ALTERNATIVE APPROXIMATIONS IN PRINCIPAL COMPONENTS ANALYSIS

Jan de Leeuw and Jacqueline Meulman
Department of Data Theory FSW/RUL
University of Leiden, The Netherlands

In this paper we present several data analysis techniques that combine features of principal component analysis (PCA) and multidimensional scaling (MDS). We concentrate on the type of approximation defined by the loss function, and on admissible optimal transformations. We use a medium-sized example to illustrate in detail the effect of various choices on low-dimensional solutions.

Keywords: Principal Component Analysis, Multidimensional Scaling, Nonlinear Multivariate Analysis, Optimal Scaling, Alternating Least Squares.

INTRODUCTION

The fact that principal component analysis and multidimensional scaling are closely related, and can both be represented in the same distance-geometrical framework, was already emphasized by Gower (1966, 1967). We shall use his approach to PCA as our starting point. A closely related formulation is given by Benzécri et al. (1973, T-II-B no 2, T-II-B no 3), and by Cailliez and Pagès (1976, especially chapters 6 and 7). Relationships between the two classes have also been emphasized by Heiser and Meulman (1983) and Meulman and Heiser (1984). It seems to us that an even more uniform presentation of PCA and MDS is possible, and we shall see that this uniform presentation unavoidably suggests a new technique in the intersection of the two classes.

CLASSICAL PCA

Suppose $z_1,\ldots,z_m$ are given elements of R^m. They can be n observations on m variables, or n time series of length m, or n discrete probability distributions on m outcomes, or n rankings of m objects, or whatever. We start our analysis with a quadratic metric on R^m, i.e. with a positive definite matrix A which defines a distance by

$$\delta_A^2(z_i,z_k) = (z_i - z_k)'A(z_i - z_k). \tag{1}$$

We sometimes write $\delta_{ik}(Z)$ or simply δ_{ik} for $\delta_A(z_i,z_k)$, if no confusion is possible. We also use (1) in the equivalent form

$$\delta_A^2(z_i,z_k) = (e_i - e_k)'ZAZ'(e_i - e_k), \tag{2}$$

where Z is the n x m matrix containing the z_i (as rows), and where e_i and e_k are unit vectors (i.e. columns i and k of the identity matrix).

In PCA we want to find points $x_1,\ldots,x_n$ in R^p, with $p \leq m$, such that the ordinary Euclidean distance between x_i and x_k is approximately equal to $\delta_A(z_i,z_k)$. Thus $d(x_i,x_k)$ or $d_{ik}(X)$ is defined by

$$d_{ik}^2(X) = (x_i - x_k)'(x_i - x_k) = (e_i - e_k)'XX'(e_i - e_k), \tag{3}$$

and ideally we want X to satisfy

$$d_{ik}(X) = \delta_A(z_i,z_k). \tag{4}$$

Exact equality in all pairs (i,k) in (4) will not be possible in general for small p, it will only be possible if $p \geq$ rank (Z). Thus for small p we have to specify what type of approximation we have in mind, and how we measure quality of approximation.

The method of approximation chosen by PCA is called quadratic approximation from below. In order to explain this concept we first observe that X is certainly not determined uniquely by conditions (4). If X satisfies (4), than any rotation XT also satisfies (4). We eliminate rotational indeterminacy first by requiring that X'BX is diagonal, where B is a known weight matrix (positive definite, of order n). Now define $C = B^{\frac{1}{2}}ZAZ'B^{\frac{1}{2}}$, and suppose $C = K\Omega^2K'$ is the eigen-decomposition of C. Thus K is square orthonormal, and Ω^2 is diagonal. The diagonal elements of Ω^2 are ordered by $\omega_{11}^2 \geq \omega_{22}^2 \geq \ldots \geq \omega_{nn}^2$. There are only rank(Z) eigenvalues which are nonzero. Let K(p) be the first p columns of K, and $\Omega(p)^2$ the corresponding submatrix of Ω^2. Define $X(p) = B^{-\frac{1}{2}}K(p)\Omega(p)$. Then $X(p)'BX(p) = \Omega(p)^2$, which is diagonal, and $X(p)X(p)' = B^{-\frac{1}{2}}K(p)\Omega(p)^2K(p)'B^{-\frac{1}{2}}$. We can also write this as $X(p)X(p)' = B^{-\frac{1}{2}}C(p)B^{-\frac{1}{2}}$, where C(p) is the best rank p approximation to C. The fact that X(p) defines a quadratic approximation from below is now expressed by the chain

$$d_{ik}^2(X(1)) \leq d_{ik}^2(X(2)) \leq \ldots \leq d_{ik}^2(X(\rho)) = \delta_{ik}^2(Z), \tag{5}$$

where ρ = rank(Z). The sequence of approximations is also nested, in the sense that the first s columns of X(t), with t > s, are X(s). A bit of care is required if there are multiple eigenvalues, but the complications are not at all essential. An obvious measure for the badness-of-fit is the sum of the n - p discarded eigenvalues.

DUALITY

In the previous section we have defined a metric on the rows of the matrix Z, and we have approximated the resulting weighted distances with ordinary Euclidan distances in a low dimensional space. But we can also proceed the other way around, and start with a metric on the columns of Z. Thus

$$\delta_B^2(z_j,z_\ell) = (e_j - e_\ell)'Z'BZ(e_j - e_\ell), \tag{6}$$

and we want to find $y_1,\ldots,y_m$ in R^p such that $\delta_B(z_j,z_\ell)$ is approximately equal to $d_{j\ell}(Y)$. To solve this by quadratic approximation from below we need the eigendecomposition $D = L\Omega^2L'$, where $D = A^{\frac{1}{2}}Z'BZA^{\frac{1}{2}}$, and Y is scaled by requiring that Y'AY is diagonal. We find $Y_p = A^{-\frac{1}{2}}L(p)\Omega(p)$ as our nested sequence of approximations.

The remarkable duality aspect of our two approximations is that the eigenvalues of both problems are the same. This is a consequence of a somewhat deeper connection between the two problems, formalized in the singular value decomposition. The singular value decomposition of our PCA can be written as

$$B^{\frac{1}{2}}ZA^{\frac{1}{2}} = K\Omega L', \tag{7}$$

or as

$$BZAY = BX\Omega, \tag{8a}$$

$$AZ'BX = AY\Omega, \tag{8b}$$

or as

$$Z = X\Omega^{-1}Y'. \tag{9}$$

Equivalence of these formulas follows from simple manipulations.

SPECIAL CASES

Because our approach to PCA is perhaps slightly unconventional, we illustrate it with some interesting special cases before we proceed. These special cases will not be illustrated with examples, because that would take us too far astray. They merely serve to make the formulas in the previous sections a bit more transparant.

The first special case is correspondence analysis (CFA). In this form of PCA we analyze a single cross table or bivariate distribution. Suppose the n x m matrix F is this table. Thus the entries of F are nonnegative, and they add up to one. Let E_1 be the diagonal matrix with row sums, and E_2 the diagonal matrix with column sums. Set $Z = E_1^{-1}FE_2^{-1}$, observe that all z_{ij} are equal to unity if the two variables are independent. The metric on R^m is chosen as $A = E_2$, and the metric

on R^n as $B = E_1$. Thus

$$\delta_A^2(z_i,z_k) = (e_i - e_k)'E_1^{-1}FE_2^{-1}F'E_1^{-1}(e_i - e_k). \tag{10}$$

The singular value problem for CFA is $E_1^{-\frac{1}{2}}FE_2^{-\frac{1}{2}} = K\Omega L'$. Equation (8) becomes

$$E_1^{-1}FY = X\Omega, \tag{11a}$$

$$E_2^{-1}F'X = Y\Omega. \tag{11b}$$

These equations are known as the transition formulas. They express the fact that the scores X and Y linearize the regression in the bivariate table. They can also be interpreted in terms of the centroid principle. This becomes more clear if we define $\bar{Y} = Y\Omega$, and rewrite (11b) as $E_2^{-1}F'X = \bar{Y}$. The column scores $\bar{Y}$ are in the centroids of the row scores X, i.e. they are conditional expectations. The centroid principle gives an obvious and natural way to make joint plots. Observe that we can also proceed the other way around, and make the row points centroids of the column points. We cannot do both at the same time, however. In CFA formula (9) is called the reconstitution formula. It is most easily understood if we write it in elementwise notation

$$f_{ij} = f_{i+}f_{+j}(1 + \sum_s \omega_s^{-1}x_{is}y_{js}). \tag{12}$$

In (12) we have used the fact that the largest singular value is always +1, and that the corresponding scores x and y also have all elements +1. Of course f_{i+} and f_{+j} are row- and column marginals of the table F. For more information about correspondence analysis we refer to Benzécri et al. (1973), Benzécri et al. (1980), Gifi (1981), and Greenacre (1984).

The second special case is homogeneity analysis, also known as multiple correspondence analysis (MCA). This can be presented as a special form of correspondence analysis applied to indicator matrices (also known as dummies). We start with n observations on m different categorical variables. Each categorical variable is coded in a binary matrix G, with n rows and k_j columns, where k_j is the number of categories (possible values) of variable j. Collect the G_j in an n x K matrix $G = (G_1 \ldots G_m)$, with K the sum of the k_j. Now apply correspondence analysis to G. We have $E_1 = mI$ and E_2 is the diagonal matrix of univariate marginals. A little reflection shows that

$$\delta_A^2(z_i,z_k) = m^{-2} \sum_{j=1}^{m} \varepsilon_{ikj}(n_{ij}^{-1} + n_{kj}^{-1}), \tag{13}$$

with $\varepsilon_{ikj} = 0$ if i and k are in the same category of variable j, and $\varepsilon_{ikj} = 1$ otherwise, and with n_{ij} the number of objects in the category of variable j in which we find object i. The centroid principle can be written for each variable separately, in the convenient form $\bar{Y}_j = E_j^{-1}G_j'X$, i.e. category quantifications are centroids of object scores of the objects in the categories. A whole system of multivariate analysis methods has been build, starting from MCA, by Gifi (1981), compare also De Leeuw (1984). We have seen that MCA is CFA on the indicator matrices. On the other hand CFA is also identical to MCA with two variables. Generalizations of MCA if there are missing data are discussed by Gifi (1981) and Meulman (1982).

An even simpler special case is standardized PCA. Here Z is an n x m data matrix, in deviations from the column mean. We set A equal to V^{-1}, with V the diagonal matrix of the variances, and we set B = I. Then $D = V^{-\frac{1}{2}}Z'ZV^{-\frac{1}{2}} = R$, the correlation matrix of the variables. Also $Y_p = V^{\frac{1}{2}}L(p)\Omega(p)$, where $R = L\Omega^2L'$. Of course if we start with Z equal to the standardized data matrix, and set A = I, then we find $Y_p = L(p)\Omega(p)$, which is different.

Standardized PCA and unweighted PCA on a standardized matrix can be thought of as techniques which transforms the columns of the data matrix (linearly) and then apply PCA. There are other techniques which transform the data and then apply PCA, and in which the transformations need not be linear. We mention some simple examples. Suppose we have an n x m matrix F, with nonnegative entries, whose rows add up to one. This could be called compositional data, they are quite common in geology (composition of rock and sediment specimens), in chemistry (composition of mixtures), in biology (composition of natural populations), and in consumer demand analysis (composition of household budgets). Many other examples are possible (Aitchison, 1982). Compositional data can be analyzed by CFA, using (10) as the distance between composites, where $E_1 = I$. Other distance measures have been proposed, however. Domenges and Volle (1980) first transform the data by taking square roots, and then use ordinary Euclidan distance. This is effectively equivalent to using the Hellinger distance between the rows, or the chordal metric of Edwards and Cavalli-Sforza (1967). Aitchison (1983) divides all elements in a row by their geometric mean, and then takes the logarithm of all entries of the resulting table. For details and interpretational consequences we refer to the cited papers; we have merely mentioned these methods to show that transformations are often combined with PCA.

The standardized PCA example can be extended in another direction. Instead of taking $A = V^{-1}$ we can also take $A = V_1^{-1} \dotplus \ldots \dotplus V_h^{-1}$, where h is the number of sets of variables and $\dotplus$ is the direct sum. Thus A has block structure, with zero off-diagonal blocks and with diagonal blocks equal to inverses of within-set dispersions. This defines generalized canonical correlation analysis. The idea of transforming the rows of a data matrix before PCA, which was used by Aitchison and Domenges & Volle, has also been applied very successfully to the analysis of curves and processes by Besse (1979). The idea there is to apply a filter or linear smoother to a row, which is a time series, and then apply PCA. Thus we see that there are a great many possibilities, all using basically the same technique. From the point of view of this paper the various special cases discussed in this paragraph all use quadratic approximation from below. If there are transformations of the data matrix they are applied before the PCA is performed.

AN APPLICATION: 50 STATES OF NORTH AMERICA

We illustrate the properties of quadratic approximation from below by analyzing the following example with classical PCA. The data consist of social indicator statistics taken from statistical abstracts of the U.S. (1977)[1]. They are given in Table 1. We first concentrate on the two-dimensional configuration for the 50 states.

As a measure to evaluate the badness-of-fit of this representation we propose the root mean square of the residuals:

$$\text{ROMRES} = \left[\frac{1}{n(n-1)} \sum_i \sum_k (\delta_{ik}(Z) - d_{ik}(X))^2 \right]^{\frac{1}{2}} \qquad (14)$$

The configuration that is depicted in Figure 1 has ROMRES equal to 1.055. The points for the states are labelled as indicated in Table 1. The fact that the configuration has a definite shape, -- the first dimension showing much more dispersion than the second - is the result of the clearly separated accompanying eigenvalues: the first one is 2.6 times as large as the second one.

Inspecting Figure 1 we find the southern states clearly separated in the lower right corner. Investigating the original data, we found the deep south to rank among the "unfortunate" half indicated by the variables 2 upto 6: low income, high illiteracy rate, low life expectancy, high homicide rate, low percentage of high school graduates. In the lower left corner we detect a mixture of states in the midwest, north-east and mountain states. Looking at the other direction in space shows California, New York, and to a smaller degree, Florida and Texas to be isolates. These 4 states rank among the 8 states with the largest population and among the 16 states with relatively few days in a year in which the temperature falls below freezing.

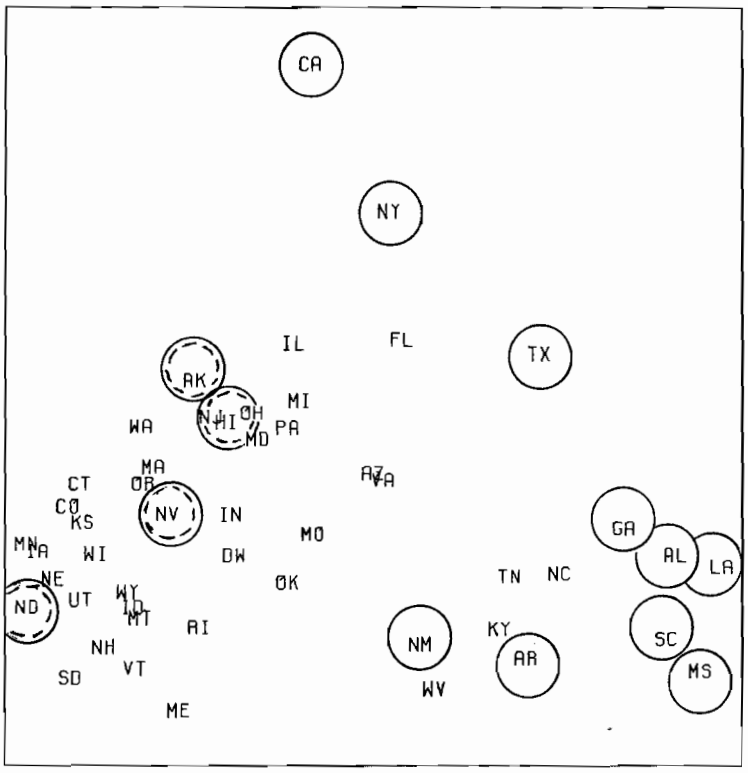


Figure 1. PCA solution for 50 states. Encircled points have dissimilarities larger than average. Dotted circles indicate more than average stress in addition.

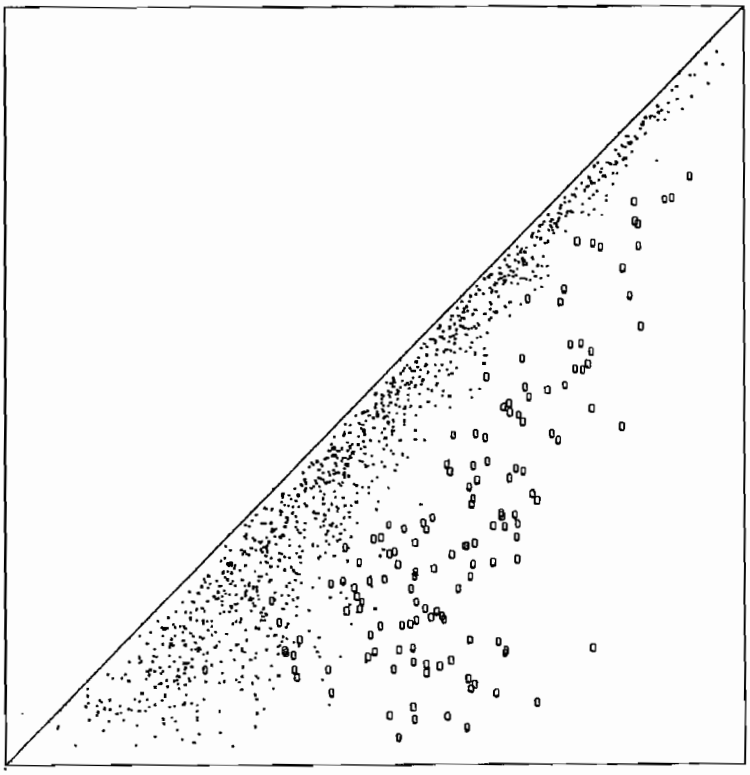


Figure 2. PCA solution for 50 states. $\delta(Z)$ (horizontal axis) versus $d(X)$ (vertical axis). Approximation from below. Ellipses refer to all pairs including AK, HI and NV.

To illustrate the quality of the representation some states have been encircled. The continuous circles are drawn around each point x_i to which applies

$$\sum_k \delta_{ik}(Z) > \frac{1}{n} \sum_i \sum_k \delta_{ik}(Z) \qquad (15)$$

so these are states with relatively large dissimilarities. It is clear that for most of the points concerned these large dissimilarities are approximated by large distances in the configuration. To support this observation we have drawn dotted circles around each of the points for which it is true, in addition to (15), that

$$\frac{1}{n-1} \sum_k (\delta_{ik}(Z) - d_{ik}(X))^2 > \frac{1}{n(n-1)} \sum_i \sum_k (\delta_{ik}(Z) - d_{ik}(X))^2 \qquad (16)$$

Thus most of the encircled points have a relatively small contribution to the sum of residuals. Alaska (AK), Hawaii (HI), Nevada (NV) and North Dakota (ND), on the contrary, do not fit into this pattern.
The figures 6.562, 5.946, 5.308 and 1.264 have been obtained respectively, i.e. for the left term of (16). So the deviation for AK, HI and NV is most serious: together they account for 32% of the total sum of squared residuals.

Approximation from below can be illustrated most clearly in the scatter plot of dissimilarities versus distances (Fig. 2). Althought the majority of dots, representing the pairs (δ_{ik}, d_{ik}), is to be found quite close to the diagonal, which symbolizes perfect fit, we detect numerous dots displaying a large approximation error. These are exactly dots portraying the approximation for AK, HI and NV. The latter are indicated by ellipses. It will be clear that we need a higher dimensional solution to approximate the dissimilarities for these three states closely. We will, however, not pursue this strategy and shall concentrate in the next section on a different approach to the scaling problem.

APPROXIMATION FROM BOTH SIDES

Since approximation from below has certain peculiarities, it is natural to look for other types of approximation of the $\delta_{ik}(Z)$ by the $d_{ik}(X)$. There are many possibilities, but we will be focussed on the explicit minimization of the loss function

$$\sigma(X) = \sum_i \sum_k (\delta_{ik}(Z) - d_{ik}(X))^2 \qquad (17)$$

over all $X \varepsilon R^{np}$. This loss function belongs to a very specific class; compare De Leeuw and Heiser (1982) for a review of the properties of these loss functions,

and for algorithms that can be used to minimize them. In this paper (17), which is called STRESS and was introduced by Kruskal (1964), will be minimized by the algorithm described by De Leeuw (1977).

This algorithm guarantees a convergent series of configurations and can be characterized as repeatedly computing the so-called Guttman transform

$$\bar{X} = \frac{1}{n} B(X)X \tag{18}$$

where the matrix B(X) is defined as

$$b_{ik}(X) = \frac{-\delta_{ik}(Z)}{d_{ik}(X)} \text{ if } i \neq k; \quad b_{ii}(X) = \sum_{i \neq k} \frac{\delta_{ik}(Z)}{d_{ik}(X)}; \quad b_{ik}(X) = 0 \text{ if } d_{ik}(X) = 0 \tag{19}$$

In the following sections, techniques that minimize STRESS will be labelled MDS. Their behavior shall be compared with various forms of PCA.

THE UNITED STATES REVISITED (PART I)

We have reanalyzed the data from our example minimizing STRESS. Figure 3 shows the two-dimensional configuration; the accompanying value for ROMRES is .636, which is an improvement of almost 40 percent compared to the PCA solution. A major difference hits you in the eye: the location of AK, HI and NV. The position of the other points shows a striking correspondence with figure 1. If we compare the two solutions with respect to the sum of squared residuals for each state, there are only 4 states for which it is true that the MDS solution is worse than the PCA solution. These states are CA, LA, SC, MS and in figure 3 they are indicated in turn with dotted circles. Inspection of the residuals showed that the increase in stress is caused by the position of AK, HI and NV. Remember that in a PCA solution two points can only be too close. In both the PCA and the MDS solution CA is located too close to the latter mentioned points. In order to minimize the overall stress, MDS is allowed to move points such that two points may become too distant. In the configuration LA, MS and SC are too remote from AK, HI and NV.

Although the location of the problem states shows individual improvements of 80%, 71% and 69% respectively, comparing the MDS residuals with the ones from PCA, these are still the points causing most of the stress, accounting for 23%. This fact will be shown again in a scatter plot, now for approximation from both sides (Figure 4). Splitting up the residuals in approximation error from above and from

below, the ratio of below sum squares to total sum of squares is obtained as .794. The ellipses are again associated with AK, HI and NV. Contrary to the scatter plot for PCA, we now detect numerous ellipses close to the diagonal, which indicates that in the MDS solution these states have obtained an appropriate distance to at least a number of other states.

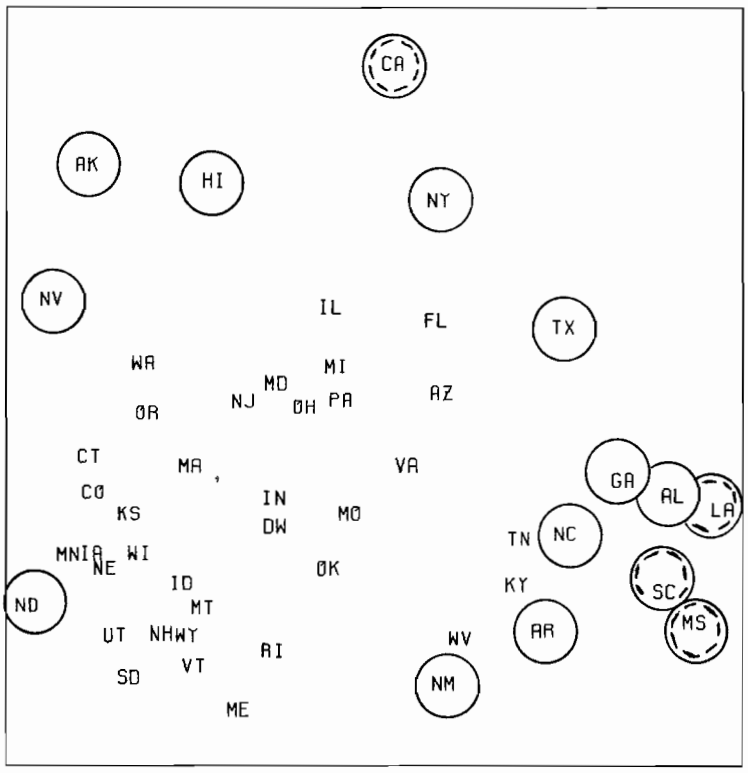


Figure 3. MDS solution for 50 states. Encircled points have dissimilarities larger than average. Dotted circles indicate more stress than PCA solution.

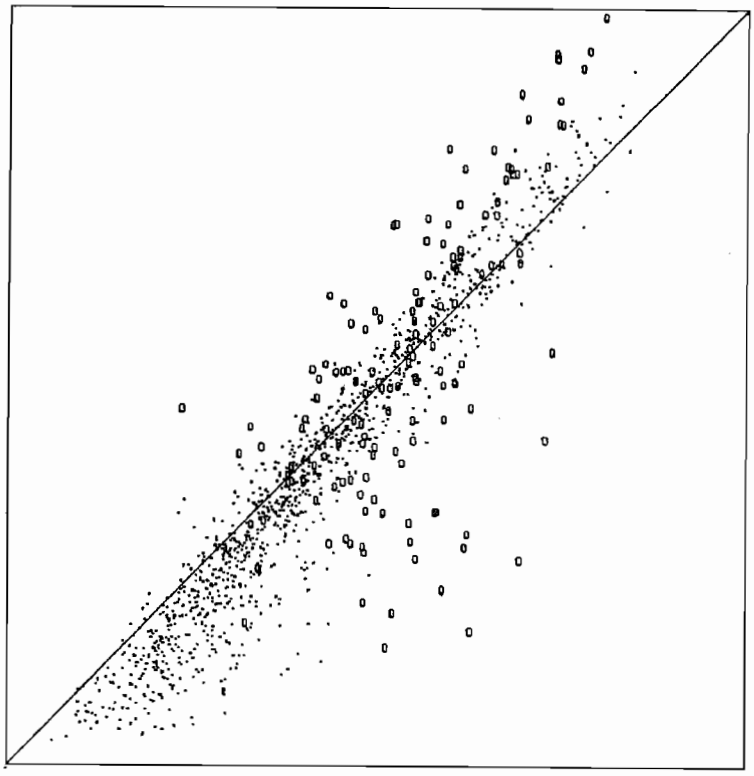


Figure 4. MDS solution for 50 states. $\delta(Z)$ horizontal axis) versus $d(X)$ (vertical axis). Approximation from both sides. Ellipses refer to all pairs including AK, HI and NV.

TRANSFORMING THE DATA

There is another way in which we can improve the fit of a PCA. We still insist on approximation from below, as in PCA, but we allow for optimal transformation of the columns of the data matrix Z. This means that we transform Z to $\underline{Z}$, from below. This means that we transform Z to $\underline{Z}$, column-wise, and we approximate the dissimilarities $\delta_{ik}(\underline{Z})$ from below. We have seen that the obvious badness-of-fit measure in case of approximation from below is the sum of the n - p discarded eigenvalues. The basic new idea in this section is to choose transformation of the columns of Z in such a way that this loss function is minimized. Of course we have to restrict the class of transformations from which we can select admissible transformations in some way or another. Complete freedom in the choice of transformation will lead to degenerate and not very interesting solutions. Thus it is often specified that the transformations of each of the columns must be monotonic, and the resulting columns of $\underline{Z}$ must have mean zero and variance unity. Different classes of transformations have also been used, but we do not go into those aspects of the problem. We now define nonlinear PCA as the technique that combines transformation of the data with quadratic approximation from below. There are various algorithms to perform a nonlinear PCA, compare Gifi (1981) or De Leeuw (1982) for a review. Most of these algorithms minimize the loss function

$$\sigma(\underline{Z},X,U) = \text{tr } (\underline{Z} - XU')'(\underline{Z} - XU') \qquad (20)$$

by using alternating least squares methods. The algorithm we have used starts with an initial estimate of $\underline{Z}$ equal to the standardized data. It then computes estimates of X and U by minimizing the loss function (20) for fixed current $\underline{Z}$. This amounts to performing an ordinary PCA. In the next step the estimate of $\underline{Z}$ is improved by finding the best fitting admissible transformation for each column of XU'. This concludes one cycle of the algorithm and cycles are repeated until convergence. (See Gifi 1982 for an alternative algorithm that is much more efficient).

THE UNITED STATES REVISITED (PART II)

Because we are aware of certain peculiarities in the data, we limited the admissible transformations to third degree polynomials instead of selecting the often chosen class of monotonic transformations, which are less restrictive. The scope of this paper narrows our interest to the performance of nonlinear PCA with respect to the anomalies detected in the previous analyses. The nonlinear PCA solution, by minimizing the sum of the n-p discarded eigenvalues, should give a better two-dimensional representation of the data. This is reflected in the figure for ROMRES, which is .724.

The two-dimensional solution is depicted in Figure 5. The configuration shows a convenient amount of similarity with Figure 1, while at the same time some major differences are apparent. By choosing optimal admissible transformations the technique has been able to replace HI, NV and most notably AK. Again circles are drawn around the points that have obtained relatively large values for $\sum_k \delta_{ik}(\underline{Z})$. NV and AK still belong to this partition, and this time seem properly located at the outskirts of the configuration. Moreover, AK does not belong any longer to the subset of states with a more than average contribution to the total sum of residuals. On the other hand, HI and, excessively, NV still contribute most. To be somewhat more specific: these states have in common that they have too small a distance to RI, ME, OK, UT, OR, WI, DW and KY, and most of all they are too close to each other.
Figure 6 shows the scatter plot of the $\delta_{ik}(\underline{Z})$ versus the $d_{ik}(X)$. In contrast with Figure 2 all large dissimilarities are quite well approximated by large distances. The major part of the composite stress is constituted by approximation errors for medium size dissimilarities. These are linked with HI and NV. When we discard the ellipses in Figure 2, we see a striking resemblance between the remaining cloud of points and Figure 6. Nonlinear PCA appears to have flattened the clearly high dimensional cloud of points Z into a low dimensional cloud $\underline{Z}$.

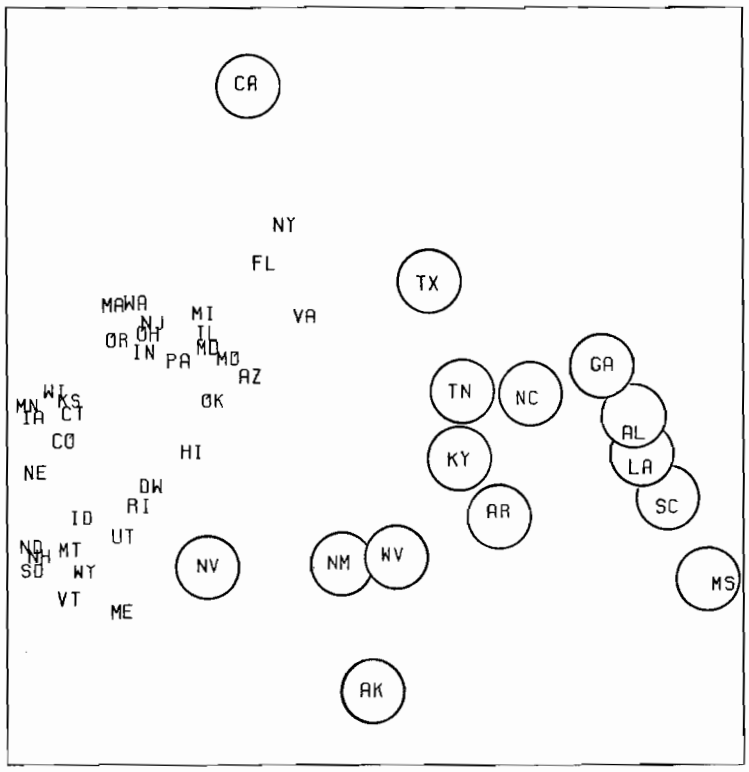


Figure 5. Nonlinear PCA solution for 50 states. Encircled points have dissimilarities larger than average.

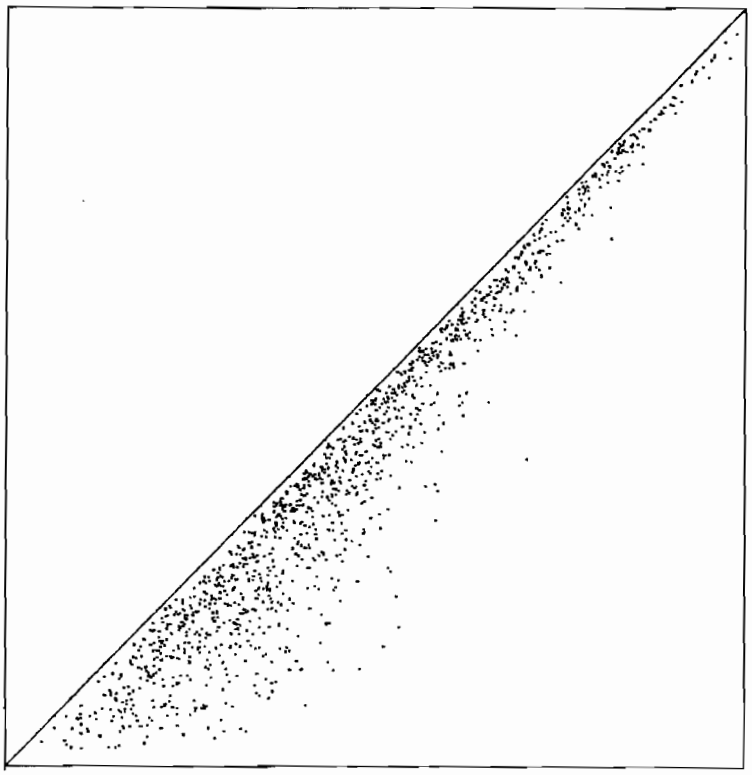


Figure 6 Nonlinear PCA solution for 50 states. $\delta(\underline{Z})$ (horizontal axis) versus d(X) (vertical axis). Approximation from below.

FURTHER IMPROVEMENT OF FIT

We have discussed two methods of improving the fit compared to a simple PCA. The first one was replacing quadratic approximation from below by approximation from both sides, the second one was transformation of the variables. The two methods can be applied independently, and it will consequently not come as a surprise that they also can be combined. In this combined technique we must minimize the loss function

$$\sigma(X,\underline{Z}) = \sum\sum_{ik} (\delta_{ik}(\underline{Z}) - d_{ik}(X))^2 \tag{21}$$

Combining the two ideas is, of course, a very natural step, at least in our framework in which PCA and MDS are treated as two instances of the same basic technique. The combined result is new, however. It seems appropriate to comment briefly on the algorithm.

The algorithm is similar to the one in the previous section in that it consists of two alternating least squares steps. In the first step we have a current best $\underline{Z}$, and we improve our solution by adjusting X. Again the unrestricted algorithm of De Leeuw (1977) is used, in which X is replaced by its Guttman transform (18) (19). In the second alternating least squares step the restricted scaling algorithm of De Leeuw and Heiser (1980) is used to improve $\underline{Z}$ for current X. This step consists in turn of two parts. First we compute the Guttman transform of the old $\underline{Z}$. Thus,

$$\bar{\underline{Z}} = \frac{1}{n} B(\underline{Z})\ \underline{Z}, \tag{22}$$

with

$$b_{ik}(\underline{Z}) = \frac{-d_{ik}(X)}{\delta_{ik}(\underline{Z})} \quad \text{if} \quad i \neq k, \tag{23}$$

and with diagonal elements filled in as (19). The second part of the adjustment of $\underline{Z}$ consists of the least squares projection of the Guttman transform on the space of admissible transformations. This implies solving a regression problem for each column of $\underline{Z}$, and after each regression the column is normalized. From the transormed variables new $\delta_{ik}(\underline{Z})$ are computed, which completes one cycle. The general theory of scaling methods of this type (De Leeuw and Heiser, 1980; De Leeuw, 1984) directly proves convergence of the combined algorithm.

THE UNITED STATES REVISITED (PART III)

The results of the technique that combines transformation of variables with approximation from both sides will be labelled nonlinear MDS, since it has to be definitely distinguished from nonmetric MDS, the notable contribution of Shepard (1962) and Kruskal (1964) to the scaling problem. Its results are quite satisfying regarding the root mean square of residuals, which shows an improvement of 61% compared to no transformation and quadratic approximation from below. Combining the results of the various analyses gives Table 2.

Table 2. Root mean square of residuals

		Transformation	
		No	Yes
Approximation	from below	1.055	.724
	from both sides	.636	.416

We conclude that the effect of approximation is slightly larger than the effect of transformation.

The configuration obtained by nonlinear MDS (Figure 7) shows that the technique has attacked the problem of AK quite drastically. The second dimension is completely dominated by the contrast large population (CA) versus small population (AK). Together with the other encircled states, having relatively large dissimilarities, they form a set almost identical to the partition in the nonlinear PCA solution.

Results for the states left from the center of the configuration look rather disappointing: a lot of states are joined in a rather tight cluster. Since ROMRES is small, we may conclude that these states have become very similar after transformation of the data.

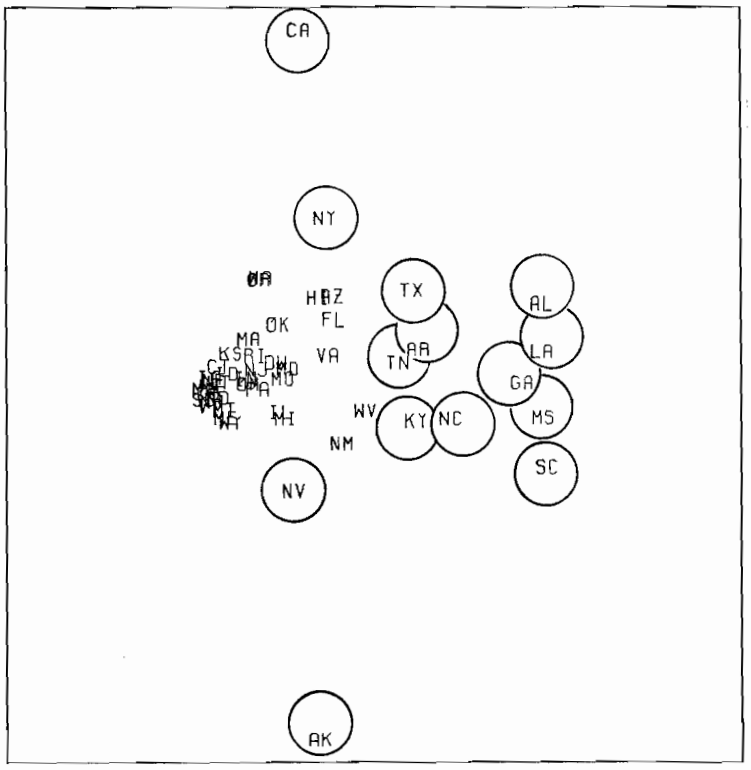


Figure 7. Nonlinear MDS solution for 50 states. Encircled points have dissimilarities larger than average.

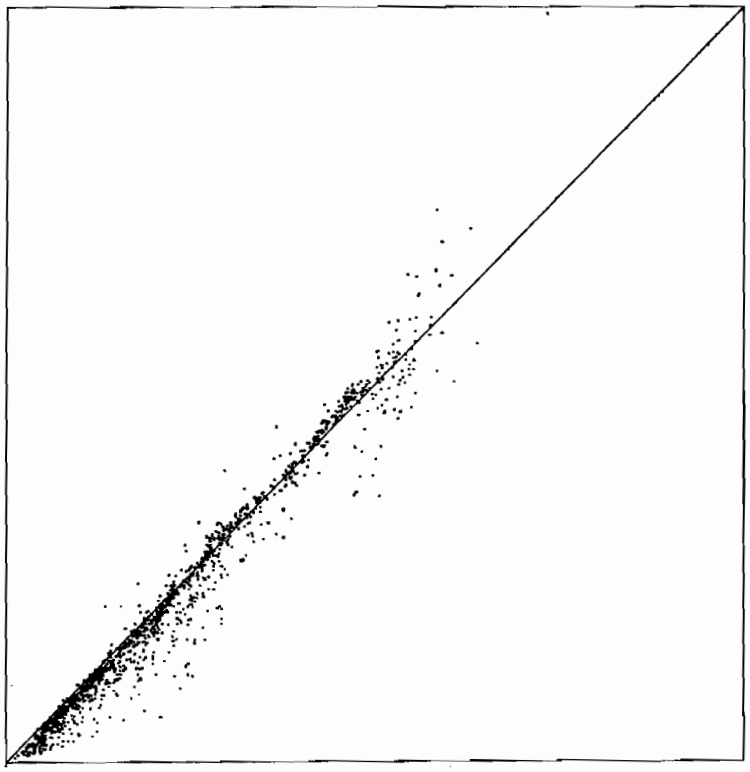


Figure 8. Nonlinear MDS solution for 50 states. $\delta(Z)$ (horizontal axis) versus $d(X)$ (vertical axis). Approximation from both sides.

AK and CA are the states with the largest contribution to the stress. It seems hard to improve upon their position in the configuration since CA is too remote from NV, SC, and NM, while AK is too close to these states. In addition CA is too close to TX, WA, OR and AL, while AK is too remote from those very same states. Moreover, CA and AK should be more close together.
The latter mentioned fact is most clearly illustrated in the scatter plot of dissimilarities versus distances (Figure 8). The complete isolated dot at the top of the figure represents the pair $\{\delta(AK,CA), d(AK,CA)\}$. Following the diagonal downwards we encounter a number of dots with a considerable amount of approximation error: these are all pairs linked with either CA or AK. The only exception is the dissimilarity between CA and NY, which is not large and is matched quite well.

Overlooking the overall results of nonlinear MDS, we might conclude that, by means of its rich resources, the technique has modelled characteristics of certain states that seemed incompatible in two-dimensional space. Compared to nonlinear PCA, nonlinear MDS seems to have failed to retain the mutual differences between the group of states that form the cluster.
But failure, sufficiently dramatized, has its delights[2)]. These are shown in the transformation plots for each variable (Figure 9). Here the n elements of z_j are plotted against the elements of $\underline{z}_j$. For each variable the function fitted by PCA is given next to the one for MDS. For variable 1 both techniques clearly model the special cases CA and NY (largest population), but MDS hardly does account for the variance concerning the rest of the states. The latter remark also applies to the MDS plot for variable 2 (Income), except for AK (rich) and, to a smaller extent, MS and AR (poor). PCA, on the other hand, transforms all income values smoothly, except for a serious anomaly: AK obtains almost the same -very low- value as AR. To cope with the apparent nonlinearity in the data, the relation between population and income, MDS retains for both variables the extreme high values, more or

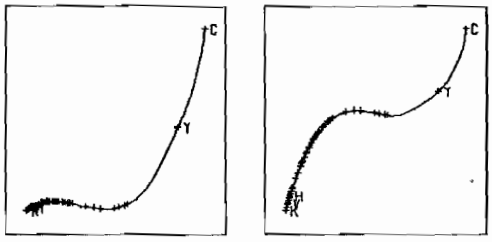
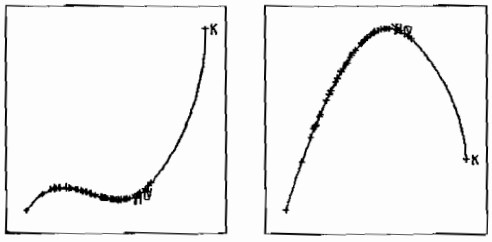
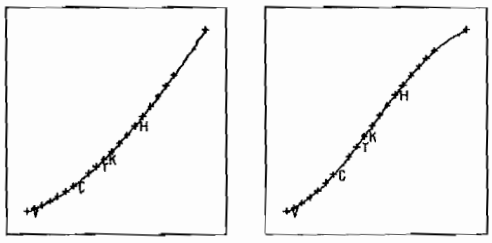


Figure 9. Transformation. MDS on the left, PCA on the right.

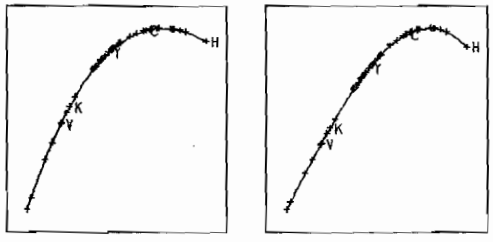
Variable 1: Population



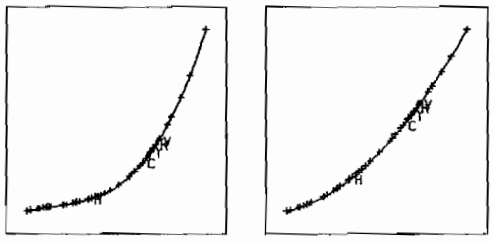
Variable 2: Income



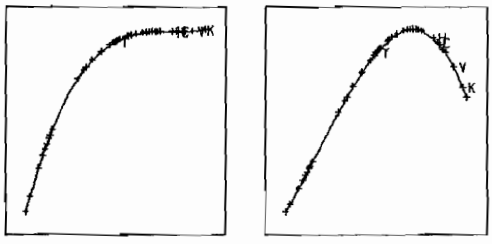
Variable 3: Illiteracy



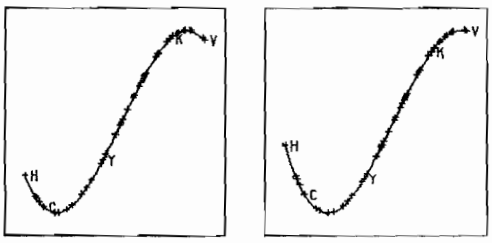
Variable 4: Life expectancy



Variable 5: Homicide



Variable 6: High school



Variable 7: Freezing

less at the expense of the rest, while PCA comes up with a nonlinear transformation. Transformations for illiteracy and homicide are very convincing for both MDS and PCA; the transformations for variable life expectancy are also very similar, both slightly nonmonotonic. Freezing has obtained a S-shaped transformation from both techniques. Variable 6, finally, presents us with another surprise. The percentage of high school graduates is, like income, nonlinearly related to population and PCA comes up with a similar nonlinear transformation. MDS, on the other hand, produces a rather smooth concave function.

CONCLUSION

PCA is a very convenient multidimensional scaling technique. But often it gives a very poor fit, and sometimes it emphasizes rather uninteresting local aspects of the data. We can improve the fit by increasing the number of dimensions, but this has obvious disadvantages from a data analysis point of view. In this paper we have shown that simple improvement of the fit, at a rather low price, is possible by going from approximation from below to approximation from both sides. This will also give a somewhat more balanced representation of the data. More dramatic improvements are possible if we allow for transformations of the data. This can be interpreted as allowing for additional dimensions (parameters), but located at a place where they can be interpreted more easily (in the transformation plots). It appears from our example, and from many other similar examples that we have analyzed, that allowing for transformations can lead to solutions which are qualitatively different. This is much more important than the comparatively trivial finding that they are quantitatively better. Allowing for transformations, especially from large families of admissable transformations, has the danger of partial or complete degeneracy, and may direct even more attention on local properties of the data matrix.

FOOTNOTES

1) We are indebted to Howard Wainer for kindly making these data available to us.
2) We thank Gore Vidal for coining this beautiful phrase.

TABLE 1. Table of social indicator statistics taken from statistical abstracts of the US (1977). U.S. Department of Commerce: Bureau of the census.
Popul : 1975 population in thousands
Incom : Per capita income in dollars
Illit : Illiteracy rate in percent of population
Life : Life expectancy in years
Homic : 1976 homicide and non-negligent manslaughter rate (per 1000)
School: Percent of the population over age 25 who are high school graduates
Freeze: Average numbers of days of the year in which temperature falls below freezing.

State		Popul	Income	Illit	Life	Homic	School	Freeze
Alabama	AB	3615	3624	2.1	69.05	15.1	41.3	20
Alaska	AK	365	6315	1.5	69.31	11.3	66.7	152
Arizona	AZ	2212	4530	1.8	70.55	7.8	58.1	15
Arkansas	AR	2110	3378	1.9	70.66	10.1	39.9	65
California	CA	21198	5114	1.1	71.71	10.3	62.6	20
Colorado	CO	2541	4884	0.7	72.06	6.8	63.9	166
Connecticut	CT	3100	5348	1.1	72.48	3.1	56.0	139
Delaware	DW	579	4809	0.9	70.06	6.2	54.6	103
Florida	FL	8277	4815	1.3	70.66	10.7	52.6	11
Georgia	GA	4931	4091	2.0	68.54	13.9	40.6	60
Hawaii	HI	868	4963	1.9	73.60	6.2	61.9	0
Idaho	ID	813	4119	0.6	71.87	5.3	59.5	126
Illinois	IL	11197	5107	0.9	70.14	10.3	52.6	127
Indiana	IN	5313	4458	0.7	70.88	7.1	52.9	122
Iowa	IA	2861	4628	0.5	72.56	2.3	59.0	140
Kansas	KS	2280	4669	0.6	72.58	4.5	59.9	114
Kentucky	KY	3387	3712	1.6	70.10	10.6	38.5	95
Louisiana	LA	3806	3545	2.8	68.76	13.2	42.2	12
Maine	ME	1058	3694	0.7	70.39	2.7	54.7	161
Maryland	MD	4122	5299	0.9	70.22	8.5	52.3	101
Massachus.	MA	5814	4755	1.1	71.83	3.3	58.5	103
Michican	MI	9111	4751	0.9	70.63	11.1	52.8	125
Minnesota	MN	3921	4675	0.6	72.96	2.3	57.6	160
Mississippi	MS	2341	3098	2.4	68.09	12.5	41.0	50
Missouri	MO	4767	4254	0.8	70.69	9.3	48.8	108
Montana	MT	746	4347	0.6	70.56	5.0	59.2	155
Nebraska	NB	1544	4508	0.6	72.60	2.9	59.3	139
Nevada	NV	590	5149	0.5	69.03	11.5	65.2	188
New Hampsh.	NH	812	4281	0.7	71.23	3.3	57.6	174
New Jersey	NJ	7333	5237	1.1	70.93	5.2	52.5	115
New Mexico	NM	1144	3601	2.2	70.32	9.7	55.2	120
New York	NY	18076	4903	1.4	70.55	10.9	52.7	82
N. Carolina	NC	5441	3875	1.8	69.21	11.1	38.5	80
N. Dakota	ND	637	5087	0.8	72.78	1.4	50.3	186
Ohio	OH	10735	4561	0.8	70.82	7.4	53.2	124
Oklahoma	OK	2715	3983	1.1	71.42	6.4	51.6	82
Oregon	OR	2284	4660	0.6	72.13	4.2	60.0	44
Pennsylv.	PA	11860	4449	1.0	70.43	6.1	50.2	126
Rh. Island	RI	931	4558	1.3	71.90	2.4	46.4	127
S. Carolina	SC	2816	3635	2.3	67.96	11.6	37.8	65
S. Dakota	SD	681	4167	0.5	72.08	1.7	53.3	172
Tennessee	TN	4173	3821	1.7	70.11	11.0	41.8	70
Texas	TX	12237	4188	2.2	70.90	12.2	47.4	35
Utah	UT	1203	4022	0.6	71.72	4.3	67.3	137
Vermont	VT	472	3907	0.6	71.64	5.5	57.1	168
Virginia	VA	4981	4701	1.4	70.08	9.5	47.8	85
Washington	WA	3559	4864	0.6	71.72	4.3	63.5	32
W. Virginia	WV	1799	3617	1.4	69.48	6.7	41.6	100
Wisconsin	WI	4589	4468	0.7	72.48	3.0	54.5	149
Wyoming	WY	376	4566	0.6	70.29	6.9	62.9	173

REFERENCES

Aitchison, J. (1982), The statistical analysis of compositional data. Journal of the Royal Statist. Soc. B, 44, 139-177.

Aitchison, J. (1983), Principal component analysis of compositional data. Biometrika, 70, 57-75.

Benzécri, J.-P. et al. (1973), L'analyse des données 1. La Taxinomie, 2. L'analyse des correspondances. Paris: Dunod.

Benzécri, J.-P. et al. (1980). Pratique de l'analyse des donnees. Paris: Dunod.

Besse, P. (1979), Etude descriptive des processus: Approximation et interpolation. Thèse de 3 cycle. Toulouse: Université Paul-Sabatier.

Cailliez, F. and Pagès, J.-P. (1976), Introduction à l'analyse des données. Paris: SMASH.

De Leeuw, J. (1977), Application of convex analysis to multidimensional scaling. In: J.R. Barra et al. (Eds.), Recent developments in statistics. Amsterdam: North Holland Publishing Company.

De Leeuw, J. (1982), Nonlinear principal component analysis. In: COMPSTAT 1982 Proceedings in computational statistics. Vienna: Physica-Verlag.

De Leeuw, J. (1984), The Gifi system of nonlinear multivariate analysis. In: E. Diday et al. (Eds.), Data analysis and informatics, III. Amsterdam: North-Holland Publishing Company.

De Leeuw, J. (1984), Convergence of the majorization algorithm for multidimensional scaling. Research Report RR-84-07. Leiden: Department of Data Theory.

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

De Leeuw, J. and Heiser, W.J. (1982), Theory of multidimensional scaling. In: P.R. Krishnaiah and L. Kanal (Eds.), Handbook of statistics, Vol. II. Amsterdam: North-Holland Publishing Company.

Domenges, D. and Volles, M. (1980), L'analyse factorielle spherique. In: E. Diday et al. (Eds.), Data analysis and informatics. Amsterdam: North-Holland Publishing Company.

Edwards, A.W.F. and Cavalli-Sforza, L.L. (1964), Reconstruction of evolutionary trees. In: V.H. Heywood & J. McNeill (Eds.), Phenetic and phylogenetic classification. London: Systematics Association Publication No. 6.

Gifi, A. (1981), Nonlinear multivariate analysis. Leiden: Department of Data Theory.

Gifi, A. (1982), PRINCALS user's guide. Leiden: Department of Data Theory.

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

Gower, J.C. (1967), Multivariate analysis and multidimensional geometry. The Statistician, 17, 13-28.

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

Heiser, W.J. and Meulman, J. (1983), Analyzing rectangular tables by joint and constrained multidimensional scaling. Journal of Econometrics, 22, 139-167.

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

Meulman, J. (1982), Homogeneity analysis of incomplete data. Leiden: DSWO Press.

Meulman, J. and Heiser, W.J. (1984), Constrained multidimensional scaling: More directions than dimensions. In: T. Havránek et al. (Eds.), COMPSTAT 1984. Proceedings in computational statistics. Vienna: Physica Verlag.

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

TABLE 1. Table of social indicator statistics taken from statistical abstracts of the US (1977). U.S. Department of Commerce: Bureau of the census.

Popul : 1975 population in thousands
Incom : Per capita income in dollars
Illit : Illiteracy rate in percent of population
Life : Life expectancy in years
Homic : 1976 homicide and non-negligent manslaughter rate (per 1000)
School: Percent of the population over age 25 who are high school graduates
Freeze: Average numbers of days of the year in which temperature falls below freezing.

State		Popul	Income	Illit	Life	Homic	School	Freeze
Alabama	AB	3615	3624	2.1	69.05	15.1	41.3	20
Alaska	AK	365	6315	1.5	69.31	11.3	66.7	152
Arizona	AZ	2212	4530	1.8	70.55	7.8	58.1	15
Arkansas	AR	2110	3378	1.9	70.66	10.1	39.9	65
California	CA	21198	5114	1.1	71.71	10.3	62.6	20
Colorado	CO	2541	4884	0.7	72.06	6.8	63.9	166
Connecticut	CT	3100	5348	1.1	72.48	3.1	56.0	139
Delaware	DW	579	4809	0.9	70.06	6.2	54.6	103
Florida	FL	8277	4815	1.3	70.66	10.7	52.6	11
Georgia	GA	4931	4091	2.0	68.54	13.9	40.6	60
Hawaii	HI	868	4963	1.9	73.60	6.2	61.9	0
Idaho	ID	813	4119	0.6	71.87	5.3	59.5	126
Illinois	IL	11197	5107	0.9	70.14	10.3	52.6	127
Indiana	IN	5313	4458	0.7	70.88	7.1	52.9	122
Iowa	IA	2861	4628	0.5	72.56	2.3	59.0	140
Kansas	KS	2280	4669	0.6	72.58	4.5	59.9	114
Kentucky	KY	3387	3712	1.6	70.10	10.6	38.5	95
Louisiana	LA	3806	3545	2.8	68.76	13.2	42.2	12
Maine	ME	1058	3694	0.7	70.39	2.7	54.7	161
Maryland	MD	4122	5299	0.9	70.22	8.5	52.3	101
Massachus.	MA	5814	4755	1.1	71.83	3.3	58.5	103
Michican	MI	9111	4751	0.9	70.63	11.1	52.8	125
Minnesota	MN	3921	4675	0.6	72.96	2.3	57.6	160
Mississippi	MS	2341	3098	2.4	68.09	12.5	41.0	50
Missouri	MO	4767	4254	0.8	70.69	9.3	48.8	108
Montana	MT	746	4347	0.6	70.56	5.0	59.2	155
Nebraska	NB	1544	4508	0.6	72.60	2.9	59.3	139
Nevada	NV	590	5149	0.5	69.03	11.5	65.2	188
New Hampsh.	NH	812	4281	0.7	71.23	3.3	57.6	174
New Jersey	NJ	7333	5237	1.1	70.93	5.2	52.5	115
New Mexico	NM	1144	3601	2.2	70.32	9.7	55.2	120
New York	NY	18076	4903	1.4	70.55	10.9	52.7	82
N. Carolina	NC	5441	3875	1.8	69.21	11.1	38.5	80
N. Dakota	ND	637	5087	0.8	72.78	1.4	50.3	186
Ohio	OH	10735	4561	0.8	70.82	7.4	53.2	124
Oklahoma	OK	2715	3983	1.1	71.42	6.4	51.6	82
Oregon	OR	2284	4660	0.6	72.13	4.2	60.0	44
Pennsylv.	PA	11860	4449	1.0	70.43	6.1	50.2	126
Rh. Island	RI	931	4558	1.3	71.90	2.4	46.4	127
S. Carolina	SC	2816	3635	2.3	67.96	11.6	37.8	65
S. Dakota	SD	681	4167	0.5	72.08	1.7	53.3	172
Tennessee	TN	4173	3821	1.7	70.11	11.0	41.8	70
Texas	TX	12237	4188	2.2	70.90	12.2	47.4	35
Utah	UT	1203	4022	0.6	71.72	4.3	67.3	137
Vermont	VT	472	3907	0.6	71.64	5.5	57.1	168
Virginia	VA	4981	4701	1.4	70.08	9.5	47.8	85
Washington	WA	3559	4864	0.6	71.72	4.3	63.5	32
W. Virginia	WV	1799	3617	1.4	69.48	6.7	41.6	100
Wisconsin	WI	4589	4468	0.7	72.48	3.0	54.5	149
Wyoming	WY	376	4566	0.6	70.29	6.9	62.9	173