Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemometrics and Intelligent Laboratory Systems 71 (2004) 83–95
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

A comparison between two robust PCA algorithms

I. Stanimirova[a], B. Walczak[a,b], D.L. Massart[a,*], V. Simeonov[c]

[a] *ChemoAC, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *on leave from Silesian University, 9 Szkolna Street, 40-006 Katowice, Poland*
[c] *Faculty of Chemistry, University of Sofia 'St. Kliment Okhridski', J. Bourchier Blvd.1, 1126 Sofia, Bulgaria*

Received 17 March 2003; received in revised form 13 November 2003; accepted 23 December 2003

Abstract

The article reports the results of a comparative study of two robust Principal Component Analysis (PCA) algorithms based on Projection Pursuit which can be used for exploratory data analysis. The first one is proposed by Croux and Ruiz-Gazen, denoted as C–R algorithm, and the second one by Hubert et al., introducing its modified version, abbreviated as RAPCA. They are applied to uniformly distributed simulated data sets, chemical data sets [environmental and near infrared (NIR) spectra] containing various numbers of variables and objects, as well as different observations' structure. Their performance and features, what they offer, are discussed in detail.

Keywords: Robust PCA; Projection pursuit; Classical scale; Robust scale

1. Introduction

Usually, when one deals with high dimensional data, the first step in the data analysis is a dimensionality reduction. There are different reasons for that. For instance, the multidimensional data sets are difficult to interpret, and their structure cannot be visualized directly. The redundant variables create empty space and computational problems. The most useful tool to solve these problems is Principal Component Analysis (PCA). The main idea of PCA is to project the data from a high dimensional space onto a lower dimensional space. If the data compression is sufficient, the large number of variables is substituted by a small number of uncorrelated latent factors which can explain sufficiently the data structure. The new latent factors, also called principal components (PCs) are obtained by maximizing the variance of the projected data, or in other words, by computing the eigenvalues of correlation or covariance matrix of the samples. The data structure can be visualized directly in a graphical way by projection of objects onto the space defined by selected PCs. In this way, it is possible to detect the main observations' distribution. It happens quite often that, in the chemical data sets, outliers are present. Because the variance is extremely sensitive to the outliers, the information explained by the principal components will be strongly influenced, and the objects' distribution observed will not describe the main bulk of the data well. In some cases, the outliers can be recognized easily on the principal components' plots, and it is necessary to remove them and to repeat the PCA analysis once again. The final PCA result will not be affected by the outlying observations, and one will gain more insight into the data structure. In many cases, it may happen that one outlier is visible but at the same time masks all the others, or several outliers can act together in such a way to diminish or even cancel each other's influence. This is so-called masking effect of outliers due to which they cannot be identified properly in the PCA space. A way might be to use directly a robust PCA technique. Many approaches of this technique are described in the literature. Some of them calculate the eigenvectors and eigenvalues using robust covariance or correlation matrix [1]. Robust PCA can also be obtained based on the Projection Pursuit approach [2–6]. Projection Pursuit, similar to PCA, is a dimensionality reduction method, in which the new latent factors are obtained by maximizing a certain projection index that describes inhomogeneity of the data. Thus, clusters and outliers are emphasized.

In this article, we focus our attention on two recent algorithms for robust PCA, based on Projection Pursuit, as

* Corresponding author. Tel.: +32-2-477-47-37; fax: +32-2-477-47-35.
E-mail address: fabi@vub.vub.ac.be (D.L. Massart).

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2003.12.011

robust PCA proposed by Croux and Ruiz-Gazen [3,4], denoted as C–R algorithm, and its two-step modified version introduced by Hubert et al. [5], denoted further as RAPCA algorithm. The aim of the paper is to compare both methods and to discuss the features they offer. In order to demonstrate the differences, if they exist, between both methods, they were applied to many data sets: uniformly distributed simulated data sets, chemical data sets [environmental, near infrared (NIR) spectra] containing various numbers of variables and objects, as well as different observations' structure. It is our conviction that this comprehensive study showing advantages and disadvantages of both algorithms will help the users to understand the methods presented here in detail. In this paper, we discuss the methods' performance on a few data sets. However, the conclusions are general and they are in agreement with the results obtained for the larger number of data sets tested.

2. Theory

From the Projection Pursuit point of view, PCA uses the variance of the projected data as a projection index. It means that, if the data contain outliers, PCA will find directions (principal components, PCs) mainly influenced by them. However, the main disadvantage of the method is the lack of robustness. The reason is that the variance itself is not robust. When in Projection Pursuit, a robust scale (not sensitive to the outliers' presence) is used as a projection index, robust principal components (RPCs) and robust dispersion matrix can be obtained. This is explored further in both algorithms for robust PCA.

3. Robust PCA algorithm of Croux and Ruiz-Gazen [3,4]

The first step in the classical PCA is to center the data around the mean. It means that the centroid of the row-pattern coincides with the origin in variables space [7]. Similarly to the variance, the mean is not also robust itself. To make this step robust, the L_1-median estimator is used. In the literature, this estimator is often called "spatial median" or "median center". It is defined as a point which minimizes the sum of Euclidian distances to all points in the data. It is highly robust and affine equivariant, which ensures that the estimate will be transformed properly when the axes are rescaled and rotated. In one-dimensional space, L_1-median is reduced to the standard median. Because of the way of estimation, in a high dimensional space (2D and more), the L_1-median center differs from the median center of the data. For more details, the authors recommend Refs. [8–10]. The next step to make the algorithm robust is to find directions in the data space, which are not influenced by outliers, once the data are centered around L_1-median. It is done by the use of a robust scale (instead of the variance as in PCA). These directions are obtained by maximizing the projection index, and they correspond to robust PCs. As the robust scale (projection index), the Q_n estimator is used [11]. For each projection, the pairwise differences between two data points are calculated which leads to a diagonal matrix. After unfolding of the upper or the lower matrix diagonal to a vector, its elements are sorted, and the value, which corresponds to the first quartile, is then taken to compute Q_n as follows:

$$Q_n = 2.2219 * c_n * \{ |z_i - z_j| ; i < j \}_{(k)} \quad (1)$$

where $k = \binom{h}{2} \approx \binom{m}{2}/4$, $h=[m/2]+1$ and c_n is a correction factor, which tends to 1, when the number of objects, m, increases.

Instead of looking for all possible directions, all objects are projected onto normalized vectors passing through each data point and the L_1-median center. The first vector, which has the highest value of the projection index, is found and the objects are then projected onto its orthogonal complement. The second vector, with the highest projection index, is then obtained, and the projection is repeated. It continues until a certain number of vectors is calculated. These vectors, similarly to PCA, are called robust eigenvectors, and the squared robust scale values of the projected data-robust eigenvalues, respectively. The algorithm can be summarized as follows:

Let $\mathbf{X}$ is the data matrix with elements x_{ij}, $i=1,\ldots, m$ (objects) and $j=1,\ldots, n$ (variables).

1. Center $\mathbf{X}$ around L_1-median. It leads to the new centered data matrix $\mathbf{X}_c$.
2. For $i=1$ to f_n, where f_n is the number of robust PCs to be extracted, construct a data matrix containing normalized rows of $\mathbf{X}_c$ (all possible eigenvectors).
3. Project all objects onto the eigenvectors.
4. Calculate the projection index of all eigenvectors.
5. Select the eigenvector with maximal value of the projection index.
6. Update the $\mathbf{X}_c$ by its orthogonal complement.
7. Go to step 2 until f_n robust PCs are found. Project all objects onto the eigenvectors found.

4. Robust PCA algorithm of Hubert et al. (RAPCA) [5]

The robust PCA algorithm, denoted as RAPCA, is proposed in order to improve the C–R algorithm. In our applications, we used the RAPCA algorithm available from [5]. Similar to the C–R algorithm, the L_1-median estimator is also used to obtain the median center of the data and the Q_n estimator as the robust scale. To speed up the algorithm, the first step is to reduce the data space using classical PCA. The data sets are reduced to the

matrix rank. The data variance can be described only by a few variables without loosing important information about the studied data. Although the data structure is described by new coordinates (PCs), it is not changed because PCA preserves the Euclidean distances among the objects. To perform PCA, many algorithms were proposed. When the number of objects is smaller than the number of variables, the kernel version of the eigenvalue decomposition is applied; otherwise, the classical singular value decomposition (SVD) is used [12,13]. The next step of the algorithm is to perform the so-called R-step, also known as Householder transformation, on the score matrix. Thus, both, robust eigenvectors and eigenvalues of the L_1-median-centered score matrix are obtained using Projection Pursuit technique and reflection. When the first eigenvector is found by the C–R algorithm, thereafter, the data are transformed by means of reflection. The reflection plane is defined by its unit normal vector in such a way that the reflection always leads to mapping of the eigenvector onto the basis vector. This reduces the dimensionality of the data space by one. All objects are then projected onto the orthogonal complement of the transformed eigenvector (for each object, the first coordinate is equal to zero). The second eigenvector is obtained by the C–R algorithm but in the reduced space. In order to interpret the results, each eigenvector is transformed back to the original multidimensional space using inverse reflection. This continues until a desired number of eigenvectors is found. The authors [5] have shown that R-step allows more numerical stability and accuracy for the studied data sets than the C–R algorithm, and the RAPCA algorithm is computationally faster than the one proposed by Croux et al.

5. Data sets description

Data set 1 contains 100 NIR spectra of wheat samples. Samples were measured in diffuse reflectance mode as $\log(1/R)$ in the range 1100–2500 nm in 2-nm intervals using Bran + Luebbe instrument. In order to eliminate the baseline shift between the spectra, the data set was preprocessed with the use of the offset correction. This data set is submitted to the database of Chemometrics and Intelligent Laboratory Systems by Kalivas [14].

Data set 2 contains 536 NIR spectra of three creams with three different concentrations of an active drug. These spectra were measured in the range 1100–2500 nm at two different temperatures (20 and 30 °C) and varying a way of cups filling (bad and good) [15].

Data set 3 is an environmental data set, and it concerns annual mean ion concentration values of 9 chemical components (H^+, NH_4^+, Na^+, K^+, Ca^{2+}, Mg^{2+}, Cl^-, NO_3^-, SO_4^{2-}) monitored during 12 years at six sampling sites (Reutte, Kufstein, Innervillgraten, Sonnblick, Nasswald and Lobau), 15 years at Haunsberg and Werfenweng sites, 10 years at Litshau and Lunz, 9 years at Nassfeld site, from the Austrian precipitation sampling network [16].

6. Results and discussion

RAPCA is introduced as a modified version of the C–R algorithm to improve its numerical stability by implementation of R-step. However, this R-step slows down roughly 10 times the RAPCA algorithm in comparison with the C–R algorithm. In order to speed it up, the PCA compression to the rank of the data is performed as a first step. Thus, the RAPCA algorithm as a whole is faster than the one proposed by Croux.

However, the C–R algorithm can also be speeded up by the PCA data compression. To compare the computation time with respect to the number of principal components that are actually computed required by RAPCA, C–R and C–R with implemented compression (denoted, further in the text, as PCA/C–R), uniformly distributed simulated data sets are

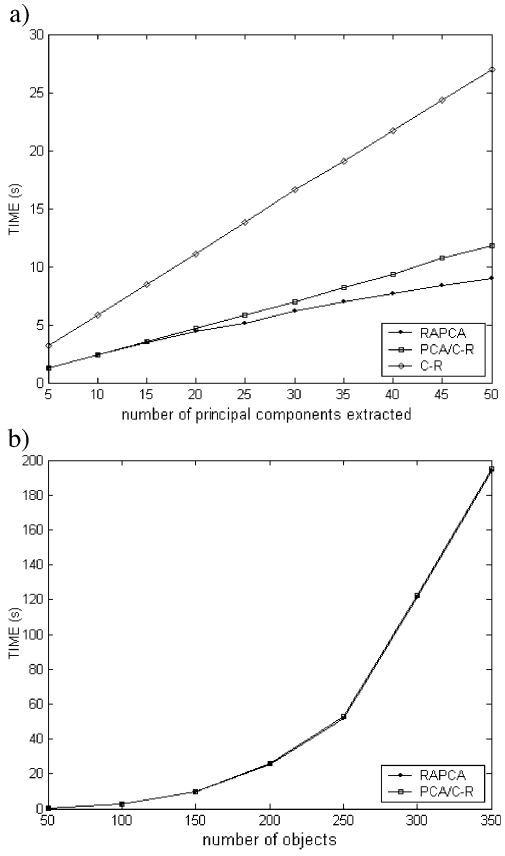

Fig. 1. Computational time comparison between: (a) RAPCA, PCA/C–R and C–R algorithms with respect to the number of extracted principal components for uniformly distributed simulated data sets containing 100 objects and 1000 variables; and (b) RAPCA and PCA/C–R algorithms with respect to the number of objects for uniformly distributed simulated data sets containing 50, 100, 150, 200, 250, 300, 350 objects and 1000 variables as well as fixed number of PCs extracted (10).

used. The computation time is averaged over 10 runs of the certain algorithm applied to the data sets of size 100 × 1000. The result is presented in Fig. 1a. Both RAPCA and PCA/C−R perform results of comparable speed up to 15 PCs extracted. When the number of PCs extracted enhances, the results are obtained faster by the RAPCA algorithm. As we expected, the C−R algorithm is the slowest one. The calculations were done in 2 GHz, 256 Mb RAM PC machine running Matlab 6.1. These results hold for the current implementation of the algorithms.

RAPCA and PCA/C−R show similar tendency of the computation time increase with respect to the number of objects. They consume much time when the number of objects increases. This is shown in Fig. 1b, for the uniformly distributed simulated data sets with different number of objects, i.e., 50, 100, 150, 200, 250, 300 and 350, and 1000 variables as well as fixed number of PCs extracted (10). The computation time is also averaged over 10 runs of the certain algorithm applied to the data sets of the same size.

While the algorithms can be speeded up according to the number of variables by the PCA compression, the computation time strongly depends on the number of objects.

The robust scale values obtained with the use of the C−R and RAPCA algorithms are compared. The results are presented in Fig. 2.

A possible deviation from the linearity (see Fig. 2) would be an indicator of difference in the results obtained by both algorithms, and it could be concluded that the performance of the C−R algorithm really leads to numerical instability caused by the lack of the reflection step (R-step).

However, no differences are observed in the robust scale values for each extracted principal component for

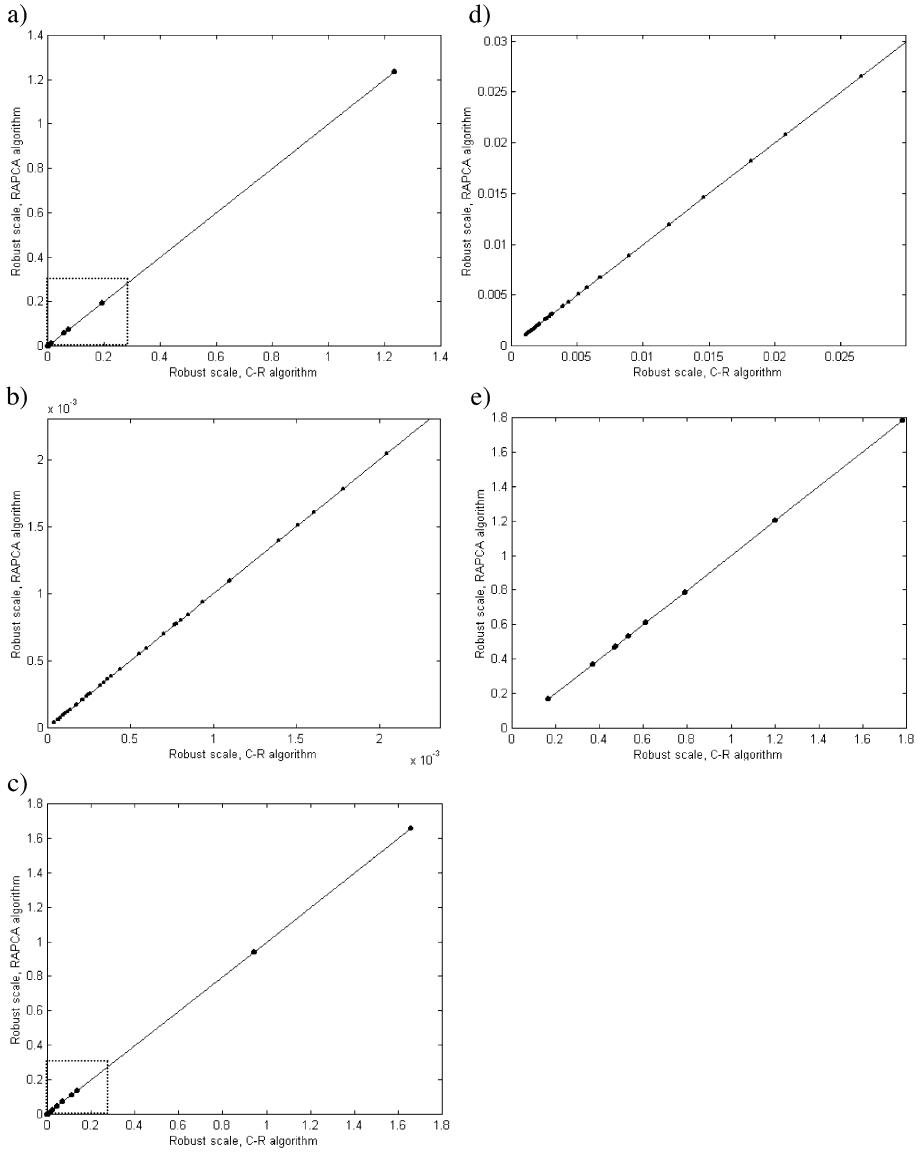

Fig. 2. Robust scale comparison between the RAPCA and C−R algorithms for: (a) data set 1; (b) enlarged region marked in subplot (a); (c) for data set 2; (d) enlarged region marked in subplot (c); and (e) data set 3.

the studied data sets. The strong linear relationship in Fig. 2 justifies the identical robust scale values obtained by both approaches. In the cases of data set 1 (see Fig. 2a and b) and data set 2 (see Fig. 2c and d), deliberately 50 principal components were extracted to detect possible differences between studied algorithms. Moreover, the extraction of so many significant principal components is usually not necessary, and it does not lead to essential improvement of the final result. The regions marked in Fig. 2a and c are shown in enlarged form in Fig. 2b and d, respectively. However, in practice, when one deals with multidimensional chemical data sets (large number of variables), it is always the first step to compress the data to a smaller number of uncorrelated significant principal components. Then, the score matrix can be used as input for other chemometrical approaches. This is already a widely applied procedure and almost considered as a preprocessing step in chemometrics [7]. The data compression speeds up the computations, enables data visualization and allows dealing with curse of dimensionality or empty space phenomenon [17]. For this purpose, the PCA compression to the rank of the data matrix is implemented in the RAPCA algorithm, and it can also be used in the C–R algorithm. Therefore, the term "robust PCA" is used to denote each of these algorithms further in the text due to the same results yielded by both of them.

Data set 3 is an example, where the number of variables is smaller than the number of objects. This happens very

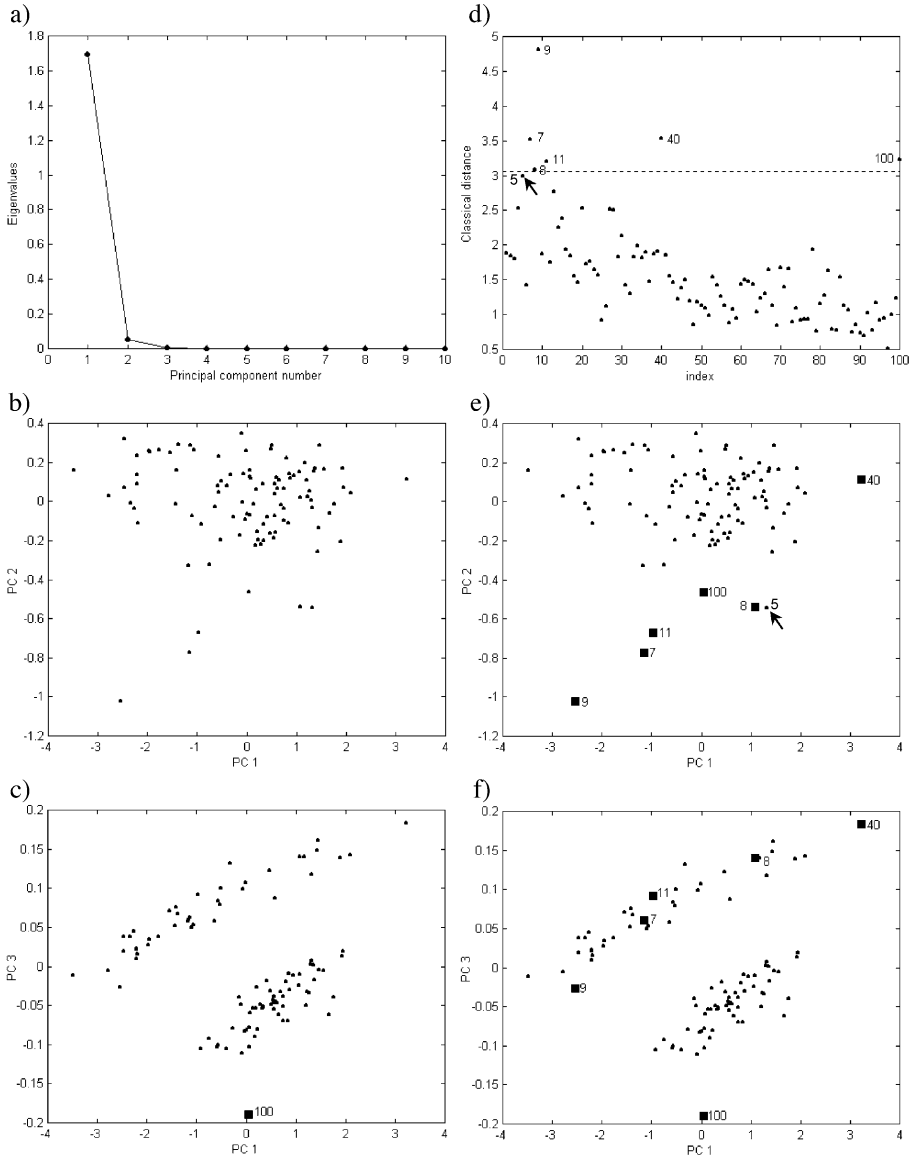

Fig. 3. PCA of data set 1: (a) eigenvalues scree plot; (b) projection of objects on the plane defined by PC 1 and PC 2; (c) projection of objects on the plane defined by PC 1 and PC 3; (d) the classical distance diagnostic plot constructed for three PCs; (e) projection of objects on the plane defined by PC 1 and PC 2 showing the position of objects exceeding the cutoff line in subplot (d); and (f) projection of objects on the plane defined by PC 1 and PC 3 showing the position of objects exceeding the cutoff line in subplot (d).

often to environmental data sets. The compression for data sets 1 and 2 (see Fig. 2a–c) is more effective in comparison to data set 3 (see Fig. 2e).

In order to compare the performance of the classical and robust PCA, Classical Distance (CD_i) and Robust Distance (RD_i) plots are proposed [5,18]. For each ith data point, CD and RD are defined as follows:

$$CD_i = \sqrt{\sum_{j=1}^{p}\left(\frac{t_{ij}^{C}}{s_j^{C}}\right)^2} \quad \text{and} \quad RD_i = \sqrt{\sum_{j=1}^{p}\left(\frac{t_{ij}^{R}}{s_j^{R}}\right)^2}, \qquad (2)$$

where t_{ij}^{C} and t_{ij}^{R} are elements of the scores matrices $\mathbf{T}^{C}$ and $\mathbf{T}^{R}$; s_j^{C} and s_j^{R} are the square root for the jth eigenvalue calculated by the classical (C) and the robust (R) PCA, and p is the number of significant principal components. Because the CD_i and RD_i are defined for the significant number of principal components, p, they are quite sensitive to this number.

Objects whose CD_i and RD_i exceed the cutoff value $\sqrt{\chi^2_{p,0.975}}$ are defined as outliers in the principal components (PCs) space or in the robust principal components (RPCs) space. The notation $\chi^2_{p,0.975}$ corresponds to chi-square distribution for p significant PCs and 97.5% level of confidence.

A disadvantage of the outlier identification based only on the RD is that it is not possible to distinguish between good and bad leverage observations as well as high residuals observations. Such opportunity arises if the robust PCA is

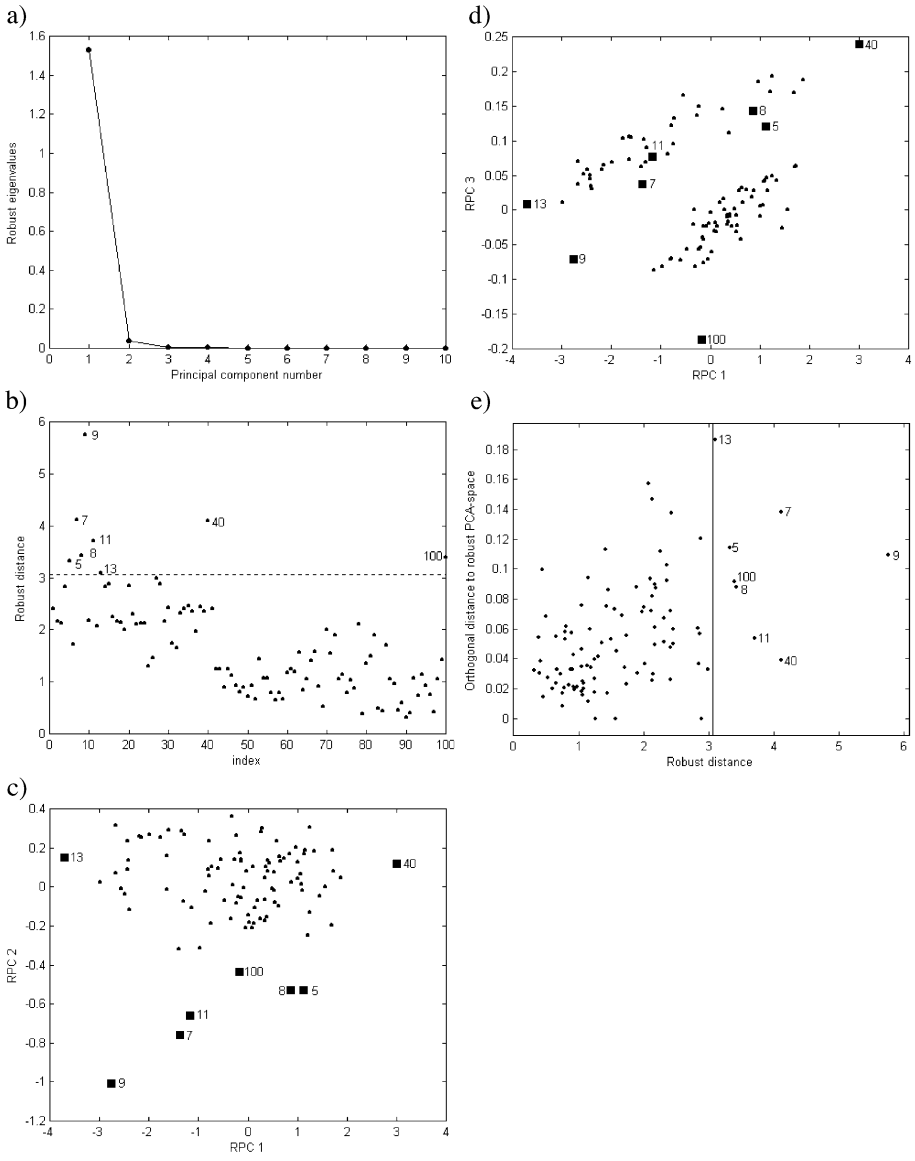

Fig. 4. Robust PCA of data set 1: (a) robust eigenvalues scree plot; (b) robust distance diagnostic plot constructed for three robust PCs; (c) projection of objects on the plane defined by RPC 1 and RPC 2 showing the position of objects exceeding the cutoff line in subplot (b); (d) projection of objects on the plane defined by RPC 1 and RPC 3 showing the position of objects exceeding the cutoff line in subplot (b); and (e) diagnostic plot: orthogonal distance to the robust PCA space vs. robust distance.

considered as a linear model similar to the robust calibration model [19]. Once the robust eigenvectors are found, for every object, its robust score values can be extracted by projecting the object onto the eigenvectors (column vectors of the loadings matrix). One then can predict this object by multiplying the scores by loadings and to compute the residuals between original object and the reconstructed one based on the robust model. Using the obtained residuals, it is possible to recognize high residuals points and to distinguish between good and bad leverage points. If some of the observations are far away from the majority of the data, but they have small residuals (they fit the model well), they are good leverage points and should not be detected as outliers. In order to be able to distinguish between good objects and these three groups of outliers, a diagnostic plot, robust distance, RD_i of each object vs. orthogonal distance, OD, of each object to the robust PCA space, proposed by Hubert and Rousseeuw [20], is used.

The OD to the robust PCA space is defined as:

$$\mathrm{OD} = \| \mathbf{X}_c - \mathbf{P}\mathbf{T}^t \|, \quad (3)$$

where, OD is a vector containing the orthogonal distance values of each object and **P** is the robust loadings matrix. The OD can be considered as a measure of deviation from the constructed model, and the outlier identification based on both OD and RD is free from disadvantage of the outlier identification performed only by the use of RD.

The same assumptions hold true for the classical PCA. A diagnostic plot constructed as classical distance of each object vs. orthogonal distance of each object to the classical PCA space can be used. Both distances are calculated using classical scores and loadings matrices as well as mean-centered data.

Let us see an example based on data set 1. The PCA results obtained for offset-corrected data are shown in Fig. 3.

The eigenvalues of the principal components are presented in Fig. 3a. The first three principal components explain 99.87% of the data variance. The objects distribution is diffused, and it is difficult to decide which objects are outliers only based on a visual inspection of the PC1 vs. PC2 score plot (see Fig. 3b). However, the data reveal clustering tendency in the PC1–PC3 subspace. There are two groups and one outlier, object no. 100 (see Fig. 3c) [14].

The classical distance plot, presented in Fig. 3d, constructed for three principal components indicates that there are several outliers in the data. They are objects nos. 7, 8, 9, 11, 40 and 100 (6% of the total amount of the objects in the data). They are farther from the bulk of the data in the PC1–PC2 subspace (see Fig. 3e), but they belong to one of the two groups in the PC1–PC3 subspace, except object no. 100 (see Fig. 3f).

Object no. 5, located very close to object no. 8 in the PC1–PC2 space (see Fig. 3e), is not detected as an outlier because it appears under the cutoff line (see Fig. 3d).

The number of outliers found strongly depends on the number of significant principal components selected. For PCA, there are tests as cross-validation, Malinowski test etc. [7,13,21,22], based on the data variance or eigenvalues, which help to select the number of significant PCs. In robust PCA, a scree plot showing the robust eigenvalues against their index can be used [5].

The robust eigenvalues strongly decrease and the first kink can be observed around two PCs in Fig. 4a. Thus, three principal components should be taken for the analysis. This is a rule proposed in Ref. [5], but it is quite subjective and it depends on the users' judgment. In some cases, it is quite

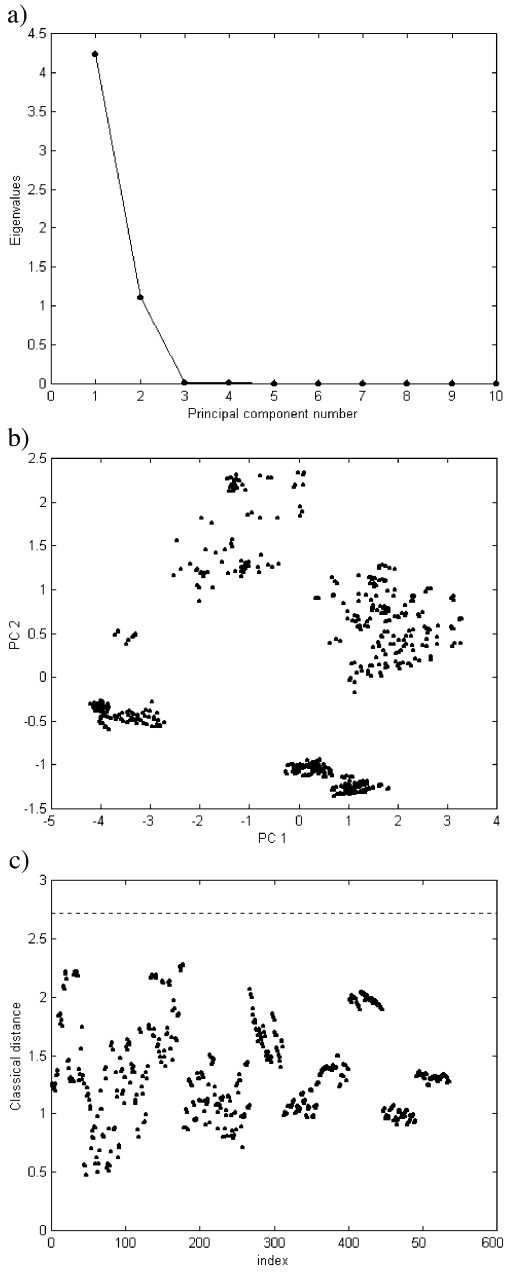

Fig. 5. PCA of data set 2: (a) eigenvalues scree plot; (b) projection of objects on the plane defined by PC 1 and PC 2; and (c) classical distance diagnostic plot constructed for two PCs.

difficult to select the proper number of PCs based only on a visual observation of the eigenvalues scree plot. For this reason, we use another selection criterion [20] based on which the p significant PCs are chosen in such a way that

$$\sum_{j=1}^{p} l_j / \sum_{j=1}^{n} l_j \geq 80\%. \quad (4)$$

For this data set, we took three PCs, where $\sum_{j=1}^{3} l_j / \sum_{j=1}^{100} l_j = 99.74\%$ according to Eq. (4). The robust distance plot designed for the selected number of PCs is shown in Fig. 4b. The outliers identified are 5, 7, 8, 9, 11, 13, 40 and 100 (8% of the total amount of the objects in the data). They are almost the same as these observed on the classical distance plot (see Fig. 3d). Two other outlying objects nos. 5 and 13 appear. These objects are also far from the main bulk of the data in the RPC1–RPC2 space (see Fig. 4c), but they also belong to one cluster in the RPC1–RPC3 space (see Fig. 4d). In this case, both objects 9 and 100 are outlying in the RPC1–RPC2 and RPC1–RPC3 spaces.

In order to be sure which objects are really outliers, the residuals from the robust PCA model were checked, and the diagnostic plot constructed for three robust PCs is shown in Fig. 4e. Object no. 13 is with a high orthogonal distance; no. 9 can be distinguished by its high robust distance and object no. 7 by both high robust and orthogonal distances. However, it is difficult to classify objects nos. 7 and 9 either as good or bad leverage points. Object no. 13 could be considered as an

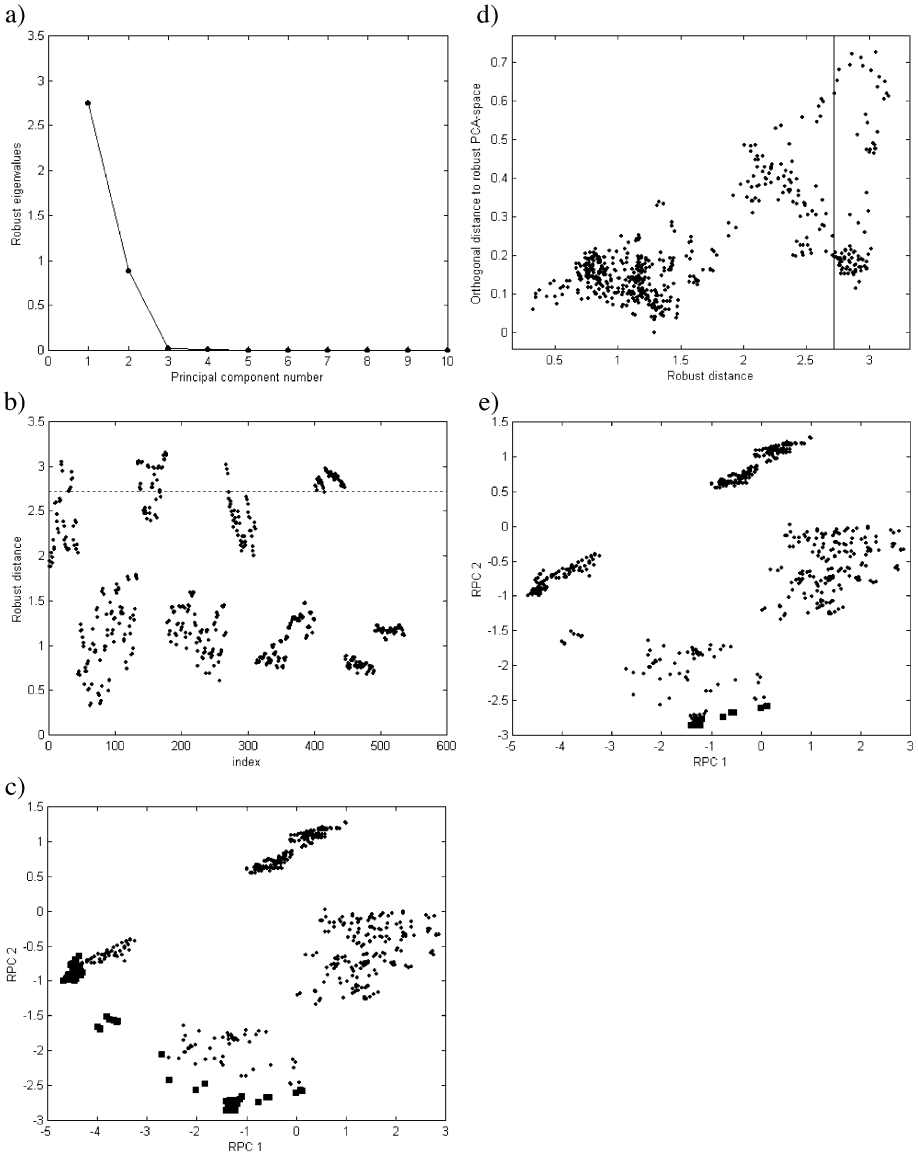

Fig. 6. Robust PCA of data set 2: (a) robust eigenvalues scree plot; (b) robust distance diagnostic plot, constructed for two PCs; (c) projection of objects on the plane defined by RPC 1 and RPC 2 showing the position of objects exceeding the cutoff line in subplot (b); (d) diagnostic plot, orthogonal distance to robust PCA space vs. robust distance constructed for two robust PCs; and (e) projection of objects on the plane defined by RPC 1 and RPC 2 showing the position of objects identified as borderline points.

"orthogonal" outlier because of its high orthogonal distance from the majority of the data. The other ones, i.e., nos. 5, 8, 11, 40 and 100 are good leverage points and should not be considered as outliers.

Another example will be illustrated by data set 2 which is with high clustering tendency. The reason why this example is presented here is that, in most cases, the chemical data sets reveal a very strong departure from normal distribution, and the outlier identification based only on RD gives misleading conclusions, while the outlier identification using RD vs. OD plot makes it possible to recognize high residual points and to distinguish between good and bad leverage points. For this data set, two significant principal components can be selected using the eigenvalues scree plot shown in Fig. 5a. They explain 99.42% of the data variance. There are four well-separated groups with a small group of inliers on the PC1–PC2 score plot (see Fig. 5b).

It is known that this data set does not contain outliers [13]. The classical distance plot built for two principal components is in agreement with this result, and there are no outliers detected (see Fig. 5c).

According to the criterion (4), $\sum_{j=1}^{2} l_j / \sum_{j=1}^{536} l_j = 99.41\%$ and the number of robust principal components which should be selected is two (see Fig. 6a).

When the robust distance plot is constructed for two principal components, there are many outliers identified (16.23% of the total amount of the objects in the data) (see Fig. 6b). They belong to parts of the clusters, and they are marked on the RPC1–RPC2 score plot (see Fig. 6c). The diagnostic plot (see Fig. 6d): robust distance vs.

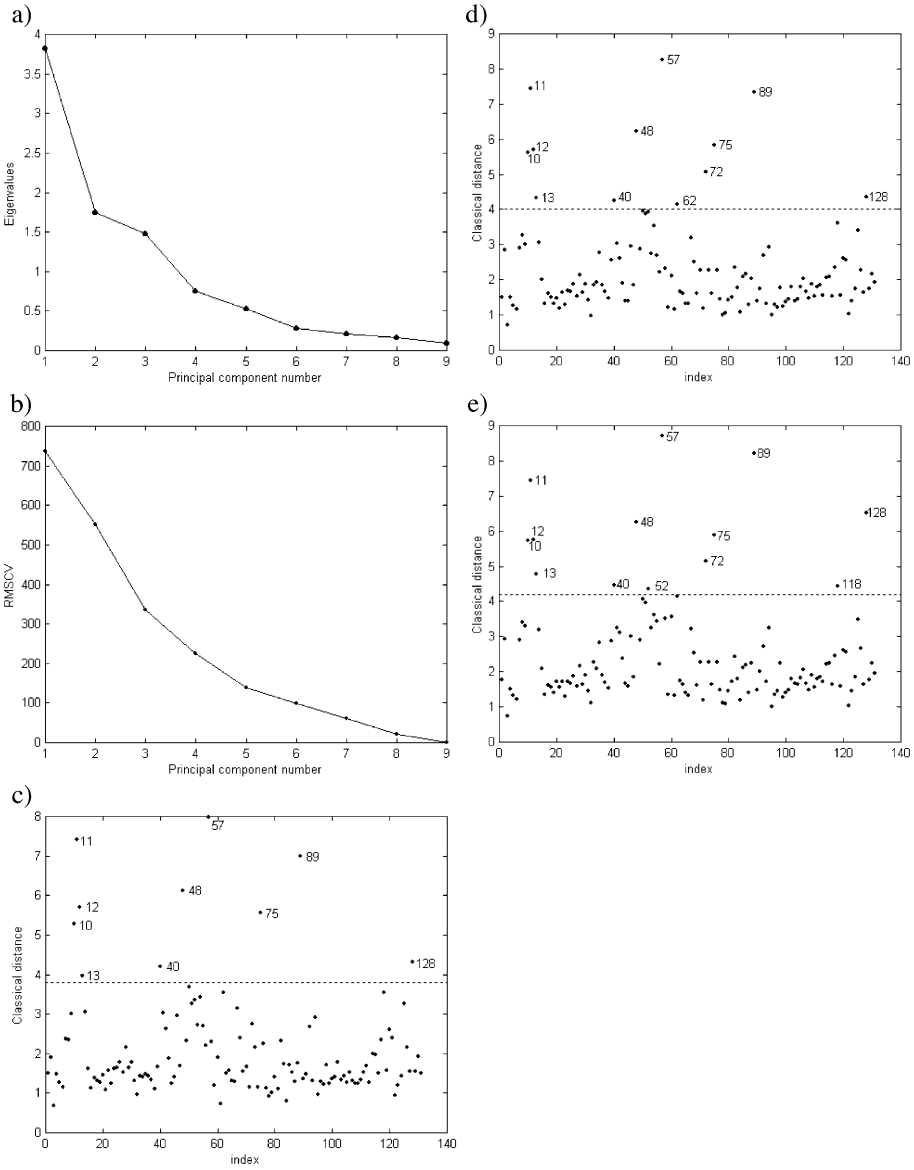

Fig. 7. PCA of autoscaled data set 3: (a) eigenvalues scree plot; (b) RMSCV scree plot; (c) classical distance diagnostic plot constructed for six PCs; (d) classical distance diagnostic plot constructed for seven PCs; and (e) classical distance diagnostic plot constructed for eight PCs.

orthogonal distance to the robust PCA space reveals additional information about the outliers identified. Most of the objects are good leverage points except for the objects marked in the RPC1–RPC2 space (see Fig. 6e) which are borderline cases. This data set does not contain bad leverage points.

The results for classical and robust PCA, based on the robust distance plot diagnostic, are quite different. Using only this outlier identification for robust PCA, many good objects are considered as outliers, whereas the same objects are detected as good leverage points when the detection based on the robust orthogonal distance is performed. Thus, if the data reveal clustering tendency, it is strongly recommended to analyze the residuals from the robust PCA model.

Usually, the measured properties for the environmental data sets have different units, and this reflects to a large difference in the range of variables. In order to remove this difference, an autoscaling procedure is applied as the first step and classical PCA is then performed as the second one. The autoscaling, also called column standardization, is a preprocessing operation which removes the differences between the variables' range and gives them the same importance in the data analysis. This is performed by subtracting the corresponding column mean of each data point and dividing by its column standard deviation [7].

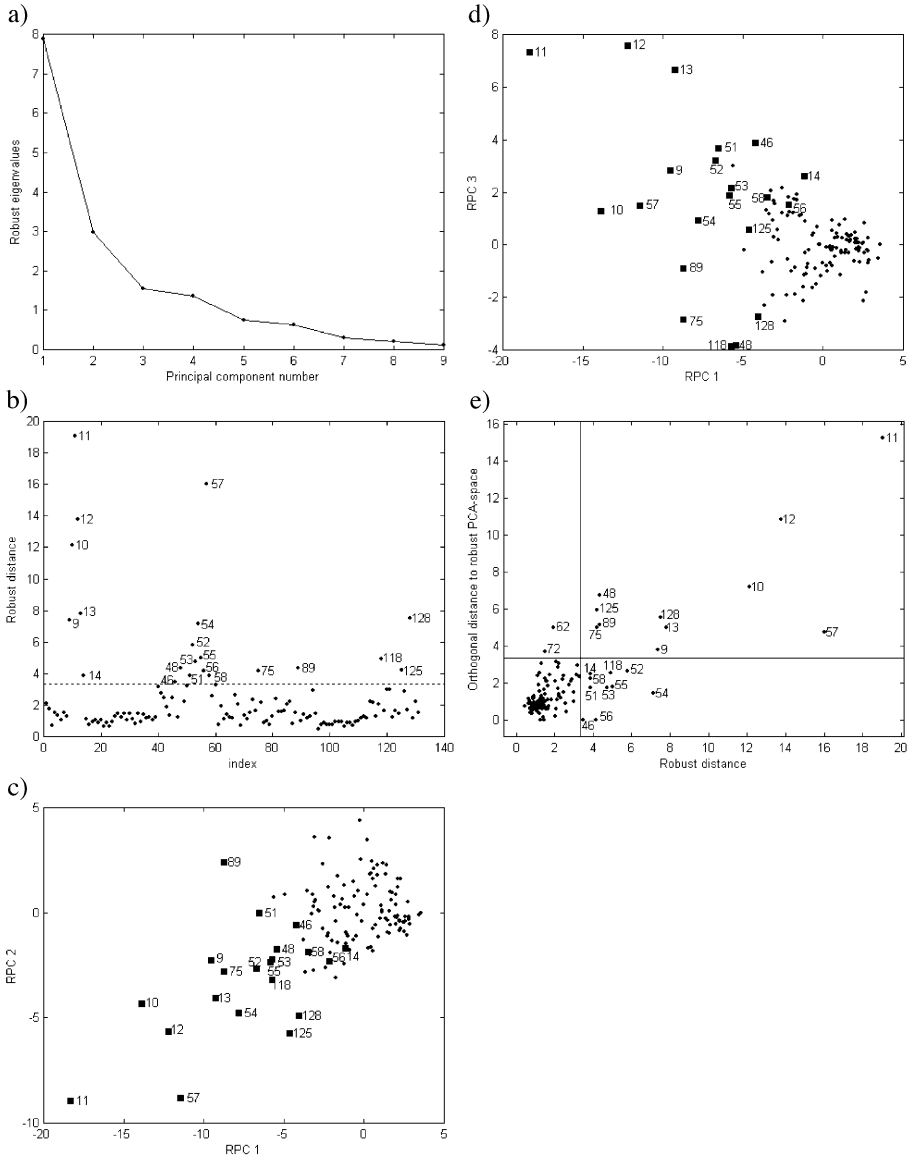

Fig. 8. Robust PCA of standardized in a robust way data set 3: (a) robust eigenvalues scree plot; (b) robust distance diagnostic plot constructed for four robust PCs; (c) projection of objects on the plane defined by RPC 1 and RPC 2 showing the position of objects exceeding the cutoff line on subplot (b); (d) projection of objects on the plane defined by RPC 1 and RPC 3 showing the position of objects exceeding the cutoff line on subplot (b); and (e) diagnostic plot, orthogonal distance to robust PCA space vs. robust distance constructed for four robust PCs.

Data set 3 is autoscaled, by mean and standard deviation, and then the classical PCA is performed. The PCA compression is not effective, and it is difficult to select the number of significant principal components using eigenvalues plot shown in Fig. 7a. For this reason, the leave-one-out object cross-validation test is used. Each object is predicted from the PCA results of the remaining ones by leaving out each object. After PCA, a given number of PCs is used to build a model. For different number of factors, the root mean square error of cross-validation (RMSCV) is obtained, and the number of optimal PCs is decided by looking for the minimum of the error. The result is presented in Fig. 7b. However, in this case, it is not very helpful.

The classical distance plot is constructed for six, seven and eight principal components (see Fig. 7c–e). Different complexity of the data leads to the different outlier identification. Ten objects exceed the cutoff line on the plot based on six PCs (see Fig. 7c). They are nos. 10–13, 40, 48, 57, 75, 89, 128. When the complexity is seven, two other objects appear as outliers—nos. 62 and 72 (see Fig. 7d). The distance plot built for eight PCs shows two new outliers—objects nos. 52 and 118. Object no. 62 now is not identified as outlier compared to the case when the outliers' detection was based on seven PCs.

When robust PCA is used, the data have to be standardized in a robust way. Instead of using mean and standard

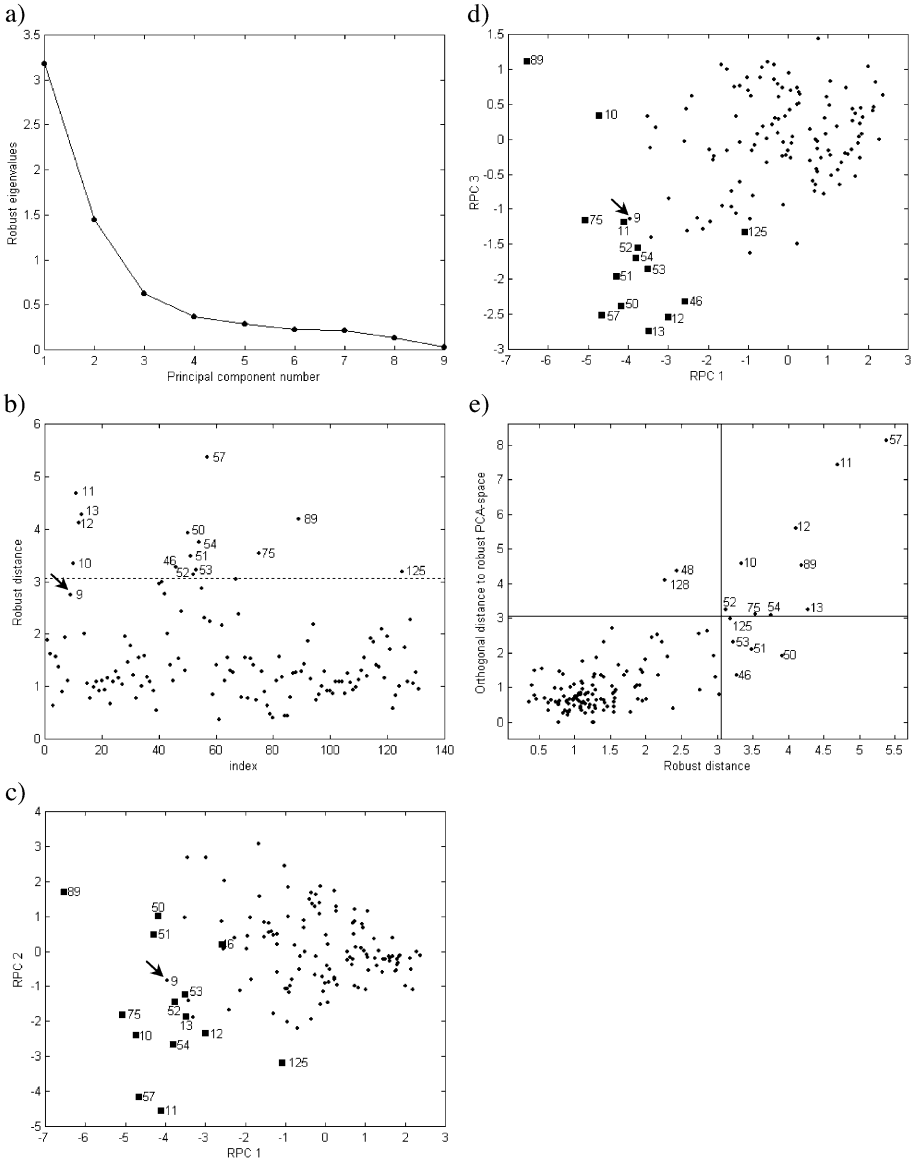

Fig. 9. Robust PCA of autoscaled data set 3: (a) robust eigenvalues scree plot; (b) robust distance plot constructed for three PCs; (c) projection of objects on the plane defined by RPC 1 and RPC 2 showing the position of objects exceeding the cutoff line on subplot (b); (d) projection of objects on the plane defined by RPC 1 and RPC 3 showing the position of objects exceeding the cutoff line on subplot (b); and (e) diagnostic plot, orthogonal distance to robust PCA space vs. robust distance constructed for three robust PCs.

deviation of each variable as in autoscaling, median and the robust scale estimate are applied. It means that, firstly, the corresponding column median is subtracted of each data point and then divided by its robust scale estimate.

Data set 3 is standardized in the robust way, and robust PCA is then performed. According to the selection criterion (4), $\sum_{j=1}^{4} l_j / \sum_{j=1}^{131} l_j = 87.21\%$, four PCs can be taken for the analysis (see Fig. 8a). However, the percentage of outliers (nos. 9–14, 46, 48, 51–58, 75, 89, 118, 125 and 128) is high—16.03—on the robust distance plot (see Fig. 8b).

Objects nos. 10, 11, 12 and 57 become far from the main bulk of the data in the RPC1–RPC2 space (see Fig. 8c), but the other objects identified as outliers seem to belong to one group. A similar picture is observed in the RPC1–RPC3 space (see Fig. 8d), where all objects are almost close to each other except for nos. 11, 12 and 13 which are far away.

When the diagnostic plot, robust distance vs. orthogonal distance to the robust PCA space, is constructed for four PCs (see Fig. 8e), one can clearly distinguish different groups of observations: objects nos. 9–13, 54, 57, 128 as bad leverage points, objects nos. 14, 46, 51–53, 55, 56, 58 and 118 as good leverage points and objects nos. 48, 75, 89, 125, together with nos. 62 and 72 as borderline cases (leverage points which differs from the model). The horizontal cutoff line on the plot (see Fig. 8e) has a value which corresponds to chi-square distribution for four significant PCs and 97.5% level of confidence $(\sqrt{\chi^2_{4,0.975}})$.

In order to demonstrate that the data standardization by mean and standard deviation leads to different outlier identification, we present here results of such investigation for data set 3. The number of significant principal components, which should be taken for the analysis according to the criterion (4), $\sum_{j=1}^{3} l_j / \sum_{j=1}^{131} l_j = 80.56\%$, is three (less than in the case when robust PCA was applied to the data standardized in the robust way by median and the robust scale estimate; see Fig. 9a). The percentage of detected outliers is also high, as in the case of the data preprocessed in the robust way, when the robust distances plot is constructed for four principal components (see Fig. 9b). It is equal to 10.69%. Objects nos. 10–13, 46, 50–54, 57, 75, 89 and 125 exceed the cutoff line on the robust distance plot and occur as outliers (see Fig. 9b). They differ from those identified, when the data was transformed in the robust way. Objects nos. 9, 14, 48 55, 56, 58, 118 and 128 are not detected now as outliers.

Diffuse objects' distribution appears on the robust score plots shown in Fig. 9c and d. Objects, which are identified as outliers, are close to objects which are not. For example, object no. 9 is close to objects nos. 52 and 53 in the RPC1–RPC2 space and to object no. 11 in the RPC1–RPC3 space (see Fig. 9c and d), but it is not detected as an outlier.

The robust orthogonal distance plot constructed for three PCs (see Fig. 9e) indicates that objects nos. 10–13, 52, 54, 57, 75 and 89 are bad leverage points, objects nos. 46, 50, 51, 53 and 125 are good leverage points, and nos. 48 and 128 are borderline cases. Object no. 125 is very close to the objects nos. 54, 75 and 52 but it is not detected as bad leverage point because it does not exceed the horizontal cutoff line. In such cases, the decision, whether the object is a good or bad leverage point, is quite difficult and can be questioned. Object no. 9 now is a regular observation, whereas when the data were standardized in the robust way, it was bad leverage point.

The autoscaled data by mean and standard deviation used as input in robust PCA lead to different outliers' identification (see Figs. 8b and 9b). For this reason, the preprocessing of the data is a very important step. In robust PCA, the input data have to be transformed in a robust way; otherwise, it leads to different conclusions.

The results for outlier identification of robust PCA depend strongly on the number of significant robust principal components selected, based on which the robust distance plot or robust orthogonal distance plot are constructed. They do not differ from the number of input PCs, which are determined by the input matrix rank.

7. Conclusions

Several conclusions can be pointed out. There are no differences in the robust scale obtained by both the RAPCA and C–R algorithms for all studied data sets. The compression step implemented in RAPCA to speed up the algorithm can also be used in the C–R algorithm for the same purpose. Once the score matrix is used as an input, we can observe comparable performance with respect to the speed of the C–R algorithm up to 15 PCs computed. The computation time increases strongly with respect to the number of objects in both RAPCA and PCA/C–R algorithms.

The number of outliers identified strongly depends on the number of the selected principal components. There are many tests helping to select a proper number of PCs while the classical PCA model is constructed, but in some cases, when the PCA compression is not effective, there are difficulties to select significant number of PCs. Different complexity of the data yields to different outliers' identification. For robust PCA, the principal components selection criterion used is very helpful. The results, based on the robust and classical distance plots diagnostic, for robust and classical PCA of homogeneous data sets seem to be in agreement, while whole groups can become outlying objects when the data sets are clustered. For this reason, it is strongly recommended the residuals, from the classical and robust PCA models, to be checked. In this way, an additional information about good and bad leverage points as well as high residuals points can be obtained which will help for the final conclusions.

It is important to choose a proper preprocessing procedure when it is necessary to remove the differences in the variables or objects range: robust standardization by median and robust scale has to be used when robust PCA is performed. Another preprocessing procedure leads to completely different results.

References

[1] C. Croux, G. Haesbroeck, Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, Biometrika 87 (2000) 603–618.

[2] G. Li, Z. Chen, Projection Pursuit approach to robust dispersion matrices and principal components: primary theory and Monte Carlo, Journal of the American Statistical Association 80 (1985) 759–766.

[3] C. Croux, A. Ruiz-Gazen, High Breakdown Estimators for Principal Components: The Projection–Pursuit Approach Revisited, vol. 29, The IMS Bulletin, 2000, p. 270.

[4] C. Croux, A. Ruiz-Gazen, A fast algorithm for robust principal components based on Projection Pursuit, COMPSTAT: Proceedings in Computational Statistics 1996, Physica-Verlag, Heidelberg, 1996, pp. 211–217.

[5] M. Hubert, P. Rousseeuw, S. Verboven, A fast method for robust principal components with application to chemometrics, Chemometrics and Intelligent Laboratory Systems 60 (2002) 101–111.

[6] L. Ammann, Robust singular value decompositions: a new approach to projection pursuit, Journal of the American Statistical Association 88 (1993) 505–514.

[7] B.M.G. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier Science B.V., Amsterdam, 1998.

[8] B.M. Brown, Statistical uses of the spatial median, Journal of the Royal Statistical Society. Series B 45 (1983) 25–30.

[9] B.M. Brown, P. Hall, G. Alastair Young, On the effect of inliers on the spatial median, Journal of Multivariate Analysis 63 (1997) 88–104.

[10] C.G. Small, A survey of multidimensional medians, International Statistical Review 58 (1990) 263–277.

[11] P.J. Rousseeuw, C. Croux, Alternatives to the median absolute deviation, Journal of the American Statistical Association 88 (1993) 1273–1283.

[12] W. Wu, D.L. Massart, S. De Jong, The kernel PCA algorithms for wide data: Part I. Theory and algorithms, Chemometrics and Intelligent Laboratory Systems 36 (1997) 165–172.

[13] W. Wu, D.L. Massart, S. De Jong, Kernel–PCA algorithms for wide data: Part II. Fast cross-validation and application in classification of NIR data, Chemometrics and Intelligent Laboratory Systems 37 (1997) 271–280.

[14] V. Centner, J. Verdu-Andres, B. Walczak, D. Jouan-Rimbaud, F. Despangne, L. Pasti, R. Poppi, D.L. Massart, O.E. de Noord, Comparison of multivariate calibration techniques applied to experimental NIR data sets, Applied Srectroscopy 54 (2000) 608–623.

[15] J. Luypaert, S. Heuerding, S. De Jong, D.L. Massart, An evaluation of direct orthogonal signal correction and other preprocessing methods for the classification of clinical study lots of a dermatological cream, Journal of Pharmaceutical and Biomedical Analysis 30 (2003) 1–14.

[16] M.F. Kalina, H. Puxbaum, A high density network for wet only precipitation chemistry sampling in Austria, Idojaras 100 (1996) 159–170.

[17] P. Huber, Projection pursuit, Annals of Statistics 13 (1985) 435–475.

[18] P. Rousseeuw, M. Leroy, Robust Regression and Outlier Detection, John Wiley and Sons, Inc., USA, 1987.

[19] B. Walczak, D.L. Massart, Robust PCR as a detection tool for outliers, Chemometrics and Intelligent Laboratory Systems 27 (1995) 41–54.

[20] M. Hubert, P. Rousseeuw, ROBPCA: A new approach to robust Principal Component Analysis, to appear in Technometrics.

[21] E. Malinowski, Factor Analysis in Chemistry, Wiley, 1991.

[22] Z.P. Chen, Y.Z. Liang, J.H. Jlang, Y. Li, J.Y. Qian, R.Q. Yu, Determination of the number of components in mixtures using a new approach incorporating chemical information, Journal of Chemometrics 13 (1999) 15–30.