Computational Statistics and Data Analysis 53 (2009) 2264–2274

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust PCA for skewed data and its outlier map

Mia Hubert [a], Peter Rousseeuw [b], Tim Verdonck [b,*]

[a] Department of Mathematics, LSTAT, Katholieke Universiteit Leuven, Belgium

[b] Department of Mathematics and Computer Science, University of Antwerp, Belgium

ARTICLE INFO

Article history:
Available online 6 June 2008

ABSTRACT

The outlier sensitivity of classical principal component analysis (PCA) has spurred the development of robust techniques. Existing robust PCA methods like ROBPCA work best if the non-outlying data have an approximately symmetric distribution. When the original variables are skewed, too many points tend to be flagged as outlying. A robust PCA method is developed which is also suitable for skewed data. To flag the outliers a new outlier map is defined. Its performance is illustrated on real data from economics, engineering, and finance, and confirmed by a simulation study.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Principal component analysis is one of the best known techniques of multivariate statistics. It is a dimension reduction technique which transforms the data to a smaller set of variables while retaining as much information as possible. These new variables, called the principal components (PCs), are uncorrelated and maximize variance (information). Once the PCs are computed, all further analysis like cluster analysis, discriminant analysis, regression, ... can be carried out on the transformed data.

When given a data matrix $\boldsymbol{X}$ with n observations and p variables, the PCs $\boldsymbol{t}_i$ are linear combinations of the data $\boldsymbol{t}_i = \boldsymbol{X}\boldsymbol{p}_i$ where

$$\boldsymbol{p}_i = \underset{\boldsymbol{a}}{\operatorname{argmax}} \{\operatorname{var}(\boldsymbol{X}\boldsymbol{a})\}$$

under the constraints

$$\| \boldsymbol{a} \| = 1 \quad \text{and} \quad \boldsymbol{a} \perp \{p_1, \ldots, p_{i-1}\}.$$

From the Lagrange multiplier method it follows that the PCs can be computed as the eigenvectors of an estimate of the covariance matrix.

Classical PCA (CPCA) uses the classical sample covariance matrix for this, but it is well known that this technique is sensitive to outliers. In order to resist these outliers, various robust alternatives have been proposed (see for example Croux and Ruiz-Gazen (2005), Higuchi and Eguchi (2004), Hubert et al. (2002), Locantore et al. (1998) and Maronna (2005)). The robust PCA technique we will focus on is called ROBPCA (Hubert et al., 2005) and will be explained in more detail in Section 2.

Since covariance matrices are intimately connected to ellipsoids, both CPCA and robust PCA methods are typically applied to data that are roughly elliptically symmetric (at least the non-outlying points). This requires the original variables to have roughly symmetric distributions. If this is not the case one often preprocesses the data by transforming the original variables

* Corresponding address: Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, B-2020 Antwerpen, Belgium. Tel.: +32 3 2653896; fax: +32 3 2653777.

E-mail addresses: Mia.Hubert@wis.kuleuven.be (M. Hubert), Peter.Rousseeuw@ua.ac.be (P. Rousseeuw), Tim.Verdonck@ua.ac.be (T. Verdonck).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.05.027

(e.g. using the Box–Cox transform), but these transformed variables may be more difficult to interpret. In such situations it may be useful to analyze the original data by means of a PCA technique that is suitable for asymmetric data. In this paper we propose a modified ROBPCA algorithm which can cope with skewed data, and we define new criteria to flag the outliers. In Section 3 the modifications are described, and the performance of the new algorithm is illustrated on real data (Section 4) and by simulation (Section 5).

2. The ROBPCA method

2.1. Description

The ROBPCA method (Hubert et al., 2005) combines ideas of both projection pursuit (PP) and robust covariance estimation. ROBPCA is very well suited for the analysis of high-dimensional data and has a.o. been applied to multivariate calibration and classification (Chiang and Colegrove, 2007; Hubert and Engelen, 2004; Hubert and Vanden Branden, 2003; Hubert and Verboven, 2003; Vanden Branden and Hubert, 2005). The algorithm consists of the following steps:

(1) Perform a singular value decomposition to restrict the observations to the space they span. This is not absolutely necessary, but when the number of variables vastly exceeds the number of observations it already yields a huge dimension reduction without losing information. (We will denote the resulting data set again as X in what follows.)
(2) Choose the coverage $\frac{1}{2} < \alpha < 1$ (the default is $\alpha = 0.75$). Set $h = [\alpha n]$ (where $[.]$ denotes the integer part). The choice of the coverage α determines the robustness as well as the efficiency of the method. The smaller α the more robust ROBPCA will be, but the less accurate.
(3) In every data point $\boldsymbol{x}_i$ (a row of $\boldsymbol{X}$), calculate the outlyingness introduced by Stahel (1981) and Donoho and Gasko (1992):

$$\mathrm{outl}(\boldsymbol{x}_i, \boldsymbol{X}) = \sup_{\boldsymbol{v} \in B} \frac{|\boldsymbol{x}_i'\boldsymbol{v} - m(\boldsymbol{x}_j'\boldsymbol{v})|}{s(\boldsymbol{x}_j'\boldsymbol{v})} \tag{1}$$

where $m(.)$ and $s(.)$ are the robust univariate Minimum Covariance Determinant (MCD) estimators of location and scale (Rousseeuw, 1984) and B is a set of 250 random directions through two data points. Then we take the set I_h of the h data points with smallest outlyingness, and compute their mean and covariance matrix:

$$\hat{\boldsymbol{\mu}}_0(\boldsymbol{X}) = \frac{1}{h} \sum_{i \in I_h} \boldsymbol{x}_i$$

$$\hat{\boldsymbol{\Sigma}}_0(\boldsymbol{X}) = \frac{1}{h-1} \sum_{i \in I_h} \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_0(\boldsymbol{X})\right) \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_0(\boldsymbol{X})\right)'.$$

(4) Reduce the dimension by projecting all data on the k-dimensional subspace V_0 spanned by the first k eigenvectors of the robust covariance estimator $\hat{\boldsymbol{\Sigma}}_0$ obtained in step 3. The choice of k can be made by looking at a scree plot of the eigenvalues, at the percentage of variance explained, or by a robust PRESS algorithm (Hubert and Engelen, 2007).
(5) For each observation, compute its orthogonal distance

$$\mathrm{OD}_i^{(0)} = \left\| \boldsymbol{x}_i - \hat{\boldsymbol{x}}_i \right\| \tag{2}$$

with $\hat{\boldsymbol{x}}_i$ the projection of $\boldsymbol{x}_i$ on the subspace V_0.

We then obtain an improved robust subspace estimate V_1 as the subspace spanned by the k dominant eigenvectors of $\hat{\boldsymbol{\Sigma}}_1$, which is the covariance matrix of all observations $\boldsymbol{x}_i$ for which $\mathrm{OD}_i^{(0)} \leq c_{OD}$ (Engelen et al., 2005). The cutoff value c_{OD} is difficult to determine because the distribution of the orthogonal distances is not known exactly. In ROBPCA this is resolved by the fact that the orthogonal distances to the power $\frac{2}{3}$ are approximately normally distributed, which can be derived from Box (1954) and Nomikos and MacGregor (1995). The cutoff value can then be determined as $c_{OD} = (\hat{\mu} + \hat{\sigma} z_{0.975})^{\frac{3}{2}}$ where $\hat{\mu}$ and $\hat{\sigma}$ are estimated by the univariate MCD and $z_{0.975}$ is the 97.5% quantile of the gaussian distribution.

Next, we project all data points on the subspace V_1.
(6) Compute a robust center and covariance matrix in this k-dimensional subspace by applying the reweighted MCD estimator (Rousseeuw and Van Driessen, 1999) to the projected data. The final principal components are the eigenvectors of this robust covariance matrix. (The MCD estimator searches for the subset of size h whose classical covariance matrix has minimal determinant. The MCD location and scatter are given by the mean and covariance matrix of that subset. After this a reweighting step is performed to increase the finite-sample efficiency. We use the fastMCD algorithm developed in Rousseeuw and Van Driessen (1999).)

Note that it is possible to stop the program without performing step 6. This reduces the computation time considerably, and still yields the same PCA subspace as the full ROBPCA, but less robust eigenvectors and eigenvalues.

When there are at least five times more observations than variables in the data and the number of variables is not too high, the ROBPCA program simply computes the reweighted MCD estimator on the original data and reports the resulting eigenvectors.

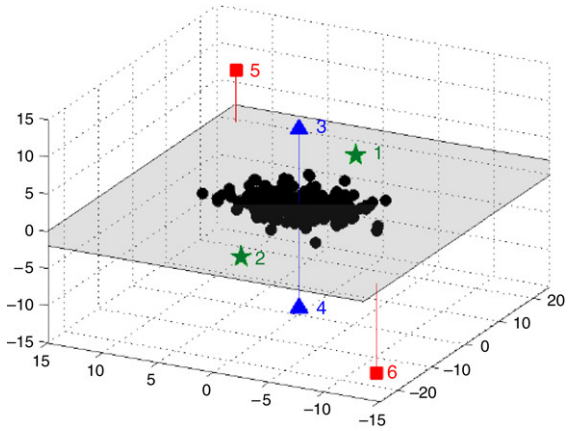

Fig. 1. Different types of PCA outliers.

2.2. Outlier map

Apart from computing principal components, ROBPCA also flags the outliers. In general, an outlier is an observation which does not obey the pattern of the majority of the data. In the context of PCA three types of outliers can be distinguished:

(1) good leverage points
(2) orthogonal outliers
(3) bad leverage points.

In Fig. 1 we see the different types of outliers in a three-dimensional data set that is projected on a robust two-dimensional PCA subspace. The data set contains 200 observations and was generated from the multivariate normal distribution with center $\boldsymbol{\mu} = [0, 0, 0]'$ and covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}\,[12, 8, 0.001]$. The first 6 observations have been replaced by different types of outliers.

Good leverage points (like observations 1 and 2) lie close to the PCA subspace but far from the regular observations, whereas orthogonal outliers (observations 3 and 4) have a large orthogonal distance to the PCA subspace while their projection on the PCA space is inlying. Bad leverage points (observations 5 and 6) have a large orthogonal distance and their projection on the PCA space is far from the regular data.

To be precise, the *orthogonal distance*, as in (2), is the distance between an observation and its projection in the k-dimensional subspace V_1:

$$\mathrm{OD}_i = \left\| \boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_x - \mathbf{P}_{p,k}\boldsymbol{t}_i \right\|$$

where $\mathbf{P}_{p,k}$ is the loading matrix with orthogonal columns (the eigenvectors), $\hat{\boldsymbol{\mu}}_x$ is the robust center, and $\boldsymbol{t}_i = \mathbf{P}'_{p,k}(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_x)$ are the robust scores. The *robust score distance* is given by

$$\mathrm{SD}_i = \sqrt{\sum_{j=1}^{k} \frac{t_{ij}^2}{l_j}}$$

where l_j are the sorted eigenvalues of the MCD scatter matrix obtained in step 6 of the algorithm.

The outlier map plots the orthogonal distances versus the score distances. Lines are drawn to distinguish between observations with a small and a large OD, and between a small and a large SD. For the orthogonal distances, the cutoff is defined as in step 5 of the algorithm. For the score distances, the cutoff value $c_{SD} = \sqrt{\chi^2_{k,0.975}}$ is used.

In the outlier map (Fig. 2) of the simulated dataset we see that ROBPCA has flagged the outliers correctly.

3. The modified algorithm for skewed data

3.1. Description

Here we propose a more general robust PCA method that can also cope with skewed data. Our approach consists of the same steps as the ROBPCA algorithm (as described in Section 2.1), but we have made three modifications.

The first is that the Stahel–Donoho outlyingness (1) in step 3 of the algorithm is replaced by a new measure, called the *adjusted outlyingness*. It is based on the adjusted boxplot (Hubert and Vandervieren, in press) and has previously been

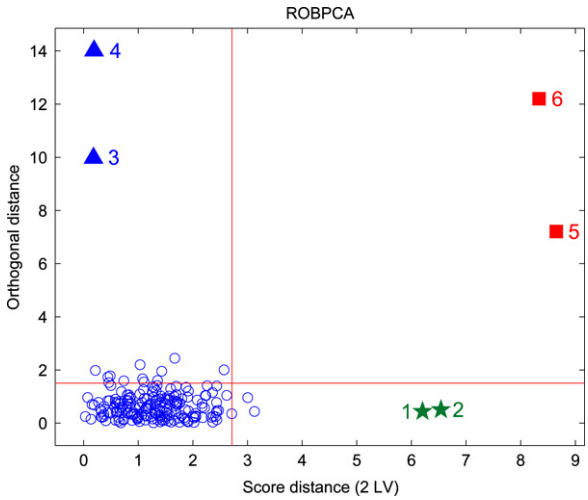

Fig. 2. ROBPCA outlier map of the data in Fig. 1.

used to robustify Independent Component Analysis (Brys et al., 2005). The good properties of this measure are shown in Hubert and Van der Veeken (2008), where it has already been extensively studied. The adjusted outlyingness has a different denominator in order to flag fewer data points at skewed distributions. It is defined as:

$$\mathrm{AO}_i = \max_{\boldsymbol{v}\in B} \frac{|\boldsymbol{x}_i'\boldsymbol{v} - \mathrm{med}(\boldsymbol{x}_j'\boldsymbol{v})|}{(c_2(\boldsymbol{v}) - \mathrm{med}(\boldsymbol{x}_j'\boldsymbol{v}))I[\boldsymbol{x}_i'\boldsymbol{v} > \mathrm{med}(\boldsymbol{x}_j'\boldsymbol{v})] + (\mathrm{med}(\boldsymbol{x}_j'\boldsymbol{v}) - c_1(\boldsymbol{v}))I[\boldsymbol{x}_i'\boldsymbol{v} < \mathrm{med}(\boldsymbol{x}_j'\boldsymbol{v})]} \tag{3}$$

where c_1 corresponds to the smallest observation which is greater than $Q_1 - 1.5\mathrm{e}^{-4\mathrm{MC}}\mathrm{IQR}$ and c_2 corresponds to the largest observation which is smaller than $Q_3 + 1.5\mathrm{e}^{3\mathrm{MC}}\mathrm{IQR}$. Here Q_1 and Q_3 are the first and third quartile of the projected data, $\mathrm{IQR} = Q_3 - Q_1$ and MC is the medcouple (Brys et al., 2004), a robust measure of skewness. Formula (3) assumes that $\mathrm{MC} \geq 0$, otherwise we replace $\boldsymbol{v}$ by $-\boldsymbol{v}$.

As in (1), B is a set of random directions through two data points. For skewed distributions it is useful to increase the size of B. More directions make the result more accurate, but also increase the computation time.

The second adjustment concerns the cutoff value for the orthogonal distances OD in step 5. We now use as cutoff value the largest OD_i smaller than $Q_3(\{\mathrm{OD}\}) + 1.5\mathrm{e}^{3\mathrm{MC}(\{\mathrm{OD}\})}\mathrm{IQR}(\{\mathrm{OD}\})$. Should the medcouple of the orthogonal distances be negative (this only happens in exceptional cases), we will take as cutoff the largest OD_i smaller than $Q_3(\{\mathrm{OD}\}) + 1.5\mathrm{IQR}(\{\mathrm{OD}\})$.By doing this the data no longer have to be transformed and the cutoff value is now dependent on the data itself instead of on some breakdown value. Taking the cutoff dependent on the breakdown value and setting this for example to 25% whereas only 5% of the data are outlying results in using too few data points in the computation, leading to less efficiency.

The third modification occurs in step 6. Instead of applying the reweighted MCD estimator, we calculate the adjusted outlyingness in the k-dimensional subspace V_1 and compute the mean and covariance matrix of the h points with the lowest adjusted outlyingness. On the horizontal axis of the outlier map we now plot the adjusted outlyingness of each observation, and the corresponding cutoff value is derived in the same way as for the orthogonal distances.

Note that the skewness-adjusted version of ROBPCA no longer uses cutoff values derived from quantiles of theoretical distributions, but instead employs adaptive cutoff values obtained from the empirical data at hand.

When we return to the data of Fig. 1 and apply the modified ROBPCA algorithm to it, we obtain the outlier map in Fig. 3. Note that the outliers are still classified correctly. Indeed, when the bulk of the data are not skewed, the modified algorithm gives similar results to the original one. Moreover, we see that the outlier cutoff lines no longer flag several regular observations as outliers.

4. Examples

In this section the regular ROBPCA method based on the Stahel–Donoho outlyingness (denoted as ROBPCA-SD) and our new approach based on the adjusted outlyingness (denoted as ROBPCA-AO) are compared with several real data sets. When there are at least five times more observations than variables in the data, we will also compute the eigenvectors of the reweighted MCD estimator (this approach is denoted as ROBPCA-MCD).

In each example we first robustly center and scale the variables by subtracting the median and dividing by the median absolute deviation. Unless otherwise mentioned, the number of random directions to search over is 1000 in both ROBPCA-AO and ROBPCA-SD.

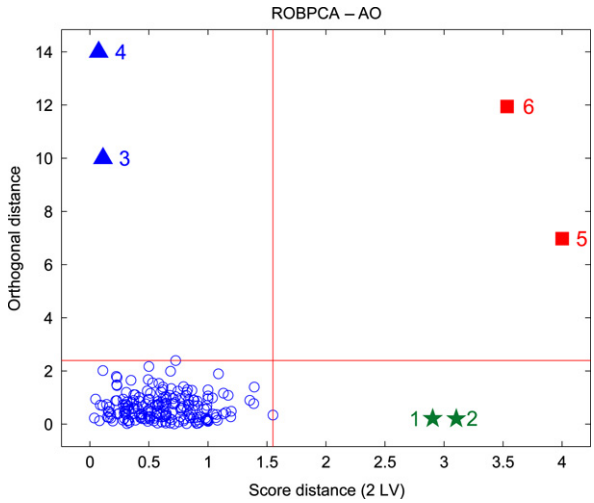

Fig. 3. Outlier map of Fig. 1 using the skewness-adjusted ROBPCA method.

Table 1
CES data with MCs and their *p*-values

Variable	Description	MC	*p*-value
EXP	Total household expenditure	0.21	<0.00001
FDHO	Food consumed at home	0.17	<0.00001
FDAW	Food consumed away from home	0.32	<0.00001
SHEL	Housing and household equipment	0.22	<0.00001
TELE	Telephone services	0.33	<0.00001
CLOT	Clothing	0.27	<0.00001
HEAL	Health care	0.24	<0.00001
ENT	Entertainment	0.37	<0.00001

4.1. Consumer expenditure survey data

This dataset stems from the Consumer Expenditure Survey (CES) of 1995 collected by the US Department of Labor, available at http://econ.lse.ac.uk/courses/ec220/G/iedata/ces/. The data contains 869 households with 8 variables described in Table 1. It also shows the skewness of each variable as measured by its medcouple. The corresponding *p*-values are all smaller than 0.00001 (using the formulas in Brys et al. (2004) for testing whether the MC is significantly different from zero). It follows that the variables are all significantly asymmetric.

Because the CES data contain missing values, we have used the expectation robust (ER) approach of Serneels and Verdonck (2008) for both ROBPCA-SD as well as ROBPCA-AO. We have decided to retain 5 components, which together explain 88%. The corresponding outlier maps are in Fig. 4.

Note that ROBPCA-SD and ROBPCA-MCD flagged 198 and 190 observations as outlying whereas ROBPCA-AO flagged only 24, which is more realistic since the data are so skewed.

4.2. Computer hardware data

The Computer Hardware (CH) data of Ein-Dor and Feldmesser are available at http://archive.ics.uci.edu/beta/datasets/Computer+Hardware.

The data set contains 209 observations, and we focus on the 8 variables in Table 2. We see that each of these variables is significantly skewed. We retained 3 components, which together explain 87%. The resulting outlier maps are given in Fig. 5. ROBPCA-SD and ROBPCA-MCD flagged 70 and 63 observations whereas ROBPCA-AO only flagged 6, which seems more reasonable.

4.3. Credit default swap (CDS) data

The CDS data were obtained from iTraxx Europe. It contains the price of credit default swaps of 125 companies over 58 weeks (Cariboni, 2007). A typical preprocessing step for financial time series is to transform the data by taking the log ratios $\log(x_{i,j}/x_{i,j-1}) = \log(x_{i,j}) - \log(x_{i,j-1})$ for every observation $\boldsymbol{x}_i$ where i is the company and j is the week. After this transformation there were some columns containing more than 125/2 zeroes, meaning that the credit default swap price of

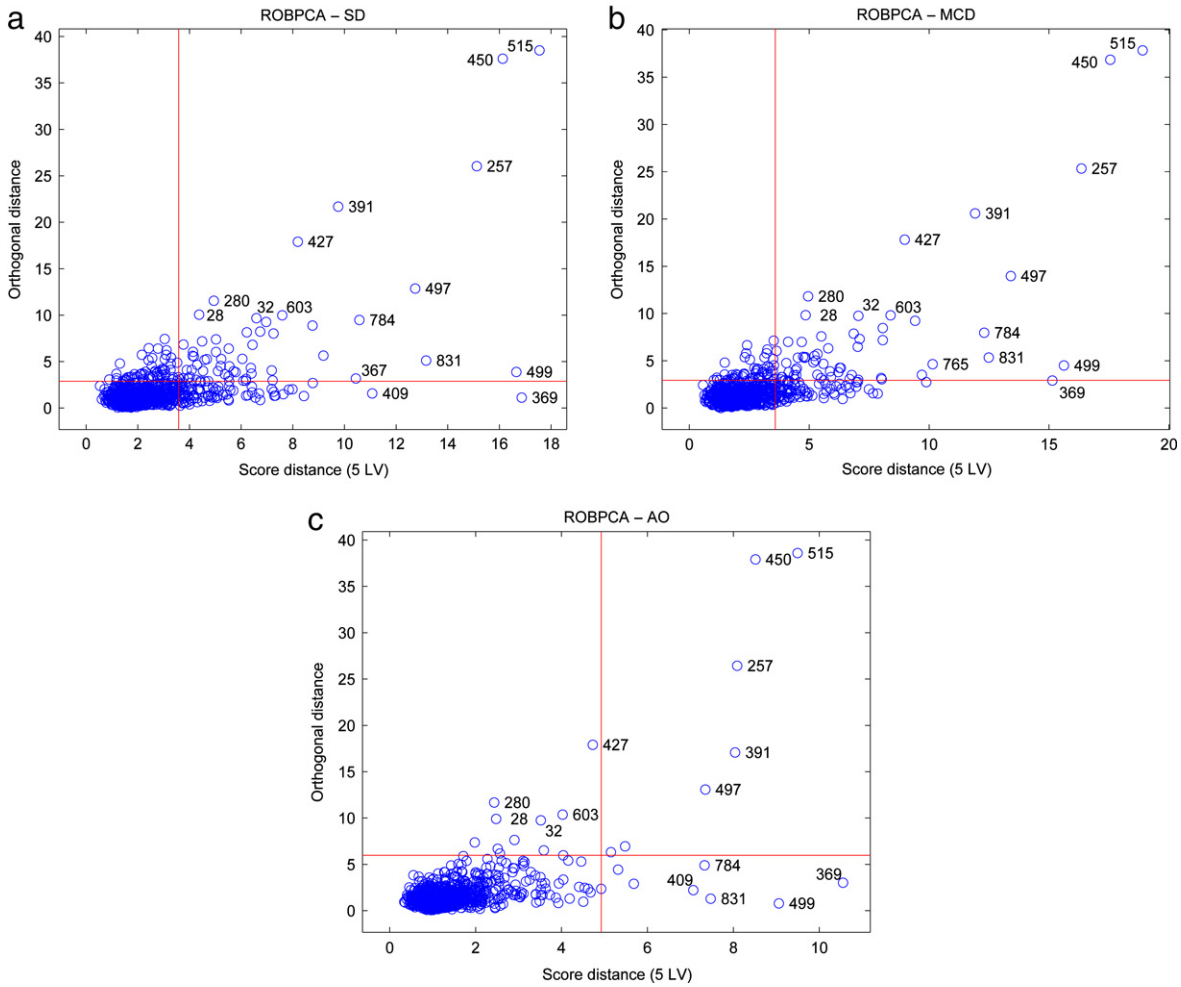

Fig. 4. Outlier maps of the CES data based on (a) ROBPCA-SD, (b) ROBPCA-MCD, and (c) ROBPCA-AO.

Table 2
Computer hardware data with MCs and their *p*-values

Variable	description	MC	*p*-value
MYCT	Machine cycle time in nanoseconds	0.42	<0.00001
MMIN	Minimum main memory in kilobytes	0.27	0.00057
MMAX	Maximum main memory in kilobytes	0.33	0.00002
CACH	Cache memory in kilobytes	0.60	<0.00001
CHMIN	Minimum channels in units	0.60	<0.00001
CHMAX	Maximum channels in units	0.60	<0.00001
PRP	Published relative performance	0.53	<0.00001
ERP	Estimated relative performance	0.57	<0.00001

more than half of the companies did not change during that week. We have left these variables out of the analysis, yielding a dataset of 125 companies and 37 log ratios.

Because the number of variables is quite high in this example, we have allowed both ROBPCA-SD and ROBPCA-AO to compute their outlyingness using 5000 random directions through two data points. In the resulting analysis we decided to retain 10 principal components, which explain 67%. The corresponding outlier maps are given in Fig. 6.

Looking at the outlier map of the ROBPCA-AO method one would flag the companies *Hilton*, *CarltonComm*, *BAA*, *VNU*, *TDCAS*, *Altadis* and *Rentokil*. In practice, researchers may not be pleased with discarding several observations out of their analysis. On the other hand it is possible that the flagged companies are only atypical during a few weeks. One possible way to tackle this problem is by doing the following steps:

(1) Plot the outlying companies to see whether they indeed have atypical values for only a few weeks and are not outlying throughout the whole period.

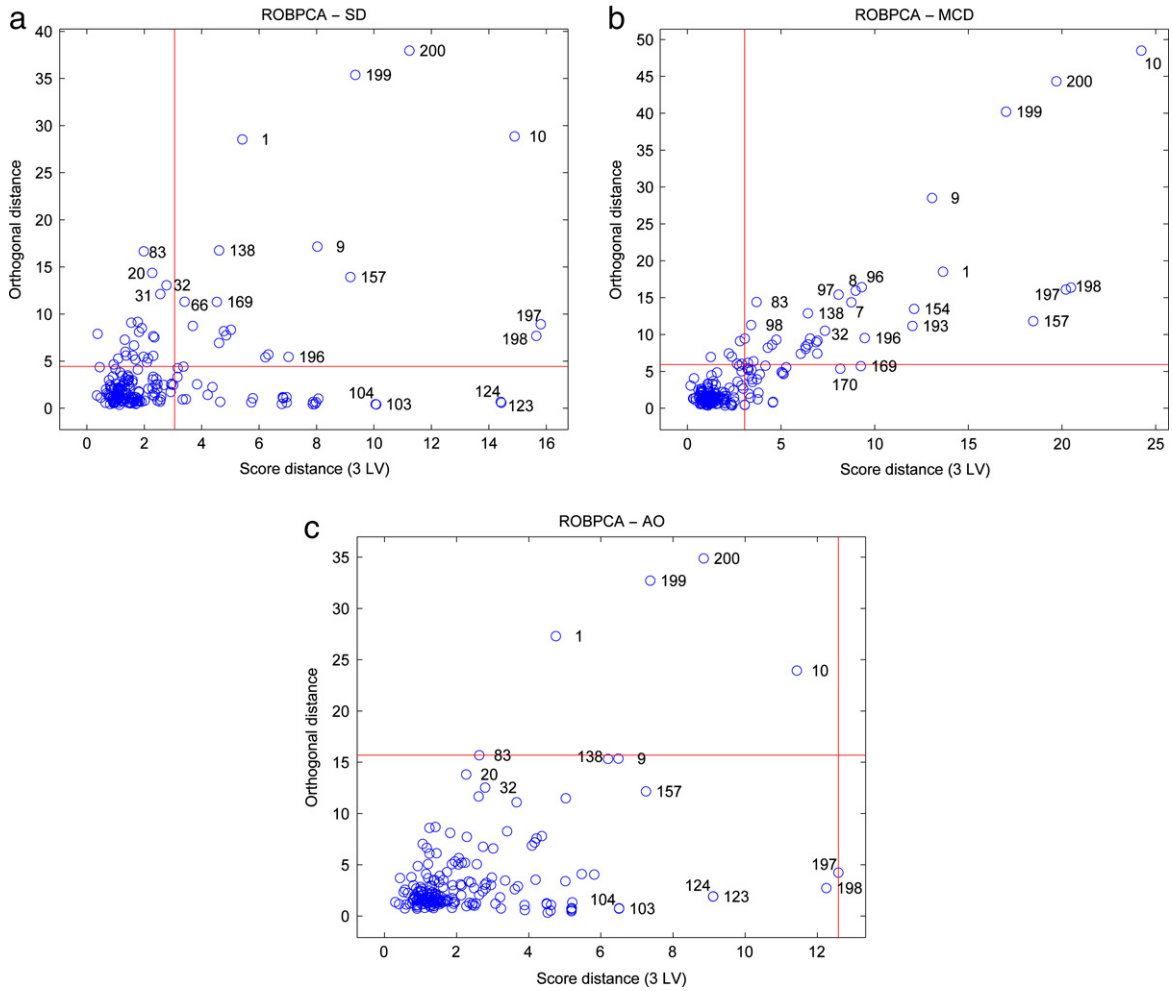

Fig. 5. Outlier maps of computer hardware data based on (a) ROBPCA-SD, (b) ROBPCA-MCD, and (c) ROBPCA-AO.

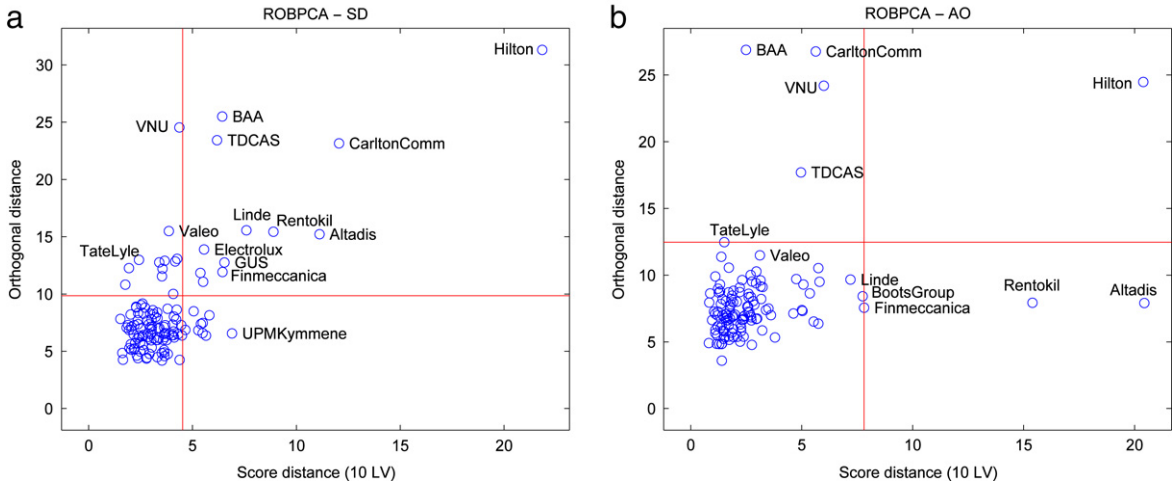

Fig. 6. Outlier map of CDS data based on (a) ROBPCA-SD and (b) ROBPCA-AO.

(2) Find the outlying columns for every company by means of the adjusted boxplot.
(3) Replace outliers in each company by missing elements (NaN).
(4) Perform ROBPCA-AO by the ER approach of Serneels and Verdonck (2008).

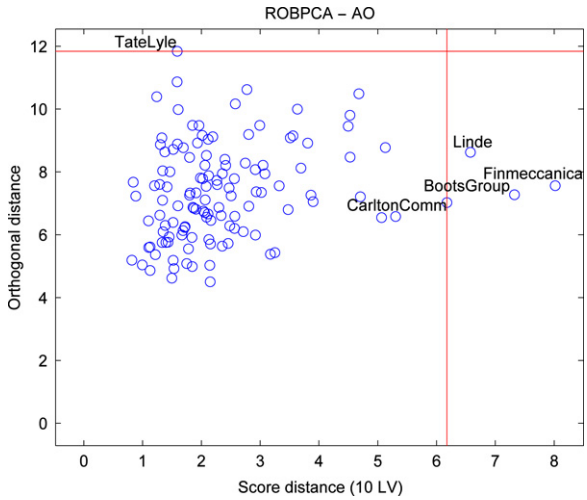

Fig. 7. Outlier map of the CDS data after adjusting the outlying companies.

By doing this it is possible to continue the analysis on the whole dataset, which is of course slightly adjusted (only a couple of weeks of the outlying companies), but still contains all the observed companies. Applying ROBPCA-AO again on this dataset gives the outlier map in Fig. 7, which confirms that the outlying behavior of the flagged companies was only due to a few weeks. By following these steps we were able to identify and to adjust the variables that lead to outlyingness.

5. Simulations

In this section we will compare the three algorithms on simulated data.

5.1. Simulation design

A straightforward way to set up a simulation for PCA is to construct data with a given complexity k as $\boldsymbol{X} = \boldsymbol{T}_k\boldsymbol{P}_k' + \boldsymbol{E}$, i.e. $\boldsymbol{X}$ is a matrix of rank k plus unstructured noise.
The orthogonal loadings $\boldsymbol{P}_k$ are taken as $\boldsymbol{P}_k = \boldsymbol{I}_{p\times k} = \delta_{ij}$ and the principal components $\boldsymbol{T}_k$ are generated as n draws from the k-variate normal inverse gaussian distribution NIG$(\gamma, \delta, \mu, \sigma)$ as defined in Barndorff-Nielsen (1997). Here μ is a location parameter, σ is a scale parameter, and γ and δ are shape parameters determining skewness and tail weight.

In our simulations we generate n observations with k variables drawn from NIG$(1, 0, 0, 1)$ and NIG$(1, 0.8, 0, 1)$. The first distribution is the standard gaussian, whereas the second distribution is skewed. We then construct the uncontaminated $n \times p$ datamatrix $\boldsymbol{X} = \boldsymbol{T}_k\boldsymbol{P}_k' + \boldsymbol{E}$ where $\boldsymbol{E}$ is gaussian noise with standard deviation 0.01.

To contaminate the data we then replace the last $100\epsilon\%$ of observations of $\boldsymbol{X}$ by outliers. We have added bad leverage points by placing the center at $\left[\eta \times \mathbf{1}_{k+1}', \mathbf{0}_{p-k-1}'\right]$, and orthogonal outliers by placing the center at $\left[\mathbf{0}_k', \eta, \mathbf{0}_{p-k-1}'\right]$, with location parameter $\eta \neq 0$ (good leverage points would not alter the PCA subspace). In our study we have considered $\eta = -8$ and 12. The outliers are generated from the multivariate normal distribution with that center and with covariance matrix $\kappa\,\mathrm{Cov}(\boldsymbol{X})$. The contamination was either concentrated ($\kappa = 0.01$) or had the same covariance structure as the regular observations ($\kappa = 1$).

In a simulation study one needs performance measures in order to compare the different techniques. We evaluated each method by considering the following measures:

- the angle between the subspace spanned by the columns of the real $\boldsymbol{P}_k = \boldsymbol{I}_{p\times k}$ and the estimated PCA subspace, which is spanned by the columns of $\hat{\boldsymbol{P}}_k$. This angle was given by Krzanowski (1979) as $\arccos(\sqrt{\lambda_k})$ with λ_k the smallest eigenvalue of $\boldsymbol{P}_k'\hat{\boldsymbol{P}}_k\hat{\boldsymbol{P}}_k'\boldsymbol{P}_k$. The optimal angle is 0.
- ND = number of unflagged outliers (its optimal value is 0).
- WD = number of flagged observations minus number of generated outliers (its optimal value is 0).

We conducted simulations of ROBPCA-MCD, ROBPCA-SD, and ROBPCA-AO and compared the results also with classical PCA (CPCA). Because all simulations were performed on data sets containing 0%, 5%, 10%, or 15% of outliers, the coverage parameter α of each robust method was set to 85%. The simulations were run in MATLAB 7.2 (The MathWorks, Natick, MA).

Table 3
Simulation results for skewed data of size 500 × 10 with $k = 2$, contaminated by bad leverage points with $\eta = -8$

ϵ (%)	Measure	CPCA		ROBPCA-MCD		ROBPCA-SD		ROBPCA-AO	
		$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$
0	Angle	0.0076	0.0071	0.0123	0.0118	0.0083	0.0079	0.0078	0.0072
	ND	0	0	0	0	0	0	0	0
	WD	39	38	76	77	83	83	13	14
5	Angle	0.4087	0.4039	0.0109	0.0105	0.0078	0.0073	0.0074	0.0068
	ND	0	0	0	0	0	0	0	0
	WD	23	21	65	65	66	67	5	4
10	Angle	0.4722	0.4668	0.0098	0.0098	0.0076	0.0081	0.0073	0.0076
	ND	3	8	0	0	0	0	0	0
	WD	21	14	51	53	50	51	0	0
15	Angle	0.5010	0.4932	0.0079	0.0081	0.1723	0.1385	0.0082	0.0083
	ND	75	52	0	0	23	4	0	0
	WD	−50	−31	33	33	19	29	0	0

Table 4
Simulation results for skewed data of size 500 × 10 with $k = 2$, contaminated by orthogonal outliers with $\eta = -8$

ϵ (%)	Measure	CPCA		ROBPCA-MCD		ROBPCA-SD		ROBPCA-AO	
		$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$
0	Angle	0.0076	0.0071	0.0123	0.0118	0.0083	0.0079	0.0078	0.0072
	ND	0	0	0	0	0	0	0	0
	WD	39	38	76	77	83	83	13	14
5	Angle	0.4456	0.3993	0.0109	0.0105	0.0079	0.0072	0.0074	0.0068
	ND	0	0	0	0	0	0	0	0
	WD	25	24	65	65	82	74	10	9
10	Angle	1.1135	1.0577	0.0098	0.0098	0.0076	0.0080	0.0073	0.0076
	ND	12	8	0	0	0	0	0	0
	WD	16	18	51	53	79	67	7	6
15	Angle	1.2729	1.2553	0.0079	0.0081	0.0082	0.0081	0.0079	0.0081
	ND	75	54	0	0	0	0	0	0
	WD	−44	−29	33	33	80	58	5	5

5.2. Simulation results

Table 3 shows the simulation results for skewed data [generated from NIG(1, 0.8, 0, 1)] with $n = 500$ points in $p = 10$ dimensions and rank $k = 2$, contaminated by bad leverage points with $\eta = -8$. For each contamination percentage ϵ, each of the four PCA methods, and each κ it shows the average angle over 50 runs, as well as the average ND and WD.

As expected, classical PCA yields the best results if there are no outliers in the data ($\epsilon = 0\%$), but breaks down as soon as there is contamination.

We see that even with contamination, ROBPCA-AO has the same angle that CPCA has without contamination. Also ROBPCA-MCD and ROBPCA-SD perform well in this regard, but from their WD we see that they flag too many non-outliers.

Similar results are obtained for orthogonal outliers (Table 4). Table 5 repeats Table 3 for much higher-dimensional data sets ($n = 750$, $p = 200$, and $k = 4$), still with the same qualitative conclusions. (Note that ROBPCA-MCD cannot be computed in this case since $n/p < 5$). And finally, Table 6 redoes the experiment of Table 3 when the uncontaminated data are not skewed (this time they are generated from the standard gaussian distribution). Also here ROBPCA-AO yields very satisfactory results.

We ran many more simulations but their results were similar, hence they are not reported here.

6. Summary and conclusions

In this paper we have developed a robust PCA method that can cope with skewed data. It combines the ROBPCA method for symmetric data with the adjusted boxplot for skewed data. Also outlier cutoff values are derived from the adjusted boxplot outlier rule. Applying the new approach to real and simulated data reveals that the method estimates the PCA subspace accurately and provides good classification between regular observations and outliers. Matlab software to apply ROBPCA-AO is available as part of LIBRA: Library for Robust Analysis (Verboven and Hubert, 2005).

Table 5
Simulation results for skewed data of size 750 × 200 with $k = 4$, contaminated by bad leverage points with $\eta = -8$

ϵ (%)	Measure	CPCA		ROBPCA-SD		ROBPCA-AO	
		$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$
0	Angle	0.0279	0.0278	0.0307	0.0307	0.0286	0.0287
	ND	0	0	0	0	0	0
	WD	80	80	146	146	10	10
5	Angle	0.3181	0.3178	0.0307	0.0311	0.0289	0.0294
	ND	0	0	0	0	0	0
	WD	59	55	122	120	3	3
10	Angle	0.3430	0.3394	0.0310	0.0310	0.0295	0.0293
	ND	75	40	0	0	0	0
	WD	−19	−13	95	95	0	0
15	Angle	0.3530	0.3496	0.0587	0.0369	0.0339	0.0302
	ND	113	92	9	0	0	0
	WD	−57	−41	121	83	1	1

Table 6
Simulation results for gaussian data of size 500 × 10 with $k = 2$, contaminated by bad leverage points with $\eta = -8$

ϵ (%)	Measure	CPCA		ROBPCA-MCD		ROBPCA-SD		ROBPCA-AO	
		$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$	$\kappa = 0.01$	$\kappa = 1$
0	Angle	0.0155	0.0148	0.0192	0.0174	0.0163	0.0155	0.0157	0.0148
	ND	0	0	0	0	0	0	0	0
	WD	35	37	52	53	58	60	6	6
5	Angle	0.5713	0.5689	0.0175	0.0180	0.0162	0.0156	0.0155	0.0150
	ND	0	0	0	0	0	0	0	0
	WD	29	26	44	43	48	46	2	2
10	Angle	0.5936	0.5915	0.0167	0.0168	0.0159	0.0160	0.0153	0.0158
	ND	0	1	0	0	0	0	0	0
	WD	31	26	38	38	40	39	0	0
15	Angle	0.6026	0.6015	0.0167	0.0169	0.0172	0.0171	0.0166	0.0169
	ND	75	57	0	0	0	0	0	0
	WD	−44	−31	30	31	30	30	0	0

References

Barndorff-Nielsen, O., 1997. Normal inverse gaussian distributions and stochastic volatility modeling. Scandinavian Journal of Statistics 24, 1–13.

Box, G.E.P., 1954. Some theorems on quadratic forms applied in the study of analysis of variance problems: Effect of inequality of variance in one-way classification. The Annals of Mathematical Statistics 25, 290–302.

Brys, G., Hubert, M., Struyf, A., 2004. A robust measure of skewness. Journal of Computational and Graphical Statistics 13, 996–1017.

Brys, G., Hubert, M., Rousseeuw, P.J., 2005. A robustification of independent component analysis. Journal of Chemometrics 19, 364–375.

Cariboni, J., 2007. Credit derivatives pricing under Lévy models. Ph.D. Thesis, Katholieke Universiteit Leuven. http://hdl.handle.net/1979/853.

Chiang, L.H., Colegrove, L.F., 2007. Industrial implementation of on-line multivariate quality control. Computational Statistics and Data Analysis 88, 143–153.

Croux, C., Ruiz-Gazen, A., 2005. High breakdown estimators for principal components: The projection-pursuit approach revisited. Journal of Multivariate Analysis 95, 206–226.

Donoho, D.L., Gasko, M., 1992. Breakdown properties of location estimates based on half-space depth and projected outlyingness. Annals of Statistics 20, 1803–1827.

Engelen, S., Hubert, M., Vanden Branden, K., 2005. A comparison of three procedures for robust PCA in high dimensions. Austrian Journal of Statistics 34, 117–126.

Higuchi, I., Eguchi, S., 2004. Robust principal component analysis with adaptive selection for tuning parameters. Journal of Machine Learning Research 5, 453–471.

Hubert, M., Engelen, S., 2004. Robust PCA and classification in biosciences. Bioinformatics 20, 1728–1736.

Hubert, M., Engelen, S., 2007. Fast cross-validation of high-breakdown resampling methods for PCA. Computational Statistics and Data Analysis 51, 5013–5024.

Hubert, M., Rousseeuw, P.J., Vanden Branden, K., 2005. ROBPCA: A new approach to robust principal component analysis. Technometrics 47, 64–79.

Hubert, M., Rousseeuw, P.J., Verboven, S., 2002. A fast method for robust principal components with applications to chemometrics. Chemometrics and Intelligent Laboratary Systems 60, 101–111.

Hubert, M., Vanden Branden, K., 2003. Robust methods for Partial Least Squares regression. Journal of Chemometrics 17, 537–549.

Hubert, M., Vandervieren, E., 2007. An adjusted boxplot for skewed distributions. Computational Statistics and Data Analysis, in press (doi:10.1016/j.csda.2007.11.008).

Hubert, M., Van der Veeken, S., 2008. Outlier detection for skewed data. Journal of Chemometrics 22, 235–246.

Hubert, M., Verboven, S., 2003. A robust PCR method for high-dimensional regressors. Journal of Chemometrics 17, 438–452.

Krzanowski, W.J., 1979. Between-groups comparison of principal components. Journal of the American Statistical Association 74, 703–707.

Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., Cohen, K.L., 1998. Principal component analysis for functional data. Test 8, 1–73.

Maronna, R., 2005. Principal components and orthogonal regression based on robust scales. Technometrics 47, 264–273.

Nomikos, P., MacGregor, J.F., 1995. Multivariate SPC charts for monitoring batch processes. Technometrics 37, 41–59.

Rousseeuw, P.J., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 871–880.

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

Serneels, S., Verdonck, T., 2008. Principal component analysis for data containing outliers and missing elements. Computational Statistics and Data Analysis 52 (3), 1712–1727.
Stahel, W.A., 1981. Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. Thesis, ETH Zürich.
Vanden Branden, K., Hubert, M., 2005. Robust classification in high dimensions based on the SIMCA method. Chemometrics and Intelligent Laboratory Systems 79, 10–21.
Verboven, S., Hubert, M., 2005. LIBRA: A Matlab library for robust analysis. Chemometrics and Intelligent Laboratory Systems 75, 127–136.