# A DISTANCE-BASED BIPLOT FOR MULTIDIMENSIONAL SCALING OF MULTIVARIATE DATA

**Jacqueline J. Meulman**

**Department of Data Theory**

**Leiden University**

**Leiden, The Netherlands**

# A Distance-Based Biplot for Multidimensional Scaling of Multivariate Data

**Summary:** Least squares multidimensional scaling (MDS) methods are attractive candidates to approximate proximities between subjects in multivariate data (Meulman, 1992). Distances in the subject space will resemble the proximities as closely as possible, in contrast to traditional multivariate methods. When we wish to represent the variables in the same display - after using MDS to represent the subjects - various possibilities exist. A major distinction is between linear and nonlinear biplots. Both types will be discussed briefly, including their drawbacks. To circumvent these drawbacks, a third alternative will be proposed. By expanding the optimal $p$-space (where $p$ denotes the dimensionality of the subject space) into an $m$-dimensional space of rank $p$ (with $m > p$), we obtain a coordinate system that is appropriate for the evaluation of the MDS solution directly in terms of the $m$ original variables. The latter are represented graphically as vectors in $p$-space, and their entries as markers that are located on these vectors. The overall approach, including the analysis of mixed sets of continuous and categorical variables, can be viewed as a distance-based alternative for the graphical display of multivariate data in Gifi (1990).

## 1. Introduction

In the approach to Multivariate Analysis (MVA) applied in this paper, the variables are used to define an observation or measurement space in which the units are located according to their scores. The distances in this observation space are regarded as proximities to be approximated by distances between subject points in a low-dimensional representation space. If the $m$-dimensional observation space is denoted by $\mathbf{Q}$, giving coordinates for $n$ points in $m$ dimensions, the proximities between all pairs of subjects are given in the proximity matrix $D(\mathbf{Q})$, where $D(\bullet)$ is the Euclidean (also called Pythagorean) distance function. So the $n \times n$ matrix $D(\mathbf{Q})$ contains proximities $d_{ik}(\mathbf{Q})$ between subject $i$ and $k$. In matrix formulation:

$$D^2(\mathbf{Q}) = \mathbf{v1}' + \mathbf{1v}' - 2\mathbf{QQ}', \tag{1}$$

where $\mathbf{v} = \text{vecdiag}(\mathbf{QQ}')$ is the $n$-vector containing the diagonal elements of $\mathbf{QQ}'$, and $\mathbf{1}$ is the $n$-vector of all 1's. Analogously, distances in the $p$-dimensional representation space $\mathbf{X}$ are defined by $D^2(\mathbf{X}) = \mathbf{v1}' + \mathbf{1v}' - 2\mathbf{XX}'$, now with $\mathbf{v} = \text{vecdiag}(\mathbf{XX}')$.

Approximation of a set of proximities by a set of distances in some low-dimensional space is usually identified as a multidimensional scaling (MDS) task. In Meulman (1986), following Gower (1966), it is shown that techniques of multivariate analysis, like principal components, canonical correlation and homogeneity analysis, are equivalent to MDS tasks applied to particular derived proximities when the so-called classical Torgerson-Gower approach to MDS (Torgerson, 1958; Gower, 1966) is used. A basic ingredient is the original Young-Householder (1938) process that transforms a squared distance matrix $D^2(\mathbf{Q})$ into an $n \times n$ scalar product matrix $\mathbf{QQ}'$, modified by locating the origin in the centroid of points through the use of the $n \times n$ centering operator $\mathbf{J} = \mathbf{I} - (\mathbf{11}'/\mathbf{1}'\mathbf{1})$, which gives

$$-1/2\mathbf{J}(D^2(\mathbf{Q}))\mathbf{J} = -1/2\mathbf{J}(\mathbf{v1}' + \mathbf{1v}' - 2\mathbf{QQ}')\mathbf{J} = \mathbf{QQ}'. \tag{2}$$

($\mathbf{I}$ is the $n \times n$ identity matrix; we assume that the variables $\mathbf{q}_j$ have zero mean.) The scalar product matrix $\mathbf{QQ}'$ is then approximated by another scalar product matrix of lower rank, using an objective function that can be written in the form:

$$\text{STRAIN}(\mathbf{X}) \;=\; ||\mathbf{QQ}' - \mathbf{XX}'||^2 = \text{tr}\,(\mathbf{QQ}' - \mathbf{XX}')'(\mathbf{QQ}' - \mathbf{XX}'), \qquad (3)$$

so $|| \bullet ||^2$ denotes a least squares discrepancy measure. (The term STRAIN is used after Carroll and Chang, 1972.) In our case, because $\mathbf{XX}' = -1/2\mathbf{J}(D^2(\mathbf{X}))\mathbf{J}$, (3) can also be written as

$$\text{STRAIN}(\mathbf{X}) \;= (1/4)||\mathbf{J}(D^2(\mathbf{Q}) - D^2(\mathbf{X}))\mathbf{J}||^2, \qquad (4)$$

so Torgerson-Gower scaling approximates double-centered squared distances. To find the coordinates in $\mathbf{X}$, first an eigenanalysis of $\mathbf{QQ}'$ is performed: $\mathbf{QQ}' = \mathbf{K\Lambda K}'$, where $\mathbf{K}$ is an $N \times t$ matrix containing $t$ eigenvectors, $\mathbf{\Lambda}$ is a $t \times t$ diagonal matrix containing the ordered positive eigenvalues, and $t$ denotes the rank of $\mathbf{Q}$ (for $t \leq m$). The optimal solution in the Torgerson-Gower procedure for obtaining a $p$-dimensional $\mathbf{X}$ (for $p \leq t$), is given by $\mathbf{X} = \mathbf{K}_p\mathbf{\Lambda}_p^{1/2}$; thus, the eigenvectors are rescaled using the eigenvalues to give coordinates $\mathbf{X}$ for the units in $p$-space with dimensions that reflect differential saliences.

The multivariate case described above, with proximities $D(\mathbf{Q})$, is usually called Principal Coordinates Analysis (Gower, 1966); using the singular value decomposition $\mathbf{Q} = \mathbf{K\Lambda L}'$, it can easily be shown that the solution for $\mathbf{X}$ that would be obtained in a principal components analysis is equivalent to the solution for $\mathbf{X}$ obtained by minimizing (3) or (4). Using the same strategy, it was shown in Meulman (1986, 1992) that Multiple Correspondence Analysis (MCA), also called homogeneity analysis, can be viewed as a classical scaling technique as well. In MCA, categorical variables are analyzed, and each categorical variable $\mathbf{h}_j$ defines a binary indicator matrix $\mathbf{G}_j$ with $n$ rows and $l_j$ columns, where $l_j$ denotes the number of categories. Elements $h_{ij}$ then define elements $g_{ir}^j$ as follows:

$$\begin{aligned} h_{ij} &= r \longrightarrow g_{ir}^j = 1; \\ h_{ij} &\neq r \longrightarrow g_{ir}^j = 0, \end{aligned} \qquad (5)$$

where $r = 1, ..., l_j$ is the running index to indicate the category number of variable $j$. In analogy with (3) and (4), MCA can be written as a classical MDS problem since its optimal solution for the subject scores $\mathbf{X}$ minimizes

$$\begin{aligned} \text{STRAIN}(\mathbf{X}) &= 1/m\sum_{j=1}^{m}||\mathbf{G}_j(\mathbf{G}_j'\mathbf{G}_j)^{-1}\mathbf{G}_j' - \mathbf{XX}'||^2 \qquad (6) \\ &= 1/m\sum_{j=1}^{m}||\mathbf{J}(D^2(\mathbf{G}_j(\mathbf{G}_j'\mathbf{G}_j)^{-1/2}) - D^2(\mathbf{X}))\mathbf{J}||^2, \end{aligned}$$

where proximities are derived simultaneously from all $\mathbf{G}_j$ separately. The columns of the indicator matrix $\mathbf{G}_j$ are divided by the square root of the marginals $\mathbf{G}_j'\mathbf{G}_j$; the latter operation defines the chi-squared metric. Finally, the proximities are approximated by Euclidean distances in $\mathbf{X}$. As before, in the classical scaling approach, $\mathbf{X}$ would not be normalized to represent the subjects in an orthonormal cloud, but instead the eigenvalues are used to give the representation space a certain shape, displaying the differential saliences.

In Meulman (1986, 1992) an alternative is proposed, which is to analyse multivariate data by minimizing a loss function that is directly defined on the distances (so it does not approximate distances through inner products). The history of least squares MDS methods can be followed from Shepard (1962), Kruskal (1964), Guttman (1968), Takane, Young, and De Leeuw (1977), De Leeuw and Heiser (1980), Ramsay (1982), a.o. Least squares MDS methods are traditionally applied to a given proximity matrix, whose proximities are then approximated through minimization of some least squares loss function that is defined on (transformations of) proximities and distances in a representation space $\mathbf{X}$. In the multivariate cases described above, we derive the proximities from the multivarate data. Then, in the distance-based modification of principal components analysis (distance-based PCA, for short), we minimize

$$\text{STRESS}(\mathbf{X}) = ||D(\mathbf{Q}) - D(\mathbf{X})||^2 \tag{7}$$

over $\mathbf{X}$. As before, $D(\bullet)$ is an $n \times n$ matrix with Euclidean distances between subjects, and $\mathbf{X}$ is the low-dimensional space in which distances should match the proximities as closely as possible. To minimize loss function (7), we have to use an iterative procedure since there is no analytic solution. In the present paper, the majorization algorithm for MDS has been used (e.g., see De Leeuw and Heiser, 1980; Groenen and Heiser, 1996). The majorization approach amounts to computing an update for $\mathbf{X}$ that reduces the value of (7) from a starting point $\mathbf{X}^o$ by:

$$\mathbf{X} = 1/nB(\mathbf{X}^o)\mathbf{X}^o. \tag{8}$$

Here, the $n \times n$ matrix $B(\mathbf{X}^o)$ is defined as

$$B(\mathbf{X}^o) = B^+(\mathbf{X}^o) - B^*(\mathbf{X}^o), \tag{9}$$

where the elements of the matrix $B^*(\mathbf{X}^o)$ are given by $b^*_{ik}(\mathbf{X}^o) = d_{ik}(\mathbf{Q})/d_{ik}(\mathbf{X}^o)$ if $i \neq k$ and $d_{ik}(\mathbf{X}^o) \neq 0$; otherwise $b^*_{ik}(\mathbf{X}^o) = 0$. The elements of the diagonal matrix $B^+(\mathbf{X}^o)$ are given by $b^+_{ik}(\mathbf{X}^o) = \mathbf{1}'B^*(\mathbf{X}^o)\mathbf{e}_i$, where $\mathbf{e}_i$ is the $i$th column of the identity matrix $\mathbf{I}$. Repeatedly computing the update $\mathbf{X}$ gives a convergent series of configurations.

A special feature of the Gifi-system is the possibility of differential treatment of variables in the analysis. For example, some variables may be treated as containing numerical scores, while others may be treated as nominal variables. The latter treatment is appropriate when a variable partitions the subjects into unordered classes. In distance-based PCA nominal treatment of variables is carried out as follows. First, the nominal variable $\mathbf{h}_j$ is replaced by a binary indicator matrix $\mathbf{G}_j$ with $n$ rows and $l_j$ columns, as above. Then, proximities are derived simultaneously from $\mathbf{Q}$ and $\mathbf{G}_j$:

$$\Delta(\mathbf{Q}; \mathbf{G}) = D(\mathbf{Q}; \mathbf{G}_j(\mathbf{G}_j'\mathbf{G}_j)^{-1/2}) \tag{10}$$

where $j = 1, ..., m$ is the running index to indicate the variables in the analysis that classify the subjects into groups (there may be more than one classifying variable, and then multiple indicator matrices should be created). As in MCA and homogeneity analysis, the columns of the indicator matrix $\mathbf{G}_j$ are divided by the square root of the marginals $\mathbf{G}_j'\mathbf{G}_j$ to give distances in the chi-squared metric. Finally, the proximities $\Delta(\mathbf{Q}; \mathbf{G})$ are approximated by Euclidean distances $D(\mathbf{X})$.

Meulman (1992) has shown that if we apply the classical scaling approach (as in Torgerson, 1958; Gower, 1966) to approximate $\Delta(\mathbf{Q}; \mathbf{G})$ then this results in a solution for $\mathbf{X}$ that is equivalent to the subject scores in Gifi's PCA, with numerical and nominal variables. (Again, apart from a scaling factor per dimension, displaying the differential saliences; PCA usually displays the subject points as an orthonormal cloud, with equal dispersions.)

## 2. Supplementary representation of variables in a MDS solution

The idea of joint representation of subjects and variables, which originates with Tucker (1960), was subsequently very succesfully applied in the analysis of preference data (Carroll, 1972), and has become well-known as the biplot (Gabriel, 1971); the classical reference to the basic idea of lower-rank approximation is Eckart and Young (1936). A recent book on biplots is Gower and Hand (1996). From the biplot point of view, principal components analysis can be regarded as a bilinear model (Kruskal, 1978). (A variable-oriented approach would regard PCA as the analysis of a correlation matrix.) In the bilinear model, we minimize

$$\sigma(\mathbf{X}) = ||\mathbf{Q} - \mathbf{X}\mathbf{A}'||^2, \tag{11}$$

over $\mathbf{X}$ and $\mathbf{A}$: the observed scores in the $m$-dimensional space $\mathbf{Q}$ are approximated by the inner product of the $p$-dimensional component scores $\mathbf{X}$ and component loadings $\mathbf{A}$ (with $p$ much smaller than $m$). Graphically, in the biplot representation, subjects are represented as points, and variables as vectors, and the orthogonal projection of the subject points onto the variable vectors gives an approximation of the observed scores.

In the framework of distance-based PCA, as in (7), variables can be considered from an *internal* and an *external* perspective. From the internal perspective, their role is to provide the proximities between the subjects. From the external perspective, the variables can be used afterwards to study whether we can account for the structure among the subjects. The latter can be done through linear and nonlinear *external* biplot methods. We call these biplots external, because the variables are fitted into the subject space in a second step, while the subject points $\mathbf{X}$ are kept fixed. By contrast, in internal biplots, the subject points and the vectors representing the variables are found simultaneously.

### 2.1 Linear external biplots through multiple regression

A straightforward way to fit a set of variables in a given configuration is through "property fitting" (e.g., Carroll, 1972; also, see Meulman, Heiser and Carroll, 1987). For distance-based PCA, this amounts to the projection of a variable $\mathbf{q}_j$ into the space of the subjects $\mathbf{X}$ by the use of multiple regression. In the regression, the columns in $\mathbf{X}$ are the independent variables, and the weights obtained from the regression determine the coordinates for the variable $\mathbf{q}_j$ in the space $\mathbf{X}$. The optimal direction $\mathbf{a}_j$for the vector representing variable $\mathbf{q}_j$ in the $p$-space $\mathbf{X}$ is thus found as

$$\mathbf{a}_j = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{q}_j. \tag{12}$$

Using $\mathbf{a}_j$ to represent the endpoint of the vector gives a linear biplot representation. This biplot is obtained, however, through the use of different rationales for fitting the subjects on the one hand and the variables on the other: the subject points are fitted through the use of least squares distance fitting in (7), and the vectors representing the variables through ordinary multiple regression in (12).

### 2.2 Nonlinear external biplots through unfolding

As a possible alternative, coherent, method, Meulman and Heiser (1993) proposed a least squares generalization of the so-called nonlinear biplot in Gower and Harding

(1988). The latter nonlinear biplot was developed to obtain nonlinear representations of variables in a space that is generated through a principal coordinates analysis. The procedure discussed in Meulman and Heiser (1993) can be described as follows. First, regard each variable as a series of $s = 1, ..., S$ supplementary points (a trajectory) in the space $\mathbf{X}$. In terms of the data, a supplementary point for variable $j$ has coordinates $\mathbf{e}_{r_s}$ in $\mathbf{X}$ that are all equal to zero, except for the $j$th variable; so when $\mathbf{e}_j$ is the $j$th column of the $m \times m$ identity matrix $\mathbf{I}$, $\mathbf{e}_{r_s} = r_s \mathbf{e}_j$, where $min(\mathbf{q}_j) \leq r_s \leq max(\mathbf{q}_j)$. Next, for each supplementary point, the distance is calculated to the $n$ original points in observation space. The vector with squared distances between supplementary point $\mathbf{e}_{r_s}$ and the subjects in $\mathbf{Q}$ is given by

$$d^2(\mathbf{e}_{r_s}; \mathbf{Q}) = \mathbf{v} + \mathbf{1}\mathbf{e}'_{r_s}\mathbf{e}_{r_s} - 2\mathbf{Q}\mathbf{e}_{r_s} = \mathbf{v} + \mathbf{1}r_s^2 - 2\mathbf{Q}\mathbf{e}_{r_s}, \tag{13}$$

with $\mathbf{v}$ the $n$-vector containing the diagonal elements of $\mathbf{QQ}'$. The $k$th element of $d^2(\mathbf{e}_{r_s}; \mathbf{Q})$ gives the squared distance between the $s$th supplementary point and the $k$th subject point in observation space, and will be written as $d^2(\mathbf{e}_{r_s}; \mathbf{q}_k)$, where $\mathbf{q}_k$ denotes the $k$th row of $\mathbf{Q}$. Mapping the trajectory for variable $\mathbf{q}_j$ involves the approximation of $d(\mathbf{e}_{r_s}; \mathbf{Q})$ by $d(\mathbf{y}_s; \mathbf{X})$, where $\mathbf{y}_s$ gives $p$-dimensional coordinates in the space of $\mathbf{X}$, for different values of $r_s, s = 1, ..., S$. Here $S$ denotes a prechosen number, appropriate to cover the range of $min(\mathbf{q}_j)$ to $max(\mathbf{q}_j)$. Each supplementary point has to be mapped separately, and the coherent method with respect to least squares multidimensional scaling as in (7) is the use of

$$\text{STRESS}(\mathbf{y}_s) = ||d(\mathbf{e}_{r_s}; \mathbf{Q}) - d(\mathbf{y}_s; \mathbf{X})||^2, \tag{14}$$

to be minimized over $\mathbf{y}_s$ for given $\mathbf{Q}$ and $\mathbf{X}$. The loss function (14) represents the least squares external unfolding problem; it is called unfolding because it fits distances between two sets of points, $\mathbf{X}$ and $\mathbf{y}$, and it is called external because $\mathbf{X}$ is known and fixed. There is no closed-form solution for STRESS-based external unfolding as in (14), so the loss function has to be minimized iteratively. The latter can be done using the SMACOF framework for unfolding (Heiser, 1981; 1987). The points $y_1, ...y_s, ..., y_S$, mapped in $\mathbf{X}$ by minimizing (14), will in general not be on a straight line, because (14) represents a nonlinear mapping. Nonlinear biplot representations have interesting properties, but some of them are not yet fully understood and are currently under study (Groenen and Meulman, 1995). In most cases, nonlinear biplots are harder to interpret than linear ones, and therefore here a biplot is presented that is linear and, unlike regression, at the same time consistent with the criterion minimized in distance-based PCA.

## 3. The Distance-Based Biplot

The basic notion to arrive at a simple, linear, and coherent biplot is that distances $D(\bullet)$ are invariant under rotation, i.e., $D(\mathbf{X}) = D(\mathbf{XA}')$ if $\mathbf{A}'\mathbf{A} = \mathbf{I}$. Usually, a rotation matrix is of the order $p \times p$; here, however, we will consider a matrix $\mathbf{A}$ of order $m \times p$. This matrix will be labeled a *rotation-expansion* matrix, because the transformation preserves the distances in $\mathbf{X}$, and at the same time expands the representation space. Thus, the coordinates $\mathbf{X}$ in $p$-space are replaced by $m$-dimensional coordinates in the space $\mathbf{XA}'$ of rank $p$, and since $\mathbf{A}'\mathbf{A} = \mathbf{I}$

$$\text{STRESS}(\mathbf{X}) \;=\; ||D(\mathbf{Q}) - D(\mathbf{X})||^2 = ||D(\mathbf{Q}) - D(\mathbf{XA}')||^2. \tag{15}$$

To obtain the biplot coordinates, we have to minimize $||\mathbf{Q} - \mathbf{XA}'||^2$ over all $\mathbf{A}$ satisfying $\mathbf{A}'\mathbf{A} = \mathbf{I}$. This amounts to an orthogonal Procrustes problem of the order $p \times p$ that is solved as follows.

Define the singular value decomposition of the $m \times p$ matrix $\mathbf{Q'X}$ as $\mathbf{Q'X} = \mathbf{K\Lambda L'}$, and the eigenvalue decomposition of the $p \times p$ matrix $\mathbf{X'QQ'X}$ as $\mathbf{X'QQ'X} = \mathbf{L\Lambda^2 L'}$. Then the rotation-expansion matrix is found by

$$\mathbf{A} = \mathbf{KL'} = \mathbf{Q'XL\Lambda^{-1}L'}. \tag{16}$$

Now the $m$-dimensional coordinate system $\mathbf{XA'}$ can be used to evaluate the MDS solution directly in terms of the original variables, with the Pearson correlation coefficient as a natural measure of association. At the same time, the $j$th row in $\mathbf{A}$ (denoted by $\mathbf{a}'_j$) gives the coordinates to display the variable $\mathbf{q}_j$ in the space $\mathbf{X}$. The scores $\{q_{ij}\}$ themselves can be represented as well; the projected coordinates in the space $\mathbf{X}$ are given by $\mathbf{q}_j\mathbf{a}'_j/\mathbf{a}'_j\mathbf{a}_j$.The latter set of quantities are called single category coordinates in Gifi's (1990) approach to PCA. Therefore, the approach proposed here can be regarded as a distance-based alternative for Gifi's biplot display of multivariate data. The series of points $\mathbf{q}_j\mathbf{a}'_j/\mathbf{a}'_j\mathbf{a}_j$ are located on the vector that represents the variable $\mathbf{q}_j$, and these are usually called markers (Gabriel, 1971; Gower and Hand, 1996).

## 4. Material and Methods

The data that are used in the application of the distance-based biplot were collected by Van Strien and Van der Ham, at the Department of Psychiatry at Utrecht University Hospital (see Van der Ham, Meulman, Van Strien and Van Engeland, 1997). The data concern 16 variables that measure well-being at four points in time for 55 patients with eating disorders. The patients were diagnosed independently into four categories (using the DSM-III-R): 1. Anorexia Nervosa ($N_1 = 25$), 2. Anorexia with Boulimia Nervosa ($N_2 = 9$), 3. Boulimia Nervosa after Anorexia ($N_3 = 14$), and 4. Atypical Eating Disorder ($N_4 = 7$). The total number of patients $N = 55$, and data are available at four different time points, and the total number of observational units would be $4 \times 55 = 220$. However, there are a few subjects with missing observations at one of the four time points, so the actual number of observational units is 55 + 53 + 54 + 55 = 217.

In addition to the 16 variables that measure well-being, two additional nominal variables are used, the first a variable (with four categories) that associates each observational unit with a point in time, and the second a variable that links each patient with a diagnosis (in one of the four eating disorder categories; the diagnosis was established before time point one, and does not change over time). In summary, the multidimensional scaling analysis is applied to a $217 \times 217$ matrix with proximities between the observational units, with the proximities derived from 16 + 4 + 4 = 20 variables (time and diagnosis each have four categories). The analysis is intended to result in a subgrouping of the eating disorders in a longitudinal perspective.

The multidimensional scaling task was performed in two dimensions; the solution as a whole is judged on two different criteria. First, the correlation between the original variable $\mathbf{q}_j$ and the fitted $\mathbf{Xa}'_j$ is considered; variables were fitted into the subject space by the biplot method discussed in Section 3. The Pearson correlation coefficient is a natural goodness-of-fit measure for the biplot representation, since standard PCA maximizes the average squared correlation. So for the distance-based biplot, we can compare this particular goodness-of-fit measure with the optimum provided by standard PCA. Second, we consider classification of subjects on the basis of the Euclidean distances in two-dimensional space. Each subject is allocated to one of the four time category points and to one of the four diagnosis points. Specifically, the centroids of the subject points in a particular class define the associated category points. In addition to Euclidean distance, the allocation used the posterior probabilities, by employing Bayes rule and the a priori distribution over the categories. The resulting assignment is compared to the original time points and diagnostic group classification.

## 5. Results

The primary result of the analysis consists of the coordinates for the observational units in two-dimensional space; the graph displaying the cloud of points is given in Figure 1, top panel. The subjects have been labeled with their diagnosis. Visual inspection immediately suggests that the second dimension is related to the diagnostic categories of eating disorder. We see that the anorexia subjects (label 1) form a group, but patients with atypical eating disorder (label 4) form a subgroup. The latter patients can be considered as anorectic patients for whom the loss of weight is unknown or less than 15%. The second dimension separates the anorectic patients (classes 1 and 4) from the boulimic patients (classes 2: anorexia nervosa with boulimia nervosa, and 3: boulimia nervosa after anorexia). Having connected the diagnosis category points over the different points in time (by computing the appropriate centroids of subject points), it is clear that the first dimension displays the development in time. The variables are displayed in the bottom panel of Figure 1. Instead of displaying the variables and subjects together, it was chosen to display the variables with the group points, since this gives a more comprehensive biplot; groups of subjects are represented by their centroid, which is associated with a particular point in time and diagnosis.

The first thing to note is that all variables have a positive correlation with the first dimension; this means there is a general factor that correlates positively with all the variables. The second dimension separates the variables. We find three bundles of variables: Bingeing (4), Vomiting (5) and Purging (6) are clearly distinguished in the vertical direction of the graph; Preoccupation (15), Body Perception (16), Hyperactivity (7), Sexual Behavior (13), and Fasting (3) correlate most with the general factor, and Weight (1), Menstruation (2), Family Relations (8), Emancipation (9), Work/School record (11) and Sexual Attitude (12) form a third bundle of variables. Friends (10) and Mood (14) do not fit very well in the overall representation. The fit of the variables, as measured by the Pearson correlation between observed scores and fitted scores, is given in Table 1. We notice that distance-based PCA gives a very decent fit for the variables, compared to the maximum that could be obtained by applying standard PCA.

As described above, we compare the classification of the subjects on the basis of the results from both analyses with the original ones. With respect to time, the grouping is given in Table 2. Here, rows indicate the original time points, and columns the fitted time points. From the percentage correctly assigned subjects in the bottom row, we conclude that distance-based PCA performs better than standard PCA, although it is obviously hard to distinguish between consecutive time points, especially between 3 and 4. Next, we inspect assignment to the eating disorders categories. Results are given in Table 3, with the rows and columns ordered according to the subgroupings. It is clear that neither approach to PCA can distinguish between groups 1 and 4 on the one hand, and groups 2 and 3 on the other hand. On the whole, distance-based PCA performs better. This becomes more clear when we combine results from Table 3 in its bottom row: now standard PCA finds 83% and 91% correctly, while distance-based PCA obtains 92% and 97%.

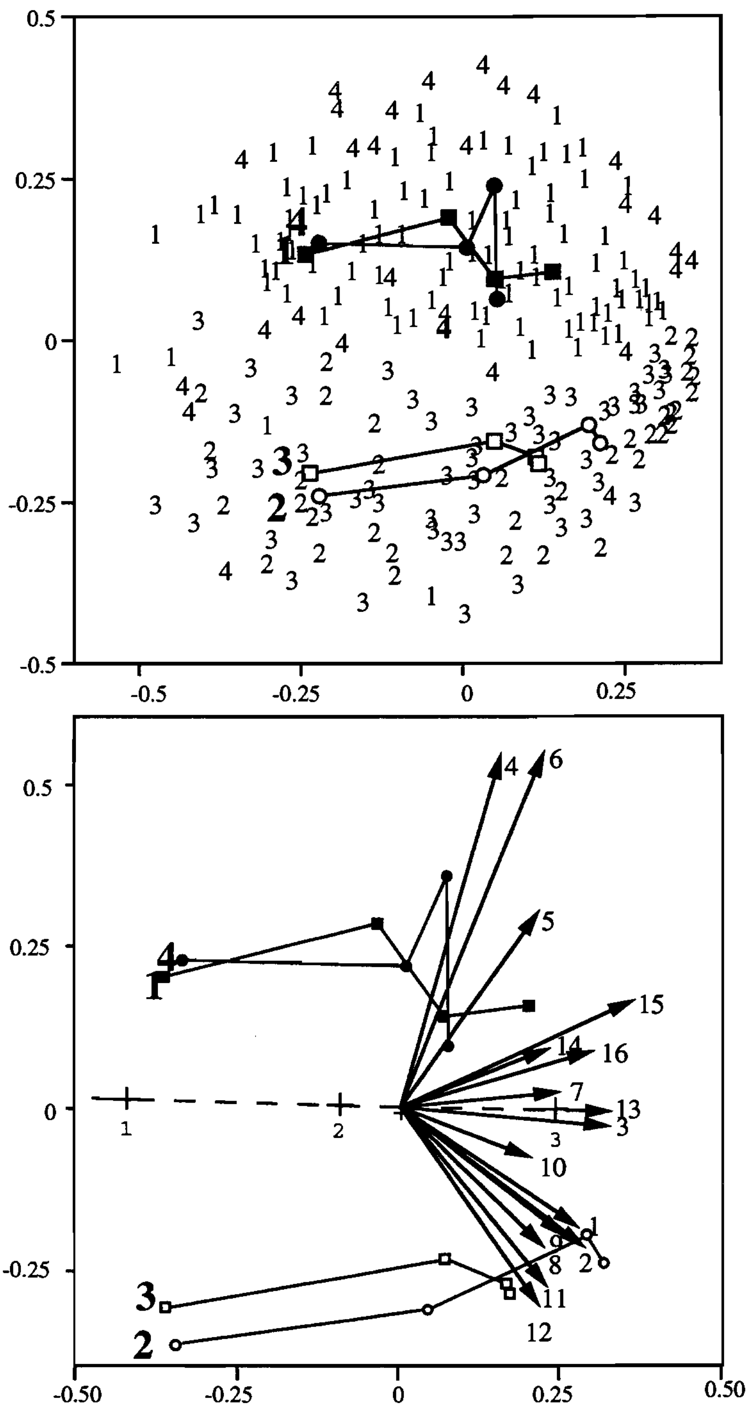


Figure 1. Top panel: Subjects represented in two-dimensional space. Labels 1: anorexia, 2: anorexia with boulimia, 3: boulimia after anorexia, 4: atypical eating disorder. Trajectories represent the diagnostic categories in time (from left to right). Lower panel: Trajectories, now displayed with the variables (described in Table 1). Markers represent the three categories for variable 13.

Table 1: Correlations between Observed Variables and Fitted Variables in Standard and Distance-Based PCA

| Variables | Standard PCA | D-Based PCA |
|---|---|---|
| Weight | 0.69 | 0.68 |
| Menstruation | 0.72 | 0.72 |
| Fasting | 0.70 | 0.72 |
| Bingeing | 0.79 | 0.71 |
| Vomitting | 0.62 | 0.55 |
| Purging | 0.88 | 0.79 |
| Hyperactivity | 0.53 | 0.53 |
| Family Relations | 0.58 | 0.58 |
| Emancipation | 0.64 | 0.64 |
| Friends | 0.46 | 0.44 |
| Work/School | 0.60 | 0.63 |
| Sexual Attitude | 0.62 | 0.62 |
| Sexual Behavior | 0.70 | 0.71 |
| Mood | 0.47 | 0.50 |
| Preoccupation | 0.76 | 0.74 |
| Body Perception | 0.63 | 0.63 |
| Mean | 0.66 | 0.63 |

Tabel 2: Classification of Subjects in Time Categories.

| | Standard PCA | | | | D-Based PCA | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | 47 | 6 | 2 | 0 | 48 | 5 | 2 | 0 |
| 2 | 13 | 15 | 10 | 15 | 13 | 16 | 5 | 19 |
| 3 | 9 | 15 | 5 | 25 | 6 | 16 | 8 | 24 |
| 4 | 6 | 10 | 4 | 35 | 5 | 12 | 2 | 36 |
| % Correct | .85 | .28 | .09 | .64 | .87 | .30 | .15 | .65 |

Table 3: Classification of Subjects in Diagnosis Categories

| | Standard PCA | | | | D-Based PCA | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 4 | 2 | 3 | 1 | 4 | 2 | 3 |
| 1 | 82 | 2 | 11 | 2 | 92 | 0 | 3 | 2 |
| 4 | 20 | 0 | 5 | 3 | 22 | 1 | 2 | 3 |
| 2 | 3 | 0 | 15 | 18 | 1 | 0 | 19 | 16 |
| 3 | 2 | 3 | 20 | 31 | 2 | 0 | 25 | 29 |
| % Correct | .85 | .00 | .42 | .55 | .95 | .04 | .53 | .52 |
| Combined | 0.83 | | 0.91 | | 0.92 | | 0.97 | |

## 6. Monte Carlo Study

By definition, the distance-based biplot is outperformed by ordinary PCA with respect to Pearson correlations between observed and fitted variables. The differences in our empirical example are small, however, and distanced-based PCA performs better with respect to the recovery of the original classification of subjects into groups.

Table 4: Correlations between True Scores and Fitted Scores in Distance-based PCA compared to Standard PCA

| Variables | D-Based PCA | Standard PCA |
|---|---|---|
| 1 | .80 | .79 |
| 2 | .81 | .79 |
| 3 | .79 | .78 |
| 4 | .78 | .75 |
| 5 | .77 | .73 |
| 6 | .73 | .69 |
| 7 | .68 | .64 |
| 8 | .65 | .60 |
| 9 | .66 | .59 |
| 10 | .57 | .54 |
| 11 | .51 | .45 |
| dim 1 | .93 | .92 |
| dim 2 | .66 | .56 |
| dim 3 | .27 | .20 |

To inspect these properties in a more general context, (replicated) artificial data were generated, with 75 subjects and 13 variables with a perfect representation in three dimensions. From this set, two variables were selected as partitioning variables, and five categories were created using an optimal discretization strategy. The remaining 11 variables were subjected to a fair amount of random error (with an average of 53%). The resulting set of variables was analyzed by distance-based PCA, and compared to standard PCA. In both cases, analyses were done in two dimensions. The number of replications was set to 100. Four criteria were inspected:

- 1. The fit per variable, as measured by the Pearson correlation between the true scores (without error) and the bilinear approximation.
- 2. The fit per dimension, as measured by the Pearson correlation between the true dimensions and the fitted dimensions.
- 3. The correct classification of subjects in two-space as compared to the original classes. The a priori distribution in the population was taken into account to compute the posterior probabilities.
- 4. The distances between the subjects in two-space as compared to the distances in true-space.

The results reported below were obtained by averaging the results after applying distance-based and standard PCA to the 100 samples of the artificially created structure described above. The first part of Table 4 gives the correlations between true scores and fitted scores; the second part reports on the results with respect to the original three dimensions (that are approximated in two-space; the correlations here are again obtained by applying the rotation-expansion strategy from Section 3, but now to the original dimensions; so if the original dimensions are denoted by $\mathbf{Z}$, we minimize $||\mathbf{Z} - \mathbf{XB}'||^2$ over all $\mathbf{B}$ satisfying $\mathbf{B}'\mathbf{B} = \mathbf{I}$, and we compute the correlations between $\mathbf{Z}$ and $\mathbf{XB}'$). We notice that distance-based PCA performs better for each variable and each dimension separately. Results for the two classification variables were combined; the aggregated results are given in Table 5, where again rows indicate the original categories and columns the fitted categories. Except for the third category (75% versus 76% correct), distance-based PCA performs better; this effect is strongest for the extreme categories 1 and 5 (88% versus 78% and 87% versus 76% correct). Finally, the overall statistics are given in Table 6, confirming the superior performance of distance-based PCA over standard PCA.

Table 5: Classification of Subjects in the Monte Carlo Study

| | D-Based PCA | | | | | Standard PCA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| 1 | .88 | .11 | .01 | .00 | .00 | .78 | .20 | .02 | .00 | .00 |
| 2 | .08 | .81 | .11 | .01 | .00 | .05 | .74 | .19 | .02 | .00 |
| 3 | .01 | .12 | .76 | .10 | .01 | .00 | .13 | .75 | .12 | .00 |
| 4 | .00 | .01 | .10 | .80 | .09 | .00 | .02 | .18 | .76 | .05 |
| 5 | .00 | .00 | .01 | .13 | .87 | .00 | .00 | .02 | .22 | .76 |

Table 6: Summary of Results of the Monte Carlo Study: Overall Statistics

| | D-Based PCA | Standard PCA |
|---|---|---|
| Distances | 0.91 | 0.88 |
| Variables | 0.74 | 0.70 |
| Dimensions | 0.62 | 0.56 |
| Classification | 0.81 | 0.75 |

## 7. Discussion

A distance-based biplot was developed for a least squares MDS analysis of multivariate data. The fitted $p$-dimensional space is expanded into an $m$-dimensional space of rank $p$ that directly represents the fitted variables, while distances between subjects are preserved. The biplot method was applied to a data concerning patients with various types of eating disorders. The variables obtained a decent fit, also when compared to standard PCA. The method was further studied in a Monte Carlo study. Here results were compared with respect to the true structure, and the method proposed performed better than standard PCA with respect to the criteria considered. The analysis allows nominal variables to be included; their categories are represented by centroids of subjects. The method could include optimal scoring of ordinal variables too. By representing nominal variables as points in the subject space, and the other variables as vectors in the same space, we have actually developed a triplot, with subjects, variables, and classes as its constituents.

### References:


Carroll, J.D., and Chang, J.J. (1972): *IDIOSCAL (Individual differences in orientation scaling): A generalization of INDSCAL allowing IDIOsyncratic reference systems as well as an analytic approximation to INDSCAL*, Paper presented at the Psychometric Society Meeting, Princeton, NJ.

Carroll, J. D. (1972): Individual differences and multidimensional scaling, In: *Multidimensional scaling: Theory and applications in the behavioral sciences*, R. N. Shepard, A. K. Romney, and S. B. Nerlove (eds.), Vol. 1, 105-155, Seminar Press, New York and London.

De Leeuw, J., and Heiser, W.J. (1980): Multidimensional scaling with restrictions on the configuration, In: *Multivariate analysis*, P.R. Krishnaiah (ed.), , Vol. V, 501-522, North-Holland, Amsterdam.

Eckart, C. and Young, G. (1936): The approximation of one matrix by another of lower rank, *Psychometrika*, **1**, 211-218.

Gabriel, K.R. (1971): The biplot graphic display of matrices with application to principal components analysis, *Biometrika*, **58**, 453-467.

Gifi, A. (1990): *Nonlinear multivariate analysis*, John Wiley and Sons, Chichester.

Gower, J.C. (1966): Some distance properties of latent roots and vector methods used in multivariate analysis, *Biometrika*, **53**, 325-338.

Gower, J.C. and Hand, D.J. (1996): *Biplots*, Chapman & Hall, London.

Gower , J.C., and Harding, S.A. (1988): Nonlinear biplots, *Biometrika*, **75**, 445-455.

Groenen, P.J.F. and Heiser, W.J. (1996): The tunneling method for global optimization, *Psychometrika*, **61**, (in press).

Groenen, P.J.F. and Meulman, J.J. (1995): *Joint nonlinear biplots through least squares MDS*, Paper presented at the 9th European Meeting of the Psychometric Society, Leiden.

Guttman, L. (1968): A general nonmetric technique for finding the smallest coordinate space for a configuration of points, *Psychometrika*, **33**, 469- 506.

Heiser, W.J. (1981): *Unfolding analysis of proximity data*, Dept. of Data Theory, Leiden.

Heiser, W.J. (1987): Joint ordination of species and sites: the unfolding technique, In: *Developments in numerical ecology*, P. Legendre and L. Legendre (eds.), 189-221, Springer, New York.

Kruskal, J.B. (1964): Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, **29**, 1-28.

Kruskal, J.B. (1978): Factor analysis and principal components analysis: bilinear methods. In: *International encyclopedia of statistics*, W.H. Kruskal and J.M. Tanur (eds.), 307-330, The Free Press, New York.

Meulman, J.J. (1986): *A distance approach to nonlinear multivariate analysis*, DSWO Press, Leiden.

Meulman, J.J. (1992): The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables, *Psychometrika*, **57**, 539-565.

Meulman, J.J., and Heiser, W.J. (1993): Nonlinear biplots for nonlinear mappings. In: *Information and Classification: Concepts, Methods and Applications*, O. Opitz, B. Lausen, and R. Klar (Eds), 201-213, Springer Verlag, Berlin.

Meulman, J.J., Heiser, W.J., and Carroll, J.D. (1987): *PREPMAP-3 User's Guide*, Bell Telephone Laboratories, Murray Hill, NJ.

Ramsay, J.O. (1982): Some statistical approaches to multidimensional scaling data, *Journal of the Royal Statistical Society, Series A*, **145**, 285-312.

Shepard, R.N. (1962): The analysis of proximities: Multidimensional scaling with an unknown distance function I and II, *Psychometrika*, **27**, 125-140, 219-246.

Takane, Y., Young, F.W., and De Leeuw, J. (1977): Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features, *Psychometrika*, **42**, 7-67.

Torgerson, W.S. (1958): *Theory and methods of scaling*, Wiley, New York.

Tucker, L. R (1960): Intra-individual and inter-individual multidimensionality. In: *Psychological Scaling: Theory and Applications*, H. Gulliksen and S. Messick (eds.), Wiley, New York.

Van der Ham, Th., Meulman, J.J., Van Strien, D.C, and Van Engeland, H. (1997): *Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective, Britisch Journal of Psychiatry*, in press.

Young, G. and Householder, A.S. (1938): Discussion of a set of points in terms of their mutual distances, *Psychometrika*, **3**, 19-22.