RR 89-05

NONLINEAR PRINCIPAL COORDINATES ANALYSIS: MINIMIZING THE SUM OF SQUARES OF THE SMALLEST EIGENVALUES OF A CORRELATION MATRIX

Jacqueline J. Meulman

Department of Data Theory

University of Leiden

NONLINEAR PRINCIPAL COORDINATES ANALYSIS: MINIMIZING THE SUM OF SQUARES OF THE SMALLEST EIGENVALUES OF A CORRELATION MATRIX

Abstract

Classical principal coordinates analysis (PCO) is a technique proposed by Gower (1966) to analyze a multivariate data matrix by approximating the squared distances between the rows (the objects) by squared distances between object points in a low-dimensional space. In this paper the technique is generalized for Euclidean distances to a nonlinear variety, i.e., nonlinear transformations of the variables (the columns of the data matrix) are incorporated that are optimal according to the criterion that is minimized. The criterion can be expressed in terms of the eigenvalues of the correlation matrix for the variables: nonlinear principal coordinates analysis minimizes the sum of squares of a pre-chosen number of small eigenvalues. The optimal nonlinear transfor-mations are obtained by using majorization.

Key words: principal coordinates analysis, principal components analysis, multi-dimensional scăling, nonlinear transformations, majorization, eigenvalues.

Introduction

Principal coordinates analysis (PCO) is a technique proposed by Gower (1966) to analyze a set of multivariate data in terms of the distances between the objects, the rows of the data matrix. The purpose of the technique is to represent the objects in a low-dimensional representation space, in which the distances between the objects should resemble the distances derived from the scores of the objects on the variables as closely as possible. Although any association measure can be considered, the use of ordinary Euclidean distances between rows of the data is crucial in the present paper. PCO will be generalized in the sense that the variables in the data matrix may be nonlinearly transformed to minimize the discrepancy between the distances in the representation space and the distances derived from the variables.

Prior to the discussion of the minimization problem that has to be solved for nonlinear PCO, a formal description of classical PCO will be given. In the classical case, where the m variables are given, a principal coordinates analysis, when it is based on the Euclidean distances, is equivalent to a principal components analysis (PCA) with respect to the scores obtained for the objects. It will be shown that in contrast to this equivalency, PCO and PCA minimize a different criterion, which results in two different minimization problems when nonlinear transformations are incorporated. Nonlinear principal components analysis has been studied quite extensively in the literature (Kruskal and Shepard, 1974; Young, Takane and De Leeuw, 1978; Gifi, 1981, 1990; Winsberg and Ramsay, 1983); nonlinear principal coordinates analysis has apparently not been pursued before.

We assume data are available for n objects or individuals and m variables. The columns of the nxm data matrix $\mathbf{Z}$ are defined by nx1 vectors $\mathbf{z}_j$, j=1,...,m, that contain observations on the variables and are assumed to have a mean of zero and a sum of squares equal to 1. The measurements on the objects for the m variables define the rows in $\mathbf{Z}$, and give the coordinates for each object in an m-dimensional *observation* space; the rows of $\mathbf{Z}$

will be denoted by $\mathbf{z}_1', \ldots, \mathbf{z}_i', \ldots, \mathbf{z}_k', \ldots, \mathbf{z}_n'$. More generally, if the columns representing the m variables in $\mathbf{Z}$ are transformed, any set of transformed variables will be denoted by $\mathbf{Q}$, while $\mathbf{Q}^*$ 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 nxp matrix $\mathbf{X}$. The rows of $\mathbf{X}$ will be denoted by $\mathbf{x}_1', \ldots, \mathbf{x}_i', \ldots, \mathbf{x}_k', \ldots, \mathbf{x}_n'$.

Given this notation, a squared distance between a pair of objects $\{i,k\}$ in the m-dimensional observation space $\mathbf{Z}$ is defined by

$$d_{ik}^2(\mathbf{Z}) = (\mathbf{z}_i - \mathbf{z}_k)'(\mathbf{z}_i - \mathbf{z}_k) = (\mathbf{e}_i - \mathbf{e}_k)'\mathbf{Z}\mathbf{Z}'(\mathbf{e}_i - \mathbf{e}_k),$$

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

$$D^2(\mathbf{Z}) = \mathbf{a}\mathbf{1}' + \mathbf{1}\mathbf{a}' - 2\mathbf{Z}\mathbf{Z}', \tag{1}$$

with $\mathbf{1}$ an nx1 vector of all 1's and $\mathbf{a}$ an nx1 vector containing the diagonal elements of $\mathbf{Z}\mathbf{Z}'$; this will be denoted as $\mathbf{a}$ = vecdiag($\mathbf{Z}\mathbf{Z}'$). For the same pair of objects $\{i,k\}$, low-dimensional distances can be defined. The distance in the representation space $\mathbf{X}$ is written as

$$d_{ik}^2(\mathbf{X}) = (\mathbf{x}_i - \mathbf{x}_k)'(\mathbf{x}_i - \mathbf{x}_k) = (\mathbf{e}_i - \mathbf{e}_k)'\mathbf{X}\mathbf{X}'(\mathbf{e}_i - \mathbf{e}_k),$$

and in matrix notation,

$$D^2(\mathbf{X}) = \mathbf{a}\mathbf{1}' + \mathbf{1}\mathbf{a}' - 2\mathbf{X}\mathbf{X}',$$

with $\mathbf{a}$ = vecdiag($\mathbf{X}\mathbf{X}'$).

Classical principal coordinates analysis

Given these preliminaries, classical principal coordinates analysis can be described as follows, where the ordinary Euclidean distances between rows of the data, as given in (1), will be used as target distances. PCO is closely related to classical multidimensional scaling (Torgerson, 1952), since both techniques are based on the approximation of a matrix of scalar products by another one that is of lower rank. A basic ingredient is the Young-Householder (1938) process that transforms a squared distance matrix $D^2(\mathbf{Z})$ into an nxn scalar product matrix $\mathbf{ZZ}'$, locating the origin in the centroid of points by using the nxn centering operator $\mathbf{J} = \mathbf{I} - (\mathbf{11}'/\mathbf{1}'\mathbf{1})$:

$$-1/2\mathbf{J}D^2(\mathbf{Z})\mathbf{J} = -1/2\mathbf{J}(\mathbf{a1}' + \mathbf{1a}' - 2\mathbf{ZZ}')\mathbf{J} = \mathbf{ZZ}', \tag{2}$$

with $\mathbf{a}$ = vecdiag($\mathbf{ZZ}'$). (It should be noted that the last equality in (2) is true only if the variables $\mathbf{z}_j$ have zero mean.) The scalar product matrix $\mathbf{ZZ}'$ must be approximated by another scalar product matrix of lower rank, which objective can be written in the form of the so-called STRAIN loss function

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

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

$$\|\mathbf{XX}' - \mathbf{QQ}'\|^2 = \text{tr}\ (\mathbf{XX}' - \mathbf{QQ}')'(\mathbf{XX}' - \mathbf{QQ}').$$

(The term STRAIN for a least squares loss function defined on the scalar products was coined by Carroll and Chang, 1972.) In PCO, first an eigenanalysis of $\mathbf{ZZ}'$ is performed:

$$\mathbf{ZZ}' = \mathbf{K}\Lambda\mathbf{K}',$$

where $\mathbf{K}$ is an nxt matrix containing the eigenvectors as columns, Λ is a txt diagonal matrix containing the ordered positive eigenvalues, $\lambda_1 \geq \ldots \geq \lambda_t$, and t denotes the rank of

$\mathbf{Z}$ (for $t \leq m$). The optimal solution in the PCO procedure for obtaining a p-dimensional $\mathbf{X}$ (for $p \leq t$), is given by $\mathbf{X} = \mathbf{K}_p\Lambda_p^{1/2}$, where the subscript p indicates the use of the first p columns in $\mathbf{K}$ (and the first p rows and columns of Λ). The minimum of the STRAIN function (3) is found as

$$\text{STRAIN}(*) = \text{tr } \mathbf{K}\Lambda^2\mathbf{K}' - \text{tr } \mathbf{K}_p\Lambda_p^2\mathbf{K}_p' = \sum_{s=p+1}^{t} \lambda_s^2(\mathbf{ZZ}') \tag{4}$$

Equation (4) shows that the minimum loss is a function of the sum of squares of the $t-p$ smallest eigenvalues of $\mathbf{ZZ}'$, which are equal to the $t-p$ smallest eigenvalues of the mxm correlation matrix $\text{R}(\mathbf{Z}) = \mathbf{Z}'\mathbf{Z}$.

Although the latter observation already points to a relation between PCO and PCA on the basis of the analysis of the correlation matrix $\text{R}(\mathbf{Z})$, this relation is more clearly established by considering the bilinear model (Kruskal, 1978), minimizing the loss function

$$\text{STRIFE}(\mathbf{X};\mathbf{A}) = \|\mathbf{XA}' - \mathbf{Z}\|^2, \tag{5}$$

by performing a singular value decomposition of the matrix $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{V}\Psi\mathbf{W}'. \tag{6}$$

A representation for the objects is obtained by choosing $\mathbf{X}$ as $\mathbf{V}_p\Psi_p$; when we wish to represent the variables at the same time, $\mathbf{A}$ must be chosen from the representation in (6) as $\mathbf{A} = \mathbf{W}_p$. Since $\mathbf{Z}$ has t left singular vectors in $\mathbf{V}_t$ that are equal to the first t eigenvectors $\mathbf{K}_t$ of the matrix $\mathbf{ZZ}'$, and $\mathbf{Z}$ has t singular values in Ψ_t that are equal to the square root of the first t eigenvalues Λ_t of $\mathbf{ZZ}'$, i.e. $\Psi_t = \Lambda_t^{1/2}$, the optimal solution in PCO $\mathbf{X} = \mathbf{K}_p\Lambda_p^{1/2}$ is identical to the optimal solution $\mathbf{X} = \mathbf{V}_p\Psi_p$ in PCA.

In Meulman (1986), however, it was observed that the minimum of (5) is not equal to the minimum of (3): the minimum of (5), using the singular value decomposition (6), is obtained as

$$\text{STRIFE}(*;*) = \text{tr}\ \mathbf{W}\Psi^2\mathbf{W}' - \text{tr}\ \mathbf{W}_p\Psi_p^2\mathbf{W}_p'$$

$$= \text{tr}\ \mathbf{K}\Lambda\mathbf{K}' - \text{tr}\ \mathbf{K}_p\Lambda_p\mathbf{K}_p' = \sum_{s=p+1}^{t} \lambda_s(\mathbf{ZZ}'),$$

so STRIFE is a function of the sum of the $t-p$ smallest eigenvalues of the scalar product matrix $\mathbf{ZZ}'$, which is equal to the sum of the $t-p$ smallest eigenvalues of the correlation matrix R($\mathbf{Z}$). The difference between these two different criteria, the sum and the sum of squares of the smallest eigenvalues, results in a different minimization process when PCA and PCO are generalized to incorporate nonlinear transformations of the variables.

Nonlinear transformations in principal coordinates analysis

The incorporation of nonlinear transformations in PCA has been discussed by various authors, including Kruskal and Shepard (1974), Young, Takane and De Leeuw (1978), Gifi (1981; 1990), and Winsberg and Ramsay (1983). The optimality of nonlinear PCA can be described as follows: if $\mathbf{Q}^*$ denotes the set of optimally transformed variables, and R($\mathbf{Q}^*$) is the correlation matrix, then the sum of the $t-p$ smallest eigenvalues of R($\mathbf{Q}^*$) will be minimum.

Similarly, nonlinear PCO aims to find optimal nonlinear transformations of the variables, but since the analysis in the classical case showed that a different criterion is involved, a different minimization problem arises. In nonlinear PCO we have to minimize

$$\text{STRAIN}(\mathbf{X};\mathbf{Q}) = \|\mathbf{XX}' - \mathbf{QQ}'\|^2 \qquad (7)$$

both over $\mathbf{X}$, and over $\mathbf{Q}$ with columns $\mathbf{q}$ satisfying the normalization constraints $\mathbf{q}'\mathbf{q} = 1$, and the transformation constraints $\mathbf{q} \in \Gamma$, where Γ denotes the set of admissible transformations: it may be the set of monotonic transformations of a given variable $\mathbf{z}$, or the set of spline transformations of a certain degree, and so on. For fixed $\mathbf{X}$, (7) can be

solved for each column of $\mathbf{Q}$ separately. First the scalar product matrix $\mathbf{QQ}'$ is partitioned into $\mathbf{QQ}' = \mathbf{Q}^{-}\mathbf{Q}^{-}{}' + \mathbf{qq}'$, where $\mathbf{q}$ is an arbitrary column in $\mathbf{Q}$ and $\mathbf{Q}^{-}$ indicates that column $\mathbf{q}$ has been omitted from $\mathbf{Q}$. This partitioning enables us to rewrite (7) in terms of a single $\mathbf{q}$ as

$$\text{STRAIN}(\mathbf{q}) = \|\mathbf{XX}' - \mathbf{Q}^{-}\mathbf{Q}^{-}{}' - \mathbf{qq}'\|^2, \tag{8}$$

and when $\mathbf{U}$ is defined as $\mathbf{U} = \mathbf{XX}' - \mathbf{Q}^{-}\mathbf{Q}^{-}{}'$, then

$$\text{STRAIN}(\mathbf{q}) = \|\mathbf{U} - \mathbf{qq}'\|^2 = \text{tr}\ (\mathbf{U}^2 + \mathbf{qq}'\mathbf{qq}' - 2\mathbf{q}'\mathbf{Uq}), \tag{9}$$

which is to be minimized over $\mathbf{q}$ satisfying both the normalization and transformation constraints. Inserting the constraint $\mathbf{q}'\mathbf{q} = 1$ in (9) gives

$$\text{STRAIN}(\mathbf{q}) = 1 + \text{tr}\ \mathbf{U}^2 - 2\mathbf{q}'\mathbf{Uq} = c - 2\mathbf{q}'\mathbf{Uq}, \tag{10}$$

where c contains all elements of (10) that are independent of $\mathbf{q}$. Equation (10) shows that we have to maximize $\mathbf{q}'\mathbf{Uq}$ over $\mathbf{q}$ satisfying $\mathbf{q}'\mathbf{q} = 1$ and $\mathbf{q} \in \Gamma$.

To attain the minimum of (7) we have to alternate between updating $\mathbf{X}$ and $\mathbf{Q}$. Updates for $\mathbf{X}$ are easily found from the eigenvalue decomposition of $\mathbf{Q}'\mathbf{Q} = \mathbf{L}\Lambda\mathbf{L}'$, since the optimal $\mathbf{X} = \mathbf{K}_p\Lambda_p^{1/2} = \mathbf{QL}$. Updating each $\mathbf{q}$ in $\mathbf{Q}$ is more complicated. Note that if $\mathbf{q}$ would be unconstrained, (10) could be solved by using the eigenvalue decomposition of $\mathbf{U}$; for constrained $\mathbf{q}$, the problem can be solved by the use of majorization, a principle that has been used earlier in multidimensional scaling (De Leeuw and Heiser, 1980; De Leeuw, 1988), in correspondence analysis with least absolute residuals (Heiser, 1987), in linear dynamical systems analysis (De Leeuw and Bijleveld, 1988), and in the distance approach to nonlinear canonical analysis (Meulman, 1986).

Majorization

The basic idea of majorization is to replace a complicated (quadratic) function by a simple quadratic function. Observe that for any $\mathbf{q}^o$ it is true that $\mathbf{q} = \mathbf{q}^o + \mathbf{q} - \mathbf{q}^o$, so STRAIN($\mathbf{q}$) in (10) becomes

$$\begin{aligned} \text{STRAIN}(\mathbf{q}) &= c - 2[\mathbf{q}^o + (\mathbf{q} - \mathbf{q}^o)]'\mathbf{U}[\mathbf{q}^o + (\mathbf{q} - \mathbf{q}^o)] \qquad (11) \\ &= \text{STRAIN}(\mathbf{q}^o) - 2(\mathbf{q} - \mathbf{q}^o)'\mathbf{U}(\mathbf{q} - \mathbf{q}^o) - 4(\mathbf{q} - \mathbf{q}^o)'\mathbf{U}\mathbf{q}^o. \end{aligned}$$

At this point we wish to find a majorizing function $\mu(\mathbf{q},\mathbf{q}^o)$ for which it is true that STRAIN($\mathbf{q}$) $\leq \mu(\mathbf{q},\mathbf{q}^o)$ and STRAIN($\mathbf{q}$) = $\mu(\mathbf{q},\mathbf{q}^o)$ = STRAIN($\mathbf{q}^o$) if $\mathbf{q} = \mathbf{q}^o$. A property of the largest eigenvalue $\lambda_{\max(\mathbf{T})}$ of any matrix $\mathbf{T}$ can be used here, applied in majorization before by Heiser (1987) in the context of robust correspondence analysis. Because $\lambda_{\max(\mathbf{T})}$ is the maximum of the Rayleigh quotient:

$$\lambda_{\max(\mathbf{T})} = \max_{\mathbf{b}} \mathbf{b}'\mathbf{T}\mathbf{b}/\mathbf{b}'\mathbf{b},$$

where the maximum is taken over all vectors $\mathbf{b}$, it follows that for any $\mathbf{b}$ we may write $\mathbf{b}'\mathbf{T}\mathbf{b} \leq \lambda_{\max(\mathbf{T})}\mathbf{b}'\mathbf{b}$, and using this result in the present problem, we see that

$$\begin{aligned} -(\mathbf{q} - \mathbf{q}^o)'\mathbf{U}(\mathbf{q} - \mathbf{q}^o) &\leq \lambda_{\max(-\mathbf{U})}(\mathbf{q} - \mathbf{q}^o)'(\mathbf{q} - \mathbf{q}^o) \\ &\leq -\lambda_{\min(\mathbf{U})}(\mathbf{q} - \mathbf{q}^o)'(\mathbf{q} - \mathbf{q}^o), \end{aligned}$$

where $\lambda_{\min(\mathbf{U})}$ denotes the smallest eigenvalue of the matrix $\mathbf{U}$.

Using this result we find the majorizing function as follows. First we replace (11) by

$$\begin{aligned} \text{STRAIN}(\mathbf{q}) &\leq \text{STRAIN}(\mathbf{q}^o) - 2\lambda_{\min(\mathbf{U})}(\mathbf{q} - \mathbf{q}^o)'(\mathbf{q} - \mathbf{q}^o) - 4(\mathbf{q} - \mathbf{q}^o)'\mathbf{U}\mathbf{q}^o \qquad (12) \\ &\leq \text{STRAIN}(\mathbf{q}^o) - 2\lambda_{\min(\mathbf{U})}\left[(\mathbf{q} - \mathbf{q}^o)'(\mathbf{q} - \mathbf{q}^o) + 2/\lambda_{\min(\mathbf{U})}(\mathbf{q} - \mathbf{q}^o)'\mathbf{U}\mathbf{q}^o\right]. \end{aligned}$$

Rewriting the right-hand term of (12) to obtain a simple quadratic function that is dependent on $\mathbf{q}$, while the remainder is independent from $\mathbf{q}$ gives

$$\text{STRAIN}(\mathbf{q}) \leq \text{STRAIN}(\mathbf{q}^o) - 2\lambda_{\min(\mathbf{U})}\left[\|\mathbf{q} - \{\mathbf{q}^o - 1/\lambda_{\min(\mathbf{U})}\mathbf{U}\mathbf{q}^o\}\|^2 + \|1/\lambda_{\min(\mathbf{U})}\mathbf{U}\mathbf{q}^o\|^2\right].$$

Now the majorizing function can be written as

$$\mu(\mathbf{q},\mathbf{q}^o) = \text{STRAIN}(\mathbf{q}^o) - 2\lambda_{\min(\mathbf{U})}\left[\|\mathbf{q} - \{\mathbf{I}-1/\lambda_{\min(\mathbf{U})}\mathbf{U}\}\mathbf{q}^o)\|^2 + \|1/\lambda_{\min(\mathbf{U})}\mathbf{U}\mathbf{q}^o\|^2\right]. \quad (13)$$

Due to the special structure of $\mathbf{U} = \mathbf{X}\mathbf{X}' - \text{tr}\,\mathbf{Q}^{-}\mathbf{Q}^{-1}$, $\lambda_{\min(\mathbf{U})} \leq 0$. So (13) is minimized by minimizing $\|\mathbf{q} - \{\mathbf{I} - 1/\lambda_{\min(\mathbf{U})}\mathbf{U}\}\mathbf{q}^o)\|^2$ over $\mathbf{q}'\mathbf{q} = 1$ and $\mathbf{q} \in \Gamma$, i.e., the problem is to find a normalized $\mathbf{q}$ that is a transformation of $\mathbf{z}$ and has at the same time minimum discrepancy to $\{\mathbf{I} - 1/\lambda_{\min(\mathbf{U})}\mathbf{U}\}\mathbf{q}^o$.

In the case of order restrictions $\mathbf{q}$ must be monotonic to the given data vector $\mathbf{z}$, and a monotonic regression of $\{\mathbf{I} - 1/\lambda_{\min(\mathbf{U})}\mathbf{U}\}\mathbf{q}^o$ upon $\mathbf{z}$ should be applied, and the result must be normalized; in the case of spline transformations we require that $\mathbf{q}$ is a linear combination of a set of basis splines $\mathbf{S}$, so $\|\mathbf{S}\mathbf{b} - \{\mathbf{I} - 1/\lambda_{\min(\mathbf{U})}\mathbf{U}\}\mathbf{q}^o\|^2$ should be minimized over $\mathbf{b}$, with $\mathbf{b}$ possibly constrained to have positive elements to ensure monotonic spline transformations.

Looking back at (10), one might wonder why a step of the much more simple power algorithm could not be used, i.e., applying the regression to $\mathbf{U}\mathbf{q}^o$. The power algorithm converges to the eigenvector associated with the eigenvalue of $\mathbf{U}$ that is *largest in absolute value*. The smallest eigenvalue of $\mathbf{U}$ may be larger in absolute value than the largest eigenvalue, and since we need a solution that is associated with the largest eigenvalue, regression of $\mathbf{U}\mathbf{q}^o$ upon $\mathbf{z}$ could give an update of the transformed variable that does not decrease the value of the STRAIN function. The properties of the majorizing function (13) guarantee that applying regression of $\{\mathbf{I} - 1/\lambda_{\min(\mathbf{U})}\mathbf{U}\}\mathbf{q}^o$ upon $\mathbf{z}$ does decrease the loss.

From a computational point of view, we do not wish to compute $\lambda_{\min(\mathbf{U})}$ itself. In stead, a lowerbound of $\lambda_{\min(\mathbf{U})}$ can be used; a particular convenient one is given by

Wolkowicz and Styan (1980), which depends only on the trace and the sum of squares of **U**:

$$\lambda_{\min(\mathbf{U})} \geq n^{-1}\text{tr } \mathbf{U} - (n-1)^{1/2}[n^{-1}\text{tr } \mathbf{U}^2 - (n^{-1}\text{tr } \mathbf{U})^2]^{1/2}.$$

The computational scheme alternates between updating **X** and **Q**. To minimize (8) for each **q** in **Q** in the iteration step that updates **Q**, inner iterations should be performed. However, it is not necessary to attain the minimum of (8); for convergence of the overall algorithm it is sufficient to update each **q** only once, decreasing the loss of (7) in each step. When the overall algorithm has converged, the final updates **q*** give the minimum value of (8) over all **q**.

Comparing the optimal solutions in nonlinear PCA and PCO.

To investigate the difference between nonlinear PCO and nonlinear PCA, various aspects of the solution can be considered. In the first place, we may look at the distribution of the eigenvalues of the correlation matrix between the transformed variables. In the second place, the transformations themselves can be compared. And finally, of course, the representation of the objects, in both cases found by performing an eigenanalysis, but based on a possibly different transformed data matrix, may differ.

In terms of the distributions of the eigenvalues, we may deduct from the criterion that is minimized that nonlinear PCO will minimize the difference between the $t-p$ smallest eigenvalues, since the sum of squares of t quantities is proportional to their sum of squared differences. So it is expected that a clear distinction in size will be found between the fist p eigenvalues and the remaining ones; the latter is usually not found in a PCA, where the decrease in size of the subsequent eigenvalues may be very gradual.

To see whether this phenomenon occurs, various gauges have been used; the term gauge is used to denote a data matrix with a special structure that is known beforehand: by applying a technique to a gauge it can be inspected whether specific properties of the data emerge in the solution. For the comparison of nonlinear PCA and PCO three gauges were studied; the variables were generated by sampling from a multinormal distribution, with a specific correlation structure establishing the true dimensionality of the gauge. By adding random error, with an error level of 10%, data matrices of full rank are obtained, with the original "true" dimensionality displayed in the distribution of the eigenvalues, i.e., there is a particular number of dominant eigenvalues.

The first two gauges consist of 50 objects and 7 variables; the first gauge is a two-dimensional structure, and the second a five-dimensional one. The third gauge consists of 50 objects and 14 variables, its true dimensionality being 7.

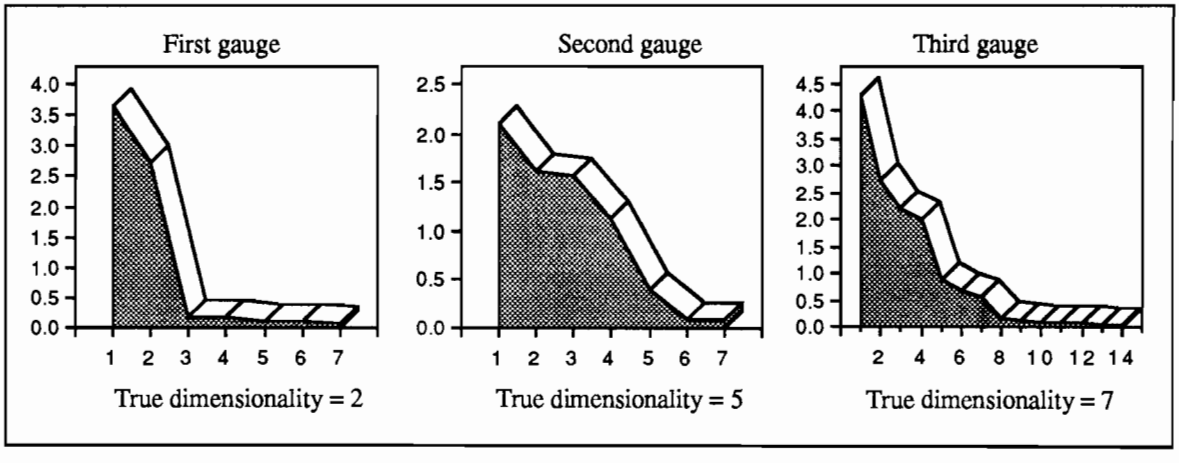

Figure 1. Distribution of the eigenvalues for the three gauges

The distribution of the eigenvalues of the correlation matrices for the three different gauges is displayed in Figure 1. The first gauge, for instance, has two dominant eigenvalues; the remaining five eigenvalues are smaller than .5. The analyses, the results of which will be given in the next section, were chosen in such of way that the true dimensionality of the gauges was deliberately neglected: the two-dimensional gauge was analyzed in one dimension, and the second and third gauge in two dimensions. Because we demand the techniques to represent a structure in a space of too few dimensions, nonlinear transformations are most likely to occur.

Results

A display of the distribution of the eigenvalues and the squared eigenvalues is given in Figure 2. Plots of the three different gauges are given next to each other; the six plots at the top give the distribution of the eigenvalues for PCA and PCO, while the six plots at the bottom give the distribution of the squared eigenvalues. The dark areas denote the sum of the residual (squared) eigenvalues, while the shaded areas denote the sum of the (squared) largest eigenvalues. By definition, the shaded areas for PCA are smaller than those for PCO in the six upper plots, while they are smaller for PCO compared to PCA in the six plots at the bottom.

The behavior of PCO is quite consistent across the three gauges: the size of the eigenvalues drops significantly after the first dimension for the first gauge, and after the second dimension of the second and third gauge; the decrease in size of the eigenvalues of PCA is much more gradual. In PCO the nonlinear transformation maximizes the difference between the largest and the remaining eigenvalues; in PCA merely the sum of the largest eigenvalues is maximized.

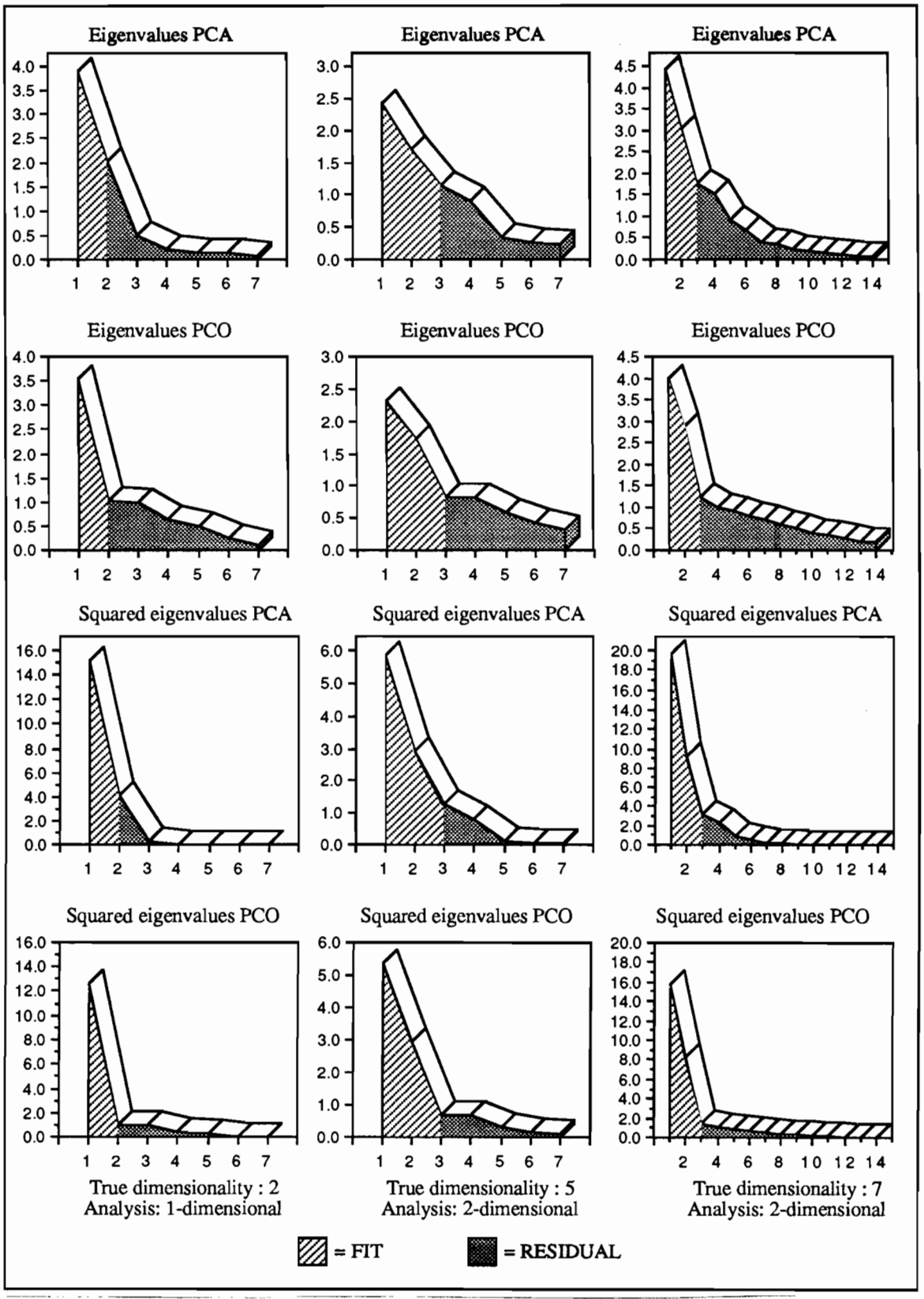

Figure 2. Distribution of eigenvalues and squared eigenvalues after nonlinear transformation in PCA and PCO

The badness-of-fit measures for PCA and PCO are found in Table 1; both the values of the function STRIFE($\mathbf{X};\mathbf{A};\mathbf{Q}$) = $\|\mathbf{XA}'-\mathbf{Q}\|^2$, and the function STRAIN($\mathbf{X};\mathbf{Q}$) = $\|\mathbf{XX}'-\mathbf{QQ}'\|^2$ are given.

TABLE 1

Badness-of-fit measures STRIFE and STRAIN
obtained for PCA and PCO

	Gauge 1		Gauge 2		Gauge 3	
Variable	PCA	PCO	PCA	PCO	PCA	PCO
STRIFE	3.10	3.44	2.88	2.96	6.52	7.16
STRAIN	4.44	2.69	2.34	1.96	7.15	5.52

Obviously, PCO performs better when STRAIN is considered, while PCA performs better according to STRIFE; the differences between PCA and PCO are smaller considering STRIFE than looking at STRAIN. Note that the value for STRIFE is not necessarily smaller than the value for STRAIN for PCA (gauge 2). This might happen when the majority of the residual eigenvalues is smaller than 1.

By looking at the correlations between the transformed variables obtained by PCA on the one hand, and by PCO on the other, it turns out that most variables are transformed very much alike, while only a few are strikingly different (Table 2). The correlations that are underlined indicate for each gauge that particular variable that was transformed most differently by the two techniques; the transformation plots for these three variables are given in Figure 3. The three plots show that the transformations in PCO are more strongly nonlinear; this observation is most pertinent to gauge 1 and 3. To see whether this conclusion holds in general, the correlations between the original variables and the transformed variables can be inspected (Table 3).

TABLE 2

Correlations between transformations obtained by PCA and PCO

Variable	Gauge 1	Gauge 2	Gauge 3	Variable	Gauge 3
1	.93	.85	.97	8	.81
2	.70	.93	.95	9	.84
3	.97	.97	.69	10	.95
4	.96	1.00	.94	11	.86
5	.99	.99	.94	12	.92
6	.08	.98	.93	13	.94
7	.98	.97	.90	14	.98

TABLE 3

Correlations between original variables and transformations obtained by PCA and PCO

	Gauge 1		Gauge 2		Gauge 3			Gauge 3	
Variable	PCA	PCO	PCA	PCO	PCA	PCO	Variable	PCA	PCO
1	.97	.86	.56	.07	.99	.96	8	.96	.74
2	.94	.43	.81	.53	.88	.76	9	.93	.60
3	.48	.36	.99	1.00	.88	.29	10	.56	.44
4	.93	.86	.90	.89	.94	.90	11	.87	.55
5	.96	.09	1.00	.99	.98	.90	12	.94	.74
6	.91	−.12	.81	.72	.95	.94	13	.91	.72
7	.88	.96	.95	.85	.97	.79	14	.97	1.00

TABLE 4

Correlations between all dimensions obtained by PCA and PCO;
Residual dimensions have been underlined

Dimension	Gauge 1	Gauge 2	Gauge 3	Dimension	Gauge 3
1	.95	.97	.95	8	.29
2	.58	.97	.94	9	.22
3	.33	.85	.83	10	.53
4	.44	.88	.71	11	.09
5	.24	.55	.68	12	.06
6	.62	.41	.45	13	.76
7	.76	.41	.66	14	.62

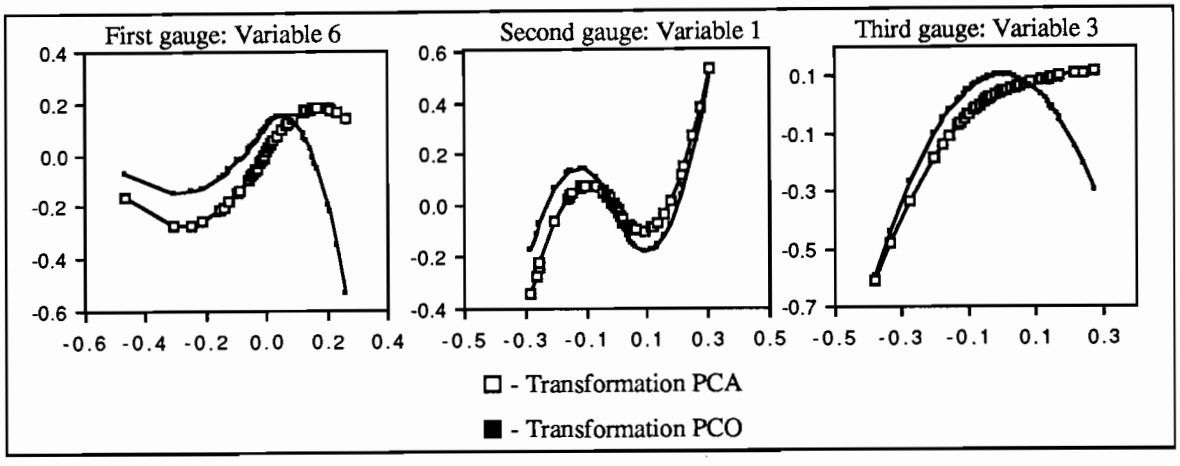

Figure 3. Transformation plots for three selected variables

Although some exceptions are noticed, the majority of the PCO variables has (much) smaller correlations with the original variables than the PCA variables have, so we conclude that the transformations must be more strongly nonlinear.

With respect to the configurations of object points, we expect that the configurations that are optimized by the two techniques will be much alike because most of the transformations are quite similar. The residual dimensions have been computed on the basis of the singular value decomposition of the transformed data matrix. The correlations between all PCA and PCO dimensions are given in Table 4; the dimensions associated with the residual eigenvalues have been underlined. It turns out that differences between the two techniques emerges in the residual dimensions; the dimensions that go with the largest eigenvalues have large correlations.

Discussion

A convergent algorithm was developed for nonlinear principal coordinates analysis based on Euclidean distances, which is an alternative for nonlinear principal components analysis. The rationale for the particular optimal transformations in PCO is that the approximation error, when a high dimensional space **Q** is approximated by a low-dimensional space **X**, is distributed as evenly as possible over the residual dimensions.

The results of the new technique have been compared to nonlinear PCA: the residual dimensions differ considerably, and some, but not all, transformations. However, although the loss functions are quite different, nonlinear PCA and PCO do not give strikingly different results for the optimal configurations in the examples studied in this paper. (A similar situation was encountered by Heiser (1988) in a completely different context. He developed a multidimensional scaling algorithm based on the L_1-norm, and compared the results with least squares multidimensional scaling.) Whether nonlinear PCO will behave more differently than nonlinear PCA for other gauges remains to be investigated.

Nonlinear PCO was originally developed in Gower (1966) to analyze *any* distance measure between objects in a multivariate data matrix. The algorithm developed in this paper is only optimal for Euclidean distances; different algorithms should be developed for different distance functions associated with different techniques in multivariate analysis (cf. the least squares distance approximation framework developed in Meulman (1986)).

References

Carroll, J.D., & 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.

De Leeuw, J. (1988). Convergence of the majorization algorithm for multidimensional scaling. *Journal of Classification, 5,* 163-180.

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

De Leeuw, J., & Bijleveld, C. (1988). *Fitting longitudinal reduced rank regresison models by alternating least squares.* Research Report RR-88-03. Leiden: Dept. of Data Theory.

Gifi, A. (1990). *Nonlinear multivariate analysis.* New York: Wiley.

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

Heiser, W.J. (1987). Correspondence analysis with least absolute residuals. *Computational Statistics and Data Analysis, 5,* 337-356.

Heiser, W.J. (1988). Multidimensional scaling with least absolute residuals. In H.H. Bock (Ed.), *Classification and Related Methods of Data Analysis,* (pp. 455-462). Amsterdam: North-Holland.

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

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

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

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

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

Wolkowicz, H., & Styan, G.P.H. (1980). Bounds for eigenvalues using traces. *Linear Algebra and its Applications, 29,* 471-506.

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

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