INFLUENCE AND STABILITY IN NONLINEAR PRINCIPAL COMPONENTS ANALYSIS

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).

INFLUENCE AND STABILITY IN NONLINEAR PCA

Abstract

The jackknife is used to study the effects of individual observations on the outcomes of nonlinear Principal Component Analyses (PCA). The main point of interest is whether influential points could be detected and how they affected, in particular, the stability of the object scores and the transformations of the categories. To perform the nonlinear PCA, three different loss functions were used: least squares, the Huber function and the biweight function; the latter two are expected to be resistant to outliers. A second point of interest was whether these functions yielded more stable results than least squares.

It was found that for the object scores the resistant functions yielded more stable results than least squares and that influential points were less likely to occur. The results for the category transformations were rather divergent.

Key words: nonlinear PCA, influence, stability, jackknife, resistance, outliers

Introduction

In applying Principal Component Analysis (PCA), it is of some interest to know that each individual observation in the data may have a different influence upon the solution. Because PCA is usually applied in the context of data exploration, the concept of influence is a very important one, like it is for many other multivariate techniques.

Observations, which will be called objects hereafter, are said to be influential, if the PCA solution changes considerably, when one of such objects is taken away from the data. Alternatively, if the solution remains approximately unchanged, the object is not influential. There is a strong relation between influential objects and outliers. For an object to be influential, it should be somewhere on the edge of the cloud of points in the high-dimensional space. If such an object is in line with the overall correlation structure in the data set, it can be an outlier of the first type, sometimes called an *inlier*, provided that its position is extreme compared to the other objects. If it is not congruent with the correlation structure of the data, it will be a type two outlier. Outliers with large influence can determine by itself an entire principal component, which could even have the largest eigenvalue.

In Verboon (1991) two resistant procedures are introduced, which can be used to perform a nonlinear PCA, in which outliers will have less influence upon the solution than in the ordinary least squares case. These procedures are based on replacing the least squares loss function by the Huber and biweight loss functions, which are known to be more resistant to outliers. Therefore, these functions are expected to decrease the influence of influential outliers. However, not much is known about influential inliers: objects that are influential, but fit into the overall correlation structure of the data.

In this paper we will systematically study the problem of influence and stability in nonlinear resistant PCA by conducting several jackknife experiments. A randomly selected dataset is used, which is thought to be representative for most empirical datasets. The main goal is to find out whether such studies can be used to detect influential points and outliers, and how these points affect the nonlinear PCA solution. Since such a nonlinear solution consists of many results, such as object scores, quantifications, eigenvalues and component loadings, which are quite difficult to compare directly, we have defined several measures to summarize the results. The derived measures are more suitable for comparison between all the different analyses. Another goal of this paper is to study the stability of the output measures under jackknifing. In particular, the interest is in the differences between the three techniques, or more specifically, do the resistant techniques yield more stable results than LS.

Influence Functions

Before giving a detailed description on how the experiment is set up, a short overview will be given about influence functions. The (theoretical) influence function, defined by Hampel (1974), is a useful device for studying the effect of observations on parameter estimates. They have been widely used, for instance, for designing robust estimators, robust tests and confidence intervals and also for developing methods for outlier detection (Devlin et al., 1975). Influence functions for a particular parameter are found by perturbing the (assumed) distribution function F by adding a small contribution from a unit mass at the point x, evaluating the parameter at the perturbed distribution function, and subtracting the parameter found at the unperturbed distribution function. In other words, the influence function gives the standardized effect of an "infinitesimal" contamination at the point x on the parameter estimate. An extensive discussion on influence functions can be found in Hampel et al. (1986). Furthermore, in Critchley (1985) the concept of influence is discussed in relation to PCA.

In the present context of exploratory data analysis, we are specifically interested in the so-called *finite sample* versions of the influence function. Three different types of these finite sample versions are usually distinguished. The first is the empirical influence function (EIF), named so by Mallows (1975), which is obtained by replacing, in the definition of the influence function, the distribution function F by the empirical distribution function F_n. The EIF can actually be defined in two different ways. In the first, an observation x is *added* to the sample and the required parameter can now be computed and plotted as a function of x. Alternatively, one could *replace* the ith observation from the sample by an arbitrary x and again plot the parameter as a function of x. The latter is sometimes called the *deleted* EIF.

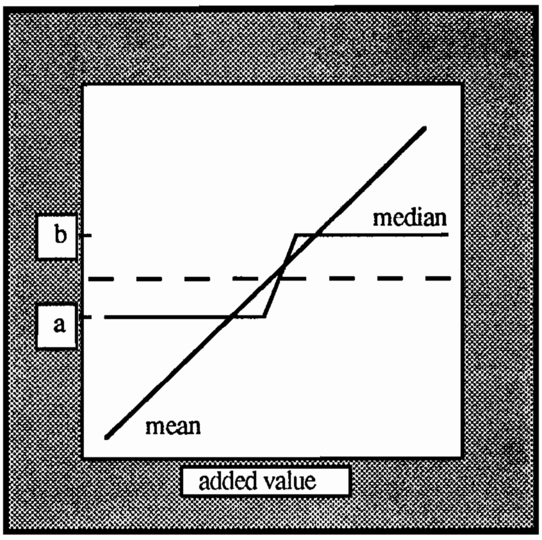

Figure 1. Sensitivity curve for the mean and the median.

The second type of finite sample versions is the *sensitivity curve* (Andrews et al., 1972), which is also defined with addition and replacement. Basically, the aim of the sensitivity curve is to show the difference between the parameter value computed for the whole sample and the parameter value computed for the sample with an additional point (or with one point replaced). In fact, the sensitivity curve is nothing more than a translated and rescaled version of the EIF. In Figure 1 an illustration is given of a sensitivity curve. In the figure two estimators of location, the mean and the median, are depicted as a function of an added value. The message of Figure 1 is that the mean is not resistant, since its value can become arbitrarily large or small by adding only one data point. The median on the contrary is resistant for its value is *bounded.* No observation can be found which brings the value of the median below *a* or above *b*.

Finally, we have the *sample influence function* (Devlin et al., 1975), arising from the jackknife method (Quenouille, 1956; Tukey, 1958; Miller, 1974). The sample influence function is obtained by plotting the *n* so-called *jackknifed pseudo-values*, defined as

$$T_{ni}^{*} = n\, T_n - (n-1)T_{n-1(i)}, \qquad (1)$$

in which the T_n represents the parameter estimate based on the full sample of n observations and $T_{n-1(i)}$ is the estimate based on the same sample with the ith observation deleted. The basic difference with the previous functions is that we are now only using the observed data points, thus not adding any external point. In this paper we construct plots that are related to the sample influence function and will be used to study stability and influence in nonlinear PCA.

Jackknifing in nonlinear PCA

Consider a datamatrix $\mathbf{Z} = \{z_{ij}\}$ of order nxm, representing the scores for n objects on m variables, for which a PCA solution must be found in p dimensions.

To study the effects of influential points and outliers in nonlinear PCA, we will use the jackknife method (Efron, 1979). In a jackknife study we examine the distribution of the required parameter estimates after repeatedly analyzing parts of the data, such as described in the previous section. In this paper, we will always leave out *one* point at the time. With one point at the time left out, n times an analysis can be performed on the remaining n-1 points, in which case each point has exactly been left out once. This yields n values $T_{n-1(i)}$, which will be plotted throughout this paper, instead of the jackknifed pseudo-values. The jackknife is a suitable method not only to examine stability of the parameters, but also to find those points that cause a possible instability.

In the present study the jackknife will be applied for three criteria, least squares (LS), the Huber loss function (HUB) and the biweight loss function (BIW). For each of these loss functions n analyses are performed, each with another row of the data matrix (object) left out. This results in n matrices for the object scores, order $(n\text{-}1, p)$, n matrices for the component loadings, order (m,p) and n matrices with category transformations, order (m,q), where q denotes the maximum number of categories over all variables. Finally, for HUB and BIW, n matrices of loss weights are obtained, order (n,m) or (n) respectively, depending on which type of weighting is used.

There are two types of weighting, which are discussed in Verboon (1991). The first type uses weights for each row of the datamatrix, hence, the rowwise approach; in the second type weights are assigned to each element in the datamatrix, which is called the elementwise approach.

Object scores

In order to study the influence of each object i (i =1,..,n) on the p-dimensional space of the object scores, the following measure has been defined:

$$\mathrm{INFL}(i) = \frac{\| \mathbf{C}_{(i)} - \mathbf{X}_{(i)} \|^2}{(n-1)p}, \tag{2}$$

where $\mathbf{X}_{(i)}$ (n,p) represents the matrix of object scores, obtained from the analysis with the ith object left out. The ith row of $\mathbf{X}_{(i)}$ is filled zero's. The matrix $\mathbf{C}_{(i)}$ is equal to $\mathbf{C}$, the centroid, with the ith row replaced with zero's. The centroid is defined as

$$\mathbf{C} = \Sigma_i \, \mathbf{X}_{(i)} \, /n \text{ -1}. \tag{3}$$

The whole term in (2) is standardized to the sum of squares of $\mathbf{X}_{(i)}$, which is $(n$-$1)p$. The lower bound for INFL is zero; there is no upper bound.

If an object is very influential with respect to the object scores, the deletion of this object from the data, is expected to cause a relatively large change in the objects scores compared to the object scores obtained from all other analyses. It follows that the object scores will have larger distances from the centroid than the object scores that are obtained when one of the non-influential objects is left out.

The stability of an object k (k=1,..,n) in the object space over all n analyses, can be computed through the following measure:

$$\mathrm{STAB}(k) = 1 - \frac{\Sigma_i \| \mathbf{x}_{(i)k} - \mathbf{c}_{(i)k} \|^2}{(n-1)p}, \tag{4}$$

where $\mathbf{x}_{(i)k}$ and $\mathbf{z}_{(i)k}$ are the kth rows of $\mathbf{X}_{(i)}$ and $\mathbf{c}_{(i)}$ respectively. The formula indicates that an object is stable, if the dispersion around its centroid, over all jackknife runs, is small. The upper bound for STAB is one; there is no lower bound. In Figure 2, the STAB and INFL measures are illustrated in a small example.

In the Figure the four jackknifed object scores of four objects are shown, together with their centroids. Here STAB of a point k is given by the sum of squares of the lengths of the lines around the kth centroid (up to some rescaling), while INFL of a point i is given by the sum of the squared lengths (up to some rescaling) of the lines, which are labeled with the same number.

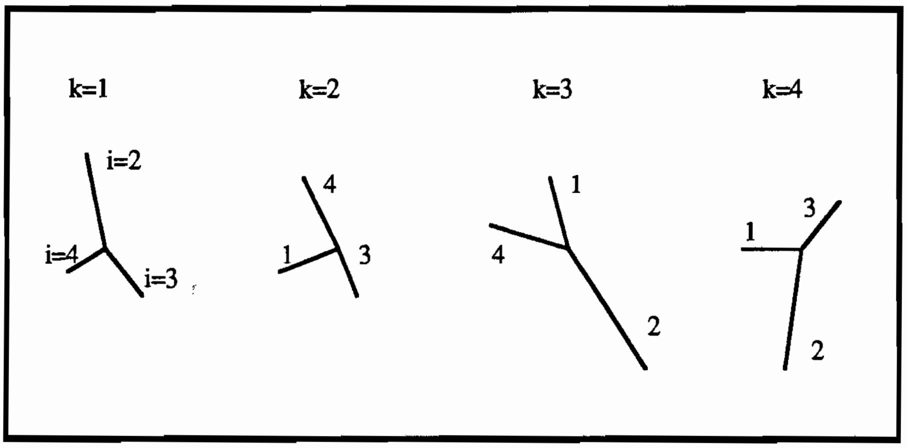

Figure 2. Illustration of INFL and STAB measures.

The STAB measure is quite similar to the one defined in De Leeuw and Meulman (1986), who used it to study the stability in multidimensional scaling solutions. However, an important difference is that we don't rotate the solution to obtain maximum agreement. In multidimensional scaling a rotation of the solution usually does not change the interpretation, in this study about PCA, however, the interest is in the exact positions of the principal axes. It follows that a slight change in orientation of the solution, should be viewed as instability and we will not correct for it by rotating. In fact, we only allow for a reflection of one or more principal axes, because this only changes the signs of the whole solution and leaves the interpretation unaltered.

Quantifications

In order to compare the quantifications, which are the values of the optimally transformed categories, it is necessary to develop one or more measures, with which the transformation functions can be compared. In this study we will use as measures of comparison the *smoothness* (SMOT) - only ordinal variables are considered - and the *shape* (SHAP) of the functions. Smoothness is defined in terms of the second order differences between the category quantifications. The smoothness of a transformation function for the jth variable (j=1,..,m) is defined as:

$$\text{SMOT}(j) = \sum_{s=1}^{nc-2} ((q_{j(s)} - q_{j(s+1)}) - (q_{j(s+1)} - q_{j(s+2)}))^2. \tag{5}$$

Here s is the index of the categories, nc is the number of categories, and $q_{j(s)}$ is the quantification of the sth category for the jth variable. If the transformation function is linear, then SMOT will be zero. High values of SMOT indicate a very whimsical pattern of the quantifications.

Smoothness is not the only characteristic on which the transformation functions will be compared. Another interesting measure is the general shape of a function, which we have defined as

$$\text{SHAP}(j) = \frac{\rho(j) - \rho_{min}}{\rho_{max} - \rho_{min}}, \tag{6}$$

where the function $\rho(j)$ is defined as

$$\rho(j) = \sum_{s=1}^{nc-1} s\ r_s. \tag{7}$$

The vector $\mathbf{r}$ $(nc\text{-}1)$ contains the index numbers (r_s) of the first order differences (or distances) between the category quantifications; these indexes order the distances between the categories from small to large. Furthermore, in the situation in which several distances are equal, we average the index numbers belonging to these equal distances. The ρ_{min} and ρ_{max} are used to the standardize the SHAP measure, with $\rho_{min} = \Sigma_s(nc\text{-}s)s$ and $\rho_{max} = \Sigma_s(s^2)$. The index SHAP has the convenient property: $0 \leq \text{SHAP} \leq 1$. If SHAP is close to zero, then the function will be *concave*-like, on the other hand, when SHAP is close to one, the function will be *convex*-like. What is stated above holds for monotone increasing functions. However, if a function is monotone decreasing, the quantifications should be multiplied by -1 or the computed SHAP value should be subtracted from 1. Some examples are given in Figure 3.

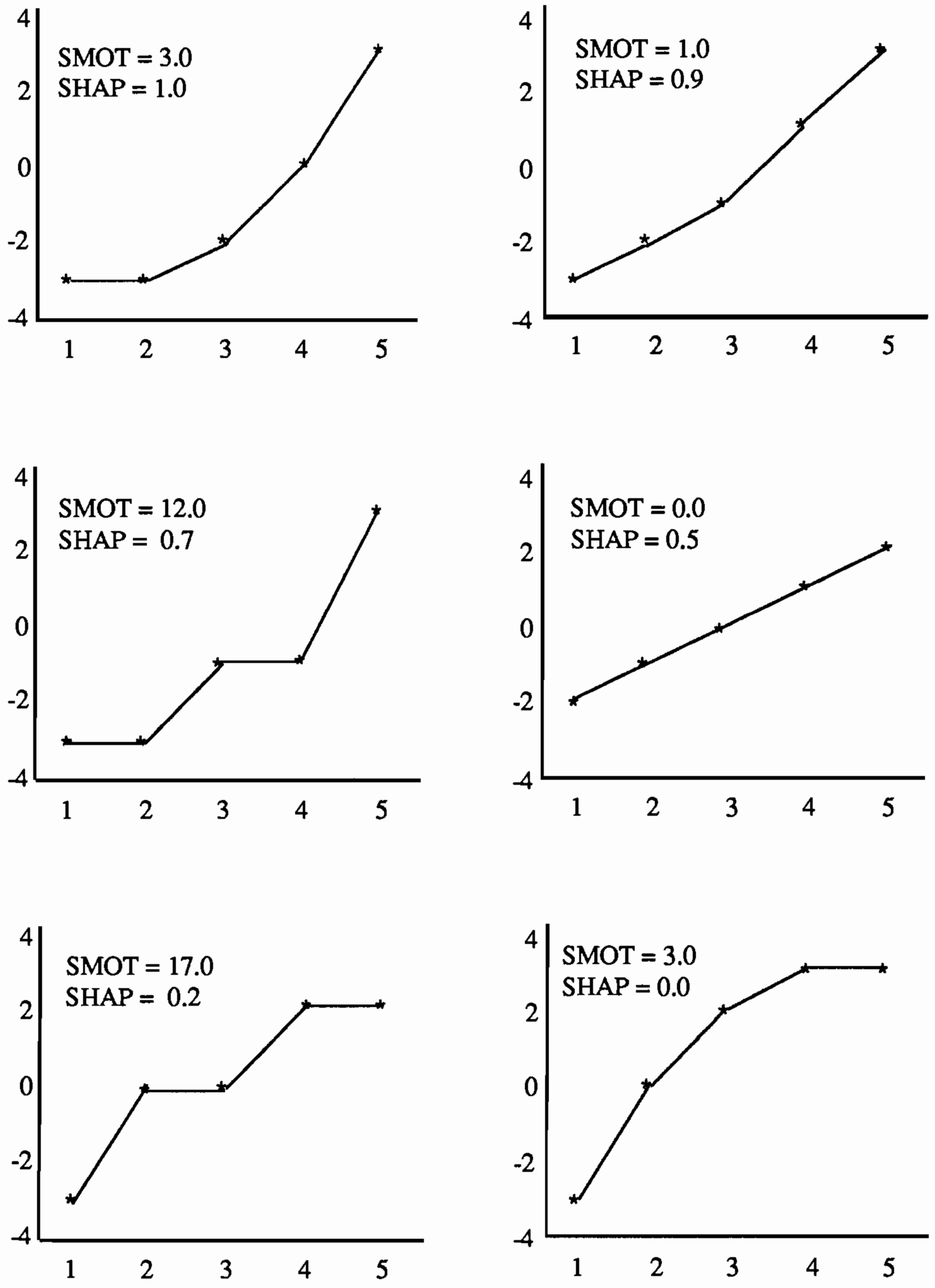

Figure 3. Illustrations of SHAP and SMOT measures.

Component loadings

Another set of output parameters that are to be compared, is the set of component loadings. To see if the component loadings have an overall similar structure in the p-dimensional space, first a generalized Procrustes analysis (GPA) was performed to rotate each set of component loadings to a centroid. We don't want an extreme configuration to have a large influence upon this centroid, so we actually used a resistant procedure to perform the GPA. For this GPA, we are only interested in orthogonal rotations of the sets. Therefore, for some optimal set of weights, we minimize the loss due to rotation,

$$\text{RLOS}(\mathbf{H}_i) = \Sigma_i \operatorname{tr} (\mathbf{C} - \mathbf{A}_i\mathbf{H}_i)'\mathbf{V}_i(\mathbf{C} - \mathbf{A}_i\mathbf{H}_i), \tag{8}$$

where $\mathbf{H}_i$ is an (p,p) rotation matrix, $\mathbf{C}$ (m,p) the centroid, $\mathbf{A}_i$ (m,p) the component loadings from the analysis, when the ith object was left out, and $\mathbf{V}_i$ (n,n) a diagonal matrix with weights that are chosen to obtain resistance (Verboon & Heiser, 1992). In this case we have chosen for the biweight weights. From (8) we can derive several measures. First, we can examine the loss for each analysis, with equal weighting of the loss components, thus

$$\text{RLOS}(i) = \operatorname{tr} (\mathbf{C} - \mathbf{A}_i\mathbf{H}_i)'(\mathbf{C} - \mathbf{A}_i\mathbf{H}_i) \,/\, mp. \tag{9}$$

If a solution is quite different, up to a rotation, from all the other solutions, then the RLOS for this analysis will be large. This means that the component loadings from this solution are quite different from all the others. In addition, we may compare the angles (ANGL), computed from the $\mathbf{H}_i$, to see if a solution has a different orientation than the centroid. These angles are relative measures, since the whole solution is arbitrarily positioned in the p-dimensional space. Furthermore, the weights $\mathbf{V}_i$, could be used to inspect, for an analysis with a large loss, which variables are responsible for this large loss.

Note that, contrary to the STAB and INFL measure, we now optimally rotate the solutions. If solutions with respect to the component loadings are approximately the same up to a rotation, we may detect it by inspecting the RLOS. The variation due to the rotation will be picked up by the STAB measure.

Loss weights

Finally, the loss weights show us if the technique considers an object or an entry in the datamatrix as an outlier. For both types of weighting, we can compute how many times a weight, that is smaller than a particular value, is assigned to the kth object, thus

$$\mathrm{WEIG}(k) = \sum_{i \neq k}^{n} g_{(\alpha)i} \qquad \text{with} \qquad \begin{array}{l} g_{(\alpha)i} = 1 \ \text{ if } \ w_{(i)k} < \alpha \\ g_{(\alpha)i} = 0 \ \text{ if } \ w_{(i)k} \geq \alpha \end{array} \tag{10}$$

where $w_{(i)k}$ is the kth row of the matrix $W_{(i)}$, order (n,m) or $(n,1)$, representing the weights obtained from the analysis, in which the ith object was left out. The ith row of this matrix has been filled with zero's. If the elementwise weighting approach is chosen, WEIG(k) is a vector (m) that indicates for each variable in the kth row of the datamatrix how many times it is assigned weights smaller than α. If the rowwise weighting approach is chosen, WEIG(k) indicates how many times a whole object is downweighted. The parameter α is chosen as .6 and .8 for HUB and .5 and .75 for BIW.

First experiment

The Data

In the first experiment the following steps have been taken. First a data matrix $\mathbf{Z}^*$ = $\{z_{ij}\}$ of order (n,m) was constructed, with n =50 and m = 6. The columns of $\mathbf{Z}^*$, representing the variables are denoted as $\mathbf{z}^*_j$ (j=1,...m). All values in $\mathbf{Z}^*$ were randomly drawn from a uniform distribution. By definition, this random process yields very low correlations. To increase some of the correlations, a new $\mathbf{z}_1$ was computed by: $\mathbf{z}_1 = \mathbf{z}^*_1 + 2\mathbf{z}^*_2 + 3\mathbf{z}^*_3$. For all other variables (j=2,..,6) we have $\mathbf{z}_j = \mathbf{z}^*_j$, yielding the data matrix $\mathbf{Z}$.

All variables were divided in 5 categories. For the first three variables this was done in such a way that the frequency distributions of these variables approximate symmetric and uni-modal distributions. Let the range of $\mathbf{z}_j$ be defined as the difference between the largest and the smallest value of $\mathbf{z}_j$. The 10% extremes on both sides of the range were scored 1 and 5, respectively. Values between 10% and 30% of the range were assigned a score 2, between 30-70% a score 3 and between 70-90% a score 4. This yields an expected frequency of the categories of respectively: 5,10, 20, 10, 5.

For the fourth and fifth variable the assignment of the categories was uniform, that is, each category has the same probability of occurrence. So, the frequency distribution of these variables becomes approximately uniform, that is, the expected frequency for all categories is 10. For the sixth variable a skew, uni-modal distribution is chosen. The categories were assigned by defining the intervals of the range of z_6 as: 0-30% (=1), 30-60% (=2), 60-80% (=3), 80-95% (=4), 95-100% (=5). This leads to expected frequencies for z_6 of: 15, 15, 10, 7.5 and 2.5. The distribution, given in Table 1, is the frequency distribution that has actually been obtained.

Table 1. Category Frequency of Random Data Matrix.

	variables					
categories	1	2	3	4	5	6
1	4	5	6	9	11	16
2	7	10	14	12	7	16
3	23	20	15	11	11	9
4	13	8	11	10	12	7
5	3	7	4	8	9	2

The correlation matrix of this artificial data set is given in Table 2.

Table 2. Correlation matrix (no outliers).

	1	2	3	4	5
2	.479				
3	.716	-.035			
4	.066	.028	.059		
5	-.045	-.199	.190	-.127	
6	.087	.127	-.064	.155	-.211

The next step is to contaminate the data matrix with outliers. Three random objects were selected as outliers (12, 36 and 42). The first two were given new values on the first three variables in order to decrease the correlations between the first variable with the second and the third. It follows that these are type II outliers that do not fit into the structure of the majority of the data. For the other outlier the values on the fourth and fifth variable were replaced by the value 6, so this object is extreme, with an unique score on two variables on the edge of the distribution; thus, a typical type I outlier.

The correlation matrix of the contaminated data set is given in Table 3.

Table 3. Correlation matrix (with outliers).

	1	2	3	4	5
2	.249				
3	.470	.035			
4	.014	.075	.063		
5	-.042	-.182	.187	-.010	
6	.033	.127	-.064	.135	-.209

It is clear that the outliers deflate most of the correlations. The third outlier slightly decreases the correlation between z_4 and z_5, because this correlation was originally negative, while the contaminated points are in the same direction.

Results

In the first experiment, we will compare the jackknife results performed for BIW (TC=3.0), for HUB (TC=1.5) and for Least Squares (LS). For both techniques, the elementwise approach for weighting the residuals will be used. First, we will show the eigenvalues for the LS analysis (see Figure 4), which are simply the sum of the squared component loadings per dimension. Note that for the elementwise weighting approach, there is no direct equivalent of eigenvalues.

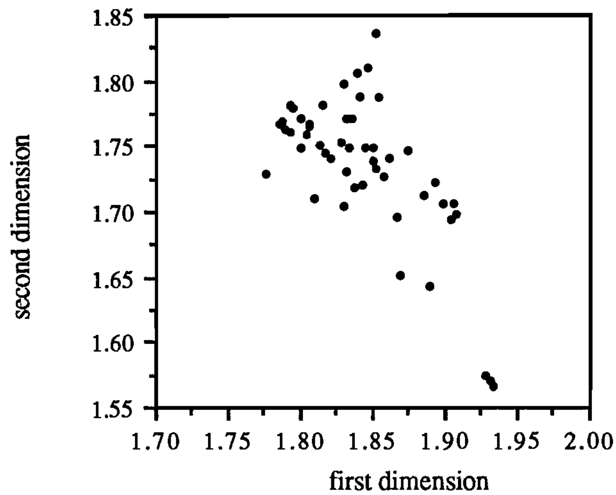

Figure 4. Eigenvalues of LS analysis.

It is clear that three objects (20,37,41) are highly influential with respect to the eigenvalues. With one of these objects deleted, the first eigenvalue increases, while the second decreases. A LS analysis on the clean data, thus without the outliers, yields as largest eigenvalues 2.04 and 1.77. This shows that for LS the outliers in particular affect the first eigenvalues, because in the jackknife experiment the first eigenvalue never exceeds the value 2.0 (see Figure 4). This is what could be expected, because the correlations are generally lower because of the outliers. The upper most point in Figure 4 is object 12, a constructed outlier, the deletion of which largely increases the second eigenvalue.

Influence and Stability

For examining the object scores we will first show the INFL indexes (see Figure 5). It appears that HUB shows three remarkable influential points (12, 39, 42), note that two of them were constructed as outliers, while all the other points have no influence at all. Therefore, for HUB, it seems that the solution is strongly dominated by these three points.

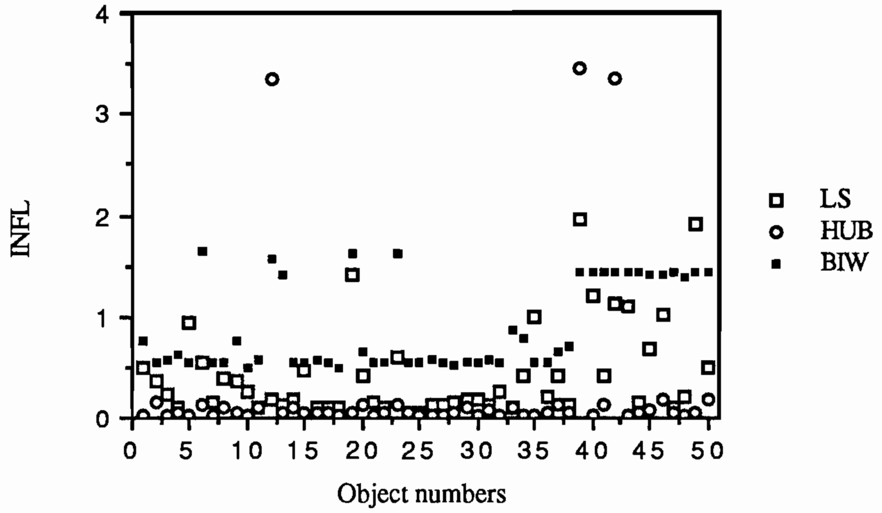

Figure 5. Influence for LS, HUB and BIW, first experiment.

Also, BIW appears to have more influential points than LS. However, after examining the solution more closely, we found that this peculiar result was due to a reflection of the solutions from the analyses with those points left out, that appeared to have a large influence. The reflections occurred both with HUB and with BIW. After reflecting these solutions, that is rotating them 180 degrees, the influence was computed again; the results are shown in Figure 6.

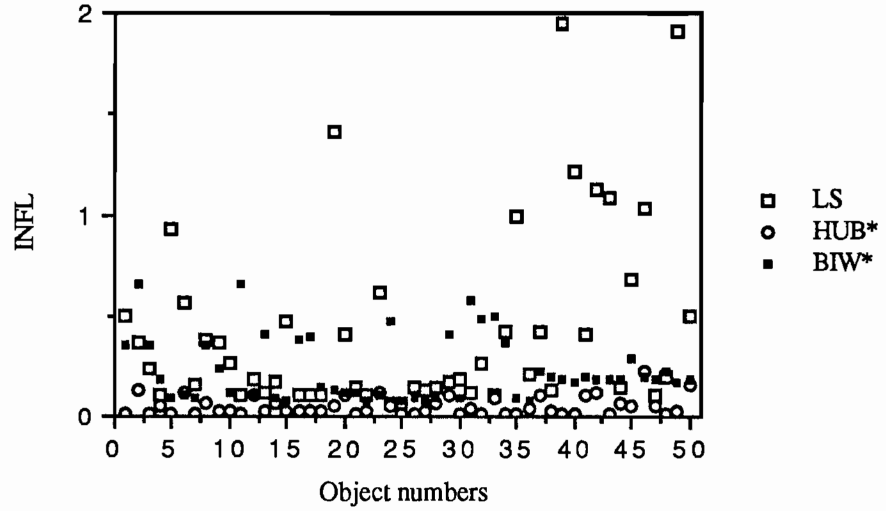

Figure 6. Influence (after reflections) for LS, HUB and BIW, first experiment.

Note that the scale of this figure is enlarged, compared to Figure 5. From Figure 6 it is immediately clear that HUB has no influential points at all, because all values are near zero. Also, for BIW there are no real influential points. However, there is a small group of objects with a slightly larger INFL value. The LS solution seems to yield quite a few influential points. These points influence the solution, because with the deletion of these points, the first and second principal axes were interchanged.

The results for the stability index (STAB), computed after the required reflections, are shown in Figure 7.

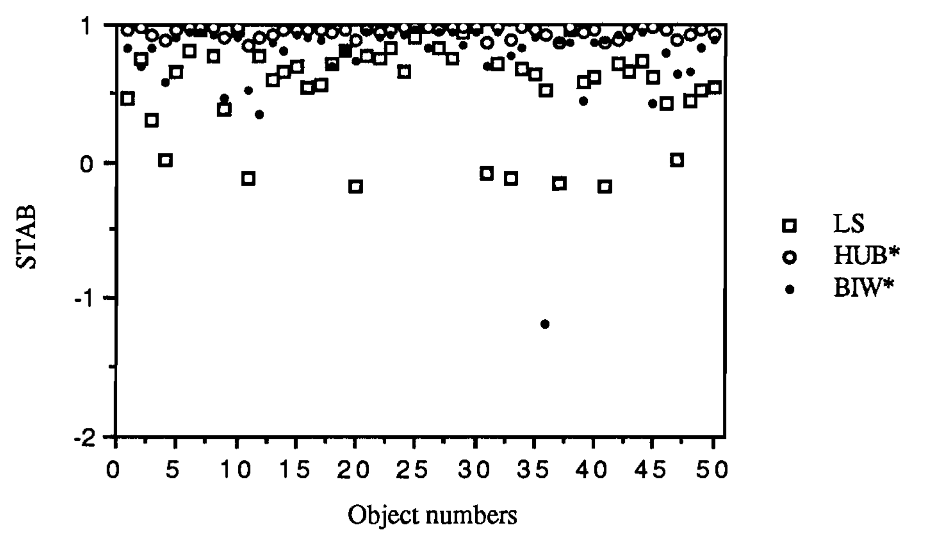

Figure 7. Stability (after reflections) for LS, HUB and BIW, first experiment.

In this figure we acknowledge that the stability for HUB is extremely good, all objects appear to be more stable than their counterparts from the other two analyses. Moreover, for most objects the stability is almost perfect, which means that there is hardly any variation of the object scores for such points under jackknifing all other points. Of course, this results could also be predicted from the INFL scores. For BIW there are a few slightly unstable points and there is one point (36) that is extremely unstable. Least squares is clearly the most unstable method. The average STAB value over all *n* runs are: .56 (LS), .77 (BIW) and .95 (HUB).

Loss weights

The loss weights will show us which points are considered as outliers by the two methods. In Table 4 the loss weights for some points are given; points are selected, which have at least been downweighted one time by HUB with the exception of point (36). This means that all cells from the data matrix which are not shown have never been assigned weights smaller than .8 in HUB. For BIW there are more cells with small weights, but we decided not to show all of them. Moreover, for reasons of clarity, we have not shown the loss weights corresponding to the second and third variable in HUB, since the cells corresponding with these variables were never downweighted.

Table 4. Frequency of loss weights, first experiment

	HUB						BIW											
	< .6		< .8				< .5						< .75					
object\vars	1	4	1	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
2	0	0	0	0	12	0	0	0	0	0	47	0	1	0	0	0	49	0
6	0	3	0	13	0	0	0	0	0	47	0	0	0	0	0	47	0	0
8	0	4	0	13	0	0	0	0	0	47	0	0	0	0	0	47	0	0
*12	34	0	48	0	0	0	49	1	0	0	1	0	49	1	0	0	35	0
13	0	2	0	35	0	0	0	0	0	2	0	0	0	0	0	2	0	0
19	0	0	0	1	0	1	0	0	0	2	0	48	0	0	0	2	1	49
22	0	2	0	35	0	0	0	0	0	2	0	0	0	0	0	2	0	0
23	0	0	0	13	0	0	0	0	0	47	0	0	0	0	0	47	0	0
26	0	0	0	0	0	32	0	0	0	0	0	24	0	0	0	0	0	45
27	0	0	0	35	0	0	0	0	0	2	0	0	0	0	0	2	0	0
*36	0	0	0	0	0	0	35	13	14	0	17	0	35	16	14	5	21	0
*42	0	13	0	49	0	0	0	0	0	49	1	0	0	0	0	49	1	0
46	0	0	0	1	0	0	0	0	0	4	6	0	0	0	0	24	25	0
50	0	0	0	8	0	0	0	0	0	33	0	0	0	0	0	47	0	0

Note: the constructed outliers are marked with a *.

The number of small weights is much larger for BIW than for HUB, which is the reason to take α somewhat smaller for BIW. For most elements both methods show the same pattern, however, there are differences. For instance, object (36) is downweighted quite drastically in BIW, while in HUB not even one element has been assigned a weight smaller than .8. Because this object is downweighted so severely in BIW, it does not contribute much to the loss of the function and one could say that it is allowed to move more freely through the object space than when it would have been fully weighted. This explains why this object is extremely unstable in BIW (see Figure 7). We should conclude that in this matter BIW outperforms HUB, because object (36) has been defined as an outlier and is detected as such by BIW, while HUB gives no indication that this object is an outlier.

The score on the first variable for object (12), which was constructed to decrease some correlations, has convincingly been detected by both analyses; also the extreme score on the fourth variable for object (42) has been detected as an outlier.

Smoothness and Shape

Finally, we will show the SHAP and SMOT values, but because it would cost too much space to show all variables, a selection of the variables has been made. Results will be presented for the variables 3 and 4, which seem to be the most interesting ones with respect to SHAP and SMOT. Box-and whisker plots (Tukey, 1977) are used to present the results. Since we compare uniform distributions, these type of plots are very suitable. Differences between the loss functions are visible and possible extreme values are visualized.

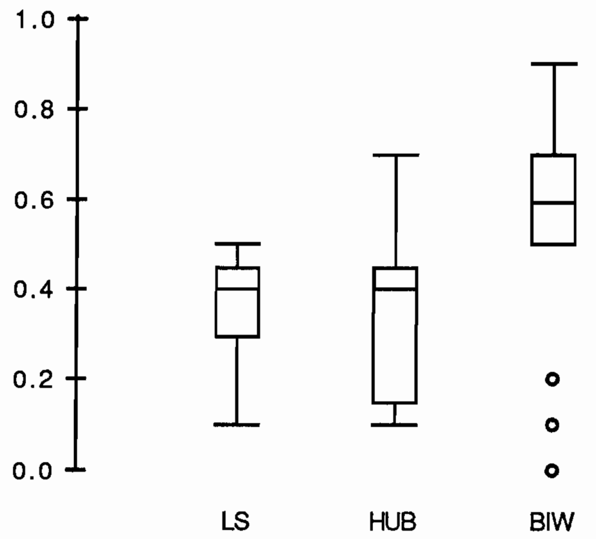

Figure 8a. Shape of variable 3 for LS, HUB and BIW, first experiment.

In Figure 8a the SHAP and SMOT for variable 3 are shown for all three techniques. The SHAP pattern for LS is rather stable under jackknifing over all objects, and is also quite similar to HUB. Object (36) causes a rather large change in SHAP for HUB, which is represented by the large upper whisker for HUB. It follows that this object has a large influence upon the quantification of variable 3. For LS the maximimum value for SHAP is also obtained under deletion of object (36), but here it is less extreme.

The SHAP pattern for BIW is very whimsical. For some points the shape of the transformation function is perfectly concave-like (shape equal to zero), for others it is convex-like. The distribution of the SHAPE measures for BIW is heterogeneous.

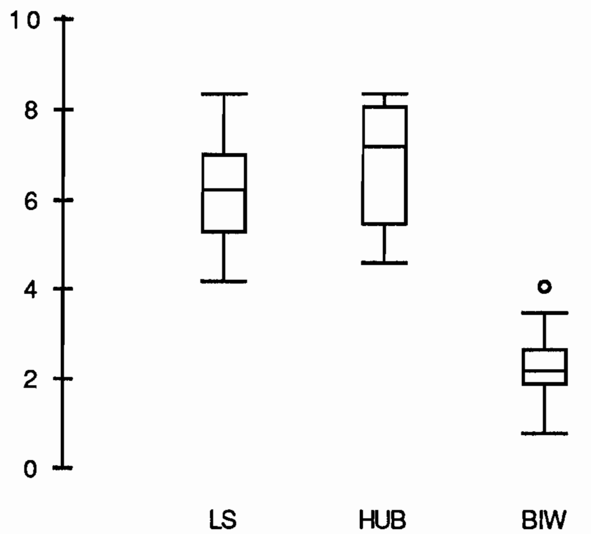

Figure 8b. Smoothness of variable 3 for LS, HUB and BIW, first experiment.

It is very clear from Figure 8b that the quantifications for the third variable in BIW are smoother than those from LS and HUB, which are quite similar with respect to the smoothness of the quantifications. Furthermore, the dispersion of the SMOT values is very small, especially for BIW.

In Figure 9 (upper plot) the SHAP values for variable 4 are shown. For this variable, BIW is much more stable with respect to SHAP than the other loss functions. The fact that some points are indicated as far out is due the extreme small interquartile range. Furthermore, Figure 9 shows that the quantifications for LS are the most concave-like, for BIW they are the most convex-like and for HUB they are somewhere in between. For identification some of the extreme points have been labeled.

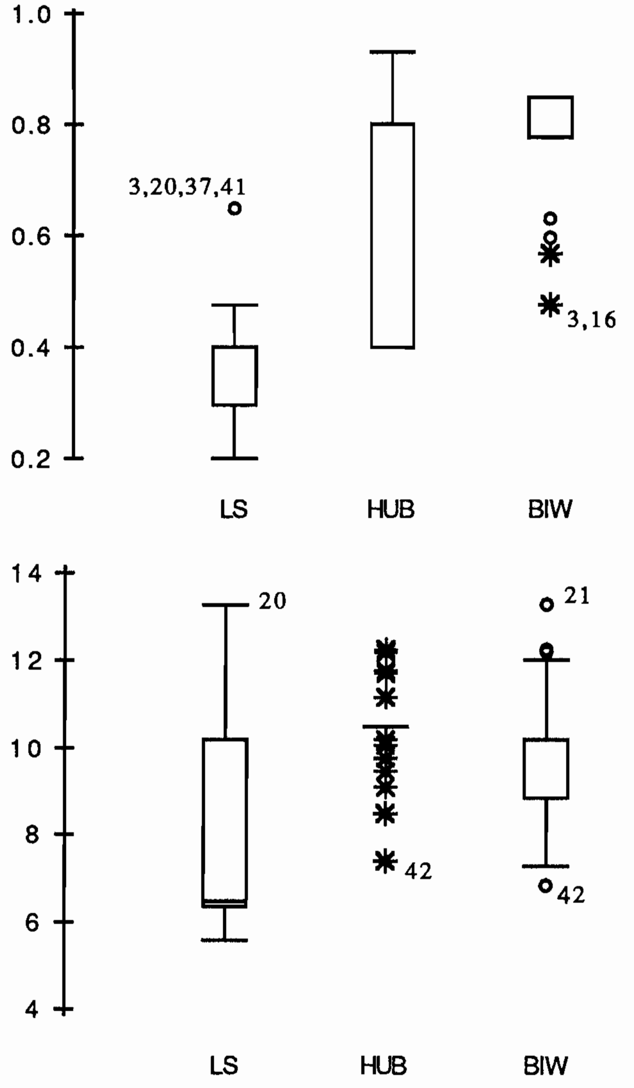

Figure 9. Shape (upper) and smoothness of variable 4 for LS, HUB and BIW, first experiment.

The SMOT measure (Figure 9, lower plot) shows that for the fourth variable LS yields somewhat smoother quantifications than the resistant functions. For HUB more than half of the analyses yield exactly the same SMOT value. Again some of the extreme values have been labeled in the plot, which shows that the constructed outlier, object (42), is positioned at the edge of the distribution for HUB and BIW.

Second experiment

Data and Results

For the second experiment the same data are used as for the first. The analyses in the second experiment were done by using the rowwise approach of weighting the residuals. Since, in the rowwise approach, which is a summation over all elementwise residuals, the rowwise residuals will be larger than in the elementwise approach, the tuning constants were also increased. For BIW the tuning constant was set to 6.0, and for HUB to 2.5. The TC for BIW was increased rather much, because in the first experiment we found many points with low weights (cf. Table 4). By increasing the TC in this way, we expect that BIW will downweight less points than before. The results for LS are, of course, the same as in the first experiment.

In Figure 10 for both HUB and BIW the eigenvalues are shown, which are computed as the sum of the squared component loadings. In the elementwise weighting approach the weighted object scores are normalized to n for both techniques. It is clear that both techniques show the same influential points as were found with LS (20,37,41). The eigenvalues for BIW are somewhat smaller than for HUB and LS.

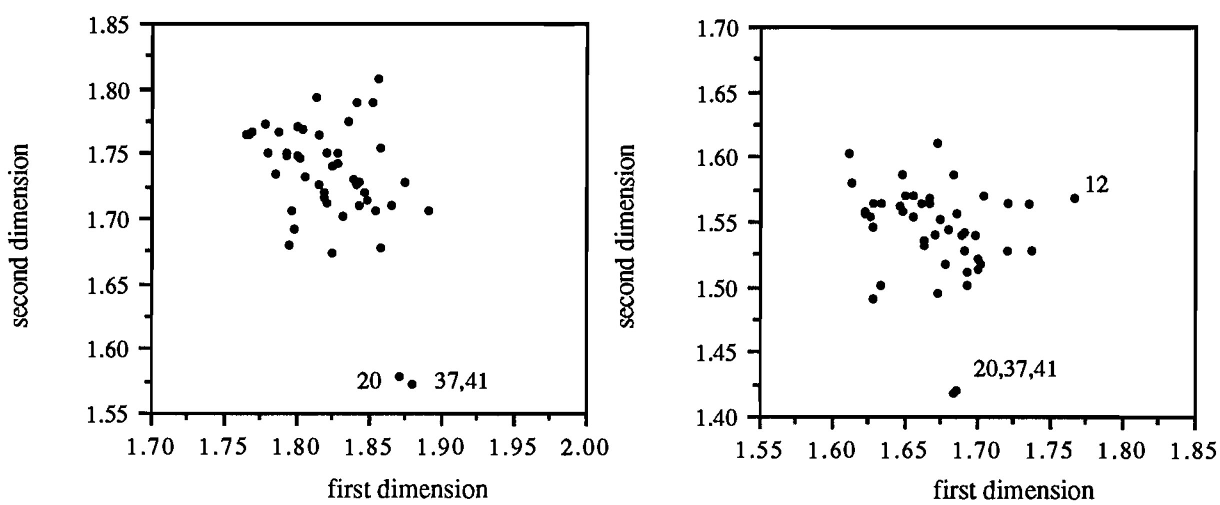

Figure 10. Eigenvalues for HUB (upper) and BIW (lower), second experiment.

Note that, contrary to LS, in BIW and HUB these eigenvalues are not optimized, which means that lower eigenvalues do not necessarily mean lower fit. Moreover, the loss values computed for each technique, according to its own loss function, are quite stable under jackknifing and do not show the same large fluctuations as for the eigenvalues.

Influence and stability

For all three techniques, the influences with respect to the object scores are shown (Figure 11). First, seven solutions from BIW have been reflected, to give all solutions the same orientation. For HUB and LS, no reflections could be applied. Figure 11 clearly shows that HUB has five influential points, while BIW has only one, moderately influential point.

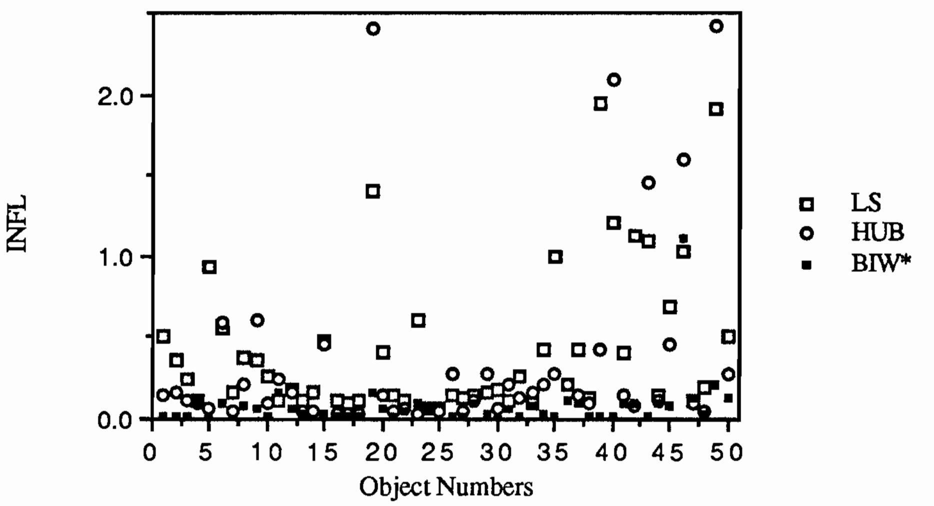

Figure 11. Influence (after reflections BIW) for LS, HUB and BIW, second experiment.

Inspecting the solutions more closely, it was found for the five influential points from HUB, that the first and second axis were almost perfectly interchanged. This is with respect to influence, of course, a very important result. It is also interesting to look at the INFL measures with these five solutions rotated 90 degrees, in order to give all solutions the same orientation. The same phenomenon was found for the LS analysis. For LS rotations have been applied for nine solutions. For BIW, there were no changes of the principal axes. The influence measures for these adjusted object scores, are given in Figure 12.

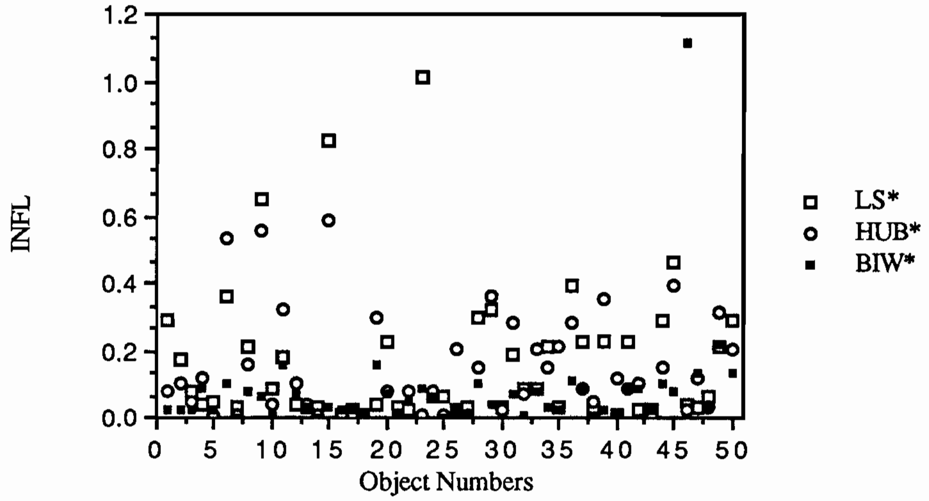

Figure 12. Influence (after rotations LS and HUB) for LS, HUB and BIW, second experiment.

It appears that for HUB, the five points only influenced the orientation of the object scores, but not its two-dimensional structure, which can be concluded from Figure 12, since there are no influential points left for HUB. The same conclusion can be drawn for LS, although for LS there are still some moderately influential points.

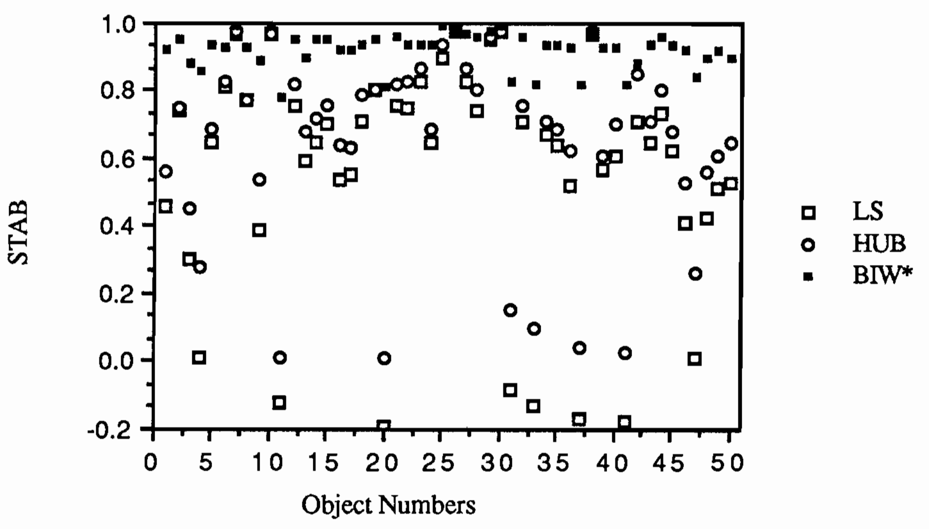

Figure 13. Stability (after reflections BIW) for LS, HUB and BIW, second experiment.

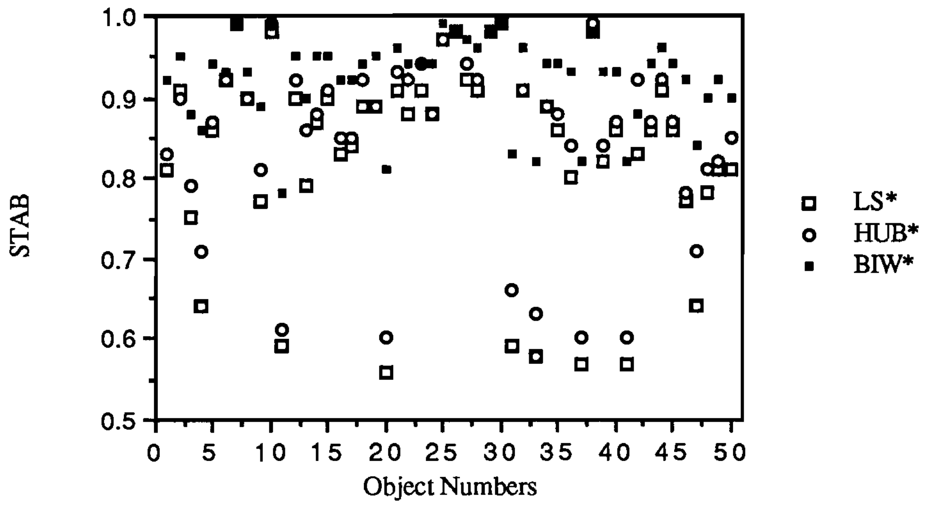

Figure 14. Stability (after rotations LS and HUB) for LS, HUB and BIW, second experiment.

In Figure 13 the stability of the object scores is shown. For each object, BIW is most stable, LS is least stable and HUB is closest to LS. If the rotations are applied, then the stability increases for all points (note the change in scale), but still the same ordening between the three techniques can be seen (Figure 14). The average STAB values for the unrotated solutions are: .56 (LS), .92 (BIW) and .65 (HUB); for the rotated solutions, we have .83 (LS) and .85 (HUB).

Loss weights

The frequency of the loss weights for the same points as in the first experiment is given in Table 5. Only two points are slightly downweighted in HUB. In fact, there are no points which are assigned weights smaller than .6. Despite the large tuning constant, there are still many points with small weights in BIW. These points are in general the same as the one from the first experiment (see Table 4). In both loss functions object (12) is assigned small weights in most of the analyses.

Table 5. Frequency of loss weights, second experiment

object	HUB < .8	BIW < .5	BIW < .75
2	0	1	1
6	0	1	6
8	0	1	6
*12	48	43	49
13	0	0	42
19	22	1	49
22	0	0	42
23	0	0	6
26	0	0	42
27	0	0	29
*36	0	0	49
*42	0	0	48
46	0	0	1
50	0	0	2

Note: the constructed outliers are marked with a *.

Smoothness and Shape

If we look at the SHAP and SMOT values, we find that for each variable the techniques show different effects. For some of the variables the results will be shown.

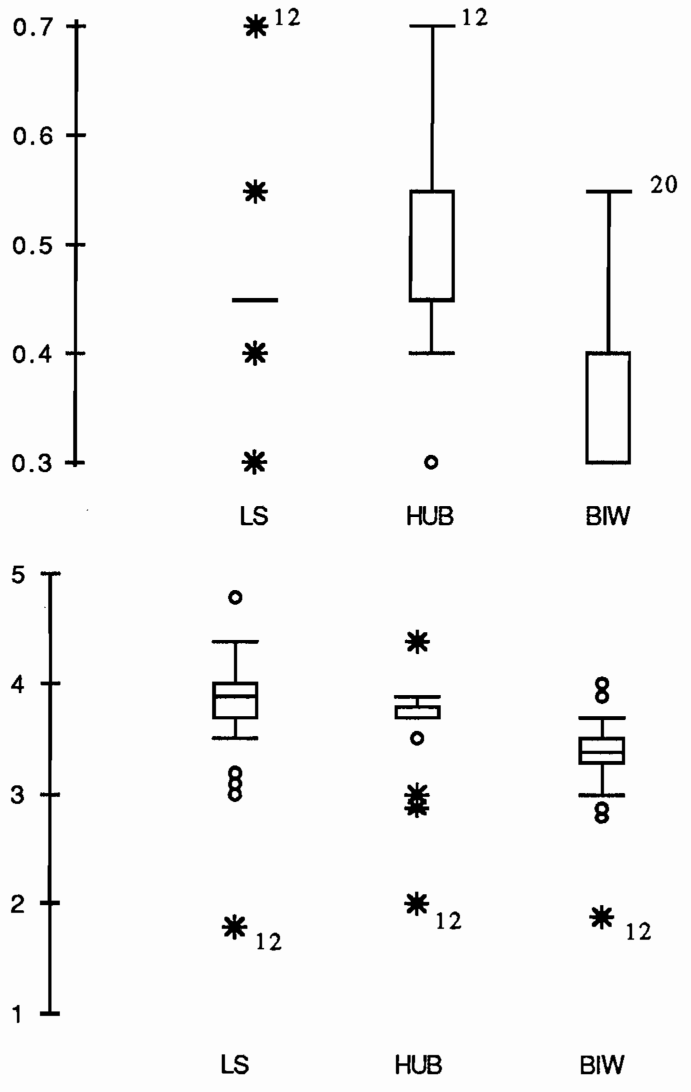

Figure 15. Shape and smoothness of variable 1 for LS, HUB and BIW, second experiment.

For the first variable (see Figure 15) BIW always has a lower value with respect to SHAP than HUB and LS, which are very much alike. The solution with object (12) deleted, yields for LS and HUB a somewhat higher SHAP value than with any other object deleted. The smoothness pattern for variable 1 is very similar for all loss functions; in case of object (36) all functions show a small decrease in SMOT.

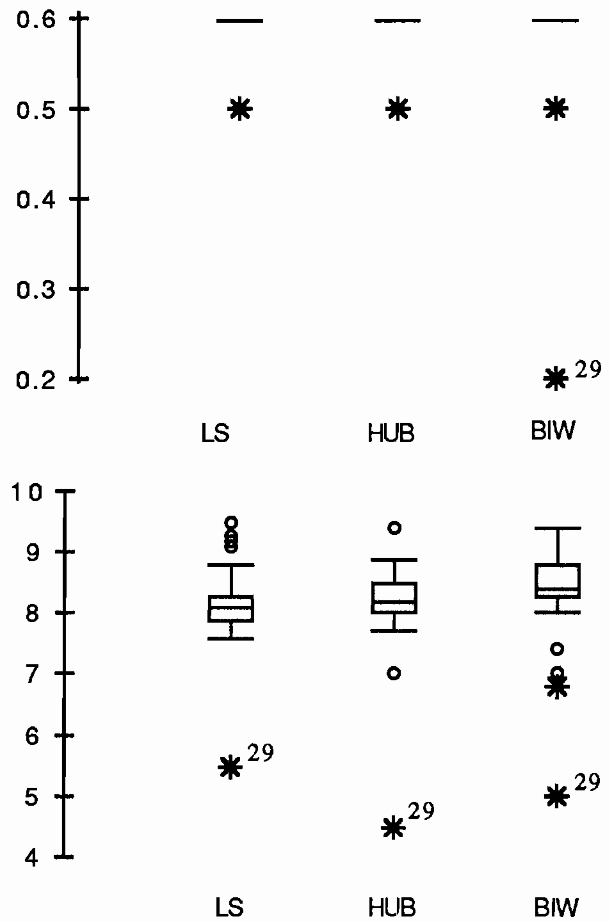

Figure 16. Shape and smoothness of variable 2 for LS, HUB and BIW, second experiment.

In Figure 16 the shape and smoothness of variable 2 are shown. It appears that the transformation of variable 2 is quite stable over all jackknifed objects, except for object (29). For this object, the SMOT values decrease for all three techniques, and furthermore, with respect to SHAP, it is BIW, which shows a significant drop. So, object (29) is very influential with respect to the transformation of variable 2.

For the shape of variable 4, LS and HUB are very similar and BIW has somewhat higher values (see Figure 17). For object (12) and (42) BIW shows deviant values, while LS yields a few slightly deviant points, furthermore HUB is quite stable. With respect to the smoothness, BIW is different from the other two techniques, which are very similar. It can also be seen that HUB and LS have a few extreme points, while BIW yields a more stable pattern, with the largest exception for object (46).

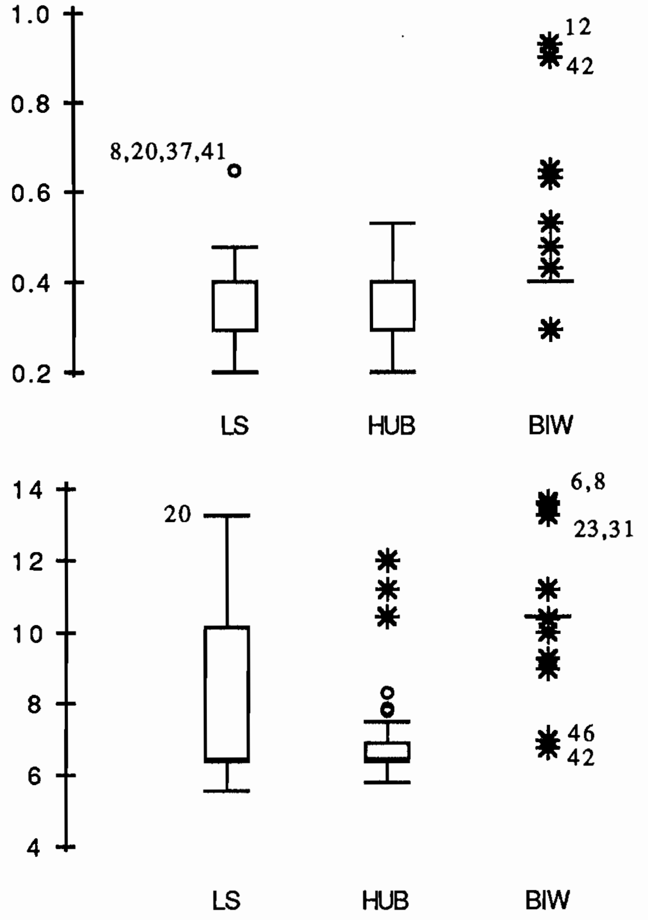

Figure 17. Shape and smoothness of variable 4 for LS, HUB and BIW, second experiment.

The shape and smoothness of variable 5 is not very interesting. There are no real deviant points and the patterns for each technique are quite similar.

Figure 18 shows the shape and smoothness of variable 6. With respect to the shape, for about half of the points a clear distinction is seen between LS on the one hand and BIW and HUB ont the other. BIW has the most stable pattern for shape. The smoothness yields the same kind of pattern, although there is consistent distinction between the three patterns, that is BIW has the smallest value of SMOT and LS the highest.

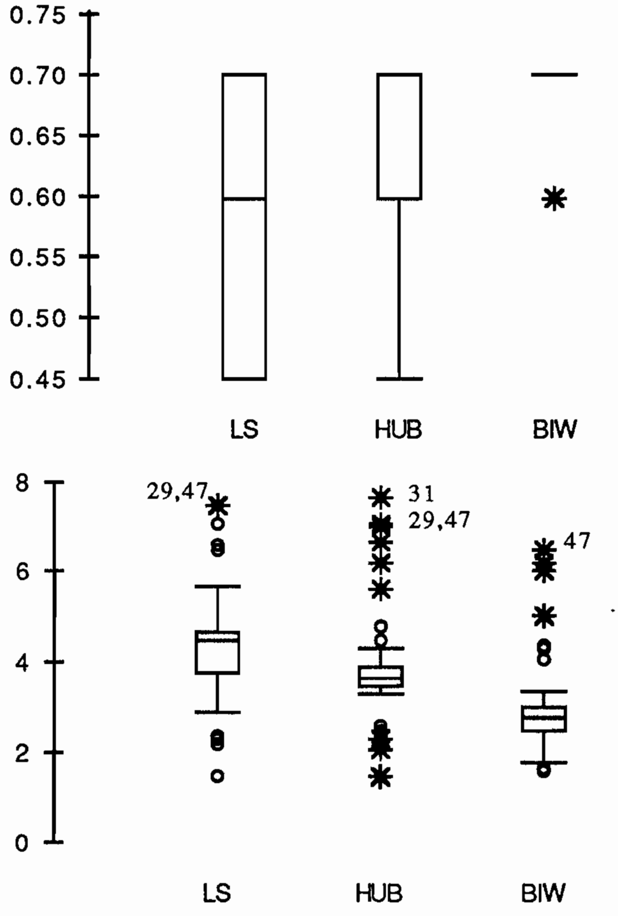

Figure 18. Shape and smoothness of variable 6 for LS, HUB and BIW, second experiment.

Angles and Rotation Loss

Finally, the angles are shown, which were found by orthogonal rotations of the component loadings to the centroid. In Figure 19 we can see that BIW (after reflection) only yields very small angles, except for object (46). So, BIW yields a very stable solution, which, of course, we already new from the STAB values. For HUB, the five objects that were acknowledged as influential points are clearly visible. These objects yield angles, which are approximately 90 degrees larger than the angles from the other objects. The nine influential points from LS can also be found, but here there are also a few points with just a rather large angle, which means that the solutions, computed with the deletion of these points, are more different from the general pattern than just a interchanging of the principal axes.

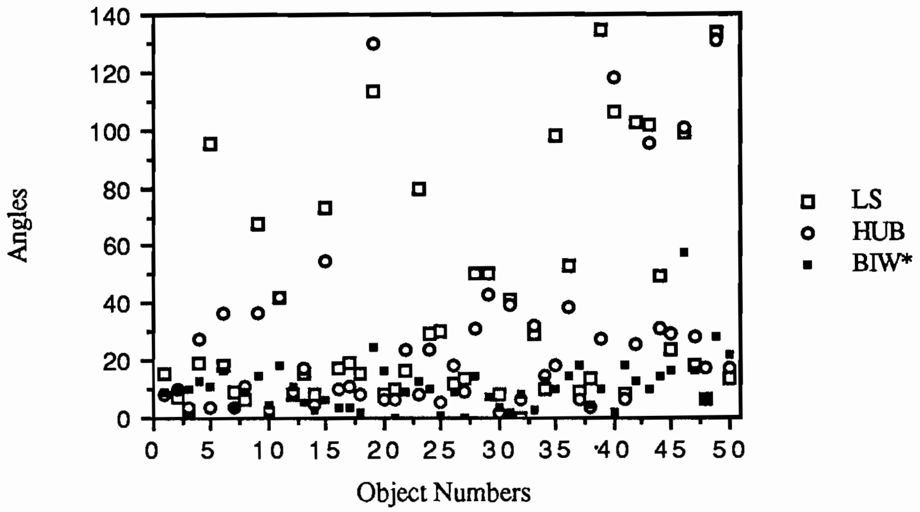

Figure 19. Optimal rotation angles for component loadings for LS, HUB and BIW, second experiment.

The RLOS values are very low for HUB, with one exception for object (8). The Procrustes weights tell us that variable four for this object didn't fit the structure of the component loadings of the centroid. The SMOT and SHAP, corresponding with the deletion of object (8), show indeed a peak. For BIW, the RLOS of the seven reflected solutions was larger than the other solutions. This was mainly due to a badly fitting variable four, as the Procrustes weights showed.

Discussion and Conclusions

A rather consistent picture appears, when we examine the stability of the object scores: BIW usually yields the most stable results, despite the many small weights in the first experiment - in particular for object (36), which decreased the stability to some extent. But especially in the second experiment it was clear that BIW is quite stable, even if we rotate some of the solutions from HUB and LS. Furthermore we noticed in both experiments that HUB is still much better with respect to the stability of the object scores than least squares.

Least squares seems to produce quite a number of influential points, more than the other techniques. In the first experiment the BIW solutions show points with slightly more influence than HUB, but in the second experiment this phenomenon is reversed. However, there is one remarkable object (46), which has a rather high influence in the BIW solution. It appears that the deletion of this object causes the quantifications of the variables three, four and five to be different from their quantifications in the other runs. In the first experiment the variables four and five were downweighted when object (46) was left out, so this differential downweighting reduced the influence of object (46), which was not possible in the second experiment.

What can be said about the transformations? It is very hard to make general statements about the category transformations with respect to the three loss criteria. For each variable, the different criteria show different results: sometimes we see that BIW is clearly distinguished from the other two, sometimes LS is quite different, but it also happens that all three criteria show different results or the same results.

However, SHAP and SMOT values can be used to show the influence of individual objects on a particular variable. It was seen that each variable may have its own influential points, which did not show up as influential for the object scores. Furthermore, we see that some variables obtain rather stable transformation functions over all jackknife runs, while other variables show whimsical patterns. So, we can conclude from this study that, even if the component loadings look quite stable under jackknifing, it may happen that one or more variables are quite differently transformed. We can also draw conclusions in the following manner: we saw that object (8) influenced the position of variable four with HUB (second experiment), furthermore, the SMOT and SHAP values tell us that this was due to a different quantification of this variable, so, object (8) influenced the quantification of variable 4, which gave it another position in the variable space. Resuming the foregoing, we may say that SMOT and SHAP are very helpful in detecting specific effects of individual observations upon the quantifications.

Finally, although we only used one dataset, which should make us careful, because some results may be data specific, it can be concluded that jackknifes can be very useful in discovering influential points in nonlinear PCA. The paper shows that different points may influence different aspects of the results, which can be made visible by making box plots or plotting sample influence functions or other curves that are derived from them.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., Tukey, J. W. (1972). *Robust estimates of location: survey and advances*. Princeton, NJ: Princeton University Press.

Critchley, F. (1985). Influence in Principal component analysis. *Biometrika, 72*, 627-636.

De Leeuw, J. & Meulman, J.J. (1986). A special jackknife for multidimensional scaling. *Journal of Classification, 3*, 97-112.

Devlin, S. J., Gnanadesikan, R., Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika, 62, 3*, 531-545.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*, 1-26.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal. Am. Statist. Assoc.*, 69, 383-393.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J.and Stahel, W.A. (1986). *Robust Statistics: the approach based on influence functions*. New York: Wiley.

Miller, R.G. (1974). The jackknife - a review. *Biometrika, 61*, 1-15.

Mallows, C. L. (1975). *On some topics in robustness*. (Research note: Bell Telephone Laboratories).

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika, 43*, 353-360.

Tukey, J.W. (1958). Bias and confidence in not-quite large samples (Abstract). *Ann. Math. Stat., 29*, 614.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading: Addison-Wesley.

Verboon, P. (1991). *Nonlinear principal component analysis: overview and new developments with respect to resistance properties*. RR 91-07. Leiden: Department of Data Theory.

Verboon, P. & Heiser, W.J. (1992). Resistant orthogonal Procrustes analysis. *Journal of Classification* (*in press*).