CANONICAL ANALYSIS

OF

CATEGORICAL DATA

JAN DE LEEUW

CANONICAL ANALYSIS OF CATEGORICAL DATA

Jan de Leeuw

PSYCHOLOGICAL INSTITUTE / UNIVERSITY OF LEIDEN

Postscript:

This book is a corrected version of my dissertation of February 1973. There are no major revisions. I simply corrected a large number of small errors. Nevertheless it is not a dissertation any more, and it should not be referred to as such.

There are three major developments in the area covered by this book. In the first place a large number of FORTRAN programs is being developed by Eeke van der Burg and Mike Tjok-Joe of the CRI, the computing center of the University of Leiden. A corresponding series of program abstracts will be issued separately. Anyone who is interested in these abstracts should contact the department of Data Theory. The project will probably not be finished before the end of 1974, and it is quite conceivable that the programs will be corrected and/or extended during the course of the project. The current programs, for example, all use the Jacobi method to compute the eigenvalues and eigenvectors of the symmetric matrices. This will almost certainly be replaced by a Householder-Givens routine, which uses the Sturm-sequence property to compute all eigenvalues in a predetermined interval.

The second development is statistical. For some special cases I computed a further term in the expansions given in section 6.8. This is a tedious job, which seems to produce nothing but very long and uninteresting formulas. In quite a number of cases encountered in practice the number of replications will still be too small for these formulas to apply. Moreover the methods do not work for some of the more interesting null hypotheses, basically because the validity of the expansions depends on the assumption that the eigenvalues are different. Recently I found out, however, that asymptotic inference is also possible in some areas if the order of the matrices tends to infinity. Under some rather stringent conditions on the distribution of the elements of the matrices it follows that the condensed distribution of the eigenvalues tends, with probability one, to the semi-circle law of Wigner. By the condensed distribution we mean the random function

$$W_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(\lambda_i^n, x) ,$$

with

$$\delta(\lambda_i^n, x) = \begin{cases} 1 & \text{if } \lambda_i^n < x, \\ 0 & \text{otherwise.} \end{cases}$$

The semi-circle law is an absolutely continuous ditribution function with density given by

$$w(x) = (2\pi\sigma^2)^{-1} \sqrt{(4\sigma^2 - x^2)}$$

on the interval $(-2\sigma,+2\sigma)$, and zero elsewhere. The most general results seem to be those of Arnold (1971), who proves that $W_n(x) \to W(x)$ a.s. for all x under conditions which seem quite strong, but which are actually met in some of the example we

discussed in chapter 4. There are three directions in which the results of Arnold should be extended. In the first place we would like to weaken the restrictions (notably those dealing with independence and equidistribution of the matrix elements), in the second place we would like to establish the weak convergence of the appropriately normed random function $W_n(x) - W(x)$ to some well-defined random function on the same interval, and in the third place we would like to obtain information about the asymptotic distribution of Kolmogorov-Smirnov type statistics such as a suitably normed version of $\sup_x |W_n(x) - W(x)|$.
I am not sure in how far this last program can actually be carried out, and I am certainly not sure that I can carry out even a small part of it. The results would, however, certainly be very useful, not only for some of the methods discussed in this book, but for psychometrics in general. If a large part of it could be carried out some of my reservations about the possibility of rigorous statistical inference in these explorative eigenvector-eigenvalue techniques would almost surely disappear, although I still think that the discussion in section 1.3 of this book remains valid.

The third development is in the direction of nonmetric canonical analysis. In section 3.19 we pointed out that ordinal restrictions on the weights could be incorporated rather easily. Nevertheless we still measure homogeneity in an essentially metric way. It is, however, possible to define homogeneity measures which have the property that the quantification is perfectly homogeneous iff all induced vectors of scores are monotone with each other. Define

$$t_{ik;jl} = (z_{ij} - z_{il})(z_{kj} - z_{kl}).$$

We now make the partition

Source	Coefficient
Between columns	$B = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{m} t_{ik;jl}.$
Within columns	$W = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{m} \{\lvert t_{ik;jl}\rvert - t_{ik;jl}\}.$
Total	$T = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{n}\sum_{l=1}^{m} \lvert t_{ik;jl}\rvert.$

We can maximize B/T over the weights by using gradient or alternating least squares methods, starting with the original metric solution as a first approximation. Because the between-component is essentially the same for both the metric and the nonmetric partitions we can derive some interesting relationships between the two approaches. If we want to we can compute additional solutions by requiring orthogonality of the different sets of weights, either in the sense used throughout this book, or in the sense of section 3.21.

CONTENTS

1 Data analysis in the social sciences

1.0 Introduction

This chapter is a short introduction to some of the current topics in psychometrics and data analysis in the social sciences in general. Its purpose is to justify to some extent the scaling techniques discussed in this dissertation. We do this by discussing a number of more or less recent controversies in the data analytic and related methodological literature (some of these controversies are more or less dramatized for the sake of the argument). In some cases the discussion of these controversial points is highly relevant for the evaluation of the particular class of techniques discussed in our later chapters.

1.1 Clinical and statistical psychologists

Once upon a time there was a heated controversy between 'clinical' and 'statistical' psychologists. The discussion (as far as there was any discussion) focused on such notorious problems and pseudo-problems as the possibility of measurement in the behavioural and social sciences, the superiority of either the clinical or the statistical methods of prediction, the possibility and the usefulness of an 'objective' or even 'objectivistic' approach to the social sciences, the usefulness of 'hermeneutic' and 'phenomenological' methods, the distinction between 'Naturwissenschaften' and 'Geisteswissenschaften', and the role of 'Verstehen' in the social sciences. Because we do not want to get mixed up in all of these problems we separate them into three classes: philosophical, empirical, and mathematical-statistical.

The main problem is a philosophical one. It has to do with the methodological value of understanding the behaviour of individuals and/or social institutions by getting to know the rules (and/or values) that govern this behaviour. This problem has not been satisfactory solved, but as it is a philosophical problem this is not exactly a surprise. We can agree that is not the function of science to give a 'complete' description of reality (cf Meehl 1954, p 130), and that it certainly is not the function of science to give a partial sensory of emotional reproduction of reality (cf Rudner 1966, p 69-70, p 81-83), but in the last analysis this only seems to mean that we (a number of scientists) agree about a particular definition of science (cf Black 1954, p 3-23). Other scientists agree that statistics is useless unless the persons applying it have a real understanding of the area under consideration (cf Winch 1958, p 113), and that in some areas it may be impossible to apply statistics at all.

Ultimately the controversy seems to amount to the fact that some scientists think that the hermeneutic, understanding, rule-knowing aspects of science belong to the context of discovery, which is only propedeutic to true science, and in which there is a complete methodological anarchy. Other scientists will argue that all the really important ideas belong to that context of discovery, and that the purely technological 'design' aspects of science are essentially trivial. The first group says that 'Verstehen' is important but certainly not scientific, the second group says that the results of the physicalistic approach to the social sciences are scientific, but certainly not important. At this point both groups get very angry. This makes the controversy a question of emphasis, a question of character, a question of taste, and above all, a question of two groups of scientists fighting a rather unreal fight over the ownership of the labels 'science' and 'psychology', which are, as labels, both not very important and not very scientific. It has been convincingly argued by Black (1954) that all definitions of ' scientific method' are essentially persuasive in the sense of Stevenson (1938). Wittgenstein has even defended the more general point of view that all statements of the form 'This is really only this' are persuasive. They try to give attractive explanations that destroy prejudices. This is true for Freud, for Darwin, for materialism, and also for behaviourism and physicalism. Of course if you succeed in persuading your client and this cures him, you win. In the same way if you succeed in persuading your friend that psychology is only badly formulated physics, and this solves his problems, then you win (Wittgenstein 1966, p 23-28). Lord Rutherford once said that science consists of physics and stamp collecting, but nobody believes that anymore. Einstein once said that it is not the task of science to give the taste to the soup, but innumerable psychologists still believe that the only respectable way to report experiences is to recreate them in the reader or hearer. Nobody denies that physics is a science, but for some reason or another there seem to be difficulties with psychology, history, ethics, esthetics, philosophy, and mathematics (the reasons are different in these different examples). Nevertheless all these disciplines are practiced by people who call themselves scientists, who are payed by scientific institutions for doing scientific work, and who are considered scientists by their friends, their colleagues and by lots of other people. This may be a better point to start than an Aristotelian definition of science which leads to useless and slightly ridiculous quarrels. Of course the empirical problems about the relative efficiencies of clinical and statistical prediction are completely irrelevant to these philosophical questions (although not everybody seems to realize that they are). It seems useful to analyze the clinical judgment

process as a psychological phenomenon, it seems interesting to evaluate the predictions of the clinical psychologist(as far as he wishes to make any predictions) by using statistical optimality criteria. Of course it is both circular and unfair to use these criteria to compare his predictions with the optimal statistical procedures, and to conclude that statistical procedures are better. It seems equally natural to investigate the statistical procedures and their predictions by using clinical criteria of optimality. The idea that clinical criteria and optimalizations are more complicated, and less exact, intersubjective, and rigourous is not a fact but a philosophical point of view (cf Wittgenstein 1953, section 88; Wittgenstein 1956, section I-5, section V-12). One should always ask what the predictions are going to be used for, if they are only going to be used to evaluate the efficiency of the clinician's predictions they have a limited value anyway.

From our point of view the most interesting aspects of the controversy are the arguments used by the clinical psychologist (and his colleages from the other social sciences) against the use of mathematical (especially statistical) methods. In the first place it has been argued that in (clinical) psychology it is the (behaviour of the)individual that counts, not the (average behaviour of a) group. Most statistical techniques use measures of central tendency computed for groups of people, and consequently are not useful for the (clinical) psychologist. This argument seems to be based on a misunderstanding of the nature of statistics, that can easily be explained by the heavy emphasis on descriptive statistics and associate tests and intervals in most behavioural science courses. What really counts is the statistical model for a particular situation. Even the clinical psychologist will admit that the behaviour of a single individual varies in certain aspects from one occasion to another, even if the conditions under which the behaviour is observed seem identical (and even if he does not want to speak of behaviour in certain situations). Part of this variation can be interesting and is capable of being understood, but certainly another part of it is uninteresting and influenced by so many factors that it cannot possibly be understood. This means that the behaviour of a subject (client) in a particular situation has a component which we may consider as random fluctuation, and for which a particular stochastic model may or may not be appropriate. It is, of course, true that in a large number of situations it has been erroneously assumed that a group of subjects has a typical behaviour and that the individual deviations from this pattern are merely random (with identical distributions). It is true that the psychological journals are filled with articles in which stochastic

independence is implicitely assumed, while a cursory inspection of the data (or even of the experimental conditions) shows that this is a patently false assumption. It is also true that even experimental psychologists use statistical techniques which are not appropriate in the situation in which they are applied (analysis of variance of distance measures, t-test on skew distributions, etc.). This does not exclude the fact that for some situations in the social sciences there exist respectable statistical techniques, that can even be used to describe the properties of a single individual. As a matter of fact I think that measurement of particular characteristics of an individual (or group) and statistical operations performed on these measurements can help the understanding process (I take the description of Winch as representative of this process), in the same way as the understanding process can help in the construction of statistical models (hypotheses). I do not believe that the object of psychology is, in some sense, essentially different from the objects of the other sciences. My conception of science lies somewhere between a particular way of thinking and a particular way of making a living; the nature of the object does not seem to be very relevant.

Another argument that has had some popularity is that measurement is essentially a reductive process in which the aspects which are most important for the psychologist are lost. That measurement is a form of reduction is obviously true, but so is language. An important function of language (both scientific and common) is classification, and classification is the basis of all measurement. Another important function of language, and especially of scientific language, is the ordering of objects in terms of several different attributes, and order is the basis of all higher forms of measurement. The fact that in psychology the concatenation operation, the basis for most of the advanced forms of measurement, is lacking has created a lot of confusion (Wittgenstein 1966, p 42). Recent developments in algebraic measurement theory have shown that concatenation operations can be introduced into the social sciences in a more abstract and roudabout way, and that these new concatenation operations then allow for most of the basic results of classical extensive measurement. The idea that in the measurement process not only some aspects, but the most important aspects of the phenomenon are lost is, again, not a fact but a philosophical point of view. Not my point of view. The difference between measurement and description seems slight. Saying that it is not possible to measure something does not mean a thing, we must specify what forms of measurement are impossible. If we weaken the definition of measurement in such a way that it applies to all classifications, then it is hard to see how we can speak about something without being able to measure it.

1.2 Qualitative and quantitative data

We use the term 'data' in its widest possible sense: data are the product of a classification process. The investigator has classified a number of objects (persons, things, situations, groups) according to a number of different criteria.

Example: We study a particular group of persons. Some of them are neurotic, others psychotic, and still others are sane. Moreover some of these persons are more than five feet long, others are less than three feet. Some of them are babies, others are adults, some are bright, others dull, and so on.

Example: We study the opinions of somebody about his relatives by asking him if he thinks that his mother is neurotic, that his sister is handsome, that his father is brilliant, that his brother is weird, his aunt Betty selfish, his grandmother selfish, and so on.

Example: We study reaction times of a single subject under different conditions. Under condition so- and-so his reaction time was approximately so-and-so many milliseconds, under condition this-and-that it was approximately so-and-so many milliseconds, and so on.

Example: We ask a number of subjects to rank a number of political parties with respect to some attributes such as preference, or left-right, of fanatism, or constructivety, or what have you.

It is clear from these examples that a rigorous distinction between qualitative and quantitative data is not necessary for our purposes. It is the classification process that is fundamental, and quantitative information is only one possible means to define the categories of the classification. We agree with Rasch (1966, p 3):'That science should require observations to be measurable quantities is a mistake, of course; even in physics observations may be qualitative (e.g. emission of radio-active particles observed as scintillations on a screen) as in last analysis they always are (e.g. reading off a point as located beween two marks on a measuring rod).' Recent developments in measurement theory have shown that supposing an attribute to be measurable (in the classical sense) is essentially a scientific hypotheses about the kind of data that can be obtained and about the structure of the universe of discourse (the objects that can be classified using this attribute). The classical physical continua suppose that a weak order canbe defined over the universe, and that a group operation consistent with the order can be applied to all pairs of objects. Most of the psychological continua (pitch, brightness, loudness, saturation, hue, etc) satisfy these axioms only approximately, and most psychological

attributes do not satisfy them at all. Sometimes there is only order and no group structure, sometimes there is not even order. We only suppose in this dissertation that classification can be unambiguously performed, i.e. a number of equivalence relations (or partitionings) are defined on the universe of classifiable objects. There has been a certain controversy in the social sciences about the value of qualitative data (nominal scales, classifications, etc). The dominant influence of experimental psychology in psychometrics, nicely illustrated in the book of Guilford (1954), has led people into thinking that data must be at least ordinal in order to be scalable. This idea is predominant, even in the more recent books of Torgerson (1958) and Coombs (1964). The psychophysical scaling methods require qualitative judgments, but these can be translated directly into (probabilistic) ordinal judgments about the underlying scale. Experimental psychologists and the methodologists inspired by experimental psychology have (or had) a tendency to ignore and belittle the more diffuse and less structured data collected by other social scientists. This attitude is somewhat beside the point. 'Areas in which measurement, and inferential classification in general, is in terms of 'weaker' relational systems are called 'soft' by some psychologists. Its usual use as a term of opprobrium for the area is quite unfair. Work in such areas need not be any less rigorous or scientific; truth statements in such weaker systems would merely tend to be less specific' (Coombs 1964, p 338). From a statistical point of view this quotation can be translated as saying that the statistical models in such 'soft' areas would tend to have more parameters (as a consequence we usually need larger samples). It would not make sense (historically) to maintain that the influence of experimental psychology on psychometry (or data analysis for the social sciences) has been harmful, but some of the habits and prejudices that survive from these earlier periods are certainly harmful and must be abandoned (the same thing is true for the structured labyrinth of bad habits called psychological factor analysis).

In multivariate analysis (especially as practised by psychometricians) there is the same tendency to downgrade the endless lists of cross tables and associate measures of classification (sometimes ten different measures for each table). Again the attitude is understandable, but harmful and unnecessary. Multivariate statistical analysis has been dominated by the multinormal distribution for a number of obvious reasons. For multinormal distributions some standard statistical optimization problems could be solved relatively easily, the algebra seemed relatively simple, and the

multinormal distribution could be applied to a very large number of data. Nevertheless not a single statistician has ever maintained that multivariate analysis is essentially multinormal, and essentially limited to interval scale data. A somewhat irritated quotation of Guttman from 1944 (I repeat:1944) is relevant here. 'It is only that most of us have been exposed exclusively to certain algebraic manipulations that we conceive such manipulations to be the essence of mathematics. A more sophisticated view is to regard mathematics as unveiling necessary relationships that arise from classifications. Much useless discussion of mathematics as a 'tool' in social research could be saved by recognition of the fact that qualitative classifications lead to just as rigourous implications as do quantitative'.(Guttman 1944, p 193). Although it is true that looking at isolated cross tables from a large body of multivariate data can be misleading and not very instructive, it certainly is not true that nominal data cannot be analyzed by multivariate techniques. In fact twenty-two years later, after the Shepard-Guttman computational breakthrough, Guttman could remark (triumphantly):'In order to comprehend great complexities, it proves to be effective and powerful to focus only on most qualitative features; from these can be derived actual metric consequences, with no special assumptions.'(Guttman 1966, p 495). In the next few sections of this chapter we shall discuss some of the techniques that can be used, both for explorative data analysis and for confirmatory analysis.

1.3 Data analysis and statistics

It is difficult to give a precise definition of the area of data analysis. In his fundamental paper Tukey lists the following things which must be included:'...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'.(Tukey 1962, p 2). Consequently he also thinks that statistics has two different roles. 'As far as statistics applies to real data it can be judged by the standards of data analysis, as far as it does not apply to real data it must be critisized according to the standards op pure mathematics.'(l.c., p 3). At other places Tukey describes the inferential (or confirmatory), the incisive (or exploratory), and the allocative (or design) aspects of data analysis. Tukey's conception of data analysis as an independent branch of science with it own (more or less explicit) standards has not received the attention it deserves.

Thinking about data analysis as an independent branch of science is an important liberalization, especially for the social sciences. I consider it a healthy modification, Tukey considers it a healthy modification, but a number of statisticians may very well consider it as a first step on the road to complete anarchy.

What I mean by liberalization is, of course, that the problems of mathematical statistics can be rigorously formulated as optimization problems and the solutions offered by the statisticians are optimal solutions to these problems (UMP tests, UMVU estimates, etc). In other cases (due to computational difficulties) the solutions are only approximate, but we have a general idea how good they are in terms of the original optimization problem. Notably there are a number of solutions which are suboptimal in finite samples, but which approach the optimal solution when the sample size tends to infinity (ML theory, LR theory). In the exploratory parts of data analysis, however, the well-defined optimization problems are missing, the criteria are vague, and the attempts at optimizing something (anything!) are not rigorous. Moreover the stochastic basis on which any statistical procedure is built, is only implicit (to put it mildly). It is consequently not very difficult to understand that statisticians frown upon Kaiser's little Jiffy & associated root-staring procedures, upon the hundreds of 'objective' rotational criteria, upon analysis of variance applied in situations where the assumptions are obviously violated. A somewhat irritated quotation from Kendall & Stuart (1966,III, p 311) illustrates the point. In their discussion of factor analysis they say: 'Once again the electronic computer has come to the aid of psychologists by enabling them to specify sundry criteria to determine rotations or structural simplification and solve the resulting equations, but even the computer may find it hard to provide accurate information about the sampling distributions of the resulting estimators'.
From the point of view of data analysis this is somewhat beside the point. The rotational criteria are not necessarily used for estimation purposes, and sampling distributions are only one of the many tools of the data analyst. The image of the dumb social scientist and his dangerous weapon, the electronic computer, is somewhat misleading.
In defence of data analysis we must also say that the objectivistic approach has the problem that different criteria, all desirable, lead to different solutions. This is called the 'objectivist's dilemma' by Schaafsma (1971) and it seems to be especially serious if we try to generalize the Neyman-Pearson approach (or the general decision theoretical approach) to multiple decision problems. Now multiple decision problems are formulated because the

Neyman-Pearson hypothesis testing situation is not very realistic, and not very representative for the diverse things that are actually done with data. It is consequently quite possible that the objectivist's dilemma will turn into an objectivist's nightmare if he tries to apply his objective criteria to the complex, multivariate exploration situations which are very common in the social sciences.

On the other hand it is important that the fact that it is still not possible to justify, say, multidimensional scaling in a statistically meaningful way does not imply that the technique has no value. In fact it does not even imply that the technique is incomplete in some sense. Usually MDS is not used for inferential purposes and the statistical questions are simply not relevant. If MDS tries to argue from the sample to the population it is a very incomplete and non-rigorous statistical technique, which can consequently make only relatively imprecise and unreliable statements about the population parameters. If it is used as a transformation which makes a given data set more easy to understand, and which gives some indications about further (possibly confirmatory, statistical) research, then MDS has already proved to work very satisfactory. Although it seems to be true that the distinction between good and bad procedures in data analysis is usually more difficult to draw than the distinction between good and bad procedures in statistics, this mainly reflects the fact that the emphasis in data analysis is more on the usefulness of procedures and not on their optimality. There are no cut-and-dried criteria for usefulness, and, in fact, in data analysis the user and his habits have a very strong and often frustrating influence on the procedures that are going to be used in particular instances (even if the data analyst prescribes other procedures which are better according to his criteria).

How does this discussion apply to our problem of the multivariate analysis of categorical data. The first investigators who applied techniques of the kind we propose were somewhat apologetic. '...if the statistician has not as yet developed an appropriate method which he can offer us, then the psychologist, however imperfectly, must set to work to devise his own.' (Bird, 1950, p 168). As usual Guttman is somewhat more self-assured: 'We cannot even begin to tackle sampling problems until we define what the best answer would be for the population, for the case where there are no sampling errors.' (Guttman 1941, p 341). We can be less apologetic for several reasons. In the first place it is possible and essentially trivial to give asymptotic sampling distributions for our 'estimates' using nothing but the assumption of independent, identically distributed

successive observations (this could easily be relaxed to Markov-dependence, as long as the CLT applies to the transition counts). The formula's are complicated (cf section 6.8) but once again the computer comes to the aid of the psychologist and computes the standard errors of his estimators. Nevertheless it must be emphasized that the statistical models used are very primitive, and that the computed approximate standard errors are only a mechanical type of additional output that may or may not be used in the exploration. This is important because since the early fifties the statisticians have developed a number of techniques for the multivariate analysis of categorical data that use more explicit probability models and concentrate on the same basic notions as we do (heterogeneity, independence, interaction). The most important references are Mitra (1956), Roy & Mitra (1956), Roy and Bhapkar (1960), Bhapkar (1961, 1966, 1969), Birch (1963, 1964, 1965), Good (1963), Lindley (1964), Caussinus (1965), Benzécri (1969), Kullback, Kupperman, & Ku (1962), Ku, Varner, & Kullback (1971), Plackett (1969), Berkson (1968), Bock (1969), Grizzle, Starmer, & Koch (1969), and especially Goodman (1968, 1970, 1971, 1972). The main disadvantage of this class of methods is that they are asymptotic as well, and use counts in all cells of the multidimensional table. If we have more than three or four variables with a moderate number of categories each this means that the number of observations has to be enormous (at least for some higher-order hypotheses). Our methods only use bivariate and univariate marginals, and these marginals are, of course, always larger than the frequencies in the body of the table. In our examples in chapter 7 we show that even if the 'proper' multivariate techniques can be applied, our procedures still can give useful additional information.

For the general problem of multivariate categorical data analysis a number of multidimensional scaling techniques have been developed by Guttman (reported by Lingoes 1968), and by De Leeuw (1969). They are based upon the pseudo-topological notions of contiguity and separation, respectively. These techniques are data analytic in an extreme sense. Minimization of the loss function (measures of, respectively, contiguity of the representation, and smoothness of the separating boundaries) requires heavy gradient-type computation, and sampling distributions are not even mentioned. There may be serious local minimum trouble, the uniqueness problem (admissible transformations) is complicated, and the very weak nonmetric requirements tend to produce degenerate solutions. Nevertheless Guttman's GL-MSA-I has been successfully applied to a number of examples, and it certainly is based on some beautiful ideas. It is clear that there is a lot of space on the

data analysis-statistics continuum between these two approaches. The procedures in this dissertation have an intermediate position, and can be used in a 'tandem' approachwhich considers data from a number of possible angles, and applies techniques from different classes (some of the examples in chapter 7 are analyzed in such a way).

1.4 Data reduction and measurement theory

Another controversy, which is closely related to the previous one, and which is of some independent interest, is the controversy between data reductionists and measurement theoreticians. A quotation from Krantz, who discusses multidimensional scaling, illustrates the point. Since Shepard's computational breakthrough 'this sort of measurement has been widely practiced, with little concern over appropriate foundations.' (Krantz,1967, p.14). By foundations Krantz means, of course, the approach to MDS based on axiomatic ordinal characterizations of Minkovski spaces or geodetic spaces. The question is, however, in how far these foundations are appropriate for the social scientist, and in what circumstances. The data reductionist approach to scaling (or the data analytic approach, which is somewhat less narrow) does not fit an explicit algebraic or stochastic model to the data by estimating free parameters, but simply tries to replace a lot of unstructured input data by a smaller amount of possibly more structured output data (without too much loss of information, or with maximum reproducibility). This data reduction point of view has been defended, with varying degrees of extremity, by people like Shepard, Torgerson, and Guttman. Even Coombs tends more to the data reductionist than to the measurement theoretical point of view in his major work. Torgerson has argued, quite convincingly, that the new MDS techniques 'will become most useful in those very areas where a purist could argue that they should not have been used at all.' (Torgerson, 1965, p 393). And indeed MDS solutions are sometimes most informative if the space turns out to be partially empty, if there are some well-defined regions in which the solution suggests that there cannot be any points (possible stimuli in the domain). This is true for the colour circle and for Dutch political parties, but even more so for solutions which show us that some of the dimensions are partly or completely qualitative. As Guttman has argued (especially since 1950, when he started working on the radex) one must always emphasize the systematic structural characteristics of the configuration and not projections on arbitrary dimensions (this is another bad habit left over from the factor analysis days). Guttman, who has defended the data reductionist

point of view in the most extreme way I know of, refuses to have anything to do with G-spaces, untestable axioms, and the like. He is also interested in representation theorems, it is true, but only for the finite case, and only as displaying a kind of perfect fit that is either trivial or nonrealistic or both (cf Guttman 1967). I think this attitude is somewhat too extreme. After all our MDS solutions may indicate that a well-filled psychological space is behind the whole thing. At some place in our theory formation process there may come a time when assumptions like solvability are necessary or convenient. At some place we may want to postulate a definite model of which solvability is a part. That this is not often the case, and that the naive data reduction point of view is still the most fruitful in MDS and related areas is clear from the testimony of people like Guttman, Torgerson, Shepard, and Coombs, who have a lot of experience with real psychological and sociological data.

The measurement theoretical 'purist' is undoubtedly more right in the case of simple polynomial models (such as the additive one). As soon as one can imagine psychological continua, one can often also imagine additive combination laws that make sense, and that have most of the properties of their classical physical brethren. In such simple additive situations we shall indeed find that even our explorative techniques and the statistics computed by our techniques have relatively straightforward statistical interpretations. In more complicated situations (in which less structure can be assumed) the interpretation in terms of a model becomes more doubtful, and we must interpret the results using other data analytic criteria. As a kind of preliminary evaluation I would consider the claims of the measurement theorist as not valid. There is no doubt that additive conjoint measurement and other simple polynomial models are a valuable addition to the tool kit of the experimental psychologist (and the social psychologist). This is sufficiently proved by the excellent papers of Tversky, Coombs and others on decision making and risk taking. But ACM seems to require controlled experimentation and balanced designs, and we have seen that techniques which require that much have a limited value for the social sciences in general. The same thing is true for the studies in multidimensional psychophysics of Krants, Tversky, Wender, and others. Again this is limited to 'uninteresting' stimulus domains with obvious dimensions, such as geometrical figures. And even in these artificial or controlled contexts the thing that really counts is the statistical model that is assumed, not a purely algebraic analysis. The axioms are now merely qualitative consequences of the overall statistical model, which can be used to test the model.

It may be the case that measurement theory loses a lot of its 'exact' or 'rigorous' appeal, if translated into these simple stochastic terms. The translation of ACM into two-way ANOVA, model I, with an arbitrary monotone transformation of the cell entries, shows what can be expected. Of course we can test this model by deriving monotonicity in rows and columns, and the various cancellation conditions, but the only advantage of this is that we can use nonparametric methods more easily. It has been argued by Krantz (1968) that algebraic measurement theory has only a philosophical relevance for the physicist, but a very real scientific value for the social scientist. This only seems to be true in a limited number of cases; in most situations algebraic measurement theory belongs to the philosophy of science for the social scientist as well.
This discussion is closely related to a distinction made by Coombs (1966). Scaling techniques can be used in two different ways. Although the difference does not seem te be very essential, we discuss it briefly. We remember that, according to Coombs, a scaling theory is a triple consisting of a measurement theory, an error theory, and an algorithm (it follows from the above discussion that the procedures discussed in this dissertation are not scaling theories, and that I think that the concept of a scaling theory is of limited value). In the first case a scaling theory can be used, according to Coombs, as a technique. It is assumed (more or less explicitely) that the deviations from the requirements of the measurement theory are errors, and a function of the errors is minimized. The main objective is data reduction, emphasis is on the algorithm. A different use of scaling theories is as criteria or tests of the measurement theory as a descriptive model. This implies a different way of looking at errors, and an emphasis on the measurement theory. In general we may use a somewhat different algorithm in the two cases, although some scaling theories may be used for both purposes. One of the main differences between a scaling technique and a scaling criterion is their degree of vulnerability (Coombs 1964, p 72-73, 82). A less vulnerable technique yields a 'best' representation, no matter how bad the data are from the measurement theoretical point of view. A highly vulnerable technique will break down in these cases. As Coombs (1964, p 81) points out, there is a certain connection between vulnerability and the theory of type I and type II errors in the Neyman-Pearson theory. A highly vulnerable technique will increase the likelihood of rejecting the hypothesis while it is true, a stronger (more robust) method of scaling will tend to increase the type II error and decrease the type I error. We have argued in the previous pages that

for all practical purposes we can replace 'measurement theory' by 'statistical model' in this discussion. The Coombsian distinction between the two cases can now be directly translated into the data analytic concepts 'exploration' and 'confirmation' with a considerable gain in clarity. It is now obvious that exploratory techniques, interpreted in terms of statistical models, have low power, while confirmatory techniques which use all the specific properties of the model have higher power. It is also obvious that highly specific confirmatory techniques break down (in terms of power) if the assumptions of the model are seriously violated. This is not exactly what Coombs had in mind. His vulnerable techniques are the qualitative scaling methods discussed in his theory of data which became out of date since the computational breakthrough.
It is well known that most of the problems of inferential statistics can be formalized as decision problems. In De Leeuw (1971) we have tried to liberalize this concept for data analytic purposes. We introduced the concept of a scaling theory in a more or less Coombsian way (without assuming the existence of an explicit measurement theory). The difference with a standard decision problem is that there may be no stochastic structure. This makes it difficult to order scaling theories in terms of a natural performance criterion such as expected loss or risk. Algebraic measurement theory is used only to describe the region of perfect fit, i.e. in the Guttman sense. This may be a useful approach when it is completely worked out. The procedures discussed in this dissertation are scaling theories in this new sense, but they are very incomplete both as Coombsian scaling theories and as classical decision problems. It is true, however, that at least some of them can be interpreted in terms of algebraic or geometrical measurement models, and in terms of definite statistical models. Of course every data analyst is only too glad when his procedures allow for a multitude of possible interpretations, although this may be bad for their power in certain specific situations.

1.5 Some minor controversies

In this final section we discuss some controversies which are essentially technical and really not very important. They correspond with three possible options for the user. He can treat his data metrically or non-metrically, he can use (multi)normal or non-parametric statistical (data analytic) techniques, and he can choose a bivariate or multivariate treatment. The first choice has provoked a considerable amount of discussion in the past, the second choice should provoke a considerable amount of discussion among social scientists but does not do so because of the

deplorable treatment of nonparametric statistics in most textbooks and the fact that multinomial multivariate analysis is not treated at all, and the third choice may provoke some discussion in the near future when people come to realize that program packages like CROSSTABS have a limited data analytic value.

There has been a lot of confusion about the relative usefulness of metric and nonmetric methods. The nonmetric methods (originally developed mainly by Guttman and Coombs) use only the ordinal properties of the data. Since Shepard's computational breakthrough (Shepard 1962a, b) it proved to be possible to derive satisfactory metric representations from ordinal data in a number of important cases. As outlined by Guttman (1967) the nonmetric methods met with a considerable amount of hostility in the early days (in fact Guttman reports triumphantly that one of his earlier papers on the subject was turned down by Psychometrika!). This has radically changed since the publications of Shepard, and Kruskal (1964a, b). Since then large numbers of very successful applications of especially MDS keep appearing in the journals. In methodological circles the enormous initial enthusiasm for these very powerful data analysis techniques begins to level off somewhat. The idea that this was all we needed for similarity data turned out to be too optimistic. As indicated already by Torgerson, one of the pioneers, the nonmetric methods have lots of advantages over the older metric methods, but sometimes they throw away information that cannot be dispensed with. The same thing applies, with even more force, to nonmetric multidimensional techniques for factor analysis, unfolding, component analysis, and scalogram analysis. In our general approach to measurement (or quantification) the choice between numerical, ordinal, or nominal treatment of the data is not very fundamental. All data are categorical, and prior numerical and ordinal information may or may not be used in the analysis. The techniques do not really change if we use this extra information, the choice is now completely in the hands of the user, which is where it belongs. The arguments for using metric/ordinal information if it is at all present are summarized by Coombs (1964, p 284). Most of these arguments are definitely out of date. Ease of computing is still of some importance, however, mainly because using metric information can reduce the storage requirements considerable, and this may be important for large data sets from survey analysis, for example. The argument that numerical results are in a form which is easier to communicate applies only to the pure (coordinate-free) nonmetric scaling methods developed by Coombs and his co-workers, and these techniques are almost completely replaced by GL-SSA, GL-MSA, MINISSA, MDSCAL, KIST, POLYCON and similar programs which generate numerical results from ordinal assumptions. The final

objection mentioned by Coombs is that using or not using the metric information gives no essential differences in the results. This argument has some popularity with people accustomed to the standard multivariate (metric) techniques. Nevertheless the idea that the results will not differ much if the model is approximately true is false (and even if it was not false, the argument would be invalid, because only nonmetric techniques can prove, in some sense, that the assumptions of metric techniques are correct). There are lots of examples from the MDS area in which using only the ordinal information in the data gives much more satisfactory representations than using the full numerical information (cf Guttman 1966). The results of Torgerson (1965) shows that the contrary case can also occur: sometimes use of the metric information improves the interpretability of the results. In our class of techniques using or not using the metric information does not change the technique, but it certainly changes the way of looking at the output. The conclusion is that we have to be careful about two bad habits which have or had some popularity. The first one is to replace nominal or ordinal information automatically by numerical information (for example by using integer scores), and the second one is to replace numerical information automatically by ordinal information. In the last case we throw away information that may be useful (or even vital), in the first case we use information that is not even there (we invent information). In some instances this may have no damaging effects at all, in other examples it can be quite misleading.

The prior information can be of another type, we may for example also know or suspect that some of the variables are approximately normally distributed in the population from which we are sampling. This information can be used to apply a more specific class of techniques than the one we discuss. In fact we can apply ML and LR methods for the multinormal distribution right away. In De Leeuw (1972) these methods are investigated from a data analytical point of view, and their exploratory aspects are emphasized. Again we may or may not have the prior information, and if we have it we may or may not use it. The choice is closely related to the choice between parametric and non-parametric methods in inferential statistics. It depends on our confidence in the multinormality of our data, but also on the questions we want to ask. The same types of errors can be made as before. We may apply multinormal techniques while they are not appropriate, and we may refuse to use our multinormal information where it could have been useful.

We have already mentioned the endless lists of cross tables in section

1.2. They stand for an essentially bivariate treatment of multivariate categorical data, corresponding with looking at a large correlation matrix in the multivariate numerical case. The point of Guttman's order analysis (1966b) is that lots of useful things can be learned by merely looking at patterns in correlation matrices, but sometimes there simply are no patterns, and all we see is a large number of pairwise relationships which are very difficult to integrate into a more complete picture of the data. Finding such patterns in large sets of bivariate contingency tables is also possible in principle, but it will be even more difficult and it can be even more misleading. In this dissertation we study the joint bivariate treatment of general categorical data, corresponding with the multinormal analysis of means and dispersions. There is, however, an essential difference. In multinormal analysis joint bivariate analysis is essentially multivariate analysis, all information about the distributions is given by the first and second moments. For general categorical data we ought to consider higher order tables too in order to speak of multivariate analysis, and this is exactly what is done in the methods mentioned in 1.3. In this sense our joint bivariate analysis, which uses the same data as the 'isolated' bivariate approach, but combines them into a single joint analysis, is somewhere between classical multinormal analysis and a CROSSTAB type of analysis. We expect that the methods will be useful both for the highly structured data sets of experimental psychology where they can supplement the statistically more direct (multinormal or multinomial) multivariate techniques and for the large, mixed data sets from survey analysis, where they can supplement the descriptive cross table analysis.

2. Indicator matrices and quantification

2.0 Notation

Capitals will be used for matrices and supermatrices, roman characters for vectors and supervectors. Greek characters will be used for scalars, except for integer constants which are denoted by k,l,m,n. Elements of the vector x will be written as x_1, x_2, ..., x_n; elements of the matrix H by h_{11}, h_{12}, ..., h_{nm} (the first subscript referring to rows, the second one to columns). If a supervector x consists of n subvectors, we write x_1, x_2, ..., x_n for these subvectors (we always take care that there is no confusion possible with elements of ordinary vectors). For supermatrices H and C we use H_i and C_{ij} for submatrices. The elements of submatrix C_{ij} are written as c_{kl}^{ij}, of subvector x_i as x_l^i. We use a prime to denote transposition, column vectors are written without a prime, row vectors with a prime. As special symbols we use I for the unit matrix, and E for a matrix with all its elements equal to unity. A vector with all its elements equal to unity will be written as e, the number of elements in e (and the order/dimensions of I and E) will always be clear from the context. A vector with all its elements equal to zero, except element i which is unity, will be written as e_i. Some abbreviations are iff for *if and only if*, wlg for *without loss of generality*, dfr for *degrees of freedom*, and df for *distribution function.*

2.1 Indicator matrices.

We shall be concerned with random samples of size m from p-variate populations. Because we deal with finite samples we can assume without loss of generality that the p random variables ϕ_i assume only a finite number of different values, even if the underlying 'population' variables are really defined in such a way that their range is a real interval. Moreover we also suppose that the m sample elements are classified according to n-p deterministic criteria; i.e. our sample is structured or stratified. Consequently there is a total of (n-p)+p=n finite sets T_i with cardinalities k_i. We record our multivariate observations in a supermatrix H of dimension Σk_i x m (= K x m, say). H is obtained by superimposing the n matrices H_i of dimension k_i x m, with $h_{lj}^i = 1$ iff variable ϕ_i assumes the value s_l^i for sample element j, and $h_{lj}^i = 0$ otherwise. H is called the *indicator matrix*, the H_i are *marginal* indicator matrices. Lingoes refers to H as the *attribute* or *trait matrix*, but I don't like these terms because they suggest particular applications

(although Yule already used the term attribute in this context in a more abstract sense). In Guttman's terminology the sets T_i are called facets, and the elements s_1^i structs. The sample elements (from ΠT_i) are called structuples.

A useful distinction in this context is the one between variates and factors. If a certain variable is a factor this simply means that the classification of the sample elements on this variable is not determined by the outcome of the experiment, is apriori, a factor is an independent variable. The classification of the sample elements on the variates is a posteriori, a variate is a dependent variable. The main point is, of course, that the variation of a variate is random, the variation of a factor or way of classification is nonrandom. This distinction plays a major role in multivariate multinomial and multinormal analysis. Instead of using facet as a single term for both vectors and variates, we use the more familiar word variable, without specifying whether it is a dependent or an independent variable we are talking about. Instead of structs we use categories.

2.2 Types of variables.

It seems as if we are dealing here exclusively with so-called nominal variables. And, indeed, the matrix H does not tell us whether there is any extra information beside the purely nominal manifold classification. Nevertheless it is quite possible to incorporate numerical variables into the analysis. If ϕ_i is numerical we simply have a k_i-element vector y_i of real numbers as extra information. If ϕ_i is ordinal, we have a different type of additional information: a partial order $<_i$ on T_i. Consequently our interpretation of multivariate measurement as manifold classification takes the nominal variables as basic (as it should do: all measurement is based, in the last analysis, on quantitative judgments, cf chapter 1). The fact that we may treat some of the variables as numerical or ordinal is, indeed, additional information which has nothing to do with the present classification but is based on prior knowledge. To put it differently: all variables are categorical, and among the things which can be used to define categories are numbers and order relations. I have previously used the word relational (in stead of categorical) to describe this type of data, because we basically study the 'belonging'-relation which partitions the sample n times (and the matrix H is the indicator of this relation, portrayed in matrix form). If $k_i = 2$ we

shall follow the conventional usage by calling ϕ_j a binary variable, inducing a twofold classification (dichotomous variable and dichotomy are also quite popular terms). In a sense the binary case is the basic one: we can always reduce a variable with k_j categories to k_j variables with two categories. Observe that we have implicitly assumed that all variables have categories which are mutually exclusive and exhaustive. This is no real restriction of generality. We can always translate non-exclusive cases into exclusive ones, and in most practical situations I can think of the so-called non-exclusive categories are simply compound binary ones in which only the 'positive' responses are recorded. Alternative terms have been proposed by Burt, who borrowed them from the logician W.E. Johnson. He calls variables determinables, categories determinates. Another possibility, which we must also reject because it is suggestive of a limited area of application, would be to use item and alternative. We do adopt Yule's term manifold classification for the process which produces the supermatrix H.

2.3 Examples.

One of the most obvious examples of such a matrix H is the one suggested by the use items for variables and alternatives for categories. A multiple choice test can obviously be scored like this. A questionaire with yes - no responses is a set of binary variables, a survey with both numerical (income, age, number of children), nominal (sex, religion, profession), and ordinal (attitude items with categories like fully agree - agree -disagree - completely disagree) variables is a more complicated mixed example. A test battery usually involves a number of numerical variables (scores for subtests), although in the last analysis these are often based on binary variables (the score is the number of correct items in a subtest). More complicated examples are paired comparisons (each pair defines a binary variable), and factorial ANOVA with a categorical dependent variable. In the data used for ordinary discriminant analysis we have one binary factor and a number of numerical variables, in canonical discriminant analysis we have a nominal factor together with a number of numerical variables. All these situations can be scored as categorical data in an indicator matrix, and in some of them we have additional information. There are, of course, more efficient ways to portray our information. If we deal with n numerical variables we can display our data as an n x m matrix of real numbers, which contains all the information in H plus the additional numerical information (the reduced matrix). A binary variable can be recorded in this reduced

matrix as a row of zeroes and ones, a nominal variable by simply numbering the categories, an ordinal variable by embedding the categories order-isomophically into the reals (supposing, of course, that we are dealing with a weak order over the categories). Because the reduced matrix is entirely numerical we only need to remember what rows refer to nominal variables, what rows to ordinal ones, and so on. We do not use H in actual computation, only for theoretical purposes.

2.4 Quantification

A direct quantification of the possible values of variable i is a real k_i-element vector x_i. Every direct quantification of T_i defines an induced quantification z_i of the m sample elements by the rule

$$z_j^i = \sum_{l=1}^{k_i} x_l^i \, h_{lj}^i. \qquad (1)$$

Observe that this merely implies that we replace every category $s_l^i \in T_i$ by a real number x_l^i. Observe that the direct quantification of all facets produces n induced quantifications of the sample elements.

A direct quantification of the sample elements is a real m-element vector z, which defines an induced quantification x_i of T_i by the rule

$$x_l^i = \sum_{j=1}^{m} z_j \, h_{lj}^i. \qquad (2)$$

Observe that a direct quantification of the sample elements produces induced quantifications of all variables. If we have direct quantifications x_i of the T_i then the n x m matrix

$$Z = \{ z_j^i \} = \left\{ \sum_{l=1}^{k_i} x_l^i \, h_{lj}^i \right\}. \qquad (3)$$

is called the induced matrix of scores. If we have a direct quantification z of the sample elements, then the supervector

$$x = \{x_l^i\} = \left\{ \sum_{j=1}^{m} z_j \, h_{lj}^i \right\}. \qquad (4)$$

is called the induced vector of weights. Consequently we may weight directly, which defines induced scores by (1). And we may score directly which defines induced weights by (2). To preserve symmetry we also define an induced vector of scores, which simply contains the column-sums of Z, the induced matrix of scores. Verbally: the induced score of a sample element is the sum of the weights of the categories it is in, the induced weight of a category is the sum of the scores of the sample elements in that category.

2.5. Historical

This particular way of scoring categorical data is due, indepently, to Guttman (1941) and Burt (1950). The term indicator matrix was coined by De Leeuw (1968), attribute and trait matrix were used by Lingoes (1968). Yule's contributions are contained in his famous textbook (1910), the facet terminology of Guttman is explained, for example, in Foa (1965) and Wish (1965). The distinction between ways of classification and variates is familiar from (factorial) analysis of variance, and from the analysis of 2 x 2 tables where we distinguish the double dichotomy and the comparatative trial tables. For the general case see, for example, Roy and Mitra (1956) or Bhapkar and Koch (1969). The parallel distinction between dependent and independent variables is familiar from regression analysis, and consequently these terms have a numerical bias. The two dual ways of quantifying are due to Guttman (1941), as are the terms 'weights' and 'scores'. The interpretation of numerical and ordinal variates in this framework is more or less explicitely contained in the work of Lingoes (1963, 1964) and Guttman (1959).

3 Principal component analysis

3.0 Introduction

As I have already outlined in the previous chapters one of the main objects of this dissertation is to extent all classical explorative multivariate techniques in such a way that they also apply to nominal variables. We start with principal component analysis (PCA), which is the most popular as well as the simplest of these techniques. It involves no partitioning of the variables in special subsets, no partitioning of the sample elements, only straightforward data reduction. In the terminology of section 2.1 all variables are variates, there are no factors. We want to explain the variance of a larger set of numbers by a smaller set of numbers with nicer properties.

3.1 Homogeneity

It is clear that we achieve a maximum amount of data reduction if we can show that all variables measure essentially the same property, or operationally, if we can assign weights to the categories in such a way that all n induced vectors of scores are identical. We call this maximum homogeneity. One way to measure homogeneity is by using concepts and notation borrowed from the analysis of variance (ANOVA, for short). The familiar decomposition

$$z_{ij} = z_{..} + (z_{i.} - z_{..}) + (z_{.j} - z_{..}) + \\ + (z_{ij} - z_{i.} - z_{.j} + z_{..}), \tag{1a}$$

can be simplified because we are not interested in the means of the induced scores for each variable. Thus we may choose our weights in such a way that $z_{i.} = 0$ for all $i=1,\ldots,n$ (and consequently also $z_{..} = 0$). The decomposition becomes

$$z_{ij} = z_{.j} + (z_{ij} - z_{.j}), \tag{1b}$$

and the corresponding ANOVA table is

Source	Sum of Squares
Between columns	$B = n \sum_{j=1}^{m} z_{.j}^2$
Within columns	$W = \sum_{i=1}^{n} \sum_{j=1}^{m} (z_{ij} - z_{.j})^2$
Total	$T = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij}^2$

(2)

We could use B and W as measures of homogeneity or heterogeneity, but

because $x_i=0$ for all i would make B,W, and T equal to zero, we measure homogeneity by the correlation ratio

$$\lambda = \frac{B}{T} \tag{3}$$

(which is scale-free in the sense that multiplying all x_i by a constant does not change λ). From (2) always $0 \leq \lambda \leq 1$, and (if $T \neq 0$)

$$\lambda = 0 \text{ iff } B = 0 \text{ iff } W = T \text{ iff } z_{.j} = z_{..} \text{ for all } j, \tag{4a}$$

$$\lambda = 1 \text{ iff } W = 0 \text{ iff } B = T \text{ iff } z_{ij} = z_{.j} \text{ for all } i,j. \tag{4b}$$

The usual variance-ratio would be

$$F = \frac{m(n-1)}{m-1} \frac{B}{W} = \frac{m(n-1)}{m-1} \frac{\lambda}{1-\lambda} \tag{5}$$

and consequently maximizing λ means maximizing F (as well as minimizing $W/B = (1-\lambda)/\lambda$ or $W/T = 1 - \lambda$). It can also be seen from (4) that maximizing the homogeneity of the induced row scores is equivalent to minimizing the homogeneity of induced column means (i.e. maximizing the difference between the m column means). Observe that

$$F < 1 \quad \text{iff } \lambda < \frac{m-1}{mn-1}, \tag{6}$$

while $t_m = \frac{m-1}{mn-1}$ is a bounded increasing sequence in m for fixed n, with limit 1/n.

3.2 Matrix formulation

If we collect the x_i in the K-element supervector x, the matrices

$$C_{ij} = H_i H_j' \tag{7}$$

in the K x K supermatrix C, and the n diagonal blocks C_{ii} in the K x K diagonal matrix D, then our homogeneity measure can be written simply as

$$\lambda = \frac{x'Cx}{nx'Dx}, \tag{8}$$

where we assume, of course, that x satisfies

$$e'D_i x_i = e'C_{ii} x_i = 0 \tag{9}$$

for all $i = 1, \ldots, n$. We have to maximize (8) over all K-element vectors x satisfying (9). The stationary equations for the maximization problem without the restriction (9) are

$$C x = n \lambda D x. \tag{10}$$

It is easy to see that $x = e$ is a solution of (10) with corresponding $\lambda = 1$, but $x = e$ obviously does not satisfy (9). If we remove this 'improper' solution by Hotelling deflation we find that the remaining solutions of (10) are the solutions of

$$\bar{C} x \equiv \left[C - \frac{1}{n} Dee'D\right] x = n\lambda D x, \tag{11}$$

and if $\lambda \neq 0$, they automatically satisfy (9). It follows that the stationary values of (8) over all x satisfying (9) are the solutions of (11).

3.3 The number of solutions

For each H_i the last row is completely determined by the other rows. The equations (10) have, therefore, in general $v \leq \min(K - n, m)$ solutions for which λ is positive. Each solution defines a different supervector of weights and, consequently, a different induced matrix of scores. The problem 'when to stop factoring' is analogous to the similar problem in ordinary PCA. In fact psychometric considerations suggest a similar bound as in PCA: we have positive generalizability as long as $\lambda > 1/n$. This follows from applying Cronbach's coefficient α in this context. We have

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{x'Dx}{x'Cx}\right) = \left(\frac{n}{n-1}\right)\left(\frac{n\lambda - 1}{n\lambda}\right), \tag{12}$$

where the rows of Z are interpreted as n parts of a test, whose total score is the induced vector of scores (the column sums of Z). It follows from (12), by the way, that we can also interprete the PCA procedure as a maximization (over weights) of the generalizability. It follows that

$$\alpha \geq 0 \text{ iff } \lambda \geq 1/n \tag{13}$$

(observe the similarity with (6) for $m \to \infty$).

Other interpretations of the generalizability are discussed in chapter 6. If we use the distinction between common and unique scores we can use

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{x'(D-U)x}{x'(C-U)x}\right), \tag{14}$$

where U is the dispersion of the unique scores. This defines α -factor analysis. If we have partitions of the variables into several subsets we can use thegeneral form of α

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{\Sigma V_s}{V}\right) \tag{15}$$

to derive

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{x'D_1x + \ldots + x'D_sx}{x'Cx}\right), \tag{16}$$

where the D_t are the within-subset parts of the matrix C (cf chapter 5). In this case the rule is

$$\alpha \geq 0 \text{ iff } \lambda \geq 1/s, \tag{17}$$

with s the number of subsets. Of course α is only a lower bound to the reliability of the composites, and the cut-off rules (13) and (17) are somewhat ad hoc. Again this can be related to the ambiguous role of statistical decision procedures in this context. From the data analytical point of view (13) and (17) may be of some help, but the interpretability of the results is much more important.

3.4 Discrimination

Consider the first variable only. It has k categories with $d_1,\ldots,d_k$ elements. We now use the ideas of discriminant analysis. The k categories define sets A_k which form a partition of the set of m sample elements. We want to assign scores in such a way that we are able to discriminate these groups as precise as possible, or (geometrically) they must be as far apart as possible. We write m_l for the mean of the scores in category l. The ANOVA table is (assuming $z_\cdot = 0$ again)

Source	Sum of squares
Between groups	$B = \sum_{l=1}^{k} d_l m_l^2$
Within groups	$W = \sum_{l=1}^{k} \sum_{j \in A_l} (z_j - m_l)^2$
Total	$T = \sum_{j=1}^{m} z_j^2$ (18)

In matrix notation for variable i

$$B_i = z'H_i'D_{ii}^{-1}H_i z, \tag{19a}$$

$$T = z'z, \tag{19b}$$

$$W_i = z'(I - H_i'D_{ii}^{-1}H_i)z. \tag{19c}$$

For all n variables at the same time we take the average correlation ratio as our scale-free index of discriminability

$$\phi = \frac{\sum B_i}{nT} = \frac{z'H'D^{-1}Hz}{n\, z'z}. \tag{20}$$

The stationary equations are

$$H'D^{-1}Hz = n\,\phi\, z \tag{21}$$

Again (21) has an improper solution z = e with $\phi = 1$. All other solutions give stationary values of the average correlation ratio. Again there are $v \leq \min(K - n, m)$ positive proper solutions.

3.5 Internal consistency

Consider the matrix $T = D^{-\frac{1}{2}}H$. By a familiar theorem T can be written uniquely as $Q\psi P'$, with Q and P square orthonormal and the K x m matrix ψ diagonal in a generalized sense (if $K \leq m$ then the first K columns are a diagonal matrix while the other m - K columns contain zeroes, if $K > m$ then the first m rows form a diagonal matrix while the other K - m rows contain zeroes). If we define $\bar{X} = D^{-\frac{1}{2}}Q$ and $\bar{Z} = P$ then we have the identities $T = D^{\frac{1}{2}}\bar{X}\psi\bar{Z}'$ and $\bar{X}'D\bar{X} = I$. It follows that

$$TT' = D^{-\frac{1}{2}}HH'D^{-\frac{1}{2}} = Q\Psi P'P\Psi'Q' = Q\Psi\Psi'Q' = D^{\frac{1}{2}}\bar{X}\Psi\Psi'\bar{X}D^{\frac{1}{2}}, \tag{22a}$$

$$T'T = H'D^{-1}H = P\Psi'Q'Q\Psi P' = P\Psi'\Psi P' = \bar{Z}\Psi'\Psi\bar{Z}'. \tag{22b}$$

Moreover

$$HH'\bar{X} = D\,\bar{X}\Psi\Psi', \tag{23a}$$

$$H'D^{-1}H\bar{Z} = \bar{Z}\Psi'\Psi\ . \tag{23b}$$

If $K \leq m$ we define Ψ as the diagonal matrix consisting of the first K columns of Ψ, and we define Z as the first K columns of $\bar{Z}$, and X as $\bar{X}$. Then

$$HH'X = DX\Psi^2, \tag{24a}$$

$$H'D^{-1}HZ = Z\Psi^2. \tag{24b}$$

If $K > m$ we define Ψ as the diagonal matrix consisting of the first m rows of Ψ, and we define X as the first m columns of $\bar{X}$, and Z as $\bar{Z}$. Again (24a) and (24b) are valid. These last two identities are exactly the stationary equations (10) and (21) and consequently, by the uniqueness properties of eigen-problems, we have proved that $n\lambda = n\Lambda = \Psi^2$. Moreover the matrices X and Z defined by the singular decomposition of $T = D^{-\frac{1}{2}}H$ are exactly the solutions of our maximization problems (10) and (21). It follows that the solutions of these two problems are related in a very simple way

$$X = D^{-\frac{1}{2}}HZ\Psi^{-1} = n^{-\frac{1}{2}}D^{-1}HZ\Lambda^{-\frac{1}{2}}\ , \tag{25a}$$

$$Z = H'D^{\frac{1}{2}}X\Psi^{-1} = n^{-\frac{1}{2}}H'X\Lambda^{-\frac{1}{2}}. \tag{25b}$$

Moreover

$$H = n^{\frac{1}{2}}DX\Lambda^{\frac{1}{2}}Z \tag{26}$$

We can also formulate our technique more directly in terms of the matrix H. A direct quantification of the attributes is a K-element vector x. Replace each element in H equal to unity by the weight corresponding to the category the sample element it is in, i.e. replace h^i_{ij} by $h^i_{ij}\,x^i_j$, and consider the resulting matrix as a bivariate distribution. The correlation ratio for weights is

$$\eta^2_x = \frac{x'Cx}{nx'Dx}. \tag{27}$$

Similarly we can directly quantify the sample elements and replace each h^i_{ij} equal to unity by z_j. The correlation ratio for scores is

$$\eta^2_z = \frac{z'H'D^{-1}Hz.}{n\,z'z} \tag{28}$$

Finally we can quantify both sample elements and attributes at the same time (simultaneous direct quantification), and replace each h^i_{ij} equal to unity by the pair (x^i_j, z_j). The correlation for the resulting bivariate distribution is

$$\rho = \frac{x'Hz}{n\,x'Dx\ z'z}. \tag{29}$$

Maximizing this correlation essentially means that, whenever a sample element is in a category, we want the quantification of these two to be similar in a least squares sense. The stationary equations for this maximization problem are

$$H z = \rho_n D x, \tag{30a}$$

$$H' x = \rho_n z. \tag{30b}$$

All three functions (27), (28), and (29) have a stationary value for the improper solution $x = e$, $z = e$, with $\eta_x^2 = \eta_z^2 = \rho^2 = 1$. We have already described the solutions of (27) and (28). Maximizing (29) gives x and z which are, considered separately, the solutions of (27) and (28) again, and which are related by (30). For the most important part of our technique we do not strictly need the matrix H. It suffices to know the bivariate marginals, and apply (10) to find the weights. The scores for any possible sample element can then be obtained by using (28b), which means that the score for a particular sample element is simply proportional to the average of the weights of the categories this element falls in. More generally: we have shown that maximization by finding optimal weights and then computing the induced scores gives the same result as finding optimal scores and computing the induced weights. Up to a proportionality factor the optimal direct scores are simply the average of the rows of the score matrix induced by a direct optimal quantification of weights, the optimal direct weight of a category is simply the average of the optimal direct scores of the sample elements in that category.

3.6 Reproducibility

If we consider all components we can, as usual in PCA, reproduce our manifold classification from the optimal weights and scores. Rewriting (26) in scalar notation and using the definition of the improper solution

$$h^i_{lj} = n^{\frac{1}{2}} d^i_{ll} \sum_{t=0}^{K-n} \lambda_t^{\frac{1}{2}} x^i_{tl} z_{tj} = n^{\frac{1}{2}} d^i_{ll} (1 + \sum_{t=1}^{K-n} \lambda_t^{\frac{1}{2}} x^i_{tl} z_{tj}) \tag{31}$$

where the extra index t refers to the number of the component.
For the reproduction of the bivariate marginals we can write

$$c^{ij}_{kl} = n\, d^i_{kk}\, d^j_{ll} \,(1 + \sum_{t=1}^{K-n} \lambda_t x^i_{tk} x^i_{tl}). \tag{32}$$

From (31)

$$h^i_{lj} > h^i_{lj'} \leftrightarrow \sum_{t=1}^{K-n} \lambda_t^{\frac{1}{2}} x^j_{tl} (z_{tj} - z_{tj'}) > 0, \tag{33}$$

and, consequently, in our geometrical discrimination model the sample elements in a category and those not in that category are separated by a hyperplane. For $p < K - n$ we approximate these linear separation boundaries. This result relates our PCA of categorical data with the techniques of Guttman and De Leeuw for categorical data mentioned in chapter 1.

3.7. Relations with chi-squares

As we have seen the system (10) has at most $K - n + 1$ solutions corresponding with nonzero $\lambda_0, \lambda_1, \ldots, \lambda_{K-n}$ with $\lambda_0 = 1$ corresponding with the improper solution. These roots have some interesting relations with the X^2-values that can be computed from our data. For the sum of the roots we find

$$\sum_{t=0}^{K-n} \lambda_t = n^{-1} \operatorname{Tr}(D^{-\frac{1}{2}} C D^{-\frac{1}{2}}) = \frac{K}{n}, \tag{34}$$

and

$$\sum_{t=1}^{K-n} \lambda_t = \frac{K - n}{n} \tag{35}$$

Interpret H as a K x m contingency table. Its chi-square is

$$X_H^2 = mn\left[\frac{1}{n} \operatorname{Tr}(D^{-\frac{1}{2}} C D^{-\frac{1}{2}}) - 1\right] = m(K-n) = mn \sum_{t=1}^{K-n} \lambda_t. \tag{36}$$

Next we compute the chi-squares for the subtables C_{ij} defined by (7). We find

$$\sum_{i=1}^{n} \sum_{j=1}^{n} X_{ij}^2 = m\left[\operatorname{Tr}(D^{-\frac{1}{2}} C D^{-1} C D^{-\frac{1}{2}}) - n^2\right] = mn^2 \sum_{t=1}^{K-n} \lambda_t^2. \tag{37}$$

For X_{ij}^2 we find

$$X_{ii}^2 = m(k_i - 1), \tag{38}$$

and

$$\sum_{i=1}^{n} X_{ii}^2 = m(K-n) = X_H^2. \tag{39}$$

Therefore

$$X_C^2 = \sum_{i \neq j} \sum X_{ij}^2 = m \sum_{t=1}^{K-N} (n\lambda_t - 1)^2 . \tag{40}$$

We can also interpret our improper solution in this framework. Consider problem (10). We shall usually solve this by computing the roots and vectors of $T_0 = \frac{1}{n} D^{-\frac{1}{2}} C D^{-\frac{1}{2}}$. The unit length vector corresponding with the dominant root $\lambda_0 = 1$ is $n^{-\frac{1}{2}} D^{\frac{1}{2}} e$, and Hotelling deflation defines

$$T_1 = \frac{1}{n} D^{-\frac{1}{2}} C D^{-\frac{1}{2}} - \frac{1}{n} D^{\frac{1}{2}} e e' D^{\frac{1}{2}}. \tag{41}$$

If we translate this by using the expected values on the hypotheses of complete bivariate independence for each of the subtables of C, we see that this is equivalent to removing the chance expectation from a X^2-analysis.

3.8 The multinormal case

If $m \to \infty$ we can translate the problem into population terminology. Although the matrix H is of little use, the matrix $\hat{C} = \frac{1}{m} C$ converges to the marginal bivariate and univariate frequency functions. Consequently we can study our technique in the case of populations too. Suppose, for example, that we are dealing with an n-variate normal distribrution. We let $k_i \to \infty$ for all i, and the weights are replaced by real valued function $\eta_i(x)$ defined on the real line(we assume that this new random variable has finite expectation and variance). Then we can expand the η_i using the Hermite-Tshebysheff polynomials $\psi_t^i(x)$ (t=0, 1, 2, ...)

$$\eta_i(x) = \sum_{t=0}^{\infty} \alpha_t^i \psi_t^i(x), \tag{42}$$

in which the series $\Sigma(\alpha_t^i)^2$ converges for each i. Then

$$B = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{s=0}^{\infty} \sum_{t=0}^{\infty} \alpha_s^i \alpha_t^j \iint \psi_s^i(x)\psi_t^j(y)\, N_{ij}(x,y)\, dx\, dy =$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{s=0}^{\infty} \alpha_s^i \alpha_s^j \gamma_{ij}^s. \tag{43}$$

$$T = n \sum_{i=1}^{n} \sum_{s=0}^{\infty} (\alpha_s^i)^2 \int \left[\psi_s^i(x)\right]^2 N_i(x)\, dx =$$

$$= n \sum_{i=1}^{n} \sum_{s=0}^{\infty} (\alpha_s^i)^2 \gamma_{ii}^s. \tag{44}$$

In formula's (43) and (44) we have used the notation N_{ij} and N_i for the bivariate and univariate normal densities, we use Y_{ij} for the covariance between variates i and j to the power s. These covariances to the power s can be collected in a matrix C^s, its diagonal is the matrix D^s, and the matrix of correlations to the power s is written as R^s. We assume that the correlations satisfy the condition $\rho_{ij}^2 \neq 1$ for all $i \neq j$. The stationary equations for maximizing $\lambda = B/T$ are

$$C^s \alpha_s = n\lambda D^s \alpha_s \qquad (45)$$

for all $s = 0,1,2,\ldots$ We can number the stationary values of (45) as λ_s^i, with λ_s^i the i^{th} largest eigenvalue of $n^{-1}R^s$. Consequently $\lambda_0^1 = 1$, and $\lambda_0^i = 0$ for all $i \neq 1$, which defines our improper solution. By considering the off-diagonal elements only we derive the remarkable result

$$\sum_{i=1}^{n} \sum_{s=1}^{\infty} (n\lambda_s^i - 1)^2 = \sum_{i \neq j} \sum (\rho_{ij}^2 + \rho_{ij}^4 + \ldots) =$$

$$\sum_{i \neq j} \sum \frac{\rho_{ij}^2}{1 - \rho_{ij}^2} = \sum_{i \neq j} \sum \phi_{ij}^2, \qquad (46)$$

with ϕ_{ij}^2 Pearson's contingency (the population analogon of X^2). It is easy to see that both $\Sigma\Sigma\lambda_s^i$ and $\Sigma\Sigma (\lambda_s^i)^2$ diverge. Under mild regularity conditions the eigenelement that gives a stationary value λ_s^i has $\alpha_t^i = 0$ for all $t \neq s$.*) Consequently the $\eta_i(x)$ corresponding with a stationary value are all polynomials of the same degree. The linear elements correspond with the principal components of the population correlation matrix. This is important exactly in so far as the multinormal distribution is a normative model. If we can trace from our components that there is one underlying approximate linear system of n components, and all other components are polynomial functions of the linear ones, then we may use a multinormal linear scoring system of just n components, reconstruct an estimate of the multinormal correlation matrix, and so on. It is an open problem in how far these effects can be detected in real data where we have sampling and grouping errors, and in how far these results can be used as a test for multinormality.

*) This is because the stationary equation (45) has to be satisfied for all s with the same value of λ. If λ_s is an eigenvalue of R^s, then it will in general not be an eigenvalue of R^t with $t \neq s$, and consequently we can only satisfy (45) by taking $\alpha_t = 0$.

3.9 Bivariate or multivariate

One of the things we have learned from the analysis in sections 3.7 and 3.8 is that our PCA, like all of classical multivariate analysis, is not strictly multivariate but actually joint bivariate analysis. Although it is perfectly possible to reproduce the (multivariate) manifold classification from our PCA, we do this by operating on the bivariate marginals. Multivariate reproducibility is merely a consequence of the fact that we can find a corresponding system of scores. The bivariate bias of classical multivariate analysis is easily explained by remembering that the multinormal distribution is completely described by its moments up to order two. For normal distributions we do not have to go any further than variances and covariances (as our analysis in section 3.8 shows). It is, consequently, not surprising that most of the work on the more general types of multivariate analysis has concentrated on the complex contingency table. Here we do not record our manifold classification in the essentially bivariate indicator matrix, but in a $k_1 \times k_2 \times \ldots \times k_n$ multiway table. These complicated and very interesting extensions of the classical cases are based again on orthonormal functions, partitionings of chi-square and the LR-statistic, and on systematic investigation of the possible hypotheses of independence in such a table. Although this approach seems to be more promising in the long run, especially for categorical data, we do not go into it in this dissertation. One should consider the techniques presented here as a practical and quite useful link between the classical, linear, joint bivariate analysis, and the new forms of multivariate analysis. In the case that we are actually dealing with multinormal distributions (or: if we can find weights which transform the margins to approximate multivariate normality) the three types of techniques should give essentially identical results, although possibly in quite a different form.

3.10 The case n = 2

The remarks in the previous two sections suggest that our technique may simplify considerably in the case n = 2. Moreover we can expect new interpretations that may generalize in some sense to the multivariate case. For ease of notation we write, in this section N for C_{12}, N_1 for D_1, N_2 for D_2, $n_{i.}$ for d^1_{ii}, n_{j} for d^2_{jj}, $n_{..}$ for m, x for x_1, y for x_2, n for k_1, and m for k_2 (and we suppose wlg that $m \leq n$). The stationary equations for the optimal weights are

$$\begin{bmatrix} N_1 & N \\ \hline N' & N_2 \end{bmatrix} \begin{bmatrix} x \\ \hline y \end{bmatrix} = 2\lambda \begin{bmatrix} N_1 & \\ \hline & N_2 \end{bmatrix} \begin{bmatrix} x \\ \hline y \end{bmatrix} . \tag{47}$$

This can be simplified by considering the two equations

$$Ny = (2\lambda - 1) N_1 x, \tag{48a}$$

$$N'x = (2 \lambda - 1) N_2 y. \tag{48b}$$

Writing ρ for $2\lambda - 1$ we see that solving (48) is equivalent to the maximization of the correlation coefficient

$$\rho = \frac{x'Ny}{(x'N_1x)^{\frac{1}{2}} (y'N_2y)^{\frac{1}{2}}} \tag{49}$$

which has $v \leq m - 1$ proper solutions and one improper one. The stationary values of (49) are $\rho_0, \rho_1, \cdots, \rho_{m-1}$, and we can agree wlg that $\rho_s \geq 0$ for all s. Then the stationary values of (47) are

$$\lambda_0 = \frac{\rho_0 + 1}{2}, \cdots, \frac{\rho_{m-1}+1}{2} = \lambda_{m-1} \tag{50a}$$

and

$$\lambda_{2m-1} = \frac{1 - \rho_0}{2}, \cdots, \frac{1 - \rho_{m-1}}{2} = \lambda_m . \tag{50b}$$

Or, for s=1,...,m,

$$\lambda_{s+m-1} = 1 - \lambda_{m-s} . \tag{50c}$$

If the solutions of (48) are $(x_0, y_0), \cdots, (x_{m-1}, y_{m-1})$, then those of (47) corresponding with the roots (50b) are $(-x_0, y_0), \cdots, (-x_{m-1}, y_{m-1})$. Finally the system (47) has n - m solutions of the form (x, 0) with eigenvalues $\lambda = \frac{1}{2}$, corresponding with the n-m N_1-orthogonal vectors x satisfying N'x = 0. This makes a total of (2m-2)+(n-m) = m+n-2 solutions with non-zero eigenvalues (of which one is improper).

The complete solution of (47) can thus be easily derived from that of (48). We also have the result from (48)

$$X_N^2 = n_{\cdot\cdot} \sum_{t=1}^{m-1} \rho_t^2 . \tag{51}$$

Another important consequence of (48) is

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}} \left(1 + \sum_{t=1}^{m-1} \rho_t x_i^t y_j^t\right) \tag{52}$$

which is the specialization of (32) to n = 2.

What happens to the results of section 3.8? It is easy to see that the roots of R^s are

$$\lambda_s^1 = \frac{1 + \rho^s}{2}, \quad \lambda_s^1 = \frac{1 - \rho^s}{2} . \tag{53}$$

Consequently (46) reduces to

$$2\sum_{s=1}^{\infty}\rho^{2s} = 2\frac{\rho^2}{1-\rho^2} = 2\phi^2. \qquad (54)$$

The α_s^1 and α_s^2 corresponding to the roots $(1+\rho^s)/2$ are all proportional to (1, 1), and consequently the solutions of (45) are the Hermite-Tshebysheff polynomials of the first degree, of the second degree, and so on. We may assume again wlg that $\rho \geq 0$, and consequently $(1+\rho^s)/2 > (1-\rho^t)/2$ for all s,t. The other eigenelements, corresponding with $\alpha \propto (-1, 1)$, need not enter into the analysis at all. The relation (52) becomes in the limit,

$$N(x,y)=(2\pi)^{-1}\exp(-\tfrac{1}{2}(x^2+y^2))\{1+\sum_{s=1}^{\infty}\rho^s\psi_s^1(x)\psi_s^2(y)\} \qquad (55)$$

which is Pearson's polychoric expansion of the binormal distribution with correlation parameter ρ. Equation (32) shows that the same thing is true for the joint bivariate marginals in the multivariate case.

Our technique linearizes the regressions in the sense that for each component (x, y)

$$y = \rho^{-1}N_2^{-1}Nx = \bar{x}_y \qquad (56a)$$

$$x = \rho^{-1}N_1^{-1}N'y = \bar{y}_x, \qquad (56b)$$

where $\bar{x}_y$ ($\bar{y}_x$) stand for the conditional expectation of x (y) for fixed values of y (x). The analogon of (56) in the joint bivariate case is (for each component x)

$$x_i = (n\lambda - 1)^{-1}\sum_{j\neq i} D_i^{-1}C_{ij}x_j = (n\lambda-1)^{-1}\sum_{j\neq i}\bar{x}_j^i, \qquad (57)$$

where $\bar{x}_j^i$ stand for the conditional expectation of x_j for a fixed value of x_i. In the bivariate case we maximize the correlation, in the multivariate case the sum of the covariances relative to the sum of the variances. This does not make all bivariate regressions linear, because that would require

$$x_i \propto D_i^{-1}C_{ij}x_j = \bar{x}_j^i \qquad (58)$$

for all i,j. Consequently our multivariate components can also be interpreted as weights which maximize the 'overall linearity' of the regressions. Once again we point out that our PCA of manifold classifications is bivariate in the sense that we analyze H as if it was a two-dimensional contingency table.

3.11 Equality constraints

It is sometimes appropriate to interprete the H_i as if they were replications of one another (in this case they clearly must have the same number of categories). We can quantify the categories under the restrictions

$$x_1 = x_2 = \ldots = x_n = x \text{ (say)}. \tag{59}$$

Then

$$\lambda = \frac{x'\tilde{C}x}{x'\tilde{D}x}, \tag{60}$$

with

$$\tilde{C} = \Sigma\Sigma C_{ij}, \tag{61a}$$

$$\tilde{D} = n\Sigma D_{ii}. \tag{61b}$$

Again maximizing λ produces an improper solution with $\lambda = 1$.

Cases in which only some of the H_i define replications suggest the generalization in which each column of each H_i defines a discrete probability distribution (which may, of course, be obtained either by 'objective' or by 'subjective' types of probability assessment methods). We define z_{ij} as before, i.e. as the expected value of the weight function on the categories, and we maximize the between-within homogeneity of these induced quantifications. Our previous PCA technique is the special case in which each of the distributions is of the one point type.

3.12 Another geometrical interpretation

We have already given a geometrical interpretation in terms of the discrimination of sample elements in section 3.4, and in terms of linear separation boundaries in section 3.6. There is a more direct way to get Euclidean distance into the picture. Let Q be any symmetric positive semi-definite (psd) matrix, and let e_i, e_j denote unit vectors with one element equal to unity and all other elements equal to zero. Define

$$\delta_{ij}^2 = (e_i - e_j)'Q(e_i - e_j), \tag{62}$$

then δ_{ij} is a pseudometric (it is a metric iff for all $i \neq j$ the vector $e_i - e_j$ is not in the null space of Q). If we write KAK' for the canonical form of Q, we can define the vector t_i by $t_i' = e_i'KA^{\frac{1}{2}}$. Then the Euclidean distance between the endpoints of the t_i is given by the square root of

$$d_{ij}^2 = (t_i - t_j)'(t_i - t_j) = \delta_{ij}^2, \tag{63}$$

which means that the 'weighted' distance δ_{ij} between the unit vectors e_i and e_j equals the ordinary Euclidean distance between the vectors t_i and t_j. Moreover $t_i't_i = q_{ii}$ and $\Sigma t_i t_i' = A$.

If K_p are the p eigenvectors of Q associated with the p largest eigenvalues (collected in A_p) then we can define $t_i^p = A_p^{\frac{1}{2}} K_p' e_i$, and

$$(d_{ij}^p)^2 = (t_i^p - t_j^p)'(t_i^p - t_j^p) =$$
$$= (e_i - e_j)' K_p \Lambda_p K_p' (e_i - e_j) =$$
$$= (e_i - e_j)' Q_p (e_i - e_j), \qquad (64)$$

where Q_p is the optimal rank-p approximation to Q in the least-squares sense. Obviously $(d_{ij}^p)^2 \leq \delta_{ij}^2$ for all i,j and for all p.

Apply these principles to a metrization of the set of sample elements. Substitute $Q = H'D^{-1}H$ in (62) and t_i is the z of (21). Geometrically we represent the sample elements as the points e_i in m-dimensional space, and define the metric (62). This metric can be interpreted quite nicely by observing that

$$\delta_{jj'}^2 = \sum_i \sum_l \frac{(h_{ij}^i - h_{ij'}^i)^2}{d_l^i} \quad . \qquad (65)$$

We interpret the set of all categories as a discrete measure space in which each element has the measure $\mu_l^i = (d_l^i)^{-1}$, then each sample element defines a subset T_j with measure equal to the sum of the measures of the elements contained in it, and

$$\delta_{jj'}^2 = \mu(T_j \cup T_{j'}) - \mu(T_j \cap T_{j'}) =$$
$$= \mu(T_{j'}) + \mu(T_j) - 2\mu(T_j \cap T_{j'}) =$$
$$= \mu(T_{j'} \Delta T_j), \qquad (66)$$

(with Δ denoting the symmetric difference). It follows that, in this particular case, not only δ_{ij} but also δ_{ij}^2 defines a pseudometric. It is a metric iff no two sample elements are exactly the same. Of course we could have modified our techniques in such a way that identical sample elements are replaced by one single element with weight equal to the number of these elements. This would mean using a D-matrix for the sample elements too, and because we usually analyze (10) and not (21) this complicates the analysis. The next step in the distance analysis is to rotate the m-dimensional representation to principal components and to discard the smaller roots, which means that we optimally approximate the set-theoretical distances δ_{ij}^2 in $p < m$ dimensions by the Euclidean distances d_{ij}^2.

A similar analysis is possible for the categories. Three possible choices are

$$Q_A = HH', \qquad A t_i = n^{\frac{1}{2}} e_i' D X \Lambda^{\frac{1}{2}}, \qquad (67a)$$
$$Q_B = D^{-\frac{1}{2}} HH' D^{-\frac{1}{2}}, \qquad B t_i = n^{\frac{1}{2}} e_i' D^{\frac{1}{2}} X \Lambda^{\frac{1}{2}}, \qquad (67b)$$
$$Q_C = D^{-1} HH' D^{-1}, \qquad C t_i = n^{\frac{1}{2}} e_i' X \Lambda^{\frac{1}{2}}. \qquad (67c)$$

With obvious notation, and with μ as counting measure, this gives

$$\delta_A^2(s_1^i, s_{1'}^{i'}) = \mu(s_1^i \Delta s_{1'}^{i'}), \tag{68a}$$

$$\delta_B^2(s_1^i, s_{1'}^{i'}) = 2\left[1 - \mu^{-\frac{1}{2}}(s_1^i)\,\mu(s_1^i \cap s_{1'}^{i'})\,\mu^{-\frac{1}{2}}(s_{1'}^{i'})\right] \tag{68b}$$

$$\delta_C^2(s_1^i, s_{1'}^{i'}) = \mu^{-1}(s_1^i)\,\mu(s_1^i \Delta s_{1'}^{i'})\,\mu^{-1}(s_{1'}^{i'}) \tag{68c}$$

Possibility C seems the most interesting one and yields our previous scaling of the solutions X. For a joint space representation we define

$$E = \begin{bmatrix} D^{-1}HH'D^{-1} & D^{-1}H \\ H'D^{-1} & H'D^{-1}H \end{bmatrix} \tag{69a}$$

$$2\Gamma = \begin{bmatrix} n\Lambda + n^{\frac{1}{2}}\Lambda^{\frac{1}{2}} & 0 \\ 0 & n\Lambda - n^{\frac{1}{2}}\Lambda^{\frac{1}{2}} \end{bmatrix} \tag{69b}$$

$$X = \begin{bmatrix} X & -X \\ Z & Z \end{bmatrix} \tag{69c}$$

Then $E = X\Gamma X'$, and we take $t_i' = e_i' X \Gamma^{\frac{1}{2}}$.

3.13 Numerical variables

Suppose next that some (or all) of the variables are numerical, i.e. for some of the i there is a vector y_i of real numbers given. We incorporate this prior information into the analysis by requiring that the weights x_i must be proportioned to the vector y_i, or $x_i = \alpha_i y_i$. It is easy to see that this makes column i of the induced score matrix Z proportional to the vector of observed scores on that variate in deviations from the mean. If all variates are numerical our homogeneity criterion reduces to

$$\lambda = \frac{\alpha' C \alpha}{n\alpha' D \alpha} \tag{70}$$

with C the sample dispersion matrix of the observed scores, and D its diagonal. Consequently the stationary values are simply the eigenvalues of the sample correlation matrix (divided by n), and our analysis reduces to scale-free ordinary PCA. This is of course one of the main reasons why we use $x_i = \alpha_i y_i$ as a restriction. Another reason can be deduced from section 3.8. In the multinormal population case the n solutions for α_1 turn out to be the principal components of the population correlation matrix. It is now also very easy to handle 'mixed' cases in which some of the variables are numerical and others are nominal. The information that some of the variables are numerical is, computationally, always a simplification of the analysis. It just means that we must replace the matrix H_i by a single row of prior scores in deviations from the mean

(obviously the internal consistency analysis applies here too). The result of this is that all optimal weight vectors for a numerical variate are linear functions of the original prior quantifications y_i. If we do not use the numerical information in the analysis they are arbitrary functions of the y_i and it can, in some cases, be very instructive not to use these prior weights right away, and see what comes out (cf section 3.8).

A less rigorous approach which may also be justifiable in some cases is to use the requirements $x_i = \alpha_i y_i$ only for the first nontrivial component. One reason for doing this is that we require $\Sigma (x_i^t)' C_{ii} x_i^s = \delta^{ts}$ for two different components x^t and x^s. If $n = 2$ we actually have $(x_i^t)' C_{ii} x_i^s = \delta^{ts}$ for all t,s and for both i=1,2. This may considered more satisfactory. We have components which are not only D-conjugate over all n-variables, but which are D_i-conjugate for each variable separately. If we require $x_i^t = \alpha_i^t y_i$ for all t, then $(x_i^t)' D_i x_i^s = n\alpha_i^t \alpha_i^s \sigma_i^2$, which is zero only when either α_i^t or α_i^s is zero. In an exploratory phase it may be quite useful to require $x_i^t = \alpha_i^t y_i$ only for the first p dominant components, because it is intuitively obvious that the smaller the eigenvalues, the less reliable the components, and the more the possible multinormal effects will be disturbed. It is also very important how 'continuous' our numerical variate is. If k_i is large (close to m) then H_i will be approximately a permutation matrix, and it will not make much sense to restrict none or only a few of the components. If k_i is much smaller than m, then it may be better not to use the restrictions right away and to do some preliminary exploring before we require $x_i^t = \alpha_i^t y_i$ for all t.

3.14 The numerical case n=2

If we have prior weights a and b for both of the attributes, and we require (in notation of section 3.10) $x = \alpha a$, $y = \beta b$, then, from (49), ρ equals (plus or minus) the sample correlation coefficient for all α, β. Consequently there is nothing to maximize in this case. What we can do is require $x = a$, $y = b$ (with a and b suitably standardized), compute the residual by deflating out a and b, and finding the unrestricted components of the residual matrix. The case becomes somewhat more interesting if we restrict only x (or only y). Suppose we require $y = b$, then the maximization of ρ leads to a unique stationary value for

$$x = \rho^{-1} N_1^{-1} N b, \tag{71a}$$

with

$$\rho^2 = n_{..}^{-1} b' N' N_1^{-1} N b \tag{71b}$$

3.15 Binary variables

If the ϕ_i are binary we can use the notation m_i^+ and m_i^- for the univariate marginals, and the notation $m_{ii'}^{++}$, $m_{ii'}^{+-}$, $m_{ii'}^{-+}$, and $m_{ii'}^{--}$, for the bivariate marginals. The two weights for attribute i are x_i^+ and x_i^-. In order that $\sum_{j=1}^{m} z_{ij} = 0$ it must be true that

$$m_i^+ x_i^+ + m_i^- x_i^- = 0, \tag{72}$$

and consequently the vector (x_i^+, x_i^-) must be proportional to the vector $m_i = (m_i^-, -m_i^+)$. Thus for all t, we require $x_i^t = \alpha_i^t m_i$. We have a situation similar to the analysis of numerical variates. If all ϕ_i are binary our homogeneity measure reduces to

$$\lambda = \frac{\Sigma\Sigma b_{ii'}\alpha_i \alpha_{i'}}{n\Sigma b_{ii}\alpha_i^2}, \tag{73a}$$

with

$$b_{ii'} = m\, m_{ii'}^{++} - m_i^+ m_{i'}^+. \tag{73b}$$

Consequently our method reduces to (scale free) PCA of phi-coefficients. We also mention a notation which is more natural from a statistical point of view, and which readily generalizes to the general case. Suppose we have a probability distribution $\{p_i\}$ on the set of all binary vectors of n elements $\{x_i\}$, and a random sameple of size m according to this distribution. This gives an observed distribution $\{\hat{p}_i\}$. We can write

$$\hat{v} = \Sigma \hat{p}_i x_i \tag{74a}$$

$$\hat{S} = \Sigma \hat{p}_i x_i x_i' \tag{74b}$$

$$\hat{C} = \hat{S} - \hat{v}\hat{v}' \tag{74c}$$

$$\hat{D} = \text{diag}\,(\hat{C}) \tag{74d}$$

and

$$\lambda = \frac{\alpha'\hat{C}\alpha}{n\alpha'\hat{D}\alpha} \tag{75}$$

Observe that in this interpretation we have a definite statistical model (independent, identically distributed observations), and in this sense the notation is somewhat less general.

Although formally it amounts to the same thing to require $x_i^t = a_i^t y_i$ for all t in the numerical case and to require $x_i^t = a_i^t m_i$ for all t in the binary case, the reasons for these requirements are quite different. In the binary case no prior information is involved, no restrictions are made, and we cannot choose whether to use these requirements or not. In the numerical case we require that x_i^t must be a linear function of the y_i, which are given numbers, in the binary case we can choose any two-elements vector y_i and require that x_i^t is a linear function of y_i, because we can always fit a straight line through two points. This reduction of the order of the matrix from 2n to n is done by a method which also generalizes to categorical items with $k_i > 2$. Suppose y_v^i are $k_i - 1$ independent vectors with $\Sigma d_{ii}^i y_{v1}^i = 0$ for all v. Then any x_i-part of a proper solution must be a linear combination of the y_v^i, and consequently we have reduced to problem to one with $k_i - 1$ linear coefficients α_v^i for each i, and $\hat{C}$ consisting of $\hat{C}_{ij} = Y_i' C_{ij} Y_j$ becomes of order K - n, and in general nonsingular. The new $D_{ii} = Y_i' D_{ii} Y_i$ will in general be nonsingular but not diagonal. Therefore we restrict ourselves somewhat more to those Y_i which have the property that $Y_i' D_{ii} Y_i$ is diagonal, and that the first column has all its elements equal to unity, i.e. $y_{.1}^i = e$. It is easy to find these Y_i by Gram-Schmidt orthogonalization (we use only the last $k_i - 1$ columns). It is a problem for further investigation which of the methods is more efficient: the one that operates on a matrix of order K which is easy to compute, or the one that operates on the matrix of oder K - n which is more difficult to find. The computational efficiency depends on the value of K/(K-n). If this is small (i.e. if $\bar{k}_i$ is small) the efficiency of the second method, compared with the first one, increases. But efficiency should not be defined exclusively in terms of computational effort. If we choose Y_i in a rational way we may get a lot of useful extra information at a relatively cheap price (cf chapter 5).

In chapter 2 we mentioned the possibility of splitting up an attribute with k_i categories into k_i attributes with two categories. In the first case we put all rows of H_i in deviations from the mean, in the second case we have k_i matrices H_i of order 2 x m, which are replaced by their first rows in deviations from the mean. These two procedures consequently give the same matrix C, the difference is in the matrix D, i.e. in the scaling of the solutions. In the first case we use the $k_i \times k_i$ matrices $H_i H_i'$, in the second case only their diagonal (because the

variance of the induced scores is defined differently). Observe that this analysis also holds for attributes with non-exclusive categories. In chapter 5 we shall study this idea in a more general sense. We shall discuss several possible groupings of the variables, show that the matrix C remains the same, and that the only effect of these groupings is changing the structure of D.

3.16 Some order-reducing methods

The techniques discussed in the previous section, especially for binary data, are also very important in the general case. In this section we mention two special choices of the matrices Y_i which are attractive from a computational point of view and/or from the point of view of interpretation.

Let $n_1, n_2, \ldots, n_k$ be the marginal frequencies of a variable with k categories. Define the k x k matrix Y by

$$\begin{matrix} n_1 - m & n_2 & n_3 & \ldots & n_k \\ n_1 & n_2 - m & n_3 & \ldots & n_k \\ n_1 & n_2 & n_3 - m & \ldots & n_k \\ \vdots & \vdots & \vdots & & \vdots \\ n_1 & n_2 & n_3 & \ldots & n_k \\ n_1 & n_2 & n_3 & \ldots & n_k - m \end{matrix} \qquad (76)$$

Let N be the diagonal matrix with elements n_i on the diagonal, n is the vector with the same elements. Then

$$Y = EN - mI, \qquad (77)$$

For any k-element vector x

$$x'Y = sn' - mx', \qquad (78)$$

with $s = \Sigma x_i$. Thus $x'Y = 0$ iff $x = \alpha n$ for some real α and Y is of rank k-1. In the same way

$$Yx = te - mx, \qquad (79)$$

with $t = \Sigma x_i n_i$. Thus $Yx = 0$ iff $x = \alpha e$ for some real α. Any k-1 columns of Y are linearly independent. It follows that any vector x with $\Sigma x_i n_i = 0$ can be written as a linear combination of k-1 columns of Y. Thus any k-1 columns will do for our purposes. Collect them in a k x (k-1) matrix $\bar{Y}$, and use the linear restrictions $\bar{Y}\alpha = x$. If H is the k x m indicator matrix of the variable, then

$$Y'H = NEH - mH = NE - mH, \qquad (80)$$

which means that the rows of Y'H are proportional to the rows of H in deviations from the mean. For the matrix $\bar{Y}$'H we simply leave out one of the rows (for example the last one). It follows that

$$Y'DY = Y'HH'Y = m^2 N - m\, NEN, \tag{81}$$

and $\bar{Y}'D\bar{Y}$ can be found by leaving out the proper row and column. Obviously it is not diagonal. The columns of Y contrast one particular category of the variable with the others.

Another procedure, in which the computations are somewhat different, is defined by taking Y as the Helmert-type matrix

$$\begin{array}{ccccccc} 1 & -n_2 & -n_3 & -n_4 & \cdots\cdots & -n_k \\ 1 & n_1 & -n_3 & -n_4 & \cdots\cdots & -n_k \\ 1 & 0 & n_1+n_2 & -n_4 & \cdots\cdots & -n_k \\ 1 & 0 & 0 & n_1+n_2+n_3 & \cdots\cdots & -n_k \\ \vdots & \vdots & \vdots & \vdots & & \\ 1 & 0 & 0 & 0 & \cdots\cdots & n_1+n_2+\ldots+n_{k-1} \end{array} \tag{82}$$

The matrix Y'DY is diagonal in this case, which is a considerable advantage in PCA. If $\bar{Y}$ contains the last k-1 columns of Y, we can again use $\bar{Y}\alpha = x$. The matrix $\bar{Y}$'H is a bit more difficult to compute than in the previous case although only addition and subtraction of integers is involved. The columns of $\bar{Y}$ contrast category i with categories 1,2,..., i-1, and consequently for the interpretation the order of the categories is relevant. The matrices $\bar{Y}_i'C_{ij}\bar{Y}_j$ can be used to obtain a Lancaster-Irwin partition of X^2. The same partition property obtains, of course, for all Y such that Y'DY is diagonal. For a numerical variable with k categories for example, it seems very interesting to take in Y the orthogonal polynomials of degree 1,2,...,k-1 with the first polynomial a linear function of the prior numerical scores. This is especially interesting if k/m is relatively small, for example in the case of rating scales.

3.17 The perfect scale

An important special case of n binary variates is the perfect Guttman-scale, which can again serve as a pseudo-normative model in the same way as the multinormal distribution. In the case of a perfect scale we have n variables, and there are only n + 1 possible score patterns. These patterns can be collected in the matrix G.

score patterns

1	2	3	.....	n	n+1		
1	1	1	.....	1	0	1	
1	1	1	.....	0	0	2	variables
.	.	.	.....	.	.	.	
.	.	.	.....	.	.	.	
1	1	0	.....	0	0	n-1	
1	0	0	.....	0	0	n	
m_1	m_2	m_3	.....	m_n	m_{n+1}	m	

(83)

Write t_i for the weighted sum of the rows of G. Then

$$t_i = \sum_{j=1}^{n+1} m_j g_{ij} = \sum_{j=1}^{n-i+1} m_j . \tag{84}$$

To obtain deviations from the mean we replace each g_{ij} by $h_{ij} = mg_{ij} - t_i$. The cross-product of the deviations in rows i and k is

$$v_{ik} = \sum_{j=1}^{n+1} m_j h_{ij} h_{kj} =$$

$$= \begin{cases} mt_i(m-t_k) & \text{for } i \geq k. \\ mt_k(m-t_i) & \text{for } i < k. \end{cases} \tag{85}$$

The product-moment correlation (phi-coefficient) is, defining

$$a_i = \sqrt{\left(\frac{t_i}{m-t_i}\right)} , \tag{86}$$

given by

$$\phi_{ik} = \begin{cases} a_i/a_k & \text{for } i \geq k. \\ a_k/a_i & \text{for } i < k. \end{cases} \tag{87}$$

Matrices with this structure (correlation simplexes) have a number of well known properties. If $a_i \neq 0$ and $a_i \neq a_k$ for all $i \neq k$ then they are nonsingular and all n latent roots are different. Moreover their inverse is tridiagonal. Their eigenvectors are the (oscillatory) discrete orthogonal polynomials. A very similar development can be used for finding the direct optimal scores.

3.18 Improper solutions

We have treated the introduction of numerical and binary variates in such a way as if the rows of the nominal variables in the set were already in deviations from the mean. This entails, however, that $H_i H_i' = D_i - m^{-1} D_i e e' D_i$, and

consequently D is singular and not diagonal. In the case in which all variables were nominal we could avoid this difficulty by introducing the improper solution which makes D both diagonal and non-singular. Fortunately such a simplification is also possible when we have a mixed case with numerical and/or binary variates. Suppose $\bar{n} \leq n$ variates are nominal with more than two categories. We assume wlg that these are the first $\bar{n}$ variables. Suppose they have a total of $\bar{K}$ categories. Replace the binary and numerical variates in H by the single column containing deviations from the mean, and do not subtract out the column means for the nominal variates. Then the product-moment matrix C has structure

$$\left[\begin{array}{ccc|c} C_{11} & \cdots\cdots & C_{1\bar{n}} & \\ C_{21} & \cdots\cdots & C_{2\bar{n}} & S \\ \cdots\cdots & & \cdots\cdots & \\ C_{\bar{n}1} & \cdots\cdots & C_{\bar{n}\bar{n}} & \\ \hline & S' & & R \end{array}\right] \qquad (88)$$

where S contains $n - \bar{n}$ columns and $\bar{K}$ rows. Moreover $e's_j = \bar{n}\, e'z_j = 0$ for all columns s_j, and even the $\bar{n}$ subvectors s_j^i with k_i elements are in deviations from the mean. Consequently, if t has its first $\bar{K}$ elements equal to unity and the other $n - \bar{n}$ equal to zero, then

$$C\ t = \bar{n}\ D\ t, \qquad (89)$$

and t satisfies (10) with $\lambda = \bar{n}/n$. Partialling out the contribution of t means, obviously, replacing C_{ij} by $C_{ij} - m^{-1} D_i e e' D_j = \bar{C}_{ij}$, and everything is back to normal again. Every other solution is proper, i.e. satisfies the constraints $\sum d_i^i x_i^i = 0$ for all $i = 1,\ldots,\bar{n}$, and there are in general $v \leq \min(\bar{K} - 2\bar{n} + n, m)$ of these proper solutions with $\lambda > 0$. It is, indeed, true that introducing numerical and/or binary variates simplifies the computation. It reduces the order from K to $\bar{K} + n - \bar{n}$, and it does not spoil any of the nice computational properties. Some people may consider it a disadvantage that the improper solution does not necessarily correspond to the dominant eigenvalue, but even this can easily be remedied. We do not, in fact, use C of (88) in the actual computations, only its diagonal D. In C we replace C_{ij} by $\bar{C}_{ij}$ right away, which means that $Ct = 0$, and the improper solution t does not enter our further considerations at all. All other solutions remain the same. In the mixed case the connections with X^2 are less clear. For nominal-nominal, nominal-binary, and binary-binary pairs we partition the X^2_{ij} in additive components.

For binary-binary paris this X^2_{ij} is, of course, $m\phi^2_{ij}$, where ϕ_{ij} is the phi-coeeficient. For binary-numerical pairs we partition the squared point-biserial correlation, for numerical-numerical pairs the squared product-moment correlation. For numerical-nominal pairs, finally, we partition the multiple discriminant correlation.

3.19 Ordinal variables.

In recent psychometric literature ordinal variables have received a great deal of attention. How do we incorporate prior ordinal information about the weights into our analysis? We only consider the case in which this prior information can be expressed as a number of homogeneous linear inequalities $Ax \geq 0$, and in which $x = e$ is a solution of these inequalities. We must solve

$$\lambda_x = \frac{x'Cx}{nx'Dx} \qquad \max ! \tag{90}$$

under the conditions

$$A x \geq 0, \tag{91}$$

and for all i

$$e'D_i x_i = 0 \tag{92}$$

We replace C_{ij} by $\tilde{C}_{ij}$ as usual. Now we can forget the improper solution maximize (90) under the condition (91) and adjust the optimal solution afterwards in such a way that it satisfies (92). The general problem of maximizing (90) under the conditions (91) is perhaps most easily solved by a cyclic-coordinate-ascend (CCA) method. Suppose x is feasible. Thus $Ax = t \geq 0$, suppose A is p x K. We now set s=1, replace the element x_s by

$$x_s^+ = x_s + \theta, \tag{93}$$

where θ is chosen in such a way that λ^+ is maximized along the coordinate direction (93) under the condition that x^+ remains feasible. The rest of the procedure is relatively straightforward: replace x_s by x_s^+, let $s = s + 1$, and compute the relevant quantities all over again. If $s = K$ we have completed a cycle. If x has changed only a little during this last cycle, we stop, if not we let $s = 1$ again and start a new cycle. It can be proved that under some mild regularity conditions (do not start in a stationary value of the unrestricted problem, assume D is non-singular) the vector x converges to the absolute maximum of (90) under the conditions (91). Consequently we have a computational procedure which gives us the absolute maximum. How do we proceed from here on ? The obvious thing to do is to compute residuals and then to decompose them with our usual unconstrained procedure. If we would maintain the

requirement $Ax \geq 0$ for all components (together with the usual orthogonality requirements) the result is not satisfactory, since the two types of requirements are more or less, contradictory. In fact, if the ordinal restrictions are $x_1^i \geq x_2^i \geq \ldots \geq x_{k_i}^i$ for all i (which will be the usual case), then for any two components $\bar{x}$ and x satisfying these restrictions

$$T(x,\bar{x}) = \sum_i \sum_l \sum_{l'} d_l^i d_{l'}^i (x_l^i - x_{l'}^i)(\bar{x}_l^i - \bar{x}_{l'}^i) \geq 0 \qquad (94)$$

with equality iff for all i, l, l' for which $x_l^i > x_{l'}^i$ it is true that $\bar{x}_l^i = \bar{x}_{l'}^i$, and for all i, l, l' for which $\bar{x}_l^i > \bar{x}_{l'}^i$, it is true that $x_l^i = x_{l'}^i$. By expanding (94) we also see that

$$\tfrac{1}{2}\, T(x,\bar{x}) = m \sum_i \sum_l d_l^i x_l^i \bar{x}_l^i - \sum_i \Big(\sum_l x_l^i d_l^i \sum_l \bar{x}_l^i d_l^i \Big) =$$

$$= m\, x'D\bar{x} = 0 \qquad (95)$$

by orthogonality. Thus requiring both orthogonality and $Ax \geq 0$ for all components means that strict inequalities in any of the p solutions correspond with ties in all other solutions. The successive solutions will contain more and more ties. If the dominant solution satisfies all inequalities strictly, then the only other vector satisfying both orthogonality and $Ax \geq 0$ is the improper solution $x = e$. If we compare this situation with the one we had in the analysis of numerical variates we find that we required $x_i^t = a_i^t y_i$, but not $a_i^t \geq 0$ for all t. In the ordinal situation this would mean requiring either $Ax_i^t \geq 0$ or $Ax_i^t \leq 0$ for all i,t. Computationally this presents us with a very difficult problem. Ordinal prior information will consequently be used in most cases for the first component only. We have warned in the numerical case against using $x_i^t = a_i^t y_i$ too quickly, we must repeat this warning even more strongly here. In quite a number of cases the inequalities $Ax \geq 0$ come out approximately if we do not use the restrictions at all. It does not make much sense to apply an elaborate maximization procedure, create boundary solutions which replace violations by ties, and so on. The violations of monotonicity may even be quite informative. I would recommend the procedure only if it is clear from an unrestricted analysis that the monotone component is there, but as a linear combination of, for example, the first few orthogonal principal components. It may then also be clear from inspection of the results whether it makes sense to require $Ax \geq 0$ or $Ax \leq 0$ for other variables and/or components. Another possible procedure, which we have consistently ignored up to now, is to find the optimal weights for one component only, compute the induced score matrix Z, and apply ordinary PCA to this matrix. This can, of course, also be done for nominal, binary, and

numerical variables. May be this is a useful ad hoc procedure, although our theoretical analysis in this chapter indicates that internal consistency is preserved by an ANOVA of Z and not by a PCA (a centroid analysis seems even more justified). Carrying this a step further we could do a separate PCA for each direct quantification of the categories, and get a whole system of quantifications.

3.20 Population models.

It is very useful to consider some rational stochastic models for categorical data and to apply our techniques to see how the principal components relate to the parameters of the model. As an example consider a set of n binary items and a one-dimensional latent structure model of the form

$$p_i(\xi) = f(\theta_i, \xi) \tag{96}$$

where ξ is the one-dimensional subject parameter, and θ_i are the one-dimensional item parameters. For the population

$$p_i = \int f(\theta_i, \xi)\, dF(\xi), \tag{97}$$

where F is the distribution of the subject parameters. Using the postulate of local independence

$$p_{ik} = \int f(\theta_i, \xi) f(\theta_k, \xi)\, dF(\xi). \tag{98}$$

For Guttman's deterministic model

$$p_i(\xi) = \begin{cases} 0 & \text{if } \xi \leq \theta_i \quad (99a) \\ 1 & \text{if } \xi > \theta_i. \quad (99b) \end{cases}$$

Consequently

$$p_i = 1 - F(\theta_i), \tag{100}$$

$$p_{ik} = \begin{cases} 1 - F(\theta_i) & \text{if } \theta_i \leq \theta_k \quad (101a) \\ 1 - F(\theta_k) & \text{if } \theta_i > \theta_k \quad (101b) \end{cases}$$

Defining

$$a_i = \sqrt{\left(\frac{F(\theta_i)}{1 - F(\theta_i)}\right)}, \tag{102}$$

We find for the phi-coefficients

$$\phi_{ik} = \begin{cases} a_k/a_i & \text{if } \theta_i \leq \theta_k \quad (103a) \\ a_i/a_k & \text{if } \theta_i > \theta_k. \quad (103b) \end{cases}$$

The matrix of phi-coefficients consequently has simplex structure. The case treated in section 3.17 is the special case in which F is the discrete rectangular df. Conversely if, for binary data, the matrix of phi-coefficients has simplex structure, then Guttman's model can be assumed to hold for some F.

Alternatively we may be able to approximate $f(\theta_i,\xi)$ in the region where there is a large population density by

$$p_i(\xi) = \theta_i \xi_i^1 + \xi_i^2 \tag{104}$$

It follows by local independence that if $i \neq k$

$$p_{ik} - p_i p_k = \theta_i \theta_k \sigma^2, \tag{105}$$

where σ^2 is the population variance of the random variable corresponding with ξ. Let $\sqrt{\{p_i - p_i^2\}} = \delta_i$, and $\mu_i = \sigma\theta_i/\delta_i$, then

$$\phi_{ik} = \begin{cases} \mu_i \mu_k, & (i \neq k) \qquad (106a) \\ 1. & (i = k) \qquad (106b) \end{cases}$$

This equivalent with the Spearman rank-one case of common factor analysis, and we have a communality problem on our hands. This can evidently be generalized to multidimensional parameters. The proper generalization is

$$p_i(\xi) = \sum_{t=1}^{r} \phi_t(\xi)\, \theta_i^t, \tag{107}$$

where the ϕ_t are r functions of the (possibly multidimensional) subject parameters ξ. Using (second order) local independence yields the common factor analysis model of rank r.

As a final example we study a one-dimensional model which has recently received a lot of attention. Suppose

$$p_i(\xi) = \frac{\xi}{\theta_i + \xi} \quad . \tag{108}$$

This is George Rasch's model. We assume that in the population of subjects $\log \xi$ has the logistic distribution with mean $\log \xi_0$. Thus

$$F(\xi) = \frac{\xi}{\xi_0 + \xi}, \tag{109}$$

and the corresponding density is given by

$$dF(\xi) = \frac{\xi_0}{(\xi_0 + \xi)^2}\, d\xi \tag{110}$$

Thus

$$p_i = \int_0^\infty \frac{\xi}{\theta_i + \xi} \frac{\xi_0}{(\xi_0 + \xi)^2}\, d\xi = \frac{\xi_0(\xi_0 - \theta_i) - \xi_0 \theta_i \log(\xi_0/\theta_i)}{(\xi_0 - \theta_i)^2}, \tag{111}$$

if $\theta_i \neq \xi_0$, and $p_i = \frac{1}{2}$ otherwise. By local independece

$$p_{ik} = \int_0^\infty \frac{\xi}{\theta_i + \xi} \frac{\xi}{\theta_k + \xi} \frac{\xi_0}{(\xi_0+\xi)^2}\, d\xi. \tag{112}$$

Defining $\mu_i = \theta_i/(\xi_0 - \theta_i)$ we find for $i \neq k$, $\theta_i \neq \theta_k$, $\theta_i \neq \xi_0, \theta_k \neq \xi_0$

$$p_{ik} - p_i p_k = \mu_i \mu_k \left\{ \xi_0 \frac{\log \theta_i - \log \theta_k}{\theta_i - \theta_k} - \right.$$

$$\left. \left\{ \xi_0 \frac{\log \xi_0 - \log \theta_i}{\xi_0 - \theta_i} \right\} \left\{ \xi_0 \frac{\log \xi_0 - \log \theta_k}{\xi_0 - \theta_k} \right\} \right\} =$$

$$= \xi_0 \mu_i \mu_k (\delta_{ik} - \xi_0 \delta_{i0} \delta_{k0}). \tag{113}$$

In the same notation

$$p_i = (1 - \mu_i)(1 - \theta_i \delta_{i0}). \tag{114}$$

We can test our PCA of binary data on various sets of variables with different ranges of θ, and see how our eigenvectors are related to the θ_i.

In a sense Guttman's deterministic model for binary variables is the basic latent trait model. It can be generalized in at least three directions. In the first place it can be made probabilistic (in the sense that the $p_i(\xi)$ are not step functions any more). This leads directly to Lazarsfeld's polynomial models, to the normal ogive model, and the Rasch-Birnbaum logistic model. In the second place it can be generalized to multi-category nominal items, and to ordered and continuous manifest variates. The general LSA model will be discussed in another publication. Using this general model one can generate all kinds of population models and apply our PCA technique to the values of ϕ_{ik} they yield.

3.21 Alternative scaling requirements

In some situations it seems advisable to maximize λ under the extra conditions that $x_i' D_i x_i = 1$ for all i. If Z is the induced score matrix, S the induced covariance matrix, and R the induced correlation matrix, then our standard technique maximizes $n^{-1} tr(ES)$ under the condition that $tr(S) = n$, and this modification maximizes $n^{-1} tr(ES)$ under the more restrictive conditions that $diag(S) = I$ or, equivalently, it maximizes $n^{-1} tr(ER)$. This gives the stationary equations

$$\Sigma C_{ik} x_k = n \lambda_i D_i x_i, \tag{115}$$

with

$$\lambda_i = \Sigma x_i' C_{ik} x_k, \tag{116}$$

$$\lambda = \Sigma \lambda_i = n^{-1} tr(ER) \tag{117}$$

Again this system has an improper solution with all x_i proportional to e, and all λ_i equal to unity. A simplification is possible by defining $y_i = D_i^{\frac{1}{2}} x_i$. Then (115) reduces to

$$\Sigma \overset{\circ}{C}_{ik} y_k = n \lambda_i y_i, \tag{118}$$

with

$$\mathcal{C}_{ik} = D_i^{-\frac{1}{2}} C_{ik} D_k^{-\frac{1}{2}} \tag{119}$$

In solving (118) we can require that $y_i'y_i = 1$. The equations can be solved by methods which strongly resemble the ordinary power method, but a convergence proof cannot easily be given. We norm all subvectors separately and the usual convergence proof based on the canonical form of the matrix does not apply. In fact one of the principal disadvantages of the method seems to be that a structural algebraic theory comparable to eigenvector-eigenvalue theory does not exist, although the method does lead to a perfectly well-defined optimization problem.

After we have found the absolute maximum we want to maximize λ again, but this time over all sets of vectors z_i satisfying $z_i'z_i = 1$ and $z_i'y_i = 0$. The orthogonality restrictions can be incorporated in the procedure by using generalized inverses. If we define

$$\mathcal{C}_{ik}^{(1)} = (I - y_i y_i') C_{ik} (I - y_k y_k') \tag{120}$$

we do not have to worry about the orthogonality requirements any more, and we can simply solve the stationary equations

$$\mathcal{C}_{ik}^{(1)} z_k = n\lambda_i z_i. \tag{121}$$

If the solution is not orthogonal with y_i we can make it orthogonal by replacing it by $z_i - (y_i'z_i)y_i$ which gives the same value of λ. After this is done we scale back to the metric of the z_i by premultiplying with $D_i^{-\frac{1}{2}}$. We can evidently continue in this way, defining 'residual' matrices $\mathcal{C}_{ik}^{(2)}$, $\mathcal{C}_{ik}^{(3)}$, ... Because

$$(I - y_i^{(s)} \underline{y}_i^{(s)'})(I - y_i^{(s-1)} \underline{y}_i^{(s-1)'}) \ldots (I - y_i^{(1)} \underline{y}_i^{(1)'}) =$$

$$= I - y_i^{(s)} \underline{y}_i^{(s)'} - y_i^{(s-1)} \underline{y}_i^{(s-1)'} - \ldots - y_i^{(1)} \underline{y}_i^{(1)'}, \tag{122}$$

it follows that $\mathcal{C}_{ik}^{(s)}$ vanishes as soon as we have found k_i solutions (including the improper one). If this happens we collaps the matrix and continue with the resulting nonzero submatrices. Other possibilities to find the $y_i^{(s)}$ are the successive method without deflation but with orthonormalization with respect to the previous solutions, the successive method with arbitrary orthonormal completion, and the simultaneous method with orthonormalization of the set after each powered-matrix iteration.

Although the scaling method discussed in this section seems more natural for several important special cases of the theory in this dissertation I would not recommend it in general. This is not only because of the lack

of structural algebraic properties but mainly because most of the interpretations in the previous sections (and most of the interpretations of principal components in general) are no longer valid any more. The basic difficulty with the method, and the reason why it leads to nonstandard matrix algebra, seems to be that our essentially bivariate treatment of the problem conflicts with the multivariate scaling requirements $X_i'D_iX_i = I$ for all i, except in the special cases with n=2 or with exact multinormality. In this last case the analysis in 3.8 shows that we find the same solutions in a more compact form. In the nonparametric generalizations of multivariate analysis mentioned the first chapter we do use requirements $X_i'D_iX_i = I$ (without using the ideas of canonical analysis) and there they cause no trouble at all. Again it can be proved that in the multinormal case the results are essentially identical to these in 3.8, although again they come out in quite a different form.

3.22 On the interpretation

The question whether it is appropriate to analyze categorical data using principal component methodology is somewhat ambiguous. In an algebraical sense it most certainly is as appropriate as for ordinary product moment matrices: our C is always positive semidefinite. In a psychometric sense there is no problem too, the concepts of homogeneity and discrimination apply as easily here as in the classical context, and the geometry of the problem is also very similar. Statistically the use of canonical variates in this context is as defendable as in any other context, and from the point of view of reproducibility again there is no difference. Consequently we can summarize our answer as: the theory of principal components can be generalized from the linear, numerical case to the situation in which some of the variates are categorical and/or ordinal; the use of PCA in this context is as appropriate as in any other context. There is, however, a more important question from the practical point of view. Can we interpret the results of our PCA of categorical data in the same way as the results of PCA for numerical data? We have taken the point of view in this study that all data are basically categorical, and that for some variates there is additional prior (ordinal, numerical) information. This information can be incorporated in the analysis by using it to define suitable restrictions on the weights. There is nothing spectacular or special about that, we can also refuse to use this prior information and treat all variables as if they were only nominal. Just as in the classical case PCA is a linear model, in the sense of internal consistency equations (25), and

in the sense that the successive components are combined in a linear way in the reproducibility equations (31) and (32). As pointed out by Bartlett (1953) and McDonald (1968) in the classical case we suppose, more or less implicitly, that the component scores are stochastically independent. If they turn out to be related by relatively simple nonlinear functions, then a more parsimoneous description can be obtained by using nonlinear principal component analysis (NLPCA). Again we must be careful not to confuse this with McDonald's nonlinear factor analysis (NLFA), we mean by NLPCA the technique outlined by Carroll (1969, 1971) under the somewhat regrettable name of polynomial factor analysis. What we can expect in the case of categorical PCA is that these nonlinearities will become more frequent (cf the discussion of the multinormal case and the perfect scale). Consequently the question whether we can interpret our components in the usual way can be answered in the affirmative. We only have to be careful, maybe even more careful than in the analysis of numerical variates. In the numerical case nonlinear systematic relationships between the component scores also mean that we cannot interpret them in the usual, naive manner. Let us apply our PCA to the analysis of binary variates for which both the Rasch model and the postulate of local independence hold. Then we have a nonlinear one-dimensional model, and our weighting functions will be orthonormal functions which are the components of the nonlinear regression of the probability of a positive response on the single latent trait. A standard interpretation of these results (i.e. a standard PCA of the matrix of phi-coefficients, with varimax rotation and the like) will be extremely misleading. In fact it will lead to many of the so-called difficulty factors, which have baffled psychometricians for a long time (cf McDonald 1965). Their conclusion was that the phi-coefficient is not the proper correlation coefficient in this case (Henryssen & Thunberg 1965). This is obviously the wrong conclusion. If we can conclude from our results that the Rasch model (or the normal ogive or arcsine model) holds , we can find estimates of all the relevant item parameters. Nevertheless it is of course perfectly true that, in the multinormal case for example, if we had a proper procedure to estimate the population correlation coefficients the results would come out more directly. We can use the tetrachoric or polychoric correlation coefficients (Lancaster & Hamdan 1964), but this has several drawbacks. In the first place we assume outright that we are dealing with a multinormal situation, in the second place some of the smaller roots

of the dispersion matrix may be negative (although I don't think that to be very serious). Consequently the question whether we should use tetrachorics or phi-coefficients in binary PCA can be very easily solved: using phi-coefficients should tell us whether the use of tetrachorics is justified. This is, of course, an idealized statement. It presupposes that there are no serious sampling errors, and that the multinormal effects can be perfectly detected. The same thing is true for general categorical data. In the cases where the multinormal effect is present, one could use the polychoric series to estimate ρ_{ij} for each pair of variables. If some of the variables do not fit in the multinormal pattern, further analysis is required. In the case of the perfect scale the regression line on the one-dimensional latent continuum is a step function, and the principal components are the orthogonal harmonics of this step function (Guttman 1950b, McDonald 1968). In this final example the standard interpretation of a PCA of phi-coefficients is again very misleading, but a proper look at the results will show where the jumps of the step functions are located, and consequently will give estimates of all the relevant parameters. The use of tetrachorics in this context is obviously absurd (cf Guttman 1950a).

3.23 Some criticisms

A critical discussion of the techniques outlined in this chapter must take into account the criticisms that have been used against classical PCA. In the first place people have questioned PCA because it does not explicitely involve a statistical model with parameters which are estimated by some conventional technique. It merely is a transformation of the data, and because it is just a transformation it is difficult to formalize the idea of some kind of error. And, since the idea of error is not very clear, it is difficult to give a rigorous justification of the idea of dropping the smaller eigenvalues. In a sense these criticisms are justified. PCA does not fit a statistical model, it is a purely algebraic transformation of the data. But this does not imply in any way that it cannot be valuable as a technique in the exploratory phase of the investigations (and a large part of the social sciences is still in such a phase). PCA must be seen as a tool helpful in structuring or even in plotting the data. A very readible account of its use in this context is Gnanadesikan & Wilk (1969), and another persuasive defense of PCA as a useful exploratory technique is given by Rao (1965).

It is well known that principal components have a large number of interesting optimality properties. Our homogeneity and discrimination indices are optimized by principal components. In the numerical case this has already been shown by Horst (1936), Edgerton & Kolbe (1936), and Wilks (1936). A large number of other optimality properties, starting with the ones discussed by Pearson (1901), are collected in Rao (1965), and Okamoto & Kanazawa (1968), and Okamoto (1969). They are very useful in this connection. In fact each optimality property of PCA can be used to define a new rationale for our approach, and implies a new way of looking at our weights and scores. The optimality properties are derived, generally, for the components corresponding with the $p \geq 1$ largest eigenvalues. A criticism of this approach is contained in Bargmann (1969). He emphasizes the fact that a principal component is a mathematical artefact, not an observed variable, and not an underlying, latent, true physical dimension. The first principal component (PC1) usually has a clear optimality interpretation, but the second one optimizes the same criterion after the elimination of this artefact. Bargmann doubts whether this is a wise procedure, and suggests to proceed stepwise in stead. We find PC1 for all the variables, find the correlations of the variables with PC1, form a subset of the variables with low correlations, compute PC1 for this subset, and so on. This may be interesting, but I doubt whether it is a true improvement. In fact I do not believe that PC1 has a different logical, mathematical, and/or statistical importance and/or meaning than the other components. PC1 gives the best rank-one optimalization, PC1 and PC2 the best rank-two optimalization, and so on. In this sense PC2 is not the optimal solution after the elimination of an artefact, but the pair (PC1, PC2) is a new artefact. The fact that the first component of this new artefact is the same as the previous artefact is a mathematical property of principal components, which does not hold, for example, in the multi-way generalizations discussed by Carroll & Chang (1970). Bargmann's hierarchical procedure may have some advantages, but we lose so much that I don't consider the gamble worthwhile (for example what happens to the geometrical interpretations we have considered?) Of course we fully agree with Bargmann that PCA is an exploratory technique which works with artificial variables, and that it is a common and very serious mistake to

consider these artefacts as physical realities. Observe, moreover, that in the ideal case in which each of the variables correlates with one and only one component the procedures of Bargmann and the deflation method are identical. This is the familiar independent cluster case of simple structure theory.

It is also possible to attack the whole idea of quantification. It can be said that the idea of representing a category by a real number implies that we suppose that basically (in some sense) we are dealing with quantitative variates; only, we do not know exactly what the correct quantifications are. This point is made by Lubin (1950), when he discusses the technique we discussed in section 3.10. I think this criticism is invalid for various reasons. In the first place it does not answer the question whether it is legitimate to quantify the sample elements. If it is (and I do not see how any psychometrician can deny that) the optimal direct weights follow by internal consistency as averages of scores in a category (and again I do not see how anybody can object to computing these averages). In the second place we have seen that we can interpret our weights as defining a real valued function on the set of categories, but we can also interpret them as defining separation boundaries in the space of scores. Finally, we have defined a set-theoretical distance measure in section 3.12, which we want to represent in low - dimensional Euclidean space in an optimal way. Again Lubin's criticism (like Bargmann's) seems to underestimate the value of exploring the data from several angles, using all kinds of transformations and representations. The point is not that we are, erroneously, representing categories as points on a linear continuum (this is only one of the possible interpretations), the point is how we arrive at these points and in what ways we can interpret them. If someone objects to the weights, then ask him if he objects to the scores. If he does not, apply the analysis and the weights come out anyway, and in such a way that he can hardly object to them any more. If he objects to both, ask if he objects to the geometry. If he does not, apply the analysis, and both weights and scores come out again. Moreover it is hard to see how anybody can object to our scores and use ordinary PCA or FA, or to our separation geometry and use MSA, or to our distance geometry and

use multidimensional scaling. We have already discussed a perfectly valid objection to our PCA in the previous section. For some particular cases the results may be misleading (essentially because we stick to a linear, bivariate model where this is not optimal). Because these cases are theoretically very important as normative models we can expect that in almost all cases some of these disturbing effects will be present. Our task is to detect them, and to apply a different technique if that really seems worthwhile. As indicated in the previous section there has been some controversy about the question whether one should 'factor' qualitative data or not. People like Guttman (1950a,b,1953) and Lazarsfeld(1950) have defended the view that FA and PCA are designed for quantitative variates only, while Burt (1950, 1953) has pleaded to treat the two cases alike, whenever possible. To defend this point of view Burt refers to such examples as the Stieltjes integral, the Daniels-Kendall generalized correlation coefficient, or modern probability theory but I think these examples miss the point. The main issue is that in the qualitative case we have only frequencies, in the quantitative case we have both scores and frequencies (and in the binary case scores and frequencies happen to be the same thing). Our categorical PCA reproduces frequencies, and partitions the frequency criterion X^2. Ordinary PCA partitions squared correlations, and reproduces scores and frequencies. Our point of view is that in the nominal case we can use as a substitute for the prior scores the posterior canonical weights and scores, which also makes X^2 into a sum of squared correlations. Guttman (1950a) already observed that factoring is joint bivariate technique, and that categorical data in general need multi-variate techniques. Partly the difference of opinion between Guttman and Burt is due to the deplorable confusion over the word factor analysis which still seemed to exist in the 1950's, but this cannot be the main issue any more. More recently Anderson (1959) and McDonald (1962) have constructed general latent structure models, which are based only on the assumption of local independence, and which apply to all kinds of latent and manifest spaces. Factor analysis is an incomplete version of a very special LSA model (incomplete exactly because it considers only the variances and covariances, not moments of higher order).

Another controversy between Burt and Guttman was if one should only consider PC1 or also the other components. Guttman argued for PC1 only, Burt advocated a complete PCA. One can ascribe this to different objectives: Guttman was trying to scale a one-dimensional attitude, Burt was factoring by weighted summation (i.e. doing a PCA). One could also argue that Guttman was extrapolating his experience with the perfect scale which is really one-dimensional (only one latent variable). Consequently, Guttman was dealing with a one-dimensional model with nonlinear regressions on the latent variable(step functions), Burt was thinking in terms of a linear model (McDonald 1969). Guttman was well aware of the dangers involved in interpreting PCA-results routinely if non-linear regressions are present, Burt seems to step over these difficulties somewhat too lightly. Since then, however, McDonald (1968) has argued that nonlinear regressions are often present in numerical data as well, and he has conjectured that it is quite probable that many of the usual routine interpretations of FA and PCA published in the literature (with simple structure relations and the like) may be questionable because of these non-linear effects.

Another objection, which may be heard from the more advanced nonmetric devotees, is why we bother with principal component type error theories if there are such splended new techniques as GL-MSA-I, and so on. In fact Guttman (1968), who used a technique equivalent to our PCA as early as 1941, now admits that he realized from the start that this was only approximate. This statement is important, in the first place because it comes from a psychometrician who ranks with Spearman, Thurstone and Burt; in the second place because the 'nonmetric' methods are becoming more and more popular. From one point of view these methods can be classified as (a) definitely very useful such as the additive and linear programs (Kruskal 1965, De Leeuw 1969a), (b) probably quite useful such as the standard MDSCAL an MINISSA programs for the complete case, (c) not very satisfactory such as the unfolding programs of Kruskal and Roskam, and as (d) rather dangerous such as the GL-MSA programs. The linear and additive models are straightforward extensions of standard statistical models, and using normal theory assumptions they produce maximum likelihood estimates of both the parameters of the models and the monotonic transformation. The standard

MDSCAL and MINISSA programs are, in their latest versions, reasonably well-behaved with respect to local minima, and the problem they try to solve is reasonably well-defined with respect to uniqueness (De Leeuw 1970b). Consequently we may consider them as useful data-reduction techniques for similarity data, although further research is needed to compare their performances with methods which apply the standard metric techniques to a set of conventional numbers (such as rank numbers, or χ^2-order statistics) The techniques for unfolding and multidimensional scalogram analysis are dangerous because they put relatively few restrictions on the data, so that a large set of quite different perfect solutions exists. Moreover the algorithms as such will concentrate on possible degenerate solutions in order to minimize their loss function, and one never knows how influential this degenerating effect is (cf De Leeuw 1970a, 1970b).

Nonmetric techniques (this is, by the way, a very unsatisfactory name, cf De Leeuw 1970a) for the analysis of indicator matrices have been produced by Guttman and Lingoes (Lingoes 1967), and by De Leeuw (1969b). It is possible to regard our PCA as an approximation to the solution of these iterative programs (as Guttman seems to do), it is also possible to regard it as an independent solution of the same problem using a different type of error theory. In the terminology of De Leeuw (1971a) iterative programs like MSA-I and MSA-II have a loss function and a solution map which are strongly consistent, in our PCA the loss function $1 - \lambda$ and the solution map are weakly consistent. This loss of consistency is compensated by a gain in determinateness, a gain in structural mathematical properties, and a gain in alternative possibilities of interpretation. I would consider Guttman's claim that he realized from the start that these methods are only approximate as somewhat premature from the point of view of methodology.

3.24 Historical remarks

Classical PCA has its mathematical roots in algebraic eigenvalue theory, which dates back to Euler, Cauchy, Jacobi, Cayley, and Sylvester. The first statistical use of these algebraic results is Karl Pearson's work on dispersion matrices (1900), but as a data reduction technique PCA must be credited to Hotelling (1933).

British psychologists like Burt and Thompson advocated it as one of the major approaches to factor analysis. The fact that PCA can be derived from ANOVA-type homogeneity and discrimination criteria was already known to Horst (1936), Edgerton & Kolbe (1936), and Wilks (1938). The pioneering paper in the analysis of categorical variates is Guttman (1941). It is a direct logical extension of the work of Horst et al to nominal attributes and it points out the relevance of Hotelling's canonical analysis (1936). Guttman derived (by using correlation ratio's) equations (9) and (21), pointed out the nature and role of improper solutions, derived the 'equations of internal consistency' (25), and investigated the relationships with X^2. In a later series of papers (1950a, 1954, 1955), he considered the application of these techniques to perfect scales and derived the results mentioned briefly in section 3.17. Especially the 1950 paper contains a brilliant mathematical exposition of the principal component properties of perfect scales. In a final paper in the series (1959a) he sketched a system of scaling methods based on facet theory, which is quite similar to the basic idea of this dissertation. The application of PCA to categorical data was discovered, independently, by Burt(1950). Compare also the discussion in Guttman (1953), Burt (1953). Useful additional references are Mosteller (1949), Lord (1958), Torgerson (1958, p 338-345), Bock (1969), Lingoes (1963,1968), and McDonald (1968). The special case n = 2 has a complicated history. Various aspects of this problem are studied by Hirschfeld (1936) who discovered (56), Fisher (1941) who contributed (52), and Lancaster (1957, 1958) who studied the more general case leading to (54) and (55). Benzécri considered special geometrical interpretations for n = 2 (cf Cordier 1963). The case n = 2 is reviewed in a recent paper of De Leeuw (1971b) and the work of Lancaster and his associates on canonical decompositions of discrete and continuous probability distributions is reviewed in Lancaster's recent book (Lancaster (1969)). Our systematic interpretation of ordinal and numerical variates in this framework is possibly new. The result in section 3.8 is new, but for n = 2 these results have been discovered by Maung (1941) and Lancaster (1957). The geometrical interpretation in 3.12 is new although inspired by Benzécri - Cordier. The results in 3.14 are due to Yates (1948) and Williams (1952). The history of the Helmert matrices (82) is reviewed by Lancaster (1965). They were applied to the partition of X^2 by Lancaster (1949) and Irwin (1949). Section 3.19 is closely related to some work by Bradley, Katti, and Coons (1962).

The problem in 3.19 can also be treated by using monotone regression (Kruskal 1965). Equations (94) and (95) were discussed in another interesting paper by Guttman (1959b). Section 3.20 is new in some respects, section 3.21 is based on some work of Horst (1961a,b, 1965). Closely related discussions of the scaling requirements and of other possible objective functions are Steel (1951), Guttman (1959b), McDonald (1968), Van de Geer (1968), Carroll (1968), and Kettenring (1971). Computational aspects of procedures like the ones in 3.16 were investigated by Doesborgh (1971), who also wrote a number of computer programs in FORTRAN for the 360 series.

4 Differencing models

4.0 Introduction

In the previous chapter we have discussed some techniques for incorporating prior ordinal information in a PCA. In this chapter we describe an alternative way of dealing with ordinal information of a somewhat different nature. In particular the techniques in this chapter deal with paired comparison data (or with ordinal data which are reduced to paired comparison form).

4.1 Paired comparisons

Consider a paired comparison experiment with n stimuli and m subjects (or occasions). This defines $N = n^2$ binary variables (forced choice procedure, also for diagonal pairs). The elements of the indicator matrix H can be coded as $h_{+j}^{(i,k)}$ and $h_{-j}^{(i,k)}$, with $h_{+j}^{(i,k)} = 1$ iff subject j prefers i to k (or thinks that i is heavier than k, or what have you). Applying the usual procedure for binary variables from 3.15 quantifies the set of all N pairs, which is not what we usually want. The familiar stochastic theories for paired comparison experiments suggest using the restrictions

$$x_{+}^{(i,k)} = y_i - y_k, \tag{1a}$$

$$x_{-}^{(i,k)} = y_k - y_i, \tag{1b}$$

for all i, k = 1, ...,n. Observe that this implies

$$\sum_j z_j^{(i,k)} = n_{+}^{(i,k)}(y_i - y_k) - n_{-}^{(i,k)}(y_i - y_k) =$$

$$= \left[n_{+}^{(i,k)} - n_{-}^{(i,k)} \right](y_i - y_k), \tag{2}$$

which is usually not equal to zero. We investigate the optimal direct quantification of the variables under the conditions (1), i.e. the finding of an optimal y. We find for the induced score vector

$$z_{.j} = \frac{1}{N} \sum_i \sum_k \left[x_{+}^{(i,k)} h_{+j}^{(i,k)} + x_{-}^{(i,k)} h_{-j}^{(i,k)} \right] =$$

$$= \frac{1}{N} \sum\sum (y_i - y_k)\left[h_{+j}^{(i,k)} - h_{-j}^{(i,k)} \right] \tag{3}$$

This can be simplified to

$$z_{.j} = N^{-1} \sum y_i (a_{ij} - b_{ij}) = \sum y_i c_{ij}, \tag{4}$$

with

$$a_{ij} = \Sigma \left[h^{(i,k)}_{+j} - h^{(i,k)}_{-j} \right], \tag{5a}$$

$$b_{ij} = \Sigma \left[h^{(k,i)}_{+j} - h^{(k,i)}_{-j} \right], \tag{5b}$$

$$c_{ij} = N^{-1}(a_{ij} - b_{ij}). \tag{5c}$$

Moreover, using (1),

$$S_T = \Sigma\Sigma\Sigma(z^{(i,k)}_j)^2 = \Sigma\Sigma(y_i - y_k)^2 \Sigma(h^{(i,k)}_{+j} - h^{(i,k)}_{-j})^2 =$$

$$= m \Sigma\Sigma(y_i - y_k)^2 = 2m\left[n\Sigma y_i^2 - (\Sigma y_i)^2 \right] = n\, y'Qy, \tag{6}$$

with

$$q_{ik} = \begin{cases} \frac{2m(n-1)}{N} & \text{if } i = k, \\ \frac{-2m}{N}, & \text{if } i \neq k. \end{cases} \tag{7}$$

Clearly

$$z_{\cdot\cdot} = y's, \tag{8}$$

with

$$s_i = m^{-1}\Sigma c_{ij}. \tag{9}$$

Applying our usual procedures we find

$$B = Ny'CC'y, \tag{10a}$$

$$W = Ny'(Q - CC')y \tag{10b}$$

$$T = Ny'Q\,y, \tag{10c}$$

provided

$$z_{\cdot\cdot} = y's = 0 \tag{11}$$

It is obvious from (5) that $\sum_j a_{ij} = \sum_i b_{ij}$, and thus $\sum_i c_{ij} = 0$. Consequently CC' is doubly centered, and so is Q. It follows that we can require $y'e = 0$, and that we can replace Q by $\hat{Q} = \frac{2mn}{N} I$.

4.2. An alternative approach

An alternative and somewhat simpler approach can be based on a slightly different way to define within and between. Define

$$t^{+}_{ij} = \Sigma h^{(i,k)}_{+j}, \tag{12a}$$

the number of times subject j judges i higher than something else, and

$$t^{-}_{ij} = \Sigma h^{(k,i)}_{+j}, \tag{12b}$$

the number of times subject j judges something else higher than i. The weighted mean of the scale values of the things subject j ranks higher than

other things is

$$u_j^+ = \Sigma t_{ij}^+ y_i \,/\, g_j, \tag{13a}$$

with $g_j = \Sigma t_{ij}^+ = \Sigma\, t_{ij}^-$. In the same way the weighted mean of the scale values of the thing subject j ranks lower than the other things is

$$u_j^- = \Sigma t_{ij}^- y_i \,/\, g_j. \tag{13b}$$

The corresponding weighted sums of squares around these means are

$$s_j^+ = \Sigma(y_i - u_j^+)^2\, t_{ij}^+ = \Sigma y_i^2 t_{ij}^+ - g_j(u_j^+)^2, \tag{14a}$$

$$s_j^- = \Sigma(y_i - u_j^-)^2\, t_{ij}^- = \Sigma y_i^2 t_{ij}^- - g_j(u_j^-)^2. \tag{14b}$$

By letting

$$e_i = \Sigma(t_{ij}^+ + t_{ij}^-), \tag{15}$$

$$e = \Sigma e_i, \tag{16}$$

we can define the overall weighted mean by

$$u = \Sigma e_i y_i \,/\, e, \tag{17}$$

and the sum of squares

$$s = \Sigma(y_i - u)^2\, e_i = \Sigma y_i^2 e_i - eu^2. \tag{18}$$

We now make the partition

Source	Sum of Squares	
Within columns	$W = \Sigma(s_j^+ + s_j^-)$	
Between columns	$B = \Sigma g_j \left\{(u_j^+ - u)^2 + (u_j^- - u)^2\right\}$	(19)
Total	$T = \Sigma(y_i - u)^2\, e_i$	

and we maximize B/T under the condition u = 0. In matrix notation we can write

$$T = y'Ey, \tag{20a}$$

$$B = y'(T_+D^{-1}T_+' + T_-D^{-1}T_-')y \tag{20b}$$

with T_+ containing the t_{ij}^+ and T_- the t_{ij}^-, while D and E are diagonal matrices having diagonal entries G_j and e_i, respectively. Again there is an improper solution y = e, and all other solutions satisfy u = 0 automatically.

4.3 Some simplifications & complications

The analysis of section 4.1 becomes slightly more complicated if we admit tied responses (equal; I don't know; not definitely larger). Each variate now has three possible values (+, -, and 0) and we require

$$x_+^{(i,k)} = y_i - y_k, \tag{21a}$$

$$x_0^{(i,k)} = 0, \tag{21b}$$

$$x_{-}^{(i,k)} = y_k - y_i, \tag{21c}$$

The formulas for $z_{.j}$ and $z_{..}$, and for a_{ij}, b_{ij}, c_{ij}, remain valid. The expression for the total sums of squares changes to

$$S_T = \Sigma\Sigma(y_i - y_k)^2 \Sigma(h_{+j}^{(i,k)} - h_{-j}^{(i,k)})^2 =$$
$$= \Sigma\Sigma(y_i - y_k)^2 (n_{+}^{(i,k)} + n_{-}^{(i,k)}). \tag{22}$$

It is easy to derive matrix expressions for this sum. Section 4.2 remains valid. Another complication arises in incomplete paired comparison experiments in which we do not compare all pairs, but only a specific set. This also includes cases in which a specific set is replicated. Nevertheless these complications are only relatively slight if the same pairs are missing for each subject. General formulas can easily be derived for balanced incomplete designs, which are familiar from the literature. There is one important kind of incompleteness which must be mentioned. If we only investigate the $\binom{n}{2}$ pairs (i,k) with i > k then we can fill in the $\binom{n}{2}$ pairs (i,k) with i < k by assuming that the subject would have given the opposite judgment here, and the n pairs (i,i) by assuming that the subject would always give tied responses here. This amount to

$$h_{+j}^{(i,k)} = 1 \quad \text{iff} \quad h_{-j}^{(k,i)} = 1 \tag{23a}$$

$$h_{0j}^{(i,k)} = 0 \quad \text{iff} \quad h_{0j}^{(k,i)} = 0 \tag{23b}$$

$$h_{0j}^{(i,i)} = 1 \tag{23c}$$

for all i,k,j which implies

$$n_{+}^{(i,k)} = n_{-}^{(k,i)}, \tag{24a}$$

$$n_{0}^{(i,k)} = n_{0}^{(k,i)}, \tag{24b}$$

$$n_{0}^{(i,i)} = m, \tag{24c}$$

for all i,k, and

$$a_{ij} = -b_{ij}, \tag{25}$$

for all i,j. If in addition

$$h_{0j}^{(i,k)} = 0 \text{ iff } i \neq k \tag{26}$$

then

$$h_{+j}^{(i,k)} = 1 - h_{-j}^{(i,k)} = h_{-j}^{(k,i)}, \tag{27}$$

for all $i \neq k$ and all j. The usual paired comparisons experiments satisfy both 23a and 26 (which obviously implies 23b and 23c). In this case the treatment in 4.2 can be simplified by observing that

$$t^-_{ij} = (n-1) - t^+_{ij}, \tag{28a}$$

$$c_{ij} = 4t^+_{ij} - 2(n-1). \tag{28b}$$

which means that the technique in 4.2 is equivalent to maximizing $y'CC'y$ on the condition $y'y = 1$.

An even more important simplification is possible if there exists a weak order $\succsim_j$ on the set of stimuli such that

$$h^{(i,k)}_{+j} = 1 \quad \text{iff} \quad i \succ_j k, \tag{29a}$$

$$h^{(i,k)}_{0j} = 1 \quad \text{iff} \quad i =_j k, \tag{29b}$$

$$h^{(i,k)}_{-j} = 1 \quad \text{iff} \quad i \prec_j k. \tag{29c}$$

This implies that

$$h^{(i,k)}_{+j} - h^{(i,k)}_{-j} = \sigma_j(i,k) \tag{30}$$

with

$$\sigma_j(i,k) = \begin{cases} 1 & \text{if } i \succ_j k, \\ 0 & \text{if } i =_j k, \\ -1 & \text{if } i \prec_j k. \end{cases} \tag{31}$$

Of course in this case we also have $a_{ij} = -b_{ij}$, but more explicit statements are possible. We have

$$a_{ij} = \Sigma\sigma_j(i,k) = 2\rho_{ij} - (n+1), \tag{32}$$

where ρ_{ij} is the rank number of stimulus i in the weak order $\succsim_j$ (ties get the average of the available rank numbers). Observe that $\Sigma_j \rho_{ij} = \frac{1}{2}n(n+1)$, and thus $a_{ij} = 2(\rho_{ij} - \rho_{\cdot j})$. The result (32) simplifies both the computations and the interpretation. Moreover it is now obvious how to apply the same techniques to experiments in which the stimuli are ranked. If $\succsim_j$ is a partial order the situation is more complicated and quite similar to the situation in which any number of pairs is missing (the same pairs for each subject). This includes all the rank k/n and pick k/n data collection methods. A further simplification, which is frequently useful, obtains if the set of stimuli can be divided into two subsets A_j and B_j (having $N(A_j)$ and $N(B_j)$ elements), and

$$i >_j k \quad \text{iff} \quad i\varepsilon A_j \wedge k\varepsilon B_j, \tag{33a}$$

$$i =_j k \quad \text{iff} \quad i,k\varepsilon A_j \vee i,k\varepsilon B_j, \tag{33b}$$

defines a weak order. In this case there exists a binary vector u_j such that $u_i^j = 1$ for all $i\varepsilon A_j$, and $u_i^j = 0$ for all $k\varepsilon B_j$. Obviously

$$a_{ij} = \begin{cases} N(B_j) & \text{if} \quad i \in A_j, \\ -N(A_j) & \text{if} \quad i \in B_j. \end{cases} \tag{34}$$

4.4 Relationships

We have discussed two different techniques in the previous sections. In the case where (23a) and (26) are true technique A, based on (1), reduces to maximization of y'CC'y on the condition that y'y = 1 and y's = 0. Technique B, described in 4.2, maximizes y'CC'y on the condition that y'y = 1. The relationships between the techniques A and B are rather obvious. In the first place it is true in most cases that the dominant eigenvector of CC' is very much like s. If this is true the two techniques give highly similar results all the way. In the second place multidimensional solutions of B can always be made to satisfy y's =0 approximately by orthogonal target rotation and dropping one dimension. Observe that the vector s is the best least squares approximation to the columns of C simultaneously. In a sense we take the first centroid in technique A and compute the principal components of the residual. It is also possible to compare A and B with the techniques known as 'nonmetric factor analysis'. There we try to find an optimal monotonic transformation and reduce by using principal components as usual, while in A and B we use a prior, fixed monotonic transformation. The relationship with classical paired comparison analysis is also interesting. In the classical analysis using discriminal dispersions and the like object i gets score x_{ij} from subject j. Each subject defines a random observation from population i, and we want to estimate the mean over subjects. We use a least squares type technique which minimizes the sum of squares of the deviations of the differences, i.e. we want to find y_i and w_j such that

$$S = \Sigma\Sigma\Sigma\left[(x_{ij} - x_{kj}) - w_j(y_i - y_k)\right]^2 \tag{35}$$

is as small as possible. The stationary equations can be written as

$$\bar{X}\, w = \alpha y, \tag{36a}$$

$$\bar{X}'y = \beta w, \tag{36b}$$

with $\bar{X} = \{\bar{x}_{ij}\} = \{x_{ij} - x_{.j}\}$, $\alpha = w'w$ and $\beta = y'y$. It follows that y is the eigenvector of $C = \bar{X}\bar{X}'$ corresponding with the largest root.
Taking expected values we find

$$E(c_{ik}) = m(\sigma_k - \sigma_{i.} - \sigma_{.k} + \sigma_{..}) + m(\mu_i - \mu_.)(\mu_k - \mu_.) \qquad (37)$$

where we have used the familiar symbols for means and covariances.
It is obvious from (37) that when σ_{ik} is constant for all i, k, then the dominant eigenvalue of C is proportional to $(\mu_i - \mu_.)$.
If we knew the values of the x_{ij} then the first eigenvector would be a consistent estimate of the means (without assuming normality of logisticity).
In analysis B we substitute $x_{ij} = \rho_{ij}$ and perform the eigen-analysis. In the same way as in section 3.20 it seems very useful to relate our techniques to stochastic latent structure type models like, for example,

$$p_j(i > k) = \frac{1}{1 + \exp(-\xi_j(\theta_i - \theta_k))}. \qquad (38)$$

4.5 Further generalizations.
An obvious and important generalization is obtained by considering paired comparisons of sets of stimuli. A comparison of the sets S, T defines a vector of weights $x^{S,T}$ which can be restricted by

$$x^{(S,T)} = \sum_{i\in S} y_i - \sum_{i\in T} y_i, \qquad (39a)$$

$$x^{(S,T)}_{-} = \sum_{i\in T} y_i - \sum_{i\in S} y_i, \qquad (39b)$$

$$x^{(S,T)}_{0} = 0. \qquad (39c)$$

In all these models we can define vectors $g^{(S,T)}$ in such a way that all elements are equal to -1, 0, 1 and $x^{(S,T)}_{+} = y'g^{(S,T)}$. The n x 3 matrix $G^{(S,T)}$ is defined as

$$G^{(S,T)} = \left(g^{(S,T)} \mid 0 \mid -g^{(S,T)} \right). \qquad (40)$$

Then

$$(x^{(S,T)})' = y'G^{(S,T)}, \qquad (41)$$

and

$$(z^{(S,T)})' = y'G^{(S,T)}H^{(S,T)}. \qquad (42)$$

Define

$$P = \sum_{(S,T)} G^{(S,T)}H^{(S,T)}(H^{(S,T)})'(G^{(S,T)})', \qquad (43a)$$

$$Q = \left[\sum_{(S,T)} G^{(S,T)}{}_H{}^{(S,T)} \right] \left[\sum_{(S,T)} G^{(S,T)}{}_H{}^{(S,T)} \right], \quad (43b)$$

and

$$\lambda = \frac{y'Qy}{y'Py}. \quad (43c)$$

Then λ is to be maximized under the condition that

$$y' \sum_{(S,T)} G^{(S,T)}{}_H{}^{(S,T)} e = y's = 0 \quad (44)$$

if we use technique A. If we use B we do not require (44). Of course $H^{(S,T)} (H^{(S,T)})' = D^{(S,T)}$ is diagonal. Simplifying somewhat further

$$P = \sum_{(S,T)} t^{(S,T)} (g^{(S,T)})(g^{(S,T)})', \quad (45)$$

with

$$t^{(S,T)} = n_+^{(S,T)} + n_-^{(S,T)}. \quad (46)$$

More generally

$$Q = \sum_{(S,T)} \sum_{(\bar{S},\bar{T})} G^{(S,T)}{}_H{}^{(S,T)} (H^{(\bar{S},\bar{T})})' (G^{(\bar{S},\bar{T})})' =$$

$$= \sum_{(S,T)} \sum_{(\bar{S},\bar{T})} G^{(S,T)} C^{(S,T)(\bar{S},\bar{T})} (G^{(\bar{S},\bar{T})})' =$$

$$= \sum_{(S,T)} \sum_{(\bar{S},\bar{T})} t^{(S,T)(\bar{S},\bar{T})} g^{(S,T)} (g^{(\bar{S},\bar{T})})', \quad (47)$$

with

$$t^{(S,T)(\bar{S},\bar{T})} = c_{11}^{(S,T)(\bar{S},\bar{T})} - c_{13}^{(S,T)(\bar{S},\bar{T})} +$$

$$c_{33}^{(S,T)(\bar{S},\bar{T})} - c_{31}^{(S,T)(\bar{S},\bar{T})}, \quad (48)$$

and in particular

$$t^{(S,T)(S,T)} = n_+^{(S,T)} + n_-^{(S,T)}. \quad (49)$$

Finally

$$s_i = \sum_{(S,T)} g_i^{(S,T)} \sum_j (h_{+j}^{(S,T)} - h_{-j}^{(S,T)}). \quad (50)$$

In the more common cases we have $e'g^{(S,T)} = 0$ for all (S,T). Then again both P and Q are singular. In case $g^{(S,T)}$ contains exactly one element equal to +1 and one equal to -1 the formulas reduce to the ones in the previous sections. This happens, for example, if S and T are of the form $U \cup \{i\}$ and $U \cup \{k\}$ with $i \neq k$ and U arbitrary (the fact that the set U does not matter is a familiar axiom in some stochastic theories of choice behaviour). We have paired comparisons if $U = \emptyset$.

The connections of the theory in this section with the theory of additive conjoint measurement and with the additive cardinal utility theory of commodity bundles is clear.

As a final generalization we mention the case in which the compared objects are arbitrary vectors of real numbers with the same dimensionality n. We require

$$x_{+}^{(i,k)} = \sum_{p=1}^{n} y_p(g_p^{(i)} - g_p^{(k)}), \tag{51a}$$

$$x_{-}^{(i,k)} = \sum_{p=1}^{n} y_p(g_p^{(k)} - g_p^{(i)}), \tag{51b}$$

$$x_{0}^{(i,k)} = 0 \tag{51c}$$

With obvious modifications all formulas of this section remain valid with this general form of g. Again this is easily seen to be related to ordinal multiple linear prediction.

4.6 Maximum sum techniques.

It can be argued that the essential part of our homogeneity coefficients is the numerator, i.e. the same numerator with a different denominator may also produce reasonable loss functions (although not necessarily variance ratios). This is the basic idea behind the maximum sum principle. If we have a 'nonmetric' theory of the type

$$i \underset{j}{\gtrsim} k \quad \text{iff} \quad \phi_j(y_i) \geq \phi_j(y_k), \tag{52}$$

then our numerator can be written as (compare formula 4)

$$\Sigma\Sigma\Sigma \left[h_{+j}^{(i,k)} - h_{-j}^{(i,k)}\right] \left[\phi_j(y_i) - \phi_j(y_k)\right]. \tag{53}$$

Add a scaling requirement which makes the set of all solutions bounded and maximize. Intuitively this means that we replace (59) by the weaker set

$$i \succ_j k \quad \rightarrow \quad \phi_j(y_i) \geq \phi_j(y_k), \tag{54}$$

which is equivalent to

$$(h_{+j}^{(i,k)} - h_{-j}^{(i,k)})\,(\phi_j(y_i) - \phi_j(y_k)) \geq 0. \tag{55}$$

An approximate solution is found by maximizing the sum of the left hand sides of these homogeneous inequalities over a compact set of vectors y. We give some examples.

IRCN: Inner product model, rectangular, conditional

Observed: m paired comparisons of n objects (j=1,...,m ; i,k=1,...,n).

Model: $$i \succeq_j k \leftrightarrow \sum_{s=1}^{t} x_{js}y_{is} \geq \sum_{s=1}^{t} x_{js}y_{ks}, \tag{56a}$$

Sum: $$\phi(x,y) = \Sigma\Sigma\Sigma\sigma_{ik}^{j} \left(\Sigma_s x_{js}y_{is} - \Sigma_s x_{js}y_{ks} \right), \tag{56b}$$

$$\sigma_{ik}^{j} = (h_{+j}^{(i,k)} - h_{-j}^{(i,k)}), \tag{56c}$$

Matrix form: $$\phi(x,y) = \text{trace } (X'ZY), \tag{56d}$$

$$z_{ji} = \Sigma(\sigma_{ik}^{j} - \sigma_{ki}^{j}) = \sigma_{i.}^{j} - \sigma_{.i}^{j}. \tag{56e}$$

IRCM: Inner product model, rectangular, complete

Observed: N paired comparisons of nm objects with product structure.

Indices: i,k=1,...,n ; j,l=1,...,m ; t=1,...,N.

Model:

$$(j,i) \succeq_t (l,k) \leftrightarrow \Sigma x_{js}y_{is} \geq \Sigma x_{ls}y_{ks}, \tag{57a}$$

Sum:

$$\phi(x,y) = \Sigma\Sigma\Sigma\Sigma\Sigma\Sigma\sigma_{ijkl}^{t}(x_{js}y_{is} - x_{ls}y_{ks}), \tag{57b}$$

$$\sigma_{ijkl}^{t} = (h_{+t}^{(i,j)(k,l)} - h_{-t}^{(i,j)(k,l)}), \tag{57c}$$

Matrix form:

$$\phi(x,y) = \text{Trace}(X'ZY), \tag{57d}$$

$$z_{ji} = \Sigma\Sigma\Sigma(\sigma_{ijkl}^{t} - \sigma_{klij}^{t}) = \sigma_{ij..}^{\cdot} - \sigma_{..ij}^{\cdot}. \tag{57e}$$

ISCN: Inner product model, square, conditional.

Observed: each of n elements orders the other n elements, N replications

Indices: i,j,k=1,...,n ; t=1,...,N.

Model:

$$(j,i) \succeq_t (j,k) \leftrightarrow \Sigma x_{js}x_{is} \geq \Sigma x_{js}x_{ks}, \tag{58a}$$

Sum:

$$\phi(x) = \Sigma\Sigma\Sigma\Sigma\Sigma\sigma_{jik}^{t}(x_{js}x_{is} - x_{js}x_{ks}), \tag{58b}$$

Matrix form:

$$\phi(x) = \text{Trace } (X'AX) = \tfrac{1}{2} \text{Trace}(X'(A + A')X), \tag{58c}$$

$$a_{ji} = \sigma_{ji.}^{\cdot} - \sigma_{j.i}^{\cdot}. \tag{58d}$$

ISCM: Inner product model, square, complete

Observed: N paired comparisons of n^2 elements with product structure.

Indices: i,j,k,l=1,...,n ; t=1,...,N.

Model:

$$(i,j) \geq_t (k,l) \leftrightarrow \Sigma x_{is}x_{js} \geq \Sigma x_{ks}x_{ls}, \tag{59a}$$

Sum:

$$\phi(x) = \Sigma\Sigma\Sigma\Sigma\Sigma\Sigma \sigma^{t}_{ijkl} (x_{is}x_{js} - x_{ks}x_{ls}), \tag{59b}$$

Matrix:

$$\phi(x) = \mathrm{Tr}(X'AX), \tag{59c}$$

$$a_{ij} = \sigma^{\cdot}_{ij..} - \sigma^{\cdot}_{..ij}. \tag{59d}$$

The next four models are very similar to the first four. The difference is that we substitute squared euclidean distances for inner products (of course this reverses the order relations, but that is not important in maximum sum models). In order to save space we only give the matrix expressions. If only the diagonal elements of a matrix are defined then the matrix is diagonal.

DRCM: Distance model, rectangular, conditional

$$\phi(x,y) = 2\mathrm{Tr}(X'ZY) - \mathrm{Tr}(Y'DY), \tag{60a}$$

$$z_{ji} = \sigma^{j}_{i.} - \sigma^{j}_{.i}, \tag{60b}$$

$$d_{ii} = \sigma^{\cdot}_{i.} - \sigma^{\cdot}_{.i}. \tag{60c}$$

DRCM: Distance model, rectangular, complete

$$\phi(x,y) = 2\mathrm{Tr}(X'ZY) - \mathrm{Tr}(Y'DY) - \mathrm{Tr}(X'EX), \tag{61a}$$

$$z_{ji} = \sigma^{\cdot}_{ij..} - \sigma^{\cdot}_{..ij}, \tag{61b}$$

$$d_{ii} = \sigma^{\cdot}_{i...} - \sigma^{\cdot}_{..i.}, \tag{61c}$$

$$e_{ii} = \sigma^{\cdot}_{.j..} - \sigma^{\cdot}_{...j}. \tag{61d}$$

DSCN: Distance model, square, conditional

$$\phi(x) = 2\mathrm{Tr}(X'AX) - \mathrm{Tr}(X'DX), \tag{62a}$$

$$a_{ji} = \sigma^{\cdot}_{ji.} - \sigma^{\cdot}_{j.i}, \tag{62b}$$

$$d_{ii} = \sigma^{\cdot}_{.i.} - \sigma^{\cdot}_{..i}. \tag{62c}$$

DSCM: Distance model, square, complete

$$\phi(x) = 2\mathrm{Tr}(X'AX) - \mathrm{Tr}(X'DX), \tag{63a}$$

$$a_{ij} = \sigma^{\cdot}_{ij..} - \sigma^{\cdot}_{..ij}, \tag{63b}$$

$$d_{ii} = \sigma^{\cdot}_{i...} + \sigma^{\cdot}_{.i..} - \sigma^{\cdot}_{..i.} - \sigma^{\cdot}_{...i}. \tag{63c}$$

Usually the simplifications in 4.3. apply here too, and the elements of the matrices are simple functions of the rank numbers. In fact IRCN is identical to our previous procedures A and B. The techniques described in 4.5 can also be interpreted as maximum sum techniques.

LIRCN: Mixed linear-inner product model, rectangular, conditional

Observed: m paired comparisons of n real N-element vectors $\mathcal{E}_i$.

Indices: i,k=1,...,n ; j=1,...,m ; t=1,...,N.

Model:

$$\mathcal{E}_i \gtrsim_j \mathcal{E}_k \leftrightarrow \sum_s x_{js} \sum_t y_{ts}\mathcal{E}_{it} \gtrsim \sum_s x_{js} \sum_t y_{ts}\mathcal{E}_{kt}, \tag{64a}$$

Sum:

$$\phi(x,y) = \Sigma\Sigma\Sigma\Sigma\Sigma\sigma^j_{ik}(x_{js}y_{ts}\mathcal{E}_{it} - x_{js}y_{ts}\mathcal{E}_{kt}), \tag{64b}$$

Matrix form:

$$\phi(x,y) = \mathrm{Tr}(X'ZY) \tag{64c}$$

$$z_{jt} = \Sigma\mathcal{E}_{it}(\sigma^j_{i.} - \sigma^j_{.i}). \tag{64d}$$

Special cases:

$\mathcal{E}_i$ unit vectors: paired comparisons (IRCN),

$\mathcal{E}_i$ are sums of a fixed number of unit vectors: additive conjoint analysis.

Up to now we have discussed a number of models in which the maximum sum principle results in the maximization of a function of the form tr(X'AX) or tr(X'ZY), subject to some scaling requirement. We did not specify these requirements, but for all models there are a number of natural choices which lead to more or less standard eigenvalue-eigenvector problems. Of course generalizations of the type discussed in 4.3 are also possible here. If we agree that $\sigma^t_{ijkl} = 0$ if the comparison by subject t of (i,j) and (k,l) is missing (either by design or by accident) then we can write down general inner product and distance models of which the four cases we discussed previously are very special examples. The formula look exactly the same as those of ISCM and DSCM with the generalized missing-data definition of σ^t_{ijkl}. We can also generalize the maximum sum method to three-way models. This also tends to give surprisingly simple formulea, although somewhat less standard arithmetic. In the models with replications the data were always pooled over replications. In the three-way extensions we represent the replications (subjects) as a set of points in Euclidean space too.

TWWD: Three-way model, weighted distances.

Each replication defines a set of weights. Dissimilarities correspond with weighted squared distances:

$$\delta^t_{ij} \approx \sum_{s=1}^{p} w_{ts}(x_{is} - x_{js})^2 \tag{65}$$

We use the generalized signature σ^t_{ijkl}. Then

$$\phi(x,w) = \Sigma\Sigma\Sigma(x_{is} - x_{js})^2 \Sigma w_{ts} (\sigma^t_{ij..} - \sigma^t_{..ij}) \tag{66}$$

Letting

$$a_{ijt} = \sigma^t_{ij..} + \sigma^t_{ji..} - \sigma^t_{..ij} - \sigma^t_{..ji}, \tag{67a}$$

$$d_{iit} = \sigma^t_{i...} + \sigma^t_{.i..} - \sigma^t_{..i.} - \sigma^t_{...i}, \tag{67b}$$

$$b_{ijt} = d_{ijt} - a_{ijt}, \tag{67c}$$

we find

$$\phi(x,w) = \Sigma\Sigma\Sigma\ b_{ijt}x_{is}x_{js}w_{ts}. \tag{68}$$

TWWI: Three-way model, weighted inner products

Again each replication defines a set of weights. Now similarities correspond with weighted inner products:

$$\delta^t_{ij} \approx \sum_{s=1}^{p} w_{ts}x_{is}x_{js}. \tag{69}$$

Again this works out as

$$\phi(x,w) = \Sigma\Sigma\Sigma b_{ijt}x_{is}x_{js}w_{ts}, \tag{70}$$

but now b_{ijt} has the much simpler form

$$b_{ijt} = \sigma^t_{ij..} - \sigma^t_{..ij} \tag{71}$$

TWTI: Three-way model, translated inner products

Each replication defines a translation of the basic space and a new origin from which the inner products are computed:

$$\delta^t_{ij} \approx \sum_{s=1}^{p} (x_{is} - z_{ts})(x_{js} - z_{ts}). \tag{72}$$

We find

$$\phi(x,z) = \mathrm{Tr}(X'AX) - \mathrm{Tr}(Z'BX) \tag{73}$$

with

$$a_{ij} = \sigma^{\cdot}_{ij..} - \sigma^{\cdot}_{..ij}, \tag{74a}$$

$$b_{ti} = \sigma^t_{i...} + \sigma^t_{.i..} - \sigma^t_{..i.} - \sigma^t_{...i}. \tag{74b}$$

Models in which n-tuples of stimuli are compared can also be constructed. An example is the following model.

TWHD: Three way model, homogeneity, measured by distances

Each replication defines a set of weights. Heterogeneity of a set of stimuli corresponds with the sum of the squared distances within the set, i.e.

$$\eta^t(I) \approx \sum_{i\in I} \sum_{j\in I} \sum_{s=1}^{p} w_{ts}(x_{is} - x_{js})^2. \tag{75}$$

Observe that the TWWD model, and thus the DSCM model, is a special case. And finally we can construct models which permit asymmetry in square complete matrices

TWDS: Three-way model, weighted distances, slide vector

Here

$$\delta_{ij}^t \approx \sum_{s=1}^{p} w_{ts}(x_{is} - x_{j\varepsilon} - z_{ts})^2. \tag{76}$$

TWWD is the special case in which all slide vectors vanish.

4.7 Generalized correlation coefficients

The maximum sum approach is only one way of approaching differencing models. We can also use generalized correlations coefficients(GCC) which are defined as follows. Let ϕ and ψ be increasing real functions that satisfy

$$\phi(y_i - y_j) = -\phi(y_j - y_i), \tag{77a}$$

$$\psi(y_i - y_j) = -\psi(y_j - y_i), \tag{77b}$$

for all i,j. This implies that

$$\phi(0) = \psi(0) = 0 \tag{78}$$

and

$$\text{sign}(\phi(y_i - y_j)) = \text{sign}(y_i - y_j), \tag{79a}$$

$$\text{sign}(\psi(y_i - y_j)) = \text{sign}(y_i - y_j). \tag{79b}$$

A GCC $\Gamma_{\psi,\phi}(x,y)$ is defined by letting

$$C_{\psi,\phi}(x,y) = \Sigma\Sigma\psi(x_i - x_j)\phi(y_i - y_j), \tag{80a}$$

$$V_{\psi}(x) = \Sigma\Sigma\psi^2(x_i - x_j), \tag{80b}$$

$$V_{\phi}(y) = \Sigma\Sigma\phi^2(y_i - y_j), \tag{80c}$$

and

$$\Gamma_{\psi,\phi}(x,y) = C_{\psi,\phi}(x,y) \,/\, (V_{\psi}^{\frac{1}{2}}(x)\, V_{\phi}^{\frac{1}{2}}(y)). \tag{81}$$

Clearly

$$-1 \leq \Gamma_{\psi\phi}(x,y) \leq +1 \tag{82}$$

and a necessary (but not sufficient)condition for $\Gamma = 1$ is that, for all i,j,

$$\text{sign}(x_i - x_j) = \text{sign}(y_i - y_j). \tag{83a}$$

In the same way if $\Gamma = -1$ then

$$\text{sign}(x_i - x_j) = -\text{sign}(y_i - y_j) \tag{83b}$$

for all i,j. In general

$$\Gamma_{\psi,\phi}(x,y) \neq \Gamma_{\psi,\phi}(y,x), \tag{84a}$$

$$\Gamma_{\psi,\phi}(x,y) \neq \Gamma_{\phi,\psi}(x,y) \tag{84b}$$

A GCC is symmetric if $\psi = \phi$. Familiar examples of symmetric GCC's are Pearson's product moment coefficient with

$$\psi(x_i - x_j) = x_i - x_j, \tag{85a}$$

Spearman's rho with

$$\psi(x_i - x_j) = \rho(x_i) - \rho(x_j), \tag{85b}$$

and Kendall's tau with

$$\psi(x_i - x_j) = \text{sign}(x_i - x_j). \tag{85c}$$

The asymmetric coefficient for which

$$\psi(x_i - x_j) = \text{sign}(x_i - x_j), \tag{86a}$$

and

$$\phi(y_i - y_j) = y_i - y_j \tag{86b}$$

is very interesting in our case. It is easy to see that maximizing the sum of squares of the GCC's that can be formed in the case of m rankings of n objects is equivalent to our procedure B if we use the functions (86). Moreover the same thing is true if we use

$$\psi(x_i - x_j) = \rho(x_i) - \rho(x_j), \tag{87a}$$

$$\phi(y_i - y_j) = y_i - y_j. \tag{87b}$$

Or, as we may also put it, Spearman and Kendall weighting of differences leads to identical results.

4.8 Some criticisms

The techniques outlined in this chapter are subject to roughly the same criticisms as the PCA techniques in chapter 3. The choice of requirements (1) may seem somewhat arbitrary, but it is also suggested by the theory of generalized Daniels-Kendall correlation coefficients, which are natural measures of disarray. Moreover the resulting maximization problems prove to be comparatively simple. The techniques in 4.6 are less well-defined in terms of rational optimality criteria or between-within variance decomposition. They are not extensions of classical componentwise multivariate analysis, they are extensions of classical metric eigen-techniques. The main criterion here is ease of computation. The general impression is that results are often as satisfactory (from the point of view of recovery) as those of iterative

gradient programs for the same measurement model and sometimes the maximum sum results are better (in terms of both recovery and interpretation). Observe that we did not specify the scaling requirements in 4.6. In general different scaling requirements lead to different solutions, and it may take some extra research to find out what the best policy is for a particular model.

4.9 Historical

The techniques discussed in this chapter are rediscovered every 10 years, presumably because there are so many different methods to derive them. The earliest reference seems to be Guttman (1946). Other discoverers and/or contributors include Slater (1960), Carroll & Chang (1964), Hayashi (1965), Benzécri (1969), De Leeuw (1968a), Bechtel (1969). Most of these authors only discuss the simplest case in which there are either m complete rankings of n objects or m complete sets of paired comparisons. Guttman also discusses the additive case from 4.8. Both the linear and the additive case were investigated in De Leeuw (1968a) both in terms of discrimination and in terms of generalized correlations coefficients. The maximum sum techniques have been applied to a number of interesting examples. An application of IRCN to adjective-noun intersection data has been reported by Levelt, De Leeuw (1968a) also has some applications of IRCN. In De Leeuw (1968b) there are some applications of DSCN and DSCM. The DSCM technique has also been described by Guttman (1968). In De Leeuw (1970 a,b) there is a more extensive discussion of these techniques. In De Klerk, De Leeuw, Oppe (1968, 1970) there are a number of interesting applications of the different versions of LIRCN. Pols and Van Der Kamp (1971) applied DSCN to confusion matrices of vowel sounds and compared it with Roskam's iterative UNFOLD program. De Leeuw (1968b) compared DSCN with the iterative NMSPOM program on politicological data. Both authors found that maximum sum methods were not worse than the iterative techniques. Carroll & Chang (1964) compared IRCN with a PCM type of technique and found that IRCN did a better job in recovering data which were not errorfree.

The model TWWD is due to Douglas Carroll (cf Carroll and Chang 1970). Carroll discusses only the metric version, and the possibility of generalizing this to a two-stage iterative nonmetric algorithm. A number of very interesting and very impressive applications has been published (Carroll & Chang 1970, Carroll & Wish 1970, Wish 1970,

Wish and Carroll 1971). The square conditional version of TWWD has been used by Van der Kloot (1969). Model TWWI is related to Tucker's multiway factor analysis (with diagonal core matrix), but it is in its metric version a straightforward generalization of PCA. It has been discussed by a number of authors. Harshman (1970) relates it to Cattell's parallel proportional profiles, Carroll & Chang (1970) discuss it in the TWWD context, and Slater (1969) mentions some work of Gower on this model. I have used it in some unpublished studies and it seems to give results which are as satisfactory as those of Carroll and Wish. TWTI is new and has never been applied (as far as I know). Generalization of TWWI and TWWD have been worked out by Harshman (1971) under the name of PARAFAC2, and by Jennrich (1971) and Carroll (mentioned in Carroll & Chang 1971 under the name of IDIOSCAL). The most general model is IDIOSCAL. In a factor analysis context this model has already been proposed by Rasch (1953) and Meredith (1964) as naturally following from the Pearson-Aitken-Lawley selection theorem. TWHD has also been used in some ad hoc studies. Its usefulness clearly depends on the question whether there exist situations in which judgments of homogeneity or heterogeneity are the natural thing to ask. The idea of using the slide vector z to explain asymmetry is due to Kruskal (personal communication). In general the mathematical properties of TWTI, TWHD, and TWDS are interesting, but the techniques are possibly not very useful. Generalized correlation coefficients are due to Daniels (1944), and have been discussed extensively by Kendall (1962). Their use in this context was first studied by De Leeuw (1968a)

5. Partitioning the variables

5.0. Introduction

In chapter 3 we discussed the generalization of PCA to categorical data, and some of the problems that arise with the applications of this generalization. In this chapter we generalize the rest of standard multinormal joint bivariate analysis to categorical data in exactly the same way. Because most of the problems have already been discussed extensively in chapter 3, we give only short indications of the extensions, and we show how the weak aspects of this class of techniques become more apparent in complicated cases.

5.1. ANOVA-formulation

Consider a partition of the n variables into N subsets, subset L containing n_L variables (L=1, ..., N). Of course $1 \leq N < n$, $\Sigma n_L = n$, $1 \leq n_L \leq n$. For ease of notation we partition the index set {1, 2, ..., n} in K subsets N_L, where N_L contains the indices of subset L. In the same way the rows of our induced score matrix Z (a function of the supervector of weights) can be partitioned in K subsets. We also construct the induced matrix of subset scores, this matrix S has K rows and m columns, and its general element is defined as

$$s_{Lj} = \sum_{i \in N_L} z_{ij} \qquad (1)$$

The sources of variation that interest us are again the variance within sample elements and the variance between sample elements, this time measured over the subset scores in S. If we apply our homogeneity definition of section 3.1. to this matrix we have (assuming $s_{..} = 0$ once again)

Source	Sum of Squares
Between columns	$B = N\Sigma s_{.j}^2$
Within columns	$W = \Sigma\Sigma(s_{Lj} - s_{.j})^2$
Total	$T = \Sigma\Sigma s_{Lj}^2$

(2)

and again we are interested in maximizing $\lambda = B/T$.

What does this mean? In the first place we do not interpret the variables within the subsets as replications of each other, which have to be homogeneous in some sense. We interpret the subsets as multidimensional variables, just as in MANOVA-type techniques. In the standard, metric cases of multivariate analysis we make linear combinations of the variables in the subsets and apply our homogeneity ideas to these linear combinations. In our categorical generalizations we make additive combinations of the induced score vectors, which are linear combinations of the original indicator vectors. In the

second place we have the identity

$$B(S) = N \sum_{j=1}^{m} (s_{.j} - s_{..})^2 = (\frac{N}{n})\, n \sum_{j=1}^{m} (z_{.j} - z_{..})^2 =$$

$$= (\frac{N}{n})\, B(Z), \qquad (3)$$

which means that the numerator of λ is essentially the same, no matter how we define our partition of the variables.

5.2. Matrix formulation

It follows from (2) and (3) that

$$\lambda = \frac{x'Cx}{Nx'Dx} \qquad (4)$$

where C is identical to our previous matrix C (section 3.2.), and D is defined in the following way: if an element of D corresponds with categories of variables in the same subset it is equal to the corresponding element of C, if it corresponds with categories of variables in different subsets it is equal to zero. Again we require $\Sigma s_{Lj} = 0$ for all L=1, ..., N. To see how these constraints can be incorporated in an easy way, observe that the vector x_0 in which each category gets a weight which is the reciprocal of the number of variables in its subset, satisfies

$$Cx_0 = N\, Dx_0, \qquad (5)$$

i.e. x_0 satisfies the basic stationary equations with $\lambda = 1$. Remove x_0 by deflation, then

$$\bar{C} = C - \frac{1}{m}\, Dx_0x_0'D, \qquad (6)$$

where D is the diagonal matrix D used in chapter 3. Thus $\bar{C}$ is equal to our previous deflated C matrix of formula 3.11. Any vector satisfying

$$\bar{C}x = \lambda N\, Dx \qquad (7)$$

automatically satisfies $\sum_{j=1}^{m} s_{Lj} = 0$ for all L=1, ..., N (although not necessary $\sum_{j=1}^{m} z_{ij} = 0$ for all i = 1, ..., n).

By now the structure of both C and D should be clear. In C the submatrices C_{ij} can be divided into two classes: those within subsets and those between subsets. In D the within matrices C_{ij} (to which the diagonal submatrices C_{ii} always belong) are copied from C at the corresponding places. Then the rest of D is filled up with zeroes. We know that the rank of C satisfies

$$r(C) \leq \min(K - n + 1, m). \qquad (8a)$$

For the rank of D we find

$$r(D) \leq \min(K - n + N, m). \tag{8b}$$

Thus $r(D) \geq r(C)$.

5.3. Linear restrictions

Consider the slightly more general problem: maximize

$$\lambda = \frac{x'Cx}{Nx'Dx} \tag{9}$$

under the conditions $Sx = 0$, where S is a p x K matrix of rank $p \leq K$. Alternatively this can be written as $x = Ty$, where T is an K x (K-p) matrix of rank K-p, satisfying $ST = 0$. In general such a T can be found by using the K-p eigenvectors of S'S corresponding with zero eigenvalues. By substitution our problem reduces to the maximization of

$$\lambda = \frac{y'T'CTy}{Ny'T'DTy}. \tag{10}$$

The matrices T'CT and T'DT are of order K-p. Moreover

$$r(T'DT) \geq r(T'CT) = \min(K - p, r(C)), \tag{11}$$

and thus both matrices are nonsingular iff $r(C) \geq K - p$. In the usual case $r(C) = K - n + 1$, and $p \geq n - 1$ suffices. The matrix T'DT is nonsingular iff $r(D) \geq K - p$, and in the usual case $p \geq n - N$ suffices. All of the theory of linear restrictions treated in sections 3.11., 3.13., 3.15., and 3.16. in a more or less disguised form falls under this section, including the desirability of choosing T in such a way that T'DT is diagonal and nonsingular, and in such a way that the vectors y are directly interpretable. Consider, for example, a set of discrete functions, orthonormal with respect to D. We can easily restrict our weights x to be polynomials of a specific degree. On the other hand the vectors in T might be related to measured 'independent' variables. In our problems with ordinal restraints of the form $Sx \geq 0$ we can also write this as $x = Ty$, where T contains the edges of the cone defined by $Sx \geq 0$, and we require, in addition, that $y \geq 0$. Of course the restrictions $\sum_{j=1}^{m} s_{Lj} = 0$ for all L can also be incorporated in this form, but also the more stringent ones $\sum_{j=1}^{m} z_{ij} = 0$ for all i. The difference is that only the restrictions on S are related in a simple way to a particular eigenproblem without restrictions.

5.4. Special effects

It is interesting to find out what happens in the general case to our 'special effects' like internal consistency, the relationship with X^2, the geometrical interpretations, and so on. It is trivial to five a formal generalization of the internal consistency equations (3.25). We have

$$H'x = (N\lambda)^{\frac{1}{2}} s, \tag{12a}$$

$$Hs = (N\lambda)^{\frac{1}{2}} Dx. \tag{12b}$$

The interpretation of these equations is more difficult than in the PCA case, but equation (12a) tells us that the optimal direct scores are again the column means of the induced score matrix S. If D^- is a generalized inverse of D, then (12b) can be rewritten as

$$x = (N\lambda)^{-\frac{1}{2}}(D^-Hs + (I - D^-D)y), \tag{13}$$

for some real vector y. If we substitute this in (12a) we find

$$H'D^-Hs = N\lambda s - H'(I - D^-D)y. \tag{14}$$

It is not difficult to check that for the Moore-Penrose inverse D^+ we have $H'(I - D^+D) = 0$, and thus (14) simplifies to the more familiar form

$$H'D^+Hs = N\lambda s \tag{15a}$$

which can be used in conjunction with

$$Cx = HH'x = N\lambda Dx. \tag{15b}$$

To derive optimal direct scores we can also use the pseudometric (in the sense of section 3.12.)

$$\delta^2_{jl} = (e_j - e_l)' H'D^+H (e_j - e_l), \tag{16}$$

and find multidimensional score vectors s_j and s_l in such a way that

$$d^2_{jl} = (z_l - z_j)'(z_l - z_j) \tag{17}$$

optimally approximates δ^2_{jl} in a (generalized) least squares sense. In most cases D is not diagonal and the simple set theoretical interpretation is not valid any more. Observe that both H and D can be partitioned into subsets corresponding with the N sets of variables. It is easy to see that

$$H'D^+H = H'_1D^+_1H_1 + \ldots + H'_ND^+_NH_N, \tag{18}$$

and thus δ^2_{ij} can be partitioned into N components too. The components $(\delta^L_{jl})^2$ are Mahanalobis-type distances between the vectors h^L_j and h^L_l in the space whose (oblique) axes are defined by the dispersion matrices of the sets of variables. Subsets with large variance contribute relatively little to the

overall distances. The reproducibility equations are

$$H = N\, D(\lambda_0^{\frac{1}{2}} x_0 s_0' + \lambda_1^{\frac{1}{2}} x_1 s_1' + \ldots + \lambda_v^{\frac{1}{2}} x_v s_v'), \tag{19}$$

and

$$C = N\, D(\lambda_0 x_0 x_0' + \ldots + \lambda_v x_v x_v') \, D. \tag{20}$$

The separating hyperplane interpretation still holds. Again

$$N^2 \Sigma\lambda^2 = \Sigma\Sigma\, \mathrm{Tr}((D_L^+)^{\frac{1}{2}} C_{LM} D_M^+ C_{ML} (D_L^+)^{\frac{1}{2}}), \tag{21}$$

with C_{LM} the cross-product matrices between subsets, $C_{LL} = D_L$. It follows that the diagonal terms of this summation are equal to the rank of the D_L, i.e.

$$N^2 \Sigma\lambda^2 = \Sigma\, R(D_L) + \sum_L \sum_{M \neq L} \mathrm{Tr}(Q_{LM}) \tag{22}$$

with Q_{LM} the scaled cross product matrices from (21). The nondiagonal terms in (22) are X^2-analogues for a complicated hypotheses of independence which says (roughly) that C_{LM} can be predicted from D_L and D_M. By our essentially bivariate treatment of the problem we have run into complications once again, in the multinomial extensions of multivariate analysis the X^2 measures and the hypotheses of independence and interaction come out much more elegantly.

By following the argument in section 3.8. it is easy to see that the stationary equations in the multinormal case can be derived from

$$B = \frac{1}{N} (\alpha_0' C^0 \alpha_0 + \alpha_1' C^1 \alpha_1 + \ldots), \tag{23a}$$

$$T = (\alpha_0' D^0 \alpha_0 + \alpha_1' D^1 \alpha_1 + \ldots), \tag{23b}$$

where C is the complete and D the within part of the population covariance matrix, and superscripts denote taking all elements to the power s (of course it only makes sense to assume multinormality of the variates, not for factors!). The stationary equations are

$$C^s \alpha_s = N\lambda\, D^s \alpha_s \tag{24}$$

for all s = 0, 1, 2, ... Again s = 0 takes care of the trivial solution, s = 1 is the multinormal solution we are interested in, and for s > 1 we find further uninformative solutions. Again we cannot expect the nice relationships with contingency to hold, but in practice it still may be possible to detect and isolate multinormal effects.

5.5. Numerical and binary variables

If the categories of a numerical variable have prior scores y_i satisfying $e'D_i y_i = 0$ we require $x^i = \alpha_i y_i$ for all i, then $\lambda = (\alpha' C \alpha)/(N \alpha' D \alpha)$, and

C and D are the ordinary dispersion matrices used in multivariate analysis. For binary variables we can require $x^j = \alpha_j(-n_{i-} \mid n_{i+})$ and we obtain the same result as in 3.15. Observe, however, that the restrictions for binary variables in the general case are real restrictions unlike those in PCA in which they were a means of simplifying the computations. This is because the natural restriction in the general case is $\Sigma s_{L_j} = 0$ for all L, or

$$\sum_{i \in N_L} (n_{i+}x_{i+} + n_{i-}x_{i-}) = 0, \tag{25a}$$

and not $\Sigma z_{ij} = 0$ for all i, or

$$n_{i+}x_{i+} + n_{i-}x_{i-} = 0. \tag{25b}$$

Of course the restrictions considered in this section are part of the general theory in 5.3..

5.6. The case N = 2

In the case that there are only two subsets the problem simplifies in the in the same way as in section 3.10. We can order the variables in such a way that the stationary equations are

$$\left[\begin{array}{c|c} C & E \\ \hline E' & D \end{array}\right]\left\{\begin{array}{c} x \\ \hline y \end{array}\right\} = 2\lambda\left[\begin{array}{c|c} C & 0 \\ \hline 0 & D \end{array}\right]\left\{\begin{array}{c} x \\ \hline y \end{array}\right\}, \tag{26}$$

and this simplifies in the usual way to the related system

$$Ey = \mu Cx, \tag{27a}$$

$$E'x = \mu Dy, \tag{27b}$$

with $\mu = 2\lambda - 1$. Suppose for the moment that E is n x m with $m \leq n$. Again any one of the m solutions of (27) corresponds to a pair of solutions of (26) whose values of λ add up to one (cf equation 3.50 and the discussion following it). Using the linear restrictions $x = T\alpha$ and $y = U\beta$ reduces the second system to

$$T'EU\beta = \mu T'CT\alpha, \tag{28a}$$

$$U'E'T\alpha = \mu U'DU\beta. \tag{28b}$$

5.7. Improper solutions

An interesting question in this context is what happens to our improper solution if only for some of the variables we use linear constraints, and not for others. The question is in how far the analysis of section 3.18. generalizes.

Define a vector t with as elements the reciprocal of the number of nominal variables in the subset on which no linear restrictions are defined. The elements corresponding to variables with linear restrictions are zero. Then t is the improper solution with $\lambda = \bar{N}/N$, where $\bar{N}$ is the number of subsets which contain variables without restrictions. Moreover $t'Dt = \bar{N}m$, and thus the deflation procedure replaces C_{ik} by $\tilde{C}_{ik}=C_{ik} - \frac{1}{m} Dtt'D$. It follows that we can replace C_{ik} by $\tilde{C}_{ik}$ right away and leave D unchanged. Observe, however, that this is not the same as transforming the corresponding rows of H to deviations from the mean, because this last procedure also changes the cross products with the restricted variables. In most cases, however, we do require $\Sigma z_{ij} = 0$ for numerical and binary variables, and in this case we can transform all rows of H to deviations from the mean.

5.8. Some familiar special cases

In this section symbols like (n_1)Nu(n_2)No(n_3)Bi(n_4)Or are used to indicate a situation in which there are 4 sets, the first set contains n_1 numerical variables, the second set n_2 nominal ones, the third set n_3 binary ones, and the fourth set n_4 ordinal ones. Examples:

Case (n-1)Nu(1)Nu : multiple linear regression.
Case (n-1)Nu(1)Bi : discriminant analysis.
Case (n-1)Nu(1)No : canonical discriminant analysis.
Case (n-1)No(1)Nu : analysis of variance.
Case (n_1)Nu(n_2)Nu : canonical analysis.
Case (n_1)No(n_2)Nu : multivariate analysis of variance.
Case (1)Nu(1)Nu...(1)Nu (n times) : principal component analysis.

This covers most of the cases usually discussed in textbooks on multivariate analysis. The nonmetric breakthrough has given us a number of techniques which result from the classical cases by replacing Nu by Or. Thus we have nonmetric multiple regression (n-1)Nu(1)Or and monotone analysis of variance (n-1)No(1)Or. In the analysis of variance techniques, by the way, the nominal variables in the first set are ways of classification. For a complete factorial design each main effect and each interaction defines a nominal variable, and the within matrices C_{ik} have a very simple structure. The same thing is true for other balanced designs like Latin squares and BIBD's. Another special case that is known in the literature is M-set matching (n_1)Nu(n_2)Nu ... (n_M)Nu.

We have shown that the classical multivariate techniques are special cases of our general subsets-of-variables set up with linear restrictions. To

make the relationships somewhat clearer we emphasize that any nominal variable is already treated as a subset of binary variables. In this sense it is indeed true that binary variables are the most basic type. A PCA of n nominal variables is already equivalent to a n-set matching of sets of binary variables (with the special property that the within set cross product matrices are diagonal). The only techniques in which there is no grouping of the variables in any sense is the PCA of n binary or of n numerical variables with rectrictions $x_i = a_i y_i$.

5.9. Relation with PCA

In this section we show that using linear restrictions is general enough to make the grouping of variables unnecessary. Suppose we have a subset of n nominal variables with $k_1, \ldots, k_n$ categories. By using the Cartesian product of the T_i we can reduce this to one nominal variable with Πk_i categories. Such a new category is indexed by $(l_1, \ldots, l_n)$ with $1 \leq l_i \leq k_i$. We impose the linear restriction that

$$x(l_1, \ldots, l_n) = x_1(l_1) + \ldots + x_n(l_n). \qquad (29)$$

It is easy to see that the matrix Z of induced scores is identical to the matrix S of subset scores and consequently the PCA analysis with linear restrictions defined by (29) gives identical results as the subsets of variables analysis outlined in this chapter. Of course a PCA without the restrictions (29) in general gives a different result, and I can imagine situations in which it is preferable not to use these restrictions at all. If we do not use the restrictions the relations with X^2 discussed in chapter 3 hold again and we 'test' the pairwise independence of sets of variables. In the multinormal case using or not using (29) makes no difference at all.

5.10. Historical

Again the main ideas in this section are not new. The fact that all forms of classical multivariate analysis can be interpreted as special cases of canonical analysis was already emphasized by Bartlett in his famous paper of 1947. Special cases of our nominal variables incorporation were already investigated by Fisher (1941), Johnson (1950), Guttman (1959b), and Lingoes (1963). The case (n-1)No(1)No was studied by Fisher (1941) who called it analysis of variance with optimal scoring and by Carroll(1968) who called it categorical conjoint measurement.

The (n-1)No(1)Or case was described by Kruskal (1965), and implemented by him and Carmone in the program MONANOVA. The special case (n-1)No(1)Or in which

the variables in the first set are all replications was studied by Bradley, Katti, and Coons (1962). The general case with (n-1)Nu(1)Or and (n-1)No(1)Or was also studied by De Leeuw (1970) using Kruskal-type techniques, the special case (n-1)Nu(1)Bi was treated in De Leeuw (1968).

A related approach to quantification, from the facet point of view, has been given by Guttman (1959a). The use of symbols like (4)Nu(1)No to describe special cases is new. For our purposes these symbols are general enough. In the computer programs written by Doesborgh (1971) different types of variables can occur in a single set. This can be described by a more general notation in which we fix the order of the types Nu, No, Bi, Or and describe each subset of variables by a quadruple of indices. Thus PCA becomes

(1, 0, 0, 0)(1, 0, 0, 0) ... (1, 0, 0, 0) (n times)

and the classical multivariate linear model becomes

$(n_1^1, n_2^1, n_3^1, 0)\ (n_1^2, 0, 0, 0)$.

6. Some special topics

6.0 Introduction

In this chapter we shall discuss the application of our general ideas to partial canonical correlation, to image analysis, and to common factor analysis and its generalizations. Again the descriptions of the techniques will be brief, partly because their usefulness has not been proven, partly because the extensions are straightforward. We conclude with some general remarks on statistical inference in situations like the ones discussed in this dissertation, and with some general criticisms.

The procedures discussed in this chapter are not yet programmed for the computer, and they have not been tried out on real data. This chapter can be interpreted as a number of suggestions for further research. Some of these suggestions have already been partially worked out in a number of related publications, others will be taken up in the future.

6.1. Partial canonical correlation

One of the things that remains to be done is a generalization of the theory of partial correlation in this framework. We shall suppose that the singularities are removed by the familiar methods, and all we need is a theory of partial canonical correlation between sets of variables. As a first example we have sets X, Y, Z and we want to partial out the contribution of X. The joint cross products matrix of X, Y, Z is

$$\begin{array}{c} X \\ Y \\ Z \end{array} \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \tag{1}$$

The multivariate regression equations are

$$Y = AX + E, \tag{2a}$$

$$Z = BX + E, \tag{2b}$$

and the least squares solution for A and B is

$$A = C_{21}\, C_{11}^{-1}, \tag{3a}$$

$$B = C_{31}\, C_{11}^{-1}, \tag{3b}$$

and thus the residuals are

$$Y = Y - C_{21}C_{11}^{-1}X, \tag{4a}$$

$$Z = Z - C_{31}C_{11}^{-1}X, \tag{4b}$$

and these residuals are used in a new canonical problem. The new dispersions are

$$E(YY') = C_{22.1} = C_{22} - C_{21}C_{11}^{-1}C_{12}, \tag{5a}$$

$$E(ZZ') = C_{33.1} = C_{23} - C_{31}C_{11}^{-1}C_{13}, \tag{5b}$$

$$E(YZ') = C_{23.1} = C_{23} - C_{21}C_{11}^{-1}C_{13}. \tag{5c}$$

The PCA problem with X partialled out has stationary equations

$$\begin{bmatrix} C_{22.1} & C_{23.1} \\ C_{32.1} & C_{33.1} \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = 2\lambda \begin{bmatrix} C_{22.1} & 0 \\ 0 & C_{33.1} \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix}, \tag{6}$$

or, equivalently,

$$C_{23.1}\, z = \mu C_{22.1}\, y, \tag{7a}$$

$$C_{32.1}\, y = \mu C_{33.1}\, z. \tag{7b}$$

Generalizations to cases in which we want to partial out the contributions of several sets of variables and study the relationship between the others are obvious. In particular this provides useful ways to generalize the analysis of covariance. From the equations in this section it is easy to see that only the sets that must be partialled out should be nonsingular. In the analysis of covariance, where we partial out numerical variables, this is likely to be the case. Again the partial correlation analogues in nonparametric multivariate analysis are quite different from the ones used here (they are X^2 measures based on conditional probabilities). And again there is a discrepancy between the psychometric approach to multivariate analysis which proceeds componentwise in order to achieve optimal data reduction and interpretation, and the statistical approach which concentrates on overall determinantal criteria to test significance of the dependencies.

6.2. Image analysis

The ideas of the previous section can also be used in extending image analysis to sets of variables. Again we suppose there are no singularities. For each subset the basic equation is

$$x_L = \bar{x}_L + \tilde{x}_L. \tag{8}$$

Here $\bar{x}_L$ is the part of x_L that can be predicted by linear regression techniques from the remaining sets, and $\tilde{x}_L$ is the residual. Of course it is also possible to do an image analysis within each of the N sets x_L. In that case we can write

$$x_L = \bar{\bar{x}}_L + \bar{\tilde{x}}_L + \tilde{\bar{x}}_L + \tilde{\tilde{x}}_L \tag{9}$$

In general the results for this fourfold partition will not be identical to those of an ordinary image analysis for the complete set of variables.

The relevant equations for the twofold partition are the following.

Images:

$$\bar{x}_L = \Sigma C_{LK} C_{KK}^{-1} x_K - x_L. \tag{10}$$

Anti-images:

$$\tilde{x}_L = 2 x_L - \Sigma C_{LK} C_{KK}^{-1} x_K. \tag{11}$$

Image dispersions:

$$\bar{C}_{KL} = C_{KL} - 2 \Sigma C_{KP} C_{PP}^{-1} C_{PL} + \Sigma\Sigma C_{KP} C_{PP}^{-1} C_{PQ} C_{QQ}^{-1} C_{QL}. \tag{12}$$

Anti-image dispersions:

$$\tilde{C}_{KL} = 4 C_{KL} - 4 \Sigma C_{KP} C_{PP}^{-1} C_{PL} + \Sigma\Sigma C_{KP} C_{PP}^{-1} C_{PQ} C_{QQ}^{-1} C_{QL}. \tag{13}$$

Mixed dispersions:

$$C_{\underline{KL}} = 2 C_{KL} - 4 \Sigma C_{KP} C_{PP}^{-1} C_{PL} + \Sigma\Sigma C_{KP} C_{PP}^{-1} C_{PQ} C_{QQ}^{-1} C_{QL}. \tag{14}$$

It follows that, for example,

$$C_{KL} = \bar{C}_{KL} - \tilde{C}_{KL} + 2 \overline{\sum_{P \neq K,L}} C_{KP} C_{PP}^{-1} C_{PL}. \tag{15}$$

Analogues of most of the other identities of image analysis can also be derived.

6.3. Linear structural models

In this section we study models of the type

$$x_L = A_L f + e_L \tag{16}$$

Here x_L is an observed, and f and e_L are hypothetical vector random variables, A_L is a known or unknown coefficient matrix. We suppose $E(f e_L') = 0$ for all L, and $E(e_L e_K') = 0$ for all $L \neq K$. This implies

$$E(x_L x_L') = A_L E(ff') A_L' + E(e_L e_L'), \tag{17a}$$

$$E(x_L x_K') = A_L E(ff') A_K. \tag{17b}$$

If we concatenate all vectors, write ϕ for $E(ff')$, Σ for $E(xx')$, and Δ for the 'diagonal' supermatrix containing all $E(e_L e_L')$, we can write this as

$$\Sigma = A \phi A' + \Delta. \tag{18}$$

Observe that a more general decomposition is also possible here. We can write

$$x_L = A_L f + B_L g_L + t_L, \tag{19}$$

and assume that $E(fg_L') = E(ft_L') = 0$ for all L, that $E(t_L t_K') = E(t_L g_K') = 0$ for all $K \neq L$, and that $E(t_L t_L')$ is diagonal for all L. Write ψ_L for $E(g_L g_L')$, and Γ_L for $E(t_L t_L')$. Then (18) is true, and for all the diagonal submatrices of Δ we have

$$\Delta_L = B_L \psi_L B_L' + \Gamma_L. \tag{20}$$

Restricting ourselves to the model (16) we want to make linear combinations $x'\alpha$ and $f'\beta$ that have maximal product moment correlation. The stationary equations turn out to be

$$A\phi\beta = \lambda(A\phi A' + \Delta)\alpha, \tag{21a}$$

$$\phi A'\alpha = \lambda\phi\beta. \tag{21b}$$

Assume that ϕ is nonsingular, then

$$\beta = \lambda^{-1} A'\alpha \tag{22a}$$

and

$$(\Sigma - \Delta)\alpha = \frac{\lambda^2}{1 - \lambda^2} \Delta\alpha. \tag{22b}$$

If A, ϕ, and Δ are all unknown, then we can require for identification purposes that $\phi = I$ and that $A'\Delta^{-1}A$ is equal to a diagonal matrix Ω. If we collect all solutions to (22b) in the matrix $\underline{A}$, then the equation

$$(AA' - \Delta)\underline{A} = \Delta\underline{A}\,\Lambda^2 (I - \Lambda^2)^{-1} \tag{22c}$$

has the solution $\underline{A} = \Delta^{-1}A$ and $\Lambda^2(I - \Lambda^2)^{-1} = \Omega - I$. The corresponding solutions to (22a) are $\underline{B} = \Lambda^{-1}\Omega$.

6.4. Error-free subsets

Suppose that there are two sets of variables x and y such that

$$x = Af + e, \tag{23a}$$

$$y = Bf. \tag{23b}$$

Then

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} A\phi A' + \Delta & A\phi B' \\ B\phi A' & B\phi B' \end{bmatrix}. \tag{24}$$

The canonical problem (22b) transforms to

$$\begin{bmatrix} \Sigma_{11} - \Delta & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \frac{\lambda^2}{1 - \lambda^2} \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}. \tag{25}$$

It follows that

$$\alpha_2 = -\Sigma_{22}^{-1} \Sigma_{21} \alpha_1, \tag{26}$$

and

$$(\tilde{\Sigma}_{11} - \Delta)\alpha_1 = \frac{\lambda^2}{1 - \lambda^2}\Delta\alpha_1, \tag{27}$$

with

$$\tilde{\Sigma}_{11} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{28}$$

Consequently if some of the variables are observed without error then we can partial these variables out right away, and restrict our attention to the others. This is of special interest if some of the sets are factors.

6.5. An alternative approach

In 6.3. and 6.4. we maximize the homogeneity of linear combinations of the observed scores x_L with linear combinations of the structural parts $A_L'f$. This is a natural approach, but an obvious alternative is to make linear combinations $\beta_L'A_L'f$ of the structural parts only, and to maximize their internal homogeneity in the sense of chapters 3 and 5. This gives the stationary equations

$$(\Sigma - \Delta)\beta = N\lambda\Theta\beta, \tag{29}$$

where Θ is the diagonal supermatrix with submatrices $\Theta_L = \Sigma_{LL} - \Delta_L$. If some of the sets are error-free we can make a partition as in 6.4., and we find

$$(\tilde{\Sigma}_{11} - \Delta_1)\beta_1 = N\lambda(\Sigma_{11} - \Delta_1)\beta_1. \tag{30}$$

6.6. Some special cases

It is obvious that if A_L is a square, nonsingular transformation, then we can simply rewrite (16) as

$$x_L = f_L + e_L, \tag{31}$$

with $E(f_L e_K') = 0$ for all L, K. This can be interpreted as the familiar decomposition into true and error components. In this interpretation it is sometimes possible to estimate the Δ_L by split-half, test-retest, or parallel forms. If each set contains only one variable the elements of (the diagonal matrix) Δ are the error variances, and applying the procedure in 6.3. means finding the components of the correlation matrix corrected for attenuation. One interpretation is thus that we are maximizing the homogeneity coefficient which is corrected for attenuation. The procedures in 6.3. finds those linear combinations of the observed variables which have maximum reliability.

If A, ϕ, and Δ are all unknown, then (22b) is the fundamental equation of canonical factor analysis, while (29) defines alpha factor analysis. Sections

6.3. - 6.5. make it possible to define CFA and AFA versions of all the multivariate procedures discussed in chapter 5, including the case in which [illegible] the sets are error-free. This last specification, for example, defines two different versions of MANOVA in the common factor or true score space.

6.7. Cluster analysis

In this section we discuss some simple versions of cluster analysis that fit neatly into the framework of chapter 5. Suppose we have a set of n nominal, numerical, or ordinal variables in the first set (matrix X), and a single nominal item with an unknown classification but a known number of k categories in the second set (matrix Y). Our general two-set rationale tells us to maximize w'XY'z over w, z, and Y under the conditions that z'YY'z = wXX'w = 1, and that Y is a binary matrix corresponding with a partition of the m objects into k subsets (we suppose that all linear restrictions are incorporated into X, and especially $\sum x_{ij} = 0$ for all i). For fixed w and Y we find that the optimal z satisfies

$$z \propto (YY')^{-1}YX'w. \tag{33}$$

Thus we want to maximize

$$w'XY'(YY')^{-1}YX'w = \lambda, \tag{33}$$

on the condition that w'XX'w = 1 (or: maximize the between cluster SSQ, with the total SSQ constant), and on the condition that Y is a k-category classification matrix. This problem can be solved for fixed w by fast and reliable integer programming algorithms using the 'string property' of optimal solutions, for fixed Y the optimal w is of course simply the dominant eigenvector of the B/T maximization problem corresponding with (33). We alternate minimization over (one-dimensional) w and over the binary Y. If the maximum is reached (i.e. if Y does not change from one cycle to the next) we compute the corresponding z, which simply contains the within cluster means of the induced score vector Xw. This is computationally a relatively simple procedure, it can be repeated by requiring orthogonality over successive direct quantifications; i.e. we solve the problem all over again, but now we also require v'XX'w = 0, and use generalized inverses to remove the effect of w right away.

A much more complicated procedure is to maximize

$$\mathrm{tr}\{(XX')^{-\frac{1}{2}}XY'(YY')^{-1}YX'(XX')^{-\frac{1}{2}}\}, \tag{34a}$$

or

$$\left|(XX')^{-\frac{1}{2}}XY'(YY')^{-1}YX'(XX')^{-\frac{1}{2}}\right| \tag{34b}$$

directly over Y, and compute all orthogonal solutions w and z for this single optimal clustering. This procedure is computationally very demanding and requires complicated and not necessarily convergent search procedures over the discrete set of all k-category classifications.

If we maximize (34a) a good initial approximation can be computed by relaxing the requirements to the single one that YY' is diagonal (and not necessarily binary). In this case it is easy to see that the optimal solution for $Y'(YY')^{-\frac{1}{2}}$ consists of the k dominant normalized eigenvectors of the matrix $X'(XX')^{-1}X$. This may provide a good starting point for our previous cluster procedures, for example by using an appropriate rounding algorithm that transforms the solution to the closest binary matrix.

Finally we observe that single partitions are the simplests forms of clusters we have. We can define much more general forms of clustering procedures by using essentially the same ideas. The second set can consist of several partitions, of trees or hierarchical clustering schemes, of lattices or partial orders, and so on. In general this defines very complicated computational procedures, which may or may not be worthwhile. Of course it is also perfectly possible that the vectors in X contain independent error in the sense outlined earlier in this chapter. This complicates the algorithms even further.

6.8. Statistical procedures

In chapter 1 we have already indicated in which sense statistical inference procedures are important in our class of techniques. We do not have very specific models, we usually do not want to assume multinormality (if we want to assume multinormality more powerful statistical techniques are available, also for exploratory purposes). This means that the model for the indicator matrix is usually a product multinomial model. The ways of classifications are used to structure the product, the variates define the multinomial distributions. The multivariate vectors of frequencies are asymptotically jointly multinormally distributed, and the elements of the matrices that enter into the generalized eigenproblems are linear functions of these frequencies (even in the general setup with linear restrictions from chapter 5). Two obvious generalizations, which will not be discussed, are Markov dependence and loglinear models.

In the first class of statistical problems we are interested in we have two random square symmetric matrices C and D whose elements satisfy the structural equations

$$c_{ij} = \Sigma a_{ij}^{k} x_{k} + a_{ij}^{0}, \tag{35a}$$

$$d_{ij} = \Sigma b_{ij}^{k} x_{k} + b_{ij}^{0}. \tag{35b}$$

The a_{ij} and b_{ij} are known real numbers, the x_k are jointly asymptotically multinormal with means μ_k and covariances σ_{kl}. We are interested in the

asymptotic distribution of the eigenvalues and the left and right eigenvectors of the matrix $D^{-1}C$. Throughout this section we assume that all eigenvalues are simple. We use the standard results from perturbation theory. Let

$$x_k^* = x_k + \varepsilon\delta^{kl} \tag{36}$$

then

$$c_{ij}^* = c_{ij} + \varepsilon a_{ij}^1, \tag{37a}$$

$$d_{ij}^* = d_{ij} + \varepsilon b_{ij}^1, \tag{37b}$$

or

$$C^* = C + \varepsilon A_1, \tag{38a}$$

$$D^* = D + \varepsilon B_1. \tag{38b}$$

If ε is small enough we may write

$$(D^*)^{-1} = D^{-1} - \varepsilon D^{-1}B_1D^{-1} + \ldots \tag{39}$$

and thus

$$(D^*)^{-1}C^* = D^{-1}C - \varepsilon D^{-1}B_1D^{-1}C + \varepsilon D^{-1}A_1 + \ldots \tag{40}$$

Suppose

$$D^{-1}C\, z_s = \lambda_s z_s, \tag{41a}$$

$$z_s'z_s = 1, \tag{41b}$$

$$CD^{-1}\, y_s = \lambda_s y_s, \tag{41c}$$

$$y_s'y_s = 1. \tag{41d}$$

It follows that

$$z_s \propto D^{-1}\, y_s. \tag{42}$$

According to standard perturbation theory

$$\frac{\partial\lambda_s}{\partial x_l} = \frac{y_s'(D^{-1}A_l - D^{-1}B_lD^{-1}C)\, z_s}{y_s'z_s} =$$

$$= \frac{z_s'(A_l - \lambda_s B_l)\, z_s}{z_s'Dz_s} \equiv g_l^s, \tag{43}$$

and

$$\frac{\partial z_{is}}{\partial\, x_l} = \sum_{t\neq s} \frac{z_t'(A_l - \lambda_s B_l)\, z_s}{(\lambda_s - \lambda_t)\, z_s'Dz_s}\, z_{it} \equiv h_l^{is}. \tag{44}$$

The δ-method gives that the $\hat{\lambda}_s$ and the $\hat{z}_{is}$ are jointly asymptotically normal. We find, for example,

$$\text{ACOV}\,(\hat{\lambda}_s, \hat{\lambda}_t) = \sum_{k=1}\sum_{l=1} g_k^s g_l^t \sigma_{kl}, \tag{45a}$$

$$\text{ACOV}\,(\hat{z}_{is}, \hat{z}_{jt}) = \sum_{k=1}\sum_{l=1} h_k^{is} h_l^{jt} \sigma_{kl}. \tag{45b}$$

In the general case we do not simplify this any further. In a more complicated case we have (C not necessarily symmetric)

$$c_{ij} = \Sigma a_{ij}^{k} x_k + a_{ij}^{0}, \tag{4.4a}$$

$$d_{ij} = \Sigma b_{ij}^{k} x_k + b_{ij}^{0}, \tag{4.4b}$$

$$e_{ij} = \Sigma f_{ij}^{k} x_k + f_{ij}^{0}, \tag{4.4c}$$

and the eigenproblem is

$$E^{-1}C'D^{-1}C\, z_s = \lambda_s z_s, \tag{4.5a}$$

$$z_s'z_s = 1, \tag{4.5b}$$

$$C'D^{-1}CE^{-1}\, y_s = \lambda_s y_s, \tag{4.5c}$$

$$y_s'y_s = 1. \tag{4.5d}$$

Again

$$z_s \propto E^{-1} y_s. \tag{4.6}$$

Letting

$$x_k^{*} = x_k + \varepsilon\delta^{kl}, \tag{4.7}$$

we now find

$$\begin{aligned}&(E^{*})^{-1}(C^{*})'(D^{*})^{-1}(C^{*}) = \\ &(E^{-1}-\varepsilon E^{-1}F_1D^{-1}+\ldots)(C'+\varepsilon A_1')(D^{-1}-\varepsilon D^{-1}B_1D^{-1}+\ldots) \\ &(C+\varepsilon A_1) = \\ &E^{-1}C'D^{-1}C + \varepsilon(E^{-1}C'D^{-1}A_1 + E'A_1'D^{-1}C - E^{-1}C'D^{-1}B_1D^{-1}C - \\ &E^{-1}F_1E^{-1}C'D^{-1}C) + \ldots .\end{aligned} \tag{4.8}$$

Letting

$$u_s = D^{-1}C\, z_s, \tag{4.9}$$

we find

$$\frac{\partial\lambda_s}{\partial x_1} = \frac{2\, u_s'A_1z_s - u_s'B_1u_s - \lambda_s z_s'F_1z_s}{z_s'E_1z_s}, \tag{4.10}$$

and so on, Again we do not simplify these formula's because very specific simplifications are possible in special cases.

As a simple example we consider a general form of the techniques discussed in chapter 4. We suppose that there is a number of fixed symmetric matrices $A_1, \ldots, A_m$. We define

$$\hat{C} = \sum_{l=1}^{m} \hat{p}_l A_l, \tag{4.11}$$

where $\hat{p}_l$ is the proportion of the n subjects (occasions, replications, etc.) that respond with (produce) the outcome corresponding with the matrix A_l.

If the A_i are the observed matrices we can obviously also write

$$\hat{C} = \frac{1}{n} \sum_{i=1}^{n} A_i. \tag{54}$$

If the y_s are the normalized eigenvectors of C and λ_s the corresponding eigenvalues, then

$$\frac{\partial \lambda_s}{\partial p_l} = y_s' A_l y_s. \tag{55}$$

Assuming a simple multinomial model over the matrices we find

$$ACOV(\hat{\lambda}_s, \hat{\lambda}_t) = \frac{1}{n} \sum_{l=1}^{m} (y_s' A_l y_s)(y_t' A_l y_t) p_l - \frac{1}{n} \sum_{l=1}^{m} (y_s' A_l y_s) p_l \sum_{l=1}^{m} (y_t' A_l y_t) p_l. \tag{56a}$$

We can estimate this covariance by

$$ECOV(\hat{\lambda}_s, \hat{\lambda}_t) = \frac{1}{n} \left\{ \sum_{i=1}^{n} (\hat{y}_s' A_i \hat{y}_s)(\hat{y}_t' A_i \hat{y}_t)/n - \hat{\lambda}_s \hat{\lambda}_t \right\}. \tag{56b}$$

Needless to say that we expect that this is only a fair approximation if n is very large (much larger than m). In the procedure of 4.2. the A_i are equal to the major product moments of the centered vectors of ranks, i.e. $A_i = r_i r_i'$, and if K is the number of ranked objects then $m = \frac{1}{2}K!$. In the same way in a paired comparison experiment the method of chapter 4 generates a set of possible matrices, and in a similarity experiment the same thing is true.

As a more complicated example we discuss the technique from 3.10. The problem is

$$CE^{-1}C' z_s = \lambda_s D z_s, \tag{57a}$$

$$z_s' z_s = 1. \tag{57b}$$

Define

$$y_s = E^{-1}C' z_s, \tag{58a}$$

$$\gamma_s = z_s' D z_s. \tag{58b}$$

Here C contains the cell probabilities p_{ij}, and D and E contain the marginal probabilities. After a lot of algebra we find

$$\frac{\partial \lambda_s}{\partial p_{kl}} = \frac{(1 - \lambda_s) z_{ks}^2 - (y_{ls} - z_{ks})^2}{\gamma_s} \tag{59}$$

It follows that

$$\sum\sum \frac{\partial \lambda_s}{\partial p_{kl}} p_{kl} = 0. \tag{60}$$

Detailed formula's for this and other special cases will be worked out in further publications. The same thing is true for the more complicated techniques in chapter 4 and 5, and for the essentially different techniques of this chapter. The asymptotic distribution of the λ_s and z_{is} can of course be used to construct asymptotically optimal tests in the sense of Wald, and asymptotic optimal confidence regions in the sense of Wald or Wilks. The hypotheses we test will mostly be of the form $\lambda_s = \lambda_t$ for all s, t in a subset of the eigenvalues, or $\lambda_s = c$ for all s in a subset. These hypotheses often have no clear interpretations in terms of a relatively simple statistical model for the observed frequencies, but this is, of course, unavoidable in explorative situations like ours. Assuming stochastic independence of the individual observations, and assuming a structure in terms of variates and ways of classification is often about all we can do.

6.9. Criticisms

The extension of image analysis and common factor analysis to sets of numerical variables seems straightforward enough, and there seem to be no extra problems with respect to interpretation. If some of the variables are nominal the interpretation becomes much more difficult, however, In the first place a purely linear model for probabilities seems to have at most a limited descriptive value, because of the natural boundaries of probabilities between zero and one. This is not as serious as it seems, because most of our procedures can also be interpreted in the framework of nonlinear factor analysis (McDonald 1962, 1967, 1968).

Nonlinear factor analysis is, however, a very problematic class of techniques. The fundamental weakness of classical Spearman-Thurstone common factor analysis is the indeterminacy in the model $x = Af + e$, investigated by Thomsom (1935), Ledermann (1938), Kestelmann (1952), Guttman (1955), Heerman (1964, 1966), and Schönemann & Wang (1972). There are no respectable statistical techniques to estimate factor loadings, factor scores, and unique variace simultaneously (except under highly restrictive assumptions), and the indeterminacy problem makes the scientific value of factor scores doubtful. This has as a practical consequence that from most points of view it seems better to work directly with the structural covariance model.

$$\Sigma = A \phi A' + \Delta, \tag{61}$$

and to remember (16) only as a possible justification for (61). The distinction between linear and nonlinear factor analysis does not make much sense without reference to the factor scores.

The problem to describe individuals in common factor space remains unsolved, although there is always the possibility to do a Q-type technique. The model (61) can be used in the context of nominal variables, it simply decomposes the deviations from bivariate independence of the nondiagonal submatrices. The idea of interpreting the diagonal submatrices differently from the rest is, of course, perfectly natural.

6.10. Historical

Partial canonical correlation analysis is a familiar extension of both canonical correlations analysis and partial regression. A satisfactory description of the various statistical aspects is the article by Rao (1969). I do not know any practical applications. Image analysis in the form in which we use it has been proposed by Guttman (1953). Important further contributions have been made by Guttman (1960), Harris (1962), and Kaiser (1963). A closely related statistical model is described by Jöreskog (1963, 1969).

The generalizations of common factor analysis discussed in 6.3. - 6.5. are partly due to McDonald. In McDonald (1968b) he discusses the general problem of canonical analysis in terms of maximizing ratios of quadratic forms that can be interpreted as variance ratios of linear combinations in models such as (16), (19), (23), and (31). In McDonald (1969a) he applies the results to the problem of defining a principal factor analysis (PFA) of n nominal variables, and in McDonald (1970) he studies the general case of groups of variables using the Lawley-Whittle equal residual variance model, the Guttman-Jöreskog image model, and the CFA model. In McDonald (1970) the PFA, CFA, and AFA approaches are discussed and contrasted for the classical case of one and only one variable in each group. His conclusion is that CFA has more useful structural properties than AFA and PFA. This is in agreement with the conclusions of McDonald & Burr (1967), and Browne (1969). In McDonald (1969b) the model (61) is generalized to the case where Δ can be a general symmetric matrix with zeroes at specified places. De Leeuw (1972) discusses several algorithms for this and even more general cases.

McDonald (1970) also describes the procedures discussed in the first paragraph of 6.6.. Meredith (1964) discusses an AFA-type matching technique which is the special case of 6.5. with $N = 2$ and Δ_1, Δ_2 both diagonal. The CFA-type procedure of finding linear combinations that maximize the reliability is due to Thomson (1940), Mosier (1943), Peel (1948), and Green (1950). The CFA model

in the classical case is due to Rao (1955), the AFA model to Kaiser & Caffrey (1965). The CFA equations turn out to be identical to those for maximum likelihood factor analysis (Lawley, 1940), and minimum determinant factor analysis (Howe, 1955; Bargmann, 1957). CFA can easily be extended to the more general individual differences models discussed in 4.9.. This has been done by Jöreskog (1971).

Structural models of the form (61) with A known have as familiar special cases the variance component models in ANOVA and MANOVA, the quasi-simplex models investigated by Mukherjee (1966, 1969) and Jöreskog (1970b), and several others of the general type discussed by Bock and Bargmann (1966), Srivastava (1966), Anderson (1966, 1968), Mukherjee (1970), and Jöreskog (1970a).

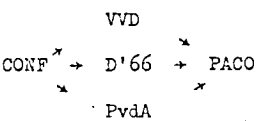
The 'string property' of optimal solutions of cluster problems is due to Fisher (1958), and it is used by Vinod (1969) and Rao(1971) to devise efficient integer programming algorithms. The alternative approach based on (34a) or (34b) is discussed by Friedman & Rubin (1967). The first eigenvector approximation to these solutions was discussed by Wilcy (1967) under the name of latent partition analysis.

Perturbation theory is discussed, for example, in Wilkinson (1965, p. 62-70). In a psychometric context it was used by Derflinger (1968) and Clarke (1970) to compute the first and second partials in ML and LS factor analysis. The δ - method is discussed by Doob (1935), Mann & Wald (1943), Hsu (1949). The relevant theorems are beautifully summarized in Rao (1965, section 6a.2, pp. 319-322). The relevant asymptotical statistical theory is almost completely discussed in Wald (1941a, 1941b, 1942, 1943).

I think that most political scientists (cf Stapel 1968, Daalder & Rusk 1971) would agree on the following order of the parties on the (conceptual) left-right dimension

PACO → PvdA → D'66 → CONF → VVD,

and the following partial order on the conceptual dimension of religious affiliation



(cf also De Leeuw 1968, De Gruyter 1967).

As a first result we have analyzed the marginal tables 1.2.b and 1.2.c with the techniques from section 3.10 (results previously reported in De Leeuw 1971a, 1971b). We used the program CARD01P. The first two components are given in tables 1.3.a and 1.3.b, and plotted in figures 1.4.a and 1.4.b. The corresponding partition of the total X^2 of the tables is given in 1.5.a and 1.5.b (observe that the three components are not independent χ^2, not even on the hypothesis of complete homogeneity of rows). The interpretation of the results is beautifully clear. If we compare the projections and the rank orders we find that in 1.3.b the first dimension is left-right, the second is religious affiliation. In 1.3.a the role of these two dimensions is interchanged, i.e. students pooled within faculties over universities use the left-right dimension in their political choices, students pooled within universities over faculties use the religious dimension. Of course the fit to the conceptual order can be improved by using oblique nonmetric Procrustus rotation (De Leeuw 1969a), but this is hardly necessary.

Lammers (1969) used prior integer scores for parties corresponding with the linear order we discussed for left-right. This defines an induced quantification for faculties which corresponds closely with our computed direct canonical quantification (plot in figure 1.6). The corresponding partition of X^2 is given in table 1.7.a. The main difference between the two quantifications is the score for theology, which is not surprising because of our second dimension. In Lammers' scoring system theology is a faculty somewhere in the middle of the scale, in our quantification it is clearly on the left side of the scale (and, in fact, this supports Lammers' conclusions from his data even better than his own scores seem to do). In a factor analysis terminology there is only one common factor, and theology has a large unique variance. Because of the results of our second canonical analysis it is also not very surprising that Lammers'

7 Examples

7.0 Introduction

In this chapter we give examples of some of the procedures discussed in the previous chapters. We used PL/I programs from the old CDARD series, from the more recent CARDOOP series, a very general new program CANONO1, and some ad hoc programs written in APL. The procedures discussed in chapter 6 are not yet programmed for the computer. We supplement some of the output of the canonical techniques by results obtained using other techniques. We tried to select a small number of interesting examples, which do not offer too much interpretational or computational difficulties, and which are consequently not completely representative for the types of data to which our procedures can be applied. Large scale applications in traffic research and in a survey of Dutch parliament members are in preparation and will be published separately. A large series of FORTRAN programs is prepared in cooperation with the computing center of the university. We want to thank professor Lammers and professor Daalder for their permission to use their data in this chapter.

7.1 Data 1: students and politics

The first set of data (provided by prof. Lammers) is given in table 1.1.a-1.1.1. The Dutch student council NSR collected in 1968 first choices among the five major political parties of 1616 university students. The sample was stratified over 12 universities and 13 faculties (i.e. the total number of students in each university-faculty combination was fixed, universities is a factor, only political choice is a random variable). The three marginal two-dimensional tables are given in 1.2.a-1.2.c (observe that 1.2.a is completely fixed, 1.2.b and 1.2.c are of the comparative trial type, i.e. one marginal is fixed). the abbreviations used are

Faculties

JUR	Law.
MED	Medicin.
W&N	Math.& physics, chemistry.
SOC	Social sciences.
LET	Literature, language.
TEC	Technical.
PSW	Political & social.
VEE	Veterinarians.
TND	Dentists.
THE	Theology.
LBW	Agriculture.
CIF	Philosophy.
ECO	Economy.

Political parties

CONF	Denominational parties (KVP, ARP, CHU, GPV, SGP).
VVD	Conservatives.
PvdA	Socialists.
PACO	Pacifists, communists.
D'66	Pragmatists.

scores perform poorly on table 1.2.c. The partition of X^2 is given in table 1.7.b.

In table 1.8.a we have formed all possible combinations of parties and computed the within group X^2 (the order here is CONF → VVD → PvdA → PACO → D'66, and thus 10010 means the group CONF - PACO). Of course the within group X^2 of 11111 is equal to the total X^2, and the within group X^2 of groups consisting of a single party is zero. In table 1.8.b these within components are used to make additive within-between partitions of X^2. In a sense the ideal situation in a parliamentary democracy is a governmental coalition that is as homogeneous as possible, and an opposition that is as homogeneous as possible. This implies that government and opposition must be as different as possible from each other. The student data imply that the best possible situation is a government consisting of CONF - PvdA - PACO - D'66, and an opposition that consists of VVD only. The second best situation is close to the actual situation in the Netherlands. We have CONF - VVD in government and PvdA - PACO - D'66 in the opposition. If the opposition is divided into two groups we find another situation close to the actual one (CONF - VVD in government, PvdA - D'66 in opposition 1, PACO in opposition 2) also satisfactory in terms of homogeneity. Due to more recent political developments it may be interesting to observe that

10101	62.489	24
01000	0.000	0
00010	0.000	0
	115.166	24
	177.655	48

is even better.

We also analyzed table 1.1 with a PCA program for three categorical variates called CARDO4P. Observe that this is not the correct probabilistic interpretation of these data, and the consequences of this are interesting. In table 1.9 all weights which are in absolute value at least .010 are given for the first seven components. The interpretation is

I TECHNICAL universities can be found in DELFT, DRIENENOORD, EINDHOVEN.

II An AGRICULTURAL university can be found in WAGENINGEN.

III ECONOMICAL universities can be found in ROTTERDAM, TILBURG.

IV POLITICAL dimension 1.

V POLITICAL dimension 2.

VI You can only become a VETERINARIAN in UTRECHT.

VII ????

Factors I, II, III, IV, and possible VII are almost completely determined by the peculiarities of the (nonrandom) table 1.2.a. Only factors IV and V seem to have some political relevance.

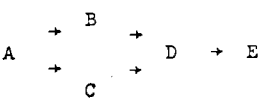
It can be seen that IV contrasts the left with the denominational right, and V contrasts the denominational right with the non-denominational right (i.e. V contrasts the two meanings of the word 'right' in political language, cf Stapel 1968).

7.2 Data 2: leaving primary school

The second set of data is given in table 2.1.a-2.1.h. In 1970 we collected data in eight GLO schools in Leiden. For each of the pupils of the 6th grade we recorded the occupation of the father and the type of secondary education the pupil was going to have after leaving primary school. Previous analysis of similar, more extensive, data indicate that all types of secondary schools could be classified without much loss of information as

A: HAVO, MAVO, VWO.
B: LTS, LHNO.
C: LAVO, LEAO.
D: BTS, INOM, MEL.
E: NO FURTHER EDUCATION.

The implied partial order (either in terms of difficulty, or intelligence, or expected later income, or social status, or what have you) is



The occupation of the father was classified into five categories (a linear order is implied).

α: Academic, directors, ...
β: Higher white collar, army officers, ...
γ: White collar, shop keepers, ...
δ: Lower white collar, skilled labour, ...
ε: Unskilled labour.

The information was collected directly by the principals of the schools who also did the classifying of the fathers (the social status scale is, of course, a rather weak point in the data gathering procedure). In table 2.1 rows refer to occupations, columns to school types. The bivariate and univariate marginals are given in 2.2.a - 2.2.c.

Previous analysis of similar data had given us the idea that almost all variation in the marginal table 2.2.c is due to the distinctions (α, β, γ) versus (δ, ε), and A versus (B, C, D, E). Accordingly we have defined the $4 \times 4 = 16$ contrasts displayed in table 2.3. The last two columns of 2.3 contain an exact and an asymptotic partition of X^2. The only significant single degree of freedom chi-squares seem to be: (α, β, γ) go more to A than (δ, ε); (α, β) go more to A

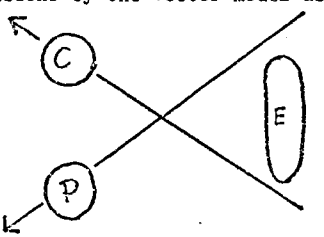
than γ; δ goes more to A than ε; and, possibly, of all (δ, ε) not going to A, δ goes more to (B, C) and ε goes more to (D, E). The simple partition A versus BCDE explains about 80% of the total X^2 , and the within BCDE component is not even significant (table 2.4.b). This seems to suggest that our school system concentrates on separating the A-candidates from the rest, and that the choice between BCDE is considered less important. On the other hand the data set is small, the number of people not going to A is even smaller, the number of zeroes in table 2.2.c is large, and the asymptotic distribution theory may be quite misleading. This means that we must be careful with our interpretations of the BCDE effects. That the linear order on the occupations is there in an overwhelming way (with respect to A-going or not-A-going) is obvious enough.

A more refined analysis, at least from the data theoretical point of view, is the canonical method of section 3.10. Using the APL program CACT01 we found that the first component had a X^2 of 85.756 (corresponding with a canonical correlation of $|\rho| = .593$). The joint scale in figure 2.5 shows that our prior orderings are reproduced quite nicely. The most remarkable features are the large distance between A and BCDE (remember that X^2 for this prior contrast was 82.374, very close to the optimum), and the large distance between γ, δ, and ε. Our 'crude' conclusion that the differentiation between BCDE is more or less random does not seem to be completely valid. Another interesting question is in how far we can predict choice of secondary education from the two remaining variables. The technique is discussed in chapter 5, the program used was CACT02, written in APL. There are two significant components with squared canonical correlations $\rho_1^2 = .516$ and $\rho_2^2 = .231$. The joint plot is given in figure 2.6. Again the main effects appear to be A vs BCDE and BC vs DE. They are closely related to αβγ vs δε, and δ vs ε. Again it is remarkable that the cluster A/αβγ/OP,LO,LL,DR is very homogeneous, much more so than the other clusters.

7.3 Data 3: political preference

In 1968 prof. Hans Daalder and his collaborators asked the 150 members of the Dutch Lower House (Second Chamber) and the 75 members of the Dutch Upper House (First Chamber) for their preference rank orders of the twelve most important Dutch political parties. Only 141 out of 150 members of the Second Chamber and only 70 out of 75 members of the First Chamber actually responded. We analyzed both sets using the CARD02P program in two dimensions, the technique is the one discussed in section 4.2. A plot of the projections on the first two principal components for the Lower House data is given in figure 3.1.a. Results for the upper house are almost identical, the plot is given in figure 3.1.b. For the interpretation it is useful to draw the two arrows representing the two large clusters PvdA - PPR - D'66 and KVP - ARP - CHU - VVD. Projections on these arrows

(directions) represent the typical preference patterns of people in these clusters. Observe that is is impossible to accomodate the preference rank orders of the members of SGP, GPV, CPN, PSP, and BP in this way. In this preference space the fact that there are two homogeneous clusters, consisting of the progressive (P) and conservative (C) parties, and a heterogeneous cluster of small extremist (E) parties leads automatically to something closely related to the curved left-right dimension discussed by De Gruyter (1967), De Leeuw (1968, 1969b), Van de Geer (1970). The major preference patterns are (P) > (C) > (E) and (C) > (P) > (E), and they can be accomodated in two dimensions by the vector model as follows



(for the distance model basically the same thing is true). Observe that the curved dimension necessarily has some holes. If we plot the induced scores for individual members of parliament the heterogeneity within parties will tend to take care of filling these holes. In figure 3.2 we have plotted the 141 members of the lower house (the scales of 3.1 and 3.2 are not directly comparable, we could make them so and draw a joint plot). Observe that the hole between the clusters is still not completely filled. The code used in 3.4 is: PSP = p; PvdA = q; D'66 = d; PPR = r; KVP = k; ARP = a; CHU = c; VVD = v; BP = b; SGP = s; GPV = g. We compare the CARDO2P output with a number of related results. In the first place Daalder and Rusk (1971) analyzed the same Lower House data with Roskam's UNFOLD program (cf Roskam 1968). The plot is given in figure 3.3. It is a nice illustration of a point we made in 3.23, and earlier in 1.3. The UNFOLD program tends to degenerate the configuration, not only by moving the very unpopular BP far from the other parties, but also by collapsing homogeneous subgroups of parties into single points. It is possible that if we increase the precision and continue iterating then BP moves even farther out, and our (P) and (C) clusters collapse even more (this is related to the fact that Kruskal's stress is very flat near the minimum). Daalder and Rusk show that by moving BP to the arbitrary point BP' the stress increases from .119 to .121. This calles for two comments. The flatness of the stress function has the advantage that the simple gardient method is relatively effective. If we try to minimize the simple transform ln S/(1-S) or S^t with t >> 1 with simple gradient methods we may be in some computational trouble. In the second place it has been

convincingly argued by Torgerson (1965) that the degenerating effects of MDS programs may be very helpful for some data sets, and in the NMSPOM programs (De Leeuw 1970) we have a special parameter which regulates collapsing and which can be used to make the MDS program into a cluster analysis program. In our particular data set we have seen that the cluster interpretation is very helpful, but the CARD02P program gives much more additional useful within-cluster information than the UNFOLD program does. A comparison of 3.1.a and 3.1.b shows that the within-cluster information is reliable. Further information on these data and their relevance for political science can be found in the article of Daalder and Rusk (since then published in S.C. Patterson (ed): Comparative Legislative Behaviour, New York, Wiley, 1972).

In the second place it may be interesting to compare our results with other political preference data analyzed by the similar CDARD2 program, and reported in De Leeuw (1968, 1969). The data from a sample of 100 psychology students collected by Dr. Leo Van der Kamp in 1968 and plotted in figure 3.4 show that the situation here is somewhat different than in parliament. The progressive students choose D'66 or PvdA, the conservative ones choose VVD. The denominational group KVP - CHU -ARP is very unpopular among students (cf also data 1). They have only 16% of first choices in this sample (in the Lower House sample 49%), the VVD has 31%, D'66 23%, and the PvdA 19%. Although the within-cluster preferences are completely different in this sample, the same clusters are still there. The situation is again completely different if we analyze the data of 80 districts in Amsterdam during general elections for parliament in february 1967. In figure 3.5 we see three different scales, one general (G), one for the establishment (E), and one for the people who are not satisfied with the current system (U). It is important to observe that CPN, BP, and PSP were all very popular in Amsterdam at that time (having about 40% of the total popular vote). A more precise analysis of the individual districts shows that some of them have a hard-core CPN vote , but other districts have a 'dissatisfied' CPN vote (is these districts the BP vote is also very high, and between elections a lot of people switch from BP to CPN and back).

7.4 Data 4: political similarity

The conditional similarity matrix in table 4.1 was obtained by summing conditional similarity rank orders for 11 subjects (also collected in 1968, subjects were students and staff members of the psychological institute). We use the data to illustrate the DSCN technique from section 4.6. There are two large eigenvalues, the projections on the first two principal components are plotted in figure 4.2. We can compare this with the output of a 'proper' nonmetric analysis by using the results form the NMSPOM program (Minkovski exponent 2, error weighting 2, cf De Leeuw 1970). This is plotted in figure 4.3. The

distances constructed by the two different techniques in two dimensions are plotted against each other in figure 4.4. The NMSPOM solution can be interpreted as a two-dimensional solution, with the major dimension standing for left-right and the second dimension for religious affiliation (in general rotation is not necessary for NMSPOM because we use floating point exponents for the power metrics). This interpretation has some unsatisfactory features, an alternative interpretation is that we have considerable variation around a single linear left-right dimension. The DSCN solution is more easily interpretable, and gives us a small amount of scatter around the familiar curved left-right dimension, already found (with varying degree of closure and curvedness) by De Gruyter (1967), and Roskam (1968, p 70-76), and since by many others.

Of course the summation of the data of individual subjects may very well be questioned in cases like this, and a conditional TWWD analysis may be much more appropriate. We used Van der Kloot's CDARD9 program (described in Van der Kloot 1969), which computes the dimensions successively with deflation after the iterations have converged (I now think that simultaneous optimalization for a fixed number of dimensions is usually better, but there is no program yet which does this). The configuration (with axes of equal length) is given in 4.4. The first two dimensions explain 53 + 13 = 68 % of the total sum of squares, which is quite a lot for such a restrictive model. It is well known that there is no rotational indeterminacy in models such as TWWD and TWWI, and therefor it is interesting to look at the projections on the dimensions. The first dimension is clearly left-right in the political sense, the second dimension can be identified with political extremism. The position of the SGP on this second dimension is a bit strange (just as in figure 4.2 we expect SGP to be closer to BP). Inspection of table 4.1 shows that the subjects tend to emphasize the fact that SGP is a denominational Calvinist party, very close to ARP and CHU, and not so much that it is also a party on the extreme right wing in the political sense (in fact our subjects think that BP is closer to VVD than to SGP). A little reflection shows that this effect is also dramatized by the fact that we have conditional data. We must also remember that the first dimension is much more important than the second, and that this can not be seen from figure 4.4. In figure 4.5 we give the weights given by the various subjects to the two dimensions (vector of weights scaled to unit length over subjects for each dimension). There is only a little variation, especially in the weights for the first dimension, and only subject 10 clearly does not fit into the general pattern. After rescaling we find that for subject 2 both dimensions are about equally important (which means that 4.4 is representative for this subject). For subjects 4 and 7 we find that the first dimension is about twice as important as the second. Consequently their configuration is better represen-

ted by figure 4.6, which is much more like 4.2.

7.5 Data 5: politics and attributes

In 1968 we selected 12 political parties and 17 attributes, and we asked eleven subjects (students and staff members of the psychological institute) for each of the 12 × 17 combinations if they thought that this particular party had that particular attribute or not. Responses for one single subject are given in table 5.1. Data of this type occur in may situations, and are somewhat difficult to analyze. They could be analyzed by unfolding type techniques, but the transpose of the matrix could equally well be analyzed by unfolding type techniques. In the GL-SSA series there seems to be a program that maintains monotonicity within rows and within columns, but this is also not exactly what we want. As a preliminary analysis we use the old CDARD1 program, in which the attributes are hyperplanes that separate the parties. The results (projections on dimensions 1 and 2) are given in table 5.2. It is clear from this table that the BP is a straggler, which is due to the fact that (according to our subject) it is the only negative and irresponsible party. In table 5.3 we have collected the signed distances, which are simply the induced scores for attributes (proportional to the direction cosines of the separating hyperplanes). In stead of using these direction cosines we have drawn in hyperplanes that seem to do the separating at least as good (figure 5.4, SGP and GPV are represented by the single point SG). The only violation of the nonmetric separting requirements is the fact that BP is classified as a left-wing party by attribute 3. By moving BP to the point (BP) we can achieve a perfect SSC solution in the sense of De Leeuw (1969c) or a perfect MSA-IV solution in the sense of Guttman (cf Lingoes 1968). Observe that (especially if we move BP to (BP)) the circular configuration of political parties is there again. It is now a small step to see that the attributes define segments of this circle, i.e. there also is a SSC(E) representation in the sense of De Leeuw (1969c), and only the attribute opportunism violates it. The representation is drawn in figure 5.5, and the use of the ideas of Shepard and Carroll (1966), Roskam (1968), Van de Geer (1970), and De Leeuw (1969c) can be used to map this representation into a single dimension. A final type of representation is the cluster representation in figure 5.6 (cf also table 5.7). In this example we have used CDARD1 as an approximation to the requirements of nonmetric programs. As such it clearly does an excellent job.

7.6 Data 6: spot patterns

This example illustrates the usefulness of our procedures in some of the standard situations treated in psychophysics. The data are Guilford's spot patterns (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, attempting to keep interpile distances psychologically equal. There were ten replications of this same experiment. The data can be collected in a 90 × 100 indicator matrix in which each replication defines a 9-category variable, and each card defines a column. We can apply our PCA procedures from chapter 3 directly to this matrix, but it seems appropriate here to require that the direct quantifications of the ten variables are identical (cf section 3.11). If we also require that the quantifications of the cards with the same number of spots must be identical we can use formulas 3.29 and 3.30 for simultaneous direct quantification to show that the problem becomes equivalent to applying the procedure of section 3.10 to the contingency table given in Guilford (of course the more complicated indicator matrices are not even given by Guilford). Our analysis consequently maximizes homogeneity over replications under these natural equality constraints, it also finds scores which maximize the correlation and the binormality of the table, which linearize the regressions, and which try to reproduce the Benzécri distances. Data analytically the homogeneity interpretation seems the most natural one by far. The analysis shows that the data can be almost completely transformed to binormality, which makes only the first component interesting. The maximal correlation is close to .93, the quantification of the patterns is plotted against the number of spots in figure 6.2, the category quantification against the category number in figure 6.3.

Two conclusions are immediately obvious. In the first place intervals between extreme categories are considerably smaller (from the homogeneity point of view) than the subject thinks they are. This is partially in agreement with results using the method of paired comparisons (Guilford l.c. p 206-207). In the second place the relation in figure 6.2 is linear and not logarithmic,, which means that if we maximize homogeneity then Fechner's law does not come out. In Guilford's analysis based on the method of equal intervals (l.c. p 204-205) Fechner's law comes out nicely, but in the paired comparison analysis the regression of the scale values on the logarithm of the number of spots was positively accelerated, i.e. the relationship also tended to linearity.

FIGURES AND TABLES

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					
	LEIDEN					A'DAM VU					A'DAM GU					UTRECHT				
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										
	WAGENINGEN					GRONINGEN					DRIENENOORD					NIJMEGEN				
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										
	TILBURG					ROTTERDAM					EINDHOVEN					DELFT				

TABLE 1.1.a - 1.1.1.
ORDER: CONF, VVD, PvdA, PACO, D'66.

	JUR	MED	W&N	SOC	LET	TEC	PSW	VEE	TND	THE	LBW	CIF	ECO	TOT
DELFT						182								182
EINDH						35								35
RODAM	5	24		7									75	111
TBURG	3			15									25	43
NYGEN	22	25	21	37	18				5	11		1		140
DRIEN						19								19
GRONI	23	26	35	36	20				5	4		3	26	178
WAGEN				2							52		1	55
UTREC	29	39	68	36	34			16	4	8		6		240
ADAMG	41	39	41	67	55		11			1		22	57	334
ADAMV	3	14	8	5	3					5		1	6	45
LEIDE	74	44	41	38	27					7		3		234
TOTAAL	200	211	214	243	157	236	11	16	14	36	52	36	190	1616

TABLE 1.2.a.

	JUR	MED	W&N	SOC	LET	TEC	PSW	VEE	TND	THE	LBW	CIF	ECO	TOT
CONF	33	37	33	28	22	36	1	3	2	19	11	2	24	251
VVD	79	61	40	34	31	80	1	6	7	1	14	4	73	431
PvdA	33	30	45	58	40	28	5	1	0	5	7	9	30	291
PACO	6	19	22	38	20	24	0	1	1	4	6	9	8	158
D'66	49	64	74	85	44	68	4	5	4	7	14	12	55	485
TOTAAL	200	211	214	243	157	236	11	16	14	36	52	36	190	1616

TABLE 1.2.b.

	DEL	EIN	ROD	TBU	NYG	DRI	GRO	WAG	UTR	ADG	ADV	LEI	TOT
CONF	24	12	20	8	33	0	20	11	40	27	27	29	251
VVD	66	7	43	12	22	7	46	15	65	59	1	88	431
PvdA	22	3	20	3	22	3	43	8	36	92	5	34	291
PACO	20	0	4	3	13	4	14	6	27	43	3	21	158
D'66	50	13	24	17	50	5	55	15	72	113	9	62	485
TOTAAL	182	35	111	43	140	19	178	55	240	334	45	234	1616

TABLE 1.2.c.

CONF	.029	-.001
VVD	-.007	-.017
PvdA	-.008	.014
PACO	-.007	.007
D'66	-.001	.005
DELFT	-.003	-.012
EINDH	.028	-.003
RODAM	.002	-.015
TBURG	.007	-.007
NYGEN	.012	.009
DRIEN	-.022	-.007
GRONI	-.006	.005
WAGEN	.006	-.003
UTREC	.002	-.002
ADAMG	-.010	.016
ADAMV	.059	.011
LEIDE	-.005	-.013

TABLE 1.3.a

CONF	-.002	.024
VVD	-.018	-.005
PvdA	.012	-.005
PACO	.020	-.002
D'66	.004	-.004
JUR	-.016	-.000
MED	-.004	.003
W&N	.008	.000
SOC	.017	-.005
LET	.010	-.002
TEC	-.008	-.000
PSW	.019	-.013
VEE	-.016	.005
TND	-.027	-.002
THE	.012	.062
LBW	-.002	.009
CIF	.027	-.013
ECO	-.014	-.006

TABLE 1.3.b

Canonical partition universities

Component 1	113.809
Component 2	66.472
Residual	22.647
Total	202.928

TABLE 1.5.a

Canonical partition faculties

Component 1	108.050
Component 2	47.870
Residual	21.730
Total	177.650

TABLE 1.5.b

01111	129.844	00111	35.516	00011	21.097
10111	88.220	01011	97.776	00101	13.688
11011	143.206	01101	90.514	00110	20.010
11101	138.063	01110	121.637	01001	50.228
11110	166.642	10011	66.903	01010	77.142
		10101	62.489	01100	72.881
		10110	74.176	10001	38.628
TABLE 1.7		11001	96.771	10010	43.558
		11010	121.124	10100	46.354
		11100	119.461	11000	56.440

10111	88.220	00111	35.516	11000	56.440
01000	0.000	11000	56.440	00101	13.688
	89.435		85.699	00010	0.000
					107.527
	177.655		177.655		177.655
10110	74.176	11001	96.771	10100	46.354
01001	50.228	00110	20.010	01001	50.228
	47.251		60.874	00010	0.000
					81.063
	177.655		177.655		177.655

TABLE 1.8.a - 1.8.f

Environment	X^2	dfr	P
Within αβγ	5.527	8	.30
Within δε	26.054	4	<.01
Between αβγ-δε	68.193	4	<.01
Total	99.774	16	<.01

School	X^2	dfr	P
Within A	0.000	0	---
Within BCDE	17.400	12	.14
Between A-BCDE	82.374	4	<.01
Total	99.774	16	<.01

TABLE 2.4.a - 2.4.b

	I	II	III	IV	V	VI	VII
JUR					-.023	-.018	
MED				-.011			
W&N			-.011			.012	
SOC							-.021
LET				.011			
TEC	.038						
PSW				.057	.020	-.020	.060
VEE			-.019	-.022	-.018	.110	.017
TND				-.019		.018	-.068
THE			-.018	-.039	.042	-.017	
LBW		.087					
CIF				.034	.011		.026
ECO			.039				
DELFT	.038						
EINDH	.038				.017		
RODAM			.043	-.011			
TBURG			.034		.012		-.045
NYGEN			-.012		.013	-.012	-.022
DRIEN	.038			.012			
GRONI							-.015
WAGEN		.084					
UTREC			-.012			.029	
ADAMG				.020			.010
ADAMV				-.036	.043	-.013	.031
LEIDE					-.022	-.017	
CONF				-.021	.017		
VVD					-.016		
PvdA				.013			
PACO				.012			
D'66							
λ	.672	.658	.539	.480	.451	.422	.397

TABLE 1.9

JL				
1	0	0	0	0
2	0	0	0	0
4	3	3	0	0
2	5	2	1	0
0	3	2	1	0

WZ				
0	0	0	0	0
3	0	0	0	0
0	0	1	1	0
6	7	1	6	2
0	0	0	3	1

LO				
5	0	0	0	0
2	0	0	0	0
13	0	0	0	0
8	1	1	1	0
0	0	0	0	0

GU				
0	0	0	0	0
0	0	0	0	0
4	2	0	0	0
6	9	1	2	0
0	0	0	1	0

DR				
3	0	0	0	0
9	0	0	0	0
9	0	0	0	0
9	3	0	0	0
0	0	0	0	0

DO				
0	0	0	0	0
0	0	0	0	0
1	0	0	0	0
2	9	2	0	0
1	6	1	1	0

OP				
16	0	0	0	0
7	0	0	0	0
5	0	0	0	0
9	2	0	0	0
1	0	0	0	0

LL				
3	0	0	0	0
6	0	0	0	0
9	1	0	0	0
7	0	0	0	1
1	0	0	1	1

TABLE 2.1.a - 2.1.h

	JL	WZ	LO	GU	DR	DO	OP	LL	
α	1	0	5	0	3	0	16	3	28
β	2	3	2	0	9	0	7	6	29
γ	10	2	13	6	9	1	5	10	56
δ	10	22	11	18	12	13	11	8	105
ε	6	4	0	1	0	11	1	3	26
	29	31	31	25	33	25	40	30	244

TABLE 2.2.a

	JL	WZ	LO	GU	DR	DO	OP	LL	
A	9	9	28	10	30	4	38	26	154
B	11	7	1	11	3	15	2	1	51
C	7	2	1	1	0	4	0	0	15
D	2	10	1	3	0	1	0	1	18
E	0	3	0	0	0	1	0	2	6
	29	31	31	25	33	25	40	30	244

TABLE 2.2.b

	A	B	C	D	E	
α	28	0	0	0	0	28
β	29	0	0	0	0	29
γ	45	6	4	1	0	56
δ	49	36	7	10	3	105
ε	3	9	4	7	3	26
	154	51	15	16	6	244

TABLE 2.2.c

	α	β	γ	δ	ε	A	B	C	D	E	exact	asymp
1	+1	+1	+1	-1	-1	+1	-1	-1	-1	-1	66.646	66.646
2	+1	+1	+1	-1	-1	0	+1	+1	-1	-1	.854	1.980
3	+1	+1	+1	-1	-1	0	+1	-1	+1	-1	.513	1.189
4	+1	+1	+1	-1	-1	0	+1	-1	-1	+1	.180	.417
5	+1	-1	0	0	0	+1	-1	-1	-1	-1	.000	
6	+1	-1	0	0	0	0	+1	+1	-1	-1	.000	
7	+1	-1	0	0	0	0	+1	-1	+1	-1	.000	
8	+1	-1	0	0	0	0	+1	-1	-1	+1	.000	
9	+1	+1	-1	0	0	+1	-1	-1	-1	-1	4.682	12.404
10	+1	+1	-1	0	0	0	+1	+1	-1	-1	.467	
11	+1	+1	-1	0	0	0	+1	-1	+1	-1	.280	
12	+1	+1	-1	0	0	0	+1	-1	-1	+1	.098	
13	0	0	0	+1	-1	+1	-1	-1	-1	-1	11.046	10.743
14	0	0	0	+1	-1	0	+1	+1	-1	-1	8.070	3.244
15	0	0	0	+1	-1	0	+1	-1	+1	-1	2.675	1.527
16	0	0	0	+1	-1	0	+1	-1	-1	+1	4.263	2.174

TABLE 2.3

	pa	ps	cp	vv	bp	ch	ar	gp	sg	d6	kv	cr
opportunistic	0	0	0	0	0	0	0	0	0	1	1	0
progressive	1	1	1	0	0	0	1	0	0	1	0	1
left-wing	0	1	1	0	0	0	0	0	0	0	0	0
dogmatic	0	0	1	0	1	0	0	1	1	0	0	0
conservative	0	0	0	1	1	0	0	1	1	0	0	0
important	0	0	0	0	0	0	0	0	0	0	0	0
clear	0	1	1	1	1	0	0	1	1	0	0	0
homogeneous	0	1	1	1	1	0	0	1	1	1	0	1
sympathetic	1	1	0	0	0	0	0	0	0	1	0	1
intelligent	0	1	0	0	0	0	0	0	0	1	0	1
democratic	1	1	0	1	0	1	1	0	0	1	1	1
tolerant	1	1	0	1	0	1	1	0	0	1	1	1
negative	0	0	0	0	1	0	0	0	0	0	0	0
constructive	1	0	0	1	0	1	1	0	0	0	1	0
up to date	0	0	0	0	0	0	0	0	0	1	0	1
responsible	1	1	1	1	0	1	1	1	1	1	1	1
consistent	0	1	1	1	1	0	0	1	1	1	0	1

TABLE 5.1

PvdA	-.310	-.172
PSP	-.202	.516
CPN	.236	.213
VVD	.030	-.287
BP	1.415	.126
CHU	-.262	-.415
AR	-.315	-.318
GPV	.379	-.113
SGP	.379	-.113
D'66	-.611	.412
KVP	-.366	-.411
CR	-.375	.561

TABLE 5.2

opportunistic	-.323	.000
progressive	-.289	.343
left-wing	.011	.371
dogmatic	.498	.036
conservative	.455	-.123
important	.000	.000
clear	.364	.050
homogeneous	.352	.263
sympathetic	-.309	.419
intelligent	-.291	.561
democratic	-.498	-.036
tolerant	-.498	-.036
negative	.850	.116
constrcutive	-.231	-.466
up to data	-.326	.495
responsible	-.850	-.116
consistent	.352	.262

TABLE 5.3

JOINT PLOT
FACULTIES AND
PARTIES

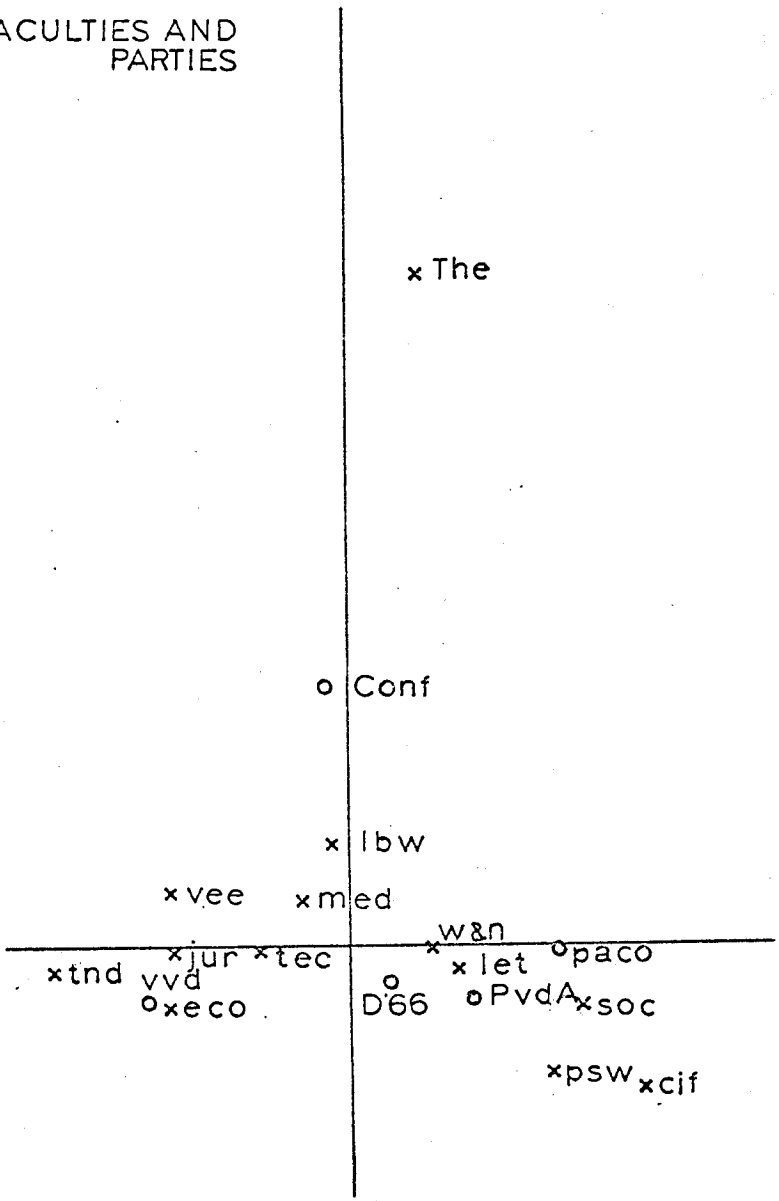


FIGURE 1.4.a

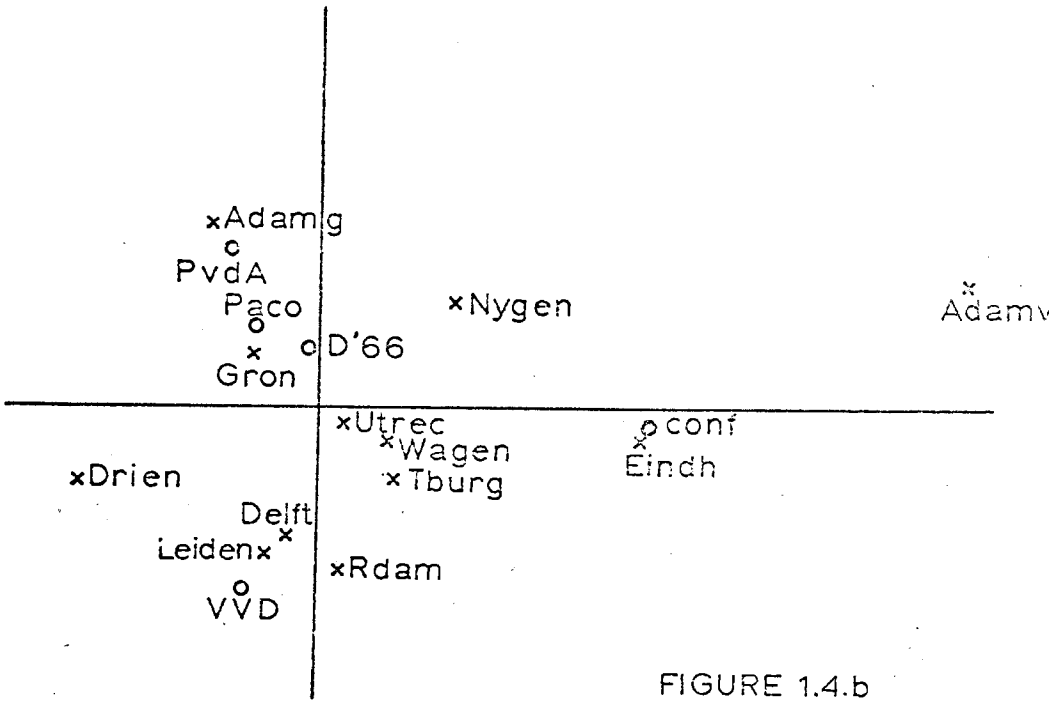

Adamg
PvdA
Paco
D'66
Gron
Nygen
Adamv
Utrec
Wagen
Tburg
conf
Eindh
Drien
Delft
Leiden
Rdam
VVD

FIGURE 1.4.b

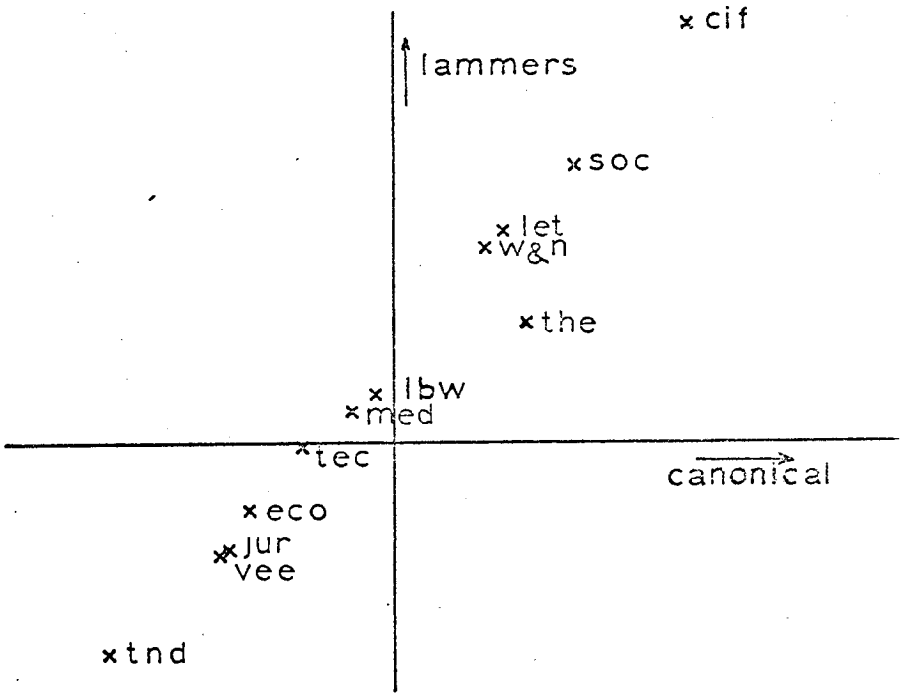


FIGURE 1.6

Lammers partition faculties

Source	X^2	dfr
Regression	105.546	12
Residual	72.109	36
Total	177.655	48

Lammers partition universities

Regression	61.656	11
Residual	141.272	33
Total	202.928	44

TABLE 1.7.

FIGURE 2.5

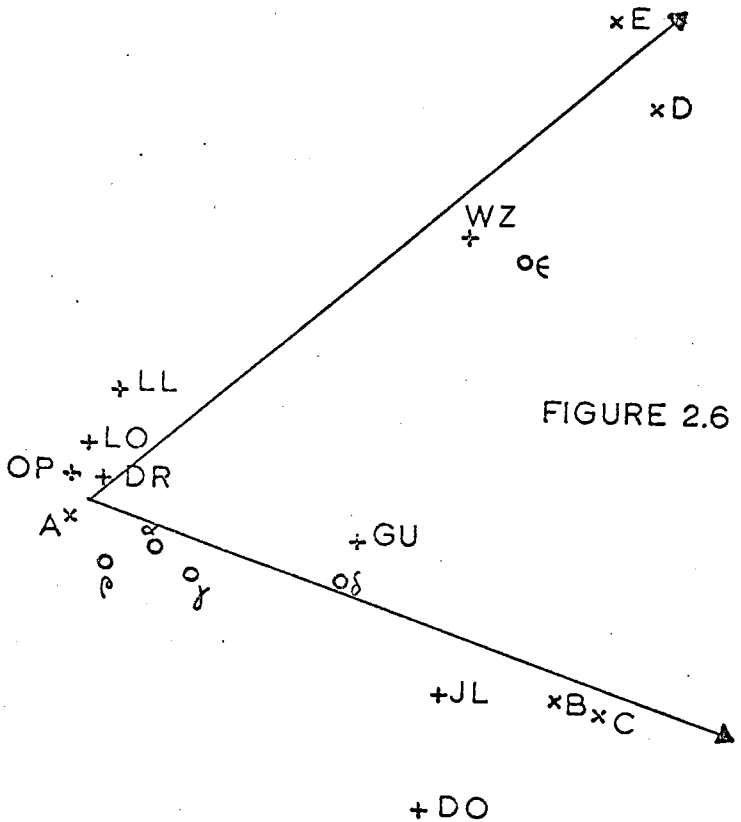


FIGURE 2.6

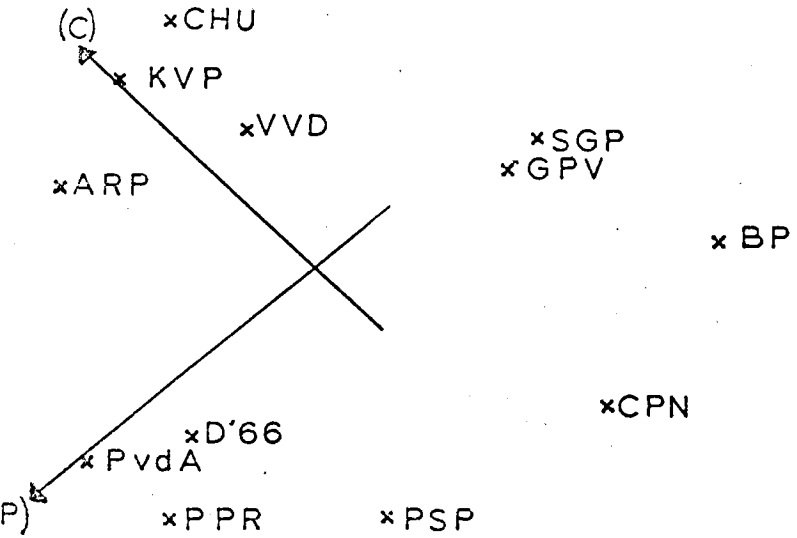


FIGURE 3.2
lowerhouse

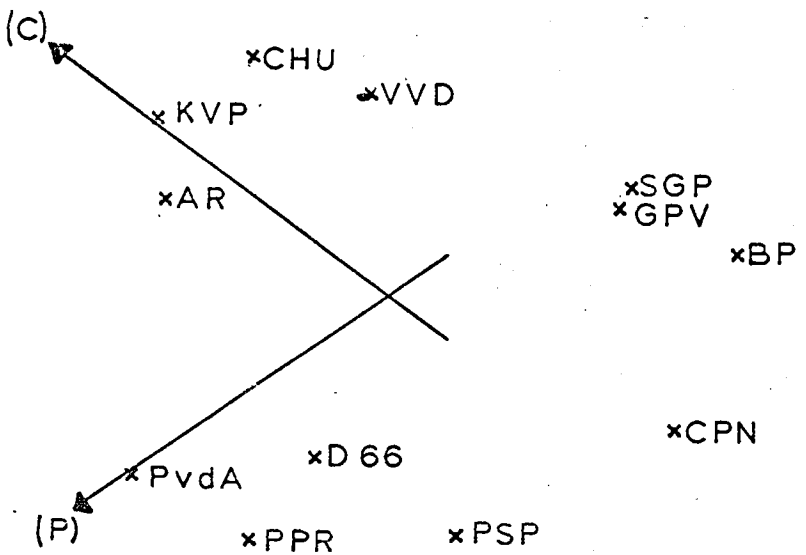
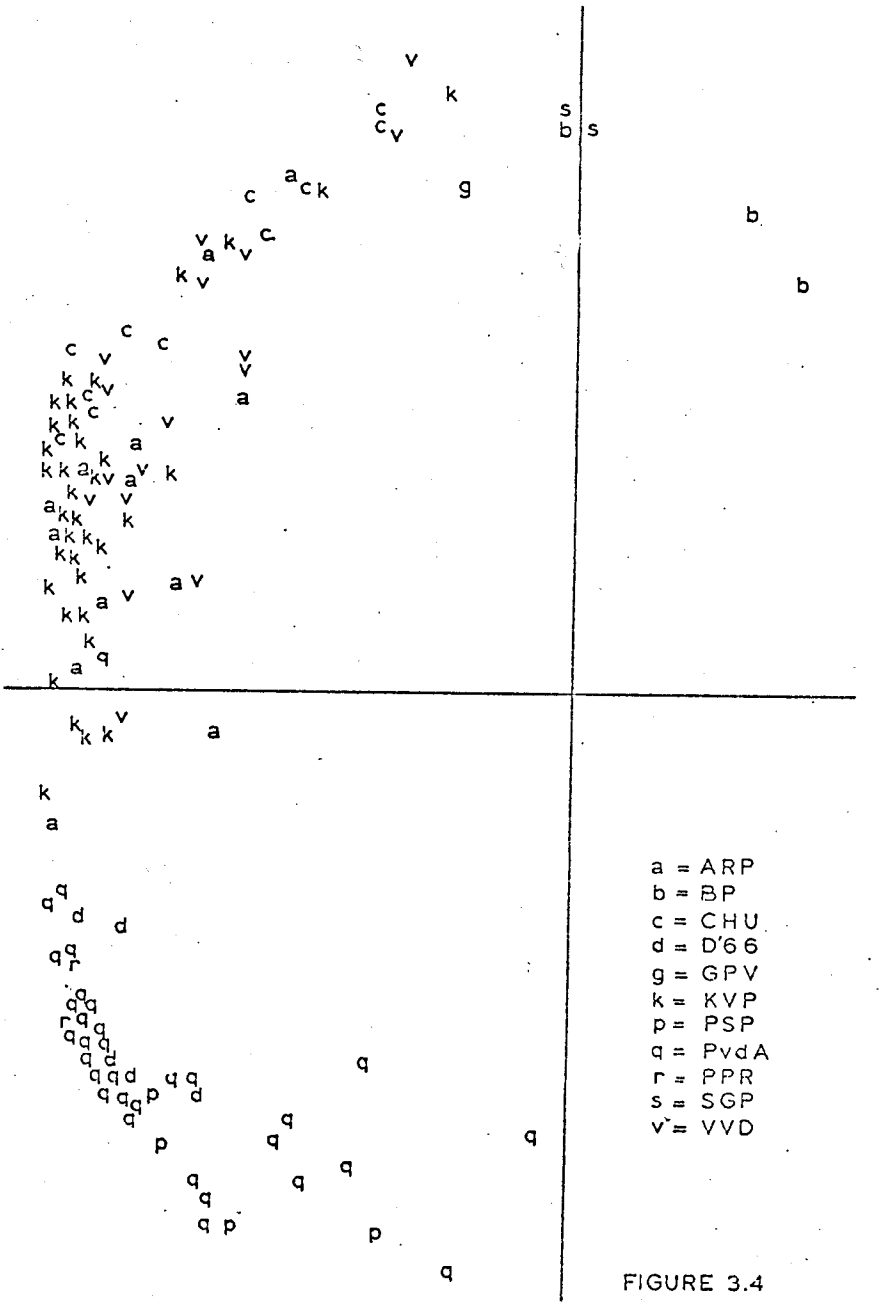


FIGURE 3.2
upperhouse



a = ARP
b = BP
c = CHU
d = D'66
g = GPV
k = KVP
p = PSP
q = PvdA
r = PPR
s = SGP
v = VVD

FIGURE 3.4

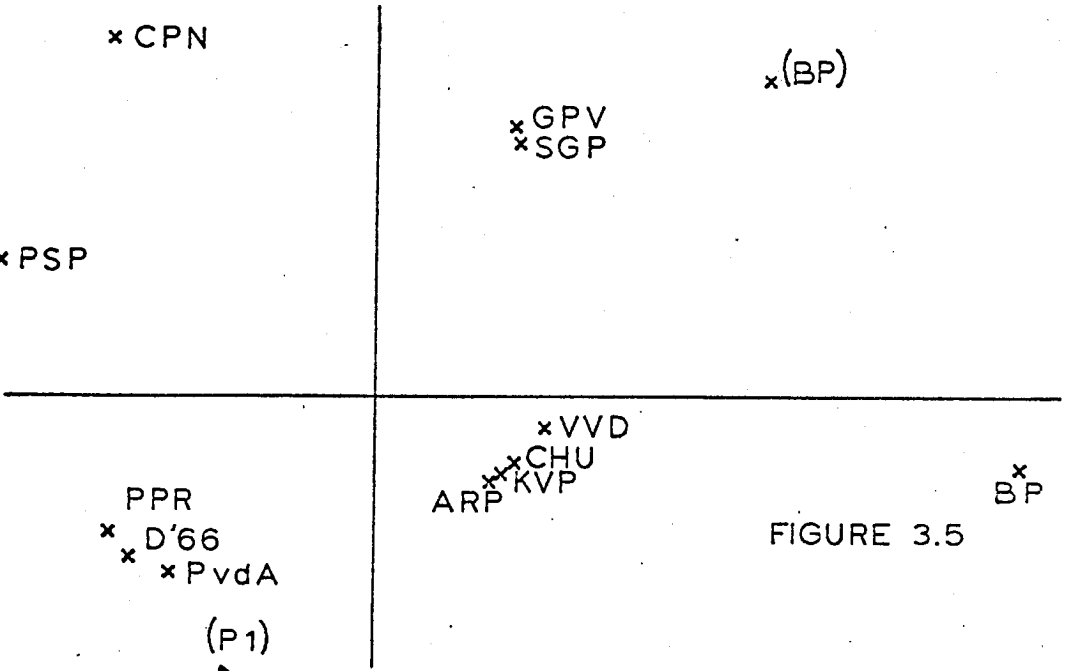


FIGURE 3.5

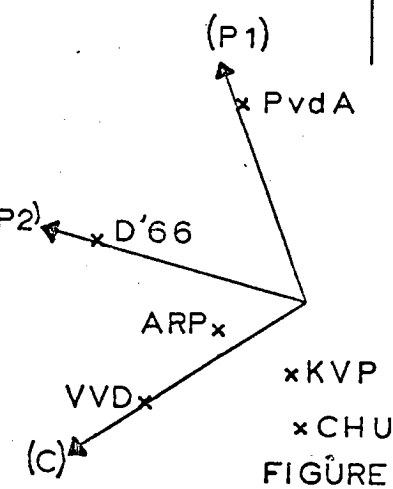


FIGURE 3.6

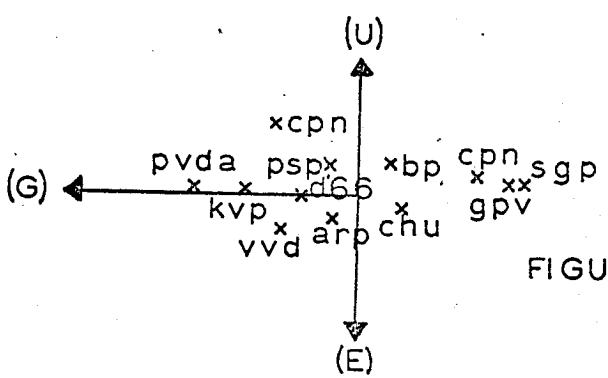


FIGURE 3.7

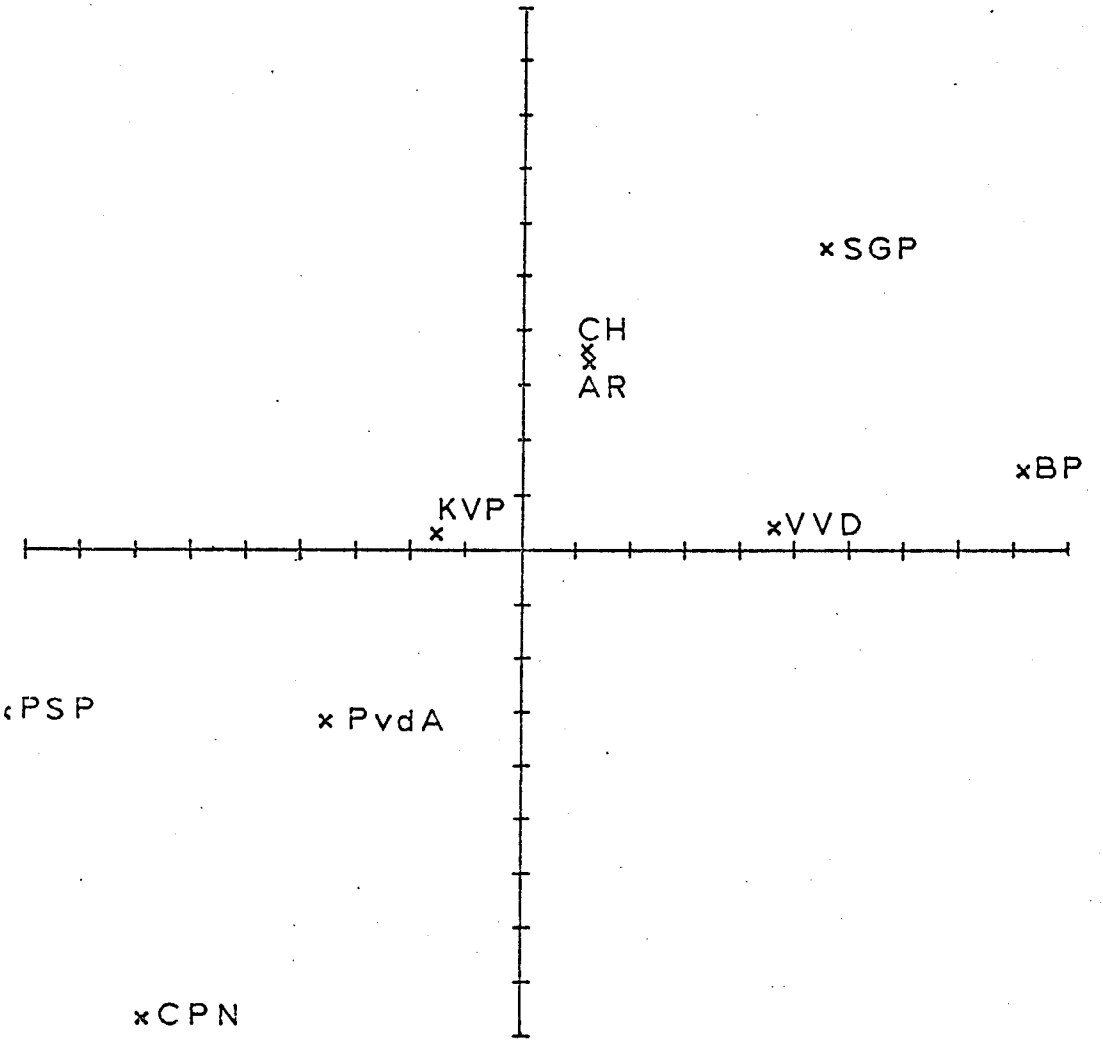


FIGURE 4.3

PvdA	11	38	90	86	40	54	66	34	76
FSP	33	11	87	83	31	60	66	49	75
BP	79	83	11	36	83	53	56	61	33
SGP	76	90	48	11	92	35	33	55	55
CPN	30	31	88	91	11	58	68	49	69
AR	62	80	73	45	94	11	28	44	54
CHU	71	85	65	54	94	28	11	46	42
KVP	39	76	83	71	81	33	47	11	54
VVD	75	83	49	53	96	44	30	54	11

TABLE 4.1.

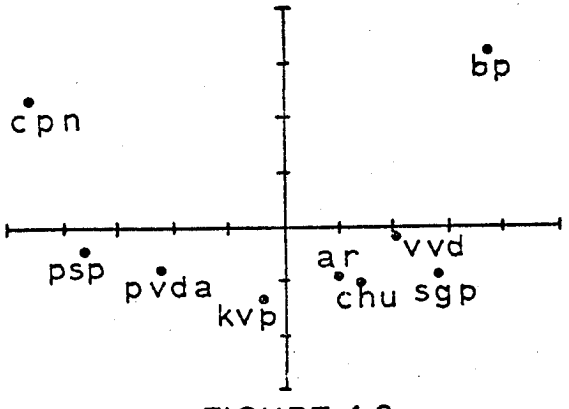


FIGURE 4.2

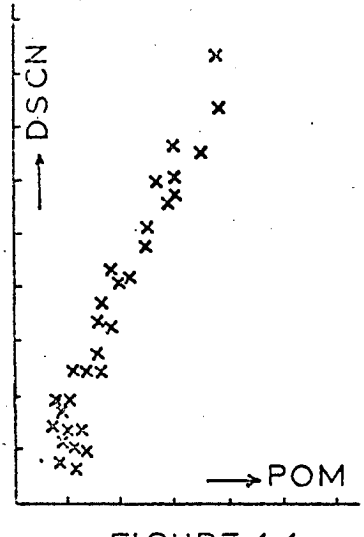


FIGURE 4.4

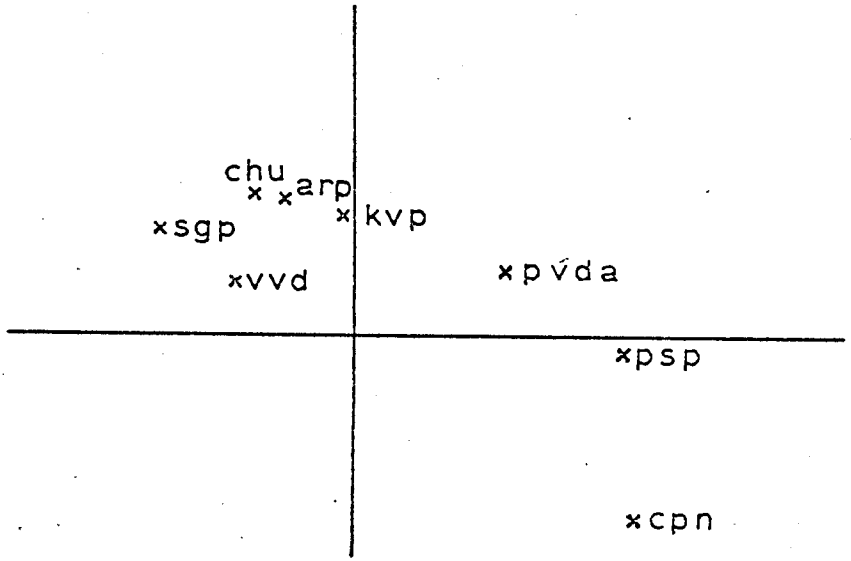


FIGURE 4.4

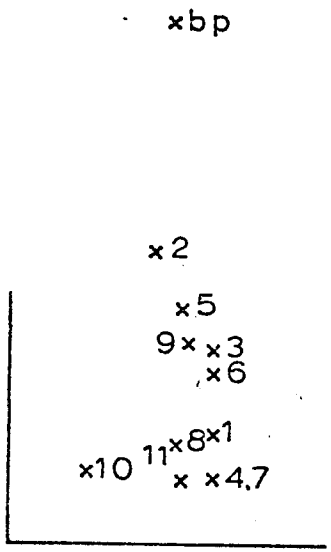


FIGURE 4.5

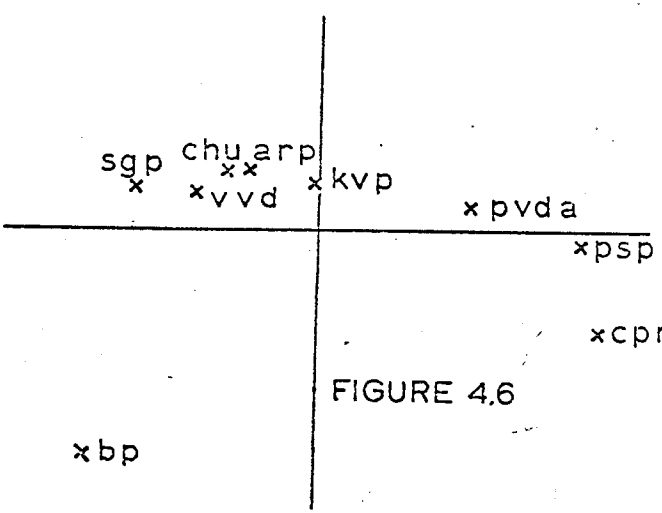


FIGURE 4.6

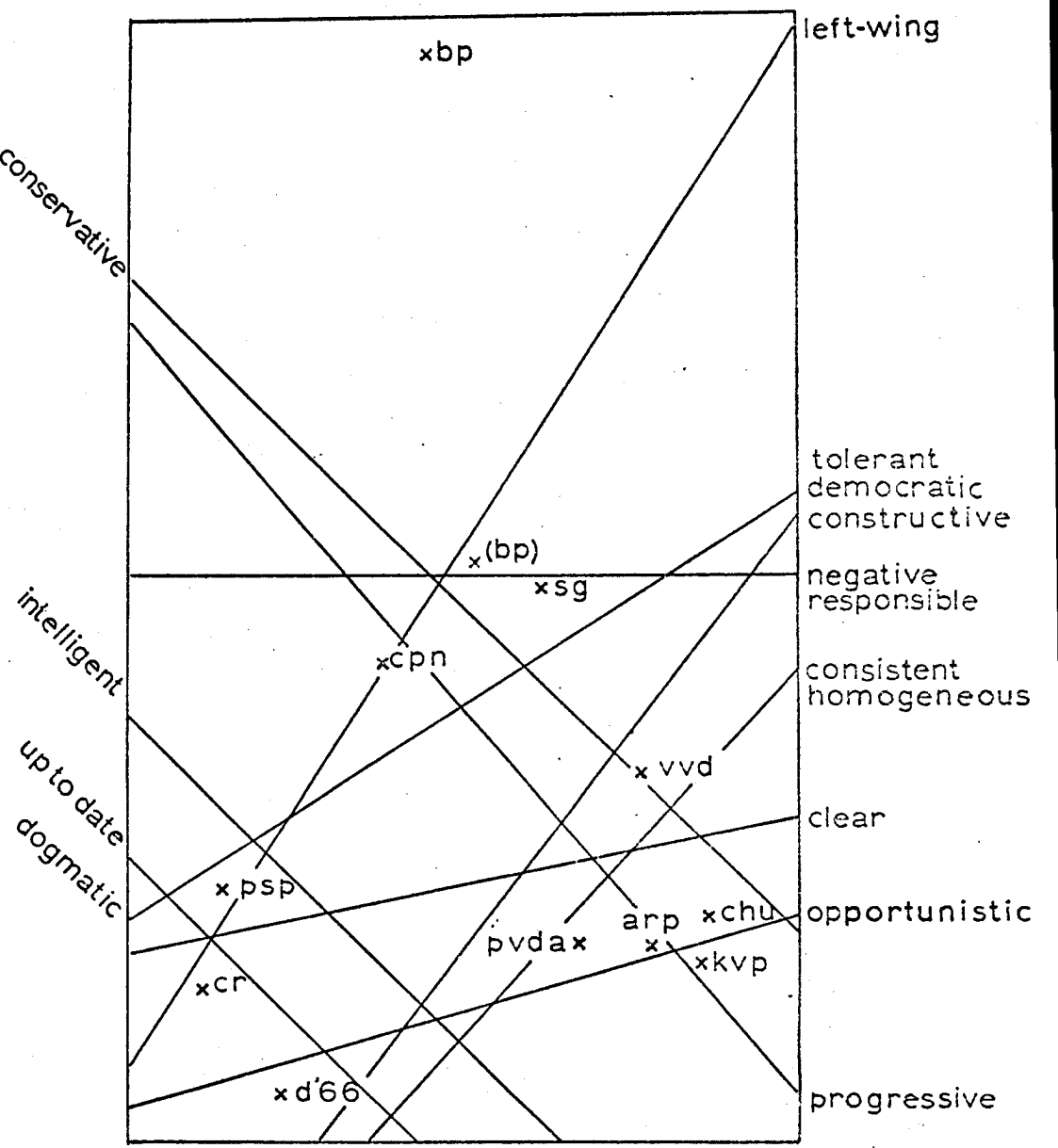

left-wing
×bp
conservative
tolerant
democratic
constructive
×(bp)
×sg
negative
responsible
intelligent
×cpn
consistent
homogeneous
×vvd
up to date
clear
dogmatic
×psp
opportunistic
arp ×chu
pvda×
×kvp
×cr
×d'66
progressive

TABLE 5.4

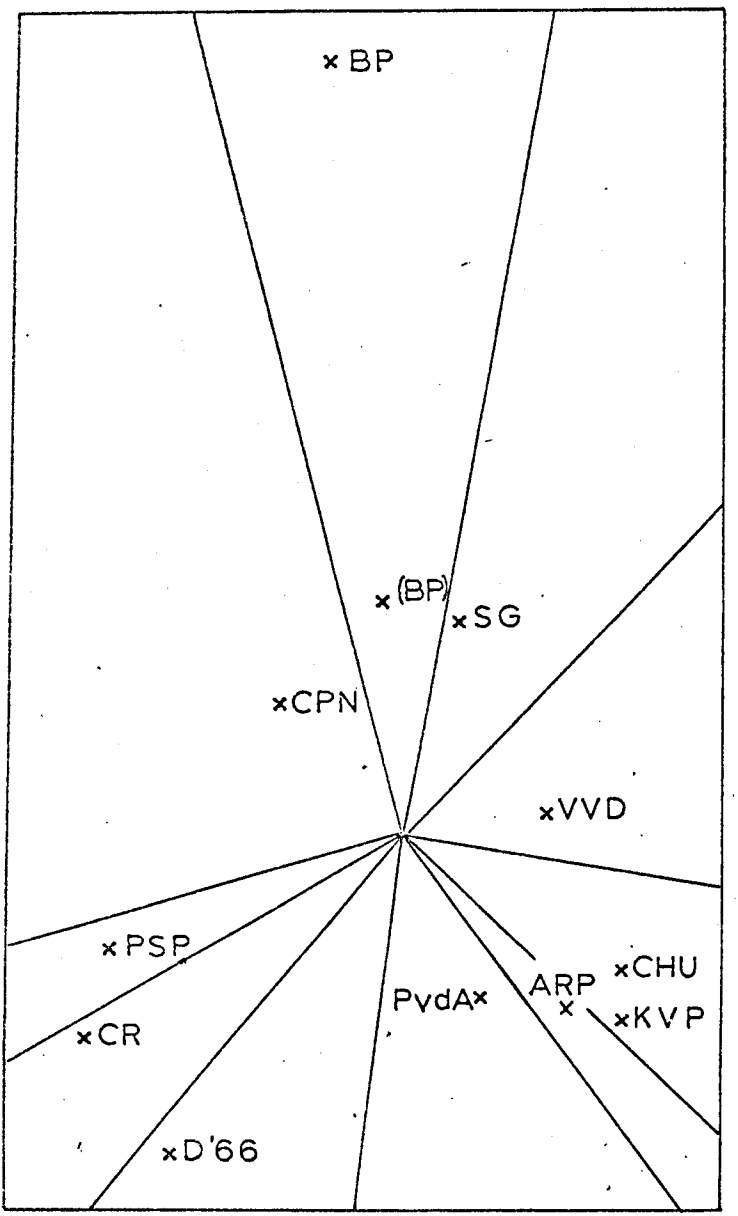


TABLE 5.5

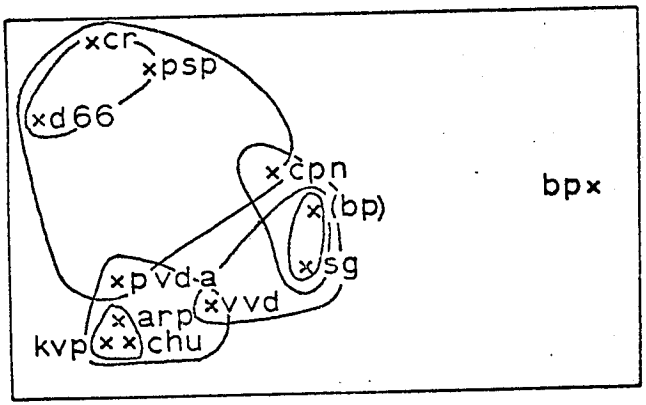


FIGURE 5.6

BP, SG	extremist right	right wing	
VVD	constructive right		
CHU, KVP, AR	middle		
PvdA	constructive left	left wing	
D'66, CR, PSP	intelligent left		
CPN	extremist left		

TABLE 5.7

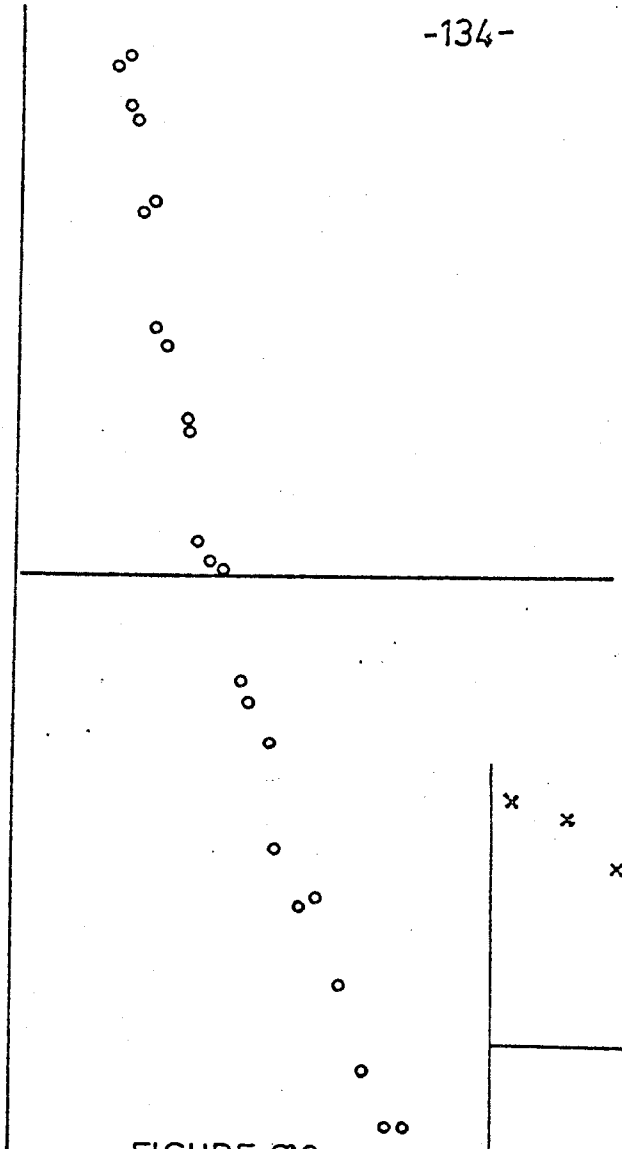


FIGURE 6.2

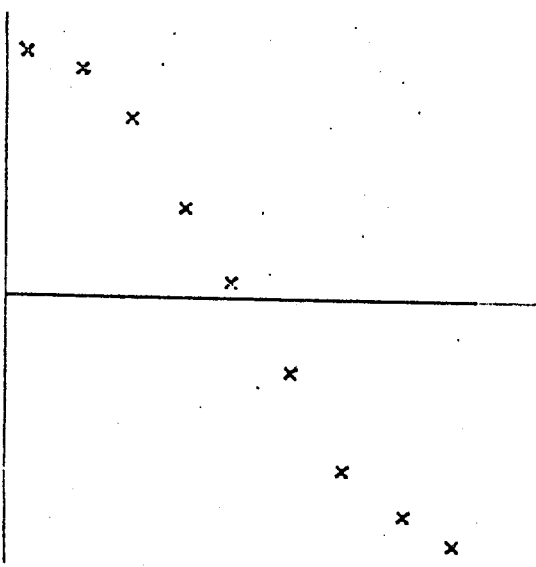


FIGURE 6.3

8.1. References to chapter 1.

J.P. Benzécri: Lois de probabilité sur un ensemble produit: les diverses notions d'indépendence et le critère d'entropie maximale. Mimeographed paper, Statistical Institute, Univ. Paris.

J. Berkson: Application of minimum logit χ^2 estimate to a problem of Grizzle with a notation on the problem of no interaction. Biometrics, 1968, 24, 75-95.

V.P. Bhapkar: Some tests for categorical data. Ann. Match. Statist., 1961, 32, 72-83.

-- A note on the equivalence of two test criteria for hypotheses in categorical data. J. Amer. Statist. Assoc., 1966, 61, 228-235.

-- Categorical data analogs of some multivariate tests. In: R.C. Bose, ed., Contributions to statistics and pr probability: essays in memory of S.N. Roy. Chapell Hill, Univ. North Carolina Press, 1969.

M.W. Birch: Maximum likelihood in three-way contingency tables. J. Royal Statist. Soc., B, 1963, 25, 220-233.

-- The detection of partial association, I: the 2 x 2 case. J. Royal Statist. Soc., B, 1964, 26, 313-324.

-- The detection of partial association, II: the general case. J. Royal Statist. Soc., B, 1965, 27, 111-123.

M. Black: Problems of analysis. London, Routledge & Kegan Paul, 1954.

R.D. Bock: Estimating multinomial response relations. In: R.C. Bose, ed., Contributions to statistics and probability: essays in memory of S.N.Roy. Chapell Hill, Univ. North Carolina Press, 1969.

C. Burt: The factorial analysis of qualitative data. British J. Statist. Psychol., 1950, 3, 166-185.

H. Caussinus: Contribution à l'analyse statistique des tableaux de corrélation. Ann. de la faculté des sciences de l'Université de Toulouse, 1965, 29, 77-182.

C.H. Coombs: A theory of data. New York, Wiley, 1964.

-- General mathematical psychology: chapter scaling and data theory. Mimeographed paper, University of Michigan, 1966.

J. de Leeuw Some contributions to the analysis of categorical data. Dep. of Data Theory, Univ. Leiden, report RN004-69, 1969.

-- The abstract structure of scaling theories. Mimeographed paper, Univ. Leiden, 1970.

-- Multinormal analysis. Mimeographed lecture notes, Univ. Leiden, 1972.

I.J. Good: Maximum entropy for hypotheses formulation, especially for multidimensional contingency tables. Ann. Math. Statist., 1963, 34, 911-934.

L.A. Goodman: The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries. J. Amer. Statist. Assoc., 1968, 63, 1091-1131.

-- The multivariate analysis of qualitative data: interactions among multiple classifications. J. Amer. Statist. Assoc., 1970, 65, 226-256.

-- The analysis of multidimensional contingency tables: stepwise procedures and direct estimation methods for building models for multiple classifications. Technomet., 1971, 13, 33-61.

-- Some multiplicative models for the analysis of cross classified data. Proc. 6th Berkely Symp.Math. Statist. and Prob., Los Angeles, Univ. Calif. Press, 1972, p 649-696.

J.E. Grizzle, C.F. Starmer, and G.G. Koch: Analysis of categorical data by linear models. Biometrics, 1969,25, 489-504.

J.P. Guilford: Psychometric methods. New York, McGrawHill, 1954.

L. Guttman: 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, Social Science Research Council, 1941.

-- A basis for scaling qualitative data. Amer. Sociol. Rev., 1944, 9, 362-369.

-- The nonmetric breakthrough for the behavioural sciences. Proc. 2nd national conf. on data processing, Information processing assoc. of Israel, 1966, pp 495-510.

-- The development of nonmetric space analysis: a letter to professor John Ross. Multivariate Behavioural Research, 1967, 2, 71-82.

M.G. Kendall & A. Stuart: The advanced theory of statistics. Volume III, London, Griffin, 1966.

D.H. Krantz: A survey of measurement theory. Michigan Math. Psychol. Program, report MMPP 67-4,1967.

H.H. Ku, R.N. Varner, and S. Kullback: On the analysis of multidimensional contingency tables. J. Amer. Statist. Assoc., 1971, 66, 55-64.

S. Kullback, M. Kupperman, and H.H. Ku: Tests for contingency tables and Markov chains. Technomet., 1962, 4, 573-608.

H.O. Lancester: Complex contingency tables treated by the partition of χ^2 J. Roy. Statist. Soc., B, 1951, 13, 242-249.

H.O. Lancester: The chi-squared distribution.
New York, Wiley, 1969.

D.V. Lindley: The baysian analysis of contingency tables.
Ann. Math. Statist., 1964, 35, 1622-1643.

J. Lingoes: The multivariate analysis of qualitative data.
Multivariate Behavioural Research, 1968, 3, 61-94.

P.E. Meehl: Clinical versus statistical prediction
Minneanapolis, Univ. Minnesota Press, 1954.

S.K. Mitra: Statistical analysis of categorical data.
Institute of statistics mimeograph series no 142,
Univ. North Carolina, 1956.

R.L. Plackett: Multidimensional contingency tables: a survey of models and methods.
Bull. Int. Statist. Inst., 1969, 43, 133-142.

G. Rasch: An informal report on the theory of objectivity in comparisons.
In: Proc. Nuffic international summer session on Psychological measurement theory.
Den Haag, 1966.

S.N. Roy and V.P. Bhapkar: Some nonparametric analogs of normal ANOVA, MANOVA, and of studies in normal association.
In: I. Olkin, ed., Contributions to probability and statistics: essays in honour of Harold Hotelling.
Stanford, Stanford Univ. Press, 1960.

S.N. Roy and S.K. Mitra: An introduction to some nonparametric generalizations of analysis of variance and multivariate analysis.
Biometrika, 1956, 43, 361-376.

R.S. Rudner: Philosophy of social science.
Englewood Cliffs, Prentice-Hall, 1966.

W. Schaafsma: The Neyman-Pearson theory for testing statistical hypotheses.
Statistica Neerlandica, 1971, 25, 1-27.

C.L. Stevenson: Persuasive definitions.
Mind, 1938, 47, 331-343.

W.S. Torgerson: Theory and methods of scaling.
New York, Wiley, 1958.

-- Multidimensional scaling of similarity.
Psychomet., 1965, 30, 379-393.

J.W. Tukey: The future of data analysis.
Ann. Math. Statist., 1962, 33, 1-79.

P. Winch: The idea of a social science.
London, Routledge & Kegan Paul, 1958.

L. Wittgenstein: Philosophical Investigations.
Oxford, Blackwell, 1953.

-- Remarks on the foundations of mathematics.
Oxford, Blackwell, 1956.

-- Lectures and conversations on aesthetics, psychology, and religious belief.
Oxford, Blackwell, 1966.

8.2 References to chapter 2.

V.P. Bhapkar & G.G. Koch: On the hypothesis of 'no interaction' in contingency tables. Biometrics 1968, 24, 567-594.

C. Burt: The factorial analysis of qualitative data. Brit. J. Psychol., Statist. Sect., 1950, 3, 166-185.

J. de Leeuw: Canonical discriminant analysis of relational data. Research Note RN 007-68, Dep. of Data theory, University of Leiden, 1968.

U.G. Foa: New developments in facet design and analysis. Psychol. Review, 1965, 72, 262-275.

L. Guttman: 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, Social Science Research Council, 1941, p 319-348.

-- An introduction to facet design and analysis. In Proc. 15th Int. Congress of Psychol. Brussels. Amsterdam, North Holland, 1959, p 130-132,

J. Lingoes: Multivariate analysis of contingencies:An IBM 7090 program for analyzing metric/nonmetric or linear/nonlinear data. Comp. Rep. 1963, 2, 1-24, University of Michigan.

-- Simultaneous linear regressions: An IBM 7090 program for analyzing metric/nonmetric or linear/nonlinear data. Behav. Science, 1964, 9, 87-88-

-- The multivariate analysis of qualitative data. Multivariate Behavioural Research, 1968, 3, 61-94.

S.N. Roy & S.K. Mitra: An introduction to some nonparametric generalizations of analysis of variance and multivariate analysis. Biometrika, 1956, 43, 361-376.

M. Wish: A facet-theoretic approach for Morse code and related signals. Michigan Mathematical Psychology Program, report MMPP 65-6.

G.U. Yule: An introduction to the theory of statistics. London, Griffin, 1910.

8.3 References to chapter 3.

T.W. Anderson: Some scaling methods and estimation procedures in the latent class model.
In: U. Grenander (ed): Probability and statistics. The Harald Cramer volume.
New York, Wiley, 1959.

R. Bargmann: Exploratory techniques involving artificial variables.
In: P.R. Krishnaiah(ed): Multivariate analysis, II.
New York, Academic Press, 1969.

M.S. Bartlett: Factor analysis in psychology as a statistician sees it.
Uppsala Symp. on Psychol. factor analysis.
Uppsala, Almqvist & Wicksell, 1953.

R.D. Bock: Methods and applications of optimal scaling.
Rep. no. 25, Psychometrics Laboratory, Univ. North Caroline, 1960.

R.A. Bradley, S.K. Katti, and I.J. Coons:
Optimal scaling for ordered categories.
Psychometrika, 27, 1962, 355-374.

C. Burt: The factorial analysis of qualitative data.
Brit. J. Psychol., Statist. Sect., 1950, 3, 166-185.

-- Scale analysis and factor analysis.
Brit. J. Statist. Psychol., 1953, 6, 5-24.

J.D. Carroll: Nonparametric multidimensional analysis of paired comparison data.
Mimeographed paper, Bell Labs, Murrey Hill, 1967.

-- Generalization of canonical correlation analysis to three or more sets of variables.
Proc. 76th Ann. Convention APA, 1968, 227-228.

-- Equations and tables for a generalization of canonical analysis to three or more sets of variables.
Mimeographed paper, Bell Labs, Murray Hill, 1968.

-- Polynomial factor analysis.
Proc. 77th Ann. Convention APA, 1969, 103-104.

-- Polynomial factor analysis.
Mimeographed paper, Bell Labs, Murray Hill, 1969.

B. Cordier: L'analyse factorielle des correspondences.
Faculté des sciences de l'université de Rennes, Mimeographed 1963.

J. de Leeuw: The linear nonmetric model.
Dep. of Data Theory, Univ. Leiden, Rep. RN003-69, 1969(a).

-- Some contributions to the analysis of categorical data.
Dep. of Data Theory, Univ. Leiden, Rep. RN004-69,1969(b).

-- The positive orthant method for nonmetric multidimensional scaling.
Dep. of Data Theory, Univ. Leiden, Rep. RN001-70, 1970(a).

J. de Leeuw: The Euclidean distance model.
Dep. of Data Theory, Univ. Leiden, Rep. RN002-70,1970 (b).

-- The abstract structure of scaling theories.
Mimeographed paper, Dep. of Data Theory, Univ.Leiden 1971(a)

-- Canonical analysis of contingency tables.
Dep. of Data Theory, Univ. Leiden, Rep. RN002-71, 1971(b).

J. Doesborgh: Some algebraic aspects of nonmetric multivariate analysis.
Mimeographed paper, Dep. of Data Theory, Leiden Univ. 1971.

H.A. Edgerton and L.E. Kolbe:
The method of minimum variation for the combination of criteria.
Psychometrika, 1, 1936, 183-187.

R.A. Fisher: The precision of discriminant functions.
Ann. Eugenics, London, 10, 1940, 422-429.

S. Henrysson and P. Thunberg:
Tetrachoric or phi-coefficients in factor analysis.
Dep. of Psychol., Univ. Uppsala, Rep. 27, 1968.

H.O. Hirschfeld: A connection between correlation and contingency.
Proc. Camb. Phil. Soc. 1935, 31, 520-524.

P. Horst: Obtaining a composite measure from a number of different measures of the same attribute.
Psychometrika, 1, 1936, 53-56.

-- Relations among m sets of measures.
Psychometrika, 26, 1961, 129-149(a).

-- Generalized canonical correlations and their applications to experimental data.
J. Clin. Psychol. (Monograph Suppl.), 1961, 14, 331-347(b)

-- Factor analysis of data matrices.
New York, Holt, Rinehart, & Winston, 1965.

H. Hotelling: Analysis of a complex of statistical variables into principal components.
J. Educ. Psychol., 1933, 24, 417-441, 498-520.

-- Relations between two sets of variables.
Biometrika, 1936, 28, 321-377.

R. Gnanadesikan and M.B. Wilk:
Data analysis methods in multivarate statistical analysis
In: P.R. Krishnaiah(ed): Multivariate analysis, II.
New York, Academic Press, 1969.

L. Guttman: Quantifying a class of attributes: a theory and method of scale construction.
In: P. Horst (ed): The prediction of personal adjustment.
New York, Social Science research council, 1941.

L. Guttman: The relation of scalogram analysis to other techniques. In Stouffer and others: Measurement and prediction. Princeton, PUP, 1950(a).

-- The principal components of scale analysis. In: S.A. Stouffer and others: Measurement and prediction. Princeton, PUP, 1950(b).

-- A note on Sir Cyril Burts 'Factorial analysis of qualitative data'. Brit. J. Statist. Psychol., 6, 1953, 1-4.

-- The principal components of scalable attitudes. In: P.F. Lazersfeld (ed): Mathematical thinking in the social sciences. Glencoe, The Free Press, 1954.

-- An additive metric from all the principal components of a perfect scale. Brit. J. Statist. Psychol., 1955, 8, 17-24.

-- Introduction to facet design and analysis. Pr c. 15th Int. congress of Psychol, Brussels, 1957. Amsterdam, North-Holland, 1959(a).

-- Metricizing rank-ordered or unordered data for a linear factor analysis. Sankhya, 1959, 21, 257-268(b).

-- The development of nonmetric space analysis: a letter to professor John Ross. Multivariate Behavioural Research, 1967, 2, 72-82.

J.O. Irwin: A note on the subdivision of X^2 into components. Biometrika, 36, 1949, 130-134.

J. Kettenring: Canonical analysis of several sets of variables. Biometrika, 58, 1971, 433-451.

J.B. Kruskal: Analysis of factorial experiments by estimating monotone transformations of the data. J. Roy. Statist. Soc., B, 27, 1965, 251-263.

H.O. Lancester: The derivation and partition of X^2 in certain discrete distributions. Biometrika, 36, 1949, 117-129.

-- Some properties of the bivariate normal distribution considered in the form of a contingency table. Biometrika, 44, 1957, 289-292.

-- The structure of bivariate distributions. Ann. Math. Statist., 1958, 29, 719-736.

-- The Helmert matrices. Amer. Math. Monthly, 1965, 72, 4-11.

-- The chi-squared distribution. New York, Wiley, 1969.

H.O. Lancester and M.A. Hamdan:
Estimation of the correlation coefficient in contingency tables with possibly nonmetrical characters.
Psychometrika, 29, 1964, 383-391.

P.F. Lazarsfeld:
The logical and mathematical foundation of latent structure analysis.
In: S.A. Stouffer and others:Measurement and prediction.
Princeton, PUP, 1950.

J. Lingoes:
Multivariate analysis of contingencies.
Comp. Rep. 1963, 2, 1-24, Univ. Michigan.

--
The multivariate analysis of qualitative data.
Multivariate Behavioural Research, 1968, 3, 61-94.

F.M. Lord:
The relationship between Guttman's principal components of scale analysis and other psychometric theory.
Psychometrika, 23, 1958, 291-296.

A. Lubin:
Linear and nonlinear discriminant functions.
Brit. J. Statist. Psychol., 1950, 3, 90-104.

K. Maung:
Measurement of association in a contingency table.
Ann. Eugen. London, 11, 1941, 189-223.

R.P. McDonald:
A note on the derivation of the general latent class model.
Psychometrika, 27, 1962, 203-206.

--
Difficulty factors and nonlinear factor analysis.
Brit. J. Math. Statist. Psychol., 18, 1965, 11-23.

--
Nonlinear factor analysis.
Psychometric monograph, no. 15, 1967.

--
A unified treatment of the weighting problem.
Psychometrika, 33, 1968, 351-381.

--
The common factor analysis of multicategory data.
Brit. J. Math. Statist. Psychol., 22, 1969, 165-175.

F. Mosteller:
A theory of scalogram analysis using noncumulative types of items.
Rep. no. 9. Laboratory of social relations, Harvard Univ. 1949.

M. Okamoto and M. Kanazawa:
Minimization of eigenvalues of a matrix and optimality of principal components.
Ann. Math. Statist., 39, 1968, 859-863.

K. Pearson:
On lines and planes of closest fit to systems of points in space.
Phil.Mag. 2, 1901, 559-572.

C.R. Rao:
The use and interpretation of principal components analysis in applied research.
Sankhya, A. 1965, 26, 329-358.

R.G.D. Steel: Minimum generalized variance for a set of linear functions. Ann. Math. Statist., 1951, 22, 456-460.

W.S. Torgerson: Theory and methods of scaling. New Yor, Wiley, 1958.

J.P. van de Geer: Matching k sets of configurations. Dep. of Data Theory, Univ. Leiden, Rep. RN005-68, 1968.

S.S. Wilks: Weighting systems for linear functions of correlated variables when there is no independent variable. Psychometrika, 3, 1938, 23-40.

E.J. Williams: Use of scores for the analysis of association in contingency tables. Biometrika 39, 1952, 275-289.

F. Yates: The analysis of contingency tables with groupings based on quantitative characters. Biometrika, 35, 1948, 176-181, 424.

8.4 References to chapter 4.

G. Bechtel: Individual differences in the linear multidimensional scaling of choice.
Mimeo, Oregon Research Institute, 1969.

J.P. Benzécri: Sur l'analyse des préférences.
Mimeo, Institute de statistique de l'Université de Paris.

J.D. Carroll and J.J. Chang:
Non-parametric multidimensional analysis of paired comparison data.
Mimeo, Bell Labs, Murray Hill, 1964.

-- -- Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition.
Psychometrika, 35, 1970, 283-319.

J.D. Carroll and M. Wish: Multidimensional scaling of individual differences in perception and judgment.
Mimeao, Bell Labs, Murray Hill, 1970.

H.A. Daniels: The relation between measures of correlation in the universe of sampling permutations.
Biometrika, 1944, 33, 129-

L.F.W. de Klerk, J. de Leeuw, and S. Oppe:
Functional learning.
Psychol. Institute, Univ. Leiden, Rep.E020-68, 1968.

-- -- --
Functional learning.
Psychol. Institute, Univ. Leiden. Rep. E024-70,1970.

J. de Leeuw: Canonical analysis of relational data.
Dep. of Data Theory, Univ. Leiden, Rep. RN 007-68,1968(a).

-- Nonmetric multidimensional scaling.
Dep. of Data Theory, Univ. Leiden, Rep. RN010-68, 1968(b).

-- The positive orthant method for nonmetric multidimensional scaling.
Dep. of Data Theory, Univ. Leiden, Rep. RN001-70,1970(a).

-- The Euclidean distance model.
Dep. of Data Theory, Univ. Leiden, Rep.RN002-70, 1970(b).

L. Guttman: An approach for quantifying paired comparisons and rank order.
Ann. Math. Statist., 17, 1946, 144-163.

R.A. Harshman: Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multi-model factor analysis.
Working papers in phonetics 16, UCLA, 1970.

-- PARAFAC 2: Mathematical and technical notes.
In: Working papers in phonetics, 22, UCLA, 1972.

C. Hayashi: Multidimensional quantifications of the data obtained by the method of paired comparison.
Ann. Inst. Statist. Match 16, 1964, 231-245; 19, 1967, 363-365; 20, 1968, 167.

R. Jennrich: A generalization of the multidimensional scaling model of Carroll and Chang.
In: Working papers in phonetics, 22, UCLA, 1972.

M.G. Kendall: Rank correlation methods.
London, Griffin, 1962

W. Meredith: Notes on factorial invariance.
Psychometrika, 29, 1964, 177-184.

G. Rasch: Simultaneous factor analysis in several populations.
In: Uppsala Symp. on Psychol. factor analysis.
Stockholm, Almqvist & Wicksell, 1953.

P. Slater: The analysis of personal preferences.
Brit. J. Stat. Psychol., 13, 1960, 119-135.

-- Review of N. Frederiksen and H. Gulliksen(eds): Contributions to mathematical Psychology.
Brit. J. Staties. Psychol., 20, 1967, 116-120.

L.J.Th. Van der Kamp and L.C.W. Pols: Perceptual analysis from confusions between vowels.
Acta Psychologica, 35, 1971, 264-278.

W. Van der Kloot: Individual differences in cognitive structure of personality traits.
Unpublished Masters thesis, Univ. Leiden, 1969.

M. Wish: An INDSCAL analysis of the Miller-Nicely consonant confusion data.
Mimea, Bell Labs, Murray Hill, 1970.

M. Wish and J.D. Carroll: Multidimensional scaling with differential weighting of dimensions.
Mimea, Bell Labs, Murray Hill, 1971.

8.5 References to chapter 5

M.S. Bartlett: Multivariate analysis.
J. Roy. Statist.Soc., Suppl., 9, 1947, 176-197.

R.A. Bradley, S.K. Katti, and I.J. Coons:
Optimal scaling for ordered categories.
Psychometrika, 27, 1962, 355-374.

J.D. Carroll: Categorical conjoint measurement.
Mimea, Bell Labs, Murray Hill, 1968.

J. de Leeuw: Nonmetric discriminant analysis.
Dep. of Data Theory, Univ. Leiden, rep. RN006-68

-- The linear nonmetric model.
Dep. of Data Theory, Univ. Leiden, Rep. RN003-69.

J. Doesborgh: Some algebraic aspects of nonmetric multivariate analysis.
Mimeaographed paper, Dep. of Data Theory, Leiden Univ, 1971.

R.A. Fischer: Statistical methods for research workers.
Edinburgh, Oliver & Boyd, 1941.

L. Guttman: An introduction to facet design and analysis.
In: Proc. 15th Int. Congress of Psychol.Brussels.
Amsterdam, North Holland, 1959(a).

-- Metricizing rank-ordered or unordered data for a linear factor analysis.
Sankhya, 21, 1959, 257-268(b).

P.O. Johnson: The quantification of qualitative data in discriminent analysis.
J. Amer. Statist. Assoc., 45, 1950, 65-76.

J.B. Kruskal: Analysis of factorial experiments by estimating monotone transformations of the data.
J. Roy. Statist. Soc. , B, 27, 1965, 251-263.

J. Lingoes: Multivariate analysis of contingencies.
Comp. Rep. 1963, 2, 1-24, Univ. Michigan.

8.6 References to chapter 6

T.W. Anderson: Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices.
In: R.C. Bose (ed): Contributions to statistics and probability: essays in memory of S.N. Roy.
Chapel Hill, Univ. North Carolina Press, 1969.

--- Statistical inference for covariance matrices with linear structure.
In: P.R. Krishnaiah (ed): Multivariate analysis II.
New York, Academic Press, 1969.

R. Bargmann: A study of dependence and independence in normal multivariate analysis.
Univ. North Carolina, Mimea series, No. 186, 1957.

R.D. Bock and R. Bargmann: Analysis of covariance structures.
Psychometrika 31, 1966, 507-534.

M.W. Browne: Fitting the factor analysis model.
Psychometrika, 34, 1969, 375-394.

M.R.B. Clarke: A rapidly convergent method for maximum likelihood factor analysis.
Brit. J. Math. Statist. Psychol., 23, 1970, 43-52.

J. de Leeuw: Factor analysis by direct and weighted least squares.
Mimeographed paper, Dep. Data Theory, Univ.Leiden, 1972.

G. Derflinger: Neue Iterationsverfahren in der Faktorenanalyse.
Biometrische Zeitschrift, 10, 1968, 58-75.

J.L. Doob: The limiting distribution of certain statistics.
Ann. Math. Statist., 6, 1935.

W.D. Fisher: On grouping for maximum homogeneity.
J. Amer. Statst. Assoc., 53, 1958, 789-798.

H.P. Friedman and J. Rubin: Some invariant criteria for grouping data.
J. Amer. Statist. Assoc., 62, 1967, 1159-1178.

B.F. Green: A note on the calculation of weights for maximim battery reliability.
Psychometrika, 15, 1950, 57-61.

L. Guttman: Image theory for the structure of quantitative variates.
Psychometrika, 18, 1953, 277-296.

-- The determinacy of factor score matrices with implications for five other basic problems of common factor theory.
Brit. J. Statist. Psychol., 8, 1955, 65-81.

-- The matrices of linear least square image analysis.
Brit. J. Math. Statist. Psychol., 13, 1960, 109-128.

C.W. Harris: Some Rao-Guttman relationships.
Psychometrika, 27, 1962, 247-264.

E.F. Heerman: The geometry of factorial indeterminacy. Psychometrika, 29, 1964, 371-381.

-- The algebra of factorial indeterminacy. Psychometrika, 31, 1966, 539-543.

W.G. Howe: Some contributions to factor analysis. Oak Ridge National Lab., Rep ORNL 1919, 1955

P.L. Hsu: Th limiting distribution of functions of sample means and applications to testing hypotheses. In: J. Neyman (ed): Proc 1st Berkeley Symp. Math.Statist. and Prob., Berkeley, Univ. California Press, 1949.

K.G. Jöreskog: Statistical estimation in factor analysis. Stockholm, Almqvist & Wicksell, 1963.

-- Efficient estimation in image factor analysis. Psychometrika, 34, 1969, 51-75.

-- A general method for the analysis of covariance structures. Biometrika, 57, 1970, 239-251

-- Estimation and testing of simplex models. Brit. J. Math. Statist. Psychol., 23, 1970, 121-145

-- Simultaneous factor analysis in several populations Psychometrika, 36, 1971, 409-420.

H.F. Kaiser: Image analysis. In: C.W. Harris (ed): Problems in measuring change. Madison, Univ. Wisconsin Press, 1963.

H.F. Kaiser and J.Caffrey: Alpha factor analysis. Psychometrika, 30, 1965, 1-14.

H. Kestelmann: The fundamental equation of factor analysis. Brit. J. Psychol., Statist. Sect., 5., 1952, 1-6.

D.N. Lawley: The estimation of factor loadings by the method of maximum likelihood. Proc. Roy. Soc. Edinb., A, 60, 1940, 64-82.

W. Ledermann: The orthogonal transformation of a factorial matrix into itself. Psychometrika, 3, 1938, 181-187.

H.B. Mann and A. Wald: On stochastic limit and order relationships. Ann. Math. Statist., 14, 1943, 217-226.

R.P. McDonald: A general approach to nonlinear factor analysis. Psychometrika, 27, 1962, 397-415.

-- Numerical methods for polynomial models in nonlinear factor analysis. Psychometrika, 32, 1967, 77-112.

-- Nonlinear factor analysis. Psychometric Monograph, No 15, 1968(a)

-- A unified treatment of the weighting problem. Psychometrika, 33, 1968, 351-381,(b)

R.P. McDonald: The common factor analysis of multicategory data. Brit. J. Math. Statist. Psychol., 22, 1969, 165-175(a).

-- A generalized common factor analysis based on residual covariance matrices of prescribed structure. Brit. J. Math. Statist. Psychol., 22, 1969, 149-163(b).

-- The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. Brit. J. Math. Statist. Psychol., 23, 1970, 1-21.

R.P. McDonald and E.J. Burr: A comparison of four methods of constructing factor scores. Psychometrika, 32, 1967, 381-401.

W. Meredith: Canonical correlations with fallible data. Psychometrika, 29, 1964, 55-65.

C.J. Mosier: On the reliability of a weighted composite. Psychometrika, 8, 1943, 381-401.

B.N. Mukherjee: Derivation of likelihood-ratio tests for Guttman quasi-simplex covariance structures. Psychometrika, 31, 1966, 97-122,

-- Invariance of the Guttman quasie-simplex linear model under selection. Brit. J. Math. Statist. Psychol., 22, 1969, 1-28.

-- Likelihood ratio tests of statistical hypotheses associated with patterned covariance matrices in psychology. Brit. J. Math. Statist. Psychol., 28, 1970, 89-120.

E.A. Peel: Prediction of a complex criterion and battery relability. Brit. J. Psychol, Statist.Sect., 1948, 1, 84-94.

B.R. Rao: Partial canonical correlations. Trabajos de Estadistica, 20, 1969, 211-219.

C.R. Rao: Estimation and tests of significance in factor analysis. Psychometrika, 20, 1955, 93-112.

-- Linear statistical inference and its applications. New York, Wiley, 1965.

M.R. Rao: Cluster analysis and mathematical programming. J. Amer. Statist. Assoc., 66, 1971, 622-626.

P.H. Schönemann: The minimum average correlation between equivalent sets of uncorrelated factors. Psychometrika, 36, 1971, 21-30.

P.H. Schönemann and M.M. Wang: Some new results on factor indeterminacy. Psychometrika, 37, 1972, 61-91.

J.N. Srivastava: On testing hypotheses regarding a class of covariance structures. Psychometrika, 31, 1966, 147-164.

G.H. Thomson: The definition and measurement of 'g'(general intelligence)
J. Educ. Psychol., 26, 1935, 241-262.

-- Weighting for battery reliability and prediction.
Brit. J. Psychol., 36, 1940, 357-366.

H.D. Vinod: Integer programming and the theory of grouping.
J. Amer. Statist. Assoc., 64, 1969, 506-519.

A. Wald: Asymptotically most powerful tests of statistical hypothes
Ann. Math. Statist., 12, 1941, 1-20.

-- Some examples of asymptotically most powerful tests.
Ann. Math. Statist., 12, 1941, 396-408.

-- Asymptotically shortest confidence intervals.
Ann. Math. Statist., 13, 1942, 127-137

-- Tests of statistical hyotheses concerning several paramete.
when the number of observations is large.
Trans. Amer. Math. Soc., 54, 1943, 426-482.

D.E. Wiley: Latent partition analysis.
Psychometrika, 32, 1967, 183-198.

J.H. Wilkinson: The algebraic eigenvalue problem.
Odford, Clarendon Press, 1965.

8.7 References to chapter 7

H. Daalder and J.G.Rusk: Party and legislator in a parliamentary system. Unpublished manuscript, Dep. Politicology, Univ. Leiden, 1971.

J. de Leeuw: Meerdimensionale analyse van politikologische gegevens. Hypothese, 13, 1968, 84-89.

-- The linear nonmetric model. Dep. of Data Theory, Univ.Leiden, Rep. RN003-69, 1969(a).

-- Nonmetric multidimensional scaling. Dep. of Data Theory, Univ. Leiden, Rep. RN010-68, 1969(b).

-- Contributions to the analysis of categorical data. Dep. of Data Theory, Univ. Leiden, Rep. RN004-69, 1969(c).

-- Canonical analysis of contingency tables. Dep. of Data Theory, Univ. Leiden, Rep. RN002-71, 1971(a).

-- Chi-square: a short review. Dep. of Data Theory, Univ. Leiden. Rep. RN003-71, 1971(b).

D.N.M. de Gruyter: The cognitive structure of Dutch political parties in 1966. Psychol. Institute, Leiden Univ., Rep E019-67, 1967.

J.P. Guilford: Psychometric methods. New York, McGrawHill, 1954.

C.L. Lammers: Is de universiteit een politieke leerschool? Universiteit en hogeschool, 15, 1969, 1-43.

E.E.Ch.I. Roskam: Metric analysis of ordinal data in psychology. Voorschoten, VAM, 1968.

R.N. Shepard and J.D. Carroll: Parametric representaion of nonlinear data structures. In: P.R. Krishnaiah(ed): Multivariate analysis I. New York, Academic press, 1966.

J. Stapel: Wie en wat staan waar tussen links en rechts? Een klein opinieonderzoek. Acta Politica, 4, 1968, 32-41.

W.S. Torgerson: Multidimensional scaling of similarity. Psychometrika, 30, 1965, 379-393.

J.P. Van de Geer: The use of non-Euclidean geometry in multidimensional scaling Mimeo, Institute for advanced Study in the Behavioral Sciences, Palo Alto, 1970.

W.A. van der Kloot: Individual differences in cognitive structure of personality traits. Unpublished Masters thesis, Univ. Leiden, 1969.