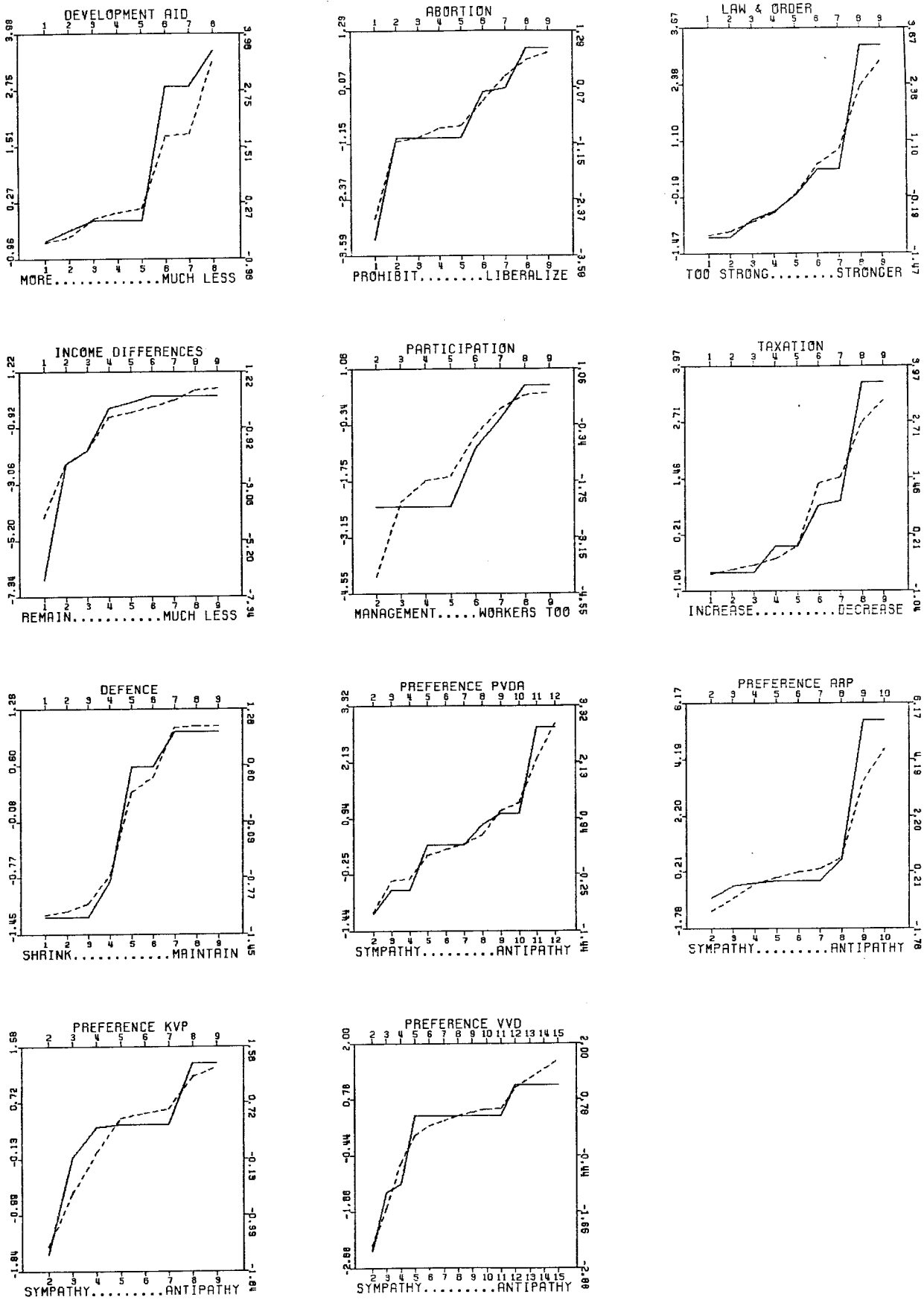


Figure 13.5.6 Optimal scaling based on ordinal CANALS (drawn lines) and averages of bootstrap results (dotted lines)

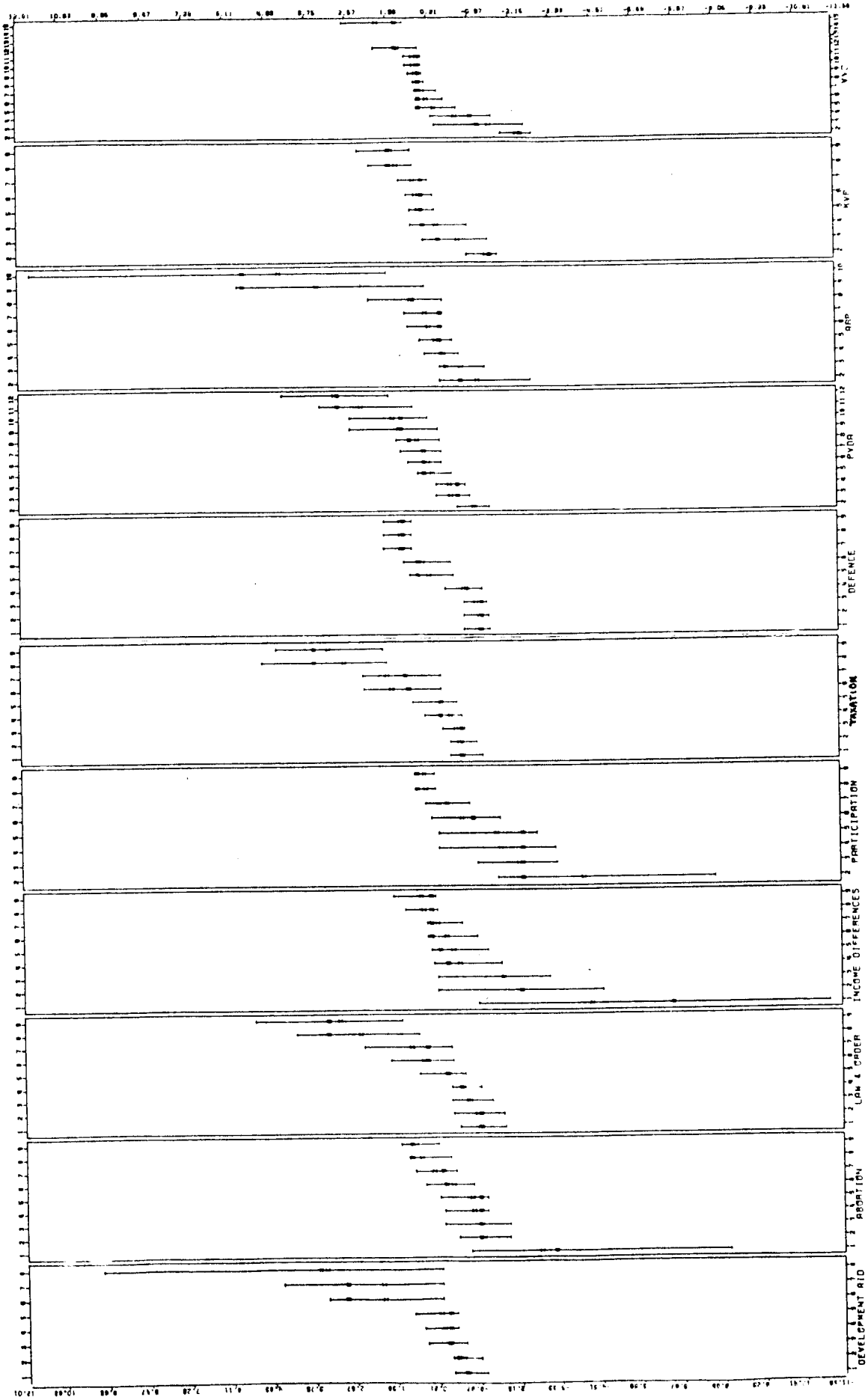


Figure 13.5.7 Bootstrap results for ordinal CANALS.

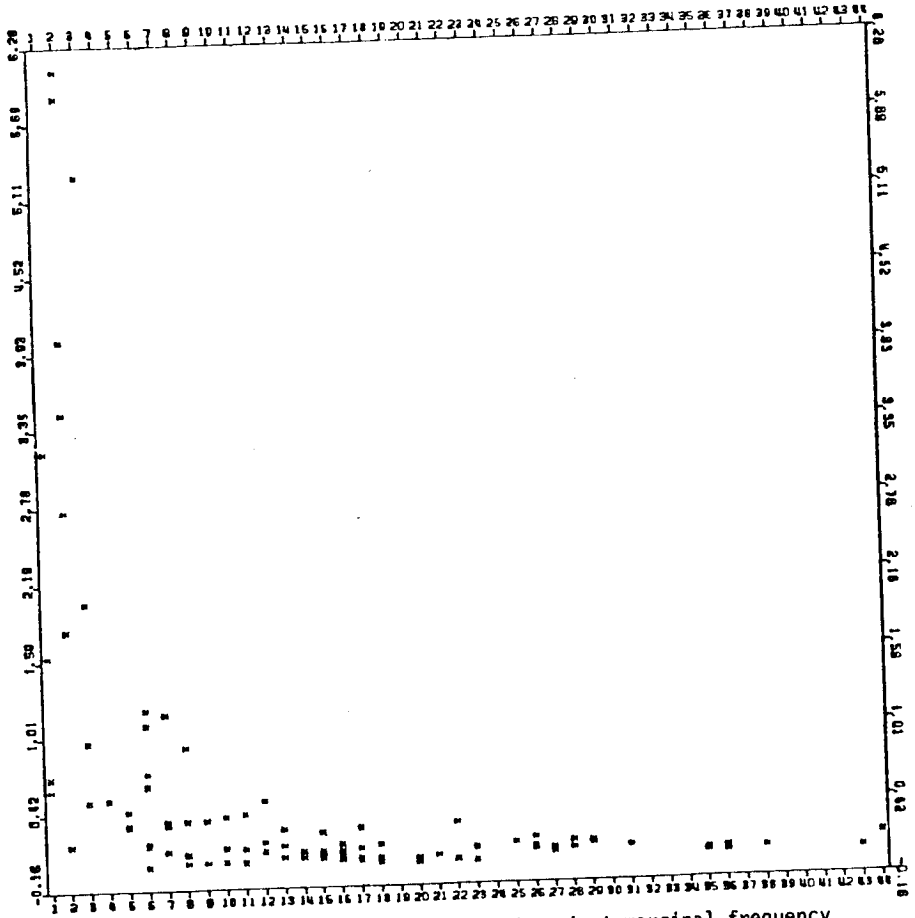


Figure 13.5.8 Bootstrap variance plotted against marginal frequency

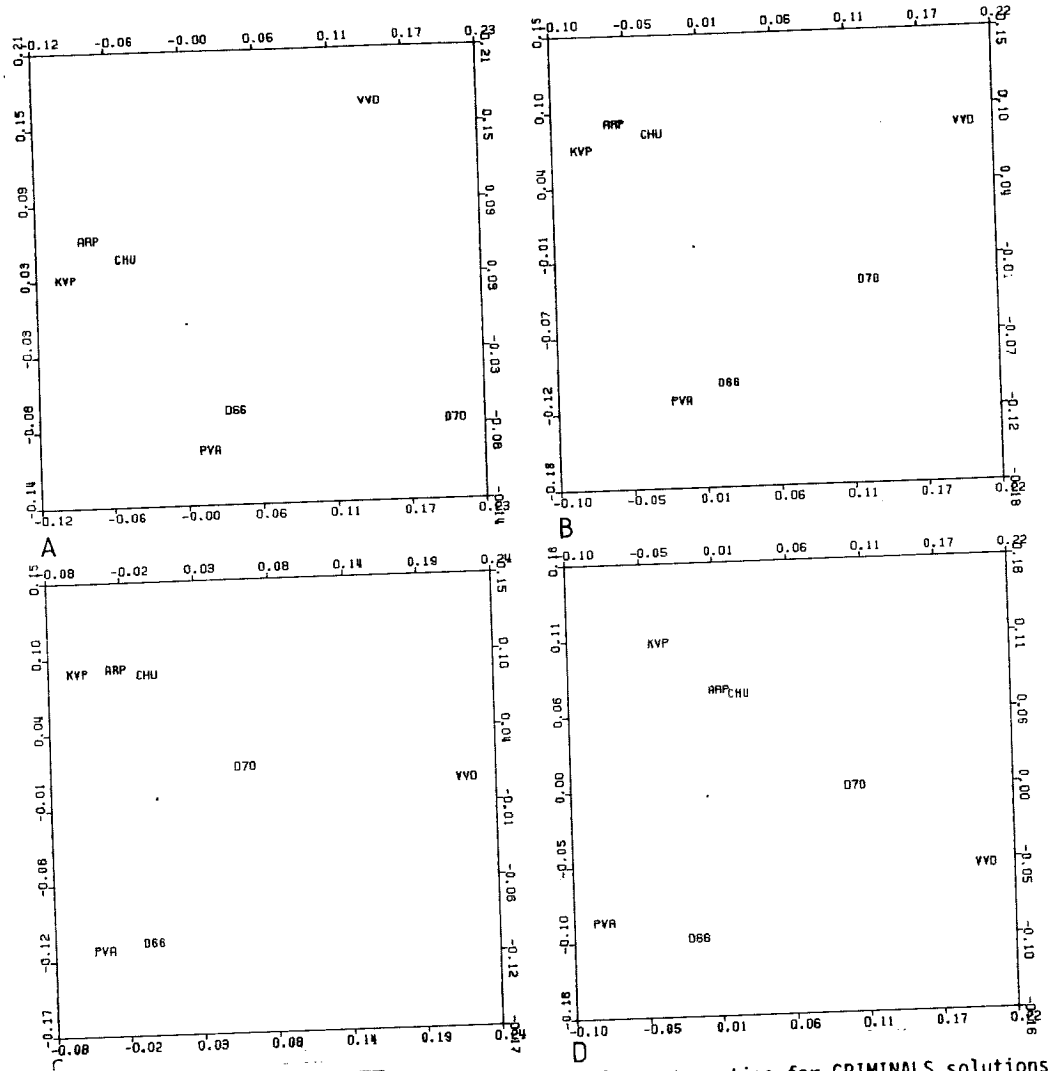


Figure 13.5.9 Canonical group means for seven largest parties for CRIMINALS solutions

A multiple nominal
B single nominal
C single ordinal
D numerical

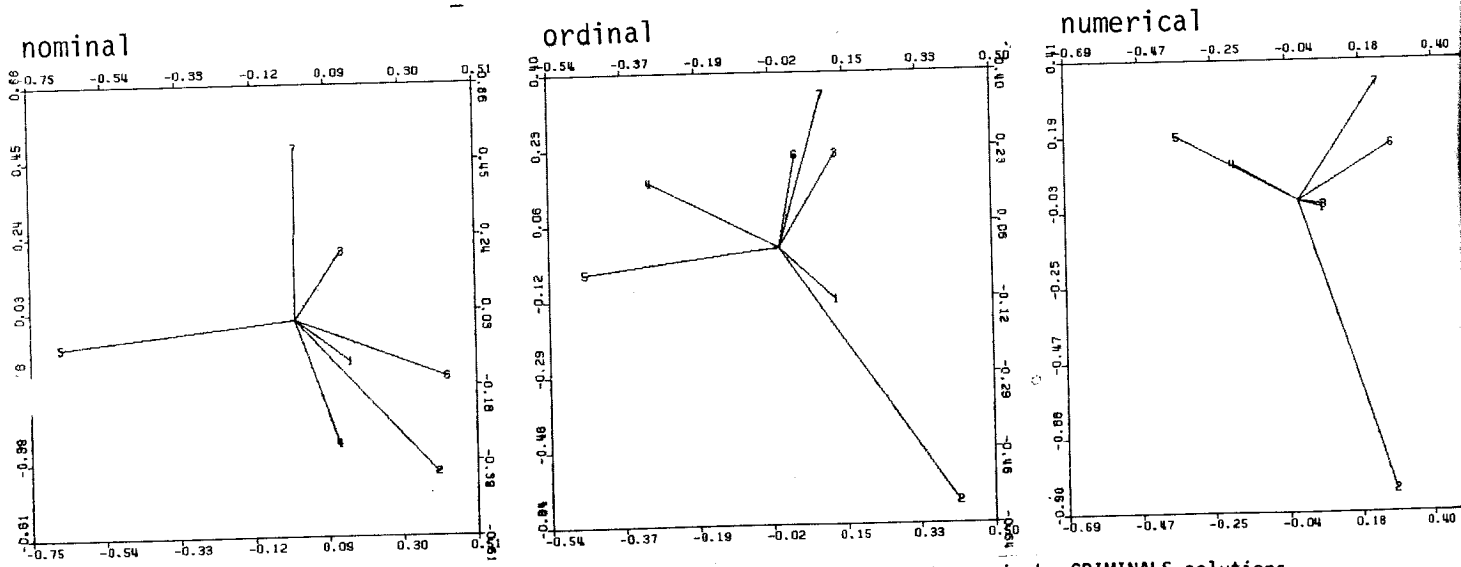
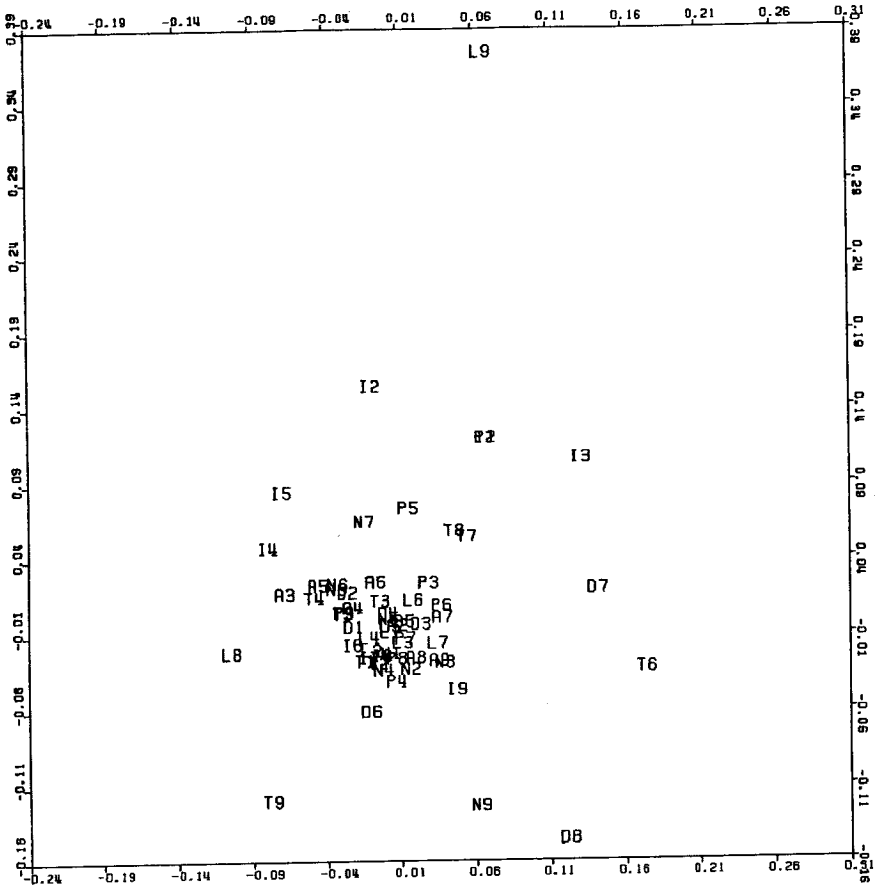
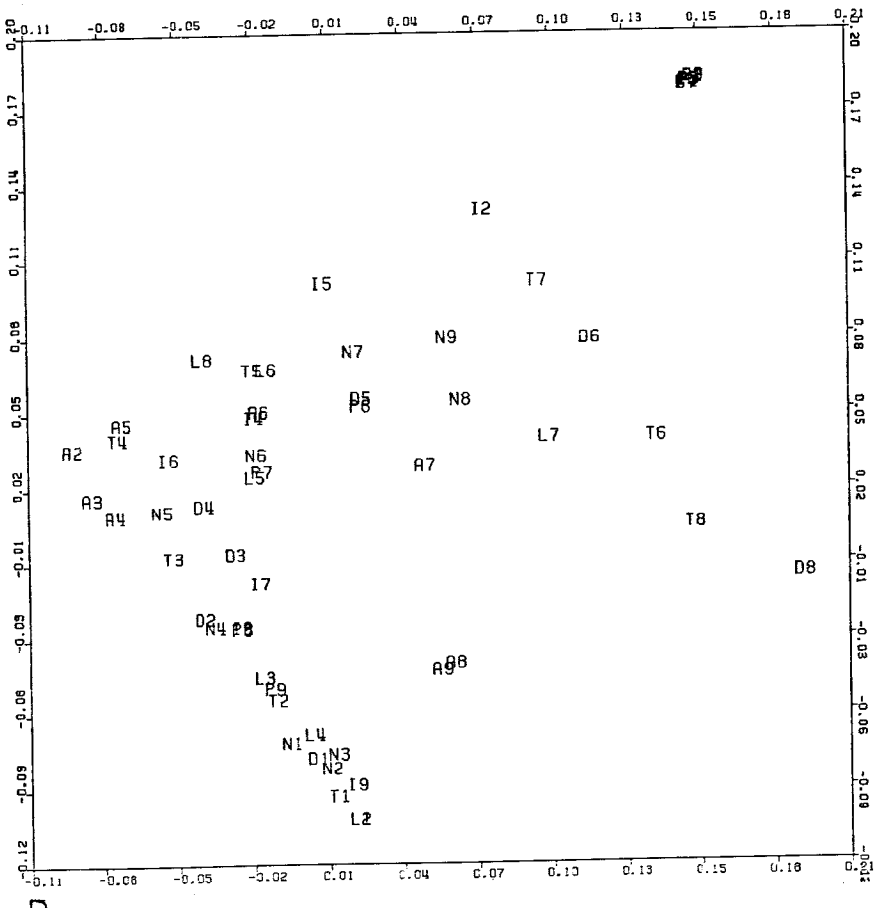


Figure 13.5.10 Correlations between issue variables and canonical axes for three single CRIMINALS solutions

1=development aid
2=abortion
3=law & order
4=income differences
5=participation
6=taxation
7=defense





D=development aid
A=abortion
L=law & order
I=income differences
P=participation
T=taxation
N=defense
1,2..=category numbe

Figure 13.5.11 Category quantification for the first two dimensions of multiple nominal CRIMINALS

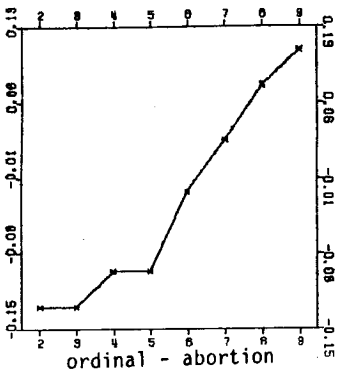
A category weights

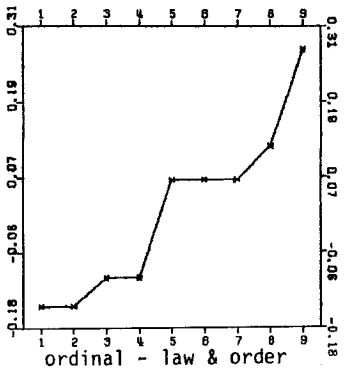
B average object score per category

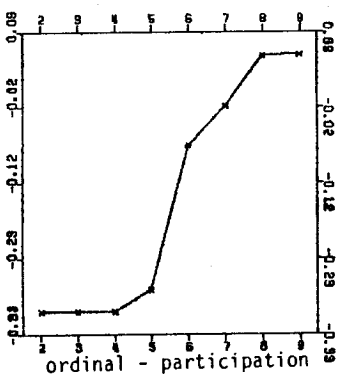
(2) Single nominal. Category quantification is the same for both dimensions. In other words, categories are optimally scaled, and the analysis gives the same results as a linear discriminant analysis on this optimally scaled data matrix, where the quantification is optimal in the sense that the sum of the first two eigenvalues is maximized.
(3) Single ordinal. Optimal quantification is restricted to ordinal transformation. Apart from that the analysis is as in (2).
(4) Numerical. Results will be identical to that of linear discriminant analysis.
Clearly, the four types of analysis, in the given order, are increasing in the degree of constraint. It follows that the measure for stress never can decrease. Results for stress are:

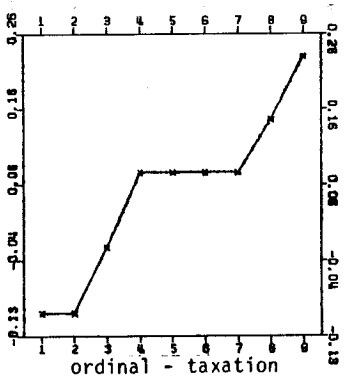
multiple nominal	.115
single nominal	.169
single ordinal	.223
numerical	.304

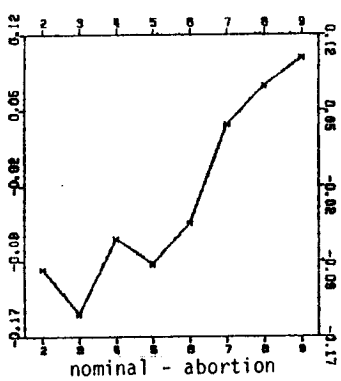
The four plots in figure 13.5.9 show canonical group means for the four solutions. Positions of parties in the four plots are roughly comparable except for DS70: this party moves from an excentric position opposite the confessional parties (multiple nominal) in a direction towards the confessional parties, and finally comes to rest about midway confessional parties and VVD (numerical). This should be interpreted in relation with the fact that there also is a change in how important the different issues are for the canonical variates. Figure 13.5.10 shows correlations between issues and canonical variates for the three 'single' solutions. In the numerical analysis ABORTION is dominant, in the ordinal solution PARTICIPATION and DEFENSE gain in influence, and in the nominal solution INCOMES and TAXATION ask for attention.
Figure 13.5.11A gives a plot of the optimally scaled categories for the first two dimensions of multiple nominal CRIMINALS. Actually, these quantifications are weights for the individual columns of the indicator matrix. As usual, especially for highly collinear variables, such weights tend to behave erratically, and their interpretation is difficult. Note that, unlike in HOMALS, category quantifications are not the averages of scores of individuals within the category. Such averages are plotted in figure 13.5.11B which figure is much more easily interpreted.
One should realize that multiple nominal CRIMINALS has a strong tendency to capitalize on ideosyncratic peculiarities of the data. Whenever some categories are used exclusively by individuals in the same group, CRIMINALS will select this feature for a separate dimensions, even if the number of individuals involved is very small. In the present example there are in

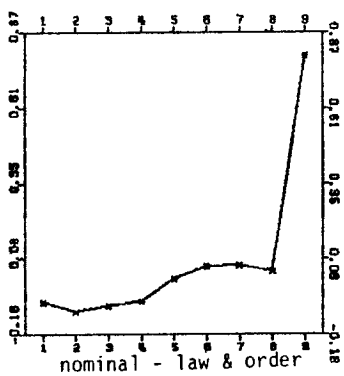


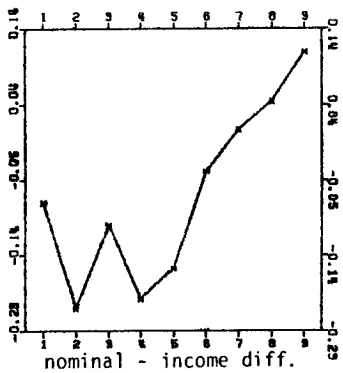












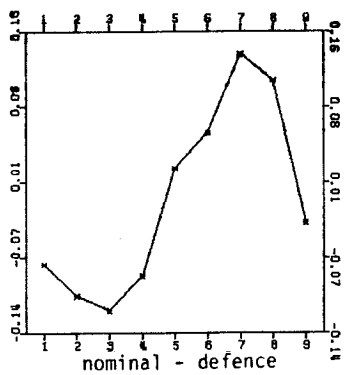
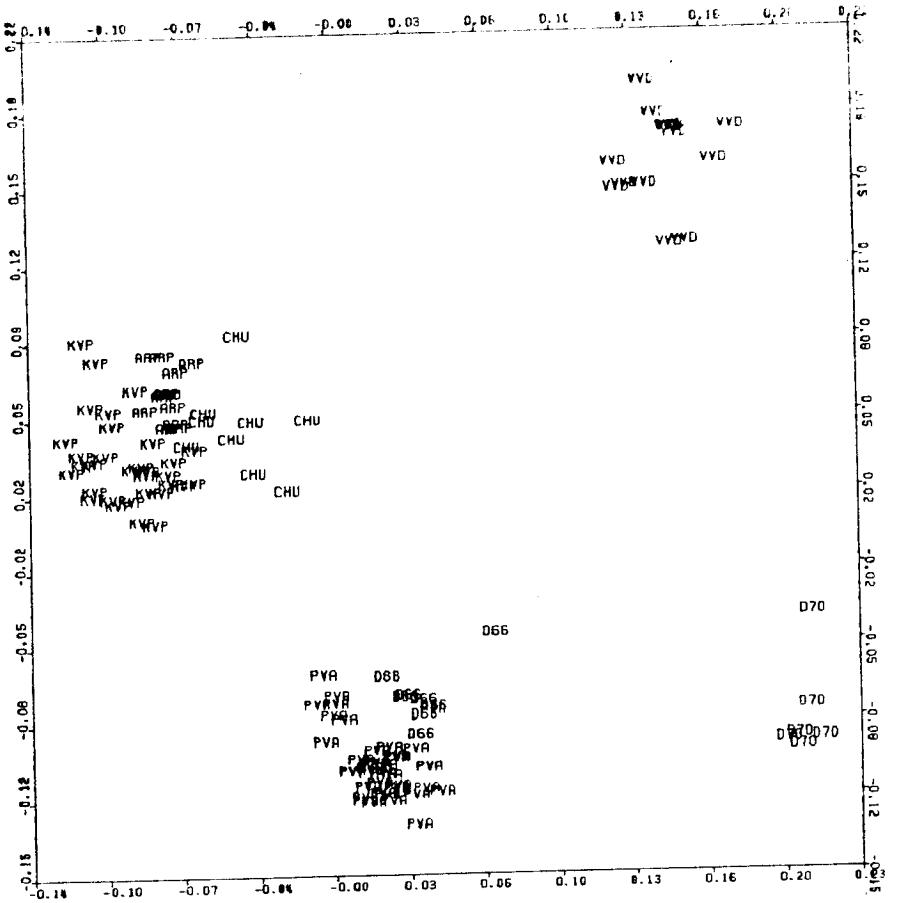
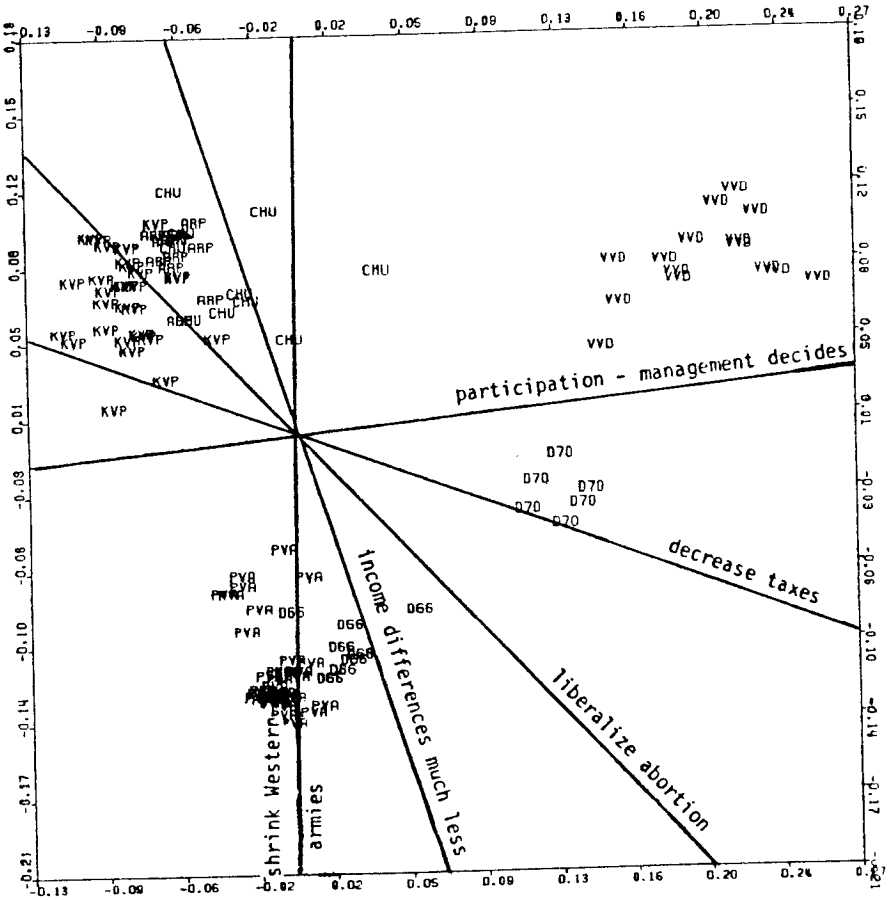


Figure 13.5.12 Some examples of category quantification in single CRIMINALS solutions



A



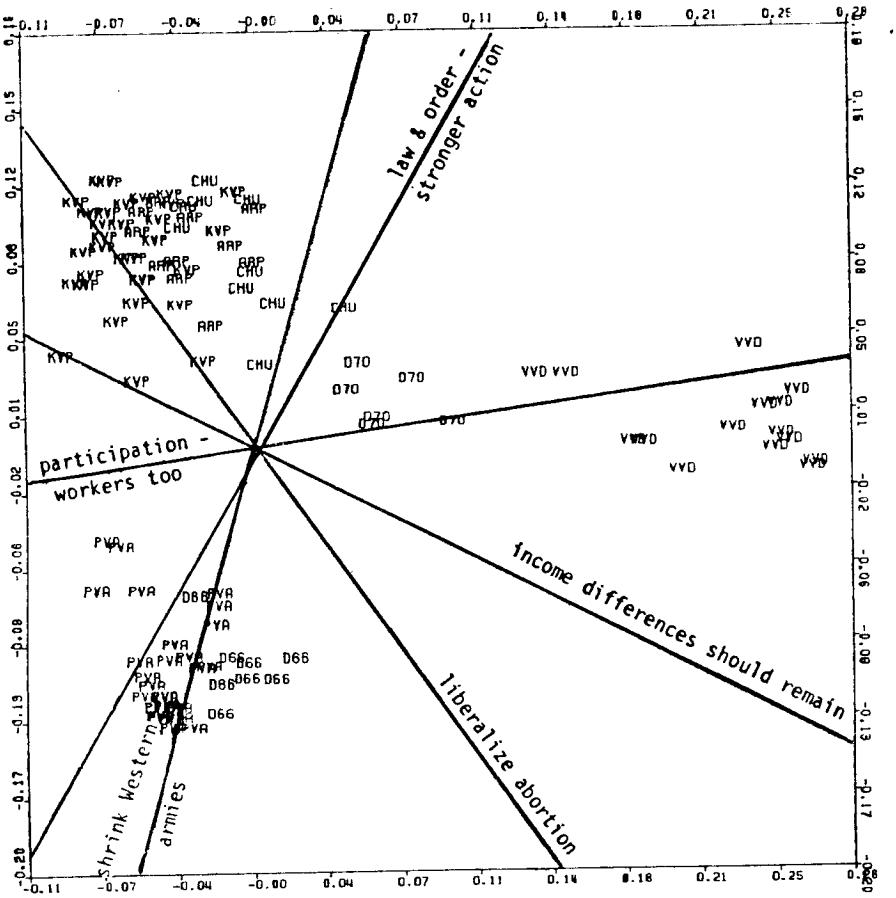
B

Figure 13.5.13 Individual scores first two dimensions CRIMINALS.

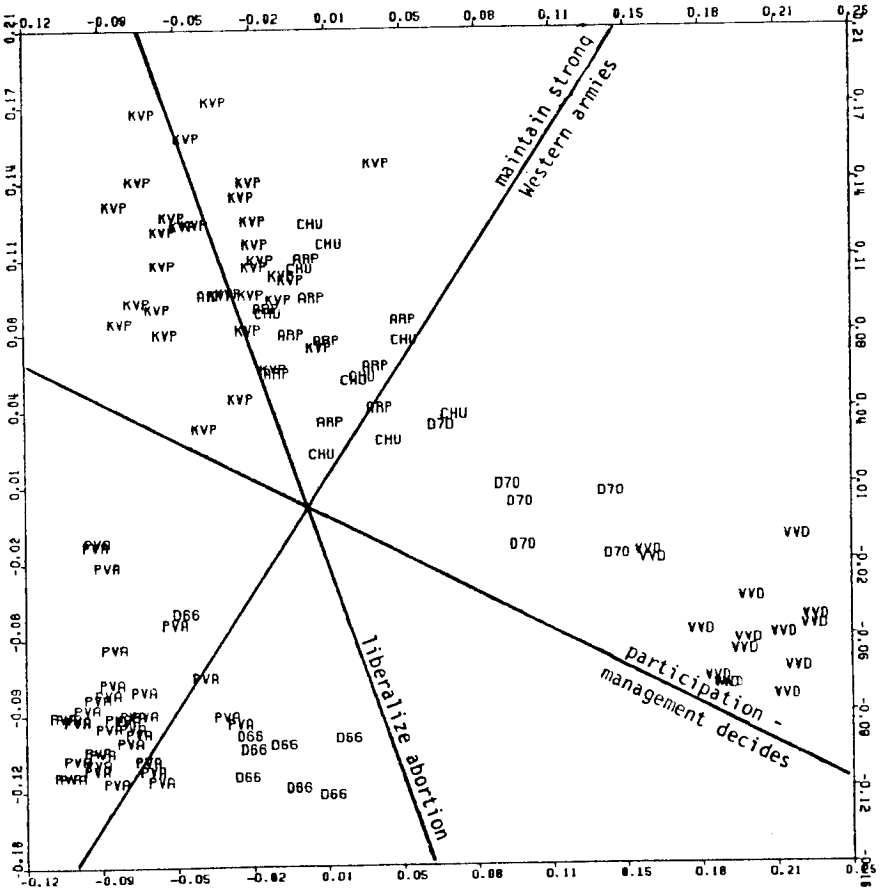
A multiple nominal

B single nominal

C single ordinal


law & order -
stronger action
participation -
workers too
income differences should remain
liberalize abortion
shrink Western armies

C


maintain strong
Western armies
participation -
management decides
liberalize abortion

D

fact a number of categories exclusively used by VVD respondents. They are DEVELOPMENT AID (7) (marginal frequency n=1); LAW & ORDER (9) (n=1); INCOME DIFFERENCES (1) (n=2) and (3) (n=5); PARTICIPATION (2) (n=1), (3) (n=1), (4) (n=2), and (5) (n=5); TAXATION (9) (n=1). Also, TAXATION (6) (n=9) and (8) (n=5), and DEVELOPMENT AID (8) (n=3) are used almost exclusively by VVD and DS70 respondents.

These results are mainly responsible for the CRIMINALS solution. It follows that an issue which is salient in the numerical solution (such as ABORTION) becomes masked in the multiple nominal solution, because its categories do not have such ideosyncratic features. One might say that to the extent the analysis is less restrictive, it will more and more tend to capture details which are, so to speak, at the periphery of the data.

Figure 13.5.12 gives some plots of the optimal scaling for the single options. Note that the single nominal solution gives special values to LAW & ORDER (9), INCOME DIFFERENCES (1) and (3), TAXATION (6), precisely such categories which also dominate in the multiple nominal solution.

Figure 13.5.13 gives plots for individual scores for the three single solutions, with the directions indicated for some of the more salient issues. Obviously, with increasing restriction individuals scatter more around their canonical group means. Nevertheless, in all three solutions respondents from PvdA and D'66, from the denomination parties, and from VVD form strikingly homogeneous clusters.

13.6 Crime and Fear

13.6.1 Introduction

In the Department of Justice a special committee has been installed to investigate the prevention of criminality. They asked the Scientific Research and Documentation Centre of the department (W.O.D.C.) to make a number of surveys. We use some of the results of one of those, on judgments and feelings about the criminality question.
These judgments were supposed to be relevant for understanding the priority of criminality prevention and for determining margins to the humanization of the administration of criminal law. (C. Cozijn, J.J.M. van Dijk, Onrustgevoelens in Nederland, W.O.D.C., juli 1976).

13.6.2 Description of the data

The survey was over 1219 respondents, and 48 questions were asked. We used the following selection of variables.

A. Six questions concerning judgments about the effectiveness of different methods to fight crime;
B. Four questions about feelings of helplessness and unrest;
C. Four variables with background information about the respondents.

We dropped three individuals because they had systematic missing observations and applying option II or III for missing data did not give satisfactory solutions. They gave perfect discrimination for variables with the missing observations.

A. Give your judgment about the effectiveness to fight crime of the following methods. (We shall call these variables CRIME henceforth).
   1. Re-education of criminals (EDUC)
   2. Locking up of criminals (LOCK)
   3. More severe punishment (PUNI)
   4. Social work, rehabilitation (SOWO)
   5. Labor-camps (CAMP)
   6. Better employment for criminals (EMPL)

Methods 1, 4 and 6 were supposed to be 'social preventive' methods, (we shall label them SOCIAL), 2, 3 and 5 to be the 'penal law' approach (PENAL).
Answer categories for SOCIAL:

1. very ineffective
2. ineffective
3. neither ineffective nor effective, and don't know
4. effective
5. very effective

Answer categories for PENAL were the other way around.

B. Feelings of helplessness and unrest (FEAR)

1. You have to watch out when you walk in the city (CITY)
2. It is unwise nowadays to go outdoors at night (DARK)
3. You cannot even rely on the police anymore (POLI)
4. When something happens to you in the street, you cannot expect aid from someone (AID)

Answer categories:

1. agree completely
2. agree
3. neither agree nor disagree, and don't know
4. disagree
5. disagree completely

Categories 1 and 2, and 4 and 5 have been combined on the basis of previous analyses.

C. Background information

C1 Religion (RELI)

1. Reformed (REFOR)
2. Protestant (PROT)
3. Catholic (CATHO)
4. Other religions (OTHER)
5. No religion (NOREL)

7 categories specifying 'reformed' were combined.

C2 Voting behavior (VOTE)

1. Labor Party (LABOR)
2. Conservative-liberal Party (CONSE)
3. Catholic denominational Party (CATHP)
4. Protestant Party (PROTP)
5. Protestant Party (PROTP)
6. Radical Party (RADIC)
7. Abstention (ABST)
8. Do not know (DOKN)
9. Do not vote (DOVO)
10. Other parties (OTHEP)

Occupational status and sex were combined to an interactive variable (see section 2.9). The first seven categories apply to males, the last seven to females.

C3 Occupational status

1. Higher ranked employees (and comparable)	(HIGHM)	8. (HIGHF)
2. Middle ranked employees	(MIDDM)	9. (MIDDF)
3. Small business	(BUSIM)	10. (BUSIF)
4. Lower ranked employees	(LOWM)	11. (LOWF)
5. Skilled workers	(SKILM)	12. (SKILF)
6. Unskilled workers	(UNSKM)	13. (UNSKF)
7. No profession	(NOPRM)	14. (NOPRF)

C4 Age

1. 16 - 17
2. 18 - 24
3. 25 - 34
4. 35 - 49
5. 50 - 64
6. 65 - 70

	Categories													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
EDUC	86	234	232	491	173									
LOCK	447	436	174	140	19									
PUNI	735	221	121	104	35									
SOWO	44	107	189	538	338									
CAMP	485	318	134	201	78									
EMPL	46	102	193	525	350									
CITY	505	164	547											
DARK	366	144	706											
POLI	359	266	591											
AID	292	213	711											
RELI	117	200	409	29	461									
VOTE	316	127	127	73	49	51	4	207	172	90				
OCCU	22	106	41	99	110	43	160	2	16	9	76	23	13	496
AGE	77	188	253	309	271	118								

Table 13.6.1 Marginal frequencies for CRIME, FEAR and background variables

13.6.3 Overview of analyses

Several ways to explore relations between variables 1-6 (CRIME), 7-10 (FEAR) and 11-14 (background variables) have been tried.
Two main approaches can be distinguished:
Non-linear analyses: HOMALS, PRINCALS, CANALS, and
Linear analyses on the basis of one-dimensional HOMALS solutions (applications of HOMALS as first step): principal components analysis (PCA), canonical correlation analysis (CCA).
Within these approaches several ways to treat background variables are discussed. They are taken as multiple nominal variables in HOMALS and PRINCALS, as single nominal variables in PRINCALS and CANALS and a comparison has been made between them having an active or a passive role in the analysis. By active we mean that they have the same status in the analysis as CRIME and FEAR variables. To say they have been treated as passive variables amounts to quantifying their categories afterwards on the basis of the individual scores obtained from the analysis of CRIME and FEAR variables only. Optimal ('active') and non-optimal ('passive') quantifications are used as transformed data to apply linear MVA afterwards.

13.6.4 Multiple join solutions over all variables

HOMALS has been applied in the first place to all variables. Transformations for CRIME and FEAR categories are monotone increasing in the first dimension. For ease of interpretation category quantifications have been split up to make up four plots. Figure 13.6.1.a (upperside) gives category points for CRIME variables and RELI plus VOTE, the bottom of the plot (13.6.1.b) gives the same points for RELI and VOTE, but now plotted with category points for FEAR variables. By inspecting positions for RELI and VOTE points, we can insert FEAR points in the plot for CRIME points. The same thing is true for figure 13.6.2, but now for the background variables OCCU and AGE, with the same category points for CRIME and FEAR as in 13.6.1.
Labeling of CRIME points: 1=very ineffective, 2..., 5=very effective;
labeling of FEAR points: 1=(completely) disagree, 3, 5=(completely) agree.
Starting in the center of figure 13.6.1.a we see a cluster of category-4 points. A lot of people think that all methods to fight crime (SOCIAL and PENAL) are effective. Going in the direction of the upper right corner, we find a cluster representing people who think of SOCIAL as being very effective (points labeled with 5's), with extreme opinions about PENAL as very ineffective when we arrive at the upper right corner. Going back to the center and making the same movements but now to the upper left corner, we encounter PENAL as very effective, with at the extreme people regarding SOCIAL as very ineffective.

We could interpret this configuration as representing a scale from SOCIAL-(very) ineffective $\overset{*5}{\rightarrow}$ PENAL-very effective $\overset{*5}{\rightarrow}$ PENAL and SOCIAL-effective $\overset{*1}{\rightarrow}$ SOCIAL-very effective $\rightarrow$ PENAL-(very) ineffective. The asterisks have been used to indicate the positions of category points for FEAR variables in the plot. It should be noticed that these variables play a far less dominant role in the analysis.

Some remarks about background variables. Inspecting category points for VOTE we find a very prominent position for the radical party (RADIC), close to PENAL-(very) ineffective. Conservative-liberals (CONSE) are more to the PENAL-(very) effective side, as are the denominational parties (PROTP and CATHP). The labor party seems to take an intermediate position, slightly closer to SOCIAL-(very) effective and very close to FEAR-1 categories (disagree). Category points for religion also indicate PENAL-(very) effective for protestants and reformed. For AGE we see 35-70 PENAL-(very) effective, 16-24 SOCIAL-very effective and PENAL-ineffective; 25-34 intermediate.

For OCCU small businessmen (BUSIM) and unskilled males (UNSKM) are very close to SOCIAL-(very) ineffective, all kinds of employees (HIGH, MIDD, LOW) close to SOCIAL-effective, females with no profession, most likely housewifes, PENAL-effective. The position of higher ranked females seems rather strange. Inspecting marginal frequencies gives only two women of the 1216 respondents in this category, which may account for the marginal position of HIGHF.

Figure 13.6.3.a gives optimal quantifications for CRIME and FEAR only, background variables have been treated as passive. We recover the main features of figure 13.6.1 (SOWO and EMPL, LOCK and PUNI 1-categories are at the extremes), but the category points for SOCIAL-very effective and PENAL ineffective on the one side are closer together. The same applies, to an even larger extent, to their counterparts on the other side.

The bottom of figure 13.6.3 gives non-optimal category quantifications, computed afterwards, for AGE and OCCU. The category points are connected with the 'optimal' points from the previous solution.

The difference between optimal and non-optimal is clear. We see points for BUSIM and UNSKM move towards the center. Those were the categories most connected with SOCIAL-ineffective categories. An interesting move is made by HIGHF; higher ranked females are very close now to EDUC-4, middle ranked females and corresponding categories for males.

13.6.5 Single join solutions over all variables

We continue our data analysis by looking at the linear PCA solutions for the first dimensions of the two preceding analyses. Figure 13.6.4 gives component

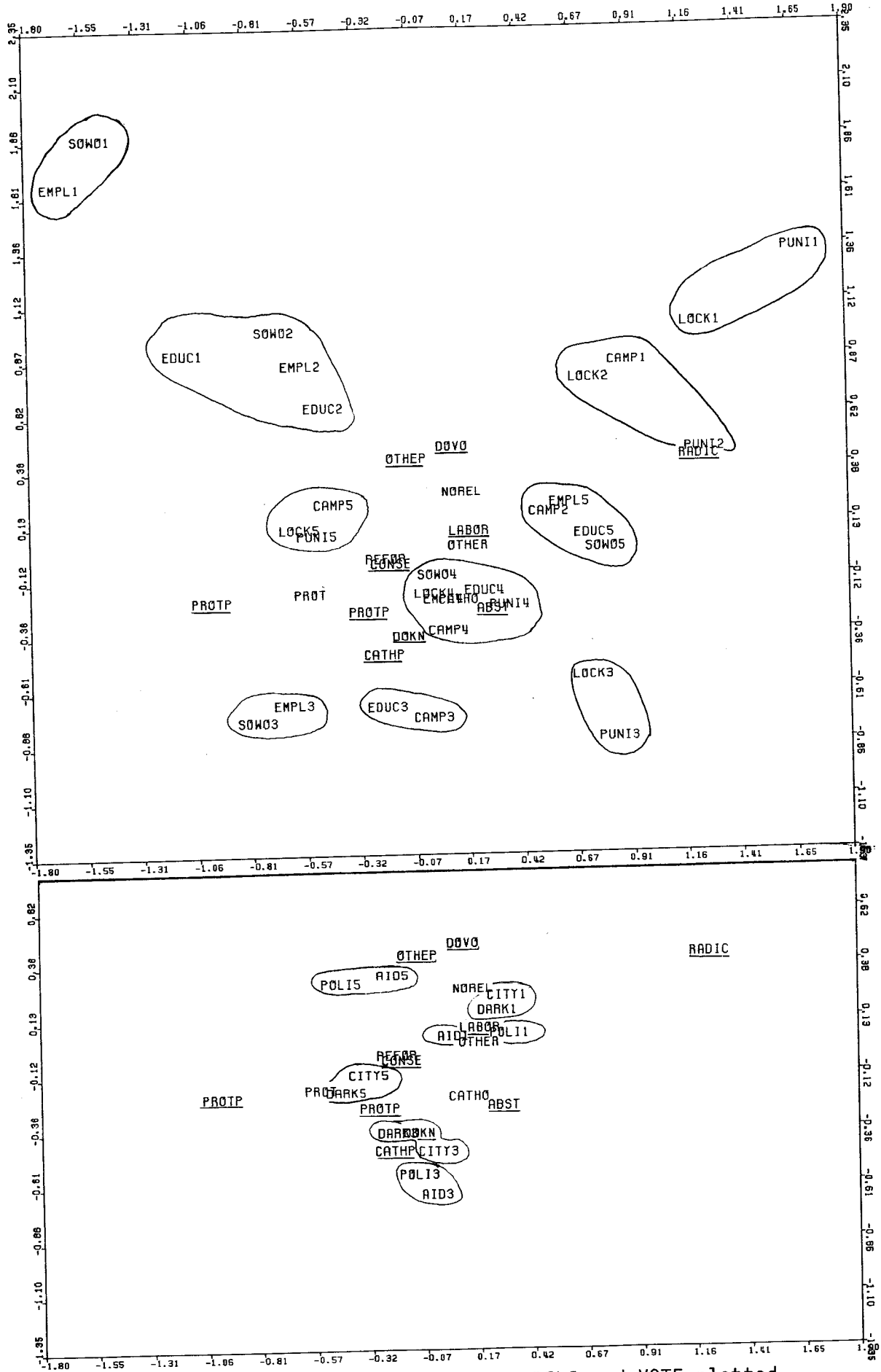


Figure 13.6.1 HOMALS over all variables, RELI and VOTE plotted, a. with CRIME, b. with FEAR

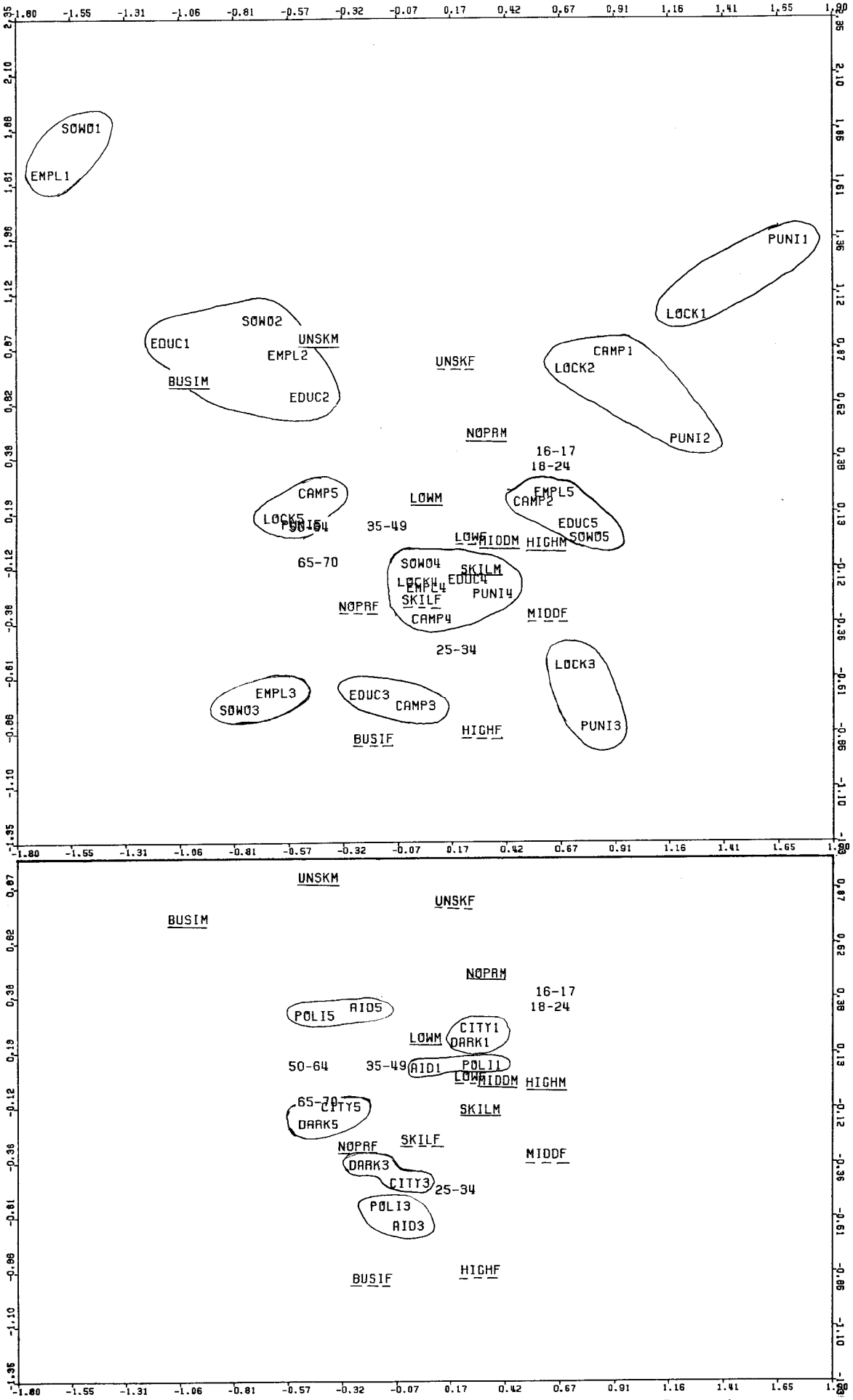


Figure 13.6.2 HOMALS over all variables, OCCU and AGE plotted, a. with CRIME, b. with FEAR

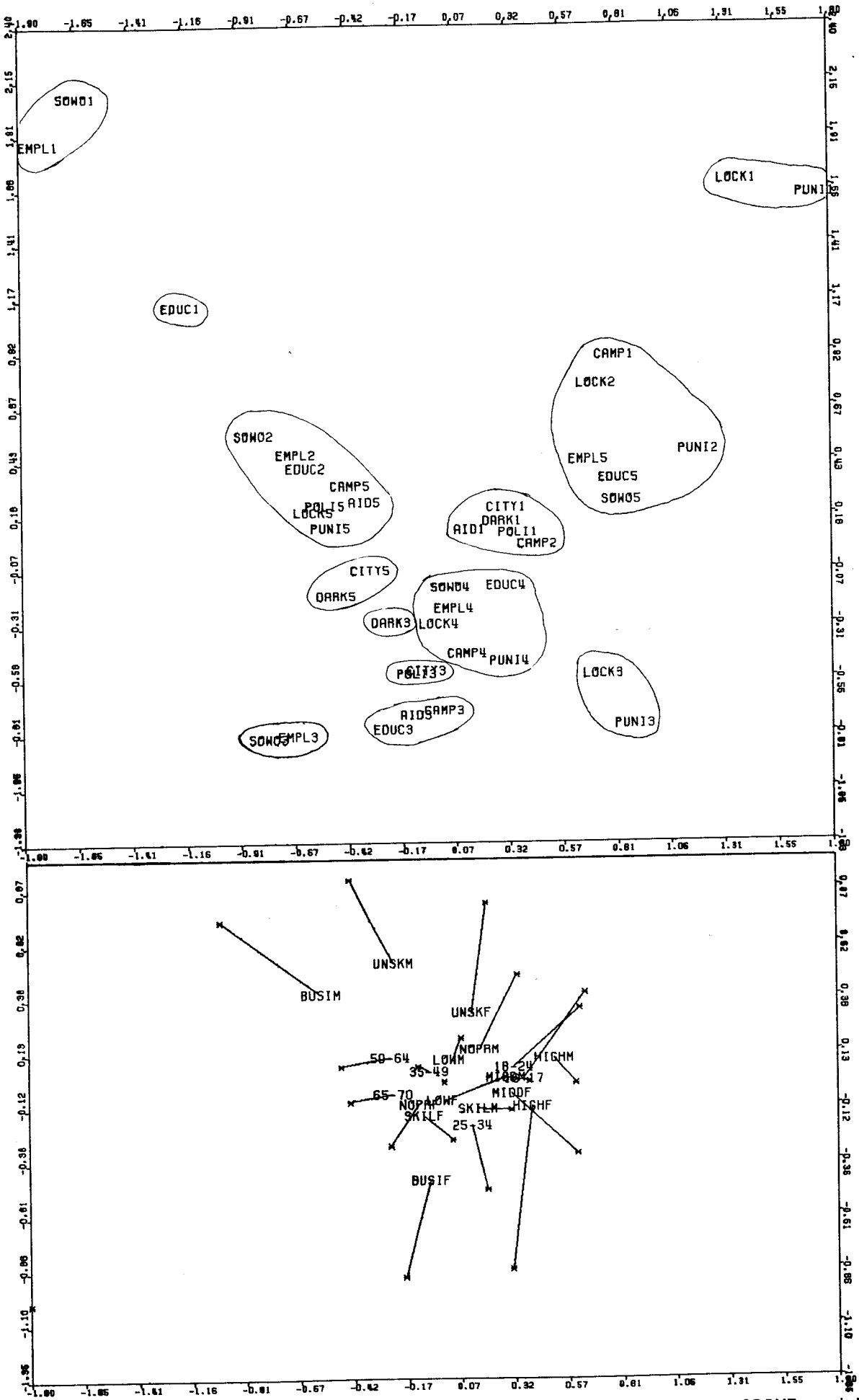


Figure 13.6.3 a. Optimal HOMALS category quantifications for CRIME and FEAR
b. Optimal and non-optimal HOMALS category quantifications for OCCU and AGE

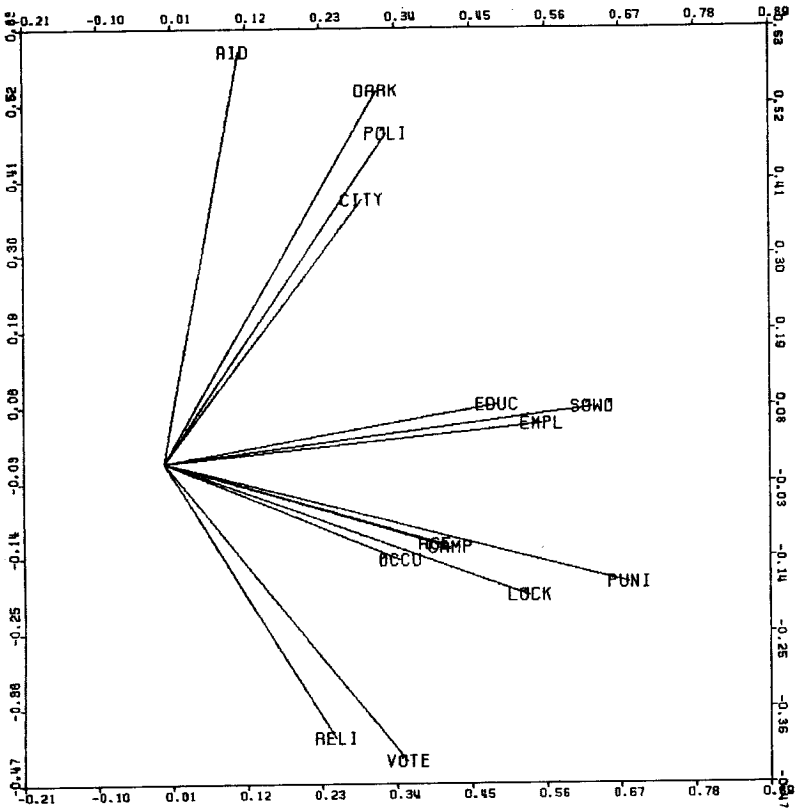


Figure 13.6.4 PCA after HOMALS, background variables active

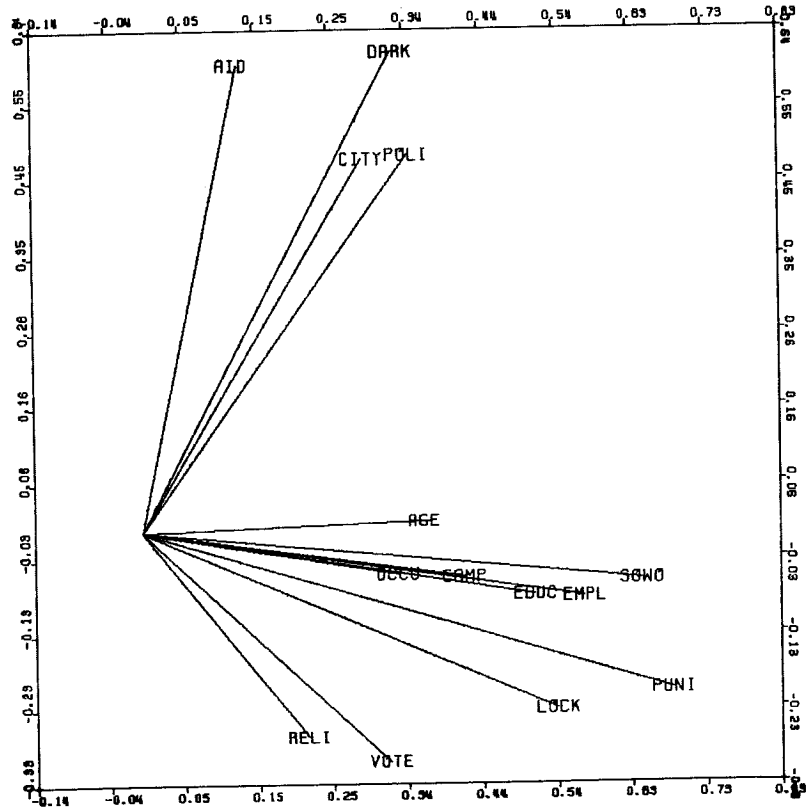


Figure 13.6.5 PCA after HOMALS, background variables passive

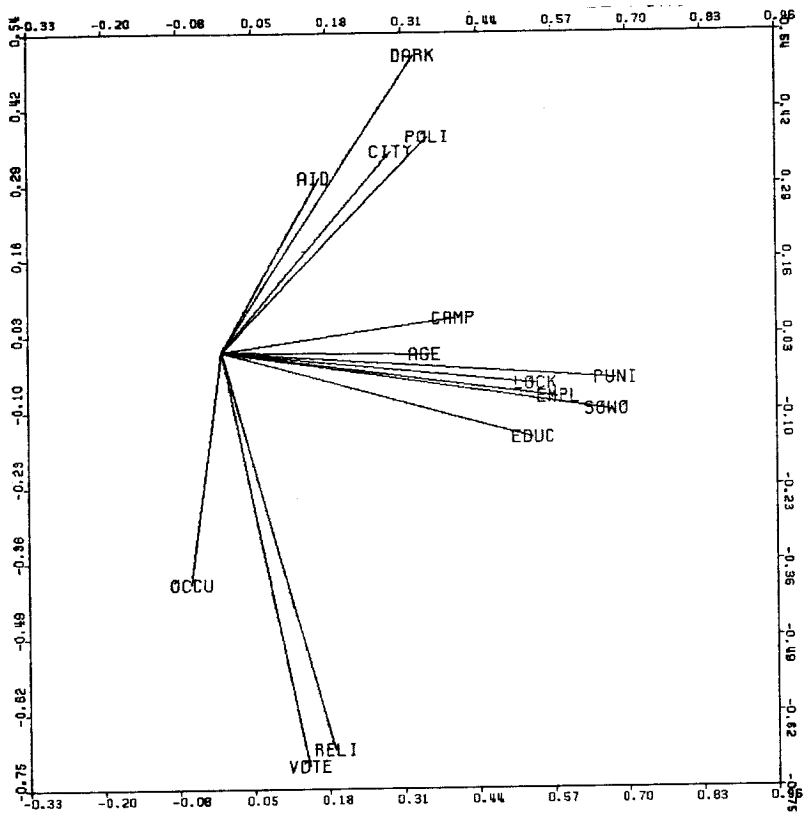


Figure 13.6.6 PRINCALS over all variables, background variables single nominal

loadings on the basis of optimal quantifications for CRIME, FEAR and background categories (eigenvalues 2.67 and 1.52), figure 13.6.5 gives loadings on the basis of non-optimal quantifications for AGE, OCCU, RELI and VOTE (eigenvalues 2.65 and 1.47).

Vectors are plotted in one direction only (but can ofcourse be mirrored) and have the following interpretation. FEAR variables-disagree, SOCIAL CRIME-very effective, PENAL CRIME-very ineffective. Transformations for the background variables give the interpretations: RELI-no religion (versus protestant), VOTE-radical party (versus protestant), AGE-young (versus old) and OCCU-higher and middle, male and female (versus no profession-female, unskilled-male and small business-male).

In both figures we see a clear distinction between FEAR and CRIME variables. The most interesting difference between 13.6.4 and 13.6.5 is the differentiation between SOCIAL and PENAL when background variables are active. With background variables passive this differentiation is more or less lost and CAMP, AGE and OCCU are more connected with SOCIAL variables.

A PRINCALS analysis has been performed with single ordinal options for CRIME and FEAR and single nominal options for the background variables. Figure 13.6.6 gives results for component loadings.

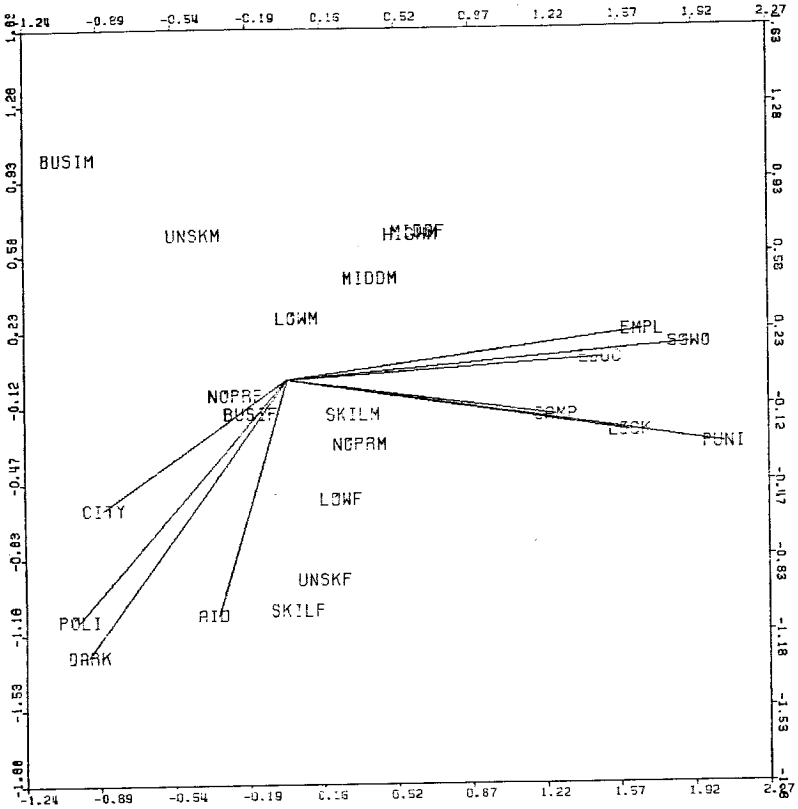


Figure 13.6.7 PRINCALS over all variables, background variables multiple nominal

At first sight this plot resembles the PCA plot 13.6.5 : a distinction between FEAR and CRIME, no distinction between SOCIAL and PENAL. Looking at the background variables however, some things have changed. RELI and VOTE are much less related to CRIME and more to DARK, POLI, CITY (and AID). According to the direction in which the vector is plotted, catholics and catholic party by far have the highest category quantifications, which is especially striking for the catholic party (see table 13.6.2).
For OCCU things have changed most. This variable has nothing left in relation to CRIME, but has a lot to do with the direction FEAR-agree (which is not plotted). Looking at the single category quantifications for OCCU this direction is related to all categories for females (except HIGHF and MIDDF), versus the plotted direction for all male categories, except SKILLM, with HIGHM, HIGHF and BUSIM at the extreme.
Another PRINCALS analysis has been performed, but now with multiple nominal options for the background variables. Results are plotted in figure 13.6.7, with multiple category points for OCCU.
We have recovered the distinction between SOCIAL and PENAL; vectors for FEAR are now plotted in the direction 'agree'. The rankorder for OCCU categories on the first dimension is almost identical to the ordering found by HOMALS. (For completeness it should be mentioned that category point HIGHF is not plotted; it had a rather extreme position on the second dimension).

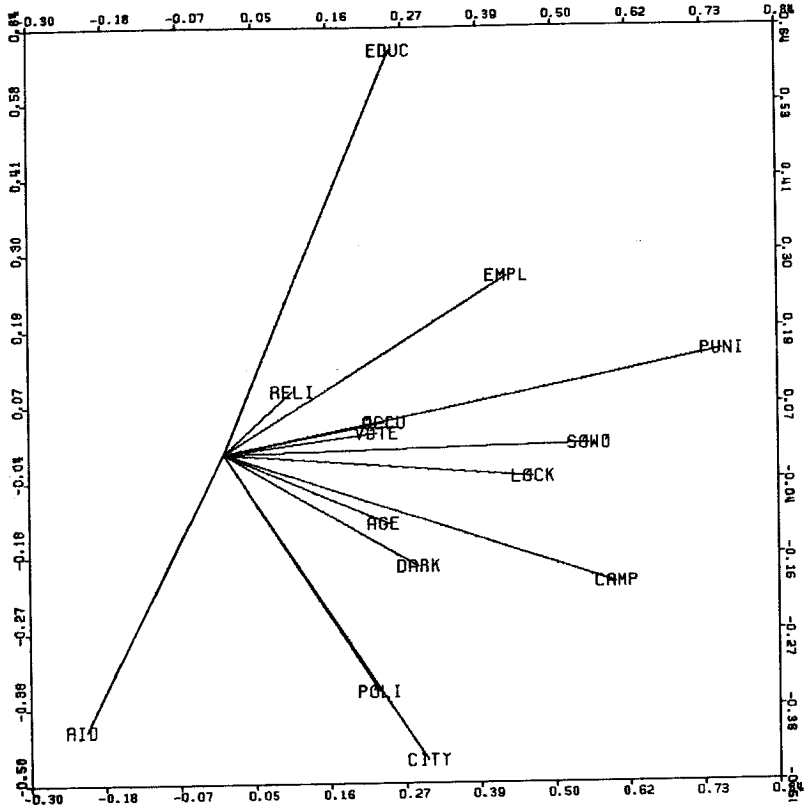


Figure 13.6.8 CCA after HOMALS over all variables

13.6.6 Single meet solutions over all variables

The transformed data matrices, which have been used for PCA in section 13.6.5, have also been used for linear canonical correlation analysis, with CRIME and FEAR in one set and the background variables (active or passive) in the other. Results are presented for CCA with non-optimal quantifications for RELI, OCCU, VOTE and AGE only, because figures are almost identical and canonical correlations are slightly higher (.383 / .159). The vectors are plotted in the plane of the canonical variates for CRIME and FEAR variables (figure 13.6.8). Very striking is the almost opposite direction of EDUC-very effective and AID-disagree. These variables have the closest connection with RELI-catholics and RELI-protestants respectively.
OCCU (HIGHM, HIGHF, MIDDM, MIDDF) is closely connected with PUNI- and LOCK-ineffective and SOWO-effective; the same applies to VOTE (radical party).
AGE (young) is closely related with CAMP-ineffective and DARK (disagree).

CANALS has been applied with single ordinal options for CRIME and FEAR and single nominal options for the background variables.
The first two dimensions of the four dimensional solution are presented; canonical correlations are .372 and .285 (figure 13.6.9).
The most interesting feature is the long vector for DARK, perfectly correlated

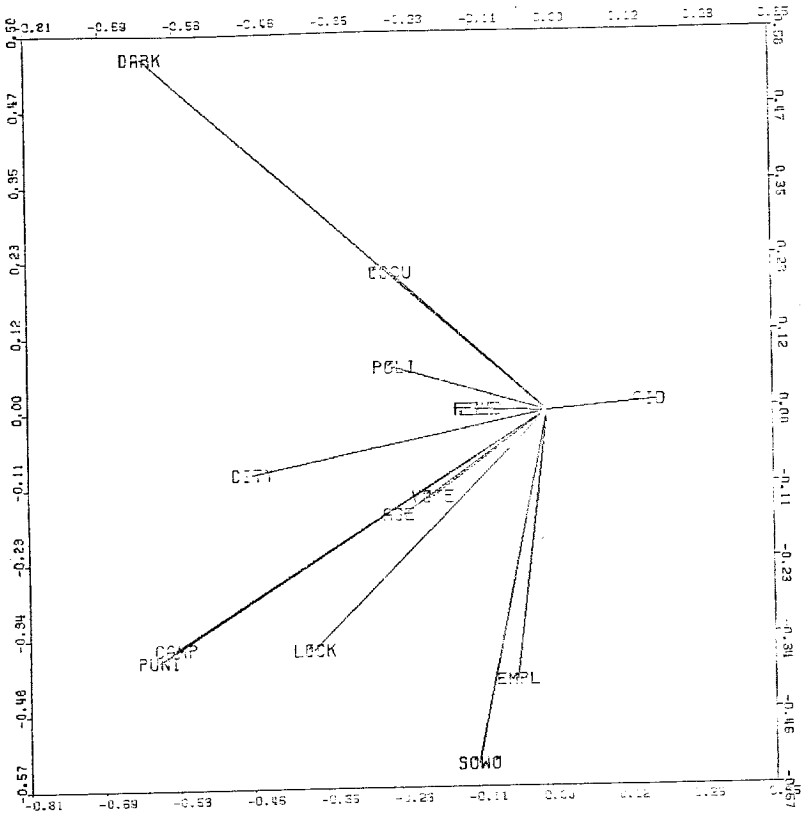


Figure 13.6.9 CANALS over all variables

with OCCU. OCCU categories have been transformed very much like the PRINCALS solution with single nominal options. (see tabel 13.6.5). Again the main division is between males and females (except HIGHF and MIDDF), MIDDM, HIGHM and BUSIM are at the other extreme.
RELI and EDUC (labels almost illegible) are opposite to AID (like in figure 13.6.8), which is pointing to the other direction compared to the other FEAR variables. AGE and VOTE have the closest connection with PENAL-ineffective. SOWO and EMPL represent very effective only. All other category transformations have a negative sign and must be thought of as lying in the opposite direction.

13.6.7 Single meet solutions over CRIME and FEAR variables

Results for CRIME and FEAR categories from the HOMALS analysis with background variables passive have been used for linear CCA with CRIME in the first set and FEAR in the second one.
Results are rather disappointing for projections of FEAR in the plane of CRIME canonical variates (figure 13.6.10). Canonical correlation on the second dimension is very small (.235 / .099).
CITY and DARK are pointing in the same direction as PENAL-ineffective; AID and POLI are more connected with SOCIAL-effective.

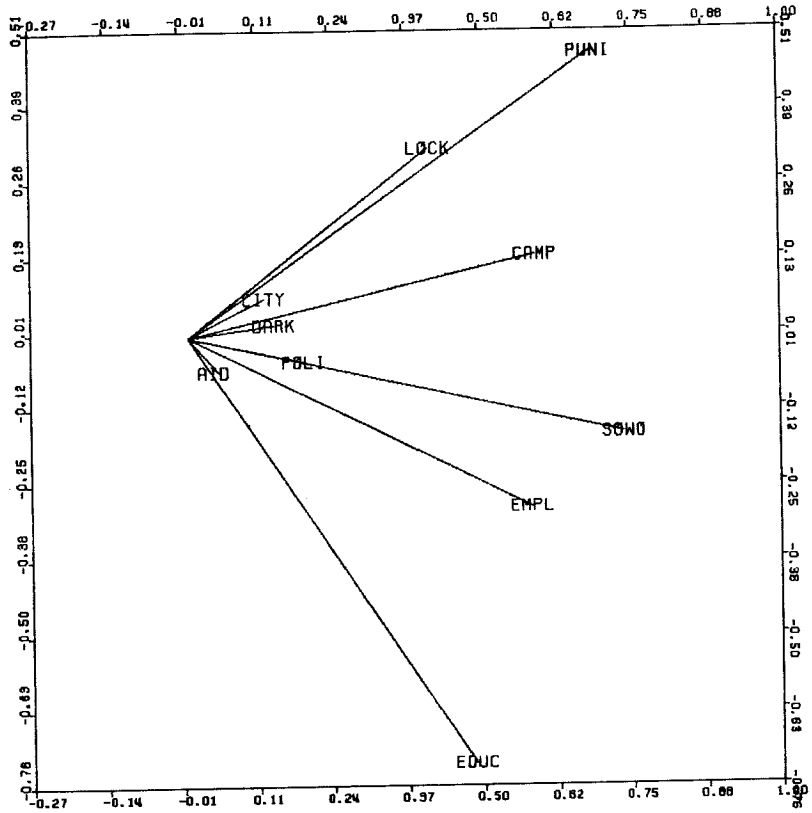


Figure 13.6.10 CCA after HOMALS over CRIME and FEAR

It seemed interesting to compare results for CCA when one-dimensional HOMALS had been applied to CRIME variables and FEAR variables separately.
Single join solutions are given in figure 13.6.11; the CCA solution is presented in figure 13.6.12. (Canonical correlations .225 / .099).
Comparing figure 13.6.12 with figure 13.6.10 we see almost identical solutions. In this case it makes no difference if we use transformed data from one HOMALS analysis or use data transformed separately.

Finally these analyses of two sets can be compared with the CANALS single ordinal solution in figure 13.6.13 (canonical correlations .255 and .130).
The relation between CRIME and FEAR variables shows the same pattern: AID and POLI related with SOCIAL, DARK and CITY with PENAL.
Relations between FEAR variables give no longer perfect correlation between LOCK and PUNI; between CRIME variables themselves SOWO and EMPL are closer connected. The rankorder of the projections on the first dimension is the same as resulting from the linear canonical correlation solutions.

13.6.8 Rankorders for the categories of the background variables

In table 13.6.2 - 13.6.5 rankorders, obtained from our several analyses, are given. For multiple join HOMALS solutions rankorders on the first dimension

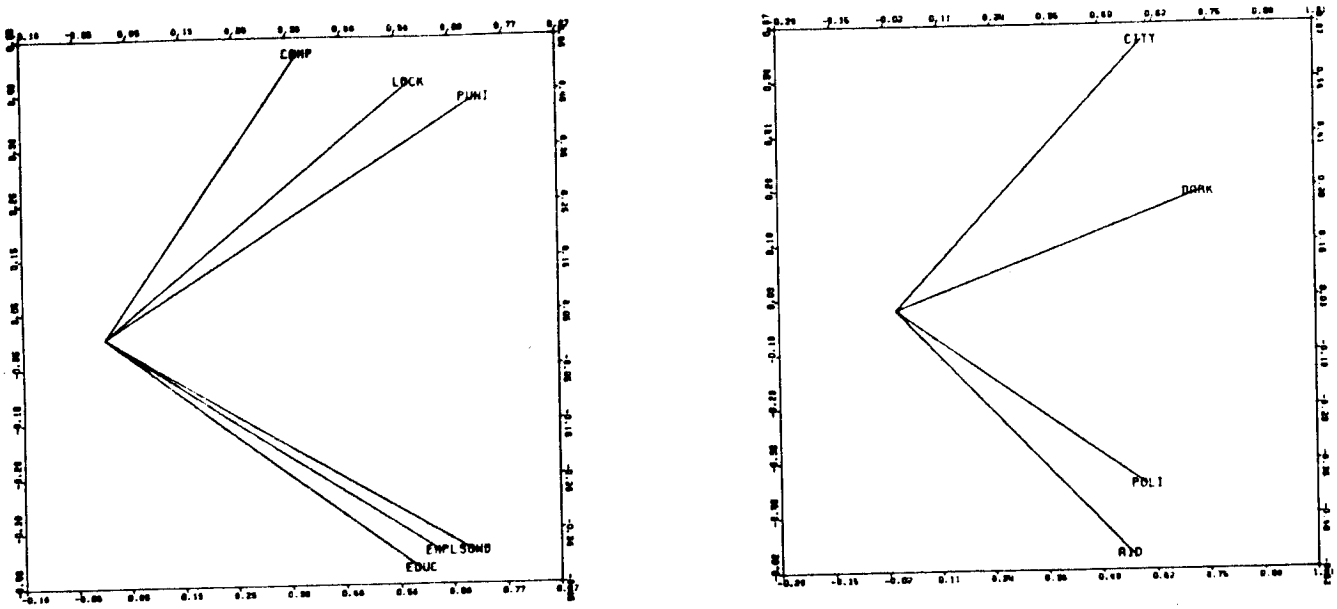


Figure 13.6.11 PCA after HOMALS over CRIME and HOMALS over FEAR

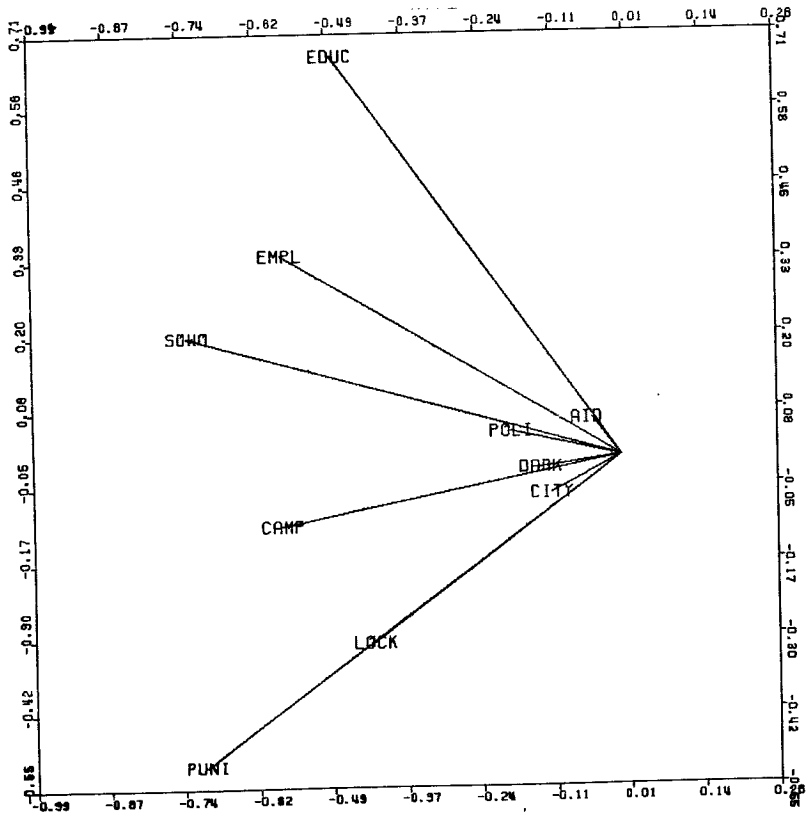


Figure 13.6.12 CCA after HOMALS over CRIME and HOMALS over FEAR

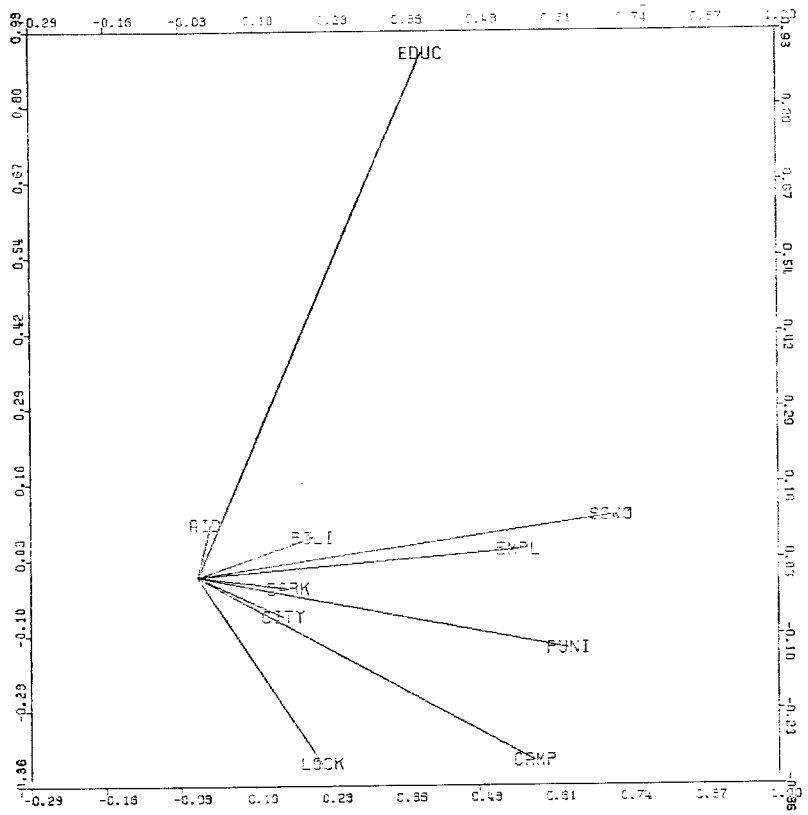


Figure 13.6.13 CANALS over CRIME and FEAR single ordinal

have been taken; the same applies to PRINCMULT, which means PRINCALS results with background variables multiple nominal (PRINCSING: these variables single nominal). HOMACT means optimal HOMALS quantifications, HOMPAS non-optimal quantifications. A dash under rankorders indicates that these categories should be thought of as lying in the direction of the background vectors plotted in figures 13.6.4 upto 13.6.13.

	PROT1	PROT2	CATHP	CONSE	LABOR	RADIC
HOMPAS	1	4	3	2	5	6
HOMACT	1	2	3	4	5	6
PRINCMULT	1	2	3	4	5	6
PRINSING	2	1	6	3	4	5
CANALS	1	2	4	3	5	6

Table 13.6.2 Rankorders for VOTE

PRINCMULT gives the same rankorders as HOMACT, HOMPAS changes ranknumbers for PROT2 and CONSE. PRINCSING gives exceptional ranknumber to CATHP.
Results for LABOR, RADIC and PROT1 are rather stable.

	REFOR	PROT	OTHER	CATHO	NOREL
HOMPAS	2	1	5	4	3
HOMACT	2	1	4	3	5
PRINCMULT	2	1	5	3	4
PRINCSING	1	2	3	5	4
CANALS	3	4	1	2	5

Table 13.6.3 Rankorders for RELI

Rankorders for REFOR and PROT are very similar. CANALS gives rather deviating ranknumbers, especially for PROT, but also for OTHER and CATHO.

	65-70	50-64	35-49	25-34	18-24	16-17
HOMPAS	2	1	3	4	5	6
HOMACT	2	1	3	4	5	6
PRINCMULT	2	1	3	4	5	6
PRINCSING	1	2	3	4	5	6
CANALS	2	1	3	4	5	6

Table 13.6.4 Rankorders for AGE

Rankorders for AGE are very stable over all analyses.

	HIGHM	MIDDM	BUSIM	LOWM	SKILM	UNSKM	NOPRM
HOMPAS	14	11	1	7	9	2	10
HOMACT	13	12	1	6	10	2	11
PRINCMULT	13	12	1	6	9	2	10
PRINCSING	12	11	13	8	6	10	7
CANALS	13	14	12	8	9	11	10
	HIGHF	MIDDF	BUSIF	LOWF	SKILF	UNSKF	NOPRF
HOMPAS	13	12	5	6	4	8	3
HOMACT	9	14	4	8	5	7	3
PRINCMULT	11	14	4	8	5	7	3
PRINCSING	14	9	5	2	1	3	4
CANALS	7	6	2	3	1	4	5

Table 13.6.5 Rankorders for OCCU

We see striking differences for BUSIM, LOWF, SKILF and UNSKF if we compare HOMPAS,HOMACT and PRINCMULT (multiple quantifications) with PRINCSING and CANALS (single quantifications).
In relation with solutions where OCCU is connected with CRIME variables, there is not much difference between males and females in general. Between males and females with the same occupational status the largest difference is between the no profession-categories, and between the unskilled-categories. Within males BUSIM and UNSKM have a deviant opinion, within females BUSIF and NOPRF.
PRINCSING and CANALS solutions are related to DARK: males are not afraid, higher and middle ranked females neither, all other women are.

Appendix A Matrix Algebra

A1 Images

A1.1 Let x be an n-tuple of n real valued numbers. Let $\mathbb{R}^n$ be the field of all such n-tuples. An interpretation of x as a vector in space becomes possible by defining in $\mathbb{R}^n$ a set of coordinate vectors. Usually one takes for these coordinate vectors the elementary vectors e_i (i=1,..,n), corresponding to the column vectors of the n x n identitity matrix X. Such elementary vectors satisfy $e_i'e_i=1$ (for all i) and $e_i'e_j=0$ (for all $i\neq j$), so that they form an orthogonal coordinate system, with unit-length coordinates, also a basis of $\mathbb{R}^n$.

The vector x is defined in space as the weighted sumvector

$$x = x_1e_1 + x_2e_2 + \ldots + x_ie_i + \ldots + x_ne_n$$

An example for n=3 is

$$x = \begin{bmatrix}2\\3\\4\end{bmatrix} = (2)\begin{bmatrix}1\\0\\0\end{bmatrix} + (3)\begin{bmatrix}0\\1\\0\end{bmatrix} + (4)\begin{bmatrix}0\\0\\1\end{bmatrix}$$

A1.2 Let A be an n x m matrix. A'x is said to be an image of x. The reason for this name becomes immediately obvious by writing

$$A'x = x_1a_1 + x_2a_2 + \ldots + x_ia_i + \ldots + x_na_n$$

which shows that A'x is constructed in the same way as a weighted sumvector of the columns a_i of A', as x was constructed from the colums of I.

The image A'x of x is an m x 1 vector, and "pictures" the n x 1 vector x as its image in $\mathbb{R}^m$. In the same way, let y be an m x 1 vector in $\mathbb{R}^m$, then its image Ay is an n x 1 vector in $\mathbb{R}^n$.

A1.3 As an example, let n=3, m=2, and

$$A = \begin{bmatrix}.5 & .3\\.3 & .1\\.4 & .5\end{bmatrix}$$

Let y be some vector in $\mathbb{R}^2$. For simplicity we take y in such a way that y'y=1, with

$$y = \begin{bmatrix}.200\\.980\end{bmatrix}$$

Figure A.1 shows how y is constructed as the weighted sumvector $y = (.200)e_1 + (.980)e_2$. Figure A.2 represents $\mathbb{R}^3$, with elementary vectors $\bar{e}_1$, $\bar{e}_2$, $\bar{e}_3$. In the figure, the two column vectors a_1 and a_2 of A have been drawn. According to the definition, a_1 is the image of e_1 (since $Ae_1 = a_1$). and a_2 is the image of e_2. Also

$$Ay = \begin{bmatrix}.394\\.158\\.570\end{bmatrix}$$

becomes the image of y, and, in figure A.2 is constructed in the same way from a_1 and a_2, as y was constructed in figure A.1 from e_1 and e_2.

We now extend Ay to a vector of unit length, and call this extended vector x. Since y'A'Ay = .505, $x = Ay.(.505)^{-\frac{1}{2}} = Ay/(.711)$. In figure A.3 x is drawn as a vector in $\mathbb{R}^3$. Note that figures A.2 and A.3 can be superimposed upon each other; both figures represent the same $\mathbb{R}^3$ in the same way. In figure A.3 x has the same direction as Ay in figure A.2. Since

$$x = Ay/(.711) = \begin{bmatrix} .554 \\ .222 \\ .802 \end{bmatrix}$$

x can be constructed in figure A.3 as the sumvector $.554\bar{e}_1+.222\bar{e}_2+.802\bar{e}_3$, as shown in figure A.3 where x appears as the body diagonal of a rectangular parallelopiped.

Figure A.4 shows the image $A'x = \begin{bmatrix} .664 \\ .589 \end{bmatrix}$ of x in $\mathbb{R}^2$. Figure A.4 also shows the images A'I of the elementary vectors $\bar{e}_1$ $\bar{e}_2$ $\bar{e}_3$ of $\mathbb{R}^3$. Their coordinates are given in the rows of A.

In figure A.4 it remains true that $A'x = .554A'\bar{e}_1+.222A'\bar{e}_2+.802A'\bar{e}_3$. This construction is shown in the figure, and results in a flat image of the three-dimensional rectangular parallelopiped of figure A.3

Figures A.1 and A.4 can be superimposed; they graph the same $\mathbb{R}^2$. If we do that, it is seen that A'x in figure A.4 does not have the same direction as y in figure A.1. But A'x has the same direction as A'Ay = (.711)A'x. It follows that where Ay is the image of y (going from $\mathbb{R}^2$ to $\mathbb{R}^3$), and A'Ay the image of Ay (going backwards from $\mathbb{R}^3$ to $\mathbb{R}^2$), y and A'Ay have not the same direction

A2 Hyperellipsoids

A2.1 Continuing the example of section A1.3, let y be an arbitrary vector in $\mathbb{R}^2$ with unit length (y'y=1). This vector describes in $\mathbb{R}^2$ a circle with unit radius. This circle is shown in figure A.5 (this figure could be superimposed upon figures A.1 or A.4).

The image of this unit circle is described in $\mathbb{R}^3$ by the vector Ay. It describes an ellipse, shown in figure A.6. Note that the vectors a_1 and a_2 are "spokes" (or "pseudo radii") of this ellipse: in figure A.5 e_1 and e_2 are radii of the unit circle, and in figure A.6 their images a_1 and a_2 are pseudo radii of the ellipse. The ellipse in figure A.6 is located in the plane spanned by a_1 and a_2 (a two-dimensional subspace of $\mathbb{R}^3$).

The proof that Ay describes an ellipse, would require that it be shown that the procedure outlined above is equivalent to the definition of an elleipse in terms of a construction rule. We omit that proof.

Conversely, in $\mathbb{R}^3$ the vector x, with x'x =1, describes a sphere with

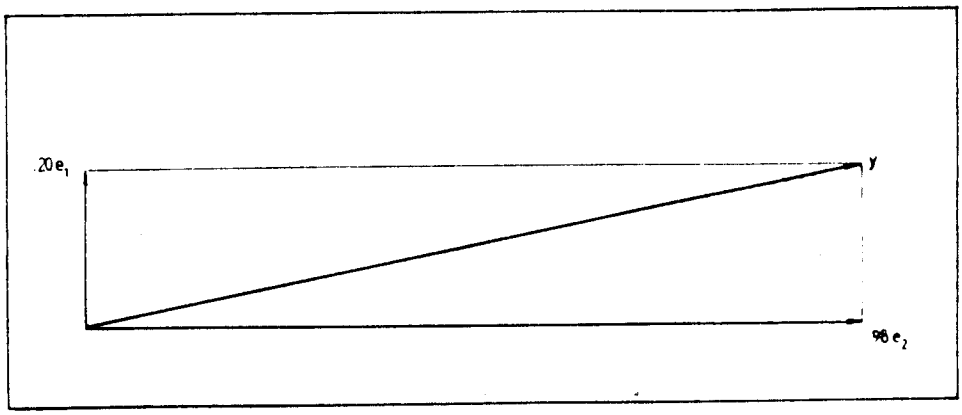


Fig. A1

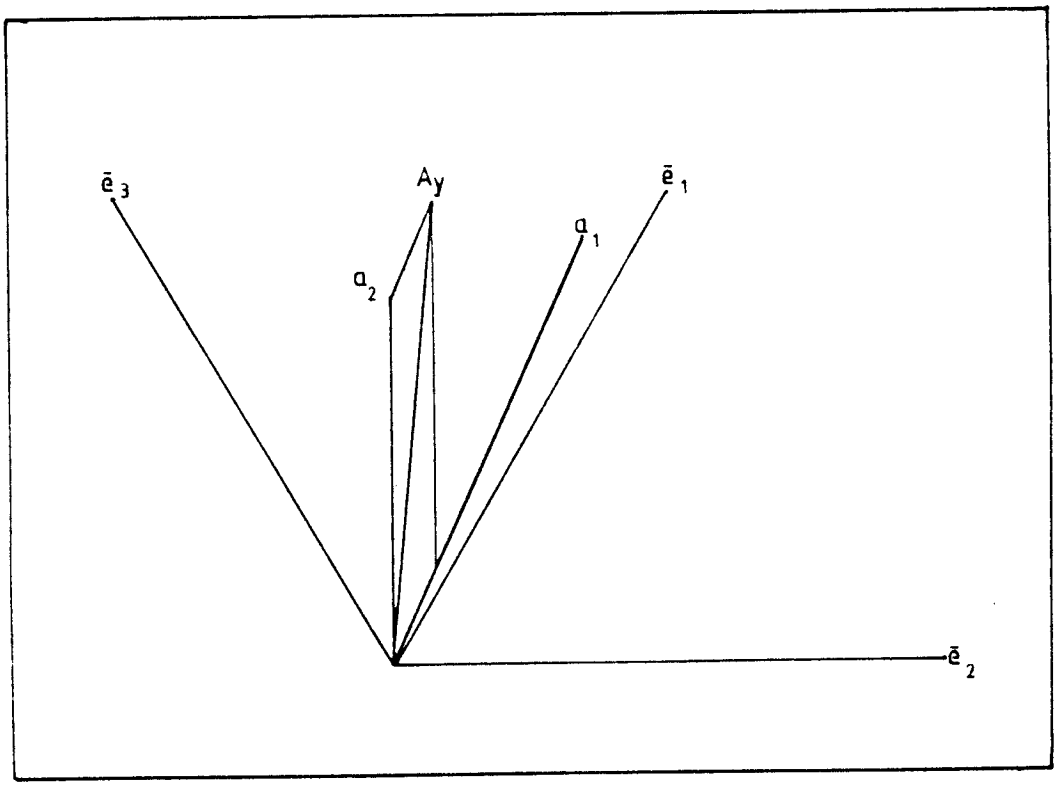


Fig. A2

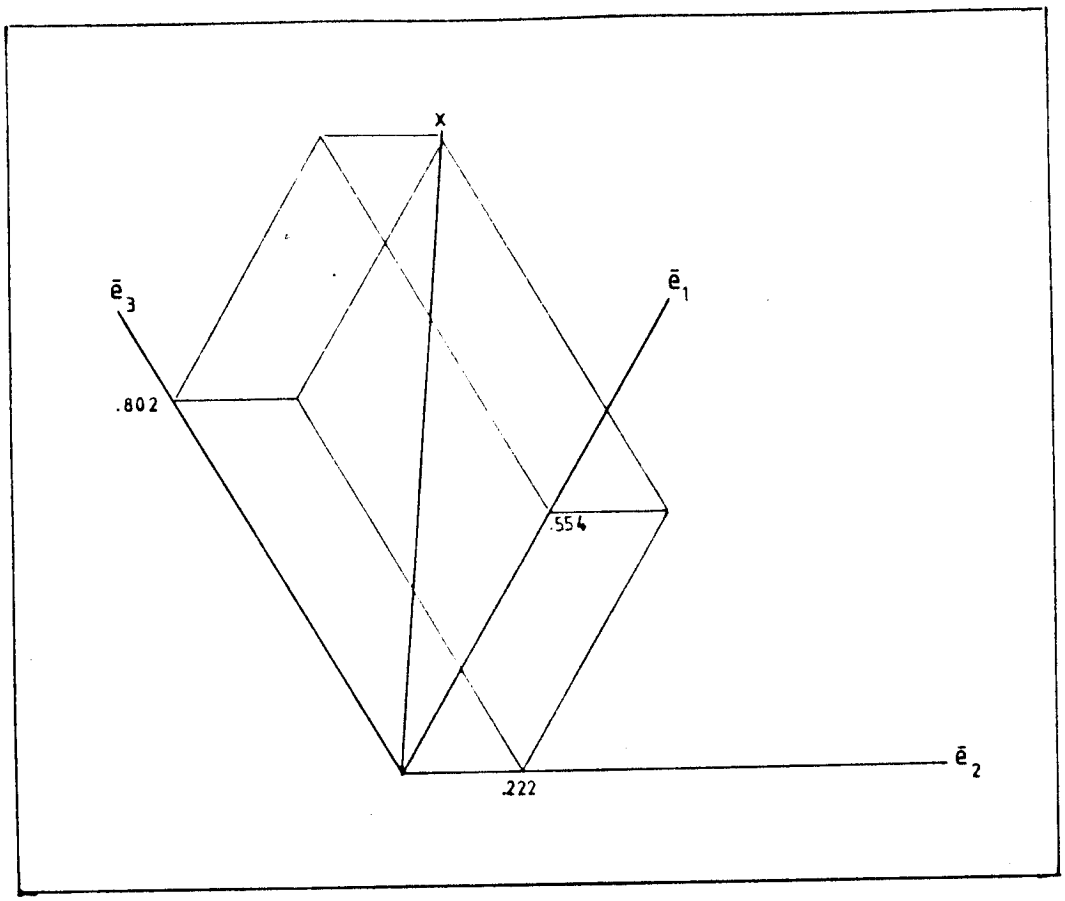
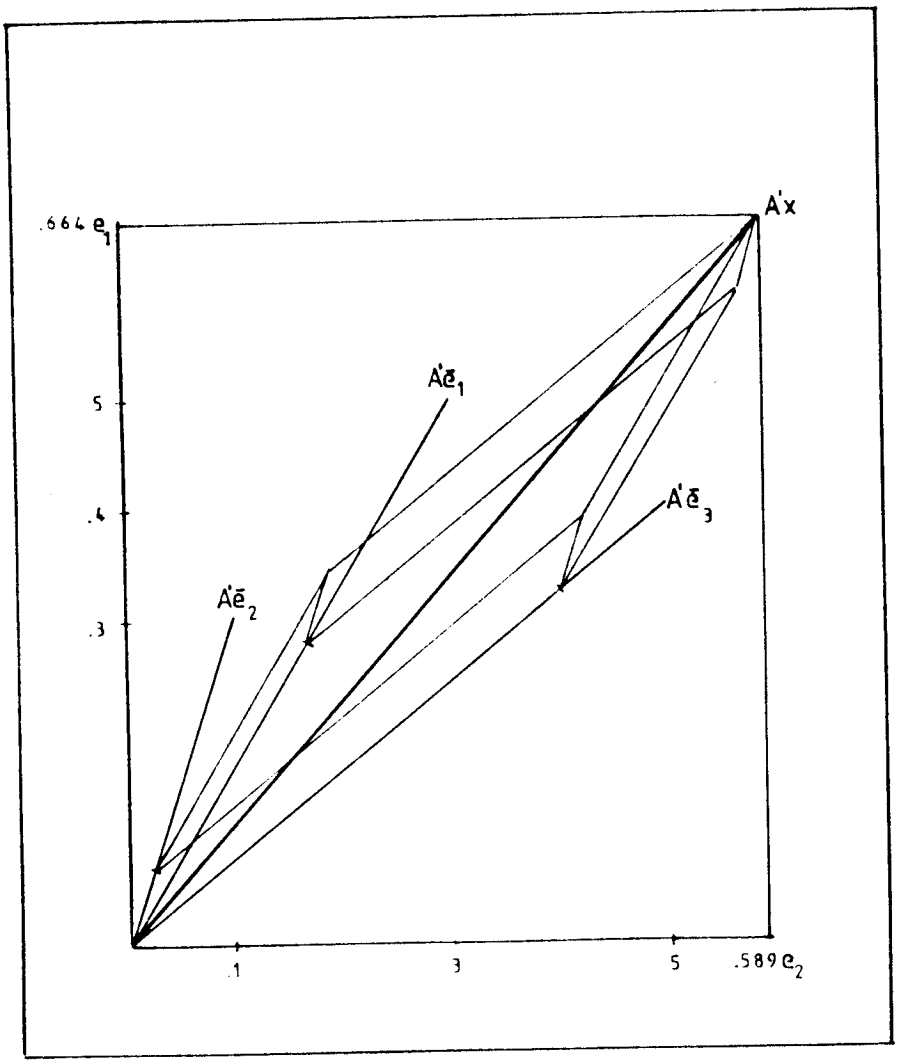

x
$\bar{e}_3$
$\bar{e}_1$
.802
.554
$\bar{e}_2$
.222

Fig. A3


A'x
.664 e_1
A'$\bar{e}_1$
5
.4
A'$\bar{e}_3$
A'$\bar{e}_2$
.3
.1
3
5
.589 e_2

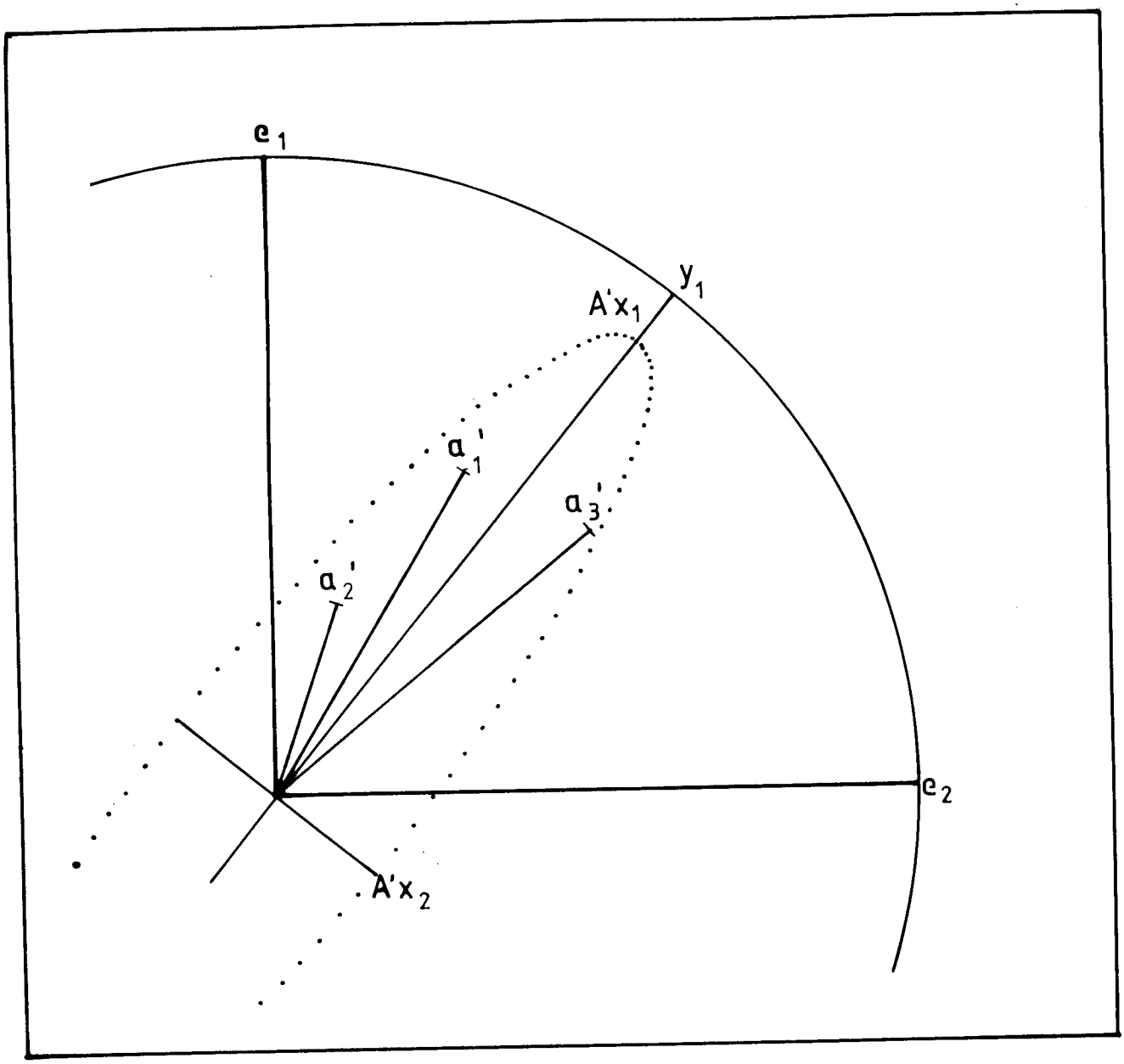


Fig. A5

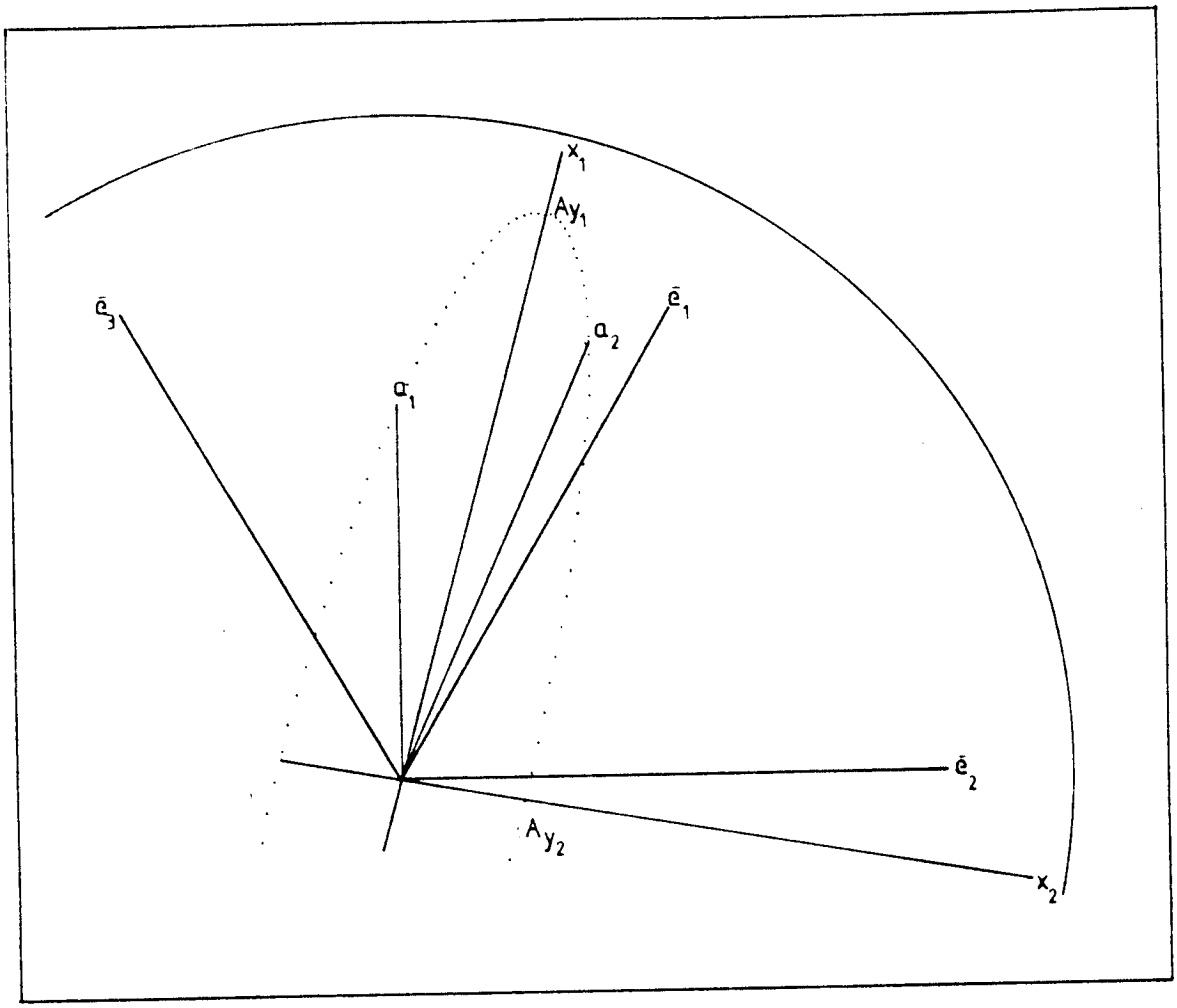


Fig. Ao

unit radius. In figure A.6 the outer contour of this sphere is drawn as a circle. The vectors $\bar{e}_1$, $\bar{e}_2$, $\bar{e}_3$ are located on the surface of the sphere, but appear in the figure as vectors interior to the outer contour (in the same way as cities, located on the surface of the earth, appear in a two-dimensional world map as points interior to a circle). The image in $\mathbb{R}^2$ of the sphere is described by A'x. It is shown in Figure A.5 as the dotted ellipse, with the vectors a_1', a_2', and a_3' (the images of $\bar{e}_1,\bar{e}_2,\bar{e}_3$) interior to the ellipse. In fact, the ellipse is a flat map of the unit sphere of $\mathbb{R}^3$.

A2.2 Generalizing to the case where A is an n x m matrix, with x a vector of unit length in $\mathbb{R}^n$, and y a vector of unit length in $\mathbb{R}^m$, and assuming that A has full column rank m, with m<n, we will have the following results. The vector x describes in $\mathbb{R}^n$ a 'hypersphere' with unit radius. Its image in $\mathbb{R}^m$ becomes a 'hyperellipsoid' with dimensionality reduced to m. As a consequence, vectors on the surface of the hypersphere in $\mathbb{R}^n$, may be mapped as interior vectors of the hyperellipsoid in $\mathbb{R}^m$. Conversely, y describes a hypersphere in $\mathbb{R}^m$. The image Ay of y describes in $\mathbb{R}^n$ a hyperellipsoid, confined to an m-dimensional subspace of $\mathbb{R}^n$, and with Ay located on the surface of the hyperellipsoid. Suppose now that A has column rank p<m. Its column vectors then span a p-dimensional subspace of $\mathbb{R}^m$. It follows that A'x will describe a hyperellipsoid also confined to the p-dimensional subspace of $\mathbb{R}^m$. Consersely, the hypersphere described by y in $\mathbb{R}^m$, becomes mapped as a p-dimensional hyperellipsoid, described by Ay in $\mathbb{R}^n$.

A3 Invariant directions

A3.1 In the example of section A1.3 it was shown that, if we take x proportional to the image Ay of y, then the image A'x of x (which must be proportional to A'Ay) is not necessarily proportional to y itself. However, there are special solutions x_i and y_i for which proportionality is retained, with identical proportionality coefficient ψ_i. This implies the equations

$$Ay_i = x_i\psi_i \qquad \text{(A3.1A)}$$

$$A'x_i = y_i\psi_i \qquad \text{(A3.1B)}$$

so that

$$A'Ay_i = y_i\psi_i^2 \qquad \text{(A3.2A)}$$

$$AA'x_i = x_i\psi_i^2 \qquad \text{(A3.2B)}$$

Such vectors x_i and y_i are said to have invariant directions. It can be shown that they correspond to the principal axes of the

hyperellipsoids described by Ay and A'x.

(Why are the proportionality constants in equations (A3.1) identical? Suppose they were not, so that

$Ay_i = x_i\lambda_i$

$A'x_i = y_i\gamma_i$

It then follows that

$y_i'A'Ay_i = y_i'A'x_i\lambda_i = y_i'y_i\gamma_i\lambda_i = \gamma_i\lambda_i$

but also

$y_i'A'Ay_i = x_i'x_i\lambda_i^2 = \lambda_i^2$

so that $\gamma_i\lambda_i=\lambda_i^2$. By a similar argument $\lambda_i\gamma_i=\gamma_i^2$. It follows that $\gamma_i=\lambda_i$.)

A3.2 For the example of section A1.3 and A2.1 (figures A.5 and A.6) let $y_1 = \begin{bmatrix}.773\\.635\end{bmatrix}$. Ay_1 is shown as a spoke of the ellipse in figure A.6, with $Ay_1 = \begin{bmatrix}.577\\.296\\.626\end{bmatrix}$. Define x_1 as Ay_1 extended to unit length (in figure A.6 x_1 remains interior to the outer contour of the unit sphere). Since $y_1'A'Ay_1 = .812$, we have

$x_1 = Ay_1/(.901) = \begin{bmatrix}.640\\.328\\.695\end{bmatrix}$. Its image is $A'x_1 = \begin{bmatrix}.696\\.572\end{bmatrix}$, and turns out to be proportional to y_1, with $A'x_1 = (.901)y_1$. We therefore have a solution for invariant directions, with $\psi_1=.901$. Figure A.5 illustrates than $A'x_1$ is the longest principal axis of the ellipse.

Conversely, in figure A.6, Ay_1 is the longest principal axis of the ellipse described by Ay (but this ellipse is in the plane of a_1 and a_2, which is not the plane of drawing, so that the principal axis of the "real" ellipse are not those of the drawn ellipse). Figure A.6 also shows the short principal axis, for

$y_2 = \begin{bmatrix}.635\\-.773\end{bmatrix}$, with $A'x_2 = \begin{bmatrix}.123\\-.150\end{bmatrix}$, and $\psi_2 = .194$.

A3.3 Note that in figure A.4 the image A'x is much closer to the first principal axis than y in figure A.1 (this could be seen by superimposing the figures A.1, A.4, and A.5). It can be shown that this result is generally valid. This at once suggests a computational algorithm for identifying the solution y_1: start with arbitrary y, find its image Ay, define x as Ay extended to unit length, find the image A'x, and redefine y as A'x extended to unit length. Repeat this cyclus, and y will become as close to y_1 as one wants. Such algorithms are described

in detail in section 3.3. They are called "alternating" algorithms because the basic cyclus has a step where y in $\mathbb{R}^m$ is mapped in $\mathbb{R}^n$, alternating with a step where x in $\mathbb{R}^n$ is mapped "backwards" in $\mathbb{R}^m$.

A4 Singular vectors and singular values

A4.1 Vectors y_i and x_i that satisfy equations (A3.1) are called singular vectors of A . They can be collected in matrices Y and X that must satisfy

$$AY = X\Psi \qquad \text{(A4.1A)}$$

$$A'X = Y\Psi \qquad \text{(A4.1B)}$$

with $X'X = I$, $Y'Y = I$, and with Ψ a diagonal matrix with non-negative diagonal elements ψ_i, called singular values of A. The equations above imply

$$X'AY = \Psi \qquad \text{(A4.2)}$$

where Ψ is called the 'canonical form' of A. The equations (A4.1) also imply

$$A = X\Psi Y' \qquad \text{(A4.3)}$$

which is called the 'singular value decomposition' (SVD) of A.

A4.2 Suppose the n x m matrix A ($n>m$) has rank k ($k<m$). The latter implies that there are m-k independent solutions for y_i so that $Ay_i=0$. They satisfy eq.(A3.1A) with $\psi_i=0$. At the same time, there will be n-k independent solutions $A'x_i=0$; they satisfy eq.(A3.1B).
On the other hand, equation (A4.3) can be written as

$$A = \sum_i x_i \psi_i y_i' \qquad \text{(A4.4)}$$

where x_i is a column of X and y_i a column of Y. It follows that singular vectors with corresponding $\psi_i=0$ can as well be omitted in eq.(A4.4), since their contribution is multiplied by the zero singular value. This in turn implies that for an n x m matrix of rank k, with $n \geq m$, and $m \geq k$, the SVD solution $A = X\Psi Y'$ can be re-defined with
X an n x k matrix, $X'X = I$
Y an m x k matrix, $Y'Y = I$
Ψ a k x k diagonal matrix with positive diagonal elements $\psi_i>0$.

A5 Eigenvectors and eigenvalues

A5.1 The equations (A4.1) imply

$$A'AY = A'X\Psi = Y\Psi^2 \qquad \text{(A5.1A)}$$

$$AA'X = AX\Psi = X\Psi^2 \qquad \text{(A5.1B)}$$

Column vectors of Y are called eigenvectors of the (square and symmetric) matrix A'A, with the diagonal elements of Ψ^2 as their corresponding eigenvalues. As with singular vectors, we maintain the normalization agreement $Y'Y = I$ (similarly for X).

When there are multiple equal eigenvalues, the eigenvector solution no longer is uniquely defined - but even then it remains feasible to take a solution that satisfies $Y'Y = I$

A5.2 The algorithm suggested in A3.3 remains applicable for eigenvectors. Let y be some arbitrary vector. Its image with respect to A'A is A'Ay. Extend A'Ay to unit length, call this vector x. The image of x becomes $(A'A)'x = A'Ax$. But this shows that the "alternating" steps become in fact undistinguishable.
Obviously, once Y and Ψ^2 are identified for A'A on the basis of an eigenvector/eigenvalue algorithm, X can be calculated afterwards from $X = AY\Psi^{-1}$, assuming that in Ψ only the p non-zero eigenvalues are retained (and in Y the p corresponding eigenvectors).

A5.3 Given that A is a real-valued matrix, its singular values ψ_i also must be real-valued, and it follows that the eigenvalues ψ_i^2 of A'A never can be negative. A'A then is said to be a positive semi-definite (or Gramian) matrix. It may happen that a square and symmetric matrix B has negative eigenvalues. An example is

$$By = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -2 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} (-2)$$

It follows that in such cases B never can be the matrix of squares and cross products of a real valued matrix A, with $A'A = B$.
A matrix B with both positive and negative eigenvalues is called indefinite. Indefinite matrices may occur in applied data analysis (e.g., a correlation matrix based on a data matrix with many missing values could become indefinite).

A6 Algebraic applications of eigenvectors and singular vectors

A6.1 Let B (square and symmetric) have eigenvectors Y and eigenvalues Ψ^2. For any c, the powered matrix B^c will be equal to

$$B^c = Y\Psi^{2c}Y'.$$

Examples are

(i) $B^2 = Y\Psi^2Y'Y\Psi^2Y' = Y\Psi^4Y'$.

(ii) $B^{-1} = Y\Psi^{-2}Y'$, since $Y\Psi^2Y'Y\Psi^{-2}Y' = YY' = I$ (if the inverse exists, B has no zero eigenvalues, and Y is a square orthonormal matrix).

(iii) $B^{\frac{1}{2}} = Y\Psi Y'$, since $Y\Psi Y'Y\Psi Y' = Y\Psi^2Y' = B$

(iv) $B^{-\frac{1}{2}} = Y\Psi^{-1}Y'$ (defined if B has an inverse, so that $YY'=I$).

A6.2 Let X be any n x m matrix ($n \geq m$) of rank $k \leq m$. Its generalized inverse is defined as a matrix X^- that satisfies

$$XX^-X = X$$

When X is square and of full rank (n=m=k), X^- becomes identical to the proper inverse X^{-1}. In all other cases, X^- is not uniquely defined. However, there is one solution for X^-, for which the notation X^+ is used, which is uniquely defined, and which also satisfies

$X^+XX^+ = X^+$

$XX^+ = (XX^+)'$

$X^+X = (X^+X)'$

X^+ is often called the Moore-Penrose generalised inverse. Given the SVD solution $X = P\Lambda Q'$, a solution for X^+ becomes $X^+ = Q\Lambda^{-1}P'$. That this solution satisfies the four requirements above, is easily verified.

The advantage of X^+ is that many algebraic expressions in which a proper inverse occurs, remain valid for cases with deficient rank, provided that the proper inverse is replaced by the generalized inverse. Example: for linear regression of y on Z, the regression vector is Zb, with $b=(Z'Z)^{-1}Z'y$. Suppose Z has deficient column rank, so that $(Z'Z)^{-1}$ does not exist. A valid solution remains $b = (Z'Z)^+Z'y$ (which in this particular example simplifies further to $b = Z^+y$).

A7 Optimization properties of SVD

A7.1 Section A3 demonstrated that under the condition y'y=1 the image Ay describes a hyperellipsoid, and that invariant directions correspond to the principal axes of such a hyperellipsoid. It follows immediately that the principal axes are related to stationary values for the sum of squares y'A'Ay. This sum of squares is absolutely maximized by taking the solution y_1 so that

$y_1'A'Ay_1 = y_1'y_1\psi_1^2 = \psi_1^2$

An unconditional minimum of y'A'Ay is represented by the shortest principal axis of the hyperellipsoid, and corresponds to

$y_m'A'Ay_m = y_m'y_m\psi_m^2 = \psi_m^2$

Intermediate solutions y_j are conditional maxima in the sense that $y_j'A'Ay_j$ is a maximum under the condition $y'y_i=0$ (i=1,..,j-1), and is a minimum under the condition $y'y_k=0$ (k=j+1,...,m).

A7.2 Section A4.2 showed that we can write

$A = \Sigma\, x_i\psi_i y_i'$

It can be proved that

$$A_{(k)} = \sum_{i=1}^{k} x_i\psi_i y_i'$$

is the best least squares rank k approximation to A, with loss function

equal to

$$\text{trace}(A'A - A_{(k)}'A_{(k)}) = \sum_{i=1}^{m} \psi_i^2 - \sum_{i=1}^{k} \psi_i^2 = \sum_{i=k+1}^{m} \psi_i^2$$

This theorem is usually called the Eckart-Young theorem (Eckart and Young, 1937), although there are earlier formulations (e.g., Schmidt, 1906).

A8 Generalized eigenvector equation

Crucial in linear MVA is the solution of the "generalized eigenvector equation"

$$Cx = Dx\phi/m \qquad \text{(A8.1)}$$

where C and D are square and symmetric matrices of dimension m x m. In section A9 examples will be given. For the present section we take the general and formal view that C and D are arbitrary square and symmetric matrices.

The generalized eigenvector equation can be reduced to a "classical" eigenvalue equation by use of the following "trick". Define $y=D^{\frac{1}{2}}x$, so that $x=D^{-\frac{1}{2}}y$. For the definition of $D^{\frac{1}{2}}$ see section A6.1; if D has no proper inverse, define $D^{-\frac{1}{2}}$ as the generalised inverse $(D^{\frac{1}{2}})^{+}$. Equation (A8.1) now can be re-written as

$$CD^{-\frac{1}{2}}y = D^{\frac{1}{2}}y\phi/m$$

or

$$D^{-\frac{1}{2}}CD^{-\frac{1}{2}}y = y\phi/m \qquad \text{(A8.2)}$$

which is a classical eigenvector equation. It can be solved for y_i and ϕ_i, after which x_i is identified from $x_i=D^{-\frac{1}{2}}y_i$. Since for the eigenvector y_i normalization $y_i'y_i=1$ has been agreed upon, the resulting normalization for x_i becomes $x_i'Dx_i=1$.

As a consequence, the solution gives stationary values for the ratio

$$\frac{x'Cx/m}{x'Dx} = \phi \quad .$$

A9 Applications in linear MVA

A9.1 All essential problems of MVA as a technique of data analysis can be brought to one and the same general format: that of the generalized eigenvector equation (A8.1). We shall illustrate this for a number of common MVA applications.

A9.2 Principal components analysis starts from an n x m datamatrix H, assumed to be in deviations from column means. In PCA the first step is to give columns of H equal normalization, say unity. Let $C = H'H$, and define D as the diagonal matrix of C. Then $HD^{-\frac{1}{2}}$ solves the first step. The matrix of sums of squares and cross-products of $HD^{-\frac{1}{2}}$ becomes

$$R = D^{-\frac{1}{2}}H'HD^{-\frac{1}{2}}$$

R is the correlation matrix. PCA is usually described in terms of the eigenvector solution of R:

$RY = Y\Psi^2$

with resulting "factor matrix"

$F = Y\Psi$

so that $R = FF'$.

But PCA could as well be formulated in terms of the SVD solution of $HD^{-\frac{1}{2}} = P\Psi Y' = PF'$, where P gives what often is called the matrix of "individual factor scores". Since $P'P=I$, it follows that $D^{-\frac{1}{2}}H'P=Y\Psi=F$ can be interpreted as a matrix of correlations between observed variables and factorscores.

Obviously, PCA fits in the format of section A8. The eigenvector equation $RY = Y\Psi^2$ is equivalent to $D^{-\frac{1}{2}}CD^{-\frac{1}{2}}Y = Y\Psi^2$. Equate $X=D^{\frac{1}{2}}Y$, and $\Psi^2=\Phi/m$, and PCA becomes quivalent to solving (A8.1).

A9.3 In canonical correlation analysis (CCA) we have the following problem. Given are two data matrices H_1 of dimension n x m_1, and H_2 of dimension n x m_2, both in deviations from column means. We want to find a solution for x_1 and x_2 in such a way that the correlation between H_1x_1 and H_2x_2 is stationary.

The problem becomes much simplified by selecting normalization conditions $x_1'H_1'H_1x_1 = 1$, and $x_2'H_2'H_2x_2 = 1$. The expression for the correlation between H_1x_1 and H_2x_2 then becomes

$$r = x_1'H_1'H_2x_2$$

Still further simplification is obtained by defining $y_1=(H_1'H_1)^{\frac{1}{2}}x_1$ and $y_2=(H_2'H_2)^{\frac{1}{2}}x_2$. This implies normalization $y_1'y_1=1$, and $y_2'y_2=1$. The expression for the correlation becomes

$$r = y_1'\{(H_1'H_1)^{-\frac{1}{2}}(H_1'H_2)(H_2'H_2)^{-\frac{1}{2}}\}y_2 = y_1'Ay_2$$

where A is introduced as a symbol for the expression between brackets. It then becomes immediately clear that y_1 and y_2 should be identified with the singular vectors of $A = Y_1\Lambda Y_2'$, since in that case the correlation becomes equal to a singular value λ_i. Maximum correlation is obtained by taking the first singular vectors y_{11} and y_{12}, with correlation λ_1.

In order to bring the CCA problem in the format of section A8, we write

$H = (H_1,H_2)$

so that H becomes the combined n x m data matrix. Define

$$C = H'H = \begin{bmatrix} H_1'H_1 & H_1'H_2 \\ H_2'H_1 & H_2'H_2 \end{bmatrix}$$

and let D be the superdiagonal matrix of C:

$$D = \begin{bmatrix} H_1'H_1 & \\ & H_2'H_2 \end{bmatrix}$$

so that

$$D^{-\frac{1}{2}} = \begin{bmatrix} (H_1'H_1)^{-\frac{1}{2}} & \\ & (H_2'H_2)^{-\frac{1}{2}} \end{bmatrix}$$

Then

$$D^{-\frac{1}{2}}CD^{-\frac{1}{2}} = \begin{bmatrix} I & (H_1'H_1)^{-\frac{1}{2}}(H_1'H_2)(H_2'H_2)^{-\frac{1}{2}} \\ (H_2'H_2)^{-\frac{1}{2}}(H_2'H_1)(H_1'H_1)^{-\frac{1}{2}} & I \end{bmatrix}$$

$$= \begin{bmatrix} I & A \\ A' & I \end{bmatrix} = \begin{bmatrix} I & Y_1 \Lambda Y_2' \\ Y_2 \Lambda Y_1' & I \end{bmatrix}$$

It then can be immediately verified that an eigenvector solution for $D^{-\frac{1}{2}}CD^{-\frac{1}{2}}$ is

$$D^{-\frac{1}{2}}CD^{-\frac{1}{2}} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} (I + \Lambda)$$

which is in the format of equation (A8.2). An equivalent expression is

$$C \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = D \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} (1 + \Lambda)$$

which is the format of equation (A8.1), with

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = D^{-\frac{1}{2}} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

(A small technical detail is that the SVD solution of A implies $Y_1'Y_1=I$, and $Y_2'Y_2=I$. It follows that in the solution above for the eigenvectors of $D^{-\frac{1}{2}}CD^{-\frac{1}{2}}$ the eigenvectors are normalized at 2.I.)

This exposition of CCA implicitly covers a number of other standard MVA techniques, such as canonical discriminant analysis (CADA). or multivariate analysis of variance (MANOVA). In CADA H_1 will be a "dummy" matrix that gives a coding for subgroups of individuals, whereas H_2 contains observations on "dependent variables". In MANOVA, H_1 contains the same sort of coding, now for the subgroups that correspond to the conditions of the systematic experimental design.

A9.4 Multiple regression is a special case of CCA, with $m_2=1$, so that H_2 becomes a single vector of observations h_2. For convenience, assume that all columns of $H = (H_1, h_2)$ are normalized to unity, so that $H'H$ is the over all correlation matrix:

$$C = H'H = \begin{bmatrix} R & r \\ r' & 1 \end{bmatrix}$$

Define

$$D = \begin{bmatrix} R & 0 \\ 0 & 1 \end{bmatrix}$$

so that

$$D^{-\frac{1}{2}}CD^{-\frac{1}{2}} = \begin{bmatrix} I & R^{-\frac{1}{2}}r \\ r'R^{-\frac{1}{2}} & 1 \end{bmatrix}$$

It is easily verified that

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} R^{-\frac{1}{2}}r \\ \rho \end{bmatrix}$$

is an eigenvector of $D^{-\frac{1}{2}}CD^{-\frac{1}{2}}$, where ρ is defined as $(r'R^{-1}r)^{\frac{1}{2}} = \rho$, so that ρ is the multiple correlation between H_1 and h_2. It then also follows that we can write

$$C \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = D \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} (1 + \rho)$$

with

$$C\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = D^{-\frac{1}{2}}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} R^{-1}r \\ \rho \end{bmatrix} = \begin{bmatrix} b \\ \rho \end{bmatrix}$$

where b is the vector of regression weights. Obviously, $D^{-\frac{1}{2}}CD^{-\frac{1}{2}}$ has more than one solution for its eigenvectors. These other solutions are unrelated to the solution for multiple regression.

Again, it should be realised that the solution for multiple regression implicitly covers discriminant analysis with two subgroups of individuals (and where h_2 contains a binary code that separates between these two groups). It also covers ANOVA (where H_1 contains a code for the subgroups prescribed by the experimental design, and where h_2 stands for the single "dependent variable").

A9.5 A generalized canonical solution (GENCAN) is indicated where the data matrix H is partitioned into K subsets

$H = (H_1 \quad H_2 \quad \ldots \quad H_K)$

where again we assume that columns of H are in deviations from means. Let $C = H'H$, and define D as the superdiagonal matrix of C with diagonal submatrices $D_{ii} = C_{ii}$, and off-diagonal submatrices $D_{ij} = 0$. It follows that $D^{-\frac{1}{2}}CD^{-\frac{1}{2}}$ will have diagonal submatrices equal to identity matrices.

The GENCAN solution then becomes the solution for y that satisfies
$D^{-\frac{1}{2}}CD^{-\frac{1}{2}}y = y\psi^2$
or
$Cx = Dx\psi^2$
with $x = D^{-\frac{1}{2}}y$. All MVA problems discussed earlier then are special cases. The PCA problem has K=m (with the implication that each partition H_j of H corresponds to a single vector h_j). The CCA problem has K=2. The multiple correlation problem has K=2, with in addition $m_2=1$.
GENCAN is closely related to HOMALS. One might say that HOMALS comes to the same as GENCAN applied to the indicator matrix G.

A9.6 In all examples above, the linear MVA problem can be phrased in terms of the SVD solution of $HD^{-\frac{1}{2}}$. The MVA problem obtains its special characteristics only from the definition of D. In PCA, D is the diagonal matrix of H'H. In CCA, D is superdiagonal matrix corresponding to the two partitions of H. In GENCAN D is a superdiagonal matrix corresponding to the K partitions of H.

A10 Joint maps

A10.1 A joint plot of rows and columns of a matrix H with SVD solution $H = X\Psi Y'$ is obtained by plotting the rows of H as points with coordinates given in the rows of $X\Psi^{\alpha}$, and the columns of H as points with coordinates given in the rows of $Y\Psi^{1-\alpha}$. Such a plot is a representation of rows and columns of H in the following way. Draw a vector through the j^{th} row point, and project the m column point on this vector. The vector of projections will be proportional to the j^{th} row of H. And, conversely: draw a vector throuhh the i^{th} column point, and project the n row points on this vector. This vector of projections will be proportional to the i^{th} column of H.
The property mentioned above is invariant under choice of α. Obvious choices for α are $\alpha=0$, $\alpha=\frac{1}{2}$, and $\alpha=1$. The choice $\alpha=\frac{1}{2}$ often results in the best graph from the aesthetic point of view. The two sets of points then will have equal spread in each dimension of the plot.
Usually, such a plot will be made for the first p dimensions of the SVD solution, with, in most cases, p=2. The properties above then are exactly true for the rank 2 approximation $H_{(2)}=x_1\psi_1y_1'=x_2\psi_2y_2'$, and will be true for rows and columns of H itself only to the extent that $H_{(2)}$ is a good approximation of H.

A10.2 The principle of a joint plot for rows and columns of a frequency matrix F is discussed in section 4.3. The same principle applies to indicator matrices G; they are, after all, also a kind of frequency

matrix. The principle is the same as in section A10.1 - however, the joint plot does not refer to rows and columns of F itself, but to a "corrected" matrix $D_r^{-1}FD_c^{-1}.n - uu'$ with elements $(f_{ij}-e_{ij})/e_{ij}$ (where e_{ij} is the 'expected frequency' on the basis of row and column marginals). The representation, in other words, focuses not on f_{ij} itself but on "deviations from what one might expect on the assumption of independence", with, in addition, a weighting inversely proportional to such expectation.

A11 Join and meet solutions

A11.1 In the general linear MVA problem (GENCAN, section A9.5) we have K sets of variables. The "best" solution is derived from the first "generalized eigenvector" y_1, satisfying

$Cy_1 = Dy_1\psi_1^2$

y_1 can be partitioned into K subvectors y_{1j} (j=1,..,K), so that for each individual set of variables H_j we have a solution $H_jy_{1j} = q_{1j}$. The solution is "good" to the extent that the vectors q_{1j} are close together. In fact, the best GENCAN solution has the property that the sum of the squared projections of $q_{1j}/(q_{1j}'q_{1j})^{\frac{1}{2}}$ on $Hy_1 = \Sigma q_{1j}$ is maximized. In non-geometrical terms: GENCAN maximizes the sum of the squared correlations between q_{1j} and Σq_{1j}.

A11.2 A perfect solution would be obtained if all q_{1j} did coincide with Hy_1. This implies that the spaces spanned by the individual H_j have a common intersection (in the case of K=2 this would imply a canonical correlation of 1). We shall say that in this case the spaces spanned by the H_j meet in Hy_1, and that the common intersection of the spaces H_j is the "perfect meet-solution".

Usually, however, there will be no common intersection. Then the best approximate meet solution will require that the q_{1j} have as much as possible small angles with Hy_1. The vectors q_{1j} form a "bundle" around their sumvector Hy_1; the solution is better to the extent that this bundle is narrower.

A11.3 A join problem is defined as follows. Let H have columns h_j (j=1,..,m). In general H spans an m-dimensional space. A perfect join-p solution would be that the vectors h_j span a p-dimensional space. This would require that H has rank p.

In general, H will not have rank p<m. We then have an approximate join-p solution that is "good" to the extent that a rank p approximation of H is closer to H itself. PCA solves this problem. In particular, let H have SVD solution $H = P\Lambda Q'$. Then the best rank p solution

of the join problem is $H_{(p)} = \sum_p p_i \lambda_i q_i'$ (cf. section A7.2). Usually, however, the solution will be formulated not in terms of this matrix $H_{(p)}$, but in terms of a basis of the join-space (such a basis is formed by the vectors p_i), or in terms of projections of the h_j on this basis (such projections are given in the factor matrix $Q\Lambda$, provided that columns of H were normalized to unity).
One could say that the join-problem is only a special case of the meet-problem, namely where K=m, so that we have K sets of variables, with only one variable in each set.

A11.4 In GENCAN (section A9.5) meet and join problems become intertwined as follows. The first meet solution defines K vectors $q_{1j} = H_j y_{1j}$, and defines the meet-solution as such as the vector Hy_1, around which the vectors q_{1j} form a bundle in K dimensions. To analyze this bundle of vectors q_{1j} further, then is a join-problem that could be approached with PCA. The meet solution Hy_1 will be the first principal component in that PCA solution, successive components give additional information about the bundle. This approach is illustrated in section 3.8, where the first HOMALS solution is a GENCAN solution for the m sets G_j, and where the optimally scaled datamatrix Q_1 is further analyzed with PCA, with the first HOMALS solution as the first principal component.
However, in general, there will be m GENCAN solutions for K sets (if H has full rank, and where m is the number of columns of H). For each GENCAN solution Hy_g (g=1,..,m) the vectors $H_j y_{gj} = q_{gj}$ will form another bundle around their sumvector Hy_g in K dimensions. Each bundle could be further analysed with PCA, with the meet solution Hy_g as one (not necessarily the first) principal component. But then we end up with more dimensions (m.K) than we started with (m). There must be a sort of 'stopping rule' for the number p of "interesting" meet solutions (e.g., $p \leq (m:K)$).

Section 5.3 illustrates an approach where especially the "worst" meet solution becomes interesting. This meetsolution Hy_m will be the last principal components of the vectors $q_{mj} = H_j y_{mj}$, and since the corresponding last eigenvalue is minimized, it follows that the first m-1 eigenvalues for the PCA solution have maximum sum.

A11.5 As usual, K=2 (CCA of section A9.3) is a very special case. Here the best meetsolution Hy_1 is the sumvector of the two vectors $q_{11} = H_1 y_{11}$, and $q_{12} = H_2 y_{12}$. The two vectors q_{1i} are as close as possible to their sumvector, but then also as close as possible to each other, so that the solution minimizes the angle between q_{11} and q_{12}, and therefore maximizes the cosine of this angle which is the canonical correlation. In this

case the two vectors q_{11} and q_{12} form a "bundle" in a K=2 dimensional plane. Also, it can be shown that the solution implies that q_{11} and q_{12} have equal length. It follows that their PCA solution has the sum vector as the first component, but has the difference vector $q_{11}-q_{12}$ as the second component. On the other hand, this difference vector is the "worst" meet solution, since this worst solution, for q_{m1} and q_{m2}, should maximize the angle between q_{m1} and q_{m2}. This will happen when $q_{m1}=q_{11}$, and $q_{m2}=-q_{12}$. In this way, for K=2, meetsolutions always come in pairs: the best one is related to the worst one, the second best one to the second worst one, etc. In stead of ending up with m.K dimensions, there are only m.

A11.6 The other special case, as was already indicated in section A11.3, is K=m, which comes to the same as K=1: there is no distinction between K sets with only one variable in each set, or one set of m variables. In this case one might say that the first principal components is the best meet solution (the vectors h_j form the closest possible bundle around their first principal component). Further analysis of this bundle leads to the subsequent principal components. But these subsequent components are at the same time the subsequent meet solutions. In stead of ending up with m.K dimensions, there are only m.

Appendix B: Notation

B1 General remarks on notation

Throughout the text the notation for a vector is a lower case letter, such as x, v, h. Without a prime, a column vector is meant; row vectors are indicated as x', v', h'.

The notation for a matrix is a capital letter, such as X, V, H. A prime indicates that X' is the transpose of X. Sometimes a matrix is indexed, so that we have matrices G_1, G_2, ... etc. Let G_k be such an indexed matrix, then the element on its i^{th} row and j^{th} column will have notation g_{ij}^k.

B2 General and specific solutions

Often some equation (like $Cy\phi = Dy/m$) has many solutions for y; very often the notation y, without index, refers to any of the possible solutions, whereas y_i is specifically the i^{th} solution. Often an index s is used for a specific solution where it does not matter which one it is in the order of such solutions. Then y_s has the same meaning as y.

B3 Stochastic variates

Stochastic variables are indicated by underlining their label, such as $\underline{x}$, $\underline{v}$, $\underline{h}$.

B4 Special usage of symbols

The following list gives symbols which systematically have a specific meaning. Sometimes however, such symbols are defined differently for a short period.

A satifies the equation X = GA' (or satisfies the approximation of X by GA') for given G and X.

C G'G

D superdiagonal matrix of C.

D_r diagonal matrix of row totals of a frequency matrix F.

D_c diagonal matrix of column totals of a frequency matrix F.

F frequency matrix.

G indicator matrix.

G_j indicator matrix for the j^{th} variable.

H data matrix.

I identity matrix.

K left singular vectors for ANACOR, with K-K = I.

L right singular vectors for ANACOR, with L'L = I.

$M_* \quad \sum_{j=1}^{m} M_j$

M_j diagonal matrix of row totals of G_j.

Q optimally scaled data matrix, with $q_j = G_j Y_j$.

R correlation matrix

V left singular vectors of $M_*^{-\frac{1}{2}} G D^{-\frac{1}{2}}$, with V'V = I

W right singular vectors of $M_*^{-\frac{1}{2}} G D^{-\frac{1}{2}}$, with W'W = I

X object scores (but often X is used temporarily for other purposes).

Y category quantifications.

Y_j category quantification for categories of j^{th} variable.

Diagonal matrices

Ψ diagonal matrix of singular values in HOMALS.

Φ Ψ^2/m

Λ singular values in ANACOR

Indices

j individual variable (j=1,..,m).

m number of variables.

n number of objects.

p number of dimensions in the solution.

k_j number of categories of j^{th} variable.

s running index (s = 1,..,p) for dimension of solution.

Vectors

e_i i^{th} vector of an identity matrix I.

u vector with unit elements

Special symbols

SSQ( ) if X is a matrix or vector, then SSQ(X) is the sum of squares of the elements of X.

AVE( ) if X is a matrix, then AVE(X) is the vector of column averages of the elements of X. If $\underline{x}$ is a random vector, then AVE($\underline{x}$) is the expected value of $\underline{x}$.

VAR( ) if X is a matrix, then VAR(X) is the vector of column variances of the elements of X. If $\underline{x}$ is a random vector, then VAR($\underline{x}$) is the vector of variances.

GRAM if X is a matrix, then GRAM(X) is the orthogonal matrix obtained by applying Gram-Schmidt orthogonalization to X.

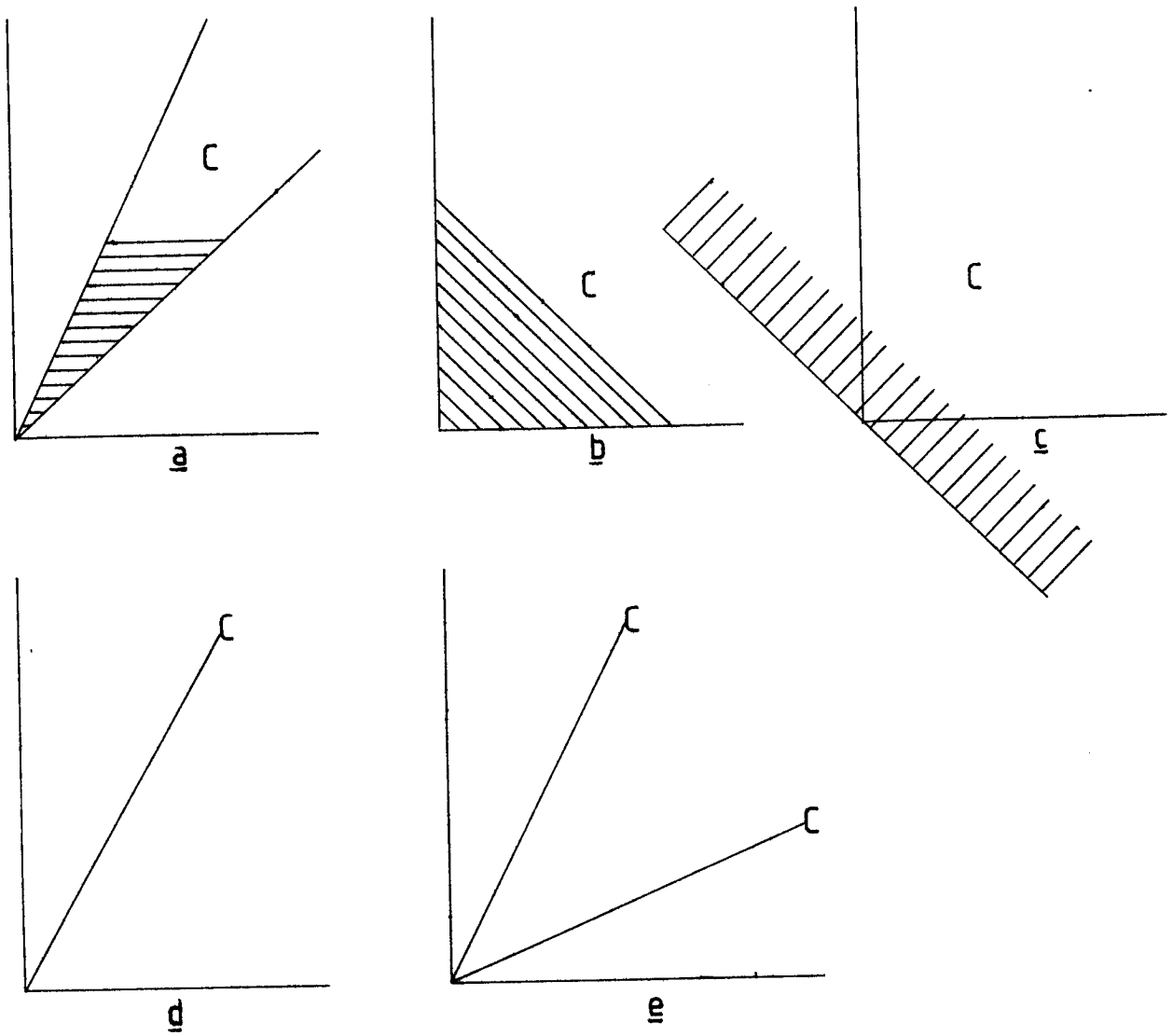


Figure C1: five cones

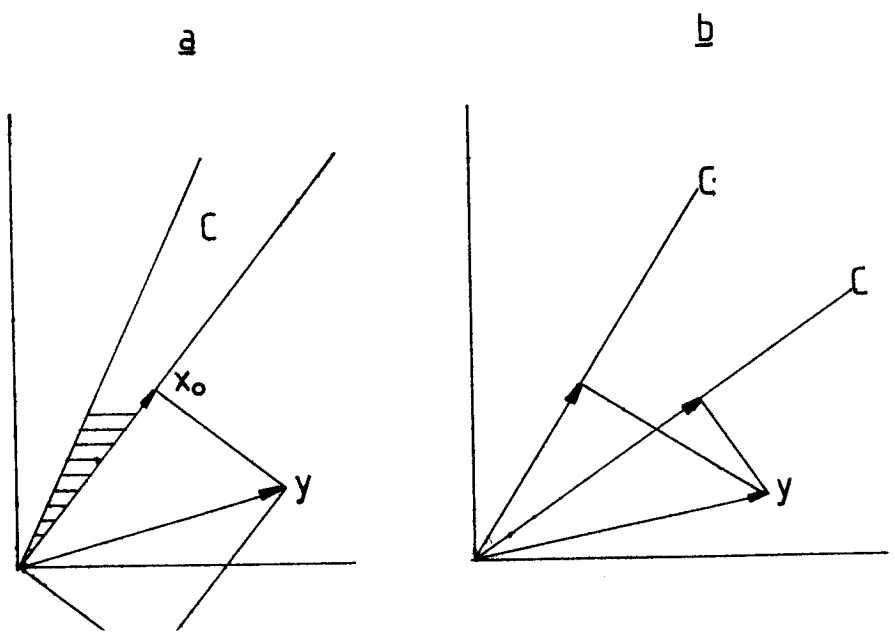


Figure C2:
projection on a cone

Appendix C: Cones and projection on cones

A set C in R^n is a cone if $x \in C$ implies that $\alpha x \in C$ for all $\alpha > 0$. In words: a set is a cone if it contains the ray through x whenever it contains x. In figure C1a-e we give some examples of cones in R^2. The cone in C1b is the positive orthant, the cone in C1c is a half-space, the cone in C1d is a one-dimensional subspace, and the cone in C1e consists of two subspaces which intersect only in the origin. The cones in C1a and C1b are pointed, by which we mean algebraically that $x \in C$ implies that $-x \notin C$. The cones in figure C1, except the one in C1e, are convex, by which we mean that $x \in C$ and $y \in C$ imply that $x + y \in C$. Many writers use the word cone in the sense of convex cone. Some convenient references on cones are Goldman and Tucker (1956) and Berman (1973). Rockafellar (1970) also contains all of the relevant material (and much more).

In our alternating least squares algorithms we often have to solve cone projection problems. Suppose C is a cone, y is a given vector in R^n, and W is a symmetric positive definite matrix. Then the projection of y on C in the metric W is that vector $x_0 \in C$ which satisfies

$$(y - x_0)'W(y - x_0) = \inf \{(y - x)'W(y - x) \mid x \in C\}.$$

We have not proved yet that such an x_0 actually exists. In fact it need not: if C is the interior of the positive orthant, W = I, then the infimum is not attained. In this case it is equal to $SSQ(y - x_0)$ with $x_0 = \max(0,y)$, but $x_0 \notin C$ if $y \notin C$. Consequently we assume that the cones we are dealing with are closed, which implies by the way that they contain the origin. It is possible to prove that the projection on a closed cone in finite-dimensional space always exists, although it may not be unique. If the cone is both closed and convex, then the projection is unique.

It is possible to derive some simple properties of the projection by straightforward calculation. If x_0 is the projection, then

$$(y - x_0)'W(y - x_0) \leq (y - \alpha x_0)'W(y - \alpha x_0)$$

for all $\alpha \geq 0$. If we simplify this, we find the result that $y'Wx_0 = x_0'Wx_0$, or $(y - x_0)'Wx_0 = 0$, or $y - x_0$ is W-perpendicular to x_0. This is illustrated, for W = I, in figure C2a and C2b for, respectively, a convex cone and a cone consisting of two rays. In this last nonconvex case there are two solutions to $(y - x_0)'Wx_0 = 0$ in fact if we have a y on the bisectrice of the acute angle formed by the two rays, then y is equally far from both rays, and the projection is not unique.

More precise statements are possible if the cone is convex. We then must have

$$((y - x_0) - z)'W((y - x_0) - z) \geq (y - x_0)'W(y - x_0)$$

for all z in C. If we simplify this we find that $(y - x_0)'Wz \leq 0$ for all z in C, which shows that $y - x_0$ makes a W-obtuse angle with all z in C. In fact for

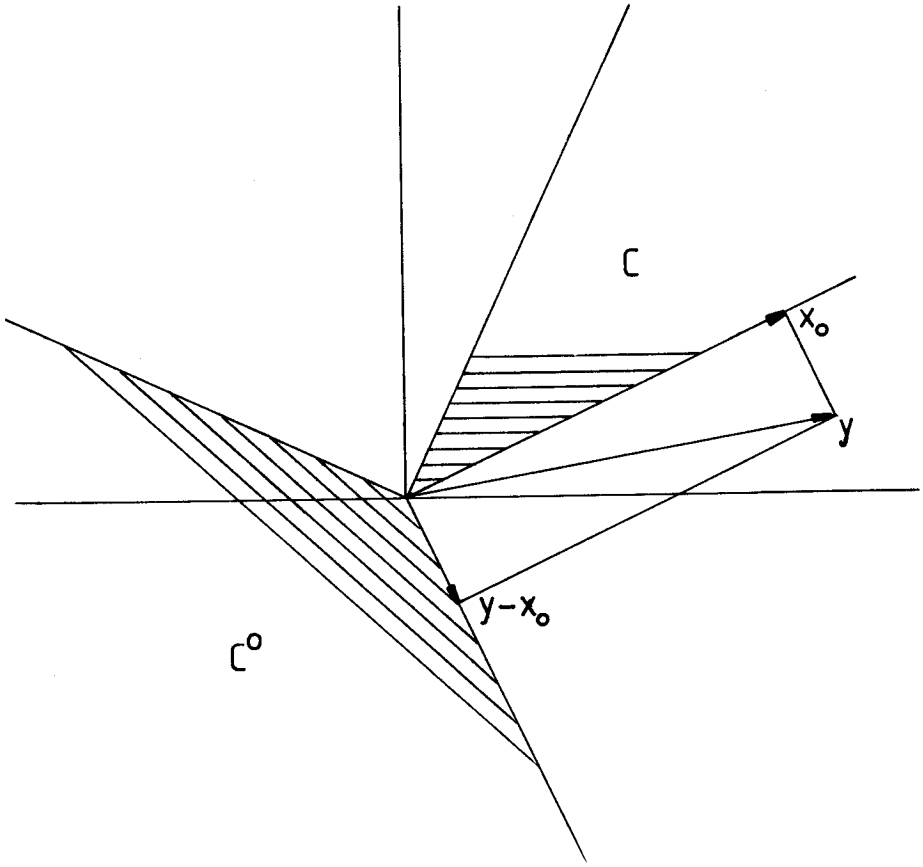


Figure C3: projection on a cone and its polar

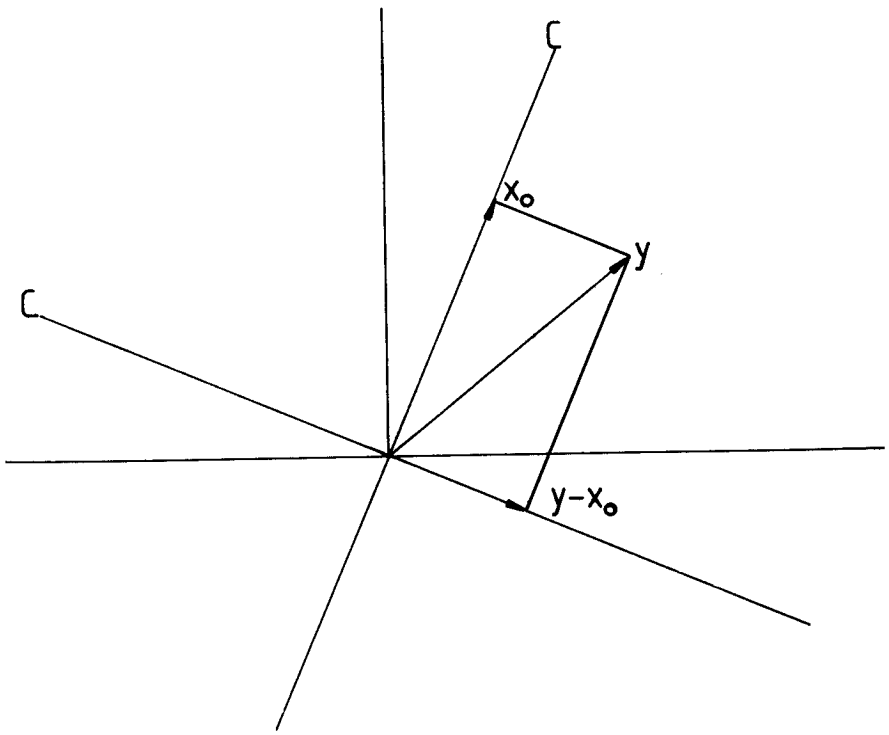


Figure C4: projection on a subspace and its complement

convex C this condition, together with $(y - x_0)'Wx_0 = 0$, characterizes a unique vector x_0 in C, which is consequently the projection we are looking for. In the nonconvex case the situation is considerably less simple.

The characterization of x_0 in the convex case implies the decomposition

$$y'Wy = x_0'Wx_0 + (y - x_0)'W(y - x_0),$$

which is true for all y in R^n. Thus the squared W-length of a vector y is equal to the squared W-length of its projection on a convex cone C and the squared W-length of the residual. Or: y, x_0, and $y - x_0$ are a W-rectangular triangle. Another way to state this is by introducing the polar cone C^o, which is the set of all x such that $x'Wz \leq 0$ for all z in C. The polar cone is illustrated in figure C3, it is convex and closed if C is convex and closed. If x_0 is the projection of y on C, then $y - x_0$ is the projection of y on C^o, and the two projections are W-orthogonal, their squared W-lengths add up to y'Wy. These results generalize more familiar results for subspaces. Observe that a subspace is a (closed, convex) cone for which $x \in C$ and $y \in C$ imply that $\alpha x + \beta y \in C$ for all real α, β. This means that x_0 is the projection of y on the subspace C in the metric W if and only if $(y - x_0)'Wz = 0$ for all $z \in C$, which means that $y - x_0$ must be in the orthogonal complement of C, which is the same thing as the polar cone of C. This is illustrated in figure C4.

An important case of cone projection in this book is monotone regression. We discuss this briefly in a comparitively simple case. Suppose

$$C = \{x \mid x_1 \leq \ldots \leq x_n\},$$

and we want to minimize $(x - y)'W(x - y)$ over x in C, where we suppose now that W is diagonal (with positive elements on the diagonal). There are two results which make the monotone regression problem comparitively easy. We discuss them briefly. Suppose the (unique) solution to the problem is $\hat{x}$.

Result A says that $y_i > y_{i+1}$ implies that $\hat{x}_i = \hat{x}_{i+1}$. The proof is by contradiction. Suppose $\hat{x}_i < \hat{x}_{i+1}$. Define a new vector x^* by $x^*_k = \hat{x}_k$ if $k \neq i$ and $k \neq i + 1$, and by $x^*_i = x^*_{i+1} = (w_i\hat{x}_i + w_{i+1}\hat{x}_{i+1})/(w_i + w_{i+1})$, where the w_i are the diagonal elements of W. Clearly the new vector x^* is also in C, and some algebra gives that

$$(y - x^*)'W(y - x^*) < (y - \hat{x})'W(y - \hat{x}),$$

which shows that $\hat{x}$ cannot be the projection.

Result B says that if we add the constraint $x_i = x_{i+1}$ to the definition of C, then the monotone regression problem can be reduced to a problem in R^{n-1}. The proof is by partitioning the sum of squares in the usual analysis of variance way. Details are in De Leeuw (1977), for example.

These two results suggest an algorithm. If y_i and y_{i+1} are in the wrong order,

then we know that $\hat{x}_i$ and $\hat{x}_{i+1}$ must be equal from result A, we then use result B to reduce the problem to one in R^{n-1}. Again we look for violations, and use them to reduce the problem even further. If there are no violations left, then we are done, ultimately we could end up with a problem in R^1, for which there are by definition no violations.

In the book we sometimes use normalized cone regression. This can be either one of two things. In the first place minimization of

$$\frac{(y - x)'W(y - x)}{x'Wx}$$

over x in a cone C. This is called implicit normalization. This name suggests that there is also something like explicit normalization. This is minimization of $(y - x)'W(y - x)$ over all x in C that satisfy in addition $x'Wx = 1$. It is a basic result of Kruskal and Carroll (1969) that in this simple case implicit normalization, explicit normalization, and no normalization all give essentially the same result. All solutions are proportional to the projection of y on the cone, with only the proportionality constant different for the different problems. This result does not use convexity. There is one exception which should be noted. If y is in C^o, i.e. $y'Wz \leq 0$ for all $z \in C$, then the origin is the projection of y on C. In the normalized problems the solution of x is one of the extreme rays of C, suitably normalized. An extreme ray is any ray in the cone which cannot be written as a nonnegative linear combination of two other rays in the cone. If C is a subspace, and y is in the W-orthogonal complement, then the infimum in the implicitly normalized problem is equal to one, not attained, but approached by letting $x \to \infty$. The infimum in the explicitly normalized problem is attained for any x in C with $x'Wx = 1$, it is equal to $1 + y'Wy$. If C is the cone used in monotone regression, then the extreme rays are the vectors with the first k elements equal to zero and the last n - k elements equal to one (k=1,...,n-1) together with the vectors u and -u which span the intersection of C and C^o. Thus the cone is not pointed. The normalized regression problem must be solved by testing all these rays and by keeping the best one.

REFERENCES

Abbreviations used in the references:

AJS	Australian Journal of Statistics
AMS	Annals of Mathemetical Statistics
AS	Annals of Statistics
BJSP	British Journal of Statistical Psychology = British Journal of Psychology, statistical section = British Journal of Mathemetical and Statistical Psychology
BK	Biometrika
CRAS	Comptes Rendues de l'Academie des Sciences (Paris)
DAN/SSSR	Doklady Akademie Nauk
EPM	Educational and Psychological Measurement
IEEE/IT	Proceedings IEEE, information theory
ISI	International Statistical Institute
ISUP	Institute de Statistique de l'Université de Paris
ITS	Instituut voor toegepaste Sociologie
JAMS	Journal of the Australian Mathemetical Society
JMAA	Journal of Mathemetical Analysis and Applications
JMP	Journal of Mathemetical Psychology
JMV	Journal of Multivariate Analysis
JRSS(A)	Journal of the Royal Statistical Society, series A (general)
JRSS(B)	idem, series B (methodological)
JRSS(C)	idem, series C (applied), also known as 'Applied Statistics'
LAA	Linear Algebra and its Applications
MBR	Multivariate Behavioural Research
PCPS	Proceedings Cambridge Philosophical Society
PM	Psychometrika
PNAS	Proceedings National Academy of Sciences (Washington)
PRSE	Proceedings of the Royal Society (Edinburgh)
PRSL	Proceedings of the Royal Society (London)
SIAM	Society for Industrial and Applied Mathematics
SIAMA	SIAM journal, Applied Mathematics
SIAMC	SIAM journal, Control and Optimization
SIAMN	SIAM journal, Numerical Analysis
SIAMR	SIAM Review
ZAMM	Zeitschrift für Angewandte Mathematik und Mechanik

Abelson, R.P.
1962 Scales derived by consideration of variance components in multiway tables. In : H. Gulliksen and S. Messick (eds.) , Psychological Scaling : Theory and application. New York, Wiley

Anderson, T.W.
1958 An introduction to multivariate statistical analysis. New York, Wiley

Anderson, T.W.
1963 Asymptotic theory for principal component analysis. AMS, 34, 122-148

Anscombe, F.J.
1967 Topics in the investigation of linear relations fitted by the method of least squares. JRSS(B), 29, 1-52

Appell, P., et J. Kampé de Fériet
1926 Fonctions hypergéometriques et hyperspheriques. Polynomes de Hermite. Paris, Gauthier-Villars

Barcikowski, R.S., and J.P. Stevens
1975 A Monte Carlo study of the stability of canonical correlations, canonical weights, and canonical variate-variable correlations. MBR, 10, 353-364

Barrett, J.F., and D.G. Lampard
1955 An expansion for some second-order probability distributions. Ieee/IT, 1, 10-15

Bartlett, M.S.
1948 Internal and external factor analysis. BJSP, 1, 73-81

Bartholomew, D.J.
1980 Factor analysis for categorical data. JRSS(B), 42, 293-321

Bechtel, G.G.
1967 The analysis of variance and pairwise scaling. PM, 32, 47-65

Bechtel, G.G.
1971 A dual scaling analysis for paired comparisons. PM, 36, 135-154

Bechtel, G.G.
1976 Multidimensional preference scaling. The Hague, Mouton

Bechtel, G.G., L.R. Tucker, W.C. Chang
1971 Scalar product model for the multidimensional scaling of choice. PM, 36, 369-388

Beckenbach, E.F., and R. Bellman
1965 Inequalities. Berlin, Springer

Beltrami, E.
1873 Sulle funzioni bilineari. Giorn. Math. Battaglin, 11, 98-106

Bennett, J.F.
1956 Determination of the number of independent parameters of a score matrix from the examination of rank orders. PM, 21, 383-393

Benzécri, J.P.
1965 Sur l'analyse des prêférences. ISUP

Benzécri, J.P.
1967 Lois de probabilité sur un ensemble produit. Les diverses notions d'independence et le critère d'entropie maximale. Mimeo, ISUP

Benzécri, J.P., e.a.
1973 Analyse des données (2 vols). Paris, Dunod

Berge, J.M.F. Ten,
1977 Optimizing factorial invariance. Unpublished doctoral dissertation. University of Groningen

Berman, A.
1973 Cones, matrices, and mathematical programming. Berlin, Springer

Bhattacharya, R.N., and J.K. Ghosh
1978 On the validity of the formal Edgeworth expansion. AS, 6, 434-451

Bickel, P.J.
1976 Another look at robustness : a review of reviews and some new developments. Scand. J. Statist., 3, 145-168

Billingsky, P.
1968 Convergence of probability measures. New York, Wiley

Birkhoff, G.
1967 Lattice Theory. Providence, American Math. Soc.

Bishop, Y.M.M., S. E. Fienberg, P.W. Holland
1975 Discrete multivariate analysis, theory and practice. Cambridge, MIT Press

Black, M.
1949 The definition of scientific method. In : R.C. Staufler , Science and Civilization. Madison, Univ. Wisconsin Press

Blalock Jr., H.M.
1964 Causal interference in nonexperimental research. Chapel Hill, University of North Carolina Press

Bock, R.D.
1960 Methods and applications of optimal scaling. University of North Carolina, L.L. Thurstone Lab., Report 25

Bock, R.D., and M. Lieberman
1970 Fitting a response model for n dichotomously scored items. PM, 35, 179-197

Boor, C. de
1978 A practical guide to splines. Berlin, Springer

Boudon, R.
1967 L'analyse mathématique des faits sociaux. Paris, Plan

Bourouche, J.M., et G. Saporta
1980 L'analyse des données. Paris, Presses Universitaires de France

Box, G.E.P.
1979 Some problems of statistics and everyday life. JASA, 74, 1-4

Box, G.E.P., and D.R. Cox
1964 An analysis of transformations. JRSS(B), 26, 211-252

Boyd, D.W.
1974 The power method for ℓ^p norms. LAA, 9, 95-102

Bunch, J.R., C.P. Nielson, D.C. Sorenson
1978 Rank-one modification of the symmetric eigenproblem. Num. Math. 31, 31-48

Burt, C.
1948 A comparison of factor analysis and analysis of variance. BJSP, 1, 3-27

Burt, C.
1948 Factor analysis and canonical correlations. BJSP, 1, 95-106

Burt, C.
1950 The influence of differential weighting. BJSP, 3, 105-128

Burt, C.
1950 The factorial analysis of qualitative data. BJSP, 3, 166-185

Burt, C.
1951 Test construction and the scaling of items. BJSP, 4, 95-129

Cambanis, S., and B. Liu
1971 On the expansion of a bivariate distribution and its relationship to the output of a nonlinearity. IEEE/IT, 17, 17-25

Caillez, F., et J.P. Pagès
1976 Introduction à l'analyse des données. Paris, SMASH

Carroll, J.D.
1968 A generalization of canonical correlation analysis to three or more sets of variables. Proc. 76th Conv. APA, 227-228

Carroll, J.D., and P. Arabie
1980 Multidimensional scaling. Ann. Rev. Psychol., 31, 607-649

Carroll, J.D., and J.J. Chang
1964 Nonmetric multidimensional analysis of paired comparison data. Psychometric Society Meeting

Chatelin, F.
1979 Approximation spectrale d'operateurs lineaires avec applications au calcul des elements propres d'operateurs differentielles et integraux. Report RR 167, Dep. Math. Appl., Univ. Grenoble

Chesson, P.L.
1976 The canonical decomposition of bivariate distributions. JMV, 6, 526-537

Christofferson, A.
1975 Factor analysis of dichotomized variables. PM, 40, 5-32

Christofferson, A.
1977 Two-step weighted least squares factor ananlysis of dichotomized variables. PM, 42, 443-438

Cliff, N.
1966 Orthogonal rotation to congruence. PM, 31, 33-42

Cochran, W.G.
1972 Observational studies. In : T.A. Bancroft (ed.), Statistical papers in honour of George Snedecor. Ames, Iowa State Univ. Press

Collaris, J.W.M., en J.A. Kropman
1978 Van Jaar tot Jaar, tweede fase. Den Haag, Staatsuitgeverij

Cooley, W.W., and P.R. Lohnes
1962 Multivariate procedures for the behavioural sciences. New York, Wiley

Cooley, W.W., and P.R. Lohnes
1971 Multivariate data analysis. New York, Wiley

Coombs, C.H.
1964 A theory of data. New York, Wiley

Cooper, R.D., M.R. Hoare, M. Rahman
1977 Stochastic processes and special functions : on the probabilistic origin of some positive kernels associated with classical orthogonal polynomials. JMAA, 61, 262-291

Cox, D.R.
1957 Note on grouping. JASA, 52, 543-547

Cozijn, C., en J.J.M. van Dijk
1976 Onrustgevoelens in Nederland. WODC

Curry, H.B., and I.J. Schoenberg
1966 On Polya frequency functions IV ; the fundamental spline functions and their limits. J. d'Analyse Math., 17, 71-107

Daalder, H., and J.G. Rusk
1972 Perceptions of party in the Dutch Parliament. In : S.C. Patterson and J.C. Wahlke (eds.) , Comparative legislative behaviour : frontiers of research. New York, Wiley

Daalder, H., en J.P. van de Geer
1977 Partijafstanden in de Tweede Kamer. Acta Politica, 12, 289-345

Dagnelie, P.
1975 Analyse statistique à plusieurs variables. Gembloux, Presses Agronomiques

Dahmen, W.
1979 Multivariate B-splines ; recurrence relations and linear combinations of truncated powers. In : W. Schempp and K. Zeller (eds.) , Multivariate approximation theory. ISNM 51. Basel, Birkhäuser

Dahmen, W.
1980 On multivariate B-splines. SIAMN, 17, 179-191

Dahmen, W.
1980 Konstruktion mehr dimensionaler B-splines und ihre Anwendung auf Approximations Probleme. In : Numerische Methoden der Approximations Theorie V. ISNM 52, Basel, Birkhäuser

Darlington, R.B., S.L. Weinberg, H.J. Walberg
1973 Canonical variate analysis and related techniques. Rev. Educ. Research, 43, 433-454

Darroch, J.N.
1974 Multiplicative and additive interaction in contingency tables. BK, 61, 207-214

Daudin, J.J.
1980 Régression qualitative : choix d'espace prédicteurs. In : E. Diday et al. (eds.) , Data analysis and informatics. Amsterdam, North Holland

Daugavet, V.A.
1968 Variant of the stepped exponential method of finding some of the first characteristic values of a symmetric matrix. USSR Comp. Math. Physics, 8, (1) 212-223

Dauxois, J., et A. Pousse,
1976 Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique. Dissertation. Université Paul Sabatier, Toulouse

Davis, A.W.
1977 Asymptotic theory for principal component analysis : nonnormal case. AJS, 19, 206-212

Davis, C., and W.M. Kahan
1970 The rotation of eigenvectors by a perturbation. SIAMN, 7, 1-46

Demjanov, V.F., and V.N. Malozemov
1974 Introduction to minimax. New York, Wiley

Dempster, A. P.
1969 Elements of continuous multivariate analysis. Reading, Addison-Wesley

Dempster, A.P.
1971 An overview of multivariate data analysis. JMV, 1, 316-346

Deville, J.C., and G. Saporta
1980 Analyse harmonique qualitative. In E. Diday et al. (eds.) , Data analysis and informatics. Amsterdam, North Holland

Divgi, D.R.
1979 Calculation of the tetrachoric correlation coefficient. PM, 44, 169-172

Dronkers, J.
1978 Manipuleerbare variabelen in de schoolloopbaan. 9th World congress of sociology. Uppsala

Dronkers, J., en J.J.M. Jungbluth
1979 Schoolloopbaan en geslacht. In : J. Peschar (ed.) , Van achteren naar voren. Den Haag, Staatsuitgeverij

Drouet d'Aubigny, G.
1975 Description statistique des données ordinales : analyse multidimensionelle. Thex, Grenoble

Eagleson, G.K.
1964 Polynomial expansions of bivariate distributions. AMS, 35, 1208-1215

Eagleson, G.K.
1969 Canonical expansions of birth and death processes. Theory Prob. Appl., 14, 209-218

Eagleson, G.K.
1969 A characterization theorem for positive definite sequences on the Krawtchouk polynomials. AJS, 11, 29-38

Eagleson, G.K., and H.O. Lancaster
1967 The regression system of sums with random elements in common. AJS, 9, 119-125

Eckart, C., and G. Young
1936 The approximation of one matrix by another of lower rank. PM, 1, 211-218

Edgeworth, F.Y.
1888 The statistics of examinations. JRSS(A), 51, 599-635

Edgerton, H.A., and L.E. Kolbe
1936 The method of minimum variation for the combination of criteria. PM, 1, 183-187

Edwards, A.L.
1957 Techniques of attitude scale construction. New York, Appleton-Century-Crofts

Efron, B.
1979 Bootstrap methods : another look at the jackknife. AS, 7, 1-26

Elias, P.
1970 Bounds on performance of optimum quantizers. IEEE/IT, 16, 172-184

Erdelyi, A.
1953 Higher transcendental functions. New York, McGraw Hill

Escofier, B., et B. Le Roux
1972 Etude de trois problemes de stabilitê en analyse factorielle. Publ. ISUP, 21, 2-48

Fan, Ky
1951 Maximum properties and inequalities for the eigenvalues of completely continuous operators. PNAS, 37, 760-766

Ferguson, G.A.
1941 The factorial interpretation of test difficulty. PM, 6, 323-329

Fienberg, S.E.
1977 The analysis of cross-classified categorical data. Cambridge, MIT

Fischer, G.H.
1974 Einführung in die Theorie psychologischer Tests. Bern, Huber

Fisher, R.A.
1940 The precision of discriminant functions. Ann. Eug., 10, 422-429

Forsythe, A. and J.A. Hartigan
1970 Efficiency of confidence intervals generated by repeated subsample calculations. BK, 57, 629-640

Galton, F.
1888 Co-relations and their measurement, chiefly from anthropometric. PRSL, 45, 135-145

Gantmacher, F.R., et M.G. Krein
1937 Sur les matrices oscillatoires et complètement non-négatives. Composition Math., 4, 445-476

Gantmacher, F.R., und M.G. Krein
1960 Oszillationsmatrizen, Oszillationskerne und kleine Schwingungen mechanischer Systeme. Berlin, Akademie Verlag

Gebelein, H.
1941 Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. ZAMM, 21, 364-379

Geer, J.P. van de
1967 Inleiding in de multivariate analyse. Arnhem, Van Loghem Slaterus

Geer, J.P. van de
1968 Matching K sets of configurations. Report RN 005-68. Department of Datatheory, Univ. Leiden

Geer, J.P. van de
1971 Introduction to multivariate analysis for the social sciences. San Francisco, Freeman

Geer, J.P. van de
1980 Introduction to multivariate linear data analysis. Part V : Relations between K sets of data. Department of Datatheory, Univ. Leiden

Gersho, A.
1979 Asymptotically optimal block quantization. IEEE/IT, 25, 373-380

Gifi, A.
1980 Niet-lineaire multivariate analyse. Department of Datatheory, Univ. Leiden

Gilbert, E.S.
1968 On discrimination using qualitative variables. JASA, 63, 1399-1412

Giri, N.C.
1977 Multivariate statistical inference. New York, Academic Press

Girschick, M.A.
1936 Principal components. JASA, 31, 519-528

Gish, H., and J.N. pierce
1968 Asymptotically optimal quantizing. IEEE/IT, 14, 676-683

Glass, D.V., (ed.)
1954 Social mobility in Britain. Glencoe, Free press

Gokhale, D.V., and S. Kullback
1978 The information in contingency tables. New York, Dekker

Goldberg, S.
1958 Introduction to difference equations : with illustrative examples from economics, psychology and sociology. New York, Wiley

Goldman, A.J., and A.W. Tucker
1956 Polyhedral convex cones. In : H.W. Kuhn and A.W. Tucker (eds.) , Linear inequalities and related systems. Princeton, Princeton University Press

Good, I.J.
1963 Maximum entropy for hypotheses formulation, especially for multidimensional contingency tables. AMS, 34, 911-934

Goodman, L.A.
1965 On the statistical analysis of mobility tables. Am. J. Sociol., 70, 564-585

Goodman, L.A.
1969 On the measurement of social mobility : an index of status persistence. Amer. Sociol. Rev., 34, 832-850

Goodman, L.A.
1978 Analyzing qualitative/categorical data, loglinear models, and latent structure analysis. Reading, Addison-Wesley

Goodman, L.A., and W.H. Kruskal
Measures of association for cross classification
1954 I : JASA, 49, 732-764
1959 II : JASA, 54, 123-163
1963 III : JASA, 58, 310-364
1972 IV : JASA, 67, 415-421

Goodman, L.A., and W.H. Kruskal
1979 Measures of association for cross classification. Berlin, Springer

Gordon, L.
1974(a) Completely separating groups in subsampling. AS, 2, 572-578

Gordon, L.
1974(b) Efficiency in subsampling. AS, 2, 739-750

Green, P.E., and J.D. Carroll
1976 Mathematical tools for applied multivariate analysis. New York, Wiley

Green, P.E., M.H. Halbert, P.J. Robinson
1966 Canonical analysis : an exposition and illustrative application. J. Marketing Research, 3, 32-39

Griffiths, R.C.
1969 The canaonical correlation coefficients of bivariate gamma distributions. AMS, 40, 1401-1408

Griffiths, R.C.
1970 Positive definite sequences and canonical correlation coefficients. AJS, 12, 162-165

Guilford, J.P.
1941 The difficulty of a test and its factor composition. PM, 6, 67-77

Guilford, J.P.
1954 Psychometric methods. New York, McGraw Hill

Gulliksen, H.
1950 Theory of mental tests, New York, Wiley

Guttman, L.
1941 The quantification of a class of attributes : a theory and method of scale construction. In : P. Horst (ed.) , The prediction of personal adjustment. New York, SSRC

Guttman, L.
1944 A basis for scaling qualitative data. Amer. Sociol. Rev., 9, 139-150

Guttman, L.
1946 An approach for quantifying paired comparisons and rank order. AMS, 17, 144-163

Guttman, L.
1950 The principal components of scale analysis. In : S.A. Stouffer et al. , Measurement and prediction. Princeton, Princeton University Press

Guttman, L.
1950 The basis for scalogram analysis. In : S.A. Stouffer et al. , Measurement and prediction. Princeton, Princeton University Press

Guttman, L.
1954 The principal components of scalable attitudes. In : P.F. Lazarsfeld (ed.) , Mathematical thinking in the social sciences. Glencoe, Free Press

Guttman, L.
1959 Metricizing rank-ordered or unordered data for a linear factor analysis. Sankhya, A, 21, 257-268

Guttman, L.
1968 A general nonmetric technique for finding the smallest coordinate space for a configuration of points. PM, 33, 469-506

Haberman, S.J.
1974 The analysis of frequency data. Chicago, University of Chicago Press

Halmos, P.R.
1948 Finite dimensional vector spaces. Princeton, Princeton University Press

Hampel, F.R.
1973 Robust estimation : a condensed partial survey. Z. Wahrscheinlichkeitstheorie verw. Geb., 27, 87-104

Hampel, F.R.
1974 The influence curve and its role in robust estimation. JASA, 69, 383-393

Hannan, E.J.
1961 The general theory of canonical correlation and its relation to functional analysis. JAMS, 2, 229-242

Hartigan, J.A.
1969 Using subsample values as typical values. JASA, 64, 1303, 1317

Hartigan, J.A.
1971 Error analysis by replaced samples. JRSS(B), 33, 98-110

Hartigan, J.A.
1975 Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. AS, 3, 573-580

Hartmann, W.
1979 Geometrische Modelle zur Analyse empirischer Daten. Berlin, Akademie Verlag

Heiser, W.J., and J. de Leeuw
1979 Metric multidimensional unfolding. MDN, 4, 26-50

Heiser, W.J., and J. de Leeuw
1979 How to use SMACOF-3. Department of Datatheory, University of Leiden

Hill, M.O.
1974 Correspondence analysis : a neglected multivariate method. JRSS(C), 23, 340-354

Hiriart-Urruty, J.B.
1978 Gradients gênéralises de functions marginals. SIAMC, 16, 301-316

Hirschfeld, H.O.
1935 A connection between correlation and contingency. PCPS, 31, 520-524

Hirshi, T., and H.C. Selvin
1973 Principles of survey analysis. Glencoe, Free Press

Hogan, W.W.
1973(a) Point-to-set maps in mathematical programming. SIAMR, 15, 591-603

Hogan, W.W.
1973(b) Directional derivation for extremal value functions, with applications to the completely convex case. Operations Research, 21, 188-209

Horst, P.
1935 Measuring complex attitudes. J. Soc. Psychol., 6, 369-374

Horst, P.
1936 Obtaining a composite measure from a number of different measures of the same attribute. PM, 1, 53-60

Horst, P.
1961(a) Relations among m sets of variables. PM, 26, 129-149

Horst, P.
1961(b) Generalized canonical correlations and their applications to experimental data. J. Clin. Psychol., 17, 331-347

Hotelling, H.
1933 Analysis of a complex of statistical variables into principal components. J. Educ. Psychol., 24, 417-441, 498-520

Hotelling, H.
1935 The most predictable criterion. J. Educ. Psychol., 26, 139-142

Huber, P.J.
1972 Robust statistics : a review. AMS, 43, 1041-1067

Hurt, J.
1976 Asymptotic expansions of functions of statistics. Appl. Math., 21, 444-456

ITS
1968 Onderzoeksvoorstel 'Van Jaar tot Jaar'. Nijmegen, ITS

Jensen, D.R.
1971 A note on positive dependence and the structure of bivariate distributions. SIAMA, 20, 749-753

Johnson, P.O.
1950 The quantification of qualitative data in discriminant analysis. JASA, 45, 65-76

Jordan, C.
1874 Mêmoire sur les formes bilinêaires. J. Math. Pures. Appl., 19, 35-54

Karlin, S.
1964 Oscillation properties of eigenvectors of strictly totally positive matrices. J. Anal. Math. Jerusalem, 9, 247-266

Karlin, S.
1968 Total positivity. Stanford, Stanford University Press

Kato, T.
1976 Perturbation theory for linear operators. Berlin, Springer

Kendall, M.G.
1957 A course in multivariate analysis. London, Griffin, 1957/1975

Kendall, M.G.
1962 Rank correlation methods. London, Griffin

Kendall, M.G.
1972 The history and future of statistics. In : T.A. Bancroft (ed.) , Statistical papers in honour of George Snedecor. Ames, Iowa State University Press

Kettenring, J.R.
1971 Canonical analysis of several sets of variables. BK, 58, 433-460

Kiefer, J.
1964 Review of M.G. Kendall and A. Stuart : The advanced theory of statistics, volume 2. AMS, 35, 1371-1380

Kolata, W.G.
1978 Approximation in variationally posed eigenvalue problems. Numer. Math., 29, 159-171

Kropman, J.A., en J.W.M. Collaris
1974 Van Jaar tot Jaar : eerste fase. Nijmegen, ITS

Kruskal, J.B.
1964(a) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. PM, 29, 1-27

Kruskal, J.B.
1964(b) Nonmetric multidimensional scaling : a numerical method. PM, 29, 115-129

Kruskal, J.B.
1965 Analysis of factorial experiments by estimating monotone transformations of the data. JRSS(B), 27, 251-263

Kruskal, J.B., and J.D. Carroll
1969 Geometric models and badness-of-fit functions. In : P.R. Krishnaiah (ed.) , Multivariate analysis, vol. II. New York, Academic Press

Kruskal, J.B., and R.N. Shepard
1974 A nonmetric variety of linear factor analysis. PM, 39, 123-157

Kullback, S.
1959 Information theory and statistics. New York, Wiley

Lafaye de Michaux, D.
1978 Approximation d'analyses canoniques non-linéaires de variables aléatoires. Dissertation. Université de Nice

Lammers, C.J.
1969 Is de universiteit een politieke leerschool? Universiteit en Hogeschool, 15, 1-43

Lancaster, H.O.
1957 Some properties of the bivariate normal distribution considered in the form of a contingency table. BK, 44, 289-292

Lancaster, H.O.
1958 The structure of bivariate distributions. AMS, 29, 719-736

Lancaster, H.O.
1959 Zero correlation and independence. AJS, 1, 53-56

Lancaster, H.O.
1960(a) On tests of independence in several dimensions. JAMS, 1, 241-254

Lancaster, H.O.
1960(b) On statistical independence and zero correlation in several dimensions. JAMS, 1, 492-496

Lancaster, H.O.
1969 The chi-squared distribution. New York, Wiley

Lancaster, H.O.
1971 The multiplicative definition of interaction. AJS, 13, 36-44

Lancaster, H.O.
1975(a) The multiplicative definition of interaction : an addendum. AJS, 17, 34-35

Lancaster, H.O.
1975(b) Joint probability distributions in the Meixner classes. JRSS(B), 37, 434-443

Lancaster, H.O., and M.A. Hamdan
1964 Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characteristics. PM, 29, 383-391

Lawley, D.N.
1944 The factorial analysis of multiple item tests. PRSE, 62, 74-82

Lazarsfeld, P.F.
1950 The logical and mathematical foundations of latent structure analysis. In : S.S. Stouffer (ed.) , Measurement and prediction. Princeton, Princeton University Press

Lazarsfeld, P.F., and N.W. Henry
1968 Latent structure analysis. New York, Houghton Mifflin

Lebart, L.
1976 The significance of eigenvalues issued from correspondence analysis of contingency tables. In : Proceedings COMPSTAT 1976 , Wien, Physika Verlag

Lee, P.A.
1971 A diagonal expansion for the 2-variate Dirichlet probability density function. SIAMA, 21, 155-165

Leeuw, J. de
1968 Canonical analysis of categorical data. Department of Datatheory, Univ. Leiden

Leeuw, J. de
1969 Some contributions to the analysis of categorical data. Report RN 004-69. Department of Datatheory, Univ. Leiden

Leeuw, J. de
1973 Canonical analysis of categorical data. Dissertation. Univ. Leiden

Leeuw, J. de
1977(a) Correctness of Kruskal's algorithms for monotone regression with ties. PM, 42, 141-144

Leeuw, J. de
1977(b) Normalized cone regression. Mimeo. Department of Datatheory, Univ. Leiden

Leeuw, J. de, and W.J. Heiser
1980 Theory of multidimensional scaling. In : P.R. Krishnaiah and L. Kanal (eds.) , Handbook of statistics. Amsterdam , North Holland

Leeuw, J. de, and J. van Rijckevorsel
1980 HOMALS & PRINCALS, some generalizations of principal components analysis. In : E. Diday et al. (eds.), Data analysis and informatics. Amsterdam, North Holland

Leeuw, J. de, en I. Stoop
1979 Sekundaire analyse 'Van Jaar tot Jaar' met behulp van niet-lineaire multivariate technieken. In : J. Peschar (ed.) , Van achteren naar voren. Den Haag, Staatsuitgeverij

Leeuw, J. de, F.W. Young, Y. Takane
1976 Additive structure in qualitative data : an alternating least squares method with optimal scaling features. PM, 41, 471-503

Levine, M.V.
1970 Transformations that render curves parallel. JMP, 7, 410-443

Levine, M.V.
1972 Transforming curves in curves with the same shape. JMP, 9, 1-16

Levine, M.V.
1975 Additive measurement with short segments of curves. JMP, 12, 212-224

Lingoes, J.C.
1968 The multivariate analysis of qualitative data. MBR, 3, 61-94.

Loevinger, J.
1947 A systematic approach to the construction and evaluation of tests of ability. Psychol. Monograph, 61, no 4.

Loevinger, J.
1948 The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. Psychol. Bull., 45, 507-530.

Lord, F.M.
1958 Some relations between Guttman's principal components of scale analysis and other psychometric theory. PM, 23, 291-296.

Lorenz, G.
1953 Bernstein polynomials. Toronto, Univ. of Toronto Press.

McDonald, R.P.
1967 Nonlinear factor analysis. Psychometric monograph, 15.

MacDonell, W.R.
1901 On criminal anthropology and the identification of criminals. BK, 1, 177-227.

McFadden, J.A.
1966 A diagonal expansion in Gegenbauer polynomials for a class of second-order probability densities. SIAMA, 14, 1433-1436.

McGraw, D.K., and Wagner, J.T.
1968 Elliptically symmetric distributions. IEEE/IT, 14, 110-120.

MacKenzie, D.
1978 Statistical theory and social interests: a case study. Social studies of science, 8, 35-83.

McKinlay, S.M.
1975 The design and analysis of the observational study. JASA, 70, 503-523.

Masson, M.
1974 Analyse non-linéaire des données. CRAS, 278, 803-806.

Maung, K.
1941 Measurement of association in a contingency table with special reference to the pigmentation of hair and eye colours of Scottish school children. Ann. Eugenics, 11, 189-223.

Maung, K.
1941 Discriminant analysis of Tocher's eye colour data. Ann. Eugenics, 11, 64-76.

Max, J.
1960 Quantizing for minimum distortion. IEEE/IT, 6, 7-12.

Miller, R.G.
1974 The jackknife: a review. BK, 61, 1-15.

Mokken, R.J.
1970 A theory and procedure of scale analysis. The Hague, Mouton.

Morrison, D.F.
1967 Multivariate statistical methods. New York, McGraw Hill, 1967, 1976.

Mosteller, F.
1949 A theory of scalogram analysis using noncumulative types of items: a new approach to Thurstone's method of scaling attitudes. Rep. no. 9, Lab. of Social Relations, Harvard University.

Muthèn, B.
1978 Contributions to factor analysis of dichotomous variables. PM, 43, 551-560.

Naouri, J.C.
1970 Analyse factorielle des correspondences continues. Publ. ISUP, 19, 1-100.

Nevels, K.
1974 Generalized canonical variates and correlations. Rep. HB-74-158-EX, Psychol. Department, University of Groningen.

Niemöller, B.
1976 Schaalanalyse volgens Mokken. Technisch Centrum FSW, University of Amsterdam.

Nishisato, S.
1980 Analysis of categorical data: dual scaling and its applications. Toronto, Univ. of Toronto Press.

Norton, B.J.
1978 Karl Pearson and statistics: the social origin of scientific innovation. Social studies of science, 8, 3-34.

Okamoto, M.
1968 Optimality of principal components. In: P.R. Krishnaiah (ed), Multivariate Analysis II. New York, Academic Press.

Olsson, U.
1979 Maximum likelihood estimation of the polychoric correlation coefficient. PM, 44, 443-460.

O'Neill, M.E.
1978 Asymptotic distribution of the canonical correlation coefficient from contingency tables. AJS, 20, 75-82.

O'Neill, M.E.
1978 Distributional expansion for canonical correlations from contingency tables. JRSS(B), 40, 303-312.

O'Neill, M.E.
1980 The distribution of higher-order interactions in contingency tables. JRSS(B), 42, 357-365.

Parzen, E.
1979 Nonparametric statistical data modelling. JASA, 74, 105-131.

Pearson, K.
1892 The Grammar of Science. London, Scott, 1892, 1900, 1910.

Pearson, K.
1901 One lines and planes of closest fit to points in space. Phil. Magazine, 2, 559-572.

Pearson, K.
1904 On the theory of contingency and its relation to association and normal correlation. Drapers company research memoirs, biometric series, no 1.

Pearson, K., and Heron, D.
1913 On theories of association. BK, 11, 159-315.

Popper, K.R.
1963 Conjectures and refutations. London, Routledge and Kegan Paul.

Puri, M.L., and Sen, P.K.
1971 Nonparametric methods in multivariate analysis. New York, Wiley.

Rao, C.R.
1964 The use and interpretation of principal component analysis in applied research. Sankhya, 26, 329-358.

Rao, C.R.
1980 Matrix approximation and reduction of dimensionality in multivariate statistical analysis. In: P.R. Krishnaiah (ed), Multivariate analysis V. Amsterdam, North Holland Publ. Co.

Rao, C.R., and Slater, P.
1949 Multivariate analysis applied to differences between neurotic groups. BJSP, 2, 17-29.

Rao, B.R.
1969 Partial canonical correlations. Trabajos de Estadistica, 20, 211-219.

Rasch, G.
1960 Probabilistic models for some intelligence and attainment tests. Copenhagen, Danish institute for educational research.

Rasch, G.
1961 On general laws and meaning of measurement in psychology. In: Proc. IV Berkeley Symp. on math. statist. and probability, Berkeley, Univ. of California Press.

Rasch, G.
1966 An item analysis which takes individual differences into account. BJSP, 19, 49-57.

Reissman, L.
1953 Levels of aspiration and social class. Amer. Sociol. Rev., 18, 233-242.

Rényi, A.
1959 On measures of dependence. Acta Math. Acad. Sc. Hungar., 10, 441-451.

Robert, F.
1967 Calcul du rapport maximal de deux normes sur $\mathbb{R}^n$. RIRO, 1, 97-118.

Rockafellar, R.T.
1970 Convex analysis. Princeton, Princeton Univ. Press.

Roe, G.M.
1964 Quantizing for minimum distortion. IEEE/IT, 10, 384-385.

Roskam, E.E.
1968 Metric analysis of ordinal data in psychology. Voorschoten, VAM.

Roskam, E.E.
1977 A survey of the Michigan-Israel-Netherlands-Integrated series. In: J.C. Lingoes (ed), Geometric representations of relational data. Ann Arbor, Mathesis Press, 1977.

Ross, J., and Cliff, N.
1964 A generalization of the interpoint distance model. PM, 29, 167-176.

Roux, B. Le, and Rouanet, H.
1979 L'analyse statistique des protocolles multidimensionelles: analyse en composantes principales. Publ. ISUP, 24, 47-74.

Rowney, D.K., and Graham, J.Q. (eds).
1969 Quantitative History. Homewood.

Roy, S.N.
1957 Some aspects of multivariate analysis. New York, Wiley.

Rozeboom, W.W.
1979 Sensitivity of a linear composite of predictor items to differential item weighting. PM, 44, 289-296.

Russett, B.M.
1964 Inequality and instability. World Politics, 21, 442-454.

Rijckevorsel, J.L.A. van, Bettonvil, B., and Leeuw, J. de.
1980 Recovery and stability in nonlinear principal component analysis. Paper presented at the European meeting of the Psychometric Society, Groningen, 1980.

Saporta, G.
1975 Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. Thèse, Université Paris VI.

Sarmanov, O.V.
1963 Investigation of stationary Markov processes by the method of eigenfunction expansions. Selected translations in Math. Statist. and Probability, 4, 245-269.

Sarmanov, O.V., and Bratoeva, Z.N.
1967 Probabilistic properties of bilinear expansions of Hermite polynomials. Theory Probality and Appl., 12, 470-481.

Sarmanov, O.V., and Zacharov, V.K.
1960 Maximum coefficients of multiple correlation. DAN/SSSR, 130, 269-271.

Scheffé, H.
1952 An analysis of variance for paired comparisons. JASA, 47, 381-400.

Sharma, D.K.
1978 Design of absolutely optimal quantizers for a wide class of distortion measures. IEEE/IT, 24, 693-702.

Shepard, R.N.
1962 The analysis of proximities: multidimensional scaling with an unknown distance function. PM, 27, 125-140 and 219-245.

Shepard, R.N.
1966 Metric structures in ordinal data. JMP, 3, 287-315.

Sheppard, W.F.
1898 On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. Proc. London Math. Soc, 29, 353-380.

Sibson, R.
1972 Order invariant methods of data analysis. JRSS(B), 34, 311-349.

Slater, P.
1960 The analysis of personal preferences. BJSP, 13, 119-135.

Spearman, C.
1913 Correlation of sums and differences. BJP, 5, 417-423.

Spingarn, J.E.
1980 Fixed and variable constraints in sesitivity analysis. SIAMC, 18, 297-310.

Steel, R.G.D.
1951 Minimum generalized variance for a set of linear functions. AMS, 22, 456-460.

Stevenson, C.L.
1938 Persuasive definitions. Mind, 47, 331-353.

Stewart, G.W.
1973 Error and perturbation bounds for subspaces associated with certain eigenvalue problems. SIAMR, 15, 727-764.

Stewart, G.W.
1973 Introduction to matrix computation. New York, Academic Press.

Stewart, G.W.
1975 Gershgorin theory for the generalized eigenvalue problem $Ax = \lambda Bx$. Math. Computation, 29, 600-606.

Stewart, G.W.
1978 Perturbation theory for the generalized eigenvalue problem. In: C. de Boor and G.H. Golub (eds), Recent advantages in numerical analysis. New York, Academic Press.

Stewart, G.W.
1979 Perturbation bounds for the definite generalized eigenvalue problem. LAA, 23, 69-85.

Stoop, I.
1980 Sekundaire analyse van de 'Van Jaar tot Jaar' data met behulp van niet-lineaire multivariate technieken: verschillen in de schoolloopbaan van meisjes en jongens. Rep. RB 001-80, Department of Data theory, Univ. of Leiden.

Styan, G.P.
1973 Hadamard products and multivariate statistical analysis. LAA, 6, 217-240.

Stigler, S.M.
1977 Do robust estimators work with real data ? AS, 5, 1055-1098.

Sugiyama, M.
1975 Religious behaviour of the Japanese. Execution of a partial order scalogram analysis based on quantification theory. US-Japan seminar on theory and application of multidimensional scaling and related techniques. La Jolla.

Sylvester, J.J.
1889 Sur la réduction biorthogonal d'une forme linéo-linéaire à sa forme canonique. CRAS, 108, 651-653.

Symm, H.J., and Wilkinson, J.H.
1980 Realistic error bounds for a simple eigenvalue and its associated eigenvector. Numerische Math., 35, 113-126.

Takane, Y., Young, F.W., De Leeuw, J.
1976 Nonmetric common factor analysis. An alternating least squares method with optimal scaling features. Behaviourmetrika.

Takane, Y., Young, F.W., De Leeuw, J.
1980 An individual differences additive model. An alternating least squares method with optimal scaling features. PM, 45, 183-209.

Tatsuoka, M.M.
1971 Multivariate analysis: techniques for educational and psychological research. New York, Wiley.

Tenenhaus, M.
1977 Analyse en composantes principales d'un ensemble de variables nominales et numériques. Revue de statist. appl., 25, 39-56.

Thompson, R.C., and Therianos, S.
1972 Inequalities connecting the eigenvalues of Hermitean matrices with the eigenvalues of complementary principal submatrices. Bull. Aust. Math. Soc., 6, 117-132.

Thompson, R.C., and Therianos, S.
1972 The eigenvalues of complementary principal submatrices of a positive definite matrix. Canadian J. Math., 24, 658-667.

Thompson, R.C., and Freede, L.J.
1970 On the eigenvalues of sums of Hermitean matrices. Aequationes Math., 5, 103-115.

Thompson, R.C., and Freede, L.J.
1971 On the eigenvalues of sums of Hermitean matrices II. LAA, 4, 369-376.

Thompson, R.C.
1975 Singular value inequalities for matrix sums and minors. LAA, 11, 251-269.

Thompson, R.C.
1976 The behaviour of eigenvalues and singular values under perturbations of restricted rank. LAA, 13, 69-78.

Thomson, G.H.
1934 Hotelling's method modified to give Spearman's g. J. Educational Psychol., 25, 366-374.

Thomson, G.H.
1947 The maximum correlation of two weighted batteries. BJSP, 1, 27-34.

Thorndike, R.M.
1978 Correlational procedures for research. New York, Gardner.

Thorndike, R.M., and Weiss, D.J.
1973 A study of the stability of canonical correlations and canonical components. EPM, 33, 123-134.

Thurstone, L.L.
1947 Multiple factor analysis. Chicago, Univ. of Chicago Press.

Timm, N.H., and Carlson, J.E.
1976 Part and bipartial canonical correlation analysis. PM, 41, 159-176.

Tintner, G.
1946 Some applications of multivariate analysis to economic data. JASA, 41, 472-500.

Topsøe, F.
1967 Preservation of weak convergence under mappings. AMS, 38, 1661-1665.

Torgerson, W.S.
1958 Theory and methods of scaling. New York, Wiley.

Tricomi, F.G.
1955 Vorlesungen über Orthogonalreihen. Berlin, Springer.

Tucker, L.R.
1960 Intra-individual and inter-individual multidimensionality. In: H. Gulliksen and S. Messick (eds), Psychological scaling: theory and application. New York, Wiley.

Tukey, J.W.
1962 The future of data analysis. AMS, 33, 1-67.

Tukey, J.W.
1977 Exploratory data analysis. Reading, Addison-Wesley.

Tukey, J.W.
1980 We need both exploratory and confirmatory. Amer. Statistician, 34, 23-25.

Tyan, S., and Thomas, J.B.
1975 Characterization of a class of bivariate distribution functions. JMV, 5, 227-235.

Vainikko, G.
1976 Funktionalanalysis der Diskretisierungsmethoden. Leipzig, Teubner.

Venter, J.H.
1966 Probability measures on product spaces. South African statist. J., 1, 3-20.

Wainer, H.
1976 Estimating coefficients in linear models: it don't make no nevermind. Psychol. Bull., 83, 213-217.

Waternaux, C.M.
1976 Asymptotic distribution of the sample roots for a nonnormal population. BK, 63, 639-645.

Weeks, D.G., and Bentler, P.M.
1979 A comparison of linear and monotone multidimensional scaling models. Psychol. Bull., 86, 349-354.

Weinberger, H.F.
1960 Error bounds in the Rayleigh-Ritz approximation of eigenvalues. J. Research National Bureau of standards, 64B, 217-225.

Weinberger, H.F.
1974 Variational methods for eigenvalue approximation. Philadelphia, SIAM.

Weiss, D.J.
1972 Canonical correlation analysis in counseling psychology research. J. Counseling Psychol., 19, 241-252.

Werner, B.
1974 Optimale Schranken für Eigenelemente selbstadjungierter Operatoren in der Hilbertraumnorm. In: ISNM no 24, Basel, Birkhauser.

Wilkinson, J.H.
1961 Rigorous error bounds for computed eigensystems. Computer J., 4, 230-241.

Wilks, S.S.
1938 Weighting systems for linear functions of correlated variables when there is no independent variable. PM, 3, 23-40.

Williams, E.J.
1952 Use of scores for the analysis of association in contingency tables. BK, 39, 274-289.

Wilson, E.B.
1926 Empiricism and rationalism. Science, 64, 47-57.

Wilson, E.B., and Worcester, J.
1939 Note on factor analysis. PM, 4, 133-148.

Winsberg, S., and Ramsay, J.O.
1980 Monotonic transformations to additivity using splines. BK, 67, 669-674.

Winsberg, S., and Ramsay, J.O.
1981 Analysis of pairwise preference data using integrated B-splines. PM, 46, in press.

Wolfowitz, J.
1969 Reflections on the future of mathematical statistics. In: R.C. Bose et al (eds), Essays in probability and statistics. Chapell Hill, Univ. of N. Carolina Press.

Wollenberg, A. van der
1977 Redundancy analysis: an alternative for canonical correlation. PM, 42, 207-219.

Wong, E., and Thomas, J.B.
1962 On polynomial expansions of second order distributions. SIAMA, 10, 507-516.

Wood, R.C.
1969 On optimum quantization. IEEE/IT, 15, 248-252.

Wood, D.A., and Erskine, J.A.
1976 Strategies in canonical correlation with applications to behavioural data. EPM, 36, 861-878.

Woodward, J.A., and Overall, J.E.
1976 Factor analysis of rank-ordered data: an old approach revisited. Psychol. Bull., 83, 864-867.

Yamamoto, T.
1980 Error bounds for computed eigenvalues and eigenvectors. Numerische Math., 34, 189-199.

Yates, F.
1933 The analysis of replicated experiments when the field results are incomplete. The Empire J. of experimental agriculture, 1, 129-142.

Yates, F.
1948 The analysis of contingency tables with grouping based on quantitative characters. BK, 35, 176-181.

Young, F.W., Leeuw, J. de, and Takane, Y.
1976 Regression with qualitative and quantitative variables: an alternating least squares method with optimal scaling features. PM, 41, 505-529.

Young, F.W., Takane, Y., and Leeuw, J. de.
1978 The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features. PM, 43, 279-281.

Zangwill, W.I.
1969 Nonlinear programming: a unified approach. Englewood Cliffs, Prentice Hall.