Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2008) 1694–1711
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Outlier identification in high dimensions

Peter Filzmoser[a,*], Ricardo Maronna[b], Mark Werner[c]

[a] *Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstraße 8-10, 1040 Vienna, Austria*
[b] *Department of Mathematics, Faculty of Exact Sciences, National University of La Plata, and C.I.C.P.B.A., La Plata, Argentina*
[c] *Department of Mathematics, The American University in Cairo, Egypt*

Received 14 September 2006; received in revised form 14 May 2007; accepted 15 May 2007
Available online 18 May 2007

Abstract

A computationally fast procedure for identifying outliers is presented that is particularly effective in high dimensions. This algorithm utilizes simple properties of principal components to identify outliers in the transformed space, leading to significant computational advantages for high-dimensional data. This approach requires considerably less computational time than existing methods for outlier detection, and is suitable for use on very large data sets. It is also capable of analyzing the data situation commonly found in certain biological applications in which the number of dimensions is several orders of magnitude larger than the number of observations. The performance of this method is illustrated on real and simulated data with dimension ranging in the thousands.

Keywords: Outlier identification; Robust estimators; High dimension; Robust principal components

1. Introduction

Accurate identification of outliers plays an important role in statistical analysis. If classical statistical models are blindly applied to data containing outliers, the results can be misleading at best. In addition, outliers themselves are often the special points of interest in many practical situations and their identification is the main purpose of the investigation. Classical tools based on the mean and covariance matrix are rarely able to detect all the multivariate outliers in a given sample due to the masking effect (Becker and Gather, 1999), with the consequence that methods based on classical measures are unsuitable for general use unless it is certain that outliers are not present. Contaminated data are commonly found in several situations, and so robust methods that identify or downweight outliers are essential tools for statisticians. The goal of this investigation is to provide an identification of outliers, prior to whatever modeling process is envisaged. Sometimes identification of outliers is the primary purpose of the analysis, other times the outliers need to be removed or downweighted prior to fitting non-robust models. We do not distinguish between the various reasons for outlier detection, we simply aim to inform the analyst of observations that are considerably different from the majority. Our procedures are therefore exploratory, and applicable to a wide variety of settings.

Most procedures with a high resistance to outliers are computationally intensive; not coincidentally, the availability of cheap computing resources has enabled this field to develop considerably in recent years. Among other advances,

* Corresponding author. Tel.: +43 1 58801 10733; fax: +43 1 58801 10799.
E-mail addresses: P.Filzmoser@tuwien.ac.at (P. Filzmoser), rmaronna@mail.retina.ar (R. Maronna), mwerner@aucegypt.edu (M. Werner).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.05.018

there currently exist a wide variety of statistical models ranging from regression to principal components (e.g. Hubert et al., 2005, among several others) that can incorporate outliers without being unduly influenced, as well as several algorithms that explicitly focus on outlier detection.

The availability of faster computers notwithstanding, the majority of robust methods are computationally intensive and still only feasible for data sets with dimensions ranging in the hundreds. Apart from S-estimators, for example the robust estimation routines mentioned in the preceding paragraph experience substantial computational difficulties on large data sets. With improved data collection and storage facilities, such data sets have become more common in recent years.

There are several applications where high-dimensional outlier identification and/or robust estimation are important. The field of Genetics, for instance, has recently received a lot of attention from statisticians (e.g. the project *Bioconductor*, http://www.bioconductor.org). Advances in computing power have enabled biologists to record and store huge databases of information. Such information tends to contain a fair amount of gross errors, however, so robust methods are needed to prevent these errors from influencing the statistical model. Clearly, algorithms that take a long time to compute are not ideal or even practical for such large data sets. In addition, there is a further complication encountered in genetic data. The number of dimensions is typically several orders of magnitude larger than the number of observations, leading to a singular covariance matrix, so the majority of statistical procedures cannot be applied in the usual way. As will be discussed later, this situation can be solved through singular value decomposition, but it does require special attention. Several other biological applications such as medical imaging and fMRI also contain very large data sets with more dimensions than observations in which the outliers are the particular values of interest. Similarly, astronomy is another field in which outlier identification is useful; with the introduction of cheap electronic recording and storage devices it is not uncommon for data sets to be measured in terms of terabytes. It can thus be seen that there are a number of important applications in which current robust statistical models are impractical.

This investigation is organized as follows. Section 2 presents a brief survey of existing outlier algorithms focusing on particular problems associated with high-dimensional data. We describe the proposed procedure in Section 3, comparing its performance in low dimensions to existing methods in Section 4. High-dimensional results are presented in Section 5 and finally, Section 6 summarizes our findings and mentions areas of future research.

2. A high-dimensional perspective on outlier detection

2.1. A brief overview of outlier detection

There are two basic approaches to outlier identification—distance-based methods and projection pursuit. Distance-based methods aim to detect outliers by computing a measure of how far a particular point is from the center of the data. The usual measure of "outlyingness" for a data point $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, is a robust version of the Mahalanobis distance,

$$RD_i = \sqrt{(\boldsymbol{x}_i - \boldsymbol{T})^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{T})}, \tag{1}$$

where $\boldsymbol{T}$ is a robust measure of location of the data set $\boldsymbol{X}$ and $\boldsymbol{C}$ is a robust estimate of the covariance matrix. Difficulties encountered by distance-based methods include (i) obtaining a reliable estimate of $\boldsymbol{C}$, as well as (ii) how large RD_i should be before a point is classified as outlying. This highlights the intimate connection between outlier identification and robust estimation—the latter is required as part of the former. Obtaining good robust estimators of $\boldsymbol{T}$ and $\boldsymbol{C}$ are prerequisites for distance-based outlier detection procedures. It is then essential to find a metric (based on $\boldsymbol{T}$ and $\boldsymbol{C}$) separating outliers from regular points. The final separation boundary generally depends on user-specified penalties for misclassification of outliers as well as regular points.

2.1.1. Robust estimation as primary goal

A simple robust estimate of location is the coordinatewise median. This estimator is not orthogonally equivariant (does not transform properly under orthogonal transformations), but if this property is important, the L1 median should

be used instead, defined as

$$\hat{\boldsymbol{\mu}}(\boldsymbol{X}) = \underset{\boldsymbol{\mu}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|, \tag{2}$$

where $\|\cdot\|$ stands for the Euclidean norm. The L1 median has maximal breakdown point, and a fast algorithm for its computation is given in Hössjer and Croux (1995).

A simple robust estimate of scale is the median absolute deviation (MAD), defined for a sample $\{x_1, \ldots, x_n\} \subset \mathbb{R}$ as

$$\mathrm{MAD}(x_1, \ldots, x_n) = 1.4826 \cdot \underset{j}{\mathrm{med}} \left| x_j - \underset{i}{\mathrm{med}}\ x_i \right|. \tag{3}$$

More complex estimators of location and scale are given by the class of S-estimators (Maronna et al., 2006), defined as the vector $\boldsymbol{T}$ and positive definite symmetric matrix $\boldsymbol{C}$ that satisfy

$$\min |\boldsymbol{C}| \quad \text{s.t.} \quad \frac{1}{n}\sum_{i=1}^{n} \rho(d_i/c) = b_0,$$

$$d_i = \sqrt{(\boldsymbol{x}_i - \boldsymbol{T})^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{T})}, \tag{4}$$

where $\rho(\cdot)$ is a non-decreasing function on $[0, \infty)$, and c and b_0 are tuning constants that can be jointly chosen to provide specific breakdown properties. It is usually easier to work with $\psi = \partial\rho/\partial d$ since ψ has a root where ρ has a minimum.

Distance-based algorithms that pursue robust estimation as a primary goal—without explicit outlier identification—include the OGK estimate (Maronna and Zamar, 2002), the minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) (Rousseeuw, 1985; Rousseeuw and van Driessen, 1999). MCD attempts to find the covariance matrix of minimum determinant including at least h data points, where h determines the robustness of the estimator; it should be at least $(n + p - 1)/2$. The MCD and MVE are examples of S-estimators with non-differentiable $\rho(d_i)$ since $\rho(d_i)$ is either 0 or 1. MCD demonstrates good performance on data sets with low dimension but on larger data sets the computational burden can be prohibitive—the exact solution requires a combinatorial search. In the latter case good starting points need to be obtained, yielding an approximately correct procedure. Equivariant methods of obtaining these starting points, however, are based on subsampling methods and the number of subsamples required to obtain an acceptable level of accuracy increases rapidly with dimension. Rousseeuw and van Driessen (1999) developed a faster version of MCD which was a sizable improvement, but is still quite computationally intensive. The OGK estimator (Maronna and Zamar, 2002) is based on pairwise robust estimates of the covariance. Gnanadesikan and Kettenring (1972) calculated a robust covariance estimate for two variables X and Y based on the identity

$$\mathrm{cov}(X, Y) = \tfrac{1}{4}(\sigma(X + Y)^2 - \sigma(X - Y)^2), \tag{5}$$

where σ is a robust estimate of the standard deviation. The matrix constructed from these pairwise estimates will not necessarily be positive semidefinite, so Maronna and Zamar (2002) proceed by performing an eigendecomposition of this matrix. Since the variables in eigenvector space are orthogonal, the covariances are zero and it is sufficient to obtain robust variance estimates of the data projected onto each eigenvector direction. The eigenvalues are then replaced with these robust variances, and the eigenvector transformation is applied in reverse to yield a positive semidefinite robust covariance matrix. If the original data matrix is robustly scaled (each component divided by its robust variance), the OGK will be scale invariant. This procedure can be iterated, although Maronna and Zamar (2002) find this is not always better. Maronna and Zamar (2002) also find that using weighted estimates is somewhat better, in which case the observations are weighted according to their robust distances d as scaled by the robust covariance matrix. They employ a weighting function of the form $I(d < d_0)$ where $I(\cdot)$ is the indicator function and d_0 is taken to be

$$d_0 = \frac{\chi^2_p(\beta)\mathrm{med}(d_1, \ldots, d_n)}{\chi^2_p(0.5)}, \tag{6}$$

where $\chi^2_p(\beta)$ is the β-quantile of the χ^2_p distribution. Observations thus receive full weight unless their robust distance $d > d_0$, in which case they receive zero weight. Maronna and Zamar (2002) note that the robust distances d can be

quickly computed in the eigenvector space without the need for matrix inversion since the p components are orthogonal in this space. That is,

$$d_i = \sum_{j=1}^{p} \left(\frac{z_{ij} - \mu(Z_j)}{\sigma(Z_j)} \right)^2, \quad i = 1, \ldots, n, \tag{7}$$

where z_{ij} are the data in the space of eigenvectors, Z_j are the p components in this space, μ is a robust location estimate and σ is a robust variance estimate.

2.1.2. Explicit outlier identification

Extending robust estimation to outlier detection requires some knowledge of the distribution of robust distances. If $\boldsymbol{X}$ follows a multivariate normal distribution, the squared classic Mahalanobis distance (based upon the sample mean and covariance matrix) follows a χ^2_p distribution (e.g. Johnson and Wichern, 1998). Also, if robust estimators $\boldsymbol{T}$ and $\boldsymbol{C}$ are applied to a large data set in which the non-outliers are normally distributed, Hardin and Rocke (2005) found that the squared distances could be described by a scaled F-distribution. However, for non-normal data, it is not clear how the outlier boundary should be determined to give optimal classification rates. These considerations form the basis for the use of d_0 by Maronna and Zamar (2002). The transformation of Eq. (6) helps the distribution of d_i's resemble that of χ^2_p for non-normal original data, leading to better results for the cutoff value than simply $\chi^2_p(\beta)$.

Promising algorithms that focus on identifying outliers include MULTOUT (Woodruff and Rocke, 1994), BACON (Billor et al., 2000) and Kurtosis1 (Peña and Prieto, 2001). MULTOUT and BACON are distance-based and accordingly, direct the majority of computational effort toward obtaining robust estimators $\boldsymbol{T}$ and $\boldsymbol{C}$. BACON starts with a small subset of observations presumed to be outlier-free, to which it iteratively adds points that have a small Mahalanobis distance based on $\boldsymbol{T}$ and $\boldsymbol{C}$ of the current subset. One reason that makes MCD unreliable for high p is that its contamination bias grows very rapidly with p (e.g. Adrover and Yohai, 2002). MULTOUT aims to reduce the computational burden by subdividing the data into cells and running MCD on each cell, i.e. reducing the number of observations that MCD operates on, with the same number of dimensions. It then combines the results from each cell to yield a starting point for an S-estimator (Davies, 1987) that solves a complex minimization problem to produce a robust estimate of the covariance matrix $\boldsymbol{C}$. S-estimators can sometimes converge to an incorrect local solution, so a good starting point is essential. However, relying on MCD in the first stage restricts MULTOUT from analyzing large data sets, especially those of high dimension. It would seem that methods based on combinatorial search and variants thereof possess an inherent inability to analyze large data sets.

2.1.3. Projection pursuit

In contrast to distance-based procedures are projection pursuit methods (Huber, 1985) which can similarly be applied to robust estimation as a primary goal or continue towards explicit outlier detection. The underlying motive of projection pursuit methods is to find suitable projections of the data in which the outliers are readily apparent and can thus be downweighted to yield a robust estimator, which in turn can be used to identify the outliers. Since they do not assume the data to originate from a particular distribution but only search for useful projections, projection pursuit procedures are not affected by non-normality and can be widely applied in diverse data situations. The penalty for such freedom comes in the form of increased computational burden, since it is not clear which projections should be examined; an exact method would require that *all* possible directions be examined. The earliest equivariant robust estimator having a high breakdown point in arbitrary dimension was the Stahel–Donoho estimator (Stahel, 1981; Donoho, 1982), studied in greater detail by Maronna and Yohai (1995). Computer approximation based on directions from random subsamples was developed by Stahel (1981), but clearly a large amount of time is required to obtain satisfactory results. To improve on the slow convergence rate, Kurtosis1 (Peña and Prieto, 2001) proposed to examine only the set of $2p$ directions that maximize or minimize the kurtosis. A small number of outliers would cause heavy tails and lead to a larger kurtosis coefficient, while a larger number of outliers would start introducing bimodality and decrease the kurtosis coefficient. Viewing the data along projections that have maximum and minimum kurtosis values would therefore seem to display the outliers in a more recognizable representation. After the two directions have been found that maximize and minimize the kurtosis value for the current data space, the search is continued in the orthogonal subspace of the remaining dimensions. Substituting $2p$ directions for a theoretically infinite set is clearly advantageous, although there is some debate about whether the kurtosis measure is always the best criterion to use. Thus, although projection pursuit

algorithms have the advantage of being applicable in unusual data situations, their computational difficulties seem formidable.

2.2. The high-dimensional situation

High-dimensional data introduce several problems to traditional statistical analysis. As previously mentioned, computation time increases more rapidly with p than with n. For combinatorial and projection pursuit algorithms, this increase is of sufficient magnitude to put in question the feasibility of such methods for high-dimensional data. Among the faster distance-based methods, computation times of algorithms increase linearly with n and cubically with p (matrix inversion is an order p^3 operation). This implies that for very high-dimensional data, the computational burden of inverting the scatter matrix is non-trivial. This is especially noticeable in iterative methods which require many iterations to converge, since the covariance matrix is inverted on every iteration. Thus, while the Mahalanobis distance is a very useful metric for finding correlated multivariate outliers, it is expensive to compute. Alternate methods of identifying outliers fare even worse, however, usually sacrificing either computational time or detection accuracy. The MCD is a good example of this trade-off in that the exact solution is very accurate but impractical to compute for all but small data sets, whereas a faster solution can be obtained if random subsampling is used to yield an approximate solution. It will be investigated in Section 5 whether the subsampling version of MCD is competitive regarding both accuracy and computation time. Projection pursuit methods including the Stahel–Donoho estimator and Kurtosis1 have computation times that increase very rapidly in higher dimensions, and are often at least an order of magnitude slower than distance-based methods since their search for appropriate projections is an inherently time-consuming task. Thus although the Mahalanobis distance may be computationally burdensome due to the matrix inversion step, the robust version—RD, as defined in Eq. (1)—is an accurate metric for outlier detection and could well be more computationally attractive than other approaches.

This is relevant to several biological applications where the data frequently have orders of magnitude more dimensions than observations. This is also the typical situation in chemometrics, which led to the development of partial least squares (PLS) (e.g. Tenenhaus et al., 2005) among other methods. Since the covariance matrix is singular the robust Mahalanobis distance cannot be computed. This is not as big a problem as initially appears, however, since the data can be transformed via singular value decomposition to an equivalent space of dimension $n - 1$ (see, e.g., Hubert et al., 2005) and the analysis conducted in the same way as $p < n$. Nevertheless, this situation requires special attention and most outlier algorithms have to be modified to process this type of high-dimensional data.

High-dimensional data have several interesting geometrical properties, described in more detail by Hall et al. (2005). One such property that is especially relevant to outlier detection is that high-dimensional data points lie near the surface of an expanding sphere. For instance, if $\|\boldsymbol{x}\|$ is the norm of $\boldsymbol{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ drawn from a normal distribution with zero mean and identity covariance matrix, then, for large p we have

$$\frac{\|\boldsymbol{x}\|}{\sqrt{p}} = \frac{\sqrt{\sum_{j=1}^{p}(x_j)^2}}{\sqrt{p}} \to 1,$$

since the summation involves a χ^2_p distribution. Thus, if the outliers have even a slightly different covariance structure from the non-outliers, they will lie on a different sphere. This does not help low-dimensional outlier detection, but if an algorithm is capable of processing high-dimensional data, it should not be too hard to discover the different spheres of the outliers and non-outliers. The hardest part, of course, is possessing the ability to analyze sufficiently high-dimensional data (within a reasonable time) to observe this phenomenon.

Principal components are a well-known method of dimension reduction that also suggest an approach to identifying high-dimensional outliers. Recall that principal components are those directions that maximize the variance along each component, subject to the condition of orthogonality. Since outliers increase the variance along their respective directions, it seems intuitive that outliers will appear more visible in principal component space than the original data space; that is, at least some of the directions of maximum variance are likely to be those that enable the outliers to "stick out" more. Searching for outliers in principal component space should at least, however, not be any worse than searching for them in the original data space. If the data originally reside in a high-dimensional space, many of these dimensions likely do not contribute significant additional information and are extraneous. Principal components thus select a handful of highly informative components (relative to the total number of components), thereby achieving a

high degree of dimension reduction and making the data set much more computationally tractable without losing a lot of information.

For high-dimensional data, a large portion of the smaller principal components are essentially noise (Jackson, 1991). Especially if $p \gg n$, the majority of principal components will indeed be noise and will not contribute to the total variance. By considering only those principal components that constitute some predetermined level of the total variance, the number of components can be substantially reduced so that only those components that are truly meaningful are retained. Practically, we found good results using a level of 99%. It can be argued this produces similar results to transforming the data via SVD to a dimension less than the minimum of n and p. Thus, instead of imposing a level of contribution to the variance such as 99%, it would also be possible to select the $n-1$ (or fewer) components with the largest variance.

As outlined in Eq. (7) in the OGK approach of Maronna and Zamar (2002), after dividing by the MAD, the Euclidean distance in principal components space is therefore equivalent to a robust Mahalanobis distance, since the off-diagonal elements of the scatter matrix are zero. Hence, it is not necessary to invert a $p \times p$ matrix when computing a measure of outlyingness for each point (i.e. the robust Mahalanobis distance), but merely to divide (or "standardize") each principal component by its respective variance element. Since eigenvector decomposition has computational complexity p^3 similar to matrix inversion, doing the robust distance calculations in principal component space is not more time-consuming than in regular data space. If this transformation helps the outliers become more visible and reduces the number of iterations required to detect them, the result will be a net savings in computational time.

It can be seen that the above concepts are based on simple inherent properties of principal components; this is another example of how principal components continue to present appealing properties to both theoretical and applied statisticians.

3. Description of the proposed procedure

The algorithm we present consists of two basic parts: a step that aims to detect location outliers, and a step that aims to detect scatter outliers. Scatter outliers possess a different scatter matrix than the rest of the data, while location outliers are described by a different location parameter.

To start, it is useful to robustly rescale or sphere each component using the coordinatewise median and the MAD, according to

$$x_{ij}^* = \frac{x_{ij} - \mathrm{med}(x_{1j}, \ldots, x_{nj})}{\mathrm{MAD}(x_{1j}, \ldots, x_{nj})}, \quad j = 1, \ldots, p. \tag{8}$$

Dimensions with a MAD of zero should be either omitted, or another scale measure has to be used instead. Starting with the rescaled data x_{ij}^*, we calculate a weighted covariance matrix, from which we compute the eigenvalues and -vectors and hence a semi-robust principal components decomposition. We retain only those eigenvectors/values that contribute to at least 99% of the total variance; call this new dimension p^*. The remaining components are generally useless noise and serve only to obscure any underlying structure. For the case $p \gg n$, this also solves the singularity problem since $p^* < n$. For the $p^* \times p^*$ matrix $\boldsymbol{V}$ of eigenvectors we thus obtain the matrix of principal components as

$$\boldsymbol{Z} = \boldsymbol{X}^* \boldsymbol{V}, \tag{9}$$

where $\boldsymbol{X}^*$ is the matrix with the elements x_{ij}^*. Rescale these principal components by the median and the MAD similar to Eq. (8),

$$z_{ij}^* = \frac{z_{ij} - \mathrm{med}(z_{1j}, \ldots, z_{nj})}{\mathrm{MAD}(z_{1j}, \ldots, z_{nj})}, \quad j = 1, \ldots, p^*. \tag{10}$$

Store $\boldsymbol{Z}^*$ for the second phase of the algorithm. After the above pre-processing steps, the location outlier phase is initiated by calculating the absolute value of a robust kurtosis measure for each component according to:

$$w_j = \left| \frac{1}{n} \sum_{i=1}^{n} \frac{(z_{ij}^* - \mathrm{med}(z_{1j}^*, \ldots, z_{nj}^*)^4}{\mathrm{MAD}(z_{1j}^*, \ldots, z_{nj}^*)^4} - 3 \right|, \quad j = 1, \ldots, p^*. \tag{11}$$

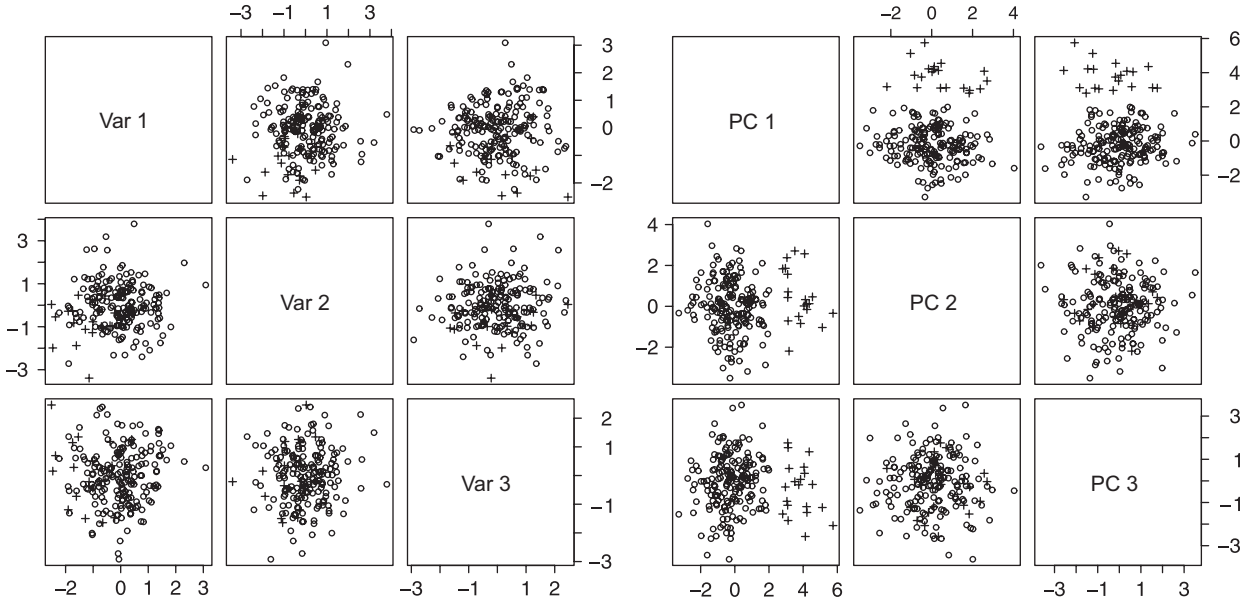

Fig. 1. Left: pairwise scatterplot of the first three original data components. Non-outliers are empty circles, outliers are marked by the symbol "+". Right: pairwise scatterplot of the first three principal components. Non-outliers are empty circles, outliers are marked by the symbol "+".

We utilize the absolute value because similar to Peña and Prieto (2001), both small and large values of the kurtosis coefficient can be indicative of outliers. This enables us to assign weights to each component according to how likely we think it is to reveal the outliers. We use relative weights $w_j/\sum_i w_i$ to provide a familiar scale $0 \leqslant w_j \leqslant 1$. If no outliers are present in a given component, we expect the principal components to be approximately normally distributed similar to the original data, yielding a kurtosis close to zero. Since the presence of outliers is likely to cause the kurtosis to become different than zero, we weight each of the p^* dimensions proportional to the absolute value of its kurtosis coefficient. Assigning equal weights to all components (during the computation of robust Mahalanobis distances) weakens the discriminatory power because if outliers clearly stick out in one component, the information in this component will be diminished unless it is given higher weight. Particularly in principal component space, outliers are more likely to be distinctly visible in one particular component than slightly visible in several components, so it is important to assign this component higher weight. Since the components are uncorrelated, we calculate a robust Mahalanobis distance utilizing the distance from the median (as scaled by the MAD), weighting each component according to the relative weights $w_j/\sum_i w_i$, with the kurtosis measure w_j defined in Eq. (11).

This idea is illustrated in Fig. 1. The left picture shows the pairwise scatterplot of the first three dimensions of a 200×20 data matrix with 20 outliers. To increase readability, only the first three dimensions are shown; the remaining dimensions do not differ substantially. The data are generated according to a multivariate standard normal distribution, the outliers have the same covariance matrix as the non-outliers but a slightly different center. The exact simulation setup will be described in Section 4 but to summarize, the data were generated according to a mixture of normals, with the mean of the outlier distribution shifted 4 units along the orthogonal direction to the first eigenvector. (In the notation of Section 4, the shift parameter is $k = 4$.) All methods except ours had difficulty with this configuration (see also Table 1 for $k = 4$). It can be seen that although a few outliers, marked by the symbol "+", are somewhat visible in the plot of the first two variables, this is not sufficient to identify the outliers. In contrast, the pairwise scatterplot in principal component space—shown in the right picture of Fig. 1—reveals many more outliers in the first component. The (raw) values of the robust kurtosis measure for these first three components are 13.43, 1.57, and 0.08, and the relative weights are 0.483, 0.056, and 0.003. The kurtosis measure defined in Eq. (11) therefore helps to ensure that important information contained in a particular component is not diluted by components which do not separate the outliers.

To finish the first phase of the algorithm, we need to determine how large the robust Mahalanobis distance should be to obtain an accurate classification between outliers and non-outliers. The kurtosis weighting scheme destroys any

Table 1
%FN and %FP; $p = 10$, $n = 1000$, $n_{out} = 100$, $\varepsilon = 10\%$, $\rho_{mult} = 0.5$, 500 simulations

Method	k	$\delta^2 = 0.1$		$\delta^2 = 0.5$		$\delta^2 = 1$		$\delta^2 = 2$		$\delta^2 = 5$	
		%FN	%FP	%FN	%FP	%FN	%FP	%FN	%FP	%FN	%FP
PCOut	0	100.00	7.15	99.96	6.81	–	5.30	61.16	4.00	8.84	3.49
Sign	0	100.00	9.14	99.99	3.97	–	2.56	66.19	1.66	10.78	1.06
OGK	0	100.00	4.30	100.00	4.06	–	2.87	61.49	2.03	6.69	1.88
Kurtosis1	0	100.00	3.38	100.00	2.66	–	1.33	16.09	1.21	0.02	1.37
MCD	0	100.00	5.51	100.00	3.74	–	2.60	63.94	1.64	7.65	1.40
PCOut	2	100.00	7.21	99.44	6.29	82.05	4.19	45.61	3.10	8.90	3.23
Sign	2	100.00	7.83	99.89	3.42	92.17	2.24	54.38	1.53	8.87	1.02
OGK	2	100.00	4.39	99.93	3.62	91.97	2.45	47.23	1.97	5.31	1.86
Kurtosis1	2	100.00	3.39	100.00	2.45	95.76	1.14	4.75	1.36	0.004	1.40
MCD	2	100.00	4.69	99.94	3.20	92.58	2.21	50.42	1.54	5.96	1.40
PCOut	5	67.27	1.60	15.29	1.59	7.25	1.65	3.77	1.74	5.92	2.11
Sign	5	91.83	8.13	46.86	3.65	26.98	2.30	12.97	1.51	3.02	0.97
OGK	5	100.00	4.62	89.74	2.12	24.54	1.91	6.82	1.87	1.28	1.86
Kurtosis1	5	100.00	3.32	100.00	2.04	72.13	1.00	0.00	1.36	0.00	1.37
MCD	5	100.00	4.00	32.16	1.43	16.46	1.41	7.20	1.40	1.52	1.40
PCOut	10	0.00	1.49	0.00	1.69	0.00	1.79	0.00	1.87	0.03	1.93
Sign	10	0.00	7.65	0.00	3.68	0.00	2.43	0.002	1.63	0.03	1.02
OGK	10	0.00	1.87	0.00	1.86	0.00	1.86	0.00	1.87	0.01	1.86
Kurtosis1	10	100.00	3.24	100.00	1.25	0.00	1.43	0.00	1.35	0.00	1.39
MCD	10	0.00	1.40	0.00	1.40	0.00	1.40	0.00	1.39	0.01	1.40

resemblance to a $\chi^2_{p^*}$ distribution that might have been present, so it is not possible use a $\chi^2_{p^*}$ quantile as a separation boundary. Nevertheless, similar to Maronna and Zamar (2002) and to Eq. (6), we found that transforming the robust distances $\{RD_i\}$ according to

$$d_i = RD_i \cdot \frac{\sqrt{\chi^2_{p^*,0.5}}}{\text{med}(RD_1, \ldots, RD_n)} \quad \text{for } i = 1, \ldots, n \tag{12}$$

helped the empirical distances $\{d_i\}$ to have the same median as the theoretical distances and thus bring the former somewhat closer to $\chi^2_{p^*}$, where $\chi^2_{p^*,0.5}$ is the $\chi^2_{p^*}$ 50th quantile. We utilize the translated biweight function (Rocke, 1996) to assign weights to each observation and use these weights as a measure of outlyingness. The translated biweight fits into the general framework of S-estimators described by Eq. (4) and is similar to Tukey's biweight function except that ψ starts rising from 0 at some point M from the origin. That is, observations closer than the scaled distance M to location estimate receive full weight of 1. The ψ function of the translated biweight is thus given by

$$\psi(d; c, M) = \begin{cases} d, & 0 \leqslant d < M, \\ d\left(1 - \left(\dfrac{d-M}{c}\right)^2\right)^2, & M \leqslant d \leqslant M + c, \\ 0, & d > M + c \end{cases} \tag{13}$$

which corresponds to the weighting function

$$w(d; c, M) = \begin{cases} 1, & 0 \leqslant d < M, \\ \left(1 - \left(\dfrac{d-M}{c}\right)^2\right)^2, & M \leqslant d \leqslant M + c, \\ 0, & d > M + c. \end{cases} \tag{14}$$

Directly assigning known non-outliers full weights of one while assigning known outliers weights of zero results in increased efficiency of the estimators (provided the classifications are correct), and is also computationally faster.

Between these extremes is a subset of points that receive weights similar to the usual biweight function. To maintain a high level of robustness, it is best to be conservative in assigning weights of one, since if any outliers enter the process with a weight of one (or close to it) that will make the other outliers harder to detect due to the masking effect. Recall that since the principal components have been scaled by the median and MAD these robust distances measure a weighted distance from the median (utilizing transformed units of the MAD). We found good empirical results assigning a weight of one to the $\frac{1}{3}$ of points possessing the smallest robust distances. At the other end of the weighting scheme we assign zero weight to points with $d_i > c$, where

$$c = \text{med}(d_1, \ldots, d_n) + 2.5 \cdot \text{MAD}(d_1, \ldots, d_n), \tag{15}$$

corresponding roughly to traditional outlier boundaries. Similar to Eq. (14), the weights for each observation are calculated by the translated biweight function according to

$$w_{1i} = \begin{cases} 0, & d_i \geqslant c, \\ \left(1 - \left(\dfrac{d_i - M}{c - M}\right)^2\right)^2, & M < d_i < c, \\ 1, & d_i \leqslant M, \end{cases} \tag{16}$$

where $i = 1, \ldots, n$ and M is the $33\frac{1}{3}$rd quantile of the distances $\{d_1, \ldots, d_n\}$. Other weighting schemes were tried; the advantage of the translated biweight is that it allows a subset of points (that we are quite sure are non-outliers) to be given full weight, while another subset of points that is likely to contain outliers can be given weights of zero, thereby precluding undesirable influence by potential outliers, and a smooth weighting curve for intermediate points. The weights $\{w_{1i}\}$ from Eq. (16) are stored; we will use them again at the end of the algorithm.

The second phase of our algorithm is similar to the first except that we do not use the kurtosis weighting scheme. Principal components focuses on those directions that have large variance, so it is perhaps not surprising that we find good results searching for scatter outliers in the semi-robust principal component space described at the beginning of this section. That is, we search for outliers in the space defined by $\boldsymbol{Z}^*$ from Eq. (10). As before, calculating the Euclidean norm for data in principal component space is equivalent to the Mahalanobis distance in the original data space, except that it is faster to compute.

Since the distribution of these distances has not been altered like it was through the kurtosis weighting scheme and assuming we start with normally distributed non-outliers, transforming the robust distances as before via Eq. (12) results in a distribution that is reasonably close to $\chi^2_{p^*}$. In setting up the translated biweight as in Eq. (16), then, satisfactory results can be obtained by setting M^2 equal to the $\chi^2_{p^*}$ 25th quantile and c^2 equal to the $\chi^2_{p^*}$ 99th quantile. This distribution is of course not exactly equal to $\chi^2_{p^*}$, so there are occasions when graphical examination of these distances could lead to a better boundary than this automated algorithm. Call the weights calculated in this way, w_{2i}, $i = 1, \ldots, n$.

Finally we combine the weights from these two steps to calculate final weights w_i, $i = 1, ..., n$, according to

$$w_i = \frac{(w_{1i} + s)(w_{2i} + s)}{(1 + s)^2}, \tag{17}$$

where typically the scaling constant $s = 0.25$. The reason for introducing s is that sometimes too many non-outliers receive a weight of 0 in only one of the two steps; setting $s \neq 0$ helps to ensure that the final weight $w_i = 0$ only if both steps assign a low weight. Outliers are then classified as points that have weight $w_i < 0.25$. These values imply that if one of the two weights w_{1i} or w_{2i} equals one, the other must be less than 0.0625 for the point $\boldsymbol{x}_i$ to be classified an outlier. Or, if $w_{1i} = w_{2i}$, then this common value must be less than 0.375 for $\boldsymbol{x}_i$ to be classified as outlying.

We will henceforth refer to this algorithm as PCOut. It is helpful to summarize the algorithm in point format:

Phase 1: Detection of location outliers.

Step 1: Robustly sphere the data according to Eq. (8). Calculate the sample covariance matrix of the transformed data $\boldsymbol{X}^*$.

Step 2: Compute a principal component decomposition of the semi-robust covariance matrix from Step 1 and retain only those p^* eigenvectors whose eigenvalues contribute to at least 99% of the total variance. Robustly sphere the transformed data as in Eq. (10).

Step 3: Compute the robust kurtosis weights for each component as in Eq. (11), and hence weighted norms for the sphered data from Step 2. Since the data have been scaled by the MAD, these Euclidean norms in principal component space are equivalent to robust Mahalanobis distances. Transform these distances according to Eq. (12).

Step 4: Determine weights w_{1i} for each robust distance according the translated biweight in Eq. (16), with M equal to the $33\frac{1}{3}$rd quantile of the distances $\{d_1, \ldots, d_n\}$ and $c = \text{med}(d_1, \ldots, d_n) + 2.5 \cdot \text{MAD}(d_1, \ldots, d_n)$.

Phase 2: Detection of scatter outliers.

Step 5: Use the same semi-robust principal component decomposition calculated in Step 2 and compute the (unweighted) Euclidean norms of the data in principal component space. Transform according to Eq. (12) to yield a set of distances for use in Step 6.

Step 6: Determine weights w_{2i} for each robust distance according to the translated biweight in Eq. (16) with c^2 equal to the $\chi^2_{p^*}$ 99th quantile and M^2 equal to the $\chi^2_{p^*}$ 25th quantile.

Combining Phase 1 *and Phase* 2: Use the weights from Steps 4 and 6 to determine final weights for all observations according to Eq. (17).

4. Comparison with methods in low dimension

Although this algorithm was designed primarily for computational efficiency at high dimension, we compare its performance against other outlier algorithms in low dimension, since a high-dimensional comparison is not feasible. In the following we examine a variety of outlier configurations in simulated data.

We follow the method of Maronna and Zamar (2002) to generate correlated multivariate normal, "worst case" data for methods based on principal components such as OGK and ours. Worst case data are important to consider since neither our method nor OGK are equivariant. We start by generating the non-outliers from a p-variate standard normal distribution $N_p(\mathbf{0}, \boldsymbol{I})$. The outliers will be generated from $N_p(\boldsymbol{y}_0, \delta^2\boldsymbol{I})$ for some δ and $\boldsymbol{y}_0 = k\boldsymbol{a}_0$ where $\boldsymbol{a}_0$ is a unit vector. We construct the simulated data $\boldsymbol{X}$ in such a way that for a highly correlated data set $\boldsymbol{X}$, the data will be concentrated around the line $\boldsymbol{a}_1 = (1, 1, \ldots, 1)^{\mathrm{T}}$, the eigenvector with the largest eigenvalue. The least favorable configuration for principal component methods will then be outliers situated orthogonally to this line. We therefore take $\boldsymbol{a}_0 = (b_1 - \bar{b}, \ldots, b_p - \bar{b})^{\mathrm{T}} / \sqrt{\sum_{j=1}^{p} (b_j - \bar{b})^2}$ where $\boldsymbol{b} = (b_1, \ldots, b_p)^{\mathrm{T}}$ consists of random draws from a uniform (0, 1) distribution, and $\bar{b}$ is the arithmetic mean of $\boldsymbol{b}$. Combining the non-outliers and outliers into a single data matrix $\boldsymbol{X}$, we then introduce correlation through multiplying by a matrix $\boldsymbol{R}$. Construct $\boldsymbol{R}$ from a diagonal consisting of 1's and off diagonal elements ρ. If no outliers are present, the multiple correlation ρ_{mult} between one coordinate of $\boldsymbol{X}$ and all the others can be calculated from ρ. We report ρ_{mult} as a measure of correlation for all of our results since it is more meaningful than ρ. Although $\rho_{\text{mult}} = 0.5$ for all of the simulations shown here, values of ρ_{mult} ranging from 0 to 0.999 were investigated; no essential differences in the relative performance of these methods were encountered compared to $\rho_{\text{mult}} = 0.5$. In this section, we show results for location outliers with k ranging from $k = 2$ to 10 and for scatter outliers with δ^2 ranging from $\delta^2 = 0.1$ to 5. Table 1 shows results for the percentage of outliers $\varepsilon = 0.1$ as a compromise between low and high levels of contamination; space does not permit us to show other values of ε, however, the results are comparable.

We present our results in the form of a table with (i) the percentage of false negatives (FN)—the outliers that were not identified, or masked outliers, and (ii) the percentage of false positives (FP)—non-outliers that were classified as outliers, or swamped non-outliers. All programs were executed in R (R Development Core Team, 2005) except for Kurtosis1, which was carried out in the Octave programming environment (http://www.octave.org), a free version mostly compatible with Matlab. We compared our algorithm to (i) a procedure based on robust principal components by Locantore et al. (1999) using the 97.5th quantile of χ^2_p as the cutoff value henceforth referred to as *Sign*, (ii) the $\text{OGK}_{(2)}(0.9)$ of Maronna and Zamar (2002) implemented in R by one of the present authors and similarly using the 97.5th quantile of χ^2_p as the cutoff value, henceforth referred to as OGK, (iii) Kurtosis1 by Peña and Prieto (2001) (Matlab code was supplied by original authors), and (iv) FastMCD with the default values as currently implemented in R, henceforth referred to as MCD. Somewhat similar to our method, the Sign procedure is also based on a type of robust principal component analysis. It obtains robust estimates of location and spread based upon projecting the data onto a sphere (or rather, an ellipsoid if the components are measured on different scales). In this way, the effects of outlying observations are limited since they are placed on the boundary of the ellipsoid and the resulting mean and covariance matrix are robust. Standard principal components can thus be carried out on the sphered data without undue influence

by any single point (or small subset of points). Unless otherwise mentioned, all of the following results represent the mean value from 500 Monte Carlo replications. Note that the case $k = 0$, $\delta^2 = 1$ corresponds to no outliers in the data, which reduces to a measurement of only the FP.

Examination of Table 1 reveals that PCOut performs well at identifying outliers (low FN), although it has a higher percentage of false positive than most of the methods. It does particularly well for location outliers, i.e. $k = 5$ ($k = 10$ is very easy for most methods and there is not much difference between them). Kurtosis1 does exceptionally well for scatter outliers ($\delta^2 = 2$ and 5), but very poorly for location outliers. Nevertheless it is apparent that with the exception of these cases involving scatter outliers, PCOut has the lowest percentage of FN, often by a considerable margin, and is a competitive outlier detection method.

5. High-dimensional results

5.1. Simulated data

In Table 2 we present simulation results in which the dimension was increased from $p = 50$–2000, based on the mean of 100 simulations at each level. In contrast to the previous simulation experiment in dimension $p = 10$, in this case we were not able to examine the performance of the other algorithms since they were not computationally feasible for these dimensions. The number of observations was held constant at $n = 2000$, as was the number of outliers at 200. The correlation coefficient ρ was calculated in such a way as to yield a multiple correlation coefficient $\rho_{\text{mult}} = 0.7$ for each dimension. The outliers consisted of pure scatter outliers (i.e. $k = 0$) with only a small amount of scatter, namely $\delta^2 = 1.2$. None of the known methods experience much success in identifying outliers with these characters for small dimensions because geometrically, the outliers are not very different from the non-outliers. However, as dimension increases, it can be seen how the outliers separate from the non-outliers and become easier to detect. At $p = 50$ dimensions, barely half of the outliers can be found, at $p = 500$ dimensions almost 90% are detected, and at $p = 2000$ dimensions more than 99% of the outliers are found.

5.2. Computation time

One of the key aspects of our algorithm is its computational speed. Accordingly, it is useful to make a detailed comparison of the various methods in this regard. We generate data in the same manner as in Section 4, keeping the number of observations fixed at $n = 2000$, with the proportion of outliers at 10%. The outliers are location and scatter outliers with $k = 10$ and $\delta^2 = 2$, the multiple correlation coefficient is taken to be $\rho_{\text{mult}} = 0$. We start with $p = 3$ dimensions, and increase p until the computation time for a method exceeds 1000 s. The experiment is repeated 10 times (using a computer with a Pentium 4 1.7 MHz processor), and the average computation time for the corresponding method is computed. These times are presented graphically in Fig. 2 using a log–log scale: The logarithm of p is shown on the horizontal axis, the logarithm of computation time (in seconds) on the vertical axis.

We see that computation times rise very rapidly for the algorithms Kurtosis1, OGK, and MCD, while PCOut and the Sign method are very similar. For $p \geqslant 100$ the computation time is very high for Kurtosis1, OGK, and MCD while PCOut and Sign are still quite feasible.

Table 2
Outliers are generated with a slightly larger covariance matrix ($\delta^2 = 1.2$)

Dimension	%FN	%FP
$p = 50$	49.5	6.92
$p = 100$	31.8	6.31
$p = 200$	18.3	4.98
$p = 500$	12.9	3.10
$p = 1000$	6.06	3.39
$p = 2000$	0.38	2.54

With increasing dimension PCOut can identify almost all outliers.

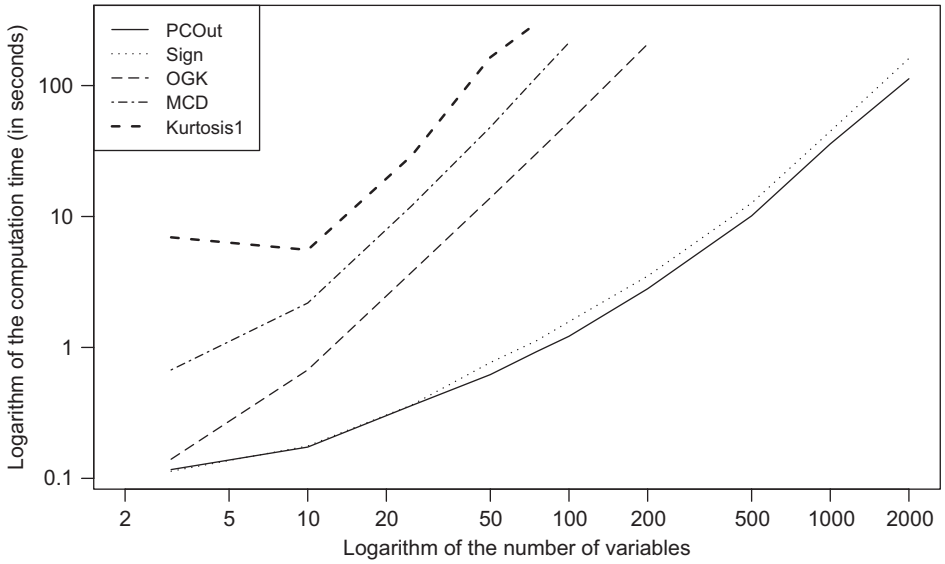

Fig. 2. Average computation time for $n = 2000$ and varying p.

5.3. Microarray data

The proposed method for multivariate outlier detection will be applied to a data set of Golub et al. (1999) with DNA microarray expression data, available as data set "golubMerge" in the R package "golubEsets". This data set comes from a study of gene expression in two types of acute leukemia: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The data comprise 47 cases of ALL and 25 cases of AML, resulting in 72 cases. Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays (HU6800 chip) containing probes for approximately 6800 human genes, but the chip actually contains 7129 different probe sets. We regard the 7129 genes as observations and the 72 cases as variables.

This data set has been used in several studies to discriminate the two patient groups (e.g. Dudoit et al., 2002). We are not aware of other outlier algorithms that examine this data set, we investigate it using PCOut to illustrate how our algorithm could assist in the analysis of high-dimensional microarray data. We will try to identify multivariate outliers among the 7129 genes, *without* using the information of the two leukemia types ALL and AML. The outlying genes will then be used for differentiating between the cases.

Similar to Golub et al. (1999) we first apply some pre-processing steps to the data: (i) thresholding, using a floor of 100 and a ceiling of 16 000; (ii) base-2 logarithmic transformation; (iii) standardization of the data so that the expression measures for each array have mean 0 and variance 1 across genes. Note that we will use all available genes, so we do not filter out those genes that are unlikely to be of interest.

Applying the proposed algorithm PCOut results in the identification of 2609 outliers among 7129 observations. Fig. 3 shows the resulting weights for each gene, together with the line at 0.25 separating outliers (below) from non-outliers (above). There is a very dense region of weights between 0.15 and 0.20. The reason is that in the original data many values are below 100, and the thresholding step sets them all equal to 100. These constant values are identified as atypical and thus appear as outliers.

It would be rather impractical to check for the meaning of all 2609 outliers in a gene expression database. Thus we will focus only on the most extreme outliers which will be those genes with weight below 0.05 in Fig. 3. In that way we extract 296 outlying genes.

Note that so far we have not used the information of the two patient groups. Nevertheless, we are interested if the identified 296 outliers allow for a separation of these groups. Applying a standard hierarchical clustering technique (we used average linkage) based on the distance matrix (specifically, the Manhattan distance in our case since it is less sensitive with respect to atypical observations) results in the dendrogram presented in Fig. 4. It can be seen that the two patient groups ALL and AML can be distinguished quite clearly only at the basis of the extreme outliers.

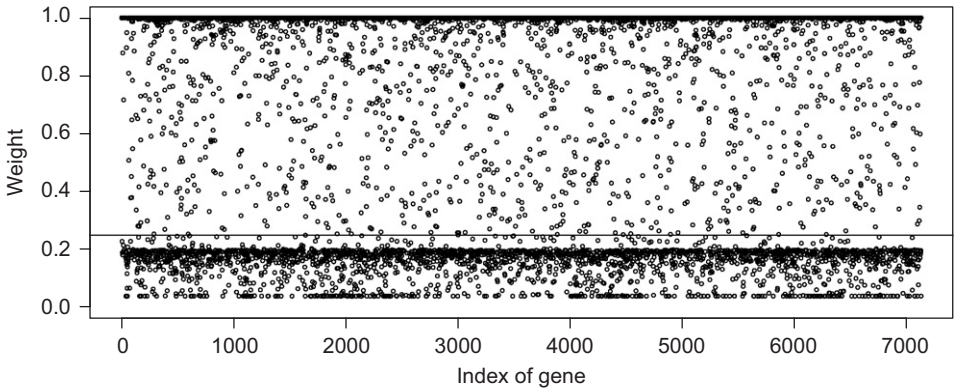

Fig. 3. Weight for each gene: genes with weight less than the boundary 0.25 are outliers.

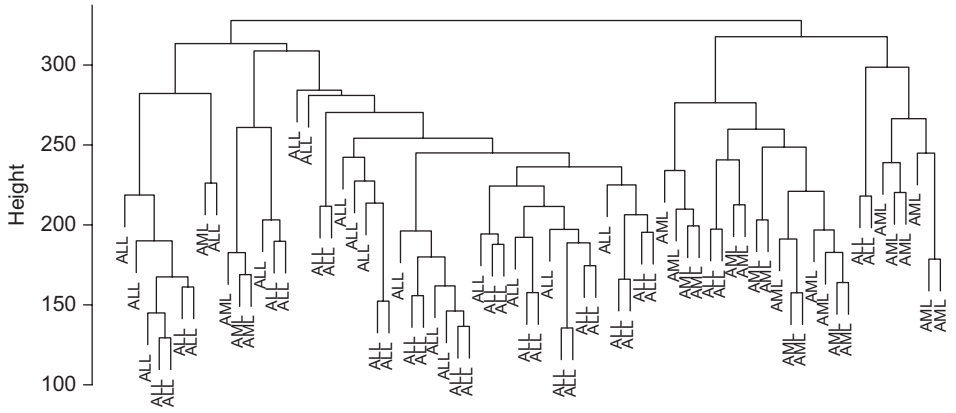

Fig. 4. Dendrogram after hierarchical clustering of the 296 most outlying genes.

The classification without using the grouping information can be of advantage if the grouping information is not reliable or if measurements are available without the corresponding information from which patient group they were taken from.

If the grouping information is used, we could be interested in deriving classification rules on the basis of the sub-data set of the 296 outliers. We could, for example, apply linear discriminant analysis (LDA) to this data set by providing the grouping information, and estimate the percentage of misclassified objects (patients). More information on the prediction quality of the 296 selected genes can be obtained by taking only a subset of observations (patients), building the LDA classification rule, and evaluating how many observations have been misclassified. In Fig. 5, subsets of 36–72 observations (50%–100% of the observations) were taken randomly, and the percentage of misclassified objects was computed over 100 replications. The results are shown as the solid line in Fig. 5. We compare the results for misclassification by selecting genes in a more traditional way: the genes are filtered according to their nominal p-value for a two-sample Welch t-test comparing expression measures in ALL and AML patients. We filter out only highly significant genes with a p-value of less than 0.0001. Four hundred and four highly significant genes are identified in this way; the dashed line in Fig. 5 shows the misclassification rates for the new subset of 404 genes. It can be seen that the misclassification rates are in general very low. There is almost no difference in the performance of both methods for gene selection if at least 75% of the observations are used for LDA. The advantage, however, of the gene selection by the multivariate outlier detection method is that the patient group information is not needed, and that the classification with LDA works well even if the group information was lost for a considerable proportion of samples.

It could be of interest to know if genes selected by the outlier identification method are contained in the 404 most significant genes, or if there is any relation between the two subsets. Fig. 6 shows the relation by a mosaic plot:

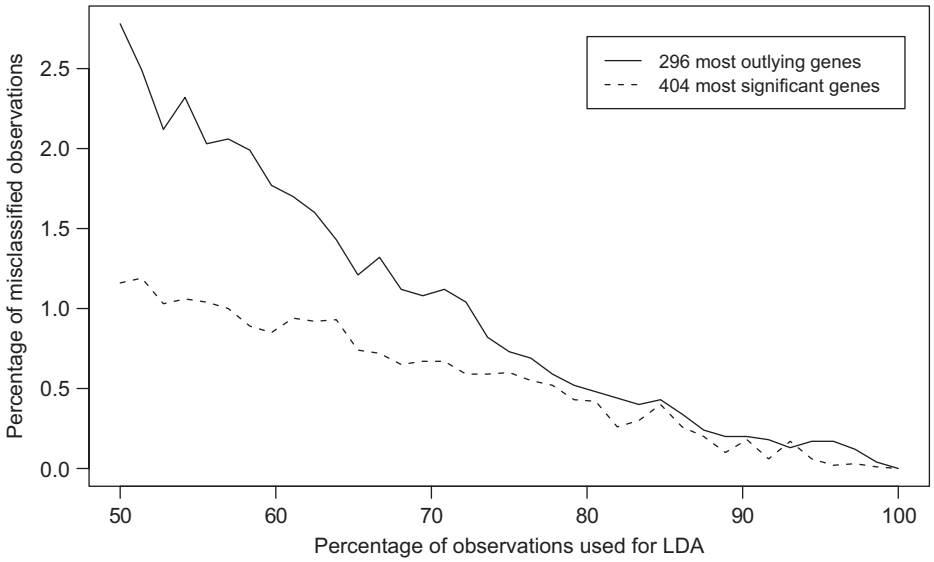

Fig. 5. Misclassification rates with LDA by using only a subset (50%–100%) of observations. The solid line is for the set of the 296 most outlying genes, the dashed line for the set of 404 most significant genes according to the two-sample Welch *t*-test.

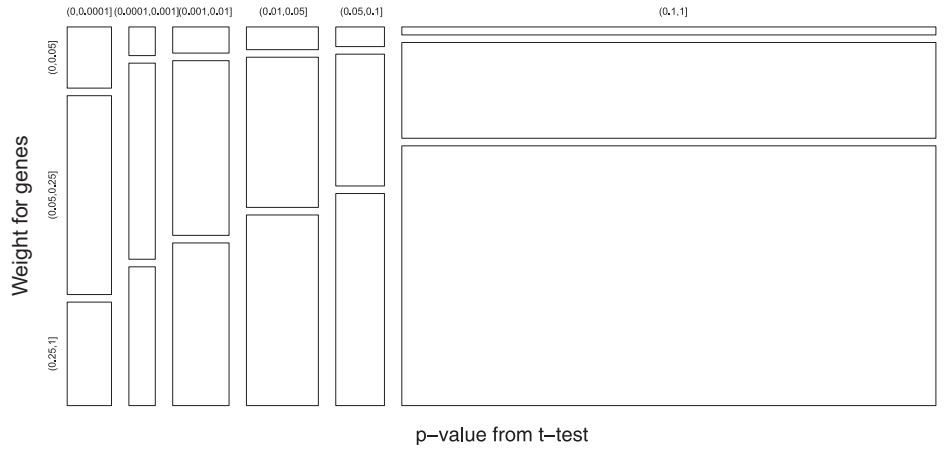

Fig. 6. Mosaic plot using classes for the *p*-values of the two-sample Welch *t*-test and classes for the weights of the multivariate outlier detection method.

the *p*-values from the two-sample Welch *t*-test are split in several classes, and also the weights from the algorithm PCOut are classified. Combining both classifications leads to a two-way table with cell frequencies, and the latter are visualized by proportional areas in the mosaic plot. The plot shows a clear relation between the *p*-values and the weights.

5.4. Geochemical data

In a large geochemical mapping project, carried out from 1992 to 1998 by the Geological Surveys of Finland and Norway, and the Central Kola Expedition, Russia, an area covering 188 000 km^2 at the peninsula Kola was sampled. In total, around 600 samples of soil were taken in four different layers (moss, humus, B-horizon, C-horizon), and

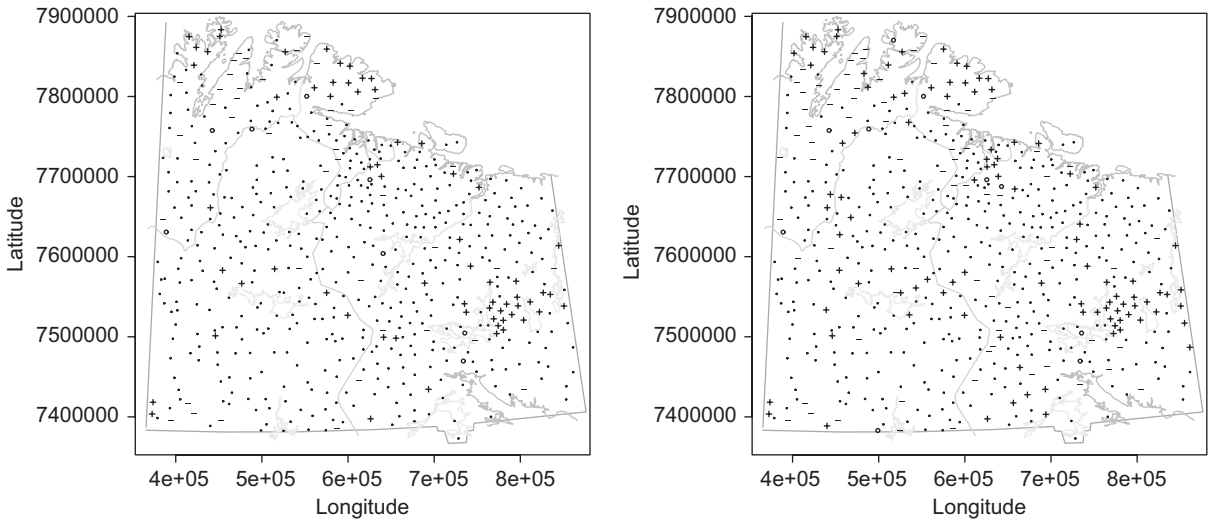

Fig. 7. Map of the Kola project area with symbols indicating outlyingness of the observations: Non-outliers are marked by the symbol •, outliers with low, moderate, and high average element concentrations by −, ∘, and +, respectively. Left: PCOut, right: Sign.

subsequently analyzed by a number of different techniques for more than 50 chemical elements. The project was primarily designed to reveal the environmental conditions in the area. More details can be found in Reimann et al. (1998) which also includes maps of the single element distributions. Many papers have been devoted to this data set, a part of the data was also used for multivariate outlier detection by Filzmoser et al. (2005). The data are available in the R package *mvoutlier*.

Similar to the microarray data, multivariate outlier detection is used here as an exploratory tool—in this case, to give the analyst an idea where in the geographical region irregularities or exceptional deviations can be expected. Therefore, we use all the variables (chemical element concentrations) except those where more than 2% of the values were below the detection limit. Since geochemical data are usually very skewed, the data were log-transformed in order to better resemble elliptical symmetry. The resulting data set for outlier detection thus consists of 581 observations and 182 variables.

Fig. 7 shows the resulting outliers embedded in a map of the project region for the methods PCOut (left) and Sign (right). Non-outliers are plotted with a small dot, ·. In order to simplify the interpretation of the outliers, we use three different symbols. We utilize the symbol − if an observation's value is smaller than the median value for at least half of the variables, the symbol + if it is larger than the median for at least half of the variables, and the symbol ∘ otherwise. Our results are consistent with geochemical knowledge of this region. For instance, there are multivariate outliers with high element concentrations in the east of the project region—this area around Monchegorsk contains big smelters causing a lot of pollution. Along the coast we also find outliers with high concentrations, caused by the sea spray from the Barents Sea. Conversely, away from the coast in the north we find an area with moderate to low element concentrations. This region is known to be pristine and is sparsely populated.

5.5. Data from chemometrics

A fast multivariate outlier detection method is particularly useful in the field of chemometrics where hundreds or even thousands of spectra need to be analyzed. We examine a situation in which archæological glass vessels from the 16th and 17th centuries were investigated through chemical analysis. The aim of this study was to learn more about the production of these vessels, particularly regarding their origin and possible trade connections between known producers. The 180 glass vessels were analyzed by an electron-probe X-ray micro-analysis (EXPMA) leading to 1920 characteristics for each vessel. Some previous studies analyzed the element concentrations of the vessels (e.g. Janssens et al., 1998). Relations between concentrations and spectra were investigated with PLS in Lemberge et al. (2000), although these results were biased due to outliers. Recently, Serneels et al. (2005) applied a robust version of PLS to

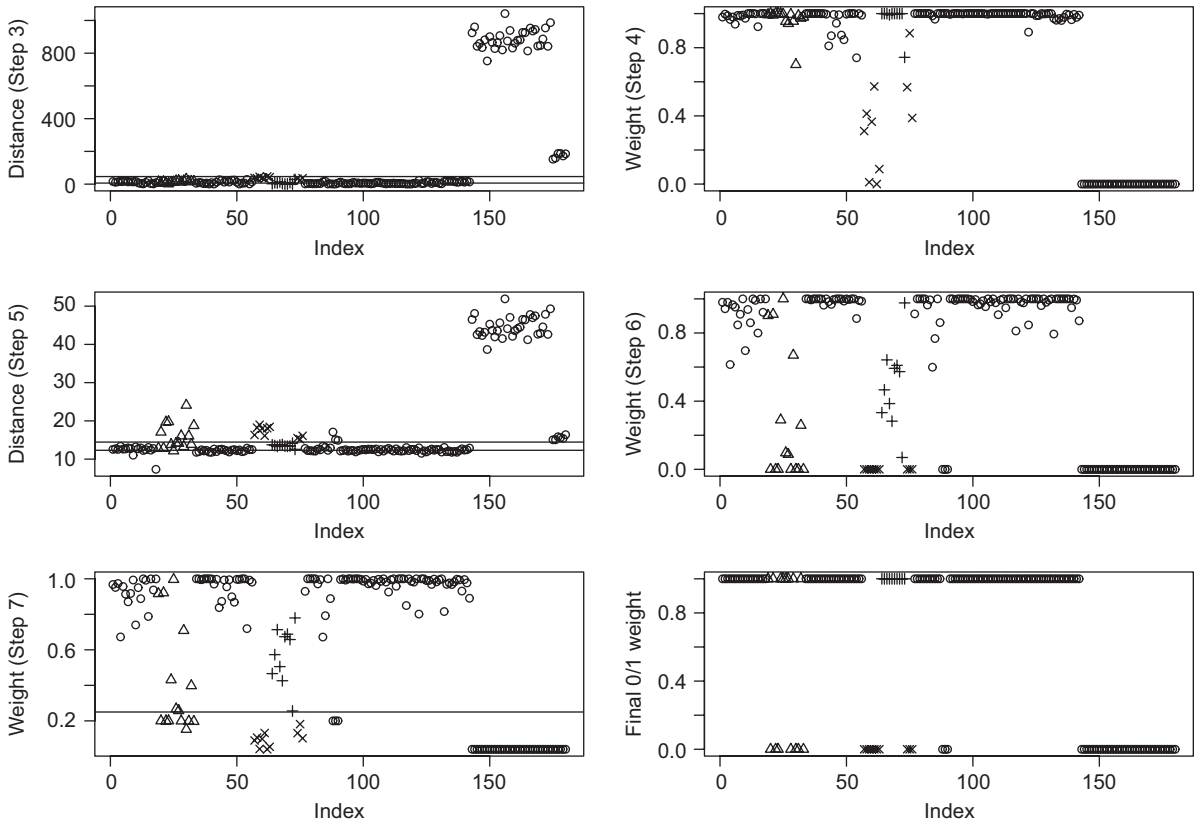

Fig. 8. The panels show the intermediate steps of the PCOut algorithms (distances and weights) for analyzing the EXPMA data of the glass vessels. The plot symbols refer to different types of glass vessels.

these data, identifying a number of important leverage points. These leverage points resulted from different detector efficiencies in the EXPMA analysis. Besides leverage points, there are also four different material compositions of the glass vessels, further increasing the inhomogeneity of the spectral data.

Here we will focus only on outlier detection among the spectra, clearly a high-dimensional data set. First, columns with MAD equal to zero were removed, with the remaining 1905 columns investigated for outliers. Fig. 8 shows the results from PCOut; the different symbols represent the four different materials comprising the vessels. The intermediate graphs provide more insight into the workflow of the method: the kurtosis weights of Step 3 of the PCOut algorithm are shown in the upper left panel, together with the weight boundaries described in Step 4, leading to the weights in the upper right panel. Similarly, the distances from Step 5 and the weights from Step 6 are, respectively, shown in the left and right panel of the second row. The lower left panel shows the combined weights (Phases 1 and 2 combined) together with the outlier boundary 0.25, which results in 0/1 weights (lower right panel). The last 38 observations are clearly visible as multivariate outliers in the intermediate steps of the algorithm; these are the previously mentioned leverage points (see Serneels et al., 2005). We can also see that the vessels produced from different materials have different multivariate behavior than those constituting the main data cloud. Some of these observations are identified as outliers using the default outlier boundary 0.25.

We compare the performance of PCOut on this data set with the Sign method; the other algorithms are not feasible due to the high-dimensionality. Fig. 9 shows the distances (left) and weights (right) as calculated by the Sign method. It is evident that the four types of vessels cannot be detected with this algorithm, although the 38 leverage points can be identified. A possible explanation for the difficulty experienced by the Sign method is a masking effect for PCA. It is evident that PCOut has better performance than the Sign method, which is also evident from the results in Table 1. We infer that PCOut is a competitive outlier detection algorithm regarding detection accuracy as well as computation time.

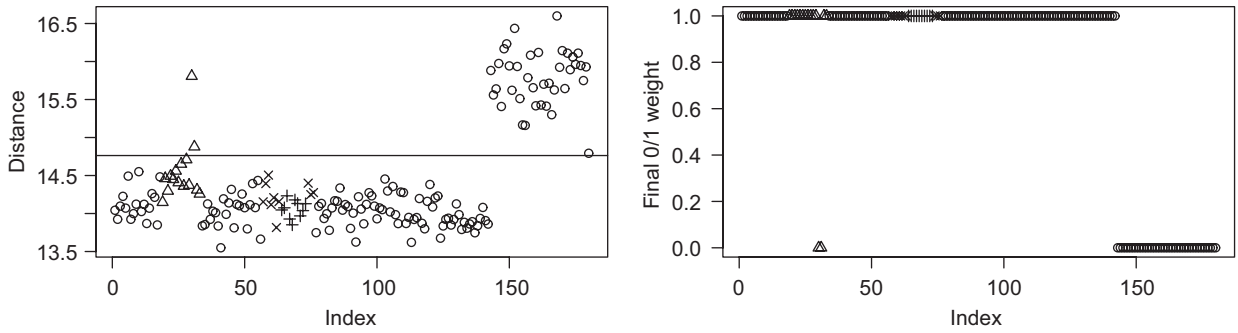

Fig. 9. Distances and weights for analyzing the EXPMA data of the glass vessels with the Sign method. The plot symbols refer to different types of glass vessels.

6. Conclusion

PCOut is a procedure for identifying outliers in multivariate data that utilizes inherent properties of principal components decomposition. It demonstrates very good performance for high-dimensional data and through the use of a robust kurtosis measure, it also obtains very good results for location outliers in any dimension. It is very fast to compute and can easily handle dimensions ranging in the thousands. Thus, it can be extended to fields such as genetics and data mining where computational feasibility of statistical routines has usually been a limiting factor. At lower dimensions, it still produces competitive results when compared to well-known outlier detection methods. Directions of future research including extending the basic algorithm to provide a fast method for detecting outliers among missing data, as well as a fast method for detecting outliers among a mixture of categorical and continuous variables. The R code for the procedure PCOut is available as the function *pcout* in the R library *mvoutlier*.

Acknowledgments

The authors would like to thank anonymous referees for many interesting and helpful comments that led to an improved version of the manuscript.

References

Adrover, J., Yohai, V., 2002. Projection estimates of multivariate location. Ann. Statist. 30, 1760–1781.
Becker, C., Gather, U., 1999. The masking breakdown point of multivariate outliers. J. Amer. Statist. Assoc. 94 (447), 947–955.
Billor, N., Hadi, A., Velleman, P., 2000. BACON: blocked adaptive computationally-efficient outlier nominators. Comput. Statist. Data Anal. 34, 279–298.
Davies, P., 1987. Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices. Ann. Statist. 15 (3), 1269–1292.
Donoho, D., 1982. Breakdown properties of multivariate location estimators. Ph.D. Thesis, Harvard University.
Dudoit, S., Fridlyand, J., Speed, T., 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Statist. Assoc. 97 (457), 77–87.
Filzmoser, P., Garrett, R., Reimann, C., 2005. Multivariate outlier detection in exploration geochemistry. Comput. Geosci. 31, 579–587.
Gnanadesikan, R., Kettenring, J.R., 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics 28, 81–124.
Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.
Hall, P., Marron, J., Neeman, A., 2005. Geometric representation of high dimension, low sample size data. J. Roy. Statist. Soc. Ser. B 67, 427–444.
Hardin, J., Rocke, D., 2005. The distribution of robust distances. J. Comput. Graphical Statist. 14, 928–946.
Hössjer, O., Croux, C., 1995. Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter. J. Nonparametric Statist. 4, 293–308.
Huber, P., 1985. Projection pursuit. Ann. Statist. 13 (2), 435–475 see also discussions by multiple authors following this article.
Hubert, M., Rousseeuw, P., Branden, K.V., 2005. Robpca: a new approach to robust principal components analysis. Technometrics 47, 64–79.
Jackson, J., 1991. A User's Guide to Principal Components. Wiley, New York.
Janssens, K., Deraedt, I., Freddy, A., Veeckman, J., 1998. Composition of 15–17th century archæological glass vessels excavated in Antwerp, Belgium. Mikrochimica Acta 15 (Suppl.), 253–267.

Johnson, R., Wichern, D.W., 1998. Applied Multivariate Statistical Analysis. Fourth ed. Prentice-Hall, Englewood Cliffs, NJ.

Lemberge, P., Raedt, I.D., Janssens, K., Wei, F., Espen, P.V., 2000. Quantitative analysis of 16–17th century archæological glass vessels using PLS regression of EPXMA and μ-XRF data. J. Chemometrics 14, 751–763.

Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., Cohen, K., 1999. Robust principal components for functional data. Test 8, 1–73.

Maronna, R., Yohai, V., 1995. The behavior of the Stahel–Donoho robust multivariate estimator. J. Amer. Statist. Assoc. 90 (429), 330–341.

Maronna, R., Zamar, R., 2002. Robust estimates of location and dispersion for high-dimensional data sets. Technometrics 44 (4), 307–317.

Maronna, R., Martin, R., Yohai, V., 2006. Robust Statistics: Theory and Methods. Wiley, New York.

Peña, D., Prieto, F., 2001. Multivariate outlier detection and robust covariance matrix estimation. Technometrics 43 (3), 286–310.

R Development Core Team, 2005. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL: ⟨http://www.R-project.org⟩.

Reimann, C., Äyräs, M., Chekushin, V., Bogatyrev, I., Boyd, R., Caritat, P.de., Dutter, R., Finne, T., Halleraker, J., Jæger, O., Kashulina, G., Lehto, O., Niskavaara, H., Pavlov, V., Räisänen, M., Strand, T., Volden, T., 1998. Environmental Geochemical Atlas of the Central Barents Region. Geological Survey of Norway (NGU), Geological Survey of Finland (GTK), and Central Kola Expedition (CKE), Special Publication, Trondheim, Espoo, Monchegorsk.

Rocke, D., 1996. Robustness properties of *S*-estimators of multivariate location and shape in high dimension. Ann. Statist. 24 (3), 1327–1345.

Rousseeuw, P., 1985. Multivariate estimation with high breakdown point. In: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (Eds.), Mathematical Statistics and Applications, vol. B. Reidel Publishing Company, Dordrecht, pp. 283–297.

Rousseeuw, P., van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Serneels, S., Croux, C., Filzmoser, P., Espen, P.V., 2005. Partial robust M-regression. Chemometrics and Intelligent Laboratory Systems 79, 55–64.

Stahel, W., 1981. Breakdown of covariance estimators. Research Report 31, Fachgruppe für Statistik, E.T.H. Zürich.

Tenenhaus, M., Vinzi, V.E., Chatelin, Y.-M., Lauro, C., 2005. Pls path modeling. Comput. Statist. Data. Anal. 48(1), 159–205.

Woodruff, D., Rocke, D., 1994. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. J. Amer. Statist. Assoc. 89 (427), 888–896.