THE INTEGRATION OF
MULTIDIMENSIONAL SCALING AND MULTIVARIATE ANALYSIS
WITH OPTIMALTRANSFORMATIONS

Jacqueline J. Meulman

Department of Data Theory
University of Leiden
THE INTEGRATION OF
MULTIDIMENSIONAL SCALING AND MULTIVARIATE ANALYSIS
WITH OPTIMAL TRANSFORMATIONS

The recent history of multidimensional data analysis suggests two distinct traditions that have developed along quite different lines. In multidimensional scaling (MDS) the available data typically describe the relationships among a set of objects in terms of similarity/dissimilarity (or (pseudo-)distances); in multivariate analysis (MVA), data usually result from observation on a collection of variables over a common set of objects. This paper starts from a very general multidimensional scaling task, defined on distances between objects derived from one or more sets of multivariate data. Particular special cases of the general problem, following familiar notions from MVA, will be discussed that encompass a variety of analysis techniques, including the possible use of optimal variable transformation. Throughout, it will be noted how certain data analysis approaches are equivalent to familiar MVA solutions when particular problem specifications are combined with particular distance approximations.

Key words: multivariate analysis, multidimensional scaling, nonlinear transformations, distance approximation, majorization.

1. Introduction

Techniques of multidimensional scaling (MDS) and multivariate analysis (MVA) are typically described using different data analysis objectives, with the concept of distance between objects being central to MDS, and the concept of covariance between variables being central to MVA. In statistical approaches to MVA, for example, the study of the covariance structure among variables is usually stated as the prime objective, and the objects on which the variables are obtained are merely regarded as a replication factor for obtaining the covariance measures. Alternatively, MDS is commonly presented as the approximation of dissimilarities between objects by distances in a low-dimensional space, and in the case where the initial dissimilarities are derived from multivariate data, the representation of objects is the main objective with variables treated as a replication factor to obtain the object dissimilarities. Although the objectives of MDS and MVA may seem
fairly distinct, the communality between these two classes of techniques is stronger than what may be apparent initially. The intent of the present paper is to detail these connections between the techniques of MDS and MVA by developing an explicit distance interpretation for the latter.

From a geometrical point of view, techniques of MVA can be interpreted as involving two types of Euclidean spaces: a single low-dimensional representation space, and one or more higher-dimensional observation spaces. Each object corresponds to a point in each of the spaces, with the coordinates in the higher-dimensional space(s) given by the measurements or scores of the objects on the variables. Given this interpretation, the objective of MVA can be rephrased as finding coordinates in a representation space so distances among the objects in the higher-dimensional space(s) are approximated well by distances in a single low-dimensional space. Clearly, this objective falls within the MDS domain. Although the variables may now seem to play a secondary role, i.e., the distances to be approximated are merely derived from them, their influence on the analysis goes much further. In particular, the covariance structure among the variables defines the metric that operates in the observation space, and depending on the analysis, distances are sometimes corrected with respect to the covariance structure, and sometimes they are not. Generally, the choice of metric determines whether we are dealing with ordinary Euclidean distances or with Mahalanobis distances, a distinction that will turn out to be crucial in specifying different techniques in a distance approach to MVA.

An important aspect of the techniques discussed in the sequel is the possible transformation of the data. Originating from usage in Shepard's (1962a,b) and Kruskal's (1964a,b) approach to nonmetric MDS, the idea of replacing the original data by some optimal transformation corresponds in MVA to the optimal transformation of the variables, and the simultaneous parameter estimation in an MVA model. In the distance approach to MVA, transformations of the variables are incorporated in much the same manner. From a
geometrical viewpoint, variable transformation involves the replacement of the coordinates for the objects in the observation space by some other coordinate system, under the restriction that there is a one-to-one nonlinear relation between the coordinates in the two systems. For example, in a nominal transformation, only categorical information from the original variable is retained; or if ordinal information for a variable should be maintained, the original variable is transformed monotonically. As an alternative interpretation, we can consider the coordinates of the observation space only partially known, and thus, for example, we only know the ordering of the objects on each of the axes. In these cases, a representation space must be solved for as well as a transformation of the observation space.

The asymmetry between the rows, the objects, and the columns, the variables, in multivariate analysis is carried through by approximating the Euclidean distances between the objects and not between the variables. This could be regarded as a quite arbitrary choice, but the point has already been stressed by Carroll and Kruskal (1968) in their discussion on the relation between multivariate analysis and multidimensional scaling on derived dissimilarities, who remark: "There is, however, a certain deceptiveness in the formal duality between rows and columns that permit us to compute covariance, correlation, Euclidean distance, or any other index between two rows as easily as between two columns. This duality suggests that any such index is equally sensible to compute in either direction. This conclusion need not be correct, however, because changes to, say, the columns that are not seen as significantly altering the data would be viewed quite differently if applied to the rows" (Carroll and Kruskal, 1968, p. 893).

The distance approach to MVA recognizes this basic asymmetry by a different treatment of objects and variables. The variables are functions defined on the objects and may be nonlinearly transformed. From the optimally transformed variables, distances between the objects are derived that are then approximated through a metric multidimensional scaling
approach. This makes the techniques discussed in the sequel distinct from applying
nonmetric multidimensional scaling in the analysis of multivariate data, where distances are
derived from fixed coordinates given by the original variables, which are then subjected to
an optimal monotonic transformation that minimizes the badness-of-fit defined by an MDS
objective function. The result of the transformation are pseudo-distances that no longer are
related in any necessary way to the original variables by a Euclidean distance function.
Because nonmetric scaling transforms the distances in observation space, a transformation
of the variables is not necessarily obtained.

2. A general MVA objective function defined on distances

A general least squares loss function will be introduced in this section that by suitable
redefinition of its constituent terms will encompass all distance-based MVA methods to be
reviewed in the sequel. Prior to presenting this definition, however, some preliminary
notation must be presented that will be used throughout the paper. Although the purpose of
the notation will become clearer as we proceed, it is convenient if it is introduced at least in
an abstract form at this point. Specifically, we assume data are available for $n$ objects or
individuals and $M \geq 1$ groups of variables with $m_J$ variables in each group, $1 \leq J \leq M$. The
columns of the $n \times m_J$ data matrix $Z_J$ are defined by $n \times 1$ vectors that contain the
observations on the variables and are assumed to have a mean of zero. The measurements
on the objects for the $m_J$ variables define the rows in $Z_J$, and give the coordinates for each
object in an $m_J$-dimensional observation space; the rows of $Z_J$ will be denoted by
$z_1, \ldots, z_i, z_k, \ldots, z_n$. In addition, a metric $W_Z$ is defined as an $m_J \times m_J$ positive-definite
matrix and $Z_J W_Z^{1/2}$ is an $n \times m_J$ matrix with the rows containing the coordinates of the
objects on a new set of $m_J$ axes. More generally, if the columns representing the $m_J$
variables in $Z_J$ are transformed, any set of transformed variables will be denoted by $Q_J$,
while $Q^*_J$ denotes the set of optimally transformed variables. Finally, the dimensionality of
the *representation* space is assumed to be \( p \), and the coordinates for the \( n \) objects in this (unknown) space are contained in the rows of the \( n \times p \) matrix \( \mathbf{X} \). The rows of \( \mathbf{X} \) will be denoted by \( \mathbf{x}_i, \ldots, \mathbf{x}_k, \ldots, \mathbf{x}_n \). Given this notation, and to develop a general class of squared distance functions based on arbitrary positive-semidefinite quadratic forms, a squared distance between a pair of objects \( \{i,k\} \) in the \( p \)-dimensional observation space \( \mathbf{X} \) is defined by

\[
d^2_{ik}(\mathbf{X}) = (\mathbf{x}_i - \mathbf{x}_k)'(\mathbf{x}_i - \mathbf{x}_k) = (\mathbf{e}_i - \mathbf{e}_k)'\mathbf{XX}'(\mathbf{e}_i - \mathbf{e}_k),
\]

where \( \mathbf{e}_i \) is the \( i \)th column of the \( n \times n \) identity matrix \( \mathbf{I} \). Applying the squared distance function \( D^2(\cdot) \) that maps coordinates into squared distances to \( \mathbf{X} \) gives the matrix formulation:

\[
D^2(\mathbf{X}) = \mathbf{a1}^\prime + \mathbf{1a}^\prime - 2\mathbf{XX}',
\]

with \( \mathbf{a} \) an \( n \times 1 \) vector containing the diagonal elements of \( \mathbf{XX}' \); this will be denoted as \( \mathbf{a} = \text{vecdiag}(\mathbf{XX}') \). For the same pair of objects \( \{i,k\} \), \( M \) high-dimensional distances can be defined. The distance in the observation space \( \mathbf{Z}_j \) in the metric \( \mathbf{W}_Z \) (i.e., based on the positive-definite matrix \( \mathbf{W}_Z \) of order \( m_j \times m_j \)) is written as

\[
d^2_{ik}(\mathbf{Z}_j\mathbf{W}_Z^{1/2}) = (\mathbf{z}_i - \mathbf{z}_k)'\mathbf{W}_Z(\mathbf{z}_i - \mathbf{z}_k) = (\mathbf{e}_i - \mathbf{e}_k)'\mathbf{Z}_j\mathbf{W}_Z\mathbf{Z}_j'(\mathbf{e}_i - \mathbf{e}_k),
\]

and in matrix notation,

\[
D^2(\mathbf{Z}_j\mathbf{W}_Z^{1/2}) = \mathbf{a}_j\mathbf{1}^\prime + \mathbf{1a}_j^\prime - 2\mathbf{Z}_j\mathbf{W}_Z\mathbf{Z}_j',
\]

with \( \mathbf{a}_j = \text{vecdiag}(\mathbf{Z}_j\mathbf{W}_Z\mathbf{Z}_j') \).

The loss functions to be minimized in the distance approach to MVA will be given the name \textsc{stress}, to acknowledge Kruskal's (1964a,b) contribution to multidimensional scaling in its metric variety. Having introduced the basic elements of the techniques pursued in this paper, the general objective function that will be used to symbolize a variety of distance methods can be written as
\[
\text{STRESS}(Q, X) = M^{-1} \sum_{J=1}^{M} \| \mathcal{D}(Q_J W_Q^{1/2}) - D(X) \|^2,
\]

where \( \| \cdot \|^2 \) denotes a least squares discrepancy measure such that

\[
\| \mathcal{D}(Q_J W_Q^{1/2}) - D(X) \|^2 = \text{tr} \left( \mathcal{D}(Q_J W_Q^{1/2}) - D(X) \right)^\text{T} \left( \mathcal{D}(Q_J W_Q^{1/2}) - D(X) \right).
\]

In the general case described by (3a), there are \( M \) sets of variables, and STRESS must be minimized over the representation space \( X \) and over the transformations of the observation spaces \( Q = \{ Q_1, ..., Q_J, ..., Q_M \} \). The distances in the space \( Q_J \) are defined in the metric \( W_Q \), which may be different for each separate \( Q_J \).

Applications of the distance approach to multivariate data analysis will be given in the next three sections, where specific choices for \( Q_J \) and \( W_Q \) in (3a) will modify familiar MVA techniques in a straightforward way. Section 3 concerns the analysis of a single set of data, where \( M = 1 \), with principal components analysis as the prototype. Section 4 deals with symmetric analyses of \( M \) different sets of variables, with the prototype of generalized canonical analysis, and homogeneity analysis or multiple correspondence analysis as a special case. Section 5 discusses the asymmetric treatment of sets, with special attention when \( M = 2 \) to redundancy analysis and multiple regression. Least squares loss functions defined on distances can be minimized using a general algorithmic framework based on majorization (De Leeuw and Heiser, 1980; De Leeuw, 1988, among others). Within this general framework the minimization of specific loss functions raises a number of problems that have to be solved.

3. Analysis of a single set of data: principal components analysis

The basic ideas and motivation behind the distance approach to multivariate analysis will be introduced by taking one of the most widely used MVA techniques, principal components analysis, as a first application. Since \( M = 1 \), the index referring to sets will be suppressed. Generally, the distance properties of PCA are well-known since Gower
(1966), and are typically considered using the ordinary Euclidean distances between rows of the data matrix, incorporating the correlations between the variables, and then approximating these by distances between the objects in low-dimensional space. The intent of the present discussion is to develop these distance interpretations explicitly, and to argue how the MVA objective function defined in (3a) may be used to provide an alternative analysis that can be justified from the type of properties a solution based on it should possess.

Since variables in PCA are commonly normalized to have equal variance, it is convenient at the outset to define $S_Z$ as an $m \times m$ diagonal matrix containing the variances of the $m$ variables in $Z$ on its main diagonal, which leads to the obvious choice of $S_Z^{-1}$ for $W_Z$, along with the identification of $Q$ in (3a) with the $n \times m$ matrix $Z$. (We assume, for convenience, throughout that $n \geq m$).

Given these preliminaries, least squares distance fitting would then reduce to the minimization of the loss function

$$\text{STRESS}(X) = \|D(ZS_Z^{-1/2}) - D(X)\|^2,$$

by an $n \times p$ matrix $X$ (where $p \leq m(\leq n)$); this characterization defines the task we propose as an alternative to the use of conventional PCA.

To provide some background as to how the use of the loss function in (4) differs from methods currently used in the literature, we begin with a review of the rationale for classical metric multidimensional scaling treating the high-dimensional distances $D(ZS_Z^{-1/2})$ as a target (Torgerson, 1958; Gower, 1966). As is usual, we begin by carrying out the Young-Householder (1938) process, locating the origin in the centroid of points by using the centering operator $J = I - 11' / 1'1$, producing an $n \times n$ scalar product matrix $ZS_Z^{-1}Z'$:

$$-1/2JD^2(ZS^{-1/2})J = -1/2J(1a' + 1a' - 2ZS_Z^{-1}Z')J = ZS_Z^{-1}Z',$$  

(5)
with \( a = \text{vecdiag}(ZS_Z^{-1}Z') \). (It should be noted that this equality is true only if the variables \( z_j \) have zero mean). Next, an eigenanalysis of the scalar product matrix is performed:

\[
ZS_Z^{-1}Z' = K \Lambda K',
\]

where \( K \) is an \( nxn \) matrix containing the eigenvectors as columns, \( \Lambda \) is a \( nxn \) diagonal matrix containing the positive eigenvalues ordered so that \( \lambda_1 \geq ... \geq \lambda_n \), and \( t \) denotes the rank of \( ZS_Z^{-1/2} \) (for \( t \leq m \)). The matrix \( ZS_Z^{-1/2}Z' \) has its first \( t \) eigenvalues in common with the \( t \) positive eigenvalues of the \( mxm \) correlation matrix \( R(Z) = S_Z^{-1/2}Z'ZS_Z^{-1/2} \), which can be decomposed as \( R(Z) = L\Lambda L' \). The \( mxt \) component loading matrix \( A \) is obtained by scaling the first \( t \) eigenvectors by the accompanying positive eigenvalues so that \( A = L\Lambda^{1/2} \), and \( R(Z) = AA' \). To obtain an explicit representation for the objects, a singular value decomposition of the matrix \( ZS_Z^{-1/2} \) is first performed:

\[
ZS_Z^{-1/2} = KA^{1/2}L'.
\]

Since this matrix has \( t \) left singular vectors in \( K \) that are equal to the first \( t \) eigenvectors of the matrix \( ZS_Z^{-1}Z' \), the optimal solution in the classical MDS procedure for obtaining a \( p \)-dimensional \( X \) (for \( p \leq t \) ) must be equivalent to the use of the first \( p \) principal components, a result that is the basis of Gower's (1966) principal coordinates analysis. Care should be given to normalization; i.e., since the representation space approximates distances, the square root of the eigenvalues of \( ZS_Z^{-1/2} \), or the singular values of \( ZS_Z^{-1/2} \), should be used to renormalize the eigenvectors or singular vectors \( K \), i.e., \( X \) is chosen as \( K_p\Lambda_p^{1/2} \), where the subscript \( p \) indicates the use of the first \( p \) columns in \( K \) (and the first \( p \) rows and columns of \( \Lambda \)). When we wish to represent the variables at the same time, \( A \) must be chosen as \( A = L_p \).

The connections between PCA and classical MDS on Euclidean distances being established, the implicit distance approximation is captured in the form of a loss function using the STRAIN framework; STRAIN is the term coined by Carroll and Chang (1972)
to distinguish scaling using scalar products from scaling directly using distances. For principal components analysis, STRAIN can be written as

\[ \text{STRAIN}(X) = \| ZS^{-1}_Z'Z - XX' \|_2^2 = (1/4) \| J(D^2(ZS^{-1/2}_Z) - D^2(X))J \|_2^2. \] (8)

If the centering operator \( J \) and the squares in the distance function \( D^2(\cdot) \) are removed from the right term in (8), we obtain (4), which thus can be considered a modification of principal components analysis that uses a different criterion to measure the discrepancy between distances in observation and representation space.

To emphasize the difference between the two approaches based on (4) and (8) we first note that STRAIN involves squared distances, a property it has in common with the SSTRESS function in multidimensional scaling proposed by Young, De Leeuw, and Takane (1977); removing only the centering operator \( J \) from (8) would give us a SSTRESS formulation of principal components analysis.

Secondly, using (8) the approximation is from below, which means that for each pair of observation and representation distances, \( d^2_{ik}(X) \leq d^2_{ik}(ZS^{-1/2}_Z) \), a property resulting from the eigenanalysis in (6). When a \( p \)-dimensional approximation is obtained,

\[ d^2_{ik}(X) = (e_i-e_k)'XX'(e_i-e_k) = (e_i-e_k)'K_p\Lambda_pK_p'(e_i-e_k) \leq (e_i-e_k)'K\Lambda K(e_i-e_k) = d^2_{ik}(ZS^{-1/2}_Z). \] (9a)

Using the fact that the eigenvalues are ordered, the sequence

\[ d^2_{ik}(X_1) \leq d^2_{ik}(X_2) \leq \ldots \leq d^2_{ik}(X_t) = d^2_{ik}(ZS^{-1/2}_Z), \] (9b)

is obtained, where \( X_t \), with \( t \) equal to the rank of \( ZS^{-1/2}_Z \), denotes the solution with maximum dimensionality.

Geometrically, approximation from below results from the projection of the points in a higher-dimensional space onto a low-dimensional subspace. Because PCA solutions are often interpreted with respect to the objects using distances and not scalar products,
conclusions based on small distances can be suspect. For instance, two object points that are close together could result from the representation of a small distance between the objects in observation space, or alternatively, of a large distance that is badly fitted, which would give a false impression of the object pair's initial similarity. Also distinct clusters of objects in observation space could be lost in representation space and be displayed as overlapping clouds of object points. With respect to approximation from below in PCA, residual plots of distances in representation space versus distances in observation space can be very illuminating (cf. De Leeuw and Meulman, 1986a; Meulman, 1986).

Least squares distance approximation in principal components analysis cannot be done by an eigenanalysis, and solutions cannot be described in terms of a projection. In contrast with approximation just from below, distances in observation space can also be approximated from above, and may be larger in representation space. By approximating some distances from above, the overall fit is improved. This feature establishes one of the major data analytical advantages of minimizing the STRESS loss function (4) compared to conventional PCA: low-dimensional solutions will generally give more reliable information with respect to small distances. In the distance approach to PCA, small distances in low-dimensional space usually represent small distances in observation space, while large distances tend to be approximated from above (Meulman, 1986). The latter is typically less distorting in the interpretation of the structure given by the analysis than when large distances are represented by small ones.

The need for approximation either from below or from above is reflected in badness-of-fit, but compared to conventional PCA, an approximation that does allow both can reduce the number of dimensions necessary to describe the distances in observation space with an acceptable badness-of-fit. This phenomenon is also recognized by Weeks and Bentler (1979), and Heiser and Meulman (1983b) as a property of metric MDS using distances compared to metric MDS using scalar products.
Optimal transformations in principal components analysis

The process of including optimal transformation of the variables in a multivariate analysis can be motivated in a number of different ways. It may be known that the variables have not been measured on an interval scale, and therefore an analysis should be invariant under functions that are more general than linear transformations. In the sequel the actual measurement level of variables will not be considered as the main rationale behind the use of optimal transformations. Instead, they are viewed as additional sources of information with respect to badness-of-fit in the fit of models to actual data. A transformation is a choice of a new variable from a subset of candidate variables that have values displaying a one-to-one relationship with the values of the original. An optimal transformation is that particular choice that minimizes the loss function used to measure the discrepancy between the observation space and the representation space. All admissible transformations of a variable $z_j$ are in a subset that is called the cone $C_j$. A cone is a particular subset of the $n$-space $\mathbb{R}^n$ if it is true that when the subset contains an arbitrary $x$ it also contains $\alpha x$, with $\alpha$ a nonnegative scale factor; the cone is closed if the subset contains the origin, and convex if it is true that when the cone contains both $x$ and $y$ it also contains $x+y$ (cf. Gifi, 1981). Thus, replacing the variables in $Z$ by a set of transformed variables $Q$ is guided by fitting a particular MVA model under the restriction $q_j \in C_j$ and $s_j^T q_j q_j = 1$.

Because goodness-of-fit is optimized with respect to $Q$, an interesting consequence ensues for the equivalence between the classical scaling approach in (8) and the conventional PCA approach. With linear transformations the solutions for $X$ are identical under the appropriate choice of normalization, although the accompanying minimum loss values are not equal. When transformations more general than linear are considered, however, the equivalence ends. To show this explicitly, we first write conventional PCA in terms of the bilinear model (cf. Kruskal, 1978), minimizing the objective function

$$\sigma(X;A) = \|ZS_Z^{-1/2} - XA\|^2,$$

(10)
over the \( nxp \) matrix \( X \) giving the component scores and the \( m \times p \) matrix \( A \) the component loadings. The minimum loss, using the singular value decomposition (7), is obtained as

\[
\sigma(\ast;\ast) = \text{tr } L \Lambda L' - \text{tr } L_p \Lambda_p L_p' = m - \sum_{s=1}^{p} \lambda_s(R(Z)),
\]

which is a function of the first \( p \) eigenvalues of the correlation matrix \( R(Z) \).

When transformations are included in (10), the loss function becomes

\[
\sigma(Q;X;A) = \| QS^{-1/2} - XA \|_2^2.
\]

Since this objective function has to be minimized also over \( Q \), with \( Q^* \) as the optimal solution, the minimum loss must be a function of the eigenvalues of \( R(Q^*) \), and the definition of optimality implies that the sum of the first \( p \) eigenvalues is as large as possible.

Now consider the minimum of the STRAIN function (8), found by using the eigenvalue decomposition (6) and the Young-Householder process displayed in (5), we find

\[
\text{STRAIN}(\ast) = \text{tr } K A^2 K' - \text{tr } K_p \Lambda_p^2 K_p'.
\]

The right term in (12) is equal to

\[
\text{tr } L \Lambda^2 L' - \text{tr } L_p \Lambda_p^2 L_p' = m - \sum_{s=1}^{p} \lambda_s^2(R(Z)),
\]

and we note that the minimum loss of conventional PCA in (11) is a function of the first \( p \) eigenvalues, but the minimum loss of STRAIN is a function of the squares of the first \( p \) eigenvalues. Optimal transformation in the STRAIN framework implies the minimization of

\[
\text{STRAIN}(Q;X) = \| J(D^2(QS^{-1/2} - D^2(X)))J \|_2^2,
\]

which maximizes the sum of squares of the first \( p \) eigenvalues of the matrix \( QS^{-1/2} Q' \), and thus also of \( R(Q) \), over \( Q \). Because this is clearly a different criterion, and transformations in both approaches are optimal by definition, this will generally lead to different transformations, and consequently to different solutions for \( X \).

The incorporation of nonlinear transformations in the STRAIN function, as given above, apparently has not been pursued in the literature. Generalizations of PCA to include
transformations have primarily used the bilinear model, but since the resulting transformations can only be suboptimal with respect to squared distance approximations from below, this weakens the distance properties of the techniques proposed by Shepard (1966), Roskam (1968), Kruskal and Shepard (1974), Young, Takane and De Leeuw (1978), Gifi (1981), and Winsberg and Ramsay (1983).

Performing a PCA by minimizing a least squares objective function defined on the distances themselves is based on yet another criterion. Combining approximation from both above and below with transformation of the variables can be rephrased through the objective function \( \text{STRESS}(Q; X) = \|D(QS_Q^{-1/2}) - D(X)\|^2 \), to be minimized over \( X \) and \( Q \), with \( q_j \in C_j \) and \( s_q^{1/2} q_j s_q^{-1/2} = 1 \). This MDS loss function implies an approximation of distances in a partially unknown space, where \( q_j \in C_j \), and the transformed variables \( q_j^* \) span an optimal space \( Q^* \) that generates distances \( D(Q^*S_Q^{-1/2}) \). The overall objective is still to replace the higher-dimensional space by a low-dimensional \( X \), but because variable transformation is guided by the distances derived from \( X \), the result is likely to be different from the two other approaches to PCA described above.

For a given representation space \( X \), the transformation problem can be dealt with by solving a constrained multidimensional scaling problem with respect to \( Q \) (cf. De Leeuw and Meulman, 1986a; Meulman, 1986). The idea is as follows. Consider \( Q \) as an admissible transformation of the observation space \( Z \), and suppose for given \( D(X) \) a space \( Y \) can be found that satisfies

\[
\|D(X) - D(Y)\|^2 \leq \|D(X) - D(Q)\|^2. 
\] (14)

Then a projection in the space that satisfies the transformation constraints is defined by

\[
P_C(Y) = \{ \hat{q}_j \in C_j \mid \| \hat{Q} - Y \|^2 = \min_{q^*_j \in C_j} \| Q^+ - Y \|^2 \}. 
\] (15)

Definition (15) says that \( \hat{q}_j \) is an admissible transformation of \( z_j \) that is as close as possible to \( y_j \), where \( y_j \), using (14), decreases the STRESS with respect to a given \( X \). Now the
convergence results of De Leeuw and Heiser (1980) can be used to show that the \( \hat{Q}_j \) will yield the desired result \( \|D(X) - D(\hat{Q})\|^2 \leq \|D(X) - D(Q)\|^2 \). Thus, when \( Y \) is found by a procedure that guarantees (14) then \( \hat{Q} \) is found by minimizing \( \|Q^+ - Y\|^2 \) under the transformation constraints, which can be done by performing the appropriate form of regression of each \( y_j \) upon \( z_j \), giving the result \( \hat{q}_j \).

In the procedure just described we neglected the fact that principal components analysis was defined in the metric \( S_Q^{-1} \). Because \( S_Q^{-1} \) is diagonal, this constraint applies to each variable separately, and thus, a transformation \( \hat{q}_j \) should be constructed under the condition \( s_j^{-1}\hat{q}_j \hat{q}_j = 1 \). This constraint is imposed by solving a so-called normalized cone regression problem, which, however, because when \( q_j \in C_j \) then \( q_j s_q^{1/2} \in C_j \), implies unrestricted regression of \( y_j \) upon \( z_j \), and identifying \( s_q^{1/2} \) as \( (\hat{q}_j/\hat{q}_j)^{-1/2} \). The fact that \( S_Q^{-1} \) is diagonal makes it possible to separate the problem of finding the optimal direction in the cone from finding the appropriate length (Kruskal and Carroll, 1969; De Leeuw, 1977).

Finally, it should be addressed how an admissible \( Y \) can be obtained, which can be done by using the majorization algorithm for MDS (De Leeuw and Heiser, 1980; De Leeuw, 1988). They define the so-called Guttman transform of the matrix \( X \), denoted by \( X^+ \), which guarantees that \( \|D(\cdot) - D(X^+)\|^2 \leq \|D(\cdot) - D(X)\|^2 \). The convergence properties of the Guttman transform can be used for solving for a \( Y \) that satisfies (14) by reversing the role of the distances in the two spaces, which implies that the Guttman transform \( Y \) is defined by

\[
Y = n^{-1} B(QS_Q^{-1/2})QS_Q^{1/2}. \tag{16}
\]

To show how the elements of the \( nxn \) matrix \( B(QS_Q^{-1/2}) \) are defined, it is convenient to create two auxiliary matrices, the \( nxn \) matrix \( B^0(QS_Q^{-1/2}) \), whose elements are:

\[
b_{ik}^0(QS_Q^{-1/2}) = d_{ik}(X) / d_{ik}(QS_Q^{-1/2}) \text{ if } i \neq k; \tag{17a}
\]

\[
b_{ik}^0(QS_Q^{-1/2}) = 0 \text{ if } d_{ik}(QS_Q^{-1/2}) = 0, \tag{17b}
\]
and the \( nxn \) diagonal matrix \( B^*(QS_Q^{-1/2}) \), with diagonal elements

\[
b^*_i(QS_Q^{-1/2}) = 1'B^o(QS_Q^{-1/2})e_i. \tag{17c}
\]

The matrix \( B(QS_Q^{-1/2}) \) itself can now be written as

\[
B(QS_Q^{-1/2}) = B^*(QS_Q^{-1/2}) - B^o(QS_Q^{-1/2}). \tag{17d}
\]

Repeatedly combining the update (16) with the metric projection (15) will give a convergent series of observation spaces with respect to a fixed representation space \( X \).

**An application**

To illustrate the distance approach to PCA with optimal transformations, a set of variables are re-analyzed that consist of 7 social indicator statistics taken from the statistical abstracts of the U.S. (1977). The data were collected by Wainer and Thissen (1981) in order to re-examine the search for "The Worst American State". The social indicator statistics are given in Table 1.

<table>
<thead>
<tr>
<th>Table 1. Social indicator statistics for the United States</th>
</tr>
</thead>
<tbody>
<tr>
<td>POPULATION</td>
</tr>
<tr>
<td>INCOME</td>
</tr>
<tr>
<td>ILLITERACY</td>
</tr>
<tr>
<td>LIFE</td>
</tr>
<tr>
<td>HOMIC</td>
</tr>
<tr>
<td>SCHOOL</td>
</tr>
<tr>
<td>FREEZING</td>
</tr>
</tbody>
</table>

Previous analyses of these data, with transformations of the variables chosen as either third degree (De Leeuw and Meulman, 1986) or second degree polynomials (Meulman,
1986) pointed to the special role in the data of the states Alaska, Nevada, and Hawaii. The transformations of several variables displayed nonlinear functions that bended back at the extremes, but this phenomenon did not occur for the same variables in a conventional PCA compared to a PCA using distances. When transformations are that different, it is difficult to distinguish the transformation effect from the effect of approximation. To inspect here both effects in the same analysis, the transformations were chosen from a more limited class, i.e., monotonic second degree polynomials, which were fitted by using the approach to integrated splines as in Winsberg and Ramsay (1983), and Ramsay (1989). Effectively, this implies that merely two parameters, the M-splines coefficients, are estimated for each transformation, and because the transformed variables are normalized, transformations in conventional PCA and PCA using distances become identical as soon as for both either the first or the second coefficient becomes zero to ensure monotonicity.

In Figure 1a the points for the fifty states are shown that resulted from the distance approach to PCA using spline transformations; Figure 1b gives them for conventional PCA. The two Figures are on the same scale, which shows immediately that the overall configuration for conventional PCA is much smaller. Except for this scale factor, the two configurations show many local resemblances. In both spaces directions for the transformed variables are fitted. In conventional PCA these are obtained from \( \mathbf{A} \) in (10); in PCA using distances they are found by a projection afterwards. In both solutions the overall structure of the USA is quite well fitted; the first dimension can be characterized by a "quality of life" factor, while the second one gives the relative positions of the states with respect to Population and Freezing. At the right hand side of the Figures we see the Southern states, scoring low on the quality of life variables, and these are contrasted with states like ND, CO, MN, CT, WY, IA, NH, NE and SD. The fit of the transformed variables in the two configurations is given in Table 2. Especially Population is much
Figure 1a. Solution for United States: PCA using the distance model

Figure 1b. Solution for United States: Conventional PCA
Table 2. Fit of the variables in the two spaces

<table>
<thead>
<tr>
<th>Variable</th>
<th>PCA Using Distances</th>
<th>PCA Conventional</th>
</tr>
</thead>
<tbody>
<tr>
<td>POPULATION</td>
<td>.72</td>
<td>.91</td>
</tr>
<tr>
<td>INCOME</td>
<td>.92</td>
<td>.86</td>
</tr>
<tr>
<td>ILLITERACY</td>
<td>.88</td>
<td>.90</td>
</tr>
<tr>
<td>LIFE</td>
<td>.75</td>
<td>.84</td>
</tr>
<tr>
<td>HOMIC</td>
<td>.87</td>
<td>.88</td>
</tr>
<tr>
<td>SCHOOL</td>
<td>.89</td>
<td>.88</td>
</tr>
<tr>
<td>FREEZING</td>
<td>.81</td>
<td>.81</td>
</tr>
</tbody>
</table>

better fitted in the conventional PCA solution. With respect to the distance approximation, however, the PCA using distances fits 2.5 times better.

Although the two representations of the structure are much alike, there are a few notable exceptions: Alaska, Nevada and Hawaii. These states do not fit very well in the structure. In the PCA using distances these point are found located in the upper left corner; in the conventional PCA solution, they are found quite near to the origin of the configuration, which follows from the projection of the points into the subspace. Consider, e.g., Alaska. It has a small population, many homicides, a high level of education, it's cold, and it has the highest income. In the conventional PCA none of these characteristics is apparent; in the PCA using distances Alaska is located in a right position according to INCOME, SCHOOL, and HOMICIDE, but in a wrong one with respect to FREEZE and especially POPUL. At this point it is worthwhile to consider the transformation plots in Figure 2. We see that the transformations for the two techniques are quite similar, except for INCOME and POPULATION. The conventional PCA analysis has given ALASKA a relative lower value for INCOME, while the distance PCA has maintained the high value. This process has been reversed for POPULATION, where in the conventional PCA the small values have been transformed linearly, while in the distance PCA this part of the transformation is flat, giving Alaska the same value as many other states. As a result,
Figure 2. Monotonic spline transformations and Residuals plot for United States analyses
Alaska has a high position for INCOME and POPULATION in the upper configuration, while the fit for INCOME is large and for POPULATION rather small. In the conventional PCA solution POPULATION fits better than INCOME.

Figure 2 also shows a plot of the residuals in the distance analysis versus the conventional one. It is clear that the distance approximation in the conventional analysis is only from below, the more a point is located to the left of the vertical line, the more serious the badness-of-fit. In the distance analysis the approximation is from above and below, the residuals are pictured respectively above and below the horizontal line. The points in the scatter diagram that fall within the two orthogonal 45 degrees lines indicate residuals that are smaller in the PCA using distances versus the conventional PCA. It may be clear that it is not true that all residuals are smaller.

5. Symmetric analysis of $M$ sets of data: generalized canonical analysis

As an application of the distance approach to the symmetric analysis of $M$ distinct sets of variables, we first consider generalized canonical analysis. There are different ways to generalize Hotelling’s (1936) canonical correlation analysis with two sets variables to the analysis of $M$ sets, and these do in general not give the same solution. Reviews are given by Kettenring (1971), Gifi (1981), Van de Geer (1984, 1986), among others.

As a starting point here, we use a particular generalization proposed by Carroll (1968). Extensions of this approach to include optimal transformation of variables can be found in Gifi (1981), Van der Burg, De Leeuw and Verdegaal (1988), and Van der Burg (1988), among others.

For the moment consider the $Q_J$ as given so that without loss of generality $Q_J = Z_J$. To investigate the relationship between $M$ groups of variables $Z_1,...,Z_J,...,Z_M$, a set of weight matrices $A = \{A_1,...,A_J,...,A_M\}$ should be found so the linear combinations $Z_1A_1,...,Z_JA_J,...,Z_MA_M$ resemble each other as closely as possible. One
way of ensuring maximal resemblance is to minimize the least squares discrepancy between each separate $Z_jA_j$ and an unknown $p$-dimensional comparison matrix $X$, i.e., minimizing

$$
\sigma(X; A) = M^{-1} \sum_{j=1}^{M} \|Z_jA_j - X\|^2,
$$

over $X$ and $A$.

Since a normalization has to be imposed to avoid a trivial solution, Gifi (1981) chooses the convenient normalization constraint, $X'X = I$. Setting partial derivatives with respect to $A_j$ equal to zero gives the optimal $A_j$ as $A_j = V_Z^1Z_j'X$, where $V_Z$ is defined as the covariance matrix $V_Z = Z_j'Z_j$. When the optimal weights $A_j$ are substituted in the objective function (19), and the normalization conditions are taken into account, the optimum is attained when the term $M^{-1}\sum Z_j \text{tr}(X'(Z_jV_Z^1Z_j)X)$ is maximized. The latter occurs when $X$ is chosen as the set of the first $p$ eigenvectors $K_p$ of the scalar product matrix $M^{-1}\sum Z_jV_Z^1Z_j$, decomposed as $M^{-1}\sum Z_jV_Z^1Z_j = K\Lambda K'$.

From the geometrical point of view, the eigenanalysis that ensures maximization shows that generalized canonical analysis involves a special metric defined in $Z_j$, since the role of $Z_j$ in the scalar product term is corrected by the term $V_Z^{-1/2}$ for the covariance structure among the variables, and $V_Z^{-1/2}Z_jZ_jV_Z^{-1/2} = I$. The $V_Z^{-1/2}$ metric, which generates the Mahalanobis distance, originates from the literature on discriminant analysis although it is being applied here in a slightly different form. The Mahalanobis distance between two objects $(i,k)$ is defined as

$$
d_{ik}^2(Z_jV_Z^{-1/2}) = (z_i - z_k)'V_Z^{-1}(z_i - z_k),
$$

and it is a Euclidean distance in the metric $V_Z^1$, giving distances between the objects in the orthonormalized space $Z_jV_Z^{-1/2}$.

The definition of the Mahalanobis distance suggests that when generalized canonical analysis is modified using least squares distance fitting, we should minimize
STRESS(X) = \( M^{-1} \sum_{j=1}^{M} \|D(Z_j V_j^{1/2}) - D(X)\|^2, \) \( (21) \)

over X. The role of \( W^{1/2}_Q \) in the general MDS objective function (3a) is taken by \( V_j^{1/2} \), and (21) implies an approximation of the Mahalanobis distances \( D(Z_j V_j^{1/2}) \) in each of the \( M \) observation spaces by Euclidean distances in a single representation space X.

The term \( V_j^{1/2} \) is not uniquely defined, and only has to be a square nonsingular matrix such that \( Z_j V_j^{1/2} \) is orthonormal. The general form \( Z_j V_j^{1/2} T_j \), with \( T_j \) a full rotation matrix, defines a class of orthonormalized matrices that differ by a rotation only. If the eigenvalue decomposition of \( Z_j Z_j' \) is written as \( Z_j Z_j' = P_j \Psi_j P_j' \), and \( T_j \) is chosen as \( T_j = I \), then \( Z_j V_j^{1/2} = Z_j P_j \Psi_j^{1/2} \), giving the principal components solution. If \( T_j = P_j' \), then \( Z_j V_j^{1/2} = Z_j P_j \Psi_j^{1/2} P_j' \), giving the Procrustes solution (Cliff, 1966); and if \( T_j \) is chosen such that \( P_j \Psi_j^{1/2} T_j \) is an upper triangular matrix, \( Z_j V_j^{1/2} \) is the Gram-Schmidt transformation of \( Z_j \). Notwithstanding this indeterminancy, distances are invariant under rotations, and \( D(Z_j V_j^{1/2}) \) will be the same no matter how we choose \( V_j^{1/2} \).

The scaling problem expressed in (21) can be simplified considerably because STRESS(X) can be partitioned into two additive components. Define an average Mahalanobis distance matrix \( D_M \) on the basis of the \( M \) matrices \( D(Z_j V_j^{1/2}) \) as

\[
D_M = M^{-1} \sum_{j=1}^{M} D(Z_j V_j^{1/2}).
\]

(22)

Then, the first component can be written as

\[
\text{STRESS}_H = M^{-1} \sum_{j=1}^{M} \|D(Z_j V_j^{1/2}) - D_M\|^2,
\]

(23)

which is stress due to heterogeneity among the \( M \) different sets using the Mahalanobis distances between the objects. The second component is called the proper stress and is defined by the expression

\[
\text{STRESS}_p(X) = \|D_M - D(X)\|^2.
\]

(24)
Because stress due to heterogeneity (23) is a constant term, the proper stress (24) is the only component dependent on \( X \). Therefore, the discrepancy measured in (21) will attain a minimum with respect to \( X \) by the minimization of the objective function (24), and the latter can be done by existing multidimensional scaling methods minimizing Kruskal's STRESS function.

To show that generalized canonical analysis in the form (19) can be viewed as a classical scaling technique, define the STRAIN loss function for generalized canonical analysis as

\[
\text{STRAIN}(X) = M^{-1} \sum_{j=1}^{M} \| J(D^2(Z_jV_{Z}^{1/2}) - D^2(X))J \| ^2. \tag{25}
\]

When (25) is partitioned in two additive parts similar to the decomposition of (21), with one dependent and one independent of \( X \), it is obvious that it is sufficient to minimize

\[
\| J(D_M^2 - D^2(X))J \| ^2,
\]

with \( D_M^2 \) denoting \( M^{-1}D^2(Z_jV_{Z}^{1/2}) \). Because the following equality holds:

\[
-\frac{1}{2}JD_M^2J = M^{-1} \sum_j Z_jV_{Z}^{1/2}Z_j',
\tag{26}
\]

generalized canonical analysis solutions, obtained either through optimal linear combinations as in (19), or through the classical scaling approach that minimizes (25), must be equivalent. In the classical scaling approach \( X \) would not be normalized so \( X'X = I \), but instead, as \( X'X = \Delta \), to give the object configuration a shape that approximates the squared distances \( D^2(Z_jV_{Z}^{1/2}) \) optimally from below.

Optimal transformations in generalized canonical analysis

As in principal components analysis, optimal transformations can be introduced in generalized canonical analysis to reduce the loss due to replacing an observation space by a representation space. Incorporating the minimization over \( Q \) in the generalized canonical analysis approaches (19) and (21) amounts to a quite different problem because of the
normalization involved. Based on (19), $X$ was normalized so that the normalization constraint is independent from the transformation constraint $q_j \in C_j$, where $q_j$ is written for any variable in set $Q_j$. In the distance approach $X$ is not normalized. Moreover, canonical analysis was identified with the Mahalanobis metric, implying that optimality must be attained with respect to the Mahalanobis distances in $Q_j$ so that the distance function $D(\cdot)$ should be applied to the orthonormalized space $Q_j V_Q^{1/2}$. When the distance loss function is written as $\text{STRESS}(Q;X) = M^{-1} \sum_j \lVert D(Q_j V_Q^{1/2}) - D(X) \rVert^2$, the transformation constraints $q_j \in C_j$ are no longer separated from the normalization constraints $V_Q^{1/2} Q_j^* V_Q^{1/2} = I$, which makes the constrained multidimensional scaling problem much more complicated than for the principal components analysis of the previous section. Like in PCA, the constrained problem can be split up into two parts, of which the first should ensure for fixed $X$ an admissible unconstrained update $Y$ that satisfies $\lVert D(X) - D(Y) \rVert^2 \leq \lVert D(X) - D(Q_j V_Q^{1/2}) \rVert^2$. Using the majorization approach again, this is true for $Y$ chosen as $Y = n^{-1} B(Q_j V_Q^{1/2}) Q_j^* V_Q^{1/2}$, where $B(Q_j V_Q^{1/2}) = B^*(Q_j V_Q^{1/2}) - B^0(Q_j V_Q^{1/2})$, and

\begin{align}
    b_{ik}^q(Q_j V_Q^{1/2}) &= d_{ik}(X) / d_{ik}(Q_j V_Q^{1/2}) \text{ if } i \neq k; \\
    b_{ik}^q(Q_j V_Q^{1/2}) &= 0 \text{ if } d_{ik}(Q_j V_Q^{1/2}) = 0; \\
    b_{ik}^*(Q_j V_Q^{1/2}) &= 1^T B^0(Q_j V_Q^{1/2}) e_i.
\end{align}

(27a) \hspace{1cm} (27b) \hspace{1cm} (27c)

The second part involves the metric projection into the space that satisfies the transformation constraints, i.e. it should minimize

\[ \lVert Q_j^+ V_Q^{1/2} - Y \rVert^2 \]

over $q_j^+ \in C_j$ and $V_Q^{1/2} Q_j^+ V_Q^{1/2} = I$.

Van der Burg and De Leeuw (1983) discussed a very closely related problem in the context of nonlinear canonical correlation analysis, where two sets of linear combinations $Q_1 A_1$ and $Q_2 A_2$ are compared directly without an intermediate comparison space $X$. 
Usually, one requires that the canonical variates, given by \( Q_1 A_1 \) and \( Q_2 A_2 \), are uncorrelated so \( A_1 Q_1^T Q_1 A_1 = A_2 Q_2^T Q_2 A_2 = I \). This condition has been called strong orthogonality by Dauxois and Pousse (1976). Because this normalization constraint when combined with the optimal transformation is difficult to deal with directly, Van der Burg and De Leeuw (1983) showed that a convergent algorithm could be obtained by a transfer of normalization, i.e., either (a) \( A_1 Q_1^T Q_1 A_1 = I \) or (b) \( A_2 Q_2^T Q_2 A_2 = I \), where (a) is true when \( Z_2 \) is transformed, and (b) when \( Z_1 \) is transformed. Unfortunately, this idea of transfer of normalization cannot be applied to (28) because \( Y \) is fixed, and thus, an alternative had to be developed.

For notational convenience the index referring to sets will be suppressed. At this point we choose for definiteness \( QV_Q^{1/2} \) as the Gram-Schmidt transformation of \( Q \), denoted as \( GS(Q) = F = QS \), with \( S \) an upper-triangular matrix. This particular choice is convenient because of the special role that is played by the last column of \( Q \) in the Gram-Schmidt transformation. When \( QS \) is partitioned into \( QS = (F_m, f_m) \), where \( F_m \) denotes that the last column \( f_m \) is omitted from \( F \), then the metric projection problem (28) can be written as

\[
\| Y - (F_m, f_m) \|_2^2 = \| y_1 - s_{11} q_1 \|_2^2 + \| y_2 - (s_{12} q_1 + s_{22} q_2) \|_2^2 + \ldots + \| y_m - (s_{1m} q_1 + s_{2m} q_2 + \ldots + s_{mm} q_m) \|_2^2.
\]

This partitioning shows that \( q_m \) is the only variable that is involved in one term only, the last one. \( s_{mm} \) is the normalization factor that gives \( f_m f_m = 1 \). The other elements of the last column of the matrix \( S \) ensure that \( f_m \) is orthogonal to \( F_m \), and they are independent of the length of \( q_m \). In addition, the Gram-Schmidt transformation makes it possible to write \( f_m \) as

\[
f_m = \alpha_m \Psi_m q_m,
\]

(31)
where $\Psi_m$ is the anti-projector $\Psi_m = I - F_m(F_m)'$, and $\alpha_m$ is defined by $\alpha_m = (q_m'\Psi_m q_m)^{-1/2}$. Thus (30) can also be written as

$$\|Y - F\|^2 = \|Y_m - F_m\|^2 + \|y_m - \alpha_m \Psi_m q_m\|^2,$$

which means we have to minimize

$$\phi(q_m) = \|y_m - \alpha_m \Psi_m q_m\|^2$$

over $q_m \in C_m$. The translation of the confounded problem (28) in two independent parts still constitutes only a minor part of the solution of the total problem. Apart from the fact that it still has to be shown how (33) is actually minimized over all admissible transformations of $z_m$, the problem has been translated for the special case where only the last variable is to be transformed. A permutation of the columns of $Q$ so the transformed variable becomes the first variable and the variable to be transformed next becomes the last one, denoted as $Q^p$, does not suffice, because $GS(Q)$ is not equal to $GS(Q^p)$ and therefore the loss measured by (32) will increase after the permutation. Since it is true that the space spanned by the columns of $Q$ is equal to the space spanned by the columns of $Q^p$, $GS(Q)$ is an orthonormal basis for the column space of $Q^p$. On the other hand, it is also true that $GS(Q^p)$ gives an orthonormal basis for the column space of $Q^p$, and therefore $GS(Q^p)$ must be a rotation of $GS(Q)$, thus $GS(Q^p) = FT$, where $T$ is an orthogonal rotation matrix. The loss measured by (32) is equal to $\|YT - FT\|^2$, and after splitting off the variable $q_m$ in $Q^p$ and redefining $\Psi_m$ as $\Psi_m = I - FT_m(T_m)'F'$, the next substep is to minimize $\phi(q_m) = \|Yt_m - \alpha_m \Psi_m q_m\|^2$ with respect to all $q_m$ in a general cone $C_m$.

**Majorization**

For some subsets of possible transformations the problem is not too complicated, for instance if we choose transformations of a fixed parametrized form, like polynomials of a certain degree. If the cone $C_m$ is defined by all possible monotonic transformations, the
problem could be solved by using a rather complicated quadratic programming procedure. Instead a different approach is preferred. To keep the notation simple, we return to the form (33) and drop reference to variable $m$. First observe that we can switch from (33) to

$$\phi_0(q) = \| \Psi y - \alpha \Psi q \|^2$$

which differs from (33) by a constant that depends on $y$ only. In addition we apply the theory of normalized cone regression mentioned above to problem (34), using the fact that when the optimal $q$ is in the cone, then $\alpha q$ is also in the cone. This comes to minimizing

$$\phi_1(q) = \| \Psi y - \Psi q \|^2 = (y - q)' \Psi (y - q)$$

over $q \in \mathbb{C}$, with $q' \Psi q$ free. At this point the major problem in minimizing (35) is caused by the projector $\Psi$, and now majorization can be applied. In majorization two types of functions are considered: $\phi_1(x)$, a function of a single variable $x$, and $\phi_2(x,y)$, a function of two variables. When $\phi_1(x)$ is a function that is not easily minimized, then we replace it by a feasible quadratic function $\phi_2(x,y)$ that majorizes $\phi_1(x)$, which means it must be true for any $x$ and $y$ that

$$\phi_1(x) \leq \phi_2(x,y).$$

In addition, it should also be true that

$$\phi_1(x) = \phi_2(x,y) = \phi_1(y) \text{ if } x = y.$$  

(37)

For the present problem (35) the majorization approach says that we should look for a replacement of the function $\phi_1(q)$ in the form of $\phi_2(q,v)$, with a minimizer $q^+$ that is independent of $\Psi$. It should be observed first that for any $q$ and $v$ it is true that

$$\phi_1(q) = [(q - v) - (y - v)]' \Psi [(q - v) - (y - v)]$$

$$= (q - v)' \Psi (q - v) + (y - v)' \Psi (y - v) - 2 (q - v)' \Psi (y - v).$$

(38)
Because \( \Psi \) is a projector, we have the inequality \((q - v)\Psi (q - v) \leq (q - v)'(q - v)\), which makes it possible to write

\[
\phi_1(q) \leq (q - v)'(q - v) + (y - v)'\Psi (y - v) - 2 (q - v)'\Psi (y - v).
\] (39)

Rewriting the right-hand term of (39) gives the function

\[
\phi_2(q,v) = \| (q - v + \Psi (y - v) \|^2,
\] (40)

and it is true that \( \phi_2(q,v) \) majorizes \( \phi_1(q) \) since (36) and (37) are satisfied. Now the minimizer of (40) \( q^* \) is obtained by the metric projection \( P_C \{ v + \Psi (y - v) \} \), which in the case of monotonic transformation is a monotone regression of \( \{ v + \Psi (y - v) \} \) upon \( z \).

Combining (36) and (37) with the fact that \( q^* \) minimizes (40) gives the crucial chain

\[
\phi_1(q^*) \leq \phi_2(q^*,v) \leq \phi_2(v,v) = \phi_1(v),
\] (41)

and by repeatedly computing the update

\[
q^* \in P_C \{ v + \Psi (y - v) \}
\] (42)

with \( v = q^{k-1} \), the result of the previous iteration, we converge to the minimum of (40), which, because here \( q^* = v \), is also the minimum \( \hat{q} \) of (35). At this point \( f \) is computed as \( f = (\hat{q}'\Psi \hat{q})^{-1/2}\Psi \hat{q} \). Next the last column of \( Q \) is replaced by \( \hat{q} \), \( Q \) is permuted and \( F \) and \( Y \) are rotated to transform the next variable in the set. More details about the algorithm and applications of the distance approach to generalized canonical analysis with transformations of the variables can be found in Meulman (1986; 1987).

*Multiple correspondence analysis*

The essentials of the technique that is commonly called multiple correspondence analysis can be found in the work of Fisher (1938), Guttmann (1941), Burt (1950), Hayashi (1952), among others, and the rediscoveries since the seventies (cf. Benzécri et al., 1973; De Leeuw, 1973; Nishisato, 1980; Gifi, 1981, under the name homogeneity analysis; Lebart, Morineau, and Warwick, 1984; Greenacre, 1984; Tenenhaus and Young, 1985).
In its history, the technique has been given many different interpretations, but here it will be viewed as a special case of generalized canonical analysis (cf. Saporta, 1975).

Multiple correspondence analysis is applied to categorical, or nominal data; i.e., each variable \( z_j \) is assumed to have \( k_j \) distinct categories. From the distance analysis viewpoint, the crucial part of the optimal transformation is performed beforehand: each of the \( m \) variables \( z_j \) is replaced by an \( n \times k_j \) orthogonal binary matrix \( G_j \), called an indicator matrix, so \( g_{jir} = 1 \) if \( z_{ij} = r \) and \( g_{jir} = 0 \) otherwise, \( r = 1, \ldots, k_j \).

In this section, \( G_j \) takes the role of \( Q_j \) in loss function (3a), and \( W_Z \) is chosen as \( C_j^{-1} = (G_j'G_j)^{-1} \), a diagonal matrix with the inverse of the column marginals of \( G_j \) on the diagonal. As in canonical analysis, a Mahalanobis metric is implicit since \( G_j \) is orthogonal by definition and the term \( C_j^{-1} \) corrects for the column marginals. In this special case of generalized canonical analysis, the Mahalanobis metric is equivalent to the \( \chi^2 \) metric, which in simple correspondence analysis involves correction for both the row and the column marginals (cf. Heiser and Meulman, 1983a). However, since the rows of \( G_j \) sum up to 1, the \( \chi^2 \) distance between two objects \( \{i,k\} \) in \( G_j \) is simply defined by

\[
\chi^2_{ik}(G_j) = (g_i - g_k')C_j^{-1}(g_i - g_k).
\]  

(43)

The distance loss function is written as

\[
\text{STRESS}(X) = m^{-1} \sum_{j=1}^{m} \|D(G_jC_j^{-1/2}) - D(X)\|^2,
\]  

(44)

but because \( G_j \) and \( C_j^{-1/2} \) are fixed it has to be minimized only over \( X \) and can be replaced by the simpler \( \text{STRESS}_p(X) = \|D_{\chi^2} - D(X)\|^2 \). Here \( D_{\chi^2} \) denotes the average \( \chi^2 \) distance matrix obtained from the \( m \) matrices \( D(G_jC_j^{-1/2}) \).

For multiple correspondence analysis, application of the Young-Householder process gives

\[
-\frac{1}{2}mD_{\chi^2}^2 J = -\frac{1}{2m} \sum_{j=1}^{m} J(a_j1^j + 1a_j^j - 2G_jC_j^{-1}G_j^j)J = m^{-1} \sum_{j=1}^{m} JG_jC_j^{-1}G_j^j J,
\]  

(45)
with $a_j = \text{vecdiag}(JG_jC_G^{-1}G_j^TJ)$. Because columns of the indicator matrix are not centered, the centering operator $J$ cannot be omitted, in short, the classical scaling approach amounts to an eigenvalue analysis after a double centering of the average projector matrix $P_0 = m^{-1} \sum_j G_jC_G^{-1}G_j^T$.

In Gifi (1981) it is shown that multiple correspondence analysis is equivalent to the minimization of the least squares loss function

$$\sigma(X;A) = m^{-1} \sum_{j=1}^m \|G_jA_j - X\|^2,$$

where the $A_j$ give weights for the binary variables in $G_j$ interpreted as $p$-dimensional coordinates for the categories of the variable $z_j$. These can be eliminated from (46) by inner minimization. The optimal $A_j$ are found as $A_j = C_G^{-1}G_j^TX$, implying that a category point is in the centroid of the appropriate object points. Substitution, taking into account the centering of $X$, shows that the optimal $X$ is found by maximizing $\text{tr} \ X'JP_0JX$, and thus optimal solutions for $X$ in multiple correspondence analysis and classical scaling defined on the $\chi^2$ distances must be equivalent. With respect to distances, they share the property of squared approximation from below for the average high-dimensional distances. Low-dimensional results of conventional multiple correspondence analysis are compared to the least squares fitting of the $\chi^2$ distances in Meulman (1986), where the latter gives a remarkable dimensionality reduction for the representation space $X$.

6. Asymmetric analysis of 2 sets of data: redundancy analysis

This section will give special attention to forms of redundancy analysis, a technique developed in Rao (1964), Fortier (1966), and Van den Wollenberg (1977), among others. A generalization to the analysis of nominal variables is given by Israëls (1984, 1987). Having discussed principal components analysis and generalized canonical analysis in the two previous sections, redundancy analysis can be described by combining several ideas.
We first note that redundancy analysis concerns two sets of variables, $Z_1$ and $Z_2$, treated asymmetrically, where $Z_1$ is considered the predictor set and $Z_2$ the criterion set. Redundancy analysis corrects for the correlations in $Z_1$ and incorporates them in $Z_2$. Combining (10) and (19) in the conventional MVA approach gives

$$\sigma(X;A) = 1/2 \{ \|XZ_1A_1^{-1}X\|^2 + \|XZ_2A_2^{-1}Z_2S_2^{-1/2}Z_2\|^2 \}. \quad (47)$$

Each of the two parts of (47) is defined on a different type of linear combination. Canonical analysis is recognized in the role that is played by $Z_1$ and principal components analysis in the role that is taken by $Z_2$, with the two parts connected through a common $X$. This special formulation of redundancy analysis is discussed in Van de Geer (1984), De Leeuw and Bijleveld (1987) and Meulman (1986, 1988), among others.

The maximization problem resulting from (47) with respect to $X$, when the optimal $A_1$ and $A_2$ are substituted, can be solved by the eigenanalysis of the matrix

$$1/2(Z_1V_{Z_1}^{-1}Z_1' + Z_2S_2^{-1/2}Z_2'),$$

where $V_{Z_1}^{-1}$ is again defined by $(Z_1'Z_1)^{-1}$, which metric spherizes the predictor space, i.e., correcting for the covariances between the variables in $Z_1$. Because $S_2^{-1/2}$ is a diagonal matrix as before, normalizing the coordinate vectors in the criterion space, redundancy analysis deals with Mahalanobis distances in the predictor space and Euclidean distances in the criterion space. Thus the STRESS pendant of (47) is written as

$$\text{STRESS}(X) = 1/2 \{ \|D(Z_1V_{Z_1}^{-1/2})D(X)\|^2 + \|D(X)D(Z_2S_2^{-1/2})\|^2 \}, \quad (48)$$

to be minimized over the comparison space $X$. Since (48) is simply a special case of (3a) - the analysis concerns two spaces with a different metric - we could generalize redundancy analysis to asymmetric analyses of more than two sets of variables. (The same is not true for the approach displayed in (47), because it is not possible to make a distinction between sets while at the same time incorporating the correlations between the variables).
By introducing the comparison space $X$, the variables in $Z_1$ are treated differently from what is more traditional in redundancy analysis. In the present version, the influence of $Z_1$ is channeled through a set of latent variables $X$, hopefully filtering out undesirable variation. This low-dimensional $X$ in turn should predict $Z_2S_2^{-1/2}$ as well as possible. In other approaches to redundancy analysis a linear combination of the predictor set $Z_1A_1$ is sought directly in order to explain as much variance as possible in the criterion set. This alternative objective can also be pursued in the distance framework by imposing additional restrictions on $X$ that satisfy $X = Z_1A_1$. This requirement amounts to replacing loss function (48) by

$$\text{STRESS}(A) = \|D(Z_1A_1) - D(Z_2A_1^{-1/2})\|^2,$$

(49)

which says that the distances in the high-dimensional criterion space $Z_2S_2^{-1/2}$ must be approximated by distances in a $p$-dimensional space given by $Z_1A_1$, a linear combination of the predictor variables in $Z_1$. This problem can be viewed as a multidimensional scaling problem with restrictions on the representation space, comparable to the type of constraints in MDS proposed by Carroll, Green and Carmone (1976), Bentler and Weeks (1978), Bloxom (1978), and De Leeuw and Heiser (1980).

Optimal transformations

The loss functions (48) and (49) can both be extended to incorporate optimal transformations for $Z_1$ and $Z_2$. For redundancy analysis using a comparison space $X$ as in (48) the optimal transformation problem is different for each of the two sets, where the solution for transforming the predictor space $Z_1$ has been discussed in the context of generalized canonical analysis in section 4, and the solution for the criterion space $Z_2$ in the context of principal components analysis in section 3. Introducing the transformations in (49), however, has to be solved by yet another approach, explored in the constrained MDS
context by Meulman and Heiser (1984). The loss function that is minimized when analyzing two sets of multivariate data is written as

\[ \text{STRESS}(Q_1; Q_2; A_1) = \| D(Q_1A_1) - D(Q_2S_{Q_2}^{-1/2}) \|^2. \] (50)

Minimization over \( Q_2 \) follows the approach in section 3. The value of the STRESS function (50) is decreased over \( Q_1 \) and \( A_1 \) by first computing the Guttman transform

\[ Y = n^{-1}B(Q_1A_1)Q_1A_1, \quad \text{with } b_{ik}^j(Q_1A_1) = d_{ik}(Q_2S_{Q_2}^{-1/2})/d_{ik}(Q_1A_1), \] (51)

and the additional definitions as in 27b and 27c. If we define a class of cones for all admissible transformations of the variables in \( Q_1 \), we next have to minimize \( \| Y - Q_1A_1 \|^2 \) over \( A_1 \) and over \( q_j \in C_j \). Because the transformation of the variable \( q_j \) is dependent on the other variables in the set \( Q_1 \), their contribution has to be corrected for. Therefore an auxiliary matrix \( U_j \) is defined as \( U_j = Y - (Q_1)_j(A_1)_j \), where \( (Q_1)_j \) indicates that the \( j \)th column-vector has been omitted from \( Q_1 \) and \( (A_1)_j \) that the \( j \)th row-vector has been omitted from the weights \( A_1 \). This enables us to rewrite the metric projection problem with respect to variable \( j \) as \( \| U_j - q_ja_j \|^2 \), which can be solved by simple alternating least squares.

Multiple regression

A modification of a multiple regression analysis along the lines of the distance approach can be derived straightforwardly from the discussion on redundancy analysis, i.e., considering the case where the predictor set \( Z \) contains \( m \) variables and criterion set contains a single normalized variable \( y \). The approach that incorporates an intermediate space between the criterion and the predictor space as in (48) has very interesting properties. The solution in the conventional framework displayed in (47) is found by an eigenanalysis of the matrix \( 1/2(ZV_Z^Z'Z' + yy') \), which matrix, however, has a very special eigenvalue structure, as is discussed in Meulman and Heiser (1988). The
largest eigenvalue turns out to be equal to $1/2\{1+(y'Z_1V^{-1}_{Z_1}Z_1'y)^{1/2}\}$, the smallest nonzero eigenvalue equals $1/2\{1-(y'Z_1V^{-1}_{Z_1}Z_1'y)^{1/2}\}$, and the $m-1$ intermediate eigenvalues are all equal to $1/2$. This eigenvalue structure shows that $p$-dimensional solutions for $X$ are not unique: many solutions are possible, which have the first dimension in common and $p-1$ subsequent dimensions to be arbitrarily chosen from $m-1$ eigenvectors that go with eigenvalues that are equal.

Minimization of $\text{STRESS}(X) = 1/2 \{ \| D(Z_1V^{-1/2}_{Z_1}) - D(X) \|^2 + \| D(X) - D(y) \|^2 \}$, however, gives a unique $p$-dimensional $X$ when the solution is rotated towards its principal axes orientation. Consequently, extending the concept of an intermediate space into the distance framework gives the possibility of interesting graphical displays through the use of two dimensions in $X$ (cf. Meulman and Heiser, 1988).

6. Discussion

In this paper we have stayed quite close to the conventional multivariate analysis domain in the sense that representation spaces in multivariate analysis turned out to be solutions to particular classical scaling problems with respect to distances between objects in given observation spaces. From this starting point the classical approach could be replaced by least squares distance fitting, giving distance approximations from above and below. Without derogating the important data analytical merits of the possibility to include optimal transformations of the variables, approximation from above and below is considered as the most important characteristic of the alternative MVA techniques discussed, as opposed to approximation from below in conventional MVA techniques.

To appreciate the difference in approximation, the objectives of parametric mapping (Shepard and Carroll, 1966) can serve as a guide. Although until recently parametric mapping has been neither a generally well known nor a frequently applied technique, its
objectives are quite fascinating (cf. Winsberg and Carroll, 1988). As the geometrical approach to multivariate analysis, parametric mapping compares an observation space with a representation space. Its leading case is the situation where the points in the observation space lie in some limited region which takes the form of a possibly curved subspace of fewer dimensions. Following Kruskal (in Kruskal and Carroll, 1969), parametric mapping tries in such cases to open out this curved subspace onto a flat subspace of the same dimensionality. The representation space serves as the platform onto which the curved region is to be flattened out. Such a flattening out is a called a mapping between the region of the observation space and some suitable region of the representation space. The mapping is guided by acknowledging features in the representation space to be agents in the observation space, and not the other way around. In fact, this is the crucial idea: to require continuity between the distances in the representation space and the distances in the observation space. Carroll proposed to measure the difference between the distances in the two spaces by a continuity index: a weighted average of the $n(n-1)$ ratios of distances in the observation space to distances in the representation space, where the weights are the inverse of some power of the distances in the representation space. (Since larger values of the index indicate less continuous mappings, the index is better called a discontinuity index). The definition of discontinuity implies that a small distance in the observation space that is represented by a small distance in the representation space contributes little to the value of the discontinuity index, which is also true when it is represented by a large distance in the representation space. The mapping between a large distance in the observation space and a large distance in the representation space is also considered to be continuous, but the mapping between a large distance in the observation space and a small distance in the representation space is considered to be highly discontinuous: the ratio between the two distances increases the value of the discontinuity value considerably because it is heavily weighted by the inverse of a power of the small distance in the
representation space. The value of the discontinuity index will always be quite large for conventional MVA solutions, which is caused by the approximation from below. In the case of a curved subspace in the observation space, conventional MVA will not be able to unfold this subspace; on the contrary, the solutions will tend to fold structures from the observation space in the representation space. By contrast, the alternative MVA methods discussed in this paper will approximate a subset of the distances in the observation space from above, which is considered to be continuous according to the parametric mapping criterion.

**Perspectives**

This line of thought basic can be extended in a number of ways. In the first place, we can combine features from the conventional MVA approach, which seeks weighted sums of variables, and the alternative approach, which seeks to fit distances. In the latter, we rely on the property that least squares distance fitting is essentially coordinate free, and consequently, the use of the distance operator \( D(\cdot) \) makes it possible to compare subspaces that are not of the same dimensionality. A first application of differentially weighting of the variables generalizes the canonical analysis from section 4. Instead of approximating distances in the full-dimensional observation spaces, distances can be fitted to reduced canonical spaces with dimensionality \( r_j \leq m_j \). The loss function that implements this objective is written as

\[
\text{STRESS}(Q;X;A) = M^{-1} \sum_{j=1}^{M} ||D(Q_j A_j) - D(X)||^2,
\]

(52)

to be minimized over \( X \), over \( Q \) satisfying the transformation restrictions, and over \( A \) satisfying the normalization restrictions \( A_j^T Q_j Q_j A_j = I \), i.e., \( Q_j A_j \) is an \( n_j \times r_j \) matrix with the rows containing the coordinates of the objects on the new set of \( r_j \) axes of an orthonormal space. In contrast with (19), each space \( Q_j A_j \) may have its own
dimensionality \( r_f \) chosen so \( p \leq r_f \leq m_f \). The objective would be to remove unimportant information from the full-dimensional observation space, and to look for aspects that the \( M \) reduced spaces have in common from the remaining information. Another possibility is to choose \( r_f \leq p \leq m_f \), with a possible application in the analysis of three-way data mentioned in Meulman (1989).

The idea of differentially weighting can also be applied to the principal components analysis from section 3. The loss function that implements distance approximation in a reduced principal space is written as \( \text{STRESS}(Q;X;A) = \|D(QA) - D(X)\|^2 \). Although this objective function is similar to (52), the constraints on \( A \) are different, because it is required that \( A \) is an \( m_f \times r_f \) rotation-annihilation matrix that first rotates the \( m_f \) axes of \( Q \) to some (optimal) position, and then annihilates the last \( m_f - r_f \) axes, where \( m_f \geq r_f \). To secure that \( A \) is a rotation-annihilation matrix, the objective function should be minimized under the condition that \( A'A = I \).

A second development that also implements weights in the loss function is in the domain of fitting individual difference models. Meulman and Verboon (1989) revisit the Points of View analysis model of Tucker and Messick (1963), where the data are a series of \( n \times n \) matrices \( \Delta_j, j = 1, \ldots, m \), with elements \( \{\delta_{ikj}\} \), which are given dissimilarity measures. It is proposed to minimize

\[
\text{STRESS}(\Delta^*_1, \ldots, \Delta^*_m; X_1, \ldots, X_r; A) = \sum_{j} \sum_{s} \sum_{i} \sum_{j} (a_{js} \delta_{ikj} - d_{ik}(X_s))^2. \tag{53}
\]

This loss function can be viewed as a particular formulation of nonlinear PCA on the basis of homogeneity analysis (cf. Gifi, 1981) and has to be minimized over three sets of parameters. The monotonic transformations of the dissimilarities, denoted by \( \delta_{ijk}^{*} \), constitute the nonlinear PCA aspect. The matrix \( A = [a_{ks}] \) contains the component loadings. These should be restricted to enforce distinct points of view. What is special in (53) is the requirement that the component scores \( d_{ik} \) are Euclidean distances between objects \( i \) and \( k \) in a low-dimensional configuration \( X_s \); therefore (53) should be minimized
over $X_1, \ldots, X_r$, the different points of view, to ensure optimality with respect to the MDS objective.

As an alternative Carroll and Chang (1970) proposed the INDSCAL and the IDIOSCAL model (Carroll and Chang, 1972). These types of models can also be fitted in the distance approach to MVA (Meulman and Heiser, 1989). Here transformations are introduced for the dimensions of the common space $X$, and the objective function is written as.

$$\text{STRESS}(X;Q;A) = \sum_j^m \|D(Q_j) - D(XA_j)\|^2,$$

(54)

where $A_j$ may be diagonal (the INDSCAL model), of full rank (the IDIOSCAL model), or may fit the reduced rank model, as in Heiser and Stoop (1986); the full and reduced rank model have also been discussed under the name general Euclidean model in Bloxom (1978) and Young (1984).

The above mentioned methods that incorporate differentially weighting need further study as to their data analytical merits. Although the techniques discussed in the sections 3-5 have all been applied to various sets of data, and compared to conventional MVA results in terms of the transformation of the variables and the approximation in representation space (De Leeuw and Meulman, 1986a; Meulman, 1986; 1987; 1988; Meulman and Heiser, 1988), little is yet known about their behavior with respect to the local minima problem that is acknowledged for MDS in general. Moreover, the existence of local minima constitutes only one aspect of possible instability of the solution. Another aspect concerns the possibility that the results of the analysis are obtained by focussing on incidental aspects of the data. For the study of this type of instability resampling methods have been shown to be fruitful in the context of nonlinear MVA (Gifi, 1981; Meulman 1982; Van Rijckevoorsel, Bettonvil and De Leeuw, 1986; Van der Burg, 1987). Since the important bootstrap method (Efron, 1979; 1982; among others) is not very feasible when the objects are not considered to be a replication factor, the special jackknife that was
developed for MDS by De Leeuw and Meulman (1986b) can be used as a form of influence analysis for the techniques discussed in this paper.

References


Carroll, J. D. (1968). Generalization of canonical correlation analysis to three or more sets of variables. *Proceedings of the 76th annual convention of the American Psychological Association, 3*, 227-228.


