UCLA Statistics Series #1 August, 1988

Multivariate Analysis with Linearizable Regressions

Jan de Leeuw
Departments of Psychology and Mathematics

UCLA Classification: Theory and methods. Social Sciences-Education

This is the text of the Presidential Address, presented at the meeting of the Psychometric Society at UCLA, June 27-29, 1988.

Abstract

We study the class of multivariate distributions in which all bivariate regressions can be linearized by separate transformation of each of the variables. This class seems more realistic than the multivariate normal or the elliptical distributions, and at the same time its study allows us to combine the results from multivariate analysis with optimal scaling and classical multivariate analysis. In particular a two-stage procedure which first scales the variables optimally, and then fits a simultaneous equations model, is studied in detail and is shown to have some desirable properties.

Keywords

Multivariate analysis, optimal scaling, correspondence analysis, structural models, simultaneous equations, factor analysis, LISREL, transformation

Introduction

In order to be well-prepared for this event I have done a small content analysis of a few earlier presidential adresses. The former presidents talked about things dear to their hearts. They were not afraid of controversy, of historical remarks, of daring methodological generalizations, and of using a shameless numbers of self-references. In short, of all the things one tends to avoid in a paper that must still be approved by a number of critical referees. I will continue in this tradition, and indulge in the same sins.

The topic dear to my heart is multivariate analysis (MVA) with optimal scaling (OS). Forrest Young (1981) has reviewed our joint work in this field until 1980 in his presidential address. Since then an enormous work of additional work has been done by the Gifi team in Leiden. This now starts to appear in a more accessible form (Van Rijckevorsel and De Leeuw, 1988; Van der Burg, De Leeuw, & Verdegaal, in press). The book by Gifi will at long last appear near the end of 1988, and the DSWO press in Leiden has already published a long series of Gifi-related books in the past four years. Also French Analyse des Données has definitely come out of the closet, with books such as Greenacre (1984), Lebart, Morineau, & Warwick (1984), and with papers such as Tenenhaus and Young (1985) and Besse and Ramsay (1986). Correspondence analysis is now discussed with great regularity in the official statistical journals, and optimal scaling techniques have become quite popular (Breiman and Friedman, 1985, Koyak, 1987).

Nevertheless some aspects of the situation remain unsatisfactory. There is little integration of MVA with OS and classical MVA in the sense of Anderson (1958), and there is very little interaction with the active and interesting field of simultaneous equations or structural covariances modelling. In the meantime a lot of polemics is going on which centers on such esoteric topics as chance capitalization, and the distinction between exploratory and confirmatory, or between inferential and descriptive. Not

much of this has actually appeared in print, but you can hear it at every conference. Twenty years ago, when I started in MVA/OS (De Leeuw, 1968), you could hear it at every street corner. Some references, with a strong Dutch bias, are De Leeuw (1988b) and Molenaar (1988). I will summarize the arguments of my learned opponents. MVA/OS techniques are exploratory and descriptive, they say. Of course exploration and description are important in the *early* stages of scientific investigation, when there is not a great deal of prior knowledge. But *ultimately* we want to generalize from our particular data set, infer from the sample to the population, and we want to test explicit hypotheses. Doing science means sticking out your neck. If you make statements which cannot be falsified, then you can never increase you knowledge of an area. In these methodological backwoods Popper still rules with an iron hand.

I have never been happy with the distinction between exploratory and confirmatory, nor with any other related Aristotelian dichotomy. It seems more realistic and productive to use the point of view that these are really two aspects of the same activity, which do not take place in linear order. It is not the case that first we explore for 30 days, then we confirm for 10 minutes. Or that inn exploration there are no rules, everything goes. While in confirmation we have to obey the very strict prescriptions of the Neyman-Pearson theory, or we have to be coherent in a very specific and somewhat peculiar sense of the word. We think both activities go on simultaneously, until the bitter end (the final revision or final rejection) and can be separated only with great difficulty. All popular statistical techniques seem to have very important descriptive components, and although sometimes statisticians and methodologists may emphasize the inferential aspects of LISREL, regression, and ANOVA, it seems quite clear, to me at least, that these techniques are mainly popular because of their descriptive properties.

In this address I shall try to indicate that MVA/OS techniques do not differ a great deal from classical MVA techniques, and from covariance structure techniques in particular. There is no gap that separates us.

The use of models in multivariate data analysis

We restrict ourselves to forms of multivariate analysis in which models for the covariances between, say, m variables are studied. Higher order moments are ignored. There is ample historical precedent for imposing this restriction. In the first place the centered multivariate normal distribution is described completely by its covariance matrix, and the multivariate normal distribution is of course the leading case in much of multivariate analysis. In the second place higher order moments can often not be determined with sufficient statistical precision, because their standard errors grow so fast. And thirdly higher order product moments are even less precisely determined, because there are so many of them. The number of 10^{th} order product moments, for instance, is m^{10}, which will usually be much larger than the number of observations. Finally second order moments have a nice interpretation in terms of linear least squares regression and prediction, and in terms of linear structural models or systems of simultaneous linear equations. This does not mean that higher order moments are completely irrelevant for multivariate analysis. In fact a lot of recent work by Bentler, Brown, De Leeuw, Mooijaart, Satorra, and Shapiro concentrates on the use of higher order moments to increase precision and/or generality, or to obtain identification. To keep things simple, and without losing too much generality in practice, we shall concentrate on models for second order moments.

These models have a certain parametric form. Thus they say that the covariance matrix (or the correlation matrix) Σ is of the form $\Sigma(\theta)$, with θ a vector of real parameters. A simple example, and by far the most interesting one in the history of psychometrics, is the Spearman model $\Sigma = aa' + \Delta^2$.

We all know that the Spearman model is false, i.e. it does not describe the covariance between batteries of tests in a satisfactory way. The fact that many of these models used in psychometric data analysis are false is often considered to be a disadvantage. David Freedman (1988, and many other papers) has attacked social science models builders because they assume things to be true which are quite obviously false. But is this really such a big disadvantage ? In order to answer this in general terms we have to discuss the role of statistical models (De Leeuw, 1984a, 1988a, 1988b).

Let us remain in the context of fitting covariance matrices. It is not difficult to find a model that is true: simply assume nothing about the form of the covariance matrix. But this model, although very true, is also very useless. In order to describe the covariance matrix we compute the covariance matrix. It is as if we have not started our analysis yet. Using a model, i.e. singling out some subset of the space of all covariance matrices, has two important functions. It makes it possible to express our results in a form in which they can be readily interpreted, i.e. related to existing theories or prejudices. And it increases the stability of the estimates.

We illustrate this with a (very) simple example. Suppose x_i (i=1,...,n) are iid $\mathcal{N}(\mu_1,\sigma^2)$ and y_i (i=1,...,n) are iid $\mathcal{N}(\mu_2,\sigma^2)$. The x_i are independent of the y_i, and $\mu_1 \neq \mu_2$. We compare two strategies: estimating μ_1 and μ_2 by the sample means m_1 and m_2 versus estimating both μ_1 and μ_2 by the pooled mean $(m_1 + m_2)/2$. The mean square error of the first procedure (which fits the correct model) is $2\sigma^2/n$, and that of the second procedure (wrong model) is $(\mu_1 - \mu_2)^2/2 + \sigma^2/n$. If the two means are very close, i.e. if $(\mu_1 - \mu_2)^2/\sigma^2 < 2/n$, then fitting the wrong model is better than fitting the correct model (at least in this sense). If μ_1 and μ_2 are $.1\sigma$ apart, then the 'wrong' procedure is better than the 'correct' procedure if $n < 200$. If we are quite sure that men and women have the same average intelligence, then the best way to estimate the average intelligence of women is to compute the average intelligence of all individuals, men and women alike.

This suggest the following conclusion. If we use very general models we reduce the bias in our estimates, because very general models tend to be true. The sample covariance matrix will be unbiased, or at least consistent, under very general conditions. On the other hand we decrease the stability of our estimates, and we decrease the interpretability of our results. But you can have too much of a good thing. Stability is desirable, but certainly not the only criterion we should take into account. If we estimate our covariance matrix by setting it equal to the identity, no matter what the data are, we have a very stable technique. All standard errors are zero. But the technique is almost always completely useless. Figure 1 illustrates the basic dilemma. S_n and T_n are two observed covariance matrices, distributed around the Truth Σ_0. We have a model $\Sigma(\theta)$, and the figure shows that projecting observed matrices on the model generally leads to smaller variability, although here clearly truth is not on the model. Whether the increased bias, due to the nonzero distance of Σ_0 and $\Sigma(\theta)$, is serious enough to offset the gain in stability cannot be seen from the figure, because it depends on the precise way of measuring these quantities. But the figure suggests that sometimes fitting the wrong model may be OK, or (a bit less provocative) that we do not want to make our models exactly right.

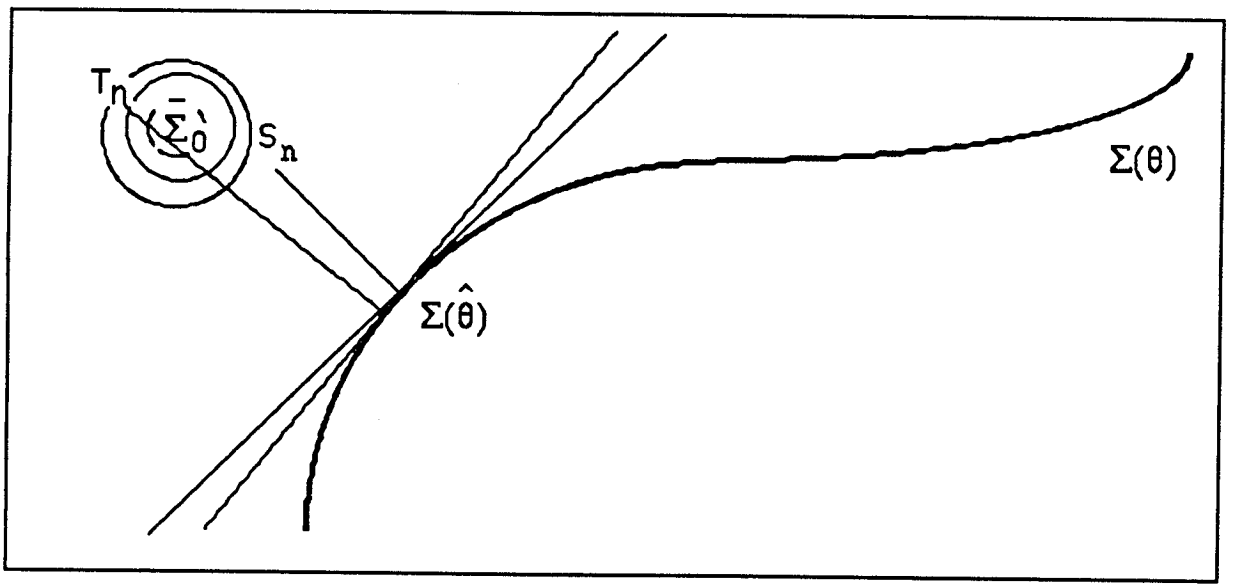


Figure 1: On the role of models for the covariance matrix

The example of the two means is perhaps a bit too simple. A far more interesting class of examples, general multinomial experiments, are analyzed in considerable detail by De Leeuw (1988a). The problem analyzed there is to estimate a vector of m probabilities π. We observe a vector of m proportions p. Both π and p are in the unit simplex $\mathbb{S}^{m-1}$, i.e. the set of all m-vectors with nonnegative elements adding up to one. We could use p directly as an estimate of π, but this is not necessarily the best procedure. We replace it by an estimate of the form $\Phi(p)$, where Φ maps $\mathbb{S}^{m-1}$ into $\mathbb{S}^{m-1}$. For example, Φ could be the maximum likelihood estimate of π based on some model Ω, where Ω is a subset of $\mathbb{S}^{m-1}$. Let us measure the closeness of our estimate to the true value by the random variable $\Delta(\pi,\Phi(p))$, more precisely by the expected estimation error $\mathbf{E}(\Delta(\pi,\Phi(p)))$. We also measure the predictive quality of our estimate by the expected prediction error $\mathbf{E}(\Delta(q,\Phi(p)))$, where q is a replication of p (i.e. it is independent of p, and it has the same distribution). Both errors cannot be observed directly because π and q are not observed, but they can be estimated by using the delta method or resampling methods such as the Jackknife and Bootstrap. De Leeuw (1988a) shows that using a model sometimes improves the quality of the estimate, even if the model is not precisely true. Or, more precisely, $\Phi(p)$ can be better than p without having to assume that the model Ω is true, i.e. that $\pi \in \Omega$.

These considerations suggest that the basic dilemma in statistical modelling (actually in any kind of mathematical modelling) is the trade-off between bias and variance. If our model is too specific (if we impose too much prior information, we use our prejudices, and model from our armchairs) then we have excellent precision (around the wrong value). If the model is too general (if we rely too much on the data, if we refuse to use our prior knowledge) then we have very bad precision around the correct value. It seems to me that this framework is not only useful to describe many recent developments in statistics, but also far more generally to describe various approaches to (inductive and deductive) modelling and to (theory-driven and data-driven) research strategies.

This way of talking about these distinctions suggests a compromise. If there is firm prior knowledge one must use it, because it enhances stability. If there are merely prejudices then one should not use them, because they introduce bias. We also want to emphasize that confirmatory statistical analysis of complicated multivariate models is tricky, because we merely test the structural core $\Sigma(\theta)$ in an implausible stochastic framework. It is nice to use the multinormal distribution in examples, because of its many beautiful properties, but from the descriptive point of view it is sadly lacking. If we do not want to assume that the structural core is true, then we certainly want to assume no such thing about the even more implausible stochastic framework. Much more useful ways of testing appropriateness of models are possible than the simple chi square, which assumes that the model is true. The AIC and cross-validation, which are special cases of the approach based on the estimated estimation and/or prediction error, seem to be more to the point (De Leeuw, 1988a, or the September 1987 issue of Psychometrika).

A multivariate analysis framework

In this talk we shall accept a somewhat nonstandard and quite general framework to discuss multivariate analysis problems. Each *variable* y is an element of a separable real Hilbert space $\mathcal{H}$. The inner product in $\mathcal{H}$ is $\langle .,. \rangle$, the norm is $\|.\|$. Thus $\|y\|^2 = \langle y,y \rangle$. The norm of a variable is called its *variance*, the inner product of variables y_1 and y_2 is their *covariance*.

There is no need to be unduly impressed by our use of Hilbert space terminology. If you keep in mind the usual interpretation of variables and their variances and covariances you can follow all the discussions. The technical aspects of our framework are discussed in De Leeuw (1988c). The notation and terminology is attractive, because it allows us to discuss various important special cases in one single framework (the same argument can also be found in Guttman, 1955). In the first place

there is the case of a finite vector of real numbers, in which $\mathcal{H} = \mathcal{R}^n$, and $\langle y_1,y_2\rangle = \Sigma_{i=1}^n y_{1i}y_{2i}$. In the second place there is the case of m (population) random variables (with finite variance) defined on the same probability space $\langle \Xi,\mathcal{B},P\rangle$. Here $\mathcal{H} = \mathbb{L}_2(\Xi,\mathcal{B},P)$ and $\langle y_1(.),y_2(.)\rangle = \int y_1(\xi)y_2(\xi)dP(\xi)$.

Multivariate analysis involves m variables $\phi_1,...,\phi_m$ defined on the same space. Collect them in a *multivariable* Φ. Just another new word. Throughout the paper we use the following example. We have measurements on four variables for 1270 pupils leaving primary education in the city of Groningen, The Netherlands, in 1959. These are the so-called GALO data, analyzed earlier by Peschar (1973). We know their SEX, their IQ, the profession of their father, and the advice the teacher gave about the most appropriate form of secondary education for this pupil. Thus the four variables are four vectors of 1270 real numbers, or alternatively four functions on $\mathcal{J} = \{1,2,...,1270\}$, equipped with counting measure. The real numbers are used only as lables, for obvious reasons. SEX is binary, IQ is fairly continuous, FATHER is a six point scale with categories {unskilled labour, skilled labour, lower white collar, small business, higher white collar, higher professions}, and ADVICE has the categories {no further education, extended ordinary primary education, lower technical education, lower agricultural education, intermediate secondary education, secondary education for girls, preparatory higher education}. Both FATHER and ADVICE are not really ordered, but sociologists usually argue that scales such as these can be treated as ordinal scales, in fact applied sociologists simply treat them as interval scales by using equally spaced quantifications of the categories. I am not saying that this is wrong. I am not the police.

Fitting covariance and correlation models

We shall compare, and to a certain extent contrast, two different approaches to multivariate analysis in this paper. In the first approach we assume that we are fitting a *model* to our variables. A model, in

this context, is an expression for the covariance matrix of the variables. We suppose that it can be written as $\Sigma(\theta)$, with $\theta \in \mathcal{R}^p$. Thus Σ is on a p-dimensional manifold in m(m+1)/2 dimensional space. Now if the y are completely known, then we know $\Sigma = E(yy')$, and we can choose parameters θ in such a way that $\Sigma(\theta)$ is approximately equal to Σ. This looks like an approximation problem, with little or no statistical content, but it may be the correct interpretation of our activities in many cases in which the standard statistical framework does not apply. If there is no question of a random sample, for instance, and we study the population of individuals we are interested in, then we approximate the population covariance matrix by a lower dimensional one based on a model. We do this because the parametric model is more parsimoneous, easier to interpret, i.e. to relate to existing theory, and to communicate. Statistics has nothing to say about fitting models to population covariance matrices. All deviations are significant. All hypotheses are rejected.

In the usual statistical interpretation we do not know Σ, but we have an *estimate* S (often based on n iid observations). We now use a distance measure Δ between the observed covariance matrix S and the parametric manifold $\Sigma(\theta)$, and we find θ in $\mathcal{R}^p$ such that $\Delta(S,\Sigma(\theta))$ is minimized. In other words: we project S on the manifold $\Sigma(\theta)$, and we evaluate the fit by looking at the distance between observed and expected (i.e. projected). The really important part is the model $\Sigma(\theta)$, statistical assumptions such as normality or iid usually only influence the choice of the distance measure Δ.

There is an important shift in emphasis here from the usual way of approaching these problems. We do not suppose that the usual statistical assumptions (such as normality, or iid observations) are part of the model (or at least, they do not belong to the *core* of the model). In the terminology of Van Praag, De Leeuw, and Kloek (1986) we decompose the problem into its population and sample aspects. Choice of loss function, or distance measure, should be seen as an independent problem. We know that some distance measures are better than others if particular stochastic models are true, but often we do not care to assume that these models are indeed true. Because it is fairly obvious, in most

cases, that assumptions such as normality and iid are false, difficult to verify, made only for technical reasons, not essential for the substantive scientific aspects of the problem, and so on. In fact it is often difficult enough to argue that the core of the model, the expression $\Sigma(\theta)$, is plausible. In the usual statistical interpretation we assume that S is not equal to Σ because of random sampling. Thus, if our sample becomes larger, we think that S will approach Σ, and we assume that this limiting Σ will satisfy the model exactly. At least we *act* as if we think this.

The second approach is quite different, and in a sense more general. We suppose the covariance matrix of the variables depends on a number of parameters (the variables are not completely known). Thus we have a function $S(\xi)$, where $\xi \in \mathcal{R}^q$. The ξ can be thought of as *transformation parameters*, but they can also be covariances involving latent variables. We now pick an *aspect* of the covariance matrix that we are interested in (the multiple correlation coefficient, the largest eigenvalue, ...). An aspect is a real valued function κ defined on the space of covariance matrices. We then optimize $\kappa(S(\xi))$ over ξ.

This may not be immediately clear, so let us illustrate it with a few examples. In Box-Cox regression we assume that there is a transformation f of the dependent variable such that the vector z with elements $z_i = f(y_i)$ is distributed as $\mathcal{N}(X\beta,\sigma^2 I)$. The transformation has the form $f(y) = (y^\lambda - 1)/\lambda$, with λ unknown. We can now compute the likelihood of the observations and maximize this over β and σ^2. The resulting expression is a function of λ, and this function can then be maximized over λ. There have been many generalizations of this approach (by Winsberg and Ramsay, 1980, or De Leeuw, 1984c, 1986, for instance). In stead of maximizing the likelihood, however, we can also minimize the residual sum of squares (or maximize the multiple correlation) over λ. This is basically the approach in optimal scaling (ALSOS by Young et al., GIFI by Gifi, ACE by Friedman et al.). Using the likelihood can be interpreted quite easily in the framework of assuming a model, and then

projecting on the model. Using the residual sum of squares means choosing an aspect which does not directly have such an interpretation.

Imputation of missing data is another example (for a recent thorough analysis of such problems we refer to Little and Rubin, 1987). We can choose a probability model (the multivariate normal distrbution), assume missing data are missing at random, write down the likelihood of the completed data (which is a function of the values chosen for the missing data), and maximize this over structural parameters and missing data. We can also integrate out the missing data from the likelihood, and use the EM algorithm to maximize the likelihood of the observed data. And we can also choose some aspect of the problem that seems interesting (for instance a measure of collinearity or of predictive power) and optimize this over missing data values. Similar approaches are possible if we do not only miss some observations, but actually complete variables (so called *latent variables*). The approach of marginalizing and optimizing the multinormal likelihood of the observed data is the usual one (Anderson, Bentler, Browne, Joreskog). Optimizing an interesting aspect of the covariance matrix over the unknown latent variables (and over missing data, and over transformations, ...) is done in ALS (Gifi), ACE (Friedman), and PLS (Wold) methods. If we only observe discreticized or ordered versions of the variables we can again follow the usual approach by computing and optimizing the likelihood of the observed data (Muthen, 1984), or we can optimize over another aspect that interests us (as in ALS, etc.).

I guess most of you have experience with the usual statistical model-based approach. In recent years many people (Bentler, Browne, De Leeuw, Mooijaart, Muthen, Shapiro) have tried to make the framework of model fitting a bit more realistic by relaxing the 'assumption of multivariate normality'. In our terminology this means that they have suggested other distances between S and $\Sigma(\theta)$ which (perhaps) behave more satisfactory in practice. Peter Bentler's presidential address of 1983 contains a discussion of much of the earlier work for continuous non-normal data, and much has been

accomplished since then. De Leeuw (1983b) reviews various extensions to discrete data. We consider most of this to be fine-tuning, useful but not very consequential. The misspecification of the structural part of the model has much more important consequences, both in terms of statististical stability and in terms of substantive interpretation, than misspecification of the stochastic framework.

On subspace and cone constraints

You are perhaps less familiar with the optimal scaling approach. We consequently specialize it a little bit, and show what it amounts to. The specialization consists of the fact that our partial knowledge of the variables can be written in the form $\phi_j \in \mathcal{L}_j$, with $\mathcal{L}_j$ a subspace of $\mathbb{L}_2(\Xi,\mathcal{B},P)$, the space of variables with finite variance. It could be a subspace of polynomials, or of splines, or whatever. Most of our results apply equally to $\phi_j \in \mathcal{K}_j$, with $\mathcal{K}_j$ a convex cone, for instance the cone of monotone functions. The interpretation of these constraints is simple. It is very similar to the Box-Cox approach to regression or to the Kruskal approach to nonmetric multidimensional scaling. We are not only interested in the variables as we observe them, but we are interested in all smooth or monotonic or polynomial or splinical transformations of them as well. If one such *reexpression* suits our purposes (i.e. our *aspect*) better than the orginal expression of the variable, then we use that reexpression.

For computational purposes we suppose the $\mathcal{L}_j$ to be finite-dimensional, and we suppose that matrix $\mathcal{G}_j$, which may have an infinite number of rows, contains a basis for $\mathcal{L}_j$. Thus all transformations that we study are of the form $\mathcal{G}_j\xi_j$. The unknowns in $S(\xi)$ are the vectors ξ_j containing the coefficients of each of the basic functions. Collect the inner products of the basis function in matrices $\mathcal{C}_{jl} = \mathcal{G}_j'\mathcal{G}_l$. This generalizes the contingency table of variables j and l. If the $\mathcal{G}_j$ are dummies (a.k.a. indicator functions, splines of degree one, step functions) then $\mathcal{C}_{jl}$ is equal to the contingency table. The covariance of the transformed variables is now simply $\xi_j'\mathcal{C}_{jl}\xi_l$. These are the elements of $S(\xi)$.

Because we work with subspaces we shall actually impose scale-freeness and work with aspects of the correlation matrix $R(\xi)$.

One way of moving a little bit closer to the model-based approach is by choosing the criterion in a particular way. In this sense we can say that the approach based on aspects is more general. We use a model $P(\theta)$ for the correlation matrix, choose a distance measure, and define as our aspect

$$\kappa(R(\xi)) = \min \{\Delta(R(\xi),P(\theta)) \mid \theta \in \Re^p \}.$$

Then minimize this aspect of the covariance matrix, which measures the fit of the correlation model $P(\theta)$, over ξ. This is done, for instance, by Takane et al. (1979) in their FACTALS program. Along these lines one could fairly easily make an optimal scaling version of LISREL or EQS or COSAN as well. Not that I suggest that anybody should really do this. This way of combining OS and fitting structural models does not seem to be very natural. We try to have the best of both worlds, but we seem to destroy the desirable properties of the first world in the process, because we obviously cannot use the statistical theory associated with LISREL etc. any more. If we maximize the likelihood over structural parameters and over transformations, we cannot expect the estimates to be efficient and we cannot expect to continue to use chi squares for our likelihood ratios. The precise definition of the model has become unclear. This can have serious consequences (Dijkstra, 1983, Little and Rubin, 1983).

What we find, in practice, is that people nevertheless like to have the best of both worlds. They want to transform their variables because they are unsure about the precise expression they need, but they also want the statistical information that comes out of LISREL. No matter how much we try to discourage them to take that information seriously, and no matter how much statistical information we present with the optimal scaling methods, they want to publish their final results as LISREL results.

Some programs are publication vehicles. Thus they first apply some optimal scaling method (such as multiple correspondence analysis). This finds the 'correct' expression of their variables. These re-expresssed variables are then fed into LISREL etc., and analyzed as usual. There are a number of nice examples in Bakker, Dronkers, and Ganzeboom (1984). This approach used to worry me a great deal, although it has obvious data analytical and didactical advantages. At a relatively cheap price we get a lot of useful additional information in the form of transformations of the variables and experience with optimal scaling techniques has taught us that these transformations are useful diagnostic tools. But in the meantime I have discovered or unearthed some pleasant results, which make this two-stage eclectic approach a bit more respectable.

Linearizable regressions

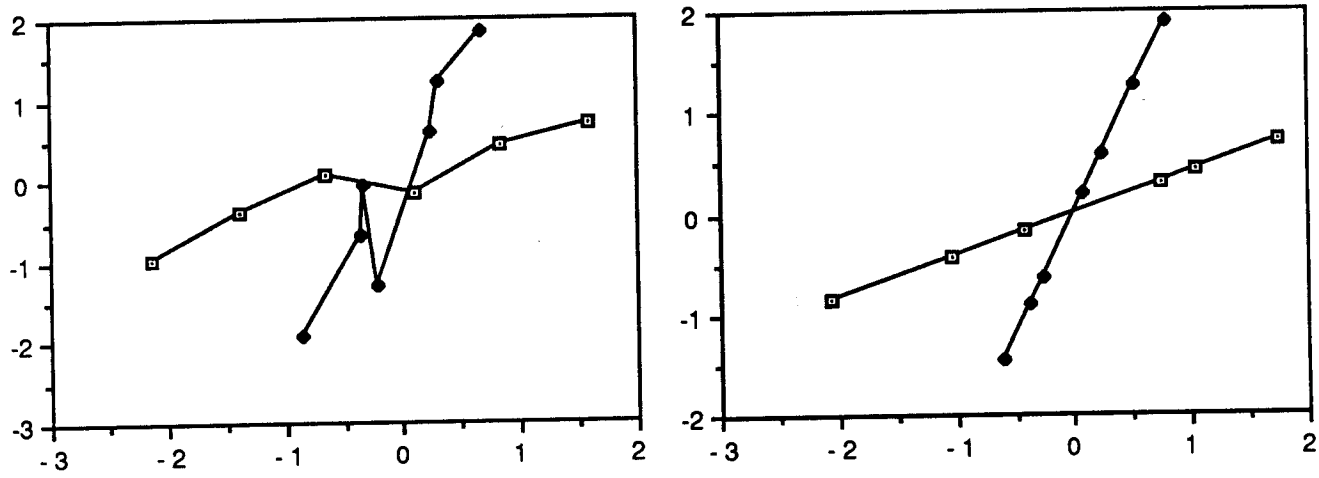
There is a simple way to introduce correspondence analysis of a two-way table. It was discovered by Hirschfeld (1935). Suppose we have two categorical variables with corresponding dummy codings $\mathcal{G}_1$ and $\mathcal{G}_2$, and with cross table $\mathcal{C} = \mathcal{G}_1'\mathcal{G}_2$. Moreover the univariate marginals are $\mathcal{D}_1 = \mathcal{G}_1'\mathcal{G}_1$ and $\mathcal{D}_2 = \mathcal{G}_2'\mathcal{G}_2$. If ξ_1 and ξ_2 are normalized scores for the categories of the variables, then we can compute the conditional expectations $\eta_1 = \mathcal{D}_1^{-1}\mathcal{C}\xi_2$ and $\eta_2 = \mathcal{D}_2^{-1}\mathcal{C}'\xi_1$. We have linear regression if η_1 is proportional to ξ_1, and η_2 is proportional to ξ_2. Or, in formulas, if

$$\mathcal{C}\xi_2 = \rho\mathcal{D}_1\xi_1, \tag{1a}$$

$$\mathcal{C}'\xi_1 = \rho\mathcal{D}_2\xi_2. \tag{1b}$$

But the two equations (1a) and (1b) can be solved by computing the singular value decomposition of the matrix $\mathcal{D}_1^{-1/2}\mathcal{C}\mathcal{D}_2^{-1/2}$. If (λ,a,b) is a triple consisting of a singular value and the corresponding left

and right singular vectors, then $\xi_1 = \mathcal{D}_1^{-1/2}a$, $\xi_2 = \mathcal{D}_2^{-1/2}b$, and $\rho = \lambda$ satisfy (1a) and (1b). And this is true for each such triple. This is Hirschfeld's theorem: we can always choose scores for the rows and columns of a contingency table in such a way that the two regressions are linear. In the table has r rows and c columns, then we can actually choose such scores in min(r,c) ways.



Figures 2a and 2b: Regression of teachers advice on fathers profession, before scaling and after scaling

Figures 2a and 2b give an example of the working of Hirschfeld's theorem. There is a very similar example, without picture, in Louis Guttman's presidential address of 1971. We have plotted fathers profession on the horizontal axis, and teachers advice on the vertical axis. The hollow squares show the regression line of advice on profession, i.e. the six average advices for the six father classes. The solid squares give the regression of profession on advice. In Figure 2a we have used equally spaced scores for the seven categories of advice and for the six categories of fathers profession. We see that the regression of advice on profession is fairly linear, at least monotonic, but the regression of father on advice is much disturbed by the fact that children with lower technical and lower agricultural advice have on the average a slightly less satisfactory father than children with extended ordinary primary education advice. Figure 2b shows the regression after optimal scoring by using the dominant

singular value. By modifying the spacing on both axes (and actually the order on the vertical axes) the regression functions become straight lines. The correlation between the scaled variables increases from 0.36 to 0.39.

If there are $m > 2$ variables the situation becomes more complicated. This was discussed in a little known, but very interesting, paper by Louis Guttman (1959). Who else. In general there do not exist scores that linearize all bivariate regressions. The precise situation has been reviewed recently by Bekker and De Leeuw (1988). Linearity of bivariate regressions imposes restrictions, which means that we have to use a model. There you are. Now we are in trouble. To quote David Freedman, we have said the M-word. Our model supposes that there exist expressions (or transformations) of the variables that make all bivariate regressions linear. Obviously this is much weaker that assuming multivariate normality or multivariate ellipticity, because in those cases the regressions are already linear without any re-expression. It is also more general than what Udny Yule has called the *strained multinormal*, which is the family you get if you apply separate monotonic distortions to all variables of a multivariate normal distribution.

For computational purposes we have to suppose that the linearizing transformations are in $\mathcal{E}_j$. The model then says that there exist ξ_j such that for all (j,l)

$$\mathcal{C}_{jl}\,\xi_l = \rho_{jl}\mathcal{D}_j\xi_j. \tag{2}$$

And perhaps the ρ_{jl} satisfy a correlation model of the form $P(\theta)$, as usual.

Now let us consider applying an optimal scaling program which optimizes an aspect of the correlation matrix, say $\kappa(R(\xi))$. At the optimum we have the stationary equations

$$\Sigma_{l=1}^{m} (\partial\kappa/\partial r_{jl})\, \mathcal{C}_{jl}\xi_l = \lambda_j \mathcal{D}_j \xi_j, \tag{3}$$

where the ξ_j are normalized by $\xi_j'\mathcal{C}_{jj}\xi_j = 1$. But let us substitute the linearizing equations (2) in this. We then find that the linearizing ξ satisfies the stationary equations (3), with

$$\lambda_j = \Sigma_{l=1}^{m} (\partial\kappa/\partial r_{jl})\rho_{jl}. \tag{4}$$

We have proved the following: if linearizing transformations exist they will be found by optimal scaling techniques. Actually we have not proved this much, it is perhaps better to say that we have shown that linearizing transformations, if they exist, *can* be found by optimal scaling techniques. Stronger results are possible by imposing additional conditions. This result is, at least implicitly, in Guttman (1959). De Leeuw (1983a) has pointed out that it generalizes a much older result of Pearson (1906). It has important consequences. Because of this the two-step procedure mentioned above, first OS and then LISREL (or factor analysis, or regression), finds consistent estimates of the structural parameters θ if all bivariate regressions can be linearized.

This is nice: the two-stage procedure reduces the bias of the existing programs, because it makes them consistent over a far larger class of distributions. We could call this a *robustness* property. The structural equation programs do not need multivariate normality or linear regressions to produce consistent estimates. If we combine them with an OS method, then they produce consistent estimates if the bivariate regressions are merely linearizable. If we apply OS to strained multinormal data in the sense of Yule, then the techniques *unstrains* them. It finds the inverse transformations, and makes the distribution of the transformed variables exactly normal.

There is another important aspect of the result. A common criticism of MVA/OS is that the quantifications of the variables depend on the aspect one has selected. If you optimize the multiple

correlation coefficient you find different quantifications from the ones you find if you are optimizing the sum of the two largest eigenvalues. And in fact a two dimensional principal component analysis with optimal scaling finds transformations which differ from those found by a three dimensional one. This is not strange, after all component loadings are also different from regression weights, but it complicates the interpretation. If income is scaled as a very flat function which accellerates suddenly very quickly at high income levels, then the interpretation of the results will have to take that into account. A transformation which is more log-like, i.e. which decelerates and spreads out the lower incomes, will lead to a different interpretation. Our previous result says that if all bivariate regressions can be linearized, then different OS techniques will find the same quantifications, namely precisely those quantifications which linearize the regressions.

There is a generalization of the result in this section which is of some importance. We have concentrated on aspects which are functions of the correlation coefficients. In the context of linearity of the regressions the correlation ratio's are also of some importance. In our notation, with normalized scores, they are defined as $\eta_{jl}^2 = \xi_j' \mathcal{C}_{jl} \mathcal{D}_l^{-1} \mathcal{C}_{lj} \xi_j$. Let us now extend our results to aspects which are functions both of the correlation coefficients and the correlation ratio's. The stationary equations (3) generalize to

$$\Sigma_{l=1}^m (\partial\kappa/\partial r_{jl})\, \mathcal{C}_{jl}\xi_l + \Sigma_{l=1}^m (\partial\kappa/\partial\eta_{jl}^2)\mathcal{C}_{jl}\mathcal{D}_l^{-1}\mathcal{C}_{lj}\xi_j = \lambda_j \mathcal{D}_j \xi_j, \qquad (5)$$

If the ξ_j and the ρ_{jl} satisfy (2), then they satisfy (5), with

$$\lambda_j = \Sigma_{l=1}^{m} (\partial\kappa/\partial r_{jl})\rho_{jl} + \Sigma_{l=1}^{m} (\partial\kappa/\partial\eta_{jl}^2)\eta_{jl}^2. \qquad (6)$$

Thus the theory also applies to this more general class of aspects. A particulary simple aspect in this class, which has been discussed by De Leeuw (1982) and Bekker and De Leeuw (1988), is

$$\kappa(\xi_1, \ldots, \xi_m) = \Sigma_{j=1}^{m}\Sigma_{l=1}^{m} (\eta_{jl}^2 - r_{jl}^2). \qquad (7)$$

This is zero if and only if all bivariate regressions are linear, and generally gives a useful and informative method to measure deviations from linearity. It is relatively simple to minimize κ of (7) over the normalized ξ_j. We optimize it over one ξ_j at the time, keeping all others fixed at current values. Each subproblem is a simple generalized eigenvalue problem, of order equal to the number of categories of variable j. And we cycle over the subproblems. This is done by using the program LINEALS, previously employed by Van Rijckevorsel (1987), and it can also be done by using the Jacobi-like plane orthogonal rotation techniques of the PREHOM program described by Bekker and De Leeuw (1988).

INSERT TABLES 1, 2, 3 ABOUT HERE

Let us now apply optimal scaling to the GALO example. If we use equally spaced normalized scores, the results are in Table 1. The most remarkable finding are the different boy/girl ratios in the various father's profession categories, which corresponds with a correlation ratio of 0.21. The squared correlation between IQ and advice is high, the correlation ratio is not much higher. The total discrepancy (7) is 0.36. Table 2 gives optimal scoring and induced correlations computed by multiple correspondence analysis (a.k.a. homogeneity analysis, Guttman's principal components of categorical variables, Hayashi's fourth method of quantification). This is the most popular OS technique, and it

does a nice job. It brings down the total discrepancy to 0.22, and linearizes most of the regressions nicely. The only remaining problem case is the regression of sex on fathers profession, which has a fixed correlation ratio not dependent on scoring. This makes it difficult to improve the situation there. And, of course, we do not expect linearization techniques to work too well with binary variables such as sex. Either linearity is trivial, in one direction, or not very natural, in the other direction. LINEALS, finally, in Table 3, does not improve much on the multiple correspondence analysis solution. The quantifications are similar, the correlations are very similar, and the total discrepancy decreases to 0.21. The similarity of Tables 2 and 3 illustrates the fact that if linearizing scores exists, then different OS techniques will find them. The optimal scores themselves look quite reasonable.

Stability of induced correlations

We have now found quantities which are consistent estimates of the correlation coefficients under the assumption of linearizability of the regressions. This links MVA/OS with the classical MVA techniques which take the correlation matrix or covariance matrix as a starting point. Anything you can do to the correlation matrix of the unscaled variables you can also do to the correlation matrix of the scaled variables. If the regressions are nonlinear, but can be linearized, then using the original variables introduces bias. From the data analysis point of view, independent of statistical considerations, the meaning of correlation coefficients is a bit doubtful in the case of nonlinear regressions. If the regressions are linear, then both matrices are consistent estimates of the same quantity. In particular for multinormal data, they both consistently estimate the population correlation matrix. For strained multinormal data OS techniques unstrain, while techniques which do not scale distort the relations.

On the other hand, from our general considerations above, we expect the stability to go down as a consequence. In a more general model, standard errors will increase. And we certainly expect the statistical information to come out of LISREL applied to optimally transformed variables to be in error. There is a second nice robustness type of result, however, which can make us feel less pessimistic in this respect. This result has been discussed, in a closely related context, by Steiger and Browne (1984).

Suppose

$$\rho_{jl} = \xi_j' \mathcal{C}_{jl} \xi_l \,/\, (\xi_j' \mathcal{D}_j \xi_j)^{1/2} (\xi_l' \mathcal{D}_l \xi_l)^{1/2} \tag{8}$$

is the induced correlation coefficient, using optimal scores. In we apply the delta method to compute the variance of its asymptotic distribution we have to look at the derivatives of ρ_{jl} with respect to the probability distribution F. To compute the derivatives of ρ_{jl} with respect to F, we need the derivatives of ξ_j and ξ_l with respect to F, and the derivatives of $\mathcal{C}_{jl}$, $\mathcal{D}_j$, and $\mathcal{D}_l$ with respect to F. The nice result is that if the ξ_j linearize the bivariate regressions, then the derivatives of ξ_j and ξ_l with respect to F drop out of the expression. This is easy to see. We find

$$\partial\rho_{jl}/\partial F = (\partial\xi_j/\partial F)'(\mathcal{C}_{jl}\xi_l - \rho_{jl}\mathcal{D}_j\xi_j) + (\partial\xi_l/\partial F)'(\mathcal{C}_{lj}\xi_j - \rho_{jl}\mathcal{D}_l\xi_l) +$$

$$+ \xi_j'(\partial\mathcal{C}_{jl}/\partial F)\xi_l - 1/2\rho_{jl}\{\xi_j'(\partial\mathcal{D}_j/\partial F)\xi_j + \xi_l'(\partial\mathcal{D}_l/\partial F)\partial\xi_l). \tag{9}$$

The first two terms on the right hand side vanish if the regressions are linear. Only the second part contributes to the standard error, and this does not involve the derivatives of the ξ_j, and is the same as it is for fixed scores. QED.

The result of Steiger and Browne is somewhat more specific, actually. In our context they show that if the scores ξ_1 and ξ_2 are chosen in such a way that they maximize the correlation coefficient, then the distribution of the optimum correlation coefficient is the same as the distribution of the ordinary correlation coefficient between $\varsigma_1\xi_1$ and $\varsigma_2\xi_2$, with the scores considered as fixed numbers. We have shown that this is true for all sets of scores that linearize the regressions. Moreover if all bivariate regressions can be linearized, then the result is true for the joint distribution of the induced correlation coefficients, with scores computed by any OS technique. Thus if the LISREL type program uses the general Isserlis (1916) distribution free formula for the covariance of correlations, then the statistical information provided by the program will be OK, even after optimal scaling of the variables.

We give the Isserlis-weights here, for completeness, using notation of De Leeuw (1983b, page 117). The correlation coefficients r_{ij} and r_{kl} are jointly asymptotically normal. A consistent estimate of the covariance in the asymptotic normal distribution is given by

$$w_{ijkl} = r_{ijkl} - \tfrac{1}{2} r_{ij}(r_{iikl} + r_{jjkl}) - \tfrac{1}{2} r_{kl}(r_{kkij} + r_{llij}) +$$

$$+ \tfrac{1}{4} r_{ij}r_{kl}(r_{iikk} + r_{iill} + r_{jjkk} + r_{jjll}), \qquad (10)$$

with

$$r_{ijkl} = s_{ijkl}s_{ii}^{-1/2}s_{jj}^{-1/2}s_{kk}^{-1/2}s_{ll}^{-1/2}, \qquad (11)$$

and

$$s_{ijkl} = n^{-1} \sum_{v=1}^{n} (x_{vi} - m_i)(x_{vj} - m_j)(x_{vk} - m_k)(x_{vl} - m_l). \qquad (12)$$

Here the x_{vi} are either the original or the optimal scores, the m_i are the sample means, and the s_{ii} are sample variances. There are no convenient matrix expressions for these quantities, unless you care to define a new matrix calculus of your own. Some people actually do this.

We can use the Isserlis formulae to compute standard errors of the induced correlation coefficients from Tables 1, 2, and 3. It turns out that these standard errors are very similar. The three correlation coefficients involving sex are a little bit more stable if we use equal interval scoring (efficiencies, i.e. ratio's of standard errors, are between .90 and .99). The three remaining correlation coefficients, which are in a sense the more interesting ones, are more stable if we use optimal scoring (efficiences between 1.00 and 1.04). If we combine this with the fact that optimal scoring reduces the bias, we see that we are in a situation in which we do not loose precision and reduce the bias, which is pretty favorable. I do not know in how far this generalizes to other examples as well, but it sure is nice.

We can combine this with the fitting of a structural model, for instance the model that sex and fathers profession are independent of teachers advice given IQ. Thus the partial correlation coefficients between sex and fathers profession on one side and advice on the other side, controlling for IQ, should be zero. We give the partial correlations, with standard errors in parentheses, and the chi square with two degrees of freedom to test the hypothesis. With equal scoring we find $r_{SAII} = -.07$ (.0275) and $r_{FAII} = .22$ (.0290), chi square is 58.5232. With multiple correspondence analysis we find -.04 (.0278) and .17 (.0304) and 35.0875. For optimal linear scaling, finally, the partial correlations are -.05 (.0276) and .14 (.0300), and chi square is 27.9716. Thus we see that our model is rather false, especially father's profession and teachers advice are not independent, even if we control for intelligence. There is a marginal loss of precision if we use optimal scaling, but, more importantly, the point estimate of the significant partial correlation coefficient does vary a lot. As a

consequence chi square after optimal scaling is less than half of chi square before optimal scaling, although we did not set out to minimize it in any sense. Although statistical theory insures that all three quantities are indeed asymptotically chi square, the interpretation of the statistics in the case of equal interval scaling, in which the regressions are clearly non-linear, is not at all clear. We can only interpret vanishing partial correlation coefficients as indices of conditional independence if regressions are indeed linear. Observe, by the way, that if one wants to use log-linear methods to test partial independence one winds up with a chi square with 9 x (7 -1) x (12 - 1) = 594 degrees of freedom. The multidimensional contingency table has 756 cells, with an average of 1.5 observations per cell. This does not seem to be a very practical alternative to correlation-based methods, although in our example it is indeed more sensible to treat sex as a genuine categorical variable which does not need quantification.

Conclusions

In this paper we have introduced a combination of optimal scaling methods with asymptotic distribution free methods to fit correlation structures. And we have found a justification for this two-step method by considering the class of distributions whose bivariate regressions can be linearized. As we have pointed out using these distributions as a *leading case* or *gauge* seems to make it possible to reduce the bias without introducing too much instability. Also it seems to be a more realistic gauge as the multinormal. Moreover the two-step method provides us with very useful additional information, and it will produce standard results if the regressions are already linear. We also think these results are interesting, because they show another relationship between optimal scaling and the

rest of the multivariate analysis world. They are also interesting because there are some sociologists who actually apply the two-step method, and it is nice to give them some theoretical comfort.

Our results so far are incomplete. They should be supplemented with a systematic investigation of linearity of regression. It is not too difficult to construct a chi square test for this purpose, provided we stay in a categorical data context. It is more complicated to construct a test for linearizability of all bivariate regressions. The discrepancy measure we have used can be used, of course, but its asymptotic distribution does not seem to be very simple (a complicated mixture of chi squares, no doubt). It is possible to construct a more satisfactory test, using the algebra in Bekker and De Leeuw (1988) to develop convenient parametrizations for the model. We shall not discuss this here, because it would take us to far astray, and because the results so far are preliminary.

Another point to emphasize is that many more techniques have appeared in the formerly barren region between classical multinormal multivariate analysis and MVA/OS. A minor statistical industry, sponsored mainly by the European Community, has developed around the idea that correspondence analysis must be related in some sense to the work in log-linear analysis. It turns out that it is indeed related in many ways. We refer in this context to the work of Van der Heijden and De Leeuw (1985), Gilula and Habermann (198.), Goodman (1987), De Leeuw and Van der Heijden (1987, 1988, in press). In the second place the so called polychoric or block multinormal models for fitting simultaneous equation models to categorical data developed among others by Muthen (1984) can be interpreted as optimal scaling techniques. In fact the same thing is true for latent variable methods in general. The easiest way to show this is to formulate these models as incomplete information models (Kiiveri, 198.), and to use the EM algorithm for optimal scaling. This similarity was emphasized for the first time by De Leeuw (1984c). More recently explicit maximum likelihood optimal scaling methods for the strained multinormal distribution have been developed by De Leeuw (1986), and Mooijaart, Meijerink, and De Leeuw (1988). Multivariate extensions of the RC model of Goodman,

which is called the point-multinormal model by De Leeuw (1983b), are also being developed by many persons.

Finally we emphasize that the results presented here are only a tiny portion of optimal scaling, and only one interpretation of the transformations computed by these techniques. One can discuss the whole theory of MVA/OS as a form of multidimensional scaling, based on making low-dimensional representations of distances, and without mentioning classical multivariate analysis at all. This is done in De Leeuw (1984b) and in De Leeuw and Van Rijckevorsel (1988). For many other optimal scaling results we refer to Gifi (1988) and to Van Rijckevorsel and De Leeuw (1988).

References

Anderson, T.W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York, Wiley.

Bakker, B.F.M., Dronkers, J., & Ganzeboom, H.B.G. (1984). *Social Stratification and Mobility in The Netherlands*. Amsterdam, SISWO.

Bekker, P. & De Leeuw, J. (1988). Relations between Various Forms of Nonlinear Principal Component Analysis. In J. van Rijckevorsel & J. de Leeuw, *Progress in Component and Correspondence Analysis*. New York, Wiley.

Bentler, P.M. (1983). Some contributions to efficient statistics in structural models: specification and estimation of moment structures. *Psychometrika*, 48, 493-518.

Besse, P. & Ramsay, J.O. (1986). Principal components analysis of sampled functions. *Psychometrika*, 51, 285-311.

Breiman, L. & Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580-598.

De Leeuw, J. (1968), *Canonical Discriminant Analysis of Relational Data*, Report RN 007-68, Department of Data Theory, University of Leiden .

De Leeuw, J. (1982), Nonlinear principal component analysis, in H. Caussinus a.o. (eds.), *COMPSTAT 1982*, page 77-86, Wien, Physika Verlag.

De Leeuw, J. (1983a), On the prehistory of correspondence analysis, *Statistica Neerlandica*, 37, 161-164.

De Leeuw, J. (1983b), Models and methods for the analysis of correlation coefficients, *Journal of Econometrics*, 22, 113-137.

De Leeuw, J. (1984a), Models of data, Kwantitatieve Methoden, 5, 17-30.

De Leeuw, J. (1984b), The Gifi-system of nonlinear multivariate analysis, in E. Diday a.o. (eds.), *Data Analysis and Informatics II*, page 415-424, Amsterdam, North Holland Publishing Company.

De Leeuw, J. (1984c), Discrete normal linear regression models, in T.K. Dijkstra (ed.), *Misspecification Analysis*, page 56-71, Berlin, Springer Verlag.

De Leeuw, J. (1986), Regression with optimal scaling of the dependent variable. In O. Bunke (ed.) *Proceedings of the 7th International Summer School on Problems of Model Choice and Parameter Estimation in Regression Analysis*. Seminarbericht nr. 84, Sektion Mathematik, Humboldt Universität zu Berlin.

De Leeuw, J. (1988a), Model selection in multinomial experiments. In T. K. Dijkstra (ed), *On Model Uncertainty and its Statistical Implications*, Berlin, Springer Verlag.

De Leeuw, J. (1988b). Models and Techniques. *Statistica Neerlandica,* 42, 91-98.

De Leeuw, J. (1988c). Multivariate analysis with optimal scaling. In S. Das Gupta (ed), *Progress in Multivariate Analysis*, Calcutta, Indian Statistical Institute.

De Leeuw , J. & Van der Heijden, P.G.M. (1987), The analysis of time budgets with a latent time budget model. In E. Diday a.o. (eds.), *Data Analysis and Informatics V,* Amsterdam, North Holland Publishing Company.

De Leeuw, J. & Van der Heijden, P.G.M. (1988), Correspondence analysis of incomplete contingency tables, *Psychometrika*, in press.

De Leeuw, J. & Van Rijckevorsel, J.L.A. (1988). Beyond homogeneity analysis.In J. van Rijckevorsel & J. de Leeuw , *Progress in Component and Correspondence Analysis*. New York, Wiley.

Dijkstra, T. (1983). Some comments on maximum likelihood and partial least squares methods. *Journal of Econometrics*, 22, 67-90.

Elffers, H., Bethlehem, J., & Gill, R. (1980). Monopolie posities horen in de wetenschap niet thuis. (Monopolies do not belong in science). *Bulletin VVS*, 13(6), 14-19.

Freedman, D.A. (1987). As Others See Us: A Case Study in Path Analysis. *Journal of Educational Statistics*, 12, 101-129.

Gifi, A. (1980). Data analyse en statistiek (Data analysis and statistics). *Bulletin VVS*, 13(5), 10-16.

Gifi, A. (1988). *Nonlinear Multivariate Analysis*. Leiden, DSWO-Press.

Gilula, Z., & Haberman, S.J. (1986) Canonical analysis of contingency tables by maximum likelihood. *Journal of the American Statistical Association*, 81, 780-788.

Goodman, L.A. (1986). Some useful extensions of the usual correspondence analysis approach and the usual loglinear models approach in the analysis of contingency tables. *International Statistical Review*, 54, 243-309.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. New York, Academic Press.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology*, 8, 65-82.

Guttman, L. (1959) Metricizing rank-ordered or unordered data for a linear factor analysis. *Sankhya*, 21, 257-268.

Guttman, L. (1971) Measurement as structural theory. *Psychometrika*, 36, 329-347.

Hirschfeld, H.O. (1935). A connection between correlation and contingency. *Proceedings of the Cambridge Philosophical Society*, 31, 520-524.

Isserlis, L. (1916) On certain probable errors and correlation coefficients of multiple frequency distributions with skew regression. *Biometrika*, 11, 185-190.

Kiiveri, H. T. (1987) An incomplete data approach to the analysis of covariance structures. *Psychometrika*, 52, 539-554.

Koyak, R.A. (1987). On measuring internal dependence in a set of random variables. *Annals of Statistics*, 15, 1215-1229.

Lebart, L., Morineau, A., & Warwick, K.M. (1984). *Multivariate Descriptive Statistical Analysis.* New York, Wiley.

Little, R.J.A., & Rubin, D.B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *American Statistician*, 37, 218-220.

Little, R.J.A., & Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York, Wiley.

Molenaar, I. (1988). Formal versus informal methods in data analysis. *Statistica Neerlandica*, 42, in press.

Mooijaart, A., Meijerink, F., & De Leeuw, J. (1988). Nonlinear path models. Submitted for publication.

Muthen, B. (1984) A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika*, 49, 115-132.

Pearson, K. (1906). On ceratin points connected with scale order in the case of a correlation of two characters which for some arrangement give a linear regression line. *Biometrika*, 5, 176-178.

Steiger, J.H. & Browne, M.W. (1984). The comparison of independent correlations between optimal linear composites. *Psychometrika*, 49, 11-24.

Takane, Y., Young, F.W. & De Leeuw, J. (1979), Nonmetric common factor analysis: an alternating least squares method with optimal scaling features, *Behaviormetrika*, 6, 45-56.

Tenenhaus, M. & Young, F.W. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, 91-119,

Van der Burg, E. , De Leeuw, J., & Verdegaal, R. (1988), Homogeneity analysis with k Sets of Variables, *Psychometrika*, in press.

Van der Heijden, P.G.M. & De Leeuw, J. (1985), Correspondence Analysis used Complementary to Loglinear Analysis, *Psychometrika*, 50, 429-447.

Van der Heijden, P.G.M. & De Leeuw, J. (1989), Correspondence Analysis with special attention to the analysis of panel data and event history data. *Sociological Methodology*, in press.

Van Praag, B.M.S., De Leeuw, J., & Kloek, T. (1986). The Population Sample Decomposition Approach to Multivariate Estimation Methods. *Applied Stochastic Models and Data Analysis*, 2, 99-120.

Van Rijckevorsel, J.L.A. (1987). *The Application of Horseshoes and Fuzzy Coding in Multiple Correspondence Analysis*. Leiden, DSWO-Press.

Van Rijckevorsel, J.L.A. & De Leeuw, J. (eds, 1988). *Progress in Component and Correspondence Analysis*. New York, Wiley.

Winsberg, S. & Ramsay, J.O. (1980). Monotonic transformations to additivity using splines. *Biometrika*, 67, 669-674.

Young, F.W. (1981). Quantitative analysis of qualitative data. *Psychometrika*, 46, 357-388.

Figure captions

Figure 1

A model $\Sigma(\theta)$, the truth Σ_0, two observed covariance matrices S_n and T_n, and the influence of projection.

Figure 2a

The regression functions in the GALO data for Teacher's Advice on Father's Profession, with equal interval scoring.

Figure 2b

The regression functions in the GALO data for Teacher's Advice on Father's Profession, with optimal scoring.

TABLE 1

Galo Solution with Equally Spaced Scores

SQUARED CORRELATIONS

+1.0000	+0.0526	+0.0058	+0.0173
+0.0526	+1.0000	+0.0862	+0.5767
+0.0058	+0.0862	+1.0000	+0.1291
+0.0173	+0.5767	+0.1291	+1.0000

CORRELATION RATIOS

+1.0000	+0.0526	+0.0058	+0.0173
+0.0564	+1.0000	+0.0917	+0.6118
+0.2116	+0.1295	+1.0000	+0.1528
+0.0438	+0.5803	+0.1438	+1.0000

LOSS

+0.3619165862

NORMALIZED VALUES

SEX	-0.9846	+1.0156						
IQ	-2.6053	-1.9283	-1.2512	-0.5742	+0.1029	+0.7799	+1.4570	+2.1340	+2.8110
FATHER	-2.1280	-1.3865	-0.6449	+0.0966	+0.8381	+1.5796			
ADVICE	-1.9299	-1.3020	-0.6741	-0.0462	+0.5817	+1.2096	+1.8374		

TABLE 2

Galo Solution with Multiple Correspondence Analysis Scores

SQUARED CORRELATIONS

+1.0000	+0.0507	+0.0690	+0.0234
+0.0507	+1.0000	+0.1292	+0.6181
+0.0690	+0.1292	+1.0000	+0.1451
+0.0234	+0.6181	+0.1451	+1.0000

CORRELATION RATIOS

+1.0000	+0.0507	+0.0690	+0.0234
+0.0564	+1.0000	+0.1328	+0.6223
+0.2116	+0.1329	+1.0000	+0.1616
+0.0438	+0.6315	+0.1514	+1.0000

LOSS

+0.2164122652

NORMALIZED VALUES

SEX	-0.9846	+1.0156						
IQ	-1.1340	-1.2521	-1.0479	-0.7426	-0.1527	+0.7697	+1.7651	+2.4577	+2.5277
FATHER	-2.4462	-0.5214	+1.1264	-0.7368	+0.8619	+1.9354			
ADVICE	-1.0871	-1.0040	-0.7336	+0.6796	+0.4266	+0.1973	+2.1027		

TABLE 3

Galo Solution with Scores Linearizing the Regressions

SQUARED CORRELATIONS

+1.0000	+0.0506	+0.0961	+0.0223
+0.0506	+1.0000	+0.1202	+0.6252
+0.0961	+0.1202	+1.0000	+0.1258
+0.0223	+0.6252	+0.1258	+1.0000

CORRELATION RATIOS

+1.0000	+0.0506	+0.0961	+0.0223
+0.0564	+1.0000	+0.1244	+0.6277
+0.2116	+0.1321	+1.0000	+0.1577
+0.0438	+0.6330	+0.1346	+1.0000

LOSS

+0.2099159169

NORMALIZED VALUES

SEX	-0.9846	+1.0156						
IQ	-1.2258	-1.2695	-1.0334	-0.7221	-0.1636	+0.7352	+1.7862	+2.5037	+2.6180
FATHER	-2.6142	-0.2722	+1.0436	-0.8474	+0.7710	+1.9489			
ADVICE	-1.0809	-1.1083	-0.6846	+0.4529	+0.3981	+0.4006	+2.117		