CLASSIFICATION AS A TOOL OF RESEARCH
W. Gaul and M. Schader (Editors)
© Elsevier Science Publishers B.V. (North-Holland), 1986

PRINCIPAL COMPONENT ANALYSIS AND RESTRICTED MULTIDIMENSIONAL SCALING

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). Because of space limitations we shall not treat algorithms and computational aspects. 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 $z_1,\dots,z_m$ 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

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

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

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

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

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

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 $\omega^2_{11} \geq \omega^2_{22} \geq \cdots \geq \omega^2_{nn}$. 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

$$d^2_{ik}(X(1)) \leq d^2_{ik}(X(2)) \leq \cdots \leq d^2_{ik}(X(\rho)) = \delta^2_{ik}(Z), \tag{5}$$

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.

Because our approach to PCA is slightly unconvential, we mention some special cases before we proceed. The first special case is correspondence analysis, the second one is homogeneity analysis, and the third one is standardized PCA. These techniques differ, essentially, in the choice of the weighting matrices A and B, but they use the same geometry, and the same type of approximation. We do not pay attention in this paper to the duality properties of classical PCA, and to the representation of the columns of the data matrix (the variables, or time points). This is mainly because the concepts involved in duality do not generalize naturally to the other data analysis techniques we discuss in this paper.

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 summarized 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:

$$\text{ROMRES} = \left[\frac{1}{n(n-1)} \sum_i \sum_k (\delta_{ik}(Z) - d_{ik}(X))^2 \right]^{\frac{1}{2}} \tag{6}$$

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.

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

$$\sum_k \delta_{ik}(Z) > \frac{1}{n} \sum_i \sum_k \delta_{ik}(Z) \tag{7}$$

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 (7), that

$$\frac{1}{n-1} \sum_k (\delta_{ik}(Z) - d_{ik}(X))^2 > \frac{1}{n(n-1)} \sum_i \sum_k (\delta_{ik}(Z) - d_{ik}(X))^2 \tag{8}$$

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 (8). 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, repre-

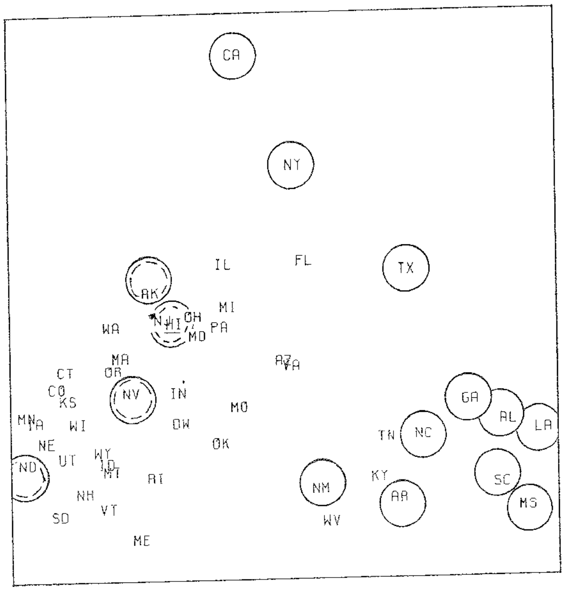


Figure 1. PCA solution for 50 states. Encircled points have dissimilarities larger than average. Dotted circles indicate more than average stress in addition.

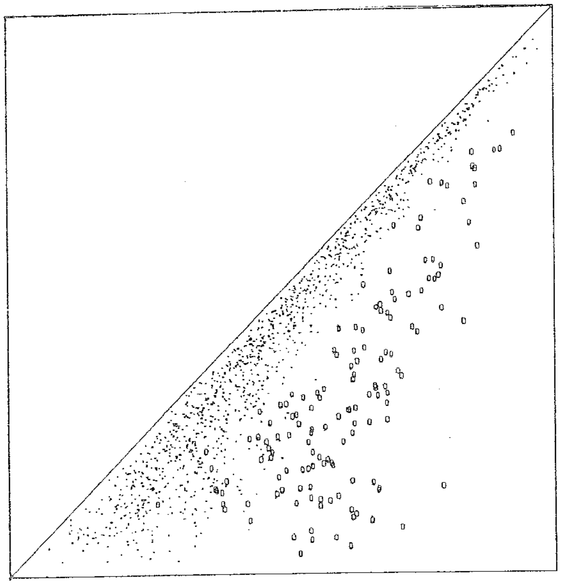


Figure 2. PCA solution for 50 states. δ(Z) (horizontal axis) versus d(X) (vertical axis). Approximation from below. Ellipses refer to all pairs including AK, HI and NV.

senting 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

$$\sigma(X) = \sum_i \sum_k (\delta_{ik}(Z) - d_{ik}(X))^2 \tag{9}$$

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 (9), which is called STRESS and was introduced by Kruskal (1964), will be minimized by the algorithm described by De Leeuw (1977).

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

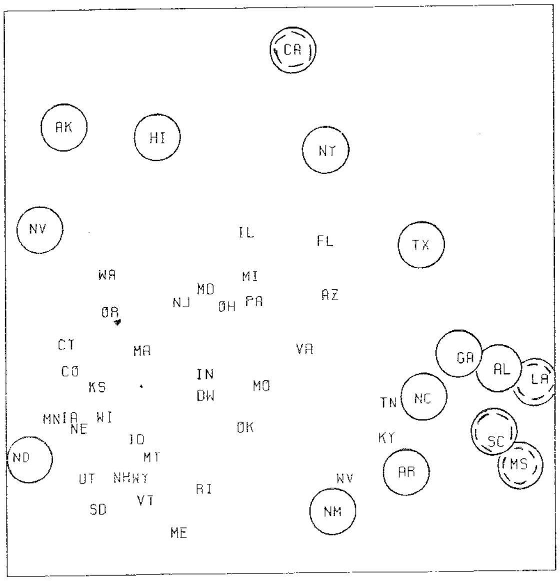


Figure 3. MDS solution for 50 states. Encircled points have dissimilarities larger than average. Dotted circles indicate more stress than PCA solution.

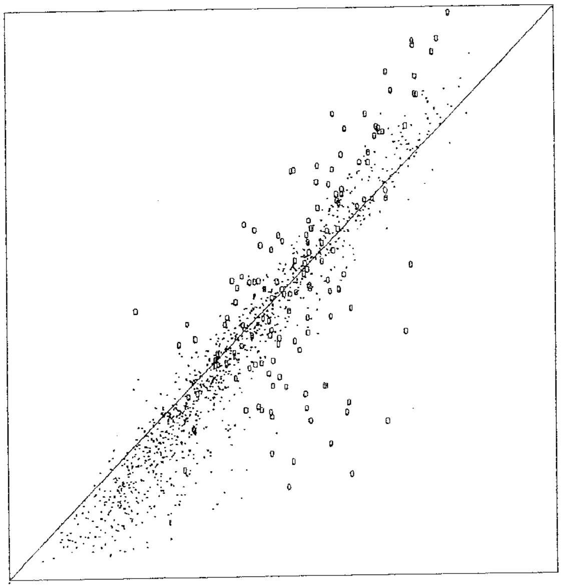


Figure 4. MDS solution for 50 states. δ(Z) horizontal axis) versus d(X) (vertical axis). Approximation from both sides. Ellipses refer to all pairs including AK, HI and NV.

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.

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 $\underline{Z}$, columnwise, 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. For algorithmic and computational details we refer to Gifi (1981, 1982).

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

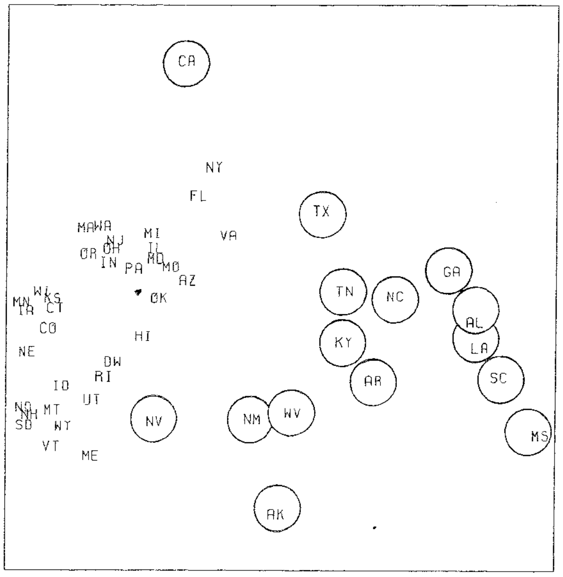


Figure 5. Nonlinear PCA solution for 50 states. Encircled points have dissimilarities larger than average.

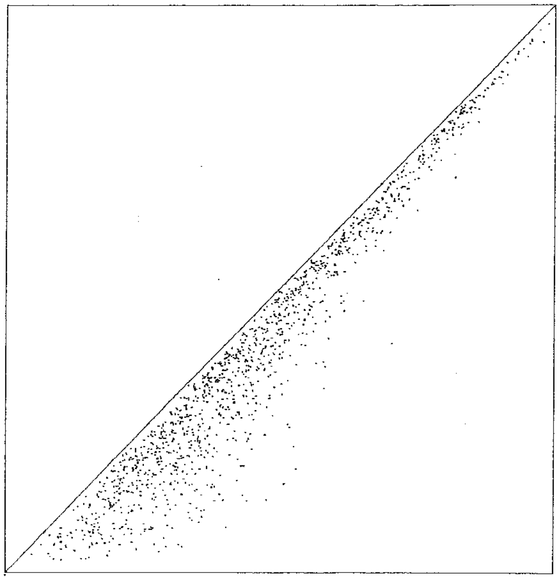


Figure 6. Nonlinear PCA solution for 50 states. $\delta(\underline{Z})$ (horizontal axis) versus d(X) (vertical axis). Approximation from below.

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 $\underline{Z}$.

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

$$\sigma(X,\underline{Z}) = \sum_{ik}\sum (\delta_{ik}(\underline{Z}) - d_{ik}(X))^2 \tag{10}$$

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.

For the algorithm we combine the unrestricted scaling algorithm of De Leeuw (1977) with the restricted scaling method of De Leeuw and Heiser (1980). The implementation is straightforward, given the general principles of algorithm construction outlined in the last mentioned paper.

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

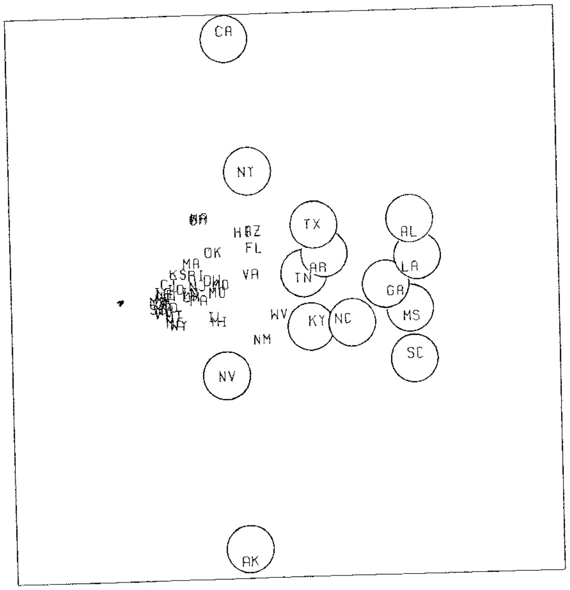


Figure 7. Nonlinear MDS solution for 50 states. Encircled points have dissimilarities larger than average.

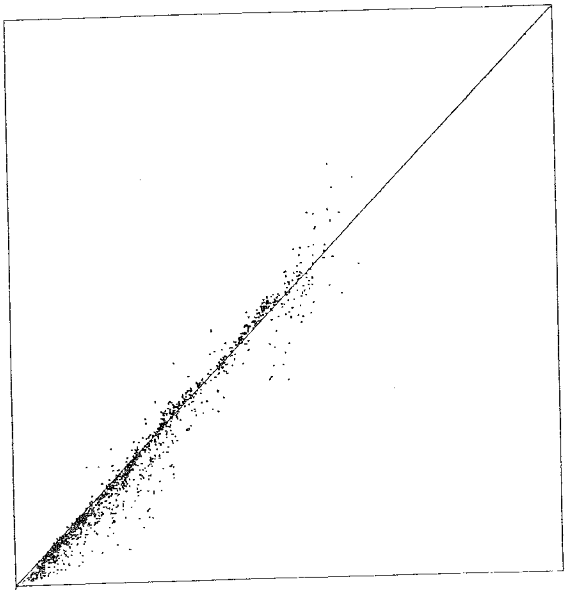


Figure 8. Nonlinear MDS solution for 50 states. $\delta(\underline{Z})$ (horizontal axis) versus $d(X)$ (vertical axis). Approximation from both sides.

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.

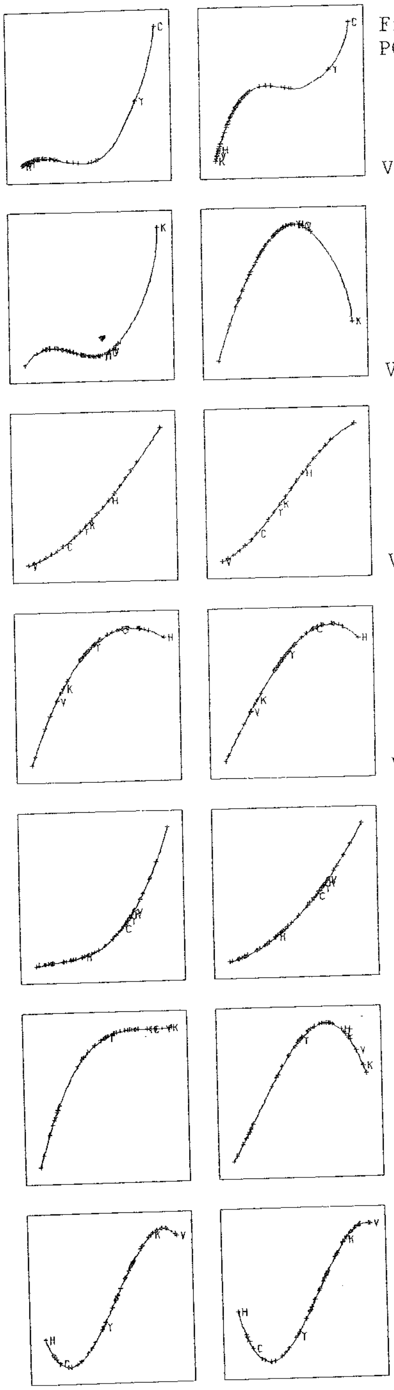
AK and CA are the states with the largest contribution to the stress. It seems hard to improve upon their position in the configuration since CA is too remote from NV, SC, and NM, while AK is too close to these states. In addition CA is too close to TX, WA, OR and AL, while AK is too remote from those 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 less at the expense of the rest, while PCA comes up with a non-

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

linear 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. The complete data matrix can be found in De Leeuw and Meulman (1985), which report also gives the algorithmic details of the various techniques discussed in the present paper.
2) We thank Gore Vidal for coining this beautiful phrase.

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

1 1975 population
2 Per capita income
3 Illiteracy rate
4 Life expectancy
5 1976 homicide and non-negligent manslaughter rate
6 Percent of the population over age 25 who are high school graduates
7 Average numbers of days of the year in which temperature falls below freezing.

States and their abbreviations

Alabama	AL	Alaska	AK	Arizona	AZ	Arkansas	AR
California	CA	Colorado	CO	Connecticut	CT	Delaware	DW
Florida	FL	Georgia	GA	Hawaii	HI	Idaho	ID
Illinois	IL	Indiana	IN	Iowa	IA	Kansas	KS
Kentucky	KY	Louisiana	LA	Maine	ME	Maryland	MD
Massachus.	MA	Michigan	MI	Minnesota	MN	Mississippi	MS
Missouri	MO	Montana	MT	Nebraska	NE	Nevada	NV
New Hampsh.	NH	New Jersey	NJ	New Mexico	NM	New York	NY
N. Carolina	NC	N. Dakota	ND	Ohio	OH	Oklahoma	OK
Oregon	OR	Pennsylv.	PA	Rh. Island	RI	S. Carolina	SC
S. Dakota	SD	Tennessee	TN	Texas	TX	Utah	UT
Vermont	VT	Virginia	VA	Washington	WA	W. Virginia	WV
Wisconsin	WI	Wyoming	WY				

REFERENCES

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

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. 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.

De Leeuw, J. and Meulman, J. (1985), Alternative approximations in principal components analysis, RR-85-13. Leiden: Department of Data Theory.

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.

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. 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.