

**NONLINEAR PRINCIPAL COMPONENTS ANALYSIS:
OVERVIEW AND NEW DEVELOPMENTS
WITH RESPECT TO RESISTANCE PROPERTIES**

Peter verboon

**Department of Data Theory
University of Leiden**

This research was supported by a PSYCHON grant (560-267-029) of the Netherlands organization for scientific research (NWO).

NONLINEAR PRINCIPAL COMPONENTS ANALYSIS: OVERVIEW AND NEW DEVELOPMENTS WITH RESPECT TO RESISTANCE PROPERTIES

Abstract

It is commonly known that Principal Component Analysis (PCA) is very sensitive to outliers in the data. Many methods have already been proposed to make PCA more resistant to outliers. Most of these methods take the covariance matrix as their starting point. In this paper another approach is used, which computes the parameters directly from the data. In this way it is possible to incorporate nonlinear transformations of the variables.

The least squares criterion to fit the PCA model is replaced by two resistant criteria, namely the Huber and biweight. An iteratively reweighted least squares algorithm to minimize these criteria is proposed. Furthermore, to avoid degenerations of the solutions, the monotonic regression is extended by defining bounds on the transformations. Finally two examples are presented, that show all the aspects discussed in the paper.

Key words: nonlinear PCA, resistance, robustness, Huber function, biweight function, iteratively reweighted least squares, bounded monotonic regression.

1. INTRODUCTION

An important technique for multivariate analysis is Principal Components Analysis (PCA). The objective of PCA is to represent high-dimensional data in a low-dimensional space. In this space of reduced dimensionality it is often more easy to find interesting relations between the variables and interesting structures for the data points.

Introduction

This paper discusses the problem of finding principal components in data that contain outliers. It is well-known, that with the classical least squares criterion, outliers can be highly disturbing for the solution. Therefore, we study alternatives for the least squares criterion, which are known to be more resistant to outliers than least squares. First an overview is given of resistant methods, that take the covariance matrix as a starting point. Next, methods will be considered that compute the solution directly from the data. Furthermore, we will extend the problem by considering nonlinear transformations of the variables (Gifi, 1990).

Consider the data matrix \mathbf{Z} as a m -dimensional cloud of n points. Finding principal components can be seen as finding directions in this cloud. The first principal component is chosen as the direction which yields the largest variance of the points projected on this direction. In other words: the average correlation between the first principal component and all observed variables is maximal. So the first principal component is the direction that explains more variance of the data than any other direction. Projections on the second principal component have maximum variance under the restriction that the second principal component is orthogonal to the first. The third component has maximum variance under the restriction that it is orthogonal to the previous two; and so on.

Having found p principal components, a weighted sum of these components yields the best representation of the data in p dimensions (Eckart & Young, 1936). The projections on the principal components are called component scores and the weights are called component loadings, furthermore, the variances of the component scores are called eigenvalues.

If there are outliers in the data and we use the least squares criterion to find the principal components, there exists a considerable danger that the general pattern in the data is not discovered. The first few principal components are especially sensitive to outliers that *inflate* the correlation between variables. This type of outlier is out of range with respect to a particular variable, that is, it has an extreme value on that variable. The variance in the direction of the outlier is increased by the outlier, therefore this direction tends to become one of the first principal components.

There is also another type of outlier. These are outliers that are not apparent with respect to the original variables but are points that do not conform to the overall correlation structure of the data and are therefore (mostly) *deflating* the correlation structure of the variables. Usually these outliers will be found in the last few principal components (Hawkins & Fatti, 1984).

These phenomena strongly resemble the corresponding sensitivity to outliers in the context of multiple regression. In fact, we could describe PCA as a multivariate multiple regression problem, where we have unobserved variables, the principal components, instead of predictor variables. However, with respect to resistance there is one important difference. In regression analysis outliers in the predictor space, leverage points, are highly disturbing. In PCA the equivalent of the predictor space is the space spanned by the principal components, which are derived from the data and therefore in principal are void of outliers. In this respect we expect that PCA is actually more simple than regression analysis because in PCA we don't have to be concerned about leverage points.

Basically there are two different strategies for performing a PCA. In the classical approach the covariance (or correlation) matrix is computed first and by some factorization of this matrix the desired components are found. The second approach is based on the direct analysis of the data matrix by a singular value decomposition or by iterative regression procedures (alternating least squares). In the alternating least squares approach it is possible to allow for nonlinear transformations of the observed variables, which after convergence are optimal with respect to the least squares criterion. (Gifi, 1990, ch. 4; Young *et al.*, 1978).

Because of the property that it shows the two types of outliers in the first or last few principal components, PCA is sometimes used as a technique to detect outliers in the high-dimensional space. Unfortunately in practice some points may actually be a mixture of the two outlier types, which makes them much harder to detect, while their influence upon the solution might still be quite considerable. Therefore a resistant version of PCA might be a convenient extension for the toolbox of multivariate analyses techniques.

2 RESISTANT PCA VIA THE COVARIANCE MATRIX

A large number of techniques in multivariate analysis take the covariance matrix as a starting point. It is obvious that instead of the covariance matrix (\mathbf{S}) the correlation matrix (\mathbf{R}) can also be taken, which is just a rescaled version of \mathbf{S} , in which the differences in variances between the variables are ignored. For PCA an eigenanalysis of the symmetric matrix \mathbf{S} yields the following so-called spectral decomposition:

$$\mathbf{S} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}' \tag{1}$$

Resistant PCA via the covariance matrix

The orthonormal columns of \mathbf{A} are the eigenvectors of \mathbf{S} and the diagonal matrix $\mathbf{\Lambda}$ contains the eigenvalues in descending order. A low-dimensional approximation of \mathbf{S} is found by taking the first p columns of \mathbf{A} together with the first p eigenvalues, thus

$$\mathbf{S}_p = \mathbf{A}_p \mathbf{\Lambda}_p \mathbf{A}_p'. \quad (2)$$

Now the component scores may be found through the singular value decomposition of the original data matrix \mathbf{Z} , written as $\mathbf{P}\mathbf{\Lambda}^{1/2}\mathbf{A}'$. Next we may choose the component scores as $\mathbf{P}\mathbf{\Lambda}^{1/2}$, thus

$$\mathbf{X}_p = \mathbf{Z}\mathbf{A}_p. \quad (3)$$

Sometimes it is convenient to require that $\mathbf{X}'\mathbf{X} = n\mathbf{I}$. In that case we must divide the columns of \mathbf{X} by its corresponding singular value.

A resistant procedure for PCA via the covariance matrix, starts with the computation of a resistant alternative of this matrix. There are several possibilities to obtain a resistant covariance matrix, denoted as \mathbf{S}^* .

One of the first ideas is to compute the matrix \mathbf{S}^* elementwise. This approach is discussed in Gnanadesikan and Kettenring (1972) and in Gnanadesikan (1977). They proposed to compute some univariate resistant measure of location, the median for instance, and then compute α -trimmed variances for each of the variables. The covariances and correlations are derived from these resistant variances. For instance, via the identity

$$\text{cov}(\mathbf{z}_1, \mathbf{z}_2) = \frac{1}{4} (\text{var}(\mathbf{z}_1 + \mathbf{z}_2) - \text{var}(\mathbf{z}_1 - \mathbf{z}_2)). \quad (4)$$

A resistant measure for the covariance between the vectors \mathbf{z}_1 and \mathbf{z}_2 can be obtained from

$$s^*(\mathbf{z}_1, \mathbf{z}_2) = \frac{1}{4} (s^*(\mathbf{z}_1 + \mathbf{z}_2) - s^*(\mathbf{z}_1 - \mathbf{z}_2)), \quad (5)$$

where $s^*(\mathbf{z}_1 + \mathbf{z}_2)$ and $s^*(\mathbf{z}_1 - \mathbf{z}_2)$ represent for instance the α -trimmed variances or some other resistant measure of variance. With these covariance and variances the correlation coefficient between \mathbf{z}_1 and \mathbf{z}_2 can easily be found. However, these measures are determined without considering the well-known Cauchy-Schwarz inequality relationship between covariance and variances, therefore there lies a considerable danger in this approach that the resulting correlation matrix \mathbf{R}^* is not positive-definite. An additional step must therefore be taken to obtain positive-definiteness. See for an extensive discussion of this approach Gnanadesikan and Kettenring (1972).

Mosteller and Tukey (1977, ch. 10) introduce another idea to arrive at the covariance matrix. They want to retain the *quadratic nature* of resistant variances and covariances. Their procedure starts with the computation of a new set of variables or components (\mathbf{Y}) by regressing variable z_j on z_1 to z_{j-1} ($j=2, \dots, m$). So \mathbf{Y} is formed by

$$\mathbf{Y} = \mathbf{Z} [\mathbf{I} - \mathbf{B}], \quad (6)$$

where \mathbf{B} is a strictly upper triangular matrix with regression coefficients. These regression coefficients are obtained by means of a resistant regression procedure, for instance via iteratively reweighted least squares using the Huber weights (Huber, 1981). Next, the variances of \mathbf{Y} are computed in a resistant way and gathered in the diagonal matrix \mathbf{D}^* .

Now the resistant covariance matrix \mathbf{S}^* is defined by

$$\mathbf{S}^* = [\mathbf{I} - \mathbf{B}']^{-1} \mathbf{D}^* [\mathbf{I} - \mathbf{B}]^{-1}. \quad (7)$$

Note that the components of \mathbf{Y} are orthogonal, thus their covariances are 0, which makes \mathbf{D}^* diagonal. Again the correlation matrix \mathbf{R}^* can simply be derived from \mathbf{S}^* , but now \mathbf{R}^* is positive-definite. A serious disadvantage of this approach is the loss of invariance with respect to permutation of the columns of \mathbf{Z} . Mosteller and Tukey (1977) comment on this as follows: "*We are acting as if invariance were nice, but can be dispensed with, while resistance and robustness of efficiency cannot be spared.*" (p. 211).

Other multivariate approaches which consider all variables simultaneously are described in Gnanadesikan and Kettenring (1972), in Devlin et al. (1975; 1981) and in Campbell (1980). The idea is to compute covariance and location estimates based on "good" points only. So the problem is to find some measure that indicates which points are outlying in the high-dimensional space.

A possibility is to compute resistant distances from each point in the high-dimensional space to the centre of the point cloud. Points which have a relatively large distance to the centre of the cloud are considered as outliers. The classical approach is to compute the Mahalanobis distances. However, the computation of Mahalanobis distances requires estimates of means and covariances to correct for variances and intercorrelations of the variables. If we use the ordinary mean and the least squares covariance matrix for computing the Mahalanobis distances, these distances will seriously be distorted when the data contain outliers. A straightforward solution for this problem is to use resistant measures of location and scale. It follows that an iterative procedure is necessary which

Resistant PCA via the covariance matrix

finds resistant Mahalanobis distances in one step and by using these distances finds resistant covariances and means in the other.

The squared resistant Mahalanobis distances are defined as

$$d_i^2 = (\mathbf{z}_i - \mathbf{m}^*)' \mathbf{S}^{*-1} (\mathbf{z}_i - \mathbf{m}^*), \quad (8)$$

where \mathbf{m}^* is a vector with resistant measures of location, for instance the coordinatewise medians, and \mathbf{S}^* the resistant covariance matrix.

Having found these distances some decreasing function can be defined, which attributes weights to the points as a function of the resistant Mahalanobis distances of these points to the centre. For instance the Huber weights would be a good choice.

$$w_i = \begin{cases} 1 & \text{if } d_i < c \\ \frac{c}{d_i} & \text{if } d_i \geq c \end{cases} \quad (9)$$

where c is some tuning constant which modifies the sensitivity of the procedure. With these weights we define the weighted mean as

$$\mathbf{m}^* = \frac{\sum w_i \mathbf{z}_i}{\sum w_i}, \quad (10)$$

and a weighted covariance matrix as

$$\mathbf{S}^* = \frac{\sum w_i (\mathbf{z}_i - \mathbf{m}^*) (\mathbf{z}_i - \mathbf{m}^*)'}{\sum w_i}. \quad (11)$$

Having found the \mathbf{m}^* and \mathbf{S}^* , new distances are computed, and so on.

A slightly different approach is described in Devlin *et al.* (1981), who use the Mahalanobis distances for a so called multivariate trimming (MVT) procedure. The $(100 - \alpha)$ percent points with the smallest distances are used to compute \mathbf{m}^* and \mathbf{S}^* . This could be seen as attributing zero weights to a fixed percentage of points, even when there are no outliers at all. Naturally this procedure is iterative too, since the \mathbf{m}^* and \mathbf{S}^* are needed for the computation of the Mahalanobis distances and conversely these distances are needed for \mathbf{m}^* and \mathbf{S}^* .

A measure for distance, which is used in (9) to obtain weights, can also be found differently. A highly resistant measure to identify outliers in the high-dimensional space is

the so called *outlyingness index* (u), which is related to projection pursuit methods (Huber, 1985).

$$u_i = \max_{\|\mathbf{a}\|=1} \frac{|\mathbf{z}_i' \mathbf{a} - \text{med}\{\mathbf{Za}\}|}{\text{MAD}\{\mathbf{Za}\}}, \quad (12)$$

where the MAD is the median absolute deviation, a resistant measure of scale. The direction vector \mathbf{a} can be found by $\mathbf{a} = \mathbf{z}_i - \mathbf{m}^+$, where \mathbf{m}^+ is the coordinatewise median of the data (Rousseeuw *et al.*, 1990). To see more clearly what the outlyingness index actually represents, we may write it in a different form as

$$u_i = \frac{\|\mathbf{z}_i\| \cos \alpha_i - \text{med}\{\|\mathbf{z}_j\| \cos \alpha_j\}}{\text{MAD}\{\|\mathbf{z}_j\| \cos \alpha_j\}}, \quad (13)$$

where $j = 1, \dots, n$. The angle α_j denotes the angle between the vector \mathbf{a} and the vector \mathbf{z}_j . With relatively homogeneous data, thus when \mathbf{z}_i is not an outlier, the angle between \mathbf{a} and the vector \mathbf{z}_i will be similar to the median of all other angles. Furthermore the length of vector \mathbf{z}_i will approximately be equal to the median of the lengths of all other vectors. Therefore the numerator of (13) will be small. The denominator removes the effect of the dispersion of the points. It follows that the outlyingness indices are small for homogeneous points, even when the total dispersion of the points is large.

Suppose there are $n-1$ homogeneous points and one outlier. The $\text{MAD}\{\|\mathbf{z}_j\| \cos \alpha_j\}$, which is known as a resistant measure of dispersion, will definitely be small and the $\text{med}\{\|\mathbf{z}_j\| \cos \alpha_j\}$ will depend on cosines and also on the length of the vectors. For all good points \mathbf{z}_i the term $(\|\mathbf{z}_i\| \cos \alpha_i)$ differs not much from the term $\text{med}\{\|\mathbf{z}_j\| \cos \alpha_j\}$ and therefore the outlyingness index will be low. The outlier, however, will have a different angle with the other vectors or a different length. It follows that its outlyingness index will be large.

Having found the outlyingness index of each point, weights can be computed like before and with these weights the weighted covariance matrix is derived. Contrary to the Mahalanobis distances these indexes do not use a resistant covariance matrix and therefore this procedure is not iterative.

Another highly resistant method to estimate the covariance matrix is based on the minimum volume ellipsoid (Rousseeuw & van Zomeren, 1990). The ellipsoid with minimum volume is obtained via a resampling algorithm. The ellipsoids are chosen in such a way that they always contain exactly half of the points. These points have the shortest averaged squared Mahalanobis distance to centre of the point cloud. Finally the

Resistant PCA via the covariance matrix

covariance is computed from the points in the ellipsoid that has a minimum volume over all sampled ellipsoids. This procedure is highly resistant since only half of the points are considered.

Finally, for sake of completeness, we refer to the approach, proposed by Li and Chen (1985), which is also based on projection pursuit ideas. In this approach a so-called projection index is maximized, which makes the approach more similar to the methods discussed in the next section.

3. RESISTANT PCA VIA THE DATA MATRIX

Another approach is to estimate the principal components directly from the data, thus, without first computing the covariance matrix. The procedure is based on alternatingly computing the component scores and component loadings by regression procedures.

The objective is to find the following factorization of \mathbf{Z} :

$$\mathbf{Z} \cong \mathbf{X}\mathbf{A}' \quad (14)$$

Here \mathbf{X} is the matrix with components scores in p dimensions and \mathbf{A} is the $m \times p$ matrix with the component loadings. To find \mathbf{X} and \mathbf{A} we must minimize the following loss function

$$\sigma(\mathbf{X}, \mathbf{A}) = \text{tr}(\mathbf{Z} - \mathbf{X}\mathbf{A}')'(\mathbf{Z} - \mathbf{X}\mathbf{A}'), \quad (15)$$

with the normalization constraint $\mathbf{X}'\mathbf{X} = n\mathbf{I}$. First consider a $p=1$ approximation of \mathbf{Z} . The loss function for the first principal component then becomes

$$\sigma(\mathbf{x}_1, \mathbf{a}_1) = \sum_{j=1}^m (\mathbf{z}_j - \mathbf{x}_1 \mathbf{a}_1')'(\mathbf{z}_j - \mathbf{x}_1 \mathbf{a}_1'). \quad (16)$$

Both unknown parameters \mathbf{x}_1 and \mathbf{a}_1 are easily computed via simple regression equations. For the first component we may iterate between \mathbf{x}_1 and \mathbf{a}_1 until the solution stabilizes, while each time we add the required normalization restrictions. After convergence the product $\mathbf{x}_1 \mathbf{a}_1'$ is the best least squares rank one approximation of \mathbf{Z} , which implies that it is the direction with the largest variance.

After having found the first principal component the process can be repeated for the second component. The data are now replaced by the residuals from the rank one approximation, by subtracting the approximation from the original data.

$$\mathbf{Z}^{(-1)} = \mathbf{Z} - \mathbf{Z}^1, \quad (17)$$

where \mathbf{Z}^1 represents the rank one approximation. In this way we can compute all p principal components.

The ordinary least squares problem may be generalized by introducing weights. For fixed weights, where each weight corresponds to each separate entry in the data matrix, this problem is studied in Gabriel and Zamir (1979). They distinguish two procedures to handle this problem: the procedure of *cyclic dyadic fits* and the *multiple criss-cross regression* procedure. In Gabriel and Odoroff (1984) some resistant aspects of PCA are discussed in relation to the procedures mentioned above. A key observation is that the above mentioned procedure of iterative fitting is in fact "reciprocal averaging". In order to obtain a resistant procedure a logical step is to replace the ordinary averaging by a form of resistant averaging. Some of the possibilities for such a resistant approach are given in Gabriel and Odoroff (1984).

3.1 The Huber and Biweight functions

In this paper we will elaborate upon another approach that uses a resistant loss criterion instead of the least squares criterion such as defined in (15). The problem to be minimized in PCA can be very generally written as:

$$\sum_{j=1}^m \sum_{i=1}^n f(z_{ij} - \mathbf{x}_i \mathbf{a}_j'). \quad (18)$$

When we represent the residual by r_{ij} then, instead of choosing for $f(r_{ij})$ the least squares criterion, we will choose for the function $f(r_{ij})$ Huber's function or Tukey's biweight function (Mosteller & Tukey, 1977). These functions have already proved to be resistant to outliers in the context of regression problems and also in the orthogonal Procrustes problem (Verboon and Heiser, 1991). The functions, together with their first derivative are shown in the Figure 1 and 2.

Resistant PCA via the data matrix

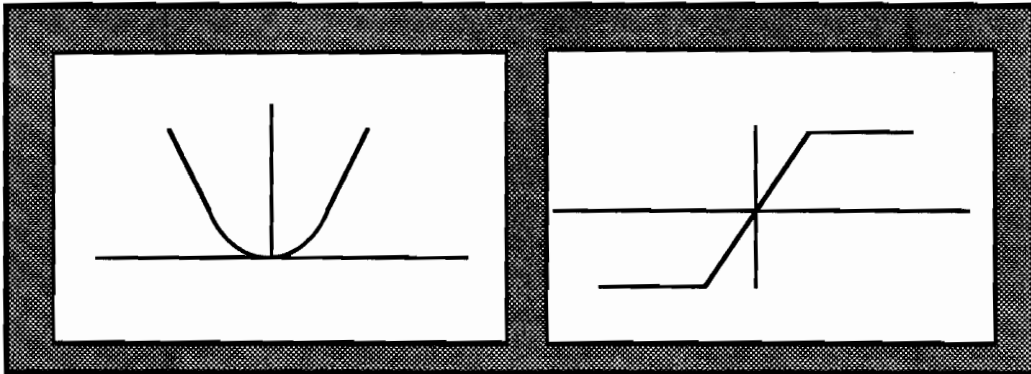


Figure 1. The Huber function and its derivative.

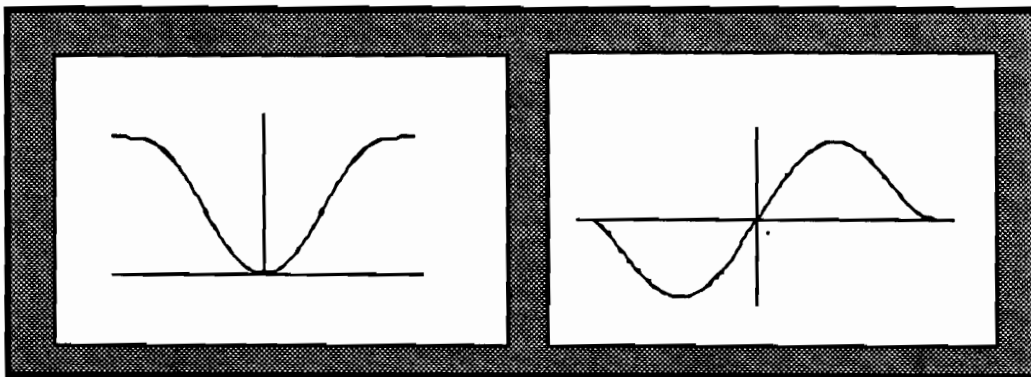


Figure 2. The Biweight function and its derivative.

Both functions can be minimized by using an iteratively reweighted least squares (IRLS) procedure (Verboon, 1990). This algorithm alternates between two steps. One step of the algorithm the problem is minimizing:

$$\sigma(\mathbf{X}, \mathbf{A}) = \sum_j (\mathbf{z}_j - \mathbf{X}\mathbf{a}_j)' \mathbf{V}_j (\mathbf{z}_j - \mathbf{X}\mathbf{a}_j), \quad (19)$$

with a fixed diagonal matrix \mathbf{V}_j . In the other step the weights are updated; they are chosen as a monotonic decreasing function of the residuals. Different choices for the weight function correspond with different resistant functions. In fact such a procedure closely resembles the IRLS procedure in multiple regression problems. A solution for the problem defined in (19) can be found in Verboon *et al.* (1991).

3.2 Aggregating the residuals

In the previous section, different weight matrices \mathbf{V}_j were considered, one for each variable. An interesting special case occurs when we assume that all \mathbf{V}_j are equal to each

other, thus $\mathbf{V}_j = \mathbf{V}$ for $j = 1, \dots, m$. This implies that we have n weights, one for each object, instead of $n \times m$ weights, one for each cell in the data matrix.

So, instead of applying the loss function to each residual element, r_{ij} , and then summing all these loss values to obtain the overall loss, we aggregate the residuals over the rows and apply the loss function to these n aggregated values, denoted as d_i ($i=1, \dots, n$). Handling the residuals rowwise, we should first compute the residuals per row, d_i . The value d_i is computed as the Euclidean distance between an object in the m -dimensional space and its model values, thus

$$d_i = \sqrt{\sum_{j=1}^m (z_{ij} - \mathbf{x}_i \mathbf{a}_j')^2}. \quad (20)$$

Now the Huber or biweight function can be applied to these values to obtain a loss per row, next a summation of these row losses yields the total loss.

For least squares both types of handling the residuals are equivalent, however, for the Huber or biweight function these two approaches lead to two different situations. Elementwise weighting is more flexible, since small weights could be assigned to separate scores of an object, leaving its other scores unaffected. On the other hand rowwise weighting might conceptually (and computationally) be more attractive, since it considers whole objects as possible outliers.

4. ADDING NONLINEAR TRANSFORMATIONS

If the variables are measured on an ordinal or nominal level, one may allow for nonlinear transformations of the variables in order to obtain better fitting results. The objective is to find optimal quantifications of the scores of \mathbf{Z} , where optimal is defined in terms of the loss function. We will denote the optimal quantified matrix as \mathbf{Q} , thus

$$\mathbf{q}_j = \tau_j(\mathbf{z}_j), \quad (21)$$

where τ_j represents the transformation function for variable j . Combining (19) and (21) gives the following loss function for nonlinear PCA:

Nonlinear transformations

$$\sigma(\mathbf{Q}, \mathbf{X}, \mathbf{A}) = \sum_{j=1}^m (\mathbf{q}_j - \mathbf{X}\mathbf{a}_j)' \mathbf{V}_j (\mathbf{q}_j - \mathbf{X}\mathbf{a}_j). \quad (22)$$

This loss function is an extension of the PRINCIPALS loss function (Young *et al.*, 1978), because of the weights matrices \mathbf{V}_j . In order to avoid degenerate solutions, it is necessary to keep the lengths of all \mathbf{q}_j fixed throughout the algorithm. If we want to minimize (22) for a fixed set of weights, then a weighted normalization, namely $\mathbf{q}_j' \mathbf{V}_j \mathbf{q}_j = \mathbf{u}' \mathbf{V}_j \mathbf{u}$ is appropriate (Verboon *et al.*, 1991). This normalization says that the weighted variables are normalized to the sum of their weights.

However, in the IRLS approach the lengths of \mathbf{q}_j should be kept fixed for different weights in order to achieve monotonic convergence. If we would use the weighted normalization, we would have a different set of variables in each step. So for the normalization the weights are ignored and the normalization $\mathbf{q}_j' \mathbf{q}_j = n$ is chosen.

Now the problem is defined in the metric \mathbf{V}_j and the normalization in the metric \mathbf{I}_j . The procedure to solve such a problem is based on iterative majorization (Heiser, 1987; Verboon *et al.*, 1991).

4.1 Optimal scaling and weighting the residuals

The concepts of optimal scaling and resistance seem to be very much entangled. To see this, let us first examine two possible situations in optimal scaling. When we are dealing with ordinal variables, the following transformation functions could occur.

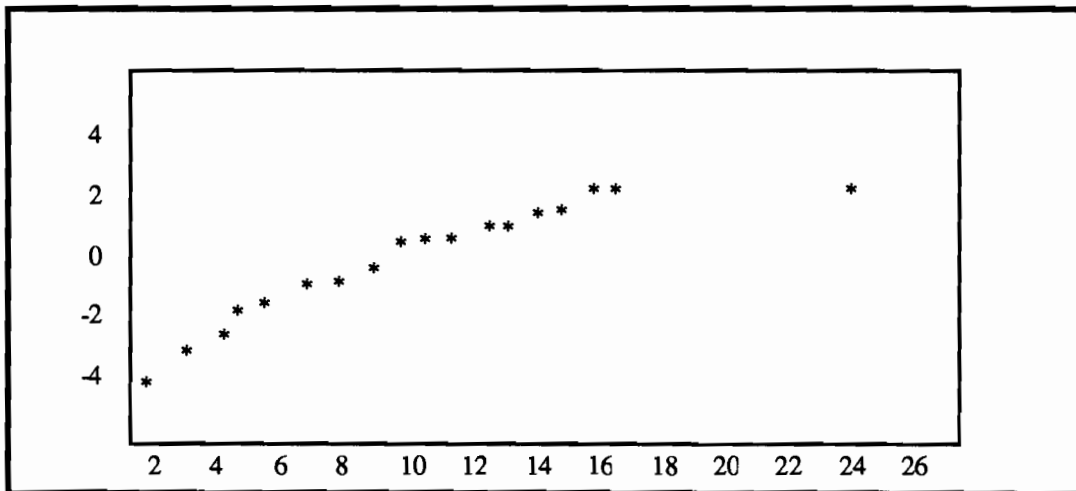


Figure 3. The smoothing effect of optimal scaling illustrated for a continuous ordinal variable (original vs. optimal values).

In the first example (Figure 3) we see that an extreme value in the original coding of a variable, which is a potential outlier of the first type, could be brought back in range by the optimal scaling. In this case optimal scaling takes care of the outliers of the first type, although the healing effect in case of ordinal variables is always limited, since the quantification must be a monotonic transformations of the original values. Despite this limitation optimal scaling could well diminish the influence of a potential outlier.

In Figure 4 quite a different situation is sketched. Now optimal scaling transforms a variable without extreme values to one with a very extreme quantification. So here we have the situation in which optimal scaling creates outliers of the first type. Now the first principal components are completely dominated by these badly transformed variables.

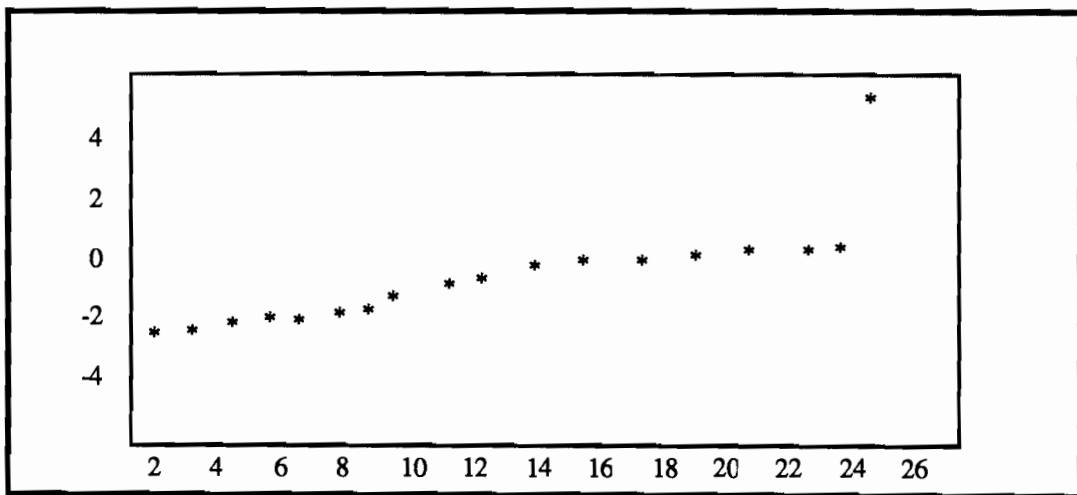


Figure 4. The "outlier-creating" effect of optimal scaling illustrated for a continuous ordinal variable (original vs. optimal values).

After some experimentation it appeared that the second situation, in which the outliers were created, occurred quite often. Using resistant functions, like the Huber and the Biweight, the frequency of occurrence was even higher. The following two situations occurred. First, for some ordinal variables the optimal scaling created outliers, which were badly fitted in the low-dimensional space. Consequently these points had a large residual, which was downweighted by the Huber (c.q. biweight) weights. Since the normalization of all variables was kept fixed, almost the complete sums of squares of these variables was consumed by the points with extreme values; consequently, all the other points had values of approximately zero. The component loadings for such a variable were also

Nonlinear transformations

(almost) zero, so that the whole variable (except for the outlier) had small residuals, but of course the solution was trivial.

In the second situation the optimal scaling also created extreme values, but now these fitted extremely well. In fact they dominated a complete principal component. This situation is also known in classical nonlinear PCA. In fact, considering the problem as a repetition of regression steps, the technique has now created leverage points.

It is clear that both situations lead to rather useless results and that the ordinary optimal scaling is not suitable anymore.

4.2 Bounded monotonic regression

A very direct way to overcome extreme transformations is to set a bound upon the transformations. The monotonic regression is replaced by a bounded form of monotonic regression, this is illustrated in Figure 5.

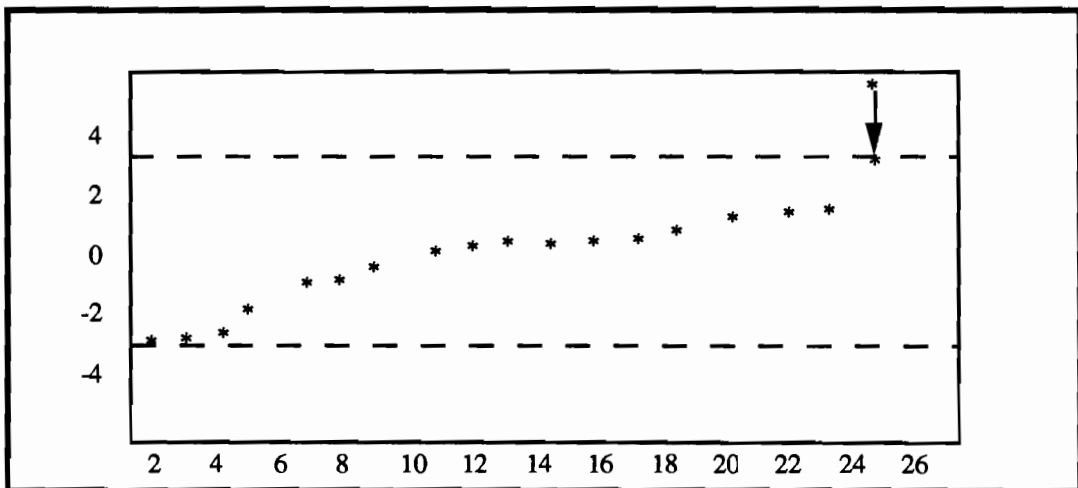


Figure 5. Illustration of bounded monotonic regression.

All points that have an absolute value larger than the bound are simply projected on the bound. So, the updates of the vector \mathbf{q}_j ($j=1,\dots,m$) are computed as: $\mathbf{q}_j^+ = \mathbf{q}_j - \boldsymbol{\delta}$ with the elements of $\boldsymbol{\delta}$ defined as ($i = 1,\dots,n$) :

$$\begin{array}{lll}
 d_i = 0 & \text{if} & |q_{ij}| \leq b \\
 d_i = q_{ij} - b & \text{if} & q_{ij} > b \\
 d_i = q_{ij} + b & \text{if} & q_{ij} < -b
 \end{array}$$

The problem to find updates for the \mathbf{q}_j , that satisfy the constraints, can be summarized in the following way. Let \mathbf{q}_j^0 be the unrestricted update, then the problem is to find a \mathbf{q}_j that minimizes $\|\mathbf{q}_j^0 - \mathbf{q}_j\|^2$ over all \mathbf{q}_j , satisfying the following constraints:

- (i) $\mathbf{q}_j \in \Gamma_j$
- (ii) $\mathbf{q}_j \in \Pi_j$
- (iii) $\mathbf{q}_j \in \Lambda_j$
- (iv) $\mathbf{q}_j \in N_j$

Here Γ_j indicates the convex set of all admissible transformations for variable \mathbf{q}_j ; Π_j is the convex set of all points with zero mean; Λ_j the convex set of all points that satisfy the bounds on the coordinates, and N_j represents the hypersphere with radius equal to the number of non-missing elements in \mathbf{q}_j . The procedure to find the constrained values starts with the unrestricted update, then projects this update on Γ_j . Projections on the other sets do not violate the monotonicity of the variable. Next the variables are alternately projected on Π_j and Λ_j . Since Π_j and Λ_j are convex sets, it can be proved (see Boyle & Dijkstra, 1986), that such an algorithm converges. Finally the projection on N_j remains to be done. This rescaling of the variable may violate the bound restrictions and therefore this step is included in the iterations between Π_j and Λ_j .

Let n_1 be the number of points in \mathbf{q}_j that exceed the bound. These points are projected on the bound and will remain fixed. After this first step there are $(n - n_1)$ free points left, which are normalized to $(n - n_1)b^2$, yielding a total sum of squares of n . If all these points remain within the bounds, we are finished, else the n_2 bound-violating points are fixed to the bound and the remaining $(n - n_1 - n_2)$ points are normalized to $(n - (n_2 + n_1)b^2)$. Such an algorithm stops if no new points violate the bound or if $\sum n_k = n$, where the summation is over the number of iterations. In this case all points are equal to the bound, which immediately gives the lower-bound for the bound value: $b^2 \geq F/n$, where F is a normalization factor. If b does not satisfy this inequality, there is no solution for the problem. This procedure can be written more formally as follows. Let $\|\mathbf{q}^k\|^2 = n$ (subscript j has been omitted) and let the bounded update in the k th iteration be given by $\mathbf{q}^{+k} = \mathbf{q}^k - \delta$, then $\|\mathbf{q}^{+k}\|^2 \leq n$. Next the variable is normalized to n by $\mathbf{q}_i^{k+1} = \alpha \mathbf{q}_i^{+k}$, with $\alpha \geq 1$, for all $q_i \leq b$. So, for each q_i a monotonic sequence is obtained,

$$|q_i^{k+1}| \geq |q_i^{k-1}|,$$

which finishes when all points are stabilized.

Nonlinear transformations

In Figure 6 a geometric representation is given of the algorithm. This presentation is intentionally simplified to communicate the idea.

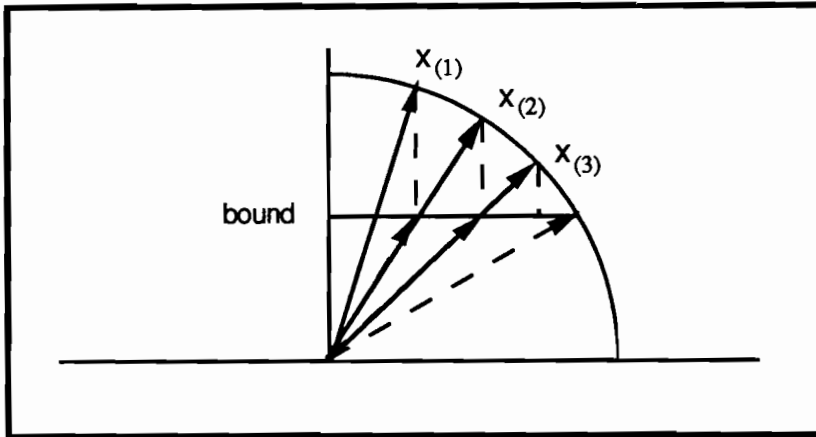


Figure 6. Geometrical illustration of the algorithm for bounded monotonic regression.

The horizontal line in the imaginary circle symbolizes the hypercube of all elements satisfying the bound constraint. The part of the circle shows the normalization constraint. In each step of the algorithm the variable is put in deviation from its mean (which is not shown in Figure 6) and is normalized, that is, the length of the vector is adjusted to the circle. Finally the dashed vector is attained, which is in deviation from its mean and satisfies both bound and normalization constraints. Note that Figure 6 is only meant to show the general idea of the algorithm.

It is easy to show that the bounded update is, given the unrestricted update, the best possible vector satisfying the restrictions. Since best is defined in terms of least squares, it must be true that

$$\| \mathbf{q}_b - \mathbf{q}_m \|^2 < \| \hat{\mathbf{q}}_b - \mathbf{q}_m \|^2, \quad (24)$$

where \mathbf{q}_m is the optimal monotonic update, \mathbf{q}_b the bounded update and $\hat{\mathbf{q}}_b$ any other update satisfying the restrictions. Working out inequality (25) we find:

$$\mathbf{q}_b' \mathbf{q}_m > \hat{\mathbf{q}}_b' \mathbf{q}_m, \quad (25)$$

which means that the angle between \mathbf{q}_m and \mathbf{q}_b must be smaller than any other angle between \mathbf{q}_m and $\hat{\mathbf{q}}_b$. Figure 6 immediately shows that $x(1)$, representing \mathbf{q}_m , has a smaller angle with the dashed vector, \mathbf{q}_b , than with any other vector in the restricted region, that is in the region below the bound, which shows that \mathbf{q}_b is the best possible update given \mathbf{q}_m .

5. The treatment of missing values

There are several options to deal with missing values (Meulman, 1982). First we will consider the case where the missing value is not considered in the minimization problem. Herefore we need an indicator matrix, \mathbf{M} (objects by variables), which indicates the missing by a 0 and the observed values by a 1. The general problem is now defined in the metric \mathbf{M} . It follows that the majorization function to be minimized becomes:

$$\sigma(\mathbf{X}, \mathbf{A}) = \sum_{j=1}^m (\mathbf{q}_j - \mathbf{X}\mathbf{a}_j)' \mathbf{M}_j \mathbf{V}_j (\mathbf{q}_j - \mathbf{X}\mathbf{a}_j). \quad (26)$$

Here \mathbf{M}_j represents a diagonal matrix, where the diagonal of each \mathbf{M}_j is the corresponding column of \mathbf{M} . Obviously, the matrices \mathbf{M}_j are fixed throughout the algorithm. The variables are normalized as $\mathbf{q}_j' \mathbf{M}_j \mathbf{q}_j = \mathbf{u}' \mathbf{M}_j \mathbf{u}$. This implies that the variables are normalized to the number of non-missing observations on that particular variable. This approach is called "missing values deleted".

A different approach is to estimate the missing values, such that they will be optimal with regard to the criterion, and are only restricted with respect to the transformation bounds. For a particular variable each missing can be considered as a separate category (*multiple* approach) or all missing can be considered together as one category (*single* approach). Giving each missing observation a separate category may lead to degenerate solutions, in which the missing categories tend to dominate the solution. However, such an approach can be useful in situations where the interest is in the estimation of optimal values for the missing.

In the next section two examples are presented. In the first, the single approach is used to handle the missing. In the second example, the missing values were deleted from the analysis by using the indicator matrix \mathbf{M} .

Applications

6. Applications

In this section we will show some results of resistant PCA for two datasets. In particular we will look if there is any gain in interpretability of the resistant procedure compared to the ordinary least squares analysis. Furthermore we inspect the differences between linear and nonlinear approaches.

6.1 Mental Disorders and Antidepressants

This example is about persons with mental disorders symptoms and the use of antidepressants. The data were collected by a group of psychologists and psychiatrists, who recorded a number of variables for 36 persons with mental disorder symptoms. A description of the ten variables that were used, is given in Table 1.

Table 1. Description of variables for MDA data

variable	description	
1. diagnosis	type of disorder:	1: anxiety disorder 2: depression 3: personality disorder
2. seriousness	7-points scale	1: not serious --- 7: very serious
3. frequency		1: daily symptoms 2: weekly symptoms 3: monthly symptoms
4. duration		1: more than two years 2: more than half a year 3: less than half a year
5. sex		1: female 2: male
6. age	sort therapy	1: without antidepressants
7. method		2: with antidepressants
8. medication	type of drug	1: anafranil 2: fevarin 3: tofranil 4: tolvon 5: tryptizol 7: no drug
9. dose	6-points scale therapy result	1: small dose --- 6: high dose (7: no dose)
10. effect		1: decrease symptoms 2: increase symptoms 3: something else

The complete data set is given in the appendix. The objective can be very generally stated as searching for relations between the variables and exploring the structure in the group of objects. The variables have mixed measurement levels: there are five ordinal and five nominal variables.

The data set contained a few missing values. These were treated as nominal categories. In fact, when there were more than one missing values for one variable, these were considered as one single category. As already mentioned before, the missing values are just like any other value restricted to the bounds, which means that they cannot have an unduly large influence upon the solution.

The data were analyzed by using the three different loss functions that have been mentioned before: ordinary least squares, the Huber and the biweight.

The first principal component clearly shows a distinction between persons that were treated with antidepressants and persons that obtained a different type of therapy in which no drugs were involved. The variables *method*, *medication* and *dose* are highly correlated and have high loadings on this dimension.

The projection of the objects onto the first two principal components (object scores) is shown in Figure 7, together with the component loadings in the first two dimensions.

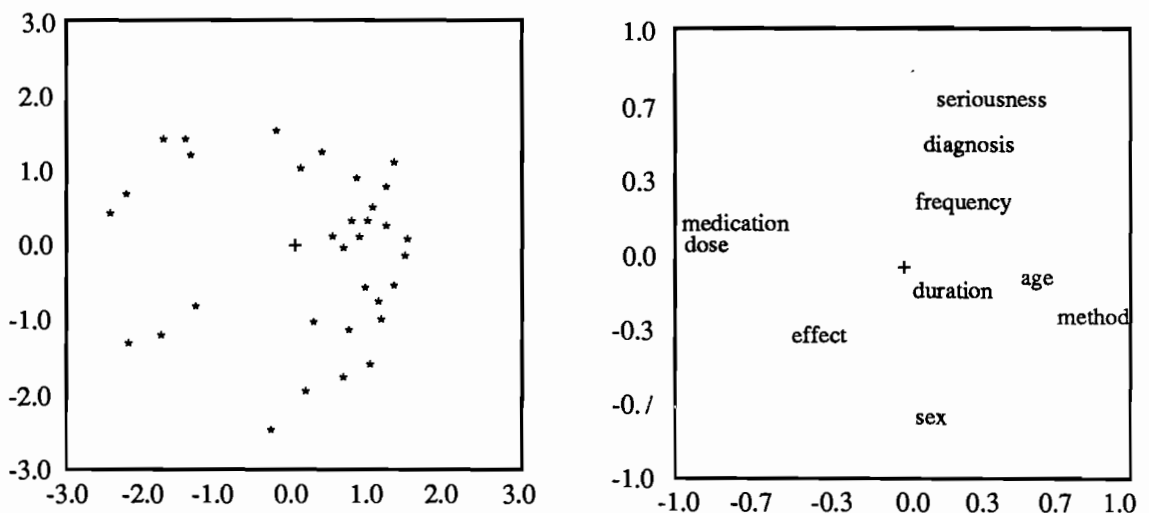


Figure 7. Object scores and component loadings for first and second dimension: least squares (MDA).

It appears that this distinction is the major source of variation in the data, which is found by all three loss functions. Let us therefore examine a plot of the second against the third dimension to see if this shows interesting differences between least squares and the resistant functions.

Applications

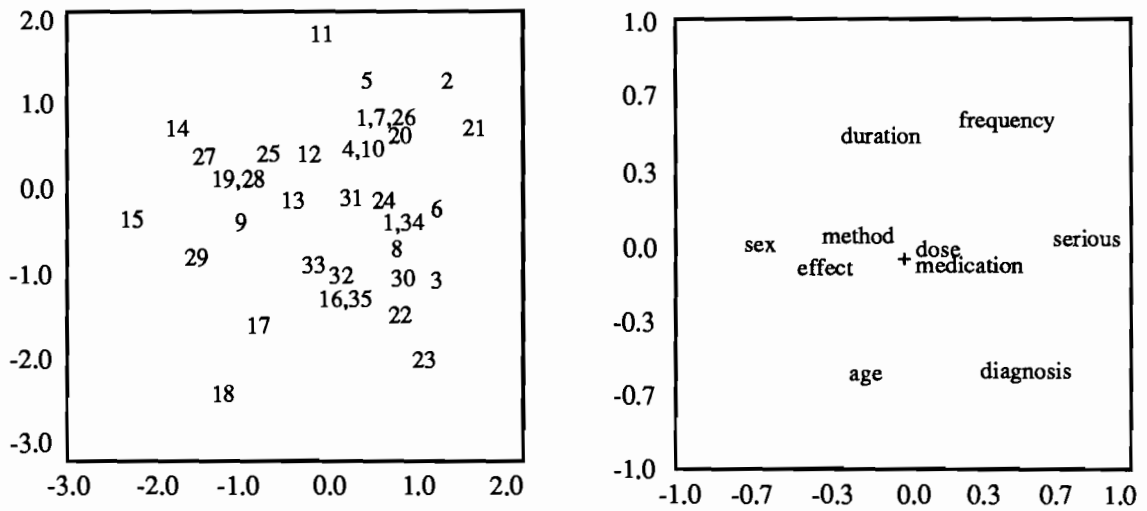


Figure 8. Object scores and component loadings for second and third dimension: least squares (MDA).

In Figure 8 the second and third dimensions are plotted for the least squares solution. There is a clear difference between the Huber solution (Figure 9). In the Huber analysis a tuning constant (TC) of 1.0 was chosen. To see whether the differences can be explained, we will have a closer look at some of the extreme points.

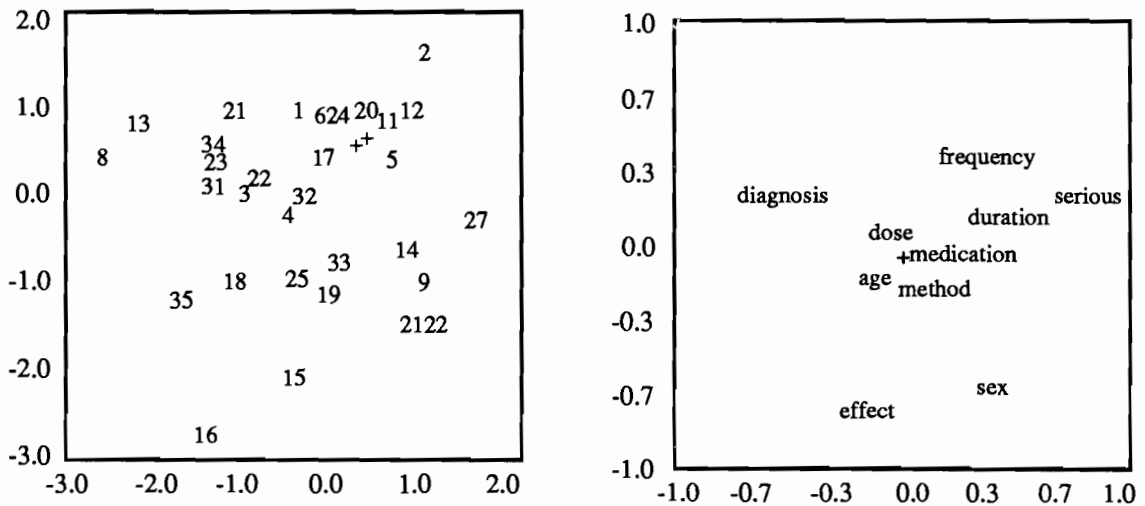


Figure 9. Object scores and component loadings for second and third dimension: Huber, TC=1.0 (MDA).

It appears that objects 8 and 13 are the only patients diagnosed with a personality disorder. Now if we inspect the category quantifications for this variable (see Figure 10) then we notice the relatively extreme quantification for the category *personality disorder* in case of the Huber function. Since the variable *diagnosis* has a rather large component loading for this dimension, it is clear that this explains at least partly the difference.

Object 16 has a missing value on the *effect* variable. The quantification of this missing value is in the Huber case much larger (namely equal to the boundary value) than with least squares. Object 15 is the most left lying point in Figure 8. So this point is somewhat outlying in both analyses. However, all weights for this object are equal to 1.0, so this point can be fitted quite well into the overall structure of the data.

There are two objects (17,18) with *daily* symptoms. This category is in both analyses quantified with the boundary value. However, in the Huber analysis these values are downweighted (with weights equal to .32 and .41), so their influence upon the solution is much less. The unweighted loss for these objects in the Huber analysis is 15.2 and 9.3, which becomes 7.7 and 4.8 respectively for the weighted loss. The least squares solution resulted in loss for these persons of 11.3 and 3.3. This illustrates the fact that the Huber criterion tolerates larger residuals than least squares in order to fit other points better. The dominating effect of these points in Figure 8 is clear.

Furthermore the duration variable has lost some influence, because it contained a missing value (for object 11, which also moved from the edge) that was down weighted by the Huber function.

In this Huber analysis the tuning constant was chosen equal to 1.0, and the function was applied to the (n by m) residual elements. If we choose the tuning constant larger, the results approach the least squares result. In this example $TC = 2.5$ gives identical results to least squares, because all residuals appear to be smaller than 2.5, which makes all weights equal to one.

Using the aggregated residuals and applying the Huber functions on these values yields in this example approximately the same results. Of course to obtain comparable results with the elementwise approach the tuning constant should be chosen larger, because the residuals are larger too. If we take $TC = 2.5$ in this approach, one object (17) is downweighted (.68). So, this analysis would identify object 17 as an outlier. The object scores and component loadings are however, almost identical to the least squares solution.

The results of the biweight function are very much like the Huber function. Although the interpretation of the results from the biweight is the same as that of the Huber analysis, the explained proportion weighted sums of squares from the biweight is somewhat higher.

Applications

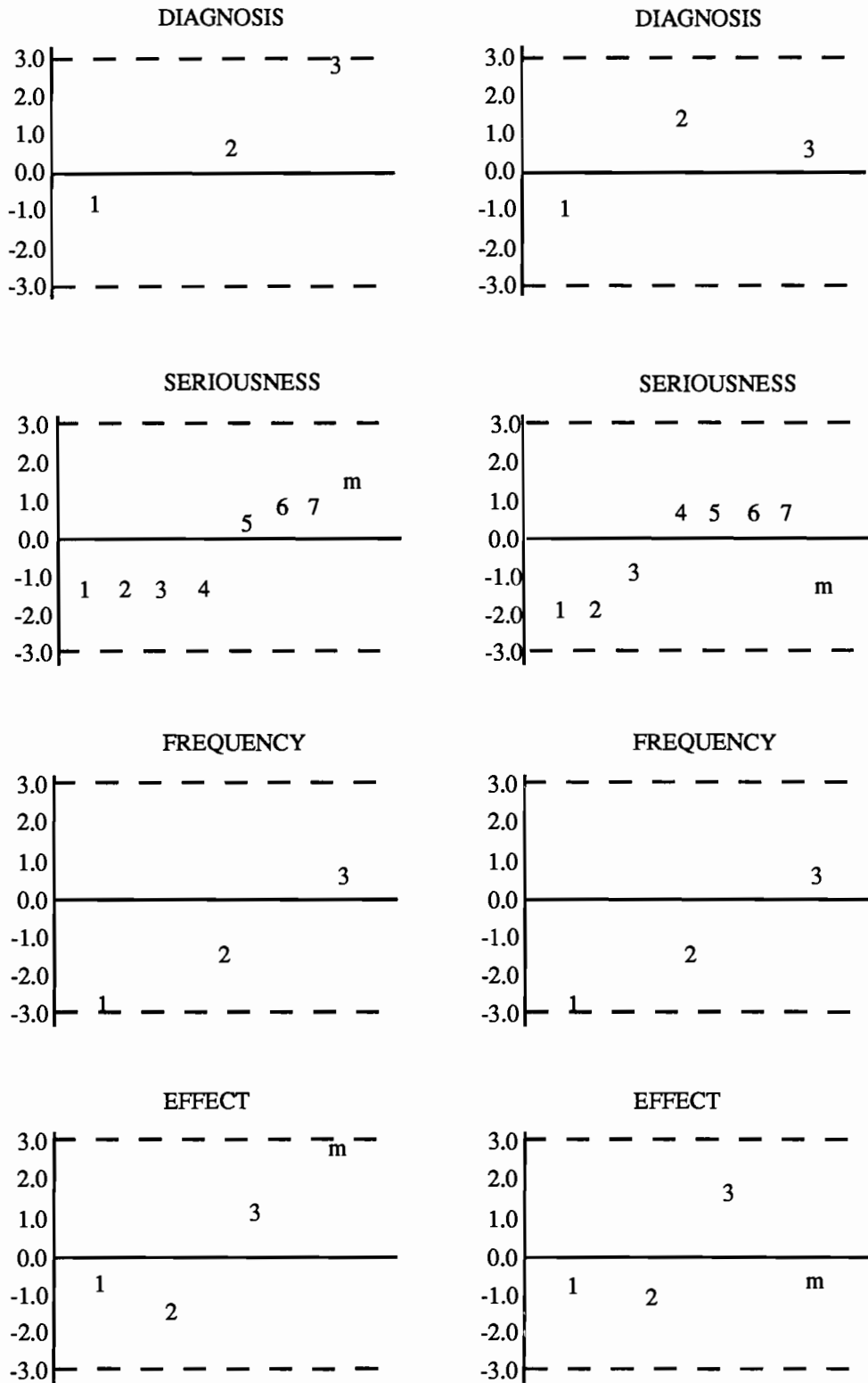


Figure 10. Optimal quantifications for four variables: Huber (left) and least squares (right).

6.2 Aesthetic Judgments

For the second example data will be used which have been presented by Davenport and Studdert-Kennedy (1972). The data are scores given by an eighteenth century art-critic, named de Piles, for 56 seventeenth century painters on four painting characteristics: *composition*, *drawing*, *colouring* and *expression*. The scores are on a scale between 0 and 20, with the latter score reserved for the "sovereign perfection, which no man has fully arrived at". The list of painters is given in Table 2.

Table 2. Names of painters in aesthetic judgment data.

1. Albani	20. Durer	39. Polidore da Caravaggio
2. Durer	21. J. Jordaens	40. Pordenone
3. Del Sarto	22. L. Jordaens	41. Pourbus
4. Barocci	23. Josepin	42. Poussin
5. Bassano	24. Giulio Romano	43. Primaticcio
6. Del Piombo	25. Lanfranco	44. Raphael
7. Bellini	26. Da Vinci	45. Rembrandt
8. Bourdon	27. Van Leyden	46. Rubens
9. Le Brun	28. Michelangelo	47. Salviata
10. Veronese	29. Caravaggio	48. Le Sueur
11. The Carracci	30. Murillo	49. Teniers
12. Corregio	31. Otho Venius	50. Testa
13. Volterra	32. Palma Vecchio	51. Tintoretto
14. Diepenbeck	33. Palma Giovane	52. Titian
15. Domenichino	34. Parmigiano	53. Van Dyck
16. Giorgione	35. Penni	54. Vanius
17. Guercino	36. Perino del Vaga	55. T. Zuccaro
18. Guido Reni	37. Cortona	56. F. Zuccaro
19. Holbein	38. Perugino	

In Davenport and Studdert-Kennedy (1972) - henceforth denoted as D&S - results of a linear PCA on the covariance matrix are reported. They justify the use of interval measurement scales of the variables by arguing that de Piles could have simply ranked his painters on each dimension if he would not have intended his scales to be interval scales.

Our analysis is different because we apply monotonic transformations of the standardized variables. Furthermore we will incorporate the two painters with one missing value. In D&S these painters were completely deleted from the analysis. Here we use the option *missing values deleted*, which means that we are using the indicator matrix \mathbf{M} , to delete the missing observations from the analysis. And of course we will use another criterion than least squares to find the principal components and potential aberrant observations.

Applications

First we will consider the numerical solution. In Table 3 the component loadings are given for the Huber solution together with the least squares solution. For the Huber criterion a tuning constant of 1.0 was used.

Table 3. Component loadings in two dimensions, aesthetic judgments (numerical).

variables	Huber		Least squares	
Composition	.72	.53	.71	.54
Drawing	.82	-.27	.83	-.28
Colour	-.38	.90	-.39	.88
Expression	.88	.17	.87	.19
Relative fit	.54	.29	.53	.29

These numbers show the same pattern as in the D&S study. The first component shows the general reaction of de Piles to the painters, with the striking feature of a negative loading for colour. D&S explain this as: "*..the importance accorded to colour by a critic of the period is likely to have been low relative to the other three dimensions. It seems to us then that the negative element in the first latent vector reflects what de Piles saw as the rarity of mastery by one painter both of colour and of the other qualities*".

The second dimension is dominated by colour. The components for the Huber solution explain a proportion of variance of 54 and 29 percent, respectively. This is slightly less than in D&S, which yields 56 and 28 percent, but almost the same as our least squares result.

The projections of the painters in the two dimensional space of the first and second component are shown in Figure 11. This plot resembles the one given by D&S, which was obtained by means of a cluster analysis. Basically we could recognize in the plot the same characteristics as in D&S. For instance, the extreme positions of Rubens (46), Raphael (44) and Penni (35) and some clusters of painters are the same. In this analysis five painters were assigned weights smaller than 1. These weights were all about .75, which implies that these painters are considered to be only mildly outlying.

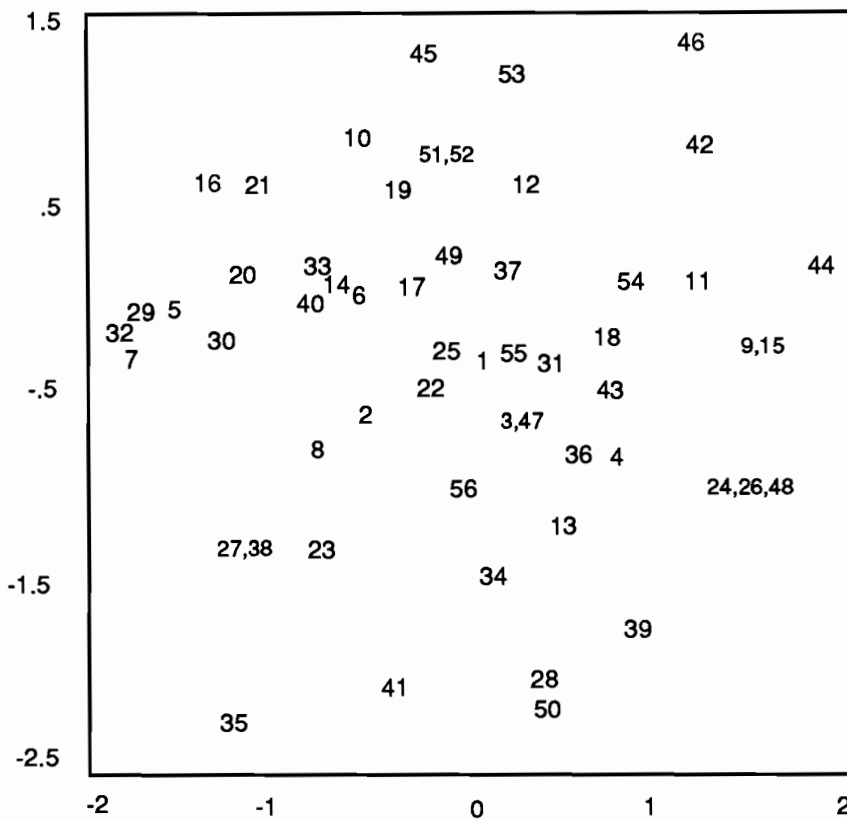


Figure 11. Object scores first and second dimension, aesthetic judgements (Huber, numerical).

The component loadings for the nonlinear (ordinal) solution are given in Table 4. obviously these values are somewhat larger than those from the numerical analysis.

Table 4. Component loadings in two dimensions, aesthetic judgment (ordinal).

variables	Huber		Least squares	
Composition	.74	.50	.75	.54
Drawing	.86	-.23	.86	-.26
Colour	-.54	.80	-.56	.78
Expression	.91	.27	.91	.26
Relative fit	.62	.26	.61	.26

In the nonlinear solution there are only two elements in the datamatrix downweighted, but now these weights are smaller than in the numerical solution, that is .48 and .64. These

Applications

weights refer to the composition of the painter Guarcino (17) and to the drawing of Van Leyden (27). These points are considered as outliers of the second type. The high score for Guarcino's composition does indeed not conform to the correlation structure in the data, since his expression and drawing is scored quite low. Also Van Leyden's score for drawing is somewhat lower than what might be expected from his other scores.

Comparing the linear and nonlinear solution with respect to the object scores, we can see that the position of Guarcino remains in the centre, but Van Leyden moves towards the centre in the nonlinear analysis. Furthermore, Figure 12 shows that in the nonlinear analysis the outer points are somewhat more accentuated.

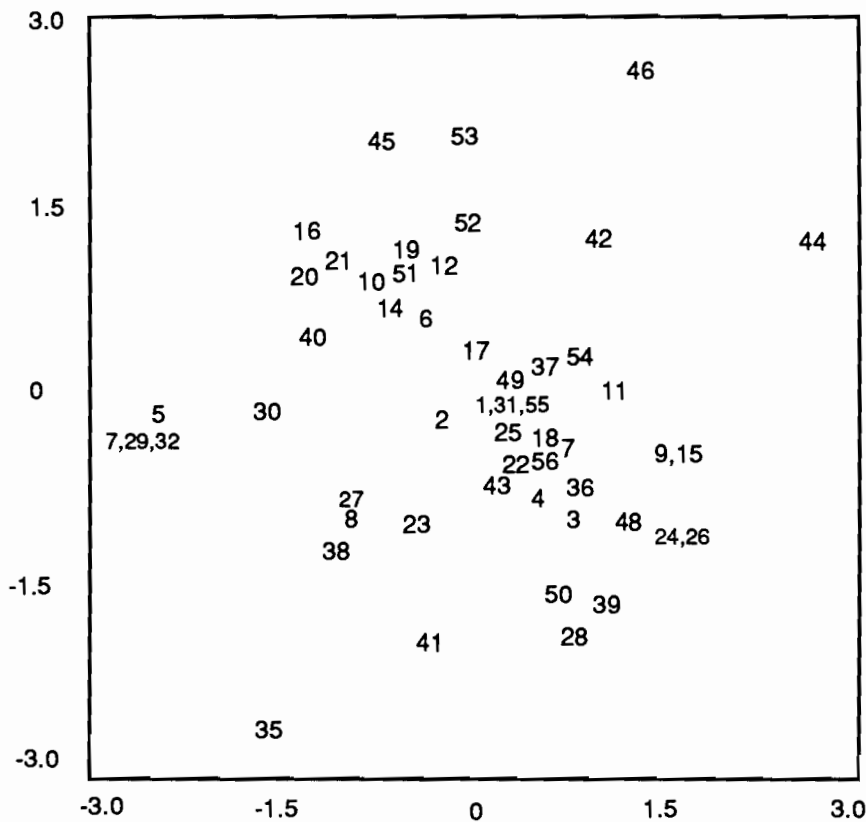


Figure 12. Object scores first and second dimension, aesthetic judgements (ordinal).

If Guarcino and Van Leyden are really type two outliers, it is possible they show up as outliers in the last principal components. To investigate this we did an analysis in four dimensions. Figure 13 shows the object scores in the third and fourth dimension. It is clear that there are some extreme points (which we have labeled), among which are Guarcino (17) and Van Leyden (27).

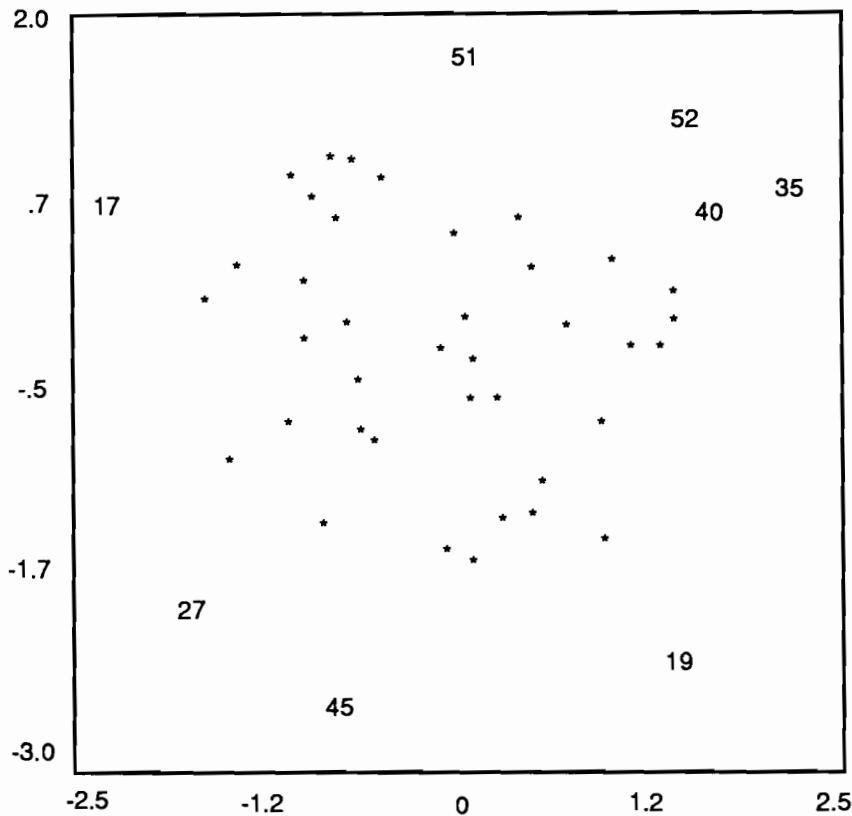


Figure 13. Object scores third and fourth dimension, aesthetic judgements (ordinal).

The figure suggests that there are more type two outliers. To see if these could be found by one of our resistant techniques, an analysis was done by using the biweight function. In Table 5 the component loadings are given for the biweight criterion with a tuning constant of 2.0.

Table 5. Component loadings in two dimensions, aesthetic judgment (biweight).

variables	numerical		ordinal	
Composition	.76	.47	.73	.28
Drawing	.89	-.14	.89	-.10
Colour	-.35	.98	-.43	.90
Expression	.90	.09	.96	.12
Relative fit	.59	.30	.69	.27

Applications

The general pattern is the same as in the previous analyses, although the fit is somewhat higher here. In this analysis five entries in the data matrix are weighted with 0.0, a few are weighted between 0.0 and 1.0 and the rest is approximately weighted by 1.0. The painters for which a variable was downweighted by zero, are: Holbein (19, expression), Penni (35, drawing), Rembrandt (45, drawing) and also of course Guarcino and Van Leyden. Compared to the overall correlation structure, the scores for Holbein and Penni on expression and drawing respectively, were too high. Rembrandt's score for drawing was too low.

In Figure 14 we can see the effect of this weighting process. Rembrandt moves somewhat to the right, towards the excellent painters, which is of course congruent with the present opinion on him. The painters Penni and Holbein move somewhat to the left, in fact Penni can now be considered as the worst painter. So, if we are willing to assume that the Piles overestimated, for instance, the drawing qualities of Penni and underestimated these qualities for Rembrandt, then the picture shown in Figure 14 is a better representation of the structure of the set of painters the other pictures.

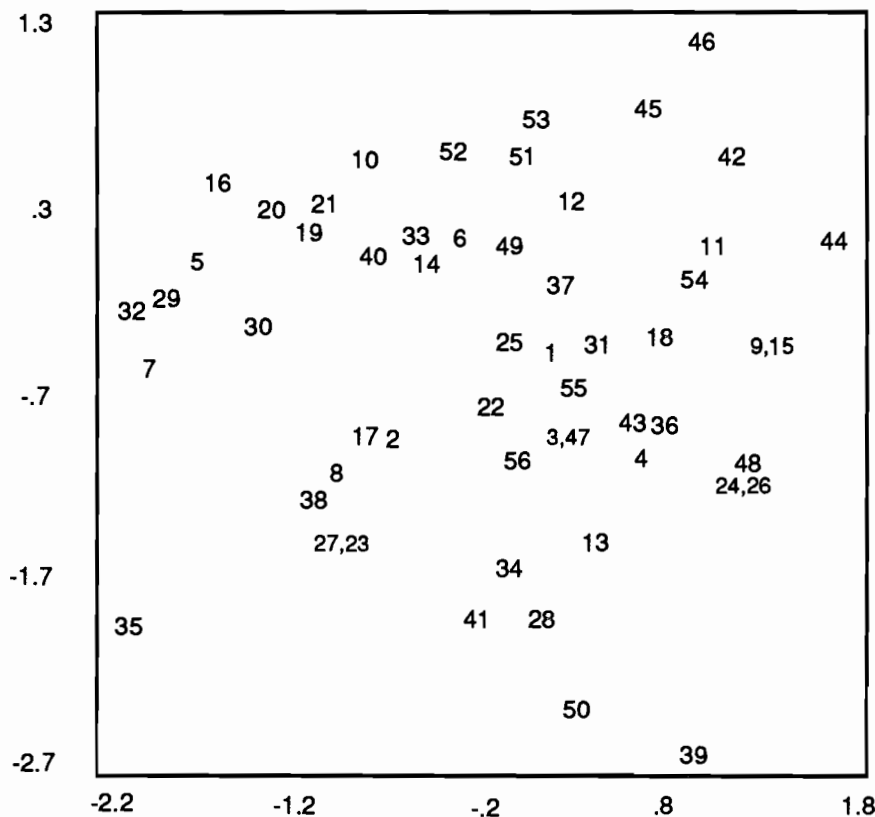


Figure 14. Object scores first and second dimension, aesthetic judgements (biweight, numerical).

For the ordinal solution the results are more extreme. There are another two painters with zero weights for expression, because their scores were quantified with -2.58, the boundary value. So here we have the situation that, despite the bounds, type one outliers are created and then downweighted. This leads to different objectscores (Figure 15).

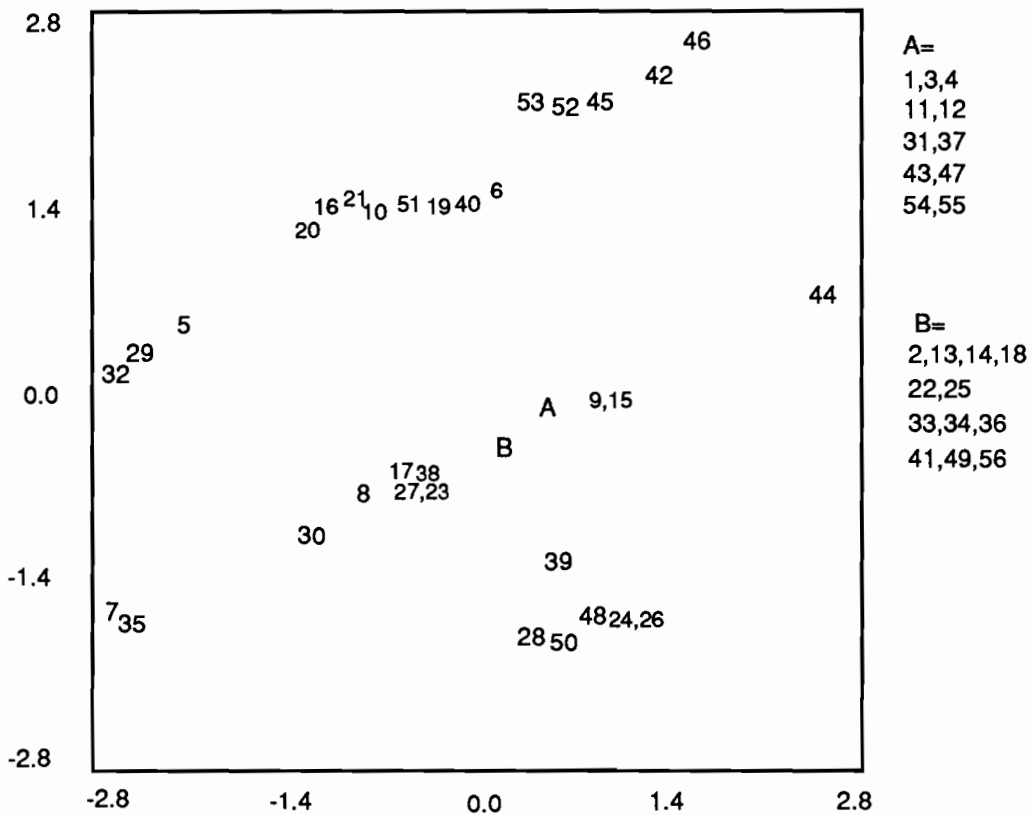


Figure 15. Object scores first and second dimension, aesthetic judgements (biweight, ordinal).

In fact we have found some more or less homogeneous groups. The painter Raphael (44) has an even extreme score on the first dimension than before, this is mainly due to its extreme quantification for expression. The optimal transformations are shown in Figure 16.

Applications

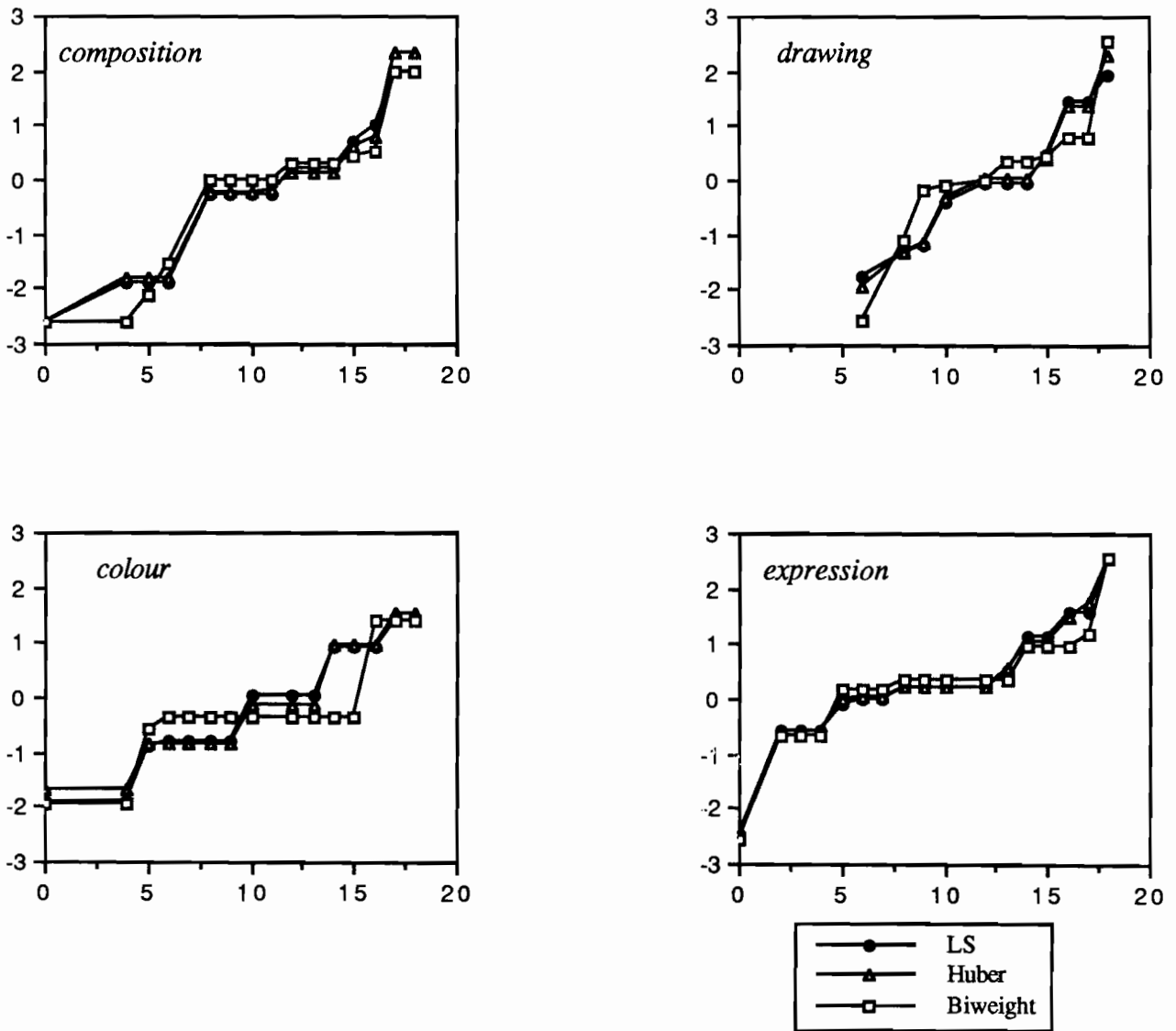


Figure 16. Optimal transformations for Least squares, Huber and biweight (aesthetic judgements).

The monotonic transformations of the original scores are very much the same for the least squares as for the Huber. The biweight has a slightly different pattern, especially for the variable colour. This explains why the plot for the ordinal biweight solution (Figure 15) is quite different from the other plots.

Choosing the tuning constant in the biweight function smaller than 2.0 causes too many points to be down weighted with zero. Of course a larger tuning constant will cause the analysis to resemble least squares.

7. Discussion

It seems that resistant alternatives for nonlinear PCA have something to offer. It is necessary to restrict the optimal scaling in order to avoid degenerate solutions. To set a bound upon the transformations works quite well although there is some arbitrariness in the choice of the bound. In the second example the boundary value was chosen to be 2.58, which of course appeals to the Gaussian distribution. In fact, if we assume that a variable is normally distributed, then only one percent of the points will exceed this boundary value. So, the error probability we have by cutting these extremes off is only one percent. Although it may sound somewhat artificial to introduce probability theory all of a sudden, it doesn't harm as a rough guideline. However, it is clear that more practical research is necessary on this subject.

The method of bounded monotonic regression can also be seen as a trimming procedure, that is, the extremes of the distribution are chopped off. Using such a trimming procedure fits neatly into the framework of robust/resistant methods.

Thus far the choice of the tuning constant seems to be no real problem. The data are standardized, so the residuals are also in a particular range. Tuning constants in the neighbourhood of 1 (for Huber) and 2 (for the biweight) will usually be a good choice. The exact choice depends on how many points one is willing to down weight; in other words, how many outliers are expected to be present. This is merely a methodological problem to be solved by the data analyst. However, it does imply that the data have to be analyzed several times, each time with a different tuning constant.

8. References

- Boyle, J.P. and Dykstra, R.L. (1986). A method for finding projections onto the intersection of convex sets in Hilbert spaces. In: R. Dykstra, T. Robertson and F.T. Wright (Eds.), *Advances on order restricted statistical inference*. Proceedings of the symposium on order restricted statistical inference, Iowa City: Springer Verlag.
- Campbell, N.A. (1980). Robust regression in multivariate analysis. I: Robust Covariance Estimation. *Applied Statistics*, 29, 3, 231-237.
- Davenport, M. and Studdert-Kennedy, G. (1972). The statistical analysis of aesthetic judgment: an exploration. *Applied Statistics*, 21, 324-333.
- Devlin, S.J., Gnanadesikan, R., Kettenring, J.R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62, 3, 531-545.
- Devlin, S.J., Gnanadesikan, R., Kettenring, J.R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76, 374, 354-362.
- Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1, 211-218.
- Gabriel, K.R. and Odoroff, L. (1984). Resistant lower rank approximation of matrices. In: E. Diday et al. (Ed.), *Data Analysis and Statistics III* (pp. 23-30). Amsterdam: North-Holland.
- Gabriel, K.R. and Zamir, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, 21, 4, 489-498.
- Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.
- Gnanadesikan, R. and Kettenring, J.R. (1972). Robust estimates, residuals and outlier detection in discriminant analysis. *Biometrics*, 28, 81-124.
- Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York: Wiley.
- Hawkins, D.M. and Fatti, L.P. (1984). Exploring multivariate data using the minor principal components. *The Statistician*, 33, 325-338
- Li, G. and Chen, Z. (1985). Projection pursuit approach to robust dispersion matrices and principal components: primary theory and Monte carlo. *Journal of the American Statistical Association*, 80, 391, 759-766.
- Heiser, W.J. (1987). Correspondence analysis with least absolute residuals. *Computational Statistics and Data Analysis*, 5, 337-356.

- Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.
- Huber, P.J. (1985). Projection Pursuit. *The Annals of Statistics*, 13, 2, 435-475.
- Meulman, J.J. (1982). *Homogeneity analysis of incomplete data*. Leiden: DSWO-press.
- Mosteller, F. and Tukey, J.W. (1977). *Data analysis and regression*. Massachusetts: Addison-Wesley.
- Rouseeuw, P.J. and Van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 114, 633-651.
- Rouseeuw, P.J., Edegem, E. and Van Zomeren, B.C. (1990). Some proposals for fst HBD regression. *Compstat Proceedings*, 1990.
- Verboon, P. (1990). *Majorization with iteratively reweighted least squares: a general approach to optimize a class of resistant loss functions*. Research Report 90-07. Leiden: Department of Data Theory.
- Verboon, P. and Heiser, W.J. (1989). *Some robust loss functions for the orthogonal Procrustes problem*. Research Report 89-03. Leiden: Department of Data Theory.
- Verboon, P. and Heiser, W.J. (1991). Resistant orthogonal Procrustes analysis. *Journal of Classification*. (in press).
- Verboon, P., Van der Lans, I. and Heiser, W. J. (1991). *The multipals algorithm*. Research Report 91-04. Leiden: Department of Data Theory.
- Young, F.W., Takane, Y. and 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.

Appendix

Appendix

Data on Mental Disorders and the use of Antidepressants

1	18331	36	2141
2	18331	24	1771
3	28311	39	1773
4	14321	19	1773
5	16331	25	1773
6	16321	41	1771
7	17331	32	2121
8	35321	60	2431
9	18232	42	2161
10	16331	70	2121
11	16391	35	2551
12	19331	68	2541
13	32331	45	2261
14	13332	26	2141
15	12232	38	2123
16	15211	55	2124
17	13131	46	1771
18	13121	54	1773
19	13312	25	1773
20	18331	39	1151
21	25331	24	1141
22	28211	31	2111
23	26211	67	1141
24	18311	44	2342
25	14331	43	2153
26	16331	41	2221
27	19332	42	2151
28	18332	65	2133
29	19322	47	2263
30	25311	52	2541
31	15311	42	2361
32	26312	50	2362
33	26232	37	2351
34	25331	58	2141
35	25231	42	2123
36	27331	55	2141

Note: the first two columns contain identification numbers