171

MULTIVARIATE ANALYSIS WITH OPTIMAL SCALING

Jan de Leeuw

Department of Data Theory
Leiden University
Middelstegracht 4
2312 TW Leiden
The Netherlands

Abstract

Suppose $K_1,\ldots,K_m$ are convex cones in a Hilbert space H, with unit sphere S and inner product $(.|.)$. For a particular choice of quantifications, transformations, or representations of a variable x_j in K_jS we can compute the correlation matrix $R(x_1,\ldots,x_m)$ by the rule $r_{ij}(x_1,\ldots,x_m) = (x_i|x_j)$. Now suppose f is a real-valued objective function, defined on the space of all correlation matrices. In this paper we study the class of techniques that choose the x_j in their feasible regions K_jS in such a way that $f(R(x_1,\ldots,x_m))$ is maximized. We discuss typical special cases, including linear and nonlinear principal component analysis, canonical correlation analysis, regression analysis. It is shown that correspondence analysis and the Breiman-Friedman ACE-methods are both special cases of this class of techniques. We discuss some choices for the cones K_j, and we indicate that the results simplify greatly if all bivariate regressions can be linearized. A class of iterative projection techniques is suggested, that produces convergent algorithms of simple structure.

Keywords

Optimal scaling, correspondence analysis, transformation, analysis of categorical variables.

Paper presented at International Conference on Advances in Multivariate Statistical Analysis, Indian Statistical Institute, Calcutta, India, december 16-20, 1985.

1: Introduction

In this paper we shall be concerned with multivariate analysis (from now on MVA) techniques that can be described as follows. We start with a number of variables. A variable is a function on a space of objects. Sometimes, but not always, the space of objects is a probability space, and the variables are random variables. If there is only a finite number of objects, and probability is defined by counting the number of elements, then random variables can be identified simply with arbitrary vectors of real numbers, of length equal to the number of objects. We use the variables, whatever their nature and origin, to compute correlations. With m variables this defines an m x m correlation matrix. The next step in our MVA is to establish how 'good' this correlation matrix is, according to some criterion or aspect. There are many possible definitions of good, corresponding with many different aspects that can be defined in terms of correlation matrices. Many of them will be reviewed below.

Different MVA techniques are associated with different aspects of correlation matrices. In multiple regression, for instance, in which we try to predict variable 1 from variables 2,...,m, the aspect we are interested in is the multiple correlation coefficient. In principal component analysis we are interested in one or several of the eigenvalues of the correlation matrix, in canonical correlation analysis we look at the canonical correlations, and in many multinormal likelihood procedures we study the determinant of the correlation matrix. Of course in computing these various criteria there can be many interesting by-products, such as component loadings, regression coefficients, and so on, but they usually do not play the part of the criterion.

Thus it is clear what an MVA technique consists of. We compute the intercorrelations of the variables, and we evaluate the correlation matrix by computing the aspect defining the technique. Thus we impose several very stringent restrictions on the class of techniques we are studying. In the first place we do not look at MVA techniques which use higher order moments than the second (or higher order marginals, for that matter). Means and variances are also not relevant. Because we look at correlations only our MVA techniques only study aspects of the variables which are invariant under separate linear transformations.

After defining the class of techniques we are interested in, we now come to the non-classical part of the paper. Suppose there is missing information, so that we cannot compute the elements of the correlation matrix, or at least not all of them. Such missing information can, again, be of various types. We can have uncorrelated measurement error, for instance, which has been studied a great deal in psychometrics and econometrics. As a consequence of this measurement error the diagonal elements of the correlation matrix of the true parts of the variables is unknown, the off-diagonal elements are equal to the observed correlations. It is clear that the aspect of the correlation matrix we are studying with our MVA technique may vary with the choice of diagonal element. For principal component analysis and regression this variation has been studied extensively. A recent review of the results that have been obtained is Bekker and De Leeuw (1985). There are other types of missing information, obviously. Some variables may not have values for some objects, for whatever reason. Filling in these missing values also influences the aspects. There is a very voluminous literature dealing with this particular type of variation. We mention only the recent review papers of Little (1982), Little and Rubin (1983), and Titterington and Jiang (1983), which are especially relevant. The problem of uncorrelated measurement errors, mentioned above, is closely related to the problem of latent variables, which is also familiar from psychometrics, econometrics, and system theory. A latent variable is, in a sense, a variable which is missing completely. We do not know anything about it, we only know its location in the system of variables, i.e. we know its role in the criterion we are computing. Related to latent variables is the theory of optimal scaling, which deals with variables about which there is partial knowledge. For instance we know how a variable orders the objects, but we do not know the precise numerical values. The theory of latent variables was reviewed recently by Aigner et. al. (1983) and by Bentler and Weeks (1982). Optimal scaling theory is in Young (1981), Gifi (1981), De Leeuw (1984c).

In this paper we shall discuss a very general approach to optimal scaling theory, in particular to algorithm construction. The work is inspired by the Gifi-system, discussed in Gifi (1981) and De Leeuw (1984c) but the presentation is much more in the tradition of classical multivariate analysis. Our techniques have numerous special cases, some of them new but a lot of them already quite old. Among the more interesting special cases we find correspondence analysis (Greenacre, 1984) and ACE (Breiman and Friedman, 1985). It is hoped that our presentation has a unifying effect on the development of this class of techniques, comparable perhaps to the effect of the Dempster, Laird, and Rubin paper about the EM-algorithm (1977).

2: Framework, notation

The variables are elements of a Hilbert space H. The inner product in H is written as $(x|y)$, and the norm as $||x||$. The correlation between variables x and y is $r(x,y)$, defined by $r(x,y) = (x|y)/(||x||.||y||)$. Thus we tacidly assume that the variables are in 'deviations from the mean'. If $x_1,\ldots,x_m$ are elements of H, then $R(x_1,\ldots,x_m)$ is the correlation matrix. It has elements $r_{ij} = r(x_i,x_j)$. It is clear that we can only compute a correlation between nonzero variables. If x and y are in S, the unit ball in H, then $r(x,y)$ is simply $(x|y)$. The two most obvious interpretations of our general symbolism are mentioned briefly here. For ordinary matrices, with n rows and m columns, the space H is the (n-1)-dimensional subspace of R^n of all centered vectors. The columns of the data matrix are the elements of H. If $x_1,\ldots,x_m$ are random variables, they are elements of $L_2(A,B,p)$, the space of all random variables on the probability space (A,B,p) with finite variances. Again we suppose the variables are centered, i.e. they have zero expectation. In the first example $(x|y)$ is the usual inner product $x'y$, in the second example $(x|y)$ is the covariance of x and y.

The aspect of the correlation matrix that we are studying is a real valued function f, defined on the space of all correlation matrices, which is a compact convex subset of the space of all m x m matrices. Actually it is a subset of the $\frac{1}{2}m(m-1)$ dimensional subspace of all symmetric m x m matrices with unit diagonal. In fact it is the intersection of this subspace and the closed convex cone of all m x m positive semidefinite matrices. It is clear that we can study the variation of f over all correlation matrices, but this will usually not be very interesting. It is not the case that we have no information at all about our variables, and the information we have restricts the set of feasible correlation matrices that we have to consider.

In this paper we suppose that the information we have is of the form that variable x_j is in a known closed convex cone K_j. Thus in our MVA problems we know $K_1,\ldots,K_m$, and we are interested in the variation of $R(x_1,\ldots,x_m)$ if x_j varies in K_j. Clearly continuity of f is sufficient to guarantee that f attains its extreme values on $R(K_1,\ldots,K_m)$, which is a closed subset of a compact set, and is consequently compact. We shall be particulary interested, in this paper, in the maximum of f over $R(K_1,\ldots,K_m)$. Of course by changing f to -f this covers the minimum as well. Looking for the maximum can be interpreted as looking for the correlation matrix which is best in the aspect we are studying.

We have not said much about the cones K_j so far. In many applications they will be subspaces. If we have a variable with information on some objects missing, then K is the subspace of all centered variables whose nonmissing part is linear with the observed nonmissing part. If we have a latent variable, then K is the whole space. For polynomials and splines we use low-dimensional subspaces. Monotone transformations define polyhedral convex cones in R^n, and more general isotone cones in L_2. General transformations of a random variable define infinite dimensional subspaces of L_2. For variables which are not to be transformed at all the cones are rays through the origin. And so on. We shall encounter various other possibilities in our historical section.

As we have already indicated above choosing an x_j in K_j can sometimes be interpreted as choosing a transformation of an observed numerical variable. In the case of missing data we often speak of imputation of the missing values. In the case of non-numerical (categorical) variables the term quantification is more appropriate. For a variable with k possible values the space of all quantifications is usually a k-1 dimensional subspace of H, but it can also be the cone of isotone functions in this subspace if the categories of the variable are ordered.

3: Some history

In this section we shall mention some of the more important examples of the class of MVA techniques we study. We shall see that some of them are already quite old. Clearly they can differ both in terms of the aspect of the correlation matrix they are studying, and in the types of cones of transformations and quantifications they admit.

For two variables there is only one correlation coefficient, which is then, almost unavoidably, the only aspect we can study. Variation of the correlation coefficient under choice of category quantification was already studied by Pearson, in the case of two categorical variables. References and discussion are in De Leeuw (1983). Further study of the case m = 2, with K_1 and K_2 finite dimensional subspaces of R^n, and with criterion r_{12}, are in Hirschfeld (1935), Fisher (1940), Maung (1941a, 1941b). Extensions to subspaces of L_2 are in Gebelein (1941), Sarmanov (1958a, 1958b), Lancaster (1958), Renyi (1959), Csaki and Fischer (1960a, 1960b, 1963). A much more complete biography of maximal correlation is contained in Lancaster (1969). The data analysis technique known as <u>canonical analysis of contingency tables</u> and also as <u>correspondence (factor) analysis</u> can be interpreted as a method which finds systems of quantifications that give stationary values of r_{12}. We do not even try to review the literature connected with this technique, but we refer the reader to De Leeuw (1973); Nishisato (1980), Gifi (1981), Greenacre (1984), Lebart et. al. (1984). For those of us who read french, we recommend Benzécri et. al. (1973, 1980) and Benzécri (1982).

A natural criterion in the case of one dependent and m-1 independent variables is the multiple correlation. In R^n, with categorical variables, the first instance of this technique is perhaps example 46.2 in Fisher (1938). He suggests the use of 'appropriate' scores for a categorical dependent variable in a simple factorial ANOVA. In the ANOVA context this was generalized to a dependent polyhedral cone by Bradley et. al. (1962) and Kruskal (1965). De Leeuw et. al. (1976) also allowed for cone transformations of the independent variables. Winsberg and Ramsay (1980) used polyhedral cones defined by monotone splines. For the general multiple regression problem Younq et. al. (1976) allowed for the choice of either the isotone cone, the linear ray, or the quantification subspace for all of the variables. Theoretical results for L_2 were obtained already by Sarmanov and Zachariv (1960), but the L_2 problem was first studied from the algorithmic point of view by Breiman and Friedman (1985). They set up their algorithm in terms of the random variables and the subspaces consisting of measurable finite-variance transformations. If translated into the practical R^n context, this leaves far too much freedom. Thus the optimal transformations found by the ACE algorithm are smoothed, which means in most cases transformed linearly. The smoother is usually not defined explicitly in terms of conditions that the optimal transformation must satisfy, which means that the practical optimization problem is not always defined unambiguously.

If m variables enter symmetrically into the criterion, we often deal with a form of principal component analysis. In such cases the aspect is usually defined in terms of the eigenvalues of the correlation matrix. The technique which maximizes the largest eigenvalue of R was introduced, in a somewhat different way, by Guttman (1941). If we look at the other stationary values of this optimization problem we are actually performing a <u>multiple correspondence analysis</u> or <u>homogeneity analysis</u>. This technique is discussed extensively in the correspondence analysis books mentioned above, and also in Hill (1974). Maximizing the sum of the first p eigenvalues was implemented by Roskam (1968) and Kruskal and Shepard (1974). They used cones of isotone transformations in R^n. Young et. al. (1978) used mixed nominal, ordinal, and numerical variables in R^n. This was combined with multiple correspondence analysis into a single principal component algorithm by De Leeuw and Van Rijckevorsel (1980). De Leeuw et. al. (1981) used subspacas defined by B-splines, and Winsberg and Ramsay (1983) used cones of monotone integrated B-splines. Koÿak (1985) extended the Breiman-Friedman ACE-algorithm to maximization of the first p eigenvalues.

Criteria defined in terms of the first p canonical correlations or squared canonical correlations were optimized by Young et. al. (1976) for p = 1 and by Van der Burg and De Leeuw (1983) for general p. Of course this has multiple regression, discriminant analysis, and MANOVA as special cases. Generalization of this approach to K sets of variables, with K arbitrary large, were proposed by Van der Burg et. al. (1984). In all these contributions the problem was formulated in R^n, and the cones were of the mixed type we have already discussed earlier in connection with the Young-De Leeuw-Takane ALSOS series or the Gifi-series of programs. The multiple regression aspect was generalized in another direction by De Leeuw (1984a),who suggested maximizing the sum of the determination coefficients for a given path model, possibly with latent variables.

The determinant of the correlation matrix, which is related to maximum likelihood estimation for the multivariate normal, was maximized over quantification subspaces in Saito (1973, 1974) and in Chang and Bargmann (1974). This was extended to mixed ALSOS-type cones in Kuhfeld et. al. (1985). For completeness we also mention Takane et. al. (1979), who find optimal scalings and modify diagonal elements at the same time to fit the factor model. Also compare Mooijaart (1984) in this context.

It is clear from this historical overview of the literature that many aspects have been studied before. The results are scattered over the theoretical and applied literature, and many more relevant references could indeed be given. Gifi (1981) has a very complete bibliography. It is not entirely clear what the relationship of the various techniques is, and how they are related to classical statistical theory. The developments, especially those in psychometrics, are strongly algorithm centered and few theoretical results are available. Thus the field is in a somewhat disorganized state, dominated by *ad hoc* porposals and solutions. The ACE procedures of Breiman, Friedman, and their students are, in some respects at least, better imbedded in mathematical theory. In other respects they are simply another parallel development. A superficial perusal of the references shows that there are various series of programs involved. The Bell-system, created by Kruskal and Shepard in the early sixties, was the first. Roskam created his own series of programs in 1968. In the early seventies Young, De Leeuw, and Takane started their ALSOS series, and in the late seventies the Gifi system got under way. In the early eighties these series were joined by Winsberg and Ramsay, who used monotone splines and likelihood-derived criteria, and by Breiman and Friedman, who used conditional expectations and worked in L_2. In most of these series the developments are started with the linear model, and very soon after that principal component analysis follows. This is not a surprising or unfortunate way to proceed, but it does mean that history seems to repeat itself many times in this field.

4: Necessary conditions for an extremum

The problem we study in this paper is to maximize $f(R(x_1,\ldots,x_m))$ over x_1 in K_1, ... , x_m in K_m. In case of infinite-dimensional H some care is needed to establish the existence of the required maxima, but usually it is not difficult to show existence. Compare Breiman and Friedman (1985), or Koyak (1985), for instance, and the many references connected with maximal correlation theory. Even if we cannot show that a maximum exists, we could still be interested in the supremum, or in the stationary values of the criterion.

In this section we study the stationary equations, assuming that f is differentiable. We write H* for the dual of H. Moreover x_j^* is the element of H* for which $x_j^*(x) = (x|x_j)$ for all x in H. Let us write g_{ij} for the partial derivative of f with respect to r_{ij}. Let us know evaluate the partials in a point $(x_1,\ldots,x_m)$ with x_j in S for all j. Then

$$\frac{\partial f}{\partial x_k} = \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\partial f}{\partial r_{ij}}\{\frac{\partial(x_i|x_j)}{\partial x_k} - r_{ij}(\frac{\partial||x_i||}{\partial x_k} + \frac{\partial||x_j||}{\partial x_k})\} =$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{m} g_{ij}\{(\delta^{jk}x_i^* + \delta^{ik}x_j^*) - r_{ij}(\delta^{ik}x_i^* + \delta^{jk}x_j^*)\}.$$

Define

$$z_k^* = \sum_{i=1}^{m} g_{ik}x_i^*,$$

and

$$\mu_k = \sum_{i=1}^{m} g_{ik}r_{ik}.$$

Then

$$\frac{\partial f}{\partial x_k} = 2(z_k^* - \mu_k x_k^*).$$

It follows that the necessary condition for an extreme value is that $z_j^* - \mu_j x_j^*$ is in K_j^*, the cone dual to K_j, for all j. Or, to put it differently, that

$$(z_j|x) - \mu_j(x_j|x) \leq 0$$

for all x in K_j. Observe that $(z_j|x_j) - \mu_j(x_j|x_j) = 0$ for all j. If K_j is a subspace, then K_j^* is the orthogonal complement. There is a much more convenient way to write the necessary condition. Suppose P_j projects on K_j. Then it is well known that $\mu_j x_j$ is the projection of z_j on K_j if and only if $(z_j - \mu_j x_j|x) \leq 0$ for all x in K_j. Consequently our necessary condition can be written as

$$P_j(z_j) = \mu_j x_j.$$

Another way to write this equation is by defining

$$\tilde{z}_j = \sum_{i \neq j} g_{ij} x_i,$$

$$\tilde{\mu}_j = \sum_{i \neq j} g_{ij} r_{ij}.$$

Then the necessary condition is

$$P_j(\tilde{z}_j) = \tilde{\mu}_j x_j.$$

Of course we also have to remember that x_j is in K_jS, the intersection of K_j and S.

5: An obvious algorithm

A nice thing about the necessary conditions we have derived in the previous section, is that they immediately suggest an algorithm. The algorithm we mean starts with feasible estimates x_j, in K_jS, and improves them one by one.

```
A1: compute $\tilde{z}_j$,
A2: compute $\tilde{x}_j = P_j(\tilde{z}_j)$,
A3: compute $x_j = \tilde{x}_j/||\tilde{x}_j||$,
A4: if $j < m$ then $j \leftarrow j + 1$ and go to A1,
A5: if $j = m$ then $j \leftarrow 1$ and go to A1.
```

The algorithm is a bit peculiar, because it neither starts nor stops, but the meaning is probably clear. There is, of course, no guarantee yet that the algorithm converges. Its behaviour will probably depend on the aspect we are maximizing and on the nature of the cones K_j. It is clear, however, that the algorithm is conceptually quite simple, because it only involves projecting the <u>targets</u> $\tilde{z}_j$ on the convex cones K_j.

We shall now discuss some simple rules for computing targets, corresponding with some of the more familiar criteria. The first one we study is the multiple correlation, or rather its square. If r is the vector of correlations between the first variable and the m-1 remaining ones, and R_1 is the correlation matrix of the last m-1 variables, then

$$1 - \rho^2_{\mathrm{mult}} = \min_{\beta} 1 - 2\beta' r + \beta' R_1 \beta.$$

This representation immediately shows two things. One minus the souared multiple correlation (SMC) is a concave function of the correlation matrix, because it is the minimum of a family of linear functions. Moreover the matrix of partials G has the form

$$G = \begin{vmatrix} 1 & -\beta' \\ -\beta & +\beta\beta' \end{vmatrix},$$

with $\beta = R_1^{-1} r$. The SMC itself is convex, and its partials are -G.

Thus, giving β elements $\beta_2,\ldots,\beta_m$,

$$\tilde{z}_1 = \sum_{j=2}^{m} \beta_j x_j,$$

$$\tilde{z}_j = \beta_j\{x_1 - \sum_{\substack{i=2\\ i\neq j}}^{m} \beta_i x_i\} \text{ for } j=2,\ldots,m.$$

Algorithm A in this case extends the ACE-algorithm to cones of transformations, and it extends the ADDALS, MORALS, MONANOVA algorithms of Young, Gifi, and Kruskal to Hilbert space.

Let us now look at the sum of the first p eigenvalues. In this case we write

$$\lambda_1 + \ldots + \lambda_p = \max_K \text{tr } K'RK,$$

where K varies of the n x p matrices satisfying $K'K = I$. Thus again the aspect is convex, and the partials are simply $G = KK'$. If $Z = XG = XKK'$, then Z is the best rank p approximation to X in the least squares sense. Algorithm A in this case is very similar to, but not identical with, the algorithms PRINCIPALS, PRINQUAL, PRINCALS suggested by Young, Gifi, and others.

As a final example, in this round, we take the determinant. We know, for example from multinormal maximum likelihood theory, that

$$\ln |R| = \min_{S>0} \ln |S| + \text{tr } S^{-1}R - m.$$

Thus $\ln |R|$ is concave, and its partials are $G = R^{-1}$. Thus $-\ln |R|$ is convex, with partials $-R^{-1}$. The target for variable 1 is, in our previous notation,

$$\tilde{z}_1 = (1 - \rho^2_{\text{mult}})^{-1} \sum_{j=2}^{m} \beta_j x_j.$$

If we call, following Guttman, the best least squares prediction of a variable from the remaining variables its image, then the target is equal to the image. We modify a variable by projecting its image on the cone. This is identical to the PRINQUAL minimûm determinant algorithm proposed by Kuhfeld et. al. (1985).

Although our examples show that the general algorithm A specializes to various known algorithms, they do not show in any sense that the algorithm really works (i.e. converges). This question will be studied in the next section.

6: Convergence under convexity assumptions

In the previous section we have demonstrated that some of the well-known aspects, such as the SMC, the sum of the largest eigenvalues, and the negative logarithm of the determinant, are convex. In this section we study our algorithm for the general class of convex aspects of the correlation matrix, and we prove that for this class of aspects, it is convergent.

Suppose f is convex. For the moment we also assume that f is continuously differentiable, but we shall see that this assumption is not really necessary. If R and S are two correlation matrices, then convexity tell us that

$$f(R) \geq f(S) + \text{tr}\, G_S(R - S),$$

with G_S the matrix of partials evaluated in S. Observe that if f is not only convex but also homogeneous, then this inequality simplifies to

$$f(R) \geq \text{tr}\, G_S R.$$

Now identify S with the correlation matrix at the start of the iteration, and suppose we want to modify x_1. We do this by maximizing tr $G_S R$ over x_1 in $K_1 S$. This gives R^+, which differs from S only in its first row and column. Because of convexity

$f(R^+) \geq f(S) + \mathrm{tr}\, G_S(R^+ - S)$,

and because of the maximizing property of R^+

$\mathrm{tr}\, G_S R^+ \geq \mathrm{tr}\, G_S S$.

Combining the two inequalities gives

$f(R^+) \geq f(S)$.

If f is strictly convex, then the subdifferential inequality is strict, and the algorithm increases f, if it changes anything at all. Even if f is simply convex, this remains true, as we now prove.

Maximizing tr $G_S R$ over x_1 in $K_1 S$ can be done by observing that, if all x_j have unit length,

$$\mathrm{tr}\, G_S R = \sum_{i=1}^{m} \sum_{j=1}^{m} g_{ij}(x_i|x_j) = \sum_{i=1}^{m} g_{ii} + \sum_{i=1}^{m} (x_i|\tilde{z}_i).$$

Thus we must maximize $(x_1|\tilde{z}_1)$ over $K_1 S$. If z_1 is in K_1^*, then $P_1(\tilde{z}_1) = 0$. If we assume that this polar situation does not occur, then the optimal x_1 is computed simply by normalizing $P_1(\tilde{z}_1)$. And this gives precisely algorithm A, discussed in the previous section. Because cone projection is unique, it follows that the aspect increases if the algorithm changes x_1. Thus it will centainly increase if we loop over x_j, and something is changed along the way. If nothing is changed, we have found a stationary $(x_1,\ldots,x_m)$, satisfying the necessary conditions for an extremum.

We have shown that if we are maximizing a convex aspect, and if z_j is during the course of the algorithm never in the polar K_j^*, then the algorithm produces an increasing sequence of aspect values, which consequently converges for an aspect which is bounded on the set of correlation matrices. This convergence is global, i.e. it occurs from any initial point. We have not proved, of course, that the sequence of transformed variables converges. For this we need to consider the existence question. What we can easily show, using general convergence theory as explained by Zangwill (1969) for instance, is that each accumulation point of the sequence of transformations is a point satisfying the necessary conditions, and that all accumulation points of the sequence have the same aspect value. Thus all correlation matrices that are accumulation points have the same aspect value, and there is at least one accummulation point in the sequence of correlation matrices generated by the algorithm. We shall, however, not study the convergence question in depth here, and neither will we discuss what should be done if $\tilde{z}_j$ wanders into the polar cone. We still have to point out, however, that our convergence proof also applies if the aspect is convex and not differentiable, because the subgradient inequality continues to apply. Compare Rockafellar (1971, part V) for details.

7: Further simplification of the algorithm

Our convergence proof we based on the general idea of majorization or minorization. In maximizing a function of a vector variable we construct an auxilary function of two vector variables, which lies below the first function and touches it if the two arguments are equal. A step of the algorithm then consists in maximizing the auxilary function over its second argument, with the first argument fixed at the current value. This is a very useful class of algorithms, because they have the property of global convergence. Majorization algorithms have been used earlier in multidimensional scaling (De Leeuw, 1977, De Leeuw and Heiser, 1980) and in maximum likelihood estimation (Dempster et. al., 1977).

Although the algorithm is simple, it may still involve a lot of computation in each cycle. Remember that we first compute the target for variable 1, then we modify variable 1 by projection, then we compute the target for variable 2, and so on. Computing the target for the second variable involves the new transformation of the first variable, which means that in general all partial derivatives must be recomputed. In many situations it would be desirable to have an algorithm which recomputes all m targets, then performs all m projections, and so on. Developments in computing make it highly desirable to have really simple algorithms, with a lot of high-level matrix manipulation, even if they converge perhaps a bit slower. Thus we suggest another algorithm, which differs from A in its loop-structure.

```
B1: Compute $Z = XG$,
B2: compute $\tilde{x}_j = P_j(z_j)$;
B3: compute $x_j = \tilde{x}_j/||\tilde{x}_j||$,
B4: if $j < m$ then $j \leftarrow j + 1$ and go to B2,
B5: if $j = m$ then go to B1.
```

Our previous results on convergence do not apply directly, but we can modify them in such a way that they do.

For this we first make the additional assumption that f is <u>monotone</u> in the sense that $f(R + S) \geq f(R)$ for all $S \geq 0$, i.e. for all positive semi-definite S. For the partials G this implies that $G \geq 0$. We now apply the minorization method a second time. Remember that in our first use of minorization f was minorized by a function linear in R, which was the tangent line in S, the current solution. A function linear in R is quadratic in the normalized x_j. If $G \geq 0$, then this quadratic is convex, and it can be minorized by a function linear in the x_j, which is again the tangent in the current solution. Thus tr G_SR, which was the first auxilary function we constructed, is now minorized by using

$$\text{tr } G_SR = \sum_{i=1}^{m}\sum_{j=1}^{m} \overline{g}_{ij}(x_i|x_j) \geq \sum_{i=1}^{m}\sum_{j=1}^{m} \overline{g}_{ij}(\overline{x}_i|\overline{x}_j) + 2\sum_{i=1}^{m}\sum_{j=1}^{m} \overline{g}_{ij}(\overline{x}_i|x_j - \overline{x}_j) =$$

$$=2\sum_{j=1}^{m} (x_j|\overline{z}_j) - \text{tr } G_SS.$$

Here $\overline{x}_j$ are the current normalized transformations, such that $s_{ij} = (\overline{x}_i|\overline{x}_j)$, and G is also evaluated at $(\overline{x}_1,\ldots,\overline{x}_m)$. We can write, slightly abusing matrix notation in the case of infinite-dimensional H, $\overline{Z} = \overline{XG}$, and in this new minorization step we find that it suffices to maximize tr $X'\overline{Z}$ over X in their cones. This exactly defines algorithm B, which is consequently globally convergent in the same sense as A.

We go back to our examples, and introduce some new ones. For the SMC we have found that $G \leq 0$, and the same thing is true for the determinant. Thus algorithm B cannot be used for these criteria. For the sum of the p largest eigenvalues however, we have already seen that $G = KK'$, and that $Z = XG = XKK'$ is the best rank-p approximation to X. In this case algorithm B becomes identical to the usual nonlinearprincipal component algorithms, which alternate cone projection and rank-p approximation. Comoare the references below.

We briefly discuss some other examples. If f is the sum of the r_{ij}^s, with s a positive integer, then f is convex if s is even, and f is monotone by Schur's theorem on powers of positive semi-definite matrices (Styan, 1973). Obviously $G = sR^{(s-1)}$, and thus Z can be taken as $XR^{(s-1)}$. For s = 1 this gives a very interesting algorithm. Then G = uu', with u a vector with m ones, and thus Z = Xuu' is formed from X by replacing each element by its row sum. All columns of Z are the same, but because they are subsequently projected on different cones they will differ again in the next X. This is perhaps the simplest algorithm in this class. It maximizes the sum of the correlation coefficients. For s = 2 we maximize the sum of squares of the correlation coefficients. This is equal to the sum of squares of the eigenvalues of R, and consequently amounts to the same thing as maximizing the variance of the eigenvalues of R. Here the target is Z = 2XR. A slightly different theory results if f is the sum of $|r_{ij}|^s$, which is always convex, but need not be monotone.

It seems somewhat unfortunate that algorithm B cannot be applied to the SMC and the determinant, which were convex but not monotone. There is a simple adaptation of B, called algorithm C, which can be applied. It is based on the fact that the diagonal of R is fixed, and that consequently maximizing tr $G_S R$ over R amounts to the same thing as maximizing tr $(G_S + \Omega)R$, with Ω a diagonal matrix which can depend on G. Now it is obviously always possible to choose Ω in such a way that $G_S + \Omega \geq 0$, and thus the algorithm works with $Z = X(G_S + \Omega)$. The effect of using Ω is making the algorithm more conservative and slow, and thus one really should try to choose Ω as small as possible. For the determinant $G = -R^{-1}$

we can take $\Omega = \lambda I$, with $\lambda \geq \lambda_{\min}^{-1}(R)$. Especially for nearly singular R this may make the algorithm hopelessly slow. For the SMC we can add $1 + \beta'\beta$ to the diagonal of G, which is not such a major modification in general.

8: Limitations

Algorithms A, B, and C are only guaranteed to work for convex functions of the correlation matrix, and B only works for monotone convex functions. We have seen that the class of aspects that are covered by these constraints is quite large. It can be made larger by various arguments. Adding various aspects, for instance, gives a new one which also satisfies the constraints. This is important, for example, in path analysis, in which the aspect is the sum of various SMC's, one for each endogeneous variable or equation. Other combination rules for convex functions can also be used.

It remains true, however, that some criteria that have been proposed in the optimal scaling literature, notable those based on canonical correlations, are not convex. In such cases we can do various things. We can develop special purpose algorithms, for instance by using the altermating least squares method. We can also go ahead with algorithm A, possibly with some safe-guards or ad hoc step-size procedures build in, because we expect it to converge at least locally. Tijssen (1985) applies algorithms similar to A in such an ad hoc way, with apparently quite satisfactory results.

There are also cases, of course, in which we may want to optimize aspects of the data which are not even functions of the correlation coeficients, or which are not solely functions of the correlation coefficients. We may want to minimize skewness, for example, or we may want to minimize the difference between correlation ratios and correlation coefficients (Bekker, 1982). For such aspects we have to develop other algorithms, either by using minorization or by using other techniques for algorithm construction.

Even in the cases in which the algorithm is globally convergent the convergence may still be too slow for practical purposes. This sometimes happens with the EM algorithm, and with the multidimensional scaling methods as well. Convergence can be to a local maximum as well. And there is the possibility of nonconvergence of the transformations, either because there are no accumulation points (which can happen in infinite dimensional cases) or because there is a continuum of accumulation points (which can even happen in finite dimensional situations). Additional research is still needed to monitor the progress of the algorithm, and to take care of various undesirable events which may happen along the way. Comparison with alternative computational methods is also needed.

For completeness we list one natural candidate for comparison. This is algorithm D, the ACE-version of algorithm A. Between steps A2 and A3 we insert the step $\tilde{x}_j \leftarrow \text{SMOOTH}(\tilde{x}_j)$, where the choice of smoother is left free. Often it will be linear, compare Breiman and Friedman (1985). In the usual ACE-implementations so far the cones K_j are very large, and the restrictions are imposed by smoothing. In the ALSOS and Gifi-series the cones are much smaller, and this takes care of the smoothing. As a consequence the ALSOS and Gifi-transformation tend to be more rigid, but the ACE-method in the finite dimensional case does not always solve a clearly defined optimization problem, and may have difficulties with convergence (or with proving convergence).

Enough about algorithms. Back to theory.

9: The importance of linear regressions

Let us go back to the necessary conditions in section 4. Suppose that the cones are linear subspaces, and that consequently the projectors P_j are linear. The stationary equations are

$$\sum_{i=1}^{m} g_{ij}P_j(x_i) = \mu_j x_j,$$

with x_j in K_jS. Now suppose $(x_1,\ldots,x_m)$ are such that all bivariate regressions are linear. This means that for all i,j we have

$$P_j(x_i) = t_{ij}x_j.$$

If we substitute this in the stationary equations we find that these are satisfied with

$$\mu_j = \sum_{i=1}^{m} g_{ij}t_{ij}.$$

This result is, of course, completely independent of G, and thus of the aspect we are optimizing. For all aspects which are functions of the correlation coefficients a system of transformations or quantifications that linearizes all bivariate regressions gives a solution to the stationary equations. This does not prove that a linearizing system always gives the maximum, in fact it need not do this at all.

We illustrate this with an example which is quite important from the theoretical point of view. Suppose we have an m-variate standard normal distribution, and K_j are the separate transformations of the variables with finite variance. Take x_i equal to the Hermite-polynomial of degree s in the subspace K_i. Then $P_j(x_i) = \rho_{ij}^{s}x_j$, with ρ_{ij} the correlation parameter of the multinormal. Thus the stationary equations are satisfied. It has been shown by Kolmogorov (1960), compare also Venter (1966), that the linear transformation optimizes the first canonical correlation, and consequently also the SMC. Gifi (1981) shows that the linear transformation optimizes the determinant and the largest and smallest eigenvalue, Koyak (1985) shows that in fact the linear polynomials maximize the sum of the p largest eigenvalues. By a familiar theorem of Ky Fan (1951) this means that they maximize all unitarily invariant matrix norms, compare Gifi (1981, page 320).

We can extend the analysis of the multinormal example or gauge a bit, by supposing that the first m_1 variables are transformed linearly and the first m_2 quadratically. Then R is the direct sum of two matrices, and for many criteria such as the determinant, the sum of squares, and the sum of the eigenvalues, G is a direct sum too. Thus the g_{ij} between sets are zero, and the stationary equations are still satisfied for this transformation system too. If all ρ_{ij} are equal to ρ, for instance, then the sum of the two largest eigenvalues is $(m\rho + (1 - \rho)) + (1 - \rho) = 2 + (m - 2)\rho$ if all transformations are linear. If m - 1 transformations are linear, and the remaining one is quadratic or otherwise orthogonal, then the sum of the two largest eigenvalues is $((m-1)\rho + (1 - \rho)) + 1 = 2 + (m - 2)\rho$ as well. The maximum given by the linear transformations need not be unique.

In general the results in this section indicate, that if a linearizing system of transformations exists, then our algorithm is often able to find it. It is shown by De Leeuw (1982) and Bekker (1982) that in many practical examples approximate linearizing systems exists, and that often optimal transformations found by any one of these algorithms are not too far from linear.

10: Statistical stability analysis

We are generally interested in the statistical stability of the transformations and of the correlation matrix that is computed. Of course we can only investigate statistical stability by introducing some kind of probabilistic model. We suppose that we are dealing with the case in which H is a space of random variables. In this case all quantities we compute are functions of the bivariate distributions or the bivariate marginals. The statistical problem arises if we do not know the (population) marginals, but we observe empirical or sample marginals. Every quantity we compute is a function of the sample bivariate marginals. If the quantity is sufficiently smooth we can apply the delta method, or its analogon in infinite dimensional H. But since actual computations are always or discreticized and finite versions, we are able, in many cases, to do the statistics in low-dimensional subspaces.

For ordinary correspondence analysis ($m = 2$, K_1 and K_2 are finite dimensional indicator- or dummy-subspaces) the stability results are well known. For multiple correspondence analysis ($m \geq 2$, K_j dummy subspaces, maximize the largest eigenvalue of the correlation matrix) they are also quite straightforward. De Leeuw (1984b) gives the necessary references. In more complicated cases the derivatives needed for the delta method tend to become unmanagable, and if the cones K_j are polyhedral projections usually are not smooth enough. In such cases we follow Gifi (1981), and we apply the Bootstrap and Jackknife. Computation-oriented methods to test significance of various aspects have been discussed by De Leeuw and Van der Burg (1985).

In the statistical context linearizability of the regressions, discussed in the previous section, is also of great importance. It was shown in De Leeuw (1984b) that if the optimal transformation linearizes all bivariate regressions then we can apply the delta method as if the transformations are actually fixed and known instead of stochastic and unknown. Thus if regressions can be linearized, it follows that we can compute the optimum transformations, and then apply the delta method as in the classical linear MVA techniques. In particular, for mutlivariate normal data, we can first scale them optimally and then apply the usual MVA techniques. This will not differ greatly from applying linear MVA directly. On the other hand if the data are, borrowing a word from Yule, strained normal, i.e. normal except possible for invertible transformations on each of the variables separately, then optimal scaling plus classical normal techniques will still perform nicely. They give consistent estimates of structural parameters, and valid chi square statistics and confidence intervals. If the data come from a strongly non-normal but strained-normal distribution, then directly applying classical techniques may lead to rather serious distortions. There are already quite a number of sucessfull applications of the combination Optimal Scaling plus LISREL or Optimal Scaling plus Factor Analysis in the applied literature.

11: Connection with likelihood theory for the multinormal

The log-likelihood for a multivariate normal sample with covariance matrix C is, except for irrelevant constants, equal to $L = \ln |\Sigma| + \text{tr}\ \Sigma^{-1}C$.

Now let f(C) be the minimum of L over a parametric manifold of matrices $\Sigma(\theta)$, i.e. over some covariance structure model. Because f is the minimum of a family of linear functions it is concave. If Σ is allowed to vary over all $\Sigma \geq 0$, then we already know that $f(C) = \ln |C|$, but for restricted models many other results can be obtained. Think of factor models, path models, simultaneous equation models, LISREL models, and so on. For all models we have $G = \Sigma^{-1}$, of course, which means that f is concave and increasing. It is now easy to think of a transformation technique for correlations. We minimize f(R), obtained in this way, over all correlation matrices R, using algorithm A or C. Observe that we have changed from the dispersion matrix C to the correlation matrix R here, in conformity with our earlier usage, but somewhat opposite to statistical considerations.

Although the approach outlined above may be appealing to some, it is not a maximum likelihood method, and it does not have the properties commonly associated with maximum likelihood. The reason is simple. In likelihood theory we maximize the likelihood of the observed data, not the likelihood of the transformed data. If the model states that some transformation of the observed variables is multivariate normal, with some structure imposed on the covariance matrix, then the likelihood of the observed data involves the usual multinormal component but also the Jacobian of the transformation. Thus the log-likelihood is similar to our previous L, but we have to add a term to it consisting of the logarithm of the derivatives of the transformations (which are supposed to be invertible and continuously differentiable). This transformation likelihood, familiar in more simple cases from Box and Cox (1964), is not a function of the correlations only any more, and our theory does not apply directly.

Nevertheless we can use minorization in this context too, by minorizing the covariance part only. Observe that the Jacobian part of the likelihood function does not involve any structural parameters. We are currently experimenting with an algorithm based on these ideas, which can be used to imbed the optimal scaling approach (usually described as exploratory or descriptive) in a more conventional statistical framework.

REFERENCES

Aigner, D.J., Hsiao, C., Kapteyn, A., & Wansbeek, T. (1983). Latent variable models in econometrics. In Z. Griliches & M. D. Intriligator (eds), Handbook of Econometrics I, Amsterdam: North Holland Publishing Co.

Bekker, P. (1982). Varianten van niet-lineaire principale komponenten analyse. Unpublished Master's Thesis, Department of Psychometrics, Leiden University.

Bekker, P., & De Leeuw, J. (1985). The rank of reduced dispersion matrices. Research Memo 187, Department of Economics, Tilburg University.

Bentler, P.M., & Weeks, D.G. (1982). Multivariate analysis with latent variables. In P.R. Krishnaiah & L. N. Kanal (eds), Handbook of Statistics II. Amsterdam: North Holland Publishing Co.

Benzécri, J.P. (1982). Histoire et Prêhistoire de l'Analyse des Données. Paris, Dunod.

Benzécri, J.P., and collaborators (1973). L'Analyse des Données (2 vols). Paris, Dunod.

Benzécri, J.P., and collaborators (1980). Pratique de l'Analye des Données (3 vols). Paris, Dunod.

Box, G.E.P., & Cox, D.R. (1964). An analysis of transformations (with discussion). Journal of the Royal Statistical Society, B26, 211-256.

Bradley, R.A., Katti, S.K., & Coons, I.J. (1962). Optimal scaling for ordered categories. Psychometrika, 27, 355-374.

Breiman, L., & Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). Journal of the American Statistical Association, 80, 580-619.

Chit-Fu Chang, J., & Bargmann, R.E. (1974). Internal multidimensional scaling of categorical variables. Technical Report 108, Department of Statistics and Computer Science, University of Georgia.

Csaki, P., & Fischer, J. (1960a). Contributions to the problem of maximum correlation. Magyar Tudomanyos Akademia, 5, 325-332.

Csaki, P., & Fischer, J. (1960b). On bivariate stochastic connection. Magyar Tudomanyos Akademia, 5, 313-323.

Csaki, P., & Fischer, J. (1963). On the general notion of maximal correlation. Magyar Tudomanyos Akademia, 8, 27-51.

De Leeuw, J. (1973). Canonical analysis of categorical data. Unpublished Doctoral Dissertation, Leiden University. Re-issued by DSWO-Press, Leiden, 1985.

De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In J.R. Barra et. al. (eds), Progress in Statistics. Amsterdam: North Holland Publishing Co.

De Leeuw, J. (1982). Nonlinear principal component analysis. In H. Caussinus et. al. (eds), COMPSTAT 1982. Vienna: Physika Verlag.

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

De Leeuw, J. (1984a). Least squares and maximum likelihood for causal models with discrete variables. Research Report RR-84-09, Department of Data Theory, Leiden University.

De Leeuw, J. (1984b). Statistical properties of multiple correspondence analysis. Research Report RR-84-06, Department of Data Theory, Leiden UNiversity.

De Leeuw, J. (1984c). The Gifi-system of nonlinear multivariate analysis. In E. Diday et. al. (eds), Data Analysis and Informatics. Amşterdam: North Holland Publishing Co.

De Leeuw, J. & Heiser, W.J. (1980). Multidimensional scaling with restrictions on the configuration. In P.R. Krishnaiah (ed), Multivariate Analysis V. Amsterdam: North Holland Publishing Co.

De Leeuw, J., & Van der Burg, E. (1985). The permutational limit distribution of generalized canonical correlations. Research Report RR-85-04, Department of Data Theory, Leiden University.

De Leeuw, J., & Van Rijckevorsel, J. (1980). HOMALS and PRINCALS: two generalizations of principal component analysis. In E. Diday et. al. (eds), Data Analysis and Informatics. Amsterdam: North Holland.

De Leeuw, J., Van Rijckevorsel, J, & Van der Wouden, H. (1981). Nonlinear principal component analysis with B-splines. Methods of Operations Research, 43, 379-393.

De Leeuw, J., Young, F.W., & Takane, Y. (1976). Additive structure in qualitative data: and alternating least squares method with optimal scaling features. Psychometrika, 41, 471-504.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM-algorithm. Journal of the Royal Statistical Society, B 39, 1-38.

Fisher, R.A. (1938). Statistical methods for research workers. 7th edition. Edinburgh: OLiver and Boyd.

Fisher, R.A. (1940). The precision of discriminant functions. Annals of Eugenics, 10, 422-429.

Gebelein, H. (1941). Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit Ausgleichsrechnung. Zeitschrift für Angewandte Mathematik und Mechanik, 21, 364-379.

Gifi, A. (1981). Nonlinear Multivariate Analysis. Leiden: Department of Data Theory.

Greenacre, M.J. (1984). Theory and applications of correspondence analysis. New York, Academic Press.

Guttman, L. (1941). The quantification of a class of attributes. A theory and method of scale construction. In P. Horst (ed), The prediction of personal adjustment. New York: SSRC.

Hill, M. (1974). Correspondence analysis: a neglected multivariate method. Applied Statistics, 3, 340-354.

Hirschfeld, H.O. (1935). A co-nection between correlation and contingency. Proceedings Cambridge Philosophical Society, 31, 520-524.

Kolmogorov, A.N. (1960). Cited by Sarmanov and Zacharov (1960).

Koyak, R. (1985). Optimal transformations for multivariate linear reduction analysis. Unpublished Ph.D. Thesis, Department of Statistics, University of California, Berkeley.

Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. Journal of the Royal Statistical Society, B 27, 251-263.

Kruskal, J.B., & Shepard, R.N. (1974). A nonmetric variety of linear factor analysis. Psychometrika, 39, 123-157.

Kuhfeld, W.F., Sarle, W.S., & Young, F.W. (1985). Methods of generating model estimates in the PRINQUAL macro. In SAS Institute (ed), SUGI-Proceedings, Cary, N.C., SAS Institute.

Ky Fan (1951). Maximum properties and inequalities for the eigenvalues of completely continuous operators. Proceedings National Academy of Sciences, 37, 760-766.

Lancaster, H.O. (1958). The structure of bivariate distributions. Annals of Mathemtical Statistics, 29, 719-736.

Lancaster, H.O. (1969). The chi-squared distribution. New York: Wiley.

Lebart, L., Morineaux, A., & Warwick, K.M. (1984). Multivariate descriptive statistical analysis. New York: Wiley.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. Journal of the American Statistical Association, 77, 237-250.

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.

Maung, K. (1941a). Measurement of association in a contingency table. Annals of Eugenics, 11, 189-223.

Maung, K. (1941b). Discriminant analysis of Tocher's eye colour data. Annals of Eugenics, 11, 64-76.

Mooijaart, A. (1984). The nonconvergence of FACTALS: a nonmetric common factor analysis. Psychometrika, 49, 143-145.

Nishisato, S. (1980). Analysis of Categorical data: Dual Scaling and its applications. Toronto: University of Toronto Press.

Renyi, A. (1959). On measures of dependence. Acta Mathematica Academia Scientia, 10, 441-451.

Rockafellar, R.T. (1971). Convex Analysis. Princeton: Princeton University Press.

Roskam, E.E.C.I. (1968). Metric analysis of ordinal data in psychology. Voorschoten: VAM.

Saito, T. (1973). Quantification of categorical data using the generalized variance. Soken Kiyo, 13, 61-80.

Saito, T. (1974). Revision and Errata on "Quantification of categorical data using the generalized variance". Soken Kiyo, 14, 113-116.

Sarmanov, O.V. (1958a). The maximum correlation coefficient (symmetrical case). DAN SSSR, 120, 1139-1143.

Sarmanov, O.V. (1958b). The maximum correlation coefficient (non-symmetrical case). DAN SSSR, 121, 52-55.

Sarmanov, O.V., & Zacharov, V.K. (1960). Maximum coefficients of multiple correlation. DAN SSSR, 130, 1960, 269-271.

Styan, G.P. (1973). Hadamard products and multivariate statistical analysis. Linear Algebra and Applications, 6, 217-240.

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.

Titterington, D.M., & Jiang, J.M. (1983). Recursive estimation procedures for missing data problems. Biometrika, 70, 613-624.

Tijssen, R. (1985). A new approach to nonlinear canonical correlation analysis. Unpublished Master's Thesis, Department of Psychometrics, Leiden University.

Van der Burg, E., & De Leeuw, J. (1983). Nonlinear canonical correlation. British Journal of Mathematical and Statistical Psychology, 36, 54-80.

Van der Burg, E., De Leeuw, J., & Verdegaal, R. (1984). Nonlinear canonical correlation with m sets of variables. Research Report RR-84-12, Department of Data Theory, Leiden University.

Venter, J.H. (1966). Probability measures on product spaces. South African Statistical Journal, 1, 3-20.

Winsberg, S., & Ramsay, J.O. (1980). Monotone transformations to additivity. Biometrika, 67, 669-674.

Winsberg, S., & Ramsay, J.O. (1983). Monotone spline transformations for dimension reduction. Psychometrika, 48, 575-595.

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

Young, F.W., De Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative data: an alternating least squares method qith optimal scaling feautures. Psychometrika, 41, 505-529.

Young, F.W., Takane, Y., & De Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features. Psychometrika, 43, 279-281.

Zangwill, W. (1969). Nonlinear programming. Englewood Cliffs, N.J., Prentice Hall.