Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 85 (2007) 203–219
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Robust statistics in data analysis — A review Basic concepts

M. Daszykowski [a], K. Kaczmarek [b,c], Y. Vander Heyden [c], B. Walczak [a,*]

[a] Department of Chemometrics, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland
[b] on leave from Department of Chemometrics, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland
[c] Department of Analytical Chemistry and Pharmaceutical Technology, Vrije Universiteit Brussel, Laarbeeklaan 103, B-1090 Brussels, Belgium

Received 27 March 2006; received in revised form 20 June 2006; accepted 21 June 2006
Available online 28 August 2006

Abstract

Presence of outliers in chemical data affects all least squares models, which are extensively used in chemometrics for data exploration and modeling. Therefore, more and more attention is paid to the so-called robust models and robust statistics that aim to construct models and estimates describing well data majority. Moreover, construction of robust models allows identifying outlying observations. The outliers identification is not only essential for a proper modeling but also for understanding the reasons for unique character of the outlying sample.

In this paper some basic concepts of robust techniques are presented and their usefulness in chemometric data analysis is stressed.

Keywords: Outliers; L1-median; Projection Pursuit; Robust covariance; Mahalanobis distance; Outlier diagnostic; Robust PCA

1. Introduction

Understanding of studied chemical phenomena or systems relies on interpretation of the analytical data obtained during experiments. The individual variables are usually described by such statistics as data mean, standard deviation or variance. However, to take into account the multivariate nature of the data, multivariate chemometric data analysis techniques ought to be applied.

Steps of data analysis usually lead through data exploration to modeling. For these purposes often different least squares techniques are employed. In chemometrics Principal Component Analysis (PCA) is probably one of the best known unsupervised techniques with least squares cost function. By means of the PCA model the data are decomposed into a set of a few orthogonal latent variables, called Principal Components (PCs), defining a new coordinate system and the so-called loadings, describing the contribution of individual variables to a given PC. The key property of PCs is that they maximize the description of the data variance, and thus, usually the first few PCs are enough to represent the data structure well. This issue makes the PCA technique well suited for multivariate data visualization and interpretation. In the arsenal of supervised modeling and classification techniques there are approaches as Principal Component Regression, Partial Least Squares regression, and more. All abovementioned least squares methods are extensively used in chemometrics in a wide range of chemical problems. However, when the fundamental assumption about normal distribution of the model residuals is not fulfilled, the least squares models are not optimal. This happens when the data contain outliers, i.e., samples with an extreme characteristic due to at least one atypical value of the measured parameters. In statistical sense outliers are samples from a different population than the data majority. The presence of outliers in the data can be due to two main reasons. One of them is an experimental error. The other reason is the unique character of a few objects. For instance, in environmental sciences outliers can represent a certain unusual environmental event that took place. In drug design, an outlier can be a molecule with a far different biological activity than the other. Therefore, it may be considered as a good starting point for the new synthesis. Both types of outliers are important to be identified, however, the reason for their identification is two fold, either, to remove them from the data in order to obtain correct results of the analysis, or to find the

* Corresponding author.
E-mail address: beata@us.edu.pl (B. Walczak).

0169-7439/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2006.06.016

explanation for their outlyingness to understand better the studied process.

To draw reliable conclusions about the contaminated data, it is necessary to consider so-called robust approaches, able to neglect the outliers presence and to represent the data majority. This relies upon finding proper estimates of the data location and scale. Up till now, many robust versions of classical estimators and classical chemometric approaches have been proposed.

Therefore, the goal of this paper is to present some fundamental concepts of robust statistics and to point out their role in the analysis of chemical data.

1.1. Notation and abbreviations

$\mathbf{X}$	a data matrix with m observations (objects) and n variables (measured parameters)
$\boldsymbol{x}_i$	the *i*-th object of the data matrix (a row vector)
$\mathbf{X}_c$	a column-wise centered data matrix
$\mathbf{C}$	a covariance matrix of $\mathbf{X}$
$\mathbf{R}$	a matrix of Pearson correlation coefficients of $\mathbf{X}$
$\mathbf{I}$	identity matrix
$\boldsymbol{I}^{\mathrm{T}}$	a column vector containing ones
$T(\boldsymbol{x})$	an estimator of variable $\boldsymbol{x}$
$\mu(\boldsymbol{x})$	mean (average) of the variable $\boldsymbol{x}$
median($\mathbf{X}$)	coordinate-wise median of data matrix $\mathbf{X}$
$\mu_{\mathrm{MCD}}(\boldsymbol{x})$	robust mean of variable $\boldsymbol{x}$ based on the MCD estimator of location and scale
$\mu_{\mathrm{L1}}(\mathbf{X})$	robust multivariate median (L1-median) of data matrix $\mathbf{X}$
$\sigma(\boldsymbol{x})$	standard deviation of the variable
$V(\boldsymbol{x})$	variance of the variable
$\sigma_{\mathrm{MAD}}(\boldsymbol{x})$	robust standard deviation of the variable $\boldsymbol{x}$ based on the median of absolute deviation scale
$\sigma_{\mathrm{Sn}}(\boldsymbol{x})$	robust standard deviation of the variable $\boldsymbol{x}$ based on the Sn scale
$\sigma_{\mathrm{Qn}}(\boldsymbol{x})$	robust standard deviation of the variable $\boldsymbol{x}$ based on the Qn scale
$\sigma_{\mathrm{MCD}}(\boldsymbol{x})$	robust standard deviation of the variable $\boldsymbol{x}$ based on the MCD estimator of location and scale
MD	Mahalanobis distance
SD	score distance
OD	orthogonal distance
PC	Principal Component
PCs	Principal Components
rPC	Robust Principal Component
rPCs	Robust Principal Components
PCA	Principal Component Analysis
EPCA	Elliptical Principal Component Analysis
SPCA	Spherical Principal Component Analysis
RPCA	Robust Principal Component Analysis
ICA	Independent Component Analysis
PP	Projection Pursuit
TOP	Trimmed Objects Projections
MAD	Median of Absolute Deviation around the median
MVT	Multivariate Trimming
MCD	Minimum Covariance Determinant
MVE	Minimum Volume Ellipsoid

2. Theory

2.1. Classical and robust estimators of the data location

The mean (average) of the data is the best-known estimate of a true value of a random variable $\boldsymbol{x}$. The mean is a location estimator characterizing a general position of the data. For a single variable $\boldsymbol{x}$ its mean, μ, is the sum of all elements divided by their number m:

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1}$$

However, in the presence of outlying observations the data mean is no longer a reliable estimate of the data location, and therefore it is said to be a non-robust estimator.

The estimators, with respect to their properties, can be divided into two categories, and namely, i.e., robust and non-robust. The robust estimators aim to describe well the data majority regardless the data contamination. The robustness of an estimator can be described by its breakdown point, a concept introduced by Hampel [1], however also other robustness criteria exist [2]. In general, one can speak about qualitative and quantitative robustness. The qualitative robustness aims to express the differences between two studied distributions by means of the Prohorov distance. When this distance is small then the difference between the distributions of estimations is also small. There are two concepts that target quantitative robustness issue (i) the breakdown point of an estimator and (ii) its influence function, expressing global and local sensitivity, respectively.

In the light of the definition for a finite sample, the breakdown point of an estimator is the maximal fraction of outlying objects in the data, that the estimator can handle yielding acceptable estimates. For instance, the breakdown point of the mean estimator equals 0% being the smallest possible. For such breakdown point a single outlier can completely bring the estimate to an abstract level. The influence function of an estimator aims to describe the influence of objects upon the estimator [3] with respect to infinitesimal perturbations.

Efficiency of an estimator expresses how good estimates the estimator yields for non-contaminated data compared to a classical estimator.

In general, a good robust estimator should have a high efficiency, high breakdown point and smoothed influence function. These requirements ensure a satisfactory performance of an estimator for the data with and without outliers. There are different types of robust estimators. Parametric estimators assume a certain data distribution, for instance that the data majority follows a normal distribution and thus such estimators simply eliminate outliers. Such parametric estimators work well for contaminated data but also offer better estimates compared to classical ones obtained for other non-normal models of the data distribution (Cauchy, t, Laplace, *etc.*). Non-parametric estimators are robust in their nature because they do not require knowledge about the data distribution at hand. Yet another possibility arises with a semi-non-parametric approach, where

any type of the non-normality can be handled. Contrary to parametric estimators, semi-non-parametric estimators do not reject outliers but transform the data.

The lack of robustness of the mean estimator, can be explained by its least squares nature. The mean of a random variable is a point minimizing the Euclidean distances to all data objects. This condition is expressed as:

$$\min_{\mu} \sum_{i=1}^{m} ||x_i - \mu(\boldsymbol{x})|| \tag{2}$$

where ||...|| is the L2-Euclidean norm.

The median of the data is a robust alternative to the mean location estimator. The median of a variable is the middle element for an odd number of sorted elements. The median of a variable with an even number of sorted elements, is the average of the two elements at the closest positions to the half-length of the variable.

The median is a very robust estimate of the data location, and its breakdown point is 50%. This is the highest possible.

Up till now, for a single variable its mean and median estimators were presented. When the data are multidimensional, i.e., the objects are described by several physico-chemical properties (variables) the data means and medians can be computed in a univariate manner, considering each data variable individually. This leads to the column means and column medians of the data (coordinatewise mean and coordinatewise median), respectively. It is also possible to consider the multidimensional nature of the data and the median as an estimate of a center of the multidimensional data cloud. One of the most popular multidimensional (spatial) median estimators is the L1-median. The concept of this estimator is due to Weber, who described a way to find an optimal position of a factory with respect to minimal transportation costs [4]. The L1-median is defined as a point, $\mu_{L1}(\mathbf{X})$, in a multidimensional space that minimizes the sum of the Euclidean distances between $\mu_{L1}(\mathbf{X})$ and any data object:

$$\min_{\mu_{L1}} \sum_{i=1}^{n} ||\boldsymbol{x}_i - \mu_{L1}(\mathbf{X})|| \tag{3}$$

where ||...|| is the L1-norm, being less influenced by the outliers compared to L2-norm.

The L1-median is a highly robust estimator of multivariate data location with a 50% breakdown point [5].

The L1-median is a generalization of the univariate median. Although the L1-median seems to be the best-known multidimensional median, some other exist as well. A good review of multidimensional medians can be found in [3,6,7] and the references therein. Robust estimates of location as well as other robust estimates can be also derived applying the fuzzy set theory [8,9].

In Fig. 1 the effect of outliers upon the data mean and L1-median is demonstrated. When outliers are present in the data, they can influence the data mean to a different degree depending on their distance from the data majority. This is contrary to the L1-median estimator, which describes well the center of the data majority, even under strong data contamination.

2.2. Classical and robust estimators of data scale

The standard deviation, σ, and the variance, V, of the random variable $\boldsymbol{x}$ are used to describe the data spread (scale). They are defined as:

$$\sigma(\boldsymbol{x}) = \sqrt{\frac{1}{(m-1)} \sum_{i=1}^{m} (x_i - \mu(\boldsymbol{x}))^2} \tag{4}$$

$$V(\boldsymbol{x}) = \frac{1}{(m-1)} \sum_{i=1}^{m} (x_i - \mu(\boldsymbol{x}))^2 = \sigma^2(\boldsymbol{x}) \tag{5}$$

Among different robust scale estimators, median of absolute deviation about the median, often called median absolute deviation (MAD) [10], seems to be best-known. The MAD estimator of a variable scale, σ_{MAD}, is defined as:

$$\sigma_{MAD} = c \cdot \text{median}_i \cdot |x_i - \text{median}_j(x_j)| \tag{6}$$

where c is a constant equal to 1.4826.

The σ_{MAD} estimator has a 50% breakdown point. Although it is easy to compute, it suffers from a low efficiency. The more efficient alternatives to σ_{MAD} estimator are the so-called Sn and Qn scale estimators [11,12]. These estimators do not take into account the spread of the data around the center but pair-wise differences between variable elements, i.e., the distances. The Sn and Qn scale estimators, σ_{Sn} and σ_{Qn}, are defined in the following way:

$$\sigma_{Sn} = \text{cf} \cdot c \cdot \text{median}_i \{ median_j |x_i - x_j| \} \tag{7}$$

$$\sigma_{Qn} = \text{cf} \cdot c \cdot \{ |x_i - x_j; i \leq j| \}_{(k)} \tag{8}$$

where cf is consistency factor depending on the data size, c is a constant factor (for σ_{Sn}, $c=1.1926$, and for σ_{Qn}, $c=2.2219$) and $k = \binom{h}{2} \approx \binom{n}{2}$, where $h=[n/2]+1$.

Such design of the σ_{Sn} and σ_{Qn} estimators makes them highly robust with maximal possible breakdown points equal to 50%. Comparison of the three non-parametric scale estimators σ_{MAD}, σ_{Sn} and σ_{Qn} favors the σ_{Qn} estimator, as that one with the highest efficiency at Gaussian distribution, being about 82% (efficiencies of MAD and Sn estimators are equal to ca. 37% and 58%, respectively) and with a smooth influence function [11]. Due to non-parametric nature of the above mentioned estimators they are very attractive, but there are also other estimators of robust scale (e.g. M-estimators, trimmed scales, *etc*. [2]). The MATLAB implementations of Qn and Sn scale estimators are a part of the TOMCAT toolbox available from [13].

2.3. Classical and robust methods of data transformation

In data analysis often the studied data require preprocessing (a transformation). Data preprocessing aims to correct undesired

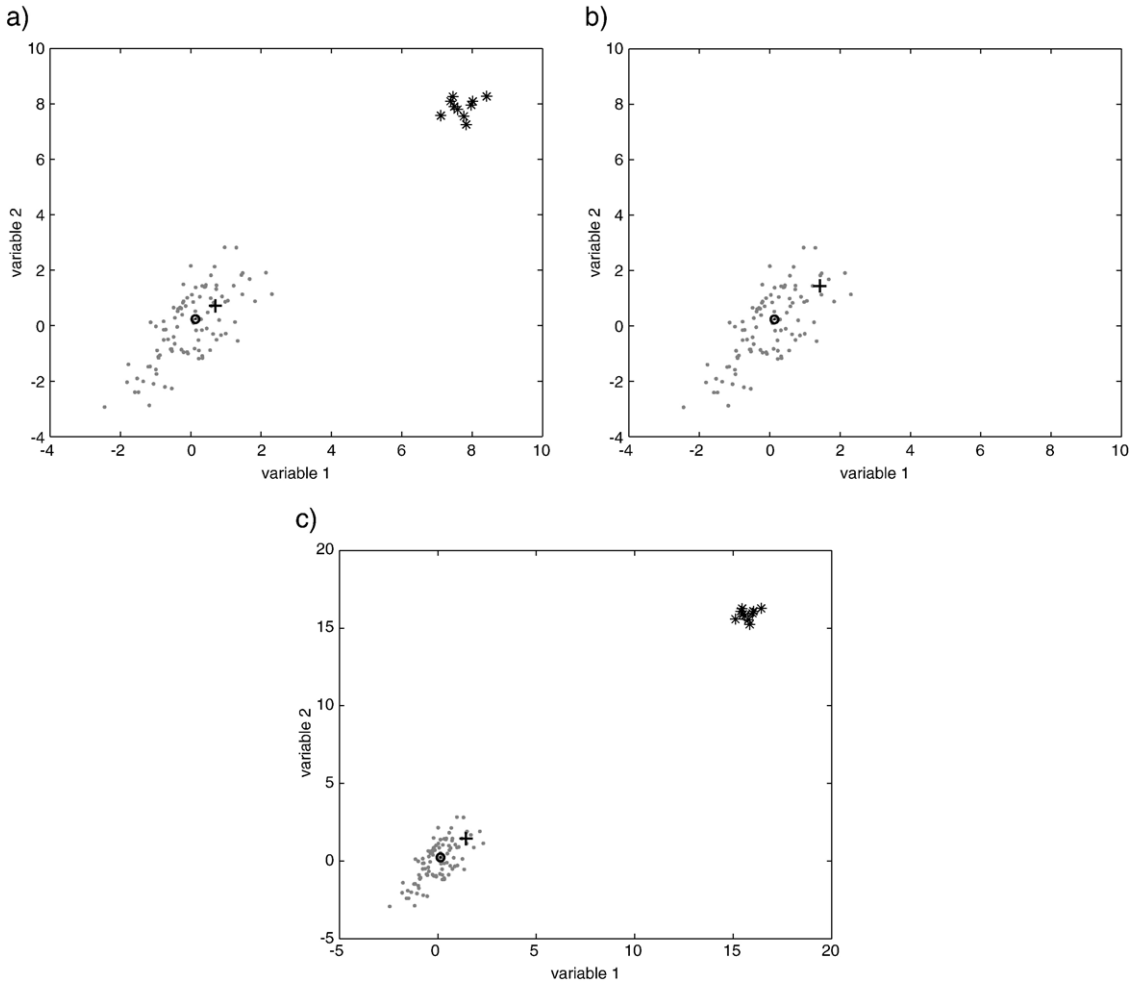

Fig. 1. Illustration of the outliers influence upon data center: a) centers of the data estimated by classical mean (+) and the L1-median (○); b) estimated centers of the data when the group of outliers (*) in the upper-right corner of the Fig. 1a was moved further from the data majority (the axes of the figure are as those in Fig. 1a); c) original axis range of the Fig. 1b.

effects, for instance, to remove offset from the data, to give equal impact of every variable in the analysis, to enhance the quality and interpretability of the obtained results. The most often applied transformations belong to the so-called affine transformations. An affine transformation is a combination of a linear transformation and translation. In geometrical sense affine transformation maps straight lines to straight lines, or in other words, a line after a transformation remains a line. The affine transformations can be presented as $\mathbf{XA}+\boldsymbol{1}^{\mathrm{T}}\boldsymbol{b}$, where $\mathbf{A}$ is any nonsingular matrix, and $\boldsymbol{b}$ is any row vector, and $\boldsymbol{1}^{\mathrm{T}}$ denotes the column vector with m ones $\boldsymbol{1}^{\mathrm{T}}=[1, 1,..., 1]^{\mathrm{T}}$.

The most popular data transformation, centering, aims to translate the multivariate data cloud to the data center. This can be expressed as:

$$\boldsymbol{x}_c = \boldsymbol{x} - \mu(\mathbf{X}) \tag{9}$$

where, $\boldsymbol{x}_{\mathrm{c}}$ denotes a column-wise centered variable of data $\mathbf{X}$.

Another very popular data transformation, autoscaling (also known as z-transformation), is applied in order to remove differences in the variables variances. Such transformation is necessary when variables of the data are in different ranges and units. The autoscaling is composed of data centering followed by scaling:

$$\boldsymbol{x}_{\mathrm{s}} = \frac{\boldsymbol{x}_{\mathrm{c}}}{\sigma(\boldsymbol{x}_{\mathrm{c}})} \tag{10}$$

When data contain outliers, the data mean as well as its standard deviation are no longer reliable estimates, and therefore, robust data preprocessing is required. Robust centering and robust autoscaling can be done using their robust variants such as median or L1-median, and the Qn, Sn or MAD scale estimators:

$$\boldsymbol{x}_{\mathrm{c}} = \boldsymbol{x} - \mathrm{median}(\boldsymbol{x}) \tag{11}$$

$$\boldsymbol{x}_{\mathrm{s}_i} = \frac{\boldsymbol{x}_{\mathrm{c}}}{\sigma_{\mathrm{Qn}}(\boldsymbol{x}_{\mathrm{c}})} \tag{12}$$

Taking into account the type of the transformation, the estimators can be divided into two groups, namely, affine and not affine.

The estimator T is said to be affine equivariant when:

$$T(\boldsymbol{x}_1\mathbf{A}+\boldsymbol{b},\boldsymbol{x}_2\mathbf{A}+\boldsymbol{b},...,\boldsymbol{x}_n\mathbf{A}+\boldsymbol{b}) = T(\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n)\cdot\mathbf{A}+\boldsymbol{b} \quad (13)$$

If the transformation $\mathbf{A}$ in Eq. (13) is orthogonal, i.e., $\mathbf{A}^{\mathrm{T}}=\mathbf{A}^{-1}$, then the estimator T is equivariant with respect to all transformations preserving the Euclidean distances. These transformations are rotation, reflection and translation.

2.4. Outlier detection by means of univariate approaches

The outliers can have either a univariate or multivariate nature. The univariate outliers are usually a result of an experimental error, and for their identification univariate approaches can be used. The multivariate outliers have a more complex nature and cannot be identified by univariate approaches. For multivariate outliers identification the exploratory tools, such as, for instance, Projection Pursuit-based techniques are considered [22].

In Fig. 2 univariate outliers are rather easy to identify, whereas the identification of multivariate outliers require multivariate techniques such as projection techniques. By projecting objects on one of the axes the outlier is located far from the data majority, and thus, it can be easily detected (see Fig. 2a-c). For the data presented in Fig. 2d none of the projections uncovers the outlier presence since it is located within the data cloud on the projections.

2.4.1. Box-plots

Box-plots give overall information about the univariate data distribution. In this plot data core is visualized as a rectangle. Its top and the bottom are the lower and the higher quartiles of the data distribution, respectively. A horizontal line in the rectangle denotes the data median. The lines above and below the rectangle, called whiskers, have a length equal to 1.5 times the interquartile range. Outliers are the objects located above or below the whiskers. Examples of box-plots constructed for two univariate distributions, with and without outliers, are presented in Fig. 3. For the first simulated example, with objects drawn from the normal distribution (Fig. 3a) no objects are above the whiskers (Fig. 3c). The second variable contains ten deliberately introduced outliers (Fig. 3b). Their presence is evident since they are located far above from the upper whisker of the box-plot (see Fig. 3d). The most extreme object, object no. 110, is the furthest from the upper whisker.

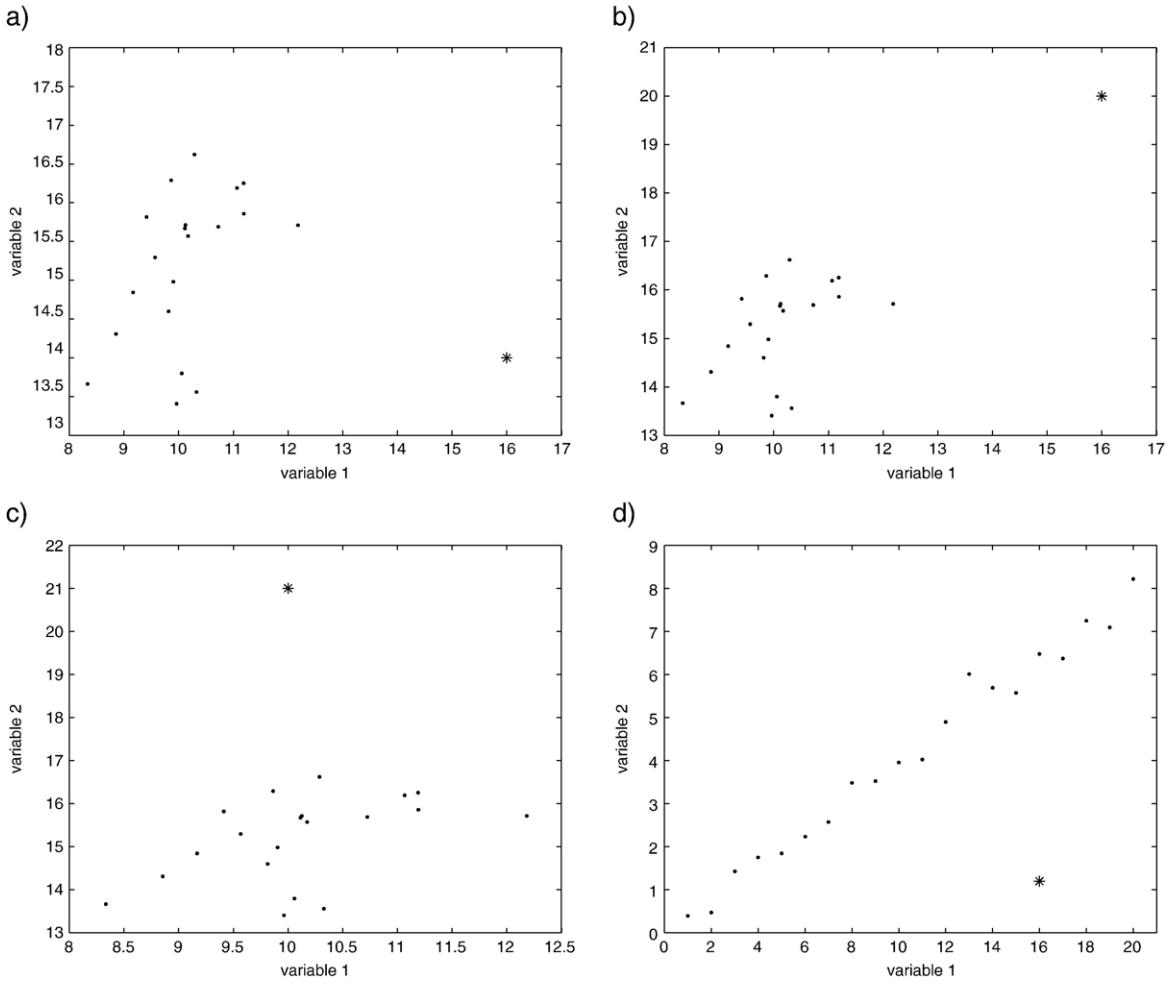

Fig. 2. Illustration of different types of outliers: a-c) univariate outliers (*) and d) a multivariate outlier (*).

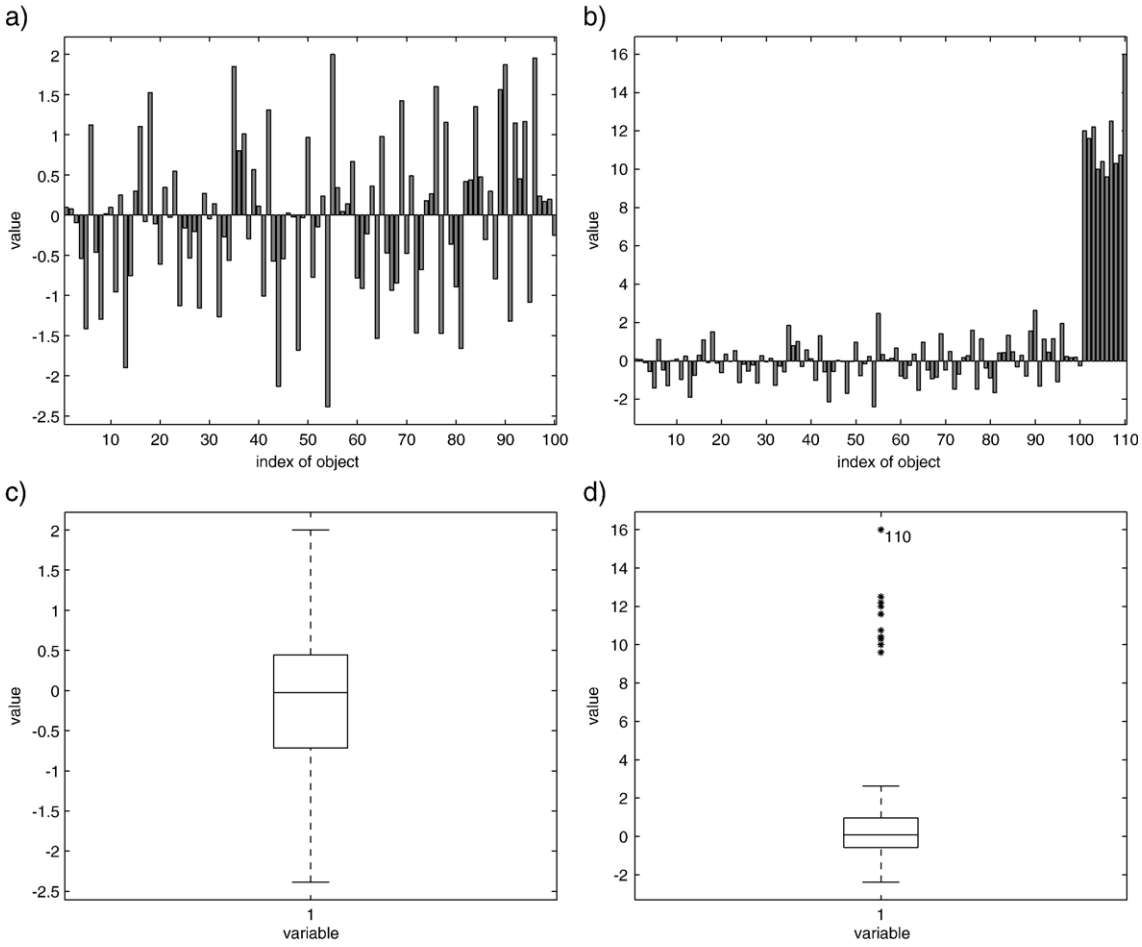

Fig. 3. Two simulated univariate distributions (a-b) and their corresponding box-plots (c-d), respectively.

2.4.2. Z-scores

Another way to identify outliers in univariate sense is to consider the so-called *z*-scores. To obtain absolute *z*-scores the elements of the variables are standardized by extracting from each element of the variable its mean and dividing it by the corresponding standard deviation:

$$z = \frac{|x - \mu(x)|}{\sigma(x)} \tag{14}$$

Then each object with a *z*-score greater than 2.5 or 3 can be identified as an outlier. The justification for these cutoff values comes from the assumption of normal distribution of the *z*-scores. Therefore, it is expected that 99.40% and 99.90% of centered objects lies within the interval of two and half and three times the standard deviation, respectively. However, the outliers influence estimates of the data mean and standard deviation, and thus also the *z*-scores. By considering robust mean of the data, i.e., median, and robust measure of the data spread, for instance σ_{Qn}, robust *z*-scores are obtained:

$$z = \frac{|x - \mathrm{median}(x)|}{\sigma_{\mathrm{Qn}}(x)} \tag{15}$$

It should be emphasized that *z*-scores are equivalent to the autoscaling transformation, also known as the *z*-transformation (see Eqs. (9)–(12)).

2.5. Covariance and correlation matrices and their robust estimates

The data covariance and correlation matrices play an important role in statistics and chemometrics. The information they give (variance of the variables, correlation between them and the covariance between the variables) is vital for the data analysis. The estimate of the covariance matrix, **C**, is defined as:

$$\mathbf{C} = \frac{1}{(m-1)} \cdot \left(\mathbf{X} - \mathbf{1}^{\mathrm{T}}\mu(\mathbf{X})\right)^{\mathrm{T}} \cdot \left(\mathbf{X} - \mathbf{1}^{\mathrm{T}}\mu(\mathbf{X})\right) \tag{16}$$

The matrix **C** is also called variance–covariance matrix since its diagonal elements are the variances of the data variables and the non-diagonal elements are the covariances between the *i*-th and the *j*-th data variables. It is said that two variables are uncorrelated, orthogonal or linearly independent when their covariance equals zero. It is noteworthy that the covariance matrix of the autoscaled data becomes a correlation matrix with Pearson's correlation coefficients as elements.

The covariance and correlation matrices are of great importance in many statistical and chemometrical techniques. For instance, the principal components are the eigenvectors of the covariance or the correlation matrix (the so-called spectral data decomposition). Since the variances of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ are the

same, their decomposition with PCA can be carried out on the smaller covariance matrix what reduces the computational time in many applications. This trick is a basis of the covariance-based PCA algorithms [14], and is also used in other chemometric approaches (SIM and WIM versions of the Partial Least Squares [15], Parafac2 [16,17], STATIS [18], *etc.*).

Due to the negative effect of outliers upon the estimates of the data mean and variance, also a covariance matrix is highly influenced by outliers. As an example let us consider a two-dimensional data set containing 100 objects. The projection of objects on the plane defined by variables 1 and 2 with a drawn 97.5% tolerance covariance ellipse is shown in Fig. 4. To illustrate the

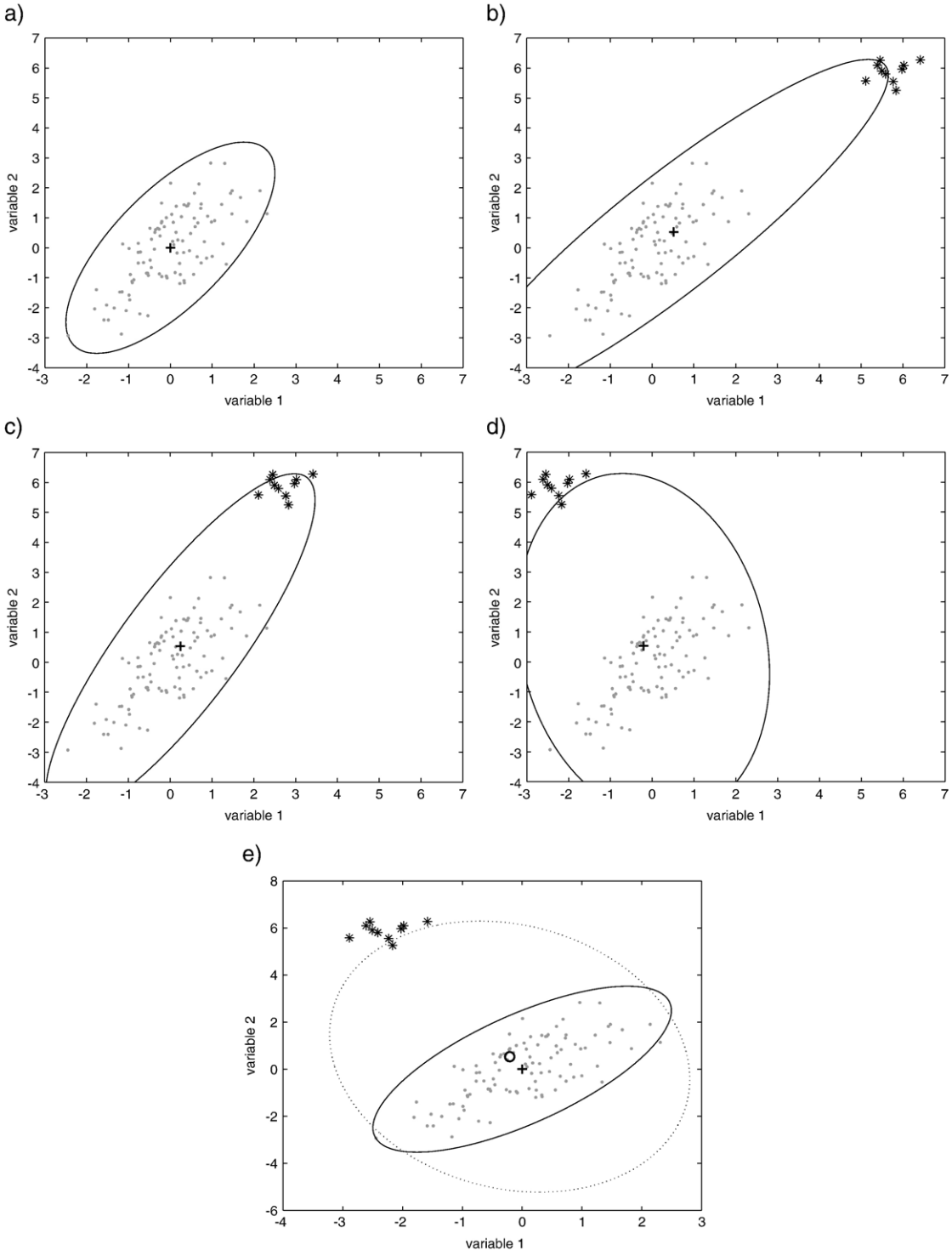

Fig. 4. Covariance ellipse with 97.5% tolerance constructed for: a) the simulated 2-dimensional data set; b-d) the data set presented in Fig. 4a additionally with ten outlying objects (*) located at different positions; e) covariance ellipses with 97.5% tolerance constructed for clean data (—) and the data with outlying observations (—), where (+) is the mean of the clean data, and (○) is the mean of the complete data.

outliers influence on the estimates of the data covariance, ten outliers were added to the data set. Their position with respect to the data majority is changed, as shown in consecutive panels of Fig. 4b–d. For comparison purposes in Fig. 4a–d the ranges of the axes are kept constant.

It can be seen that depending on the position of the outliers, the covariance ellipse strongly changes its direction (i.e., the correlation between variables is affected) and the ellipse volume. In Fig. 4e two covariance ellipses are simultaneously plotted. One for the clean data set (solid line), and the second one, represented by a dashed line, for the data with outliers. By means of this example it can be seen that the outliers affect the data covariance, and thus, change the true picture of correlation between variables and give a wrong impression about data mean and variance.

2.5.1. Robust estimators of the covariance matrix

As already shown, the data covariance is very sensitive to outliers. The use of the robust estimates of the data covariance gives possibilities to construct different robust approaches. From chemometrical perspective, the most welcome robust approaches are robust variants of Principal Component Analysis, robust pattern recognition techniques [19] (e.g. Soft Independent Modeling of Class Analogy (SIMCA), Unequal Class Modeling (UNEQ), different variants of discriminant analysis: Linear Discriminant Analysis (LDA), Regularized Discriminant Analysis (RDA), *etc*.) and robust calibration methods such as Partial Least Squares regression [20]).

In the literature, several approaches to obtain robust covariance matrices are proposed. Let us shortly present the most fundamental.

2.5.1.1. Multivariate trimming. The pioneer work on the robust data covariance is by Gnanadesikan et al. [21], who developed Multivariate Trimming (MVT). The method is based on calculating the Mahalanobis distances and the estimates of the covariance matrix iteratively. At every step of the method, an assumed fraction of objects with the highest Mahalanobis distances is removed from the data. The algorithm stops when the estimates of the mean of the retained subset of objects and covariance converge. The final mean and covariance estimates are derived for the clean subset of objects, i.e., objects with the smallest Mahalanobis distances. Although the MVT estimator is affine equivariant, nevertheless, its breakdown point depends strongly on the data dimensionality, being around $1/n$ only [22].

2.5.1.2. High breakdown point estimators of covariance. The high breakdown point estimators form another group of covariance estimators with affine equivariant properties. In the Minimum Volume Estimator (MVE), proposed by Rousseeuw [23,24], an ellipsoid of the smallest volume with a subset of p objects (non-contaminated data) is constructed. In one of the proposed iterative algorithms, MINVOL [22,25], in each iteration $n+1$ objects are selected iteratively at random and their mean and covariance are determined. Then, the ellipsoid containing exactly p data objects is found by deflating or expanding the data covariance. The steps of the algorithm are repeated until the subset of p objects yielding the smallest volume of the covariance ellipsoid is found. Also other algorithms for computing the MVE estimator are available but they are computationally exhaustive [26–28].

Although the MVE estimator has the 50% breakdown point, over time, it was replaced by another highly robust Minimum Covariance Determinant (MCD) estimator [22]. The aim of MCD is to find a subset of data objects with the smallest determinant of the covariance matrix. Key steps of this approach are presented in the Appendix A of this paper.

Compared to MVE, the MCD estimator of covariance offers a better statistical efficiency and accuracy. Moreover, the MCD estimator, computed with the FAST–MCD algorithm [29], is computationally more efficient than the MVE estimator. The FAST–MCD algorithm in MATLAB code is available from [30].Although the MCD estimator is often used, the other highly robust ones exist, being more efficient than MCD and with smooth influence function [31–33]. The most efficient robust estimators for covariance are MM estimators [34].

2.6. Leverage and Mahalanobis distance for outliers identification

The leverage, h_i, is a measure explaining the influence of certain objects on their own prediction:

$$\boldsymbol{h} = \mathrm{diag}(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}) \tag{17}$$

where 'diag' denotes the diagonal of matrix $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}$, whereas the matrix $\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}$ is the so-called leverage, hat or influence matrix.

The leverage values close to one, correspond to the objects located very close to the model hyperplane. Thus, the leverage is a measure of outlyingness and inform about the distance of the object from the data center.

There are many different distance measures, but the best known are the Euclidean and Mahalanobis distances. The distance of the i-th object from the data center is defined as:

$$d_i^2 = (\boldsymbol{x}_{\mathrm{c}_i}) \cdot \mathbf{Q} \cdot (\boldsymbol{x}_{\mathrm{c}_i})^{\mathrm{T}} \tag{18}$$

where $\boldsymbol{x}_{\mathrm{c}_i}$ denotes a centered i-th object of the data matrix $\mathbf{X}$, and $\mathbf{Q}$ is a diagonal matrix with ones on the diagonal for the Euclidean distance, whereas it is the inverse of the variance–covariance matrix $\mathbf{C}$ for the Mahalanobis distance.

The Mahalanobis distance takes into the account correlation between the data variables, contrary to Euclidean distance. A relationship between leverage and Mahalanobis distance exists, which can be expressed as:

$$\mathrm{MD}_i^2 = (m-1)\left(h_i - \frac{1}{m}\right) \tag{19}$$

where h_i is the i-th leverage computed with Eq. (17).

Although the Mahalanobis diagnostic is often used for detecting outliers in the data it should be kept in mind that it works well only for single case outlier problems [22]. This is due to the so-called masking effect that may occur when more outliers are in the data, affecting the variance–covariance matrix.

Further, a large outlier is able to mask the presence of other outliers in the data.

The use of the leverage and the Mahalanobis distance for outlier detection in chemical data is additionally limited by the fact that the data contain usually a large number of correlated variables. In such case the matrix ($\mathbf{X}^T\mathbf{X}$) becomes singular or nearly singular. To overcome this problem, the original data variables can be replaced by orthogonal principal components [35].

In Fig. 5 the effect of outliers upon the Mahalanobis distance is illustrated with simulated one-dimensional data with and without ten outlying samples. For the distribution presented in Fig. 5a its corresponding robust *z*-scores of the Mahalanobis distances (standardized Mahalanobis distances) are plotted in Fig. 5b. The standardized Mahalanobis distances computed for contaminated data uncover a presence of five outliers, whereas the remaining five are located close to the cutoff line. The impact of outliers upon the Mahalanobis distances is very clear especially when the Mahalanobis distances are compared with those obtained by constructing a covariance matrix for clean data only (see Fig. 5c).

In order to reduce the negative effect of outliers upon the Mahalanobis distances it is necessary to replace the classical estimates of the covariance matrix by its robust counterpart, for instance, robust covariance obtained with the MCD approach. Then, the outliers in the multivariate data are exposed to a higher extent [36].

2.7. From principal component analysis to its robust variants

2.7.1. Principal component analysis

Principal Component Analysis is the most widespread technique for data compression and visualization [37]. It aims to describe the data variance by constructing a set of new orthogonal features, called principal components (PCs). The PCs are a linear combination of the data variables and are mutually orthogonal. Every new PC describes a part of the data variance not explained by the previous ones. Due to this fact, usually a few first PCs are enough to represent well the data variance, and thus the visualization of the data structure is possible. The original data is decomposed by means of PCA into matrix of scores, **T**, and matrix of loadings, **P**:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T \tag{20}$$

The orthogonal character of the PCA as the data transformation ensures that the Euclidean distances among objects in

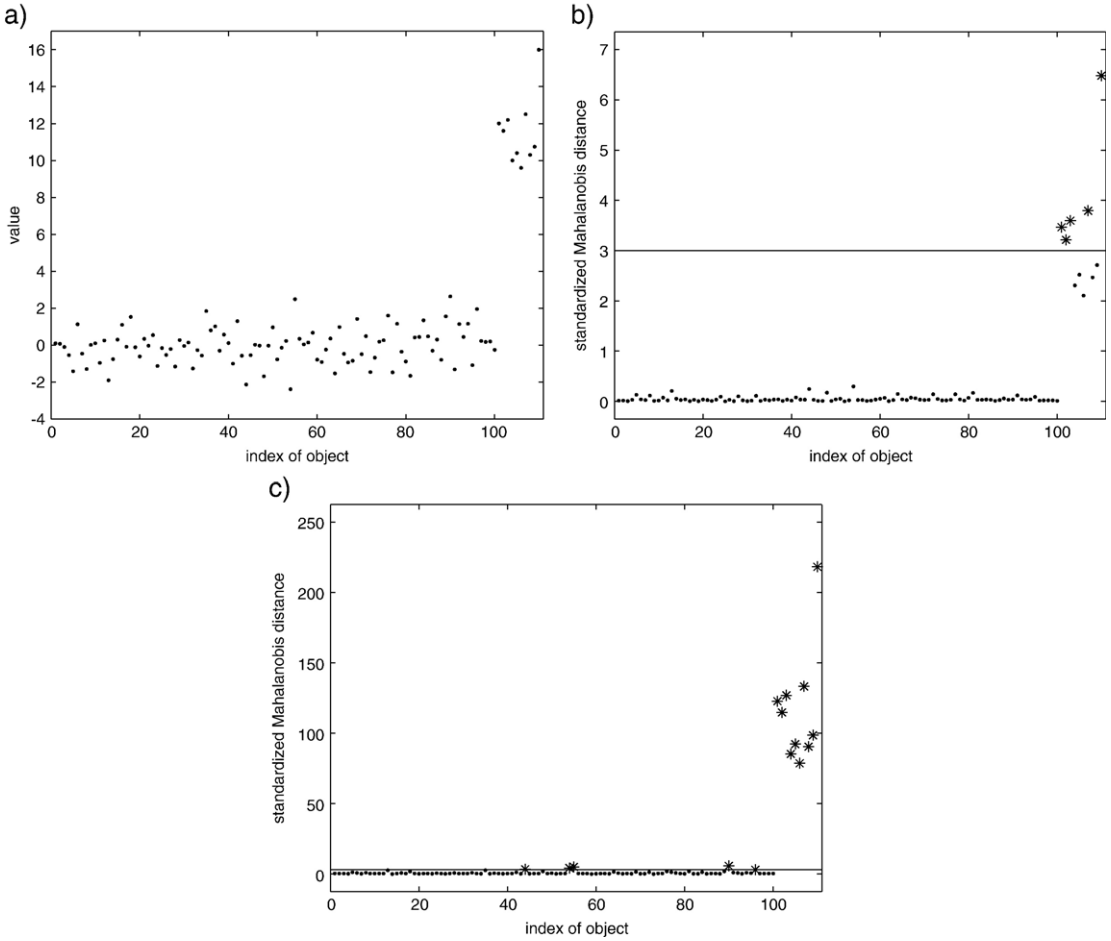

Fig. 5. a) a variable containing 90 objects drawn from normal distribution and 10 outliers, b) z-transformed classical Mahalanobis distances (standardized) - regular objects (·) and outliers (*), c) z-transformed Mahalanobis distances (standardized) computed using only clean subset - regular objects (·) and outliers (*).

the space of the PCs are preserved. The PCA properties make it a very attractive method not only for compression and visualization. Therefore, the PCA is usually applied as a first step of the data analysis and the PCs often serve as the input data for other approaches. PCA is also a core part of other chemometrical techniques including multivariate regression, curve resolution methods, discriminant analysis, *etc*. The need for PCA transformation in chemistry arises from the fact that most of the chemical data, such as spectral data, contain more variables than samples. This leads to the problem of processing correlated variables and singular (or close to singular) matrix $\mathbf{X}^T\mathbf{X}$. To overcome this problem the data variables are substituted by a set of orthogonal principal components. If the data matrix is decomposed by PCA to its full rank the information explained by the PCs is the same as the one contained in the original data set.

On another hand, PCA minimizes least squares cost function, and thus, is vulnerable to outliers. The outliers, and even a single one, tend to rotate the PCs axes towards them what changes the correlation structure of the data majority. Such effect of outliers is presented in Fig. 6. In Fig. 6a a simulated two-dimensional data set is presented as a projection of objects on the plane defined by the two data variables. The two solid lines are the first two principal components axes (a longer one it is PC 1 and the shorter one, orthogonal to PC 1 it is PC 2). Fig. 6b shows projections of objects on the plane defined by the two first principal components given in left panel. The next panels of Fig. 6 present a situation when the data contains outlying objects. Ten outliers were deliberately introduced into the data. Their location with respect to the data majority was changed (see Fig. 6c) in order to illustrate their influence upon PC axes. It is evident by means of this example that the outliers strongly rotate PC axes towards them (see Fig. 6d). It can be seen that the negative effect of outliers upon the construction of the principal components axes depends strongly on their position in the data space.

2.7.2. Robust PCA approaches

There are two classes of robust approaches proposed to diminish a negative effect of outliers upon PCs. The first group of approaches is based on PCA on a robust covariance matrix. This seems to be a straightforward way since the PCs are the eigenvectors of the covariance matrix. Therefore, replacing the covariance matrix by its robust variant leads to robust PCA. The differences among PCA approaches performed on robust

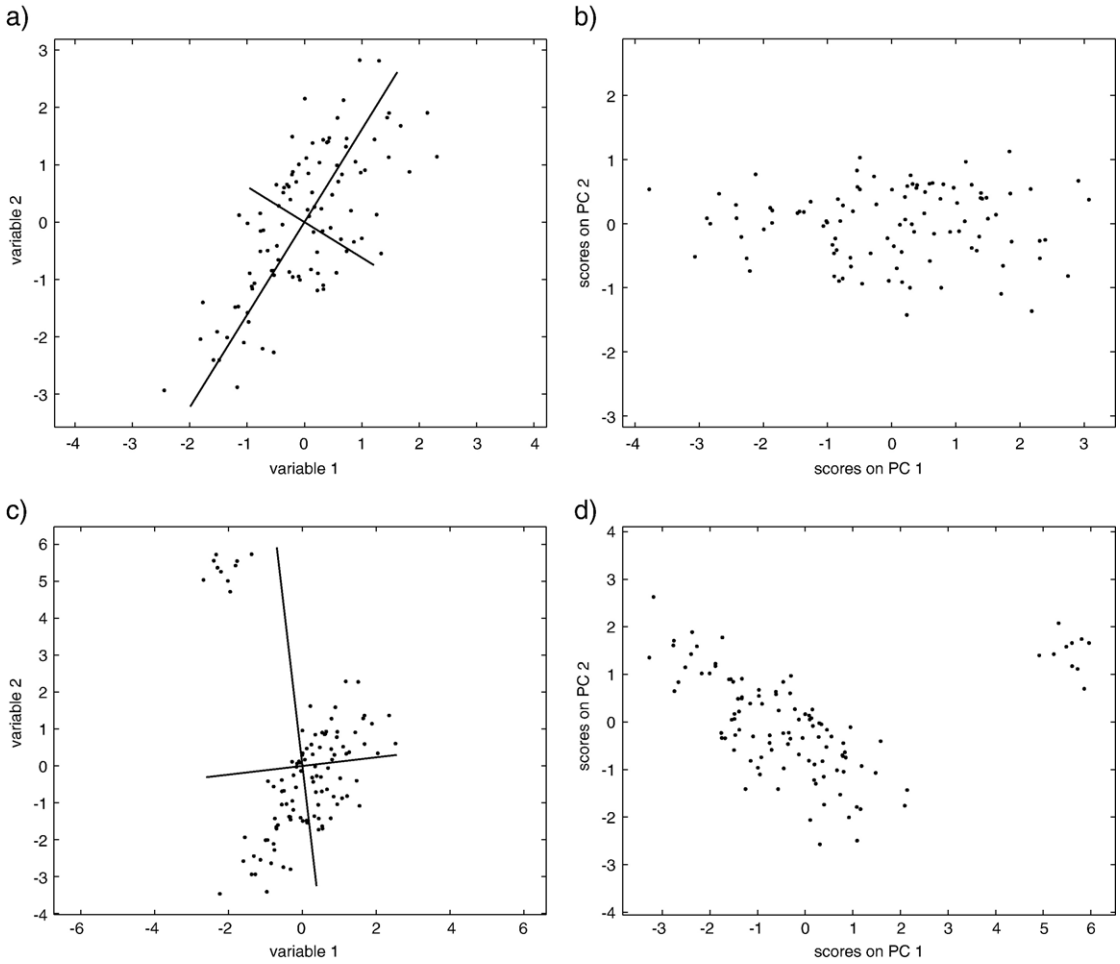

Fig. 6. Influence of outlying samples on the first two principal components:a) and c) directions of two first principal components in the space of two original variables; b) and d) projection of objects on the space defined by two first principal components.

covariance matrix are due to different construction of robust covariance and not the decomposition algorithm. Several robust estimators of the covariance matrix are proposed in the literature. Among the most popular there are MVT [38], MVE and MCD [22]. The earliest attempts to robustify the PCA method were due to Maronna [39] and Cambell [40]. Unfortunately, their approach does not allow handling larger fraction of outliers. Another proposal of robust PCA is due to Singh and Sing and Nocerino [41,42]. The next group of approaches is based on the idea of searching the projections of the data most exposing outliers. This is essentially the concept of the Projection Pursuit, PP [43]. In fact, all of the projection methods, such as for instance PCA, can be viewed from the perspective of projection pursuit. Replacing the objective function in Projection Pursuit gives a rise to other known techniques. For instance, PP with a variance as objective function leads to PCA, PP maximizing a measure of non-Gaussianity (for instance kurtosis) leads to Independent Component Analysis [44], and PP maximizing robust measure of scale brings us to robust PCA. The first robust PCA approach based on the idea of PP was developed by Li and Chen [45]. Later on, the work on robust PCA was continued by Ammann [46], Xie et al. [47] and Hove et al. [48], and Hubert et al. [49]. The last publications brought more attention of the chemometrical community to problem of outliers in PCA. Another robust method was described by Croux and Ruiz-Gazen [50,51], who exploits the idea of Projection Pursuit applying robust measure of the data scale as a projection pursuit criterion. A similar approach with kurtosis projection index is described by Peña and Prieto [52].

2.7.2.1. PCA on the robust covariance matrix. A more recent proposal suggests using high breakdown estimators such as MCD [53] to derive a robust covariance matrix. Since the chemical data often contain more variables than objects, it appears that robust estimators such as MCD, MVT, or MVE, cannot be directly applied in majority of chemical applications. However, the data can be compressed to orthogonal principal components by means of PCA. As already mentioned, this orthogonal data transformation keeps unchanged the Euclidean distances among objects in the PCs space. The principal components can serve later to derive robust estimates. This is essentially proposed in reference [54]. The same philosophy of using orthogonal variables is followed by the others, for instance, the robust PCA based on the MCD estimator of the covariance matrix has been recently proposed [55].

2.7.2.2. Projection pursuit with robust estimate of data scale. The goal of Projection Pursuit, PP, is to find directions in the data space, such that the projections of the data onto these directions maximize a certain criterion. Usually, this criterion is a measure of non-Gaussianity, like entropy. Since distribution of objects with outliers on the projection is far different from the projection revealing normal distribution, selection of entropy as a projection index seems to be a straightforward choice. The concept of PP is fundamental in constructing robust PCA technique.

As reported by Stahel [56] and Donoho [57], who independently developed the so-called outlyingness-weighted mean, a highly robust location estimator can be obtained by finding a low-dimensional projection of the data, which emphasizes the outlying objects. For this purpose they propose a measure of outlyingness given as:

$$w_i = \max_{||\boldsymbol{p}||=1} \frac{|\boldsymbol{x}_i\boldsymbol{p}^{\mathrm{T}} - \underset{j}{\mathrm{median}}(\boldsymbol{x}_j\boldsymbol{p}^{\mathrm{T}})|}{\mathrm{median}_k|\boldsymbol{x}_k\boldsymbol{p}^{\mathrm{T}} - \mathrm{median}_j(\boldsymbol{x}_j\boldsymbol{p}^{\mathrm{T}})|} \tag{21}$$

where $\boldsymbol{p}$ is a direction and expression $\boldsymbol{x}_j\boldsymbol{p}^{\mathrm{T}}$ denotes a projection of the i-th object $\boldsymbol{x}_i$ on the direction $\boldsymbol{p}$.

Please notice that the Eq. (21) presents in fact robust z-scores, already introduced in Section 2.3, where a projection ($\boldsymbol{x}\boldsymbol{p}^{\mathrm{T}}$) is taken as argument. In robust z-scores instead of centering and standardizing the variable with classical mean and standard deviation estimates, median and the MAD are used. Of course, one can also use other robust measure of the data scale, for instance, σ_{Qn}. What is essential in this concept is that projections without outliers have small w_i values, contrary to projections that expose outliers.

It is noteworthy that for a univariate projection such as e.g. $\boldsymbol{x}_i\boldsymbol{p}^{\mathrm{T}}$, the Eq. (21) is closely related to the Mahalanobis distance, which can be also expressed as:

$$\mathrm{MD}_i = \max_{||\boldsymbol{p}||=1} \frac{|\boldsymbol{x}_i\boldsymbol{p}^{\mathrm{T}} - \frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{x}_i\boldsymbol{p}^{\mathrm{T}})|}{\sigma(\boldsymbol{x}_1\boldsymbol{p}^{\mathrm{T}}, \ldots, \boldsymbol{x}_m\boldsymbol{p}^{\mathrm{T}})} \tag{22}$$

Therefore, by introducing in Eq. (22) robust estimates of data center and scale and using this modified criterion in PP, robust PCs are obtained. This is essentially done in the Croux and Ruiz-Gazen robust PCA algorithm [50]. One problem which remains is how to find optimal directions. Searching of such directions is an optimization problem, and therefore it is computationally exhaustive. To facilitate this task looking for an optimal direction can be restricted to the directions defined by the data objects. One should however keep in mind that looking for the best direction only among the directions defined by the data objects is just an approximation of the true solution. The risk of finding a solution being far from the optimal one grows when the number of objects in the data is small. One way to improve the quality of the obtained results is to enlarge a set of examined directions as done by Serneels et al. [58]. A more precise algorithm for PP-PCA has been proposed in [59]. It can be done by considering additionally a set of randomly generated directions. The algorithm relays on projecting the data objects onto a set of directions, and thus it is computationally efficient. Another trick to speed up the algorithm is to use PCs obtained by compressing original data to the full rank using kernel PCA.

2.7.2.3. Trimmed objects projections. The principle of the Trimmed Objects Projections approach (TOP) [48] is to find directions in the data such that the projections of objects onto these directions have maximal sum of absolute values of projections. At the beginning, the data objects are weighted with weights w corresponding to their distances from the data center defined by coordinate-wise median. Each object with the distance larger than the median of all distances receives weight w_i, equal to 0.01, or otherwise, equal to 1. Such weighting of

objects diminishes the outliers influence, and at the same time, the use of median ensures 50% break down point of the method.

Then, the cosines of the weighted (with weights w) projections of the data on the directions defined by the data objects are computed. The robust principal component is constructed by finding a so-called principal object with a maximal sum of absolute values of projections. In the next step, to the coordinates of the principal object (its variables), sum of coordinates of the remaining objects, weighted by the squares of their projections onto the principal object are added. In this way, the objects representing the data majority have the largest projection weights and thus, contribute the most in constructing a robust PC. The new directions are found in the space of residuals. A more detailed description of the algorithm can be found in reference [48] and in the article Appendix A.

2.7.2.4. Robust PCA with Croux and Ruiz-Gazen algorithm. In the robust PCA algorithm proposed by Croux and Ruiz-Gazen [50] the principal components are defined as projections of the data onto directions maximizing a robust scale estimate. For this task robust Qn scale can be used, since the Qn estimator compared to the other ones offers a good efficiency. In original version of the algorithm, in every step, all directions that are formed by the data objects are examined. The direction that maximizes the Qn scale of the projected objects is regarded as the optimal one and the projection is a robust principal component. Consecutive robust PCs are found in the residual data space. A more detailed description of the algorithm is presented in the Appendix A of the article.

2.7.2.5. ROBPCA algorithm. The ROBPCA algorithm, proposed by Hubert et al. [55,60], combines the idea of the Projection Pursuit and the robust estimation of the covariance and location. The similarity of the ROBPCA approach and PP is due to optimizing the outlyingness measure (which in PP terminology is a projection index), and namely, the directions that maximize the outlyingness measure are found. The ROBPCA can be summarized in two steps. As in the Croux's algorithm, in ROBPCA, the directions pass through two individual data objects. In order to speed up computations, first the data is compressed to principal components defining potential directions. Then, each i-th direction is scored by its corresponding value of outlyingness, w_i:

$$w_i = \underset{||p||=1}{\operatorname{argmax}} \frac{|\boldsymbol{x}_i\boldsymbol{p}^{\mathrm{T}} - \mu_{\mathrm{MCD}}(\boldsymbol{x}_i\boldsymbol{p}^{\mathrm{T}})|}{\sigma_{\mathrm{MCD}}(\boldsymbol{x}_i\boldsymbol{p}^{\mathrm{T}})} \tag{23}$$

In the second step, an assumed fraction of objects, with the smallest values of w_i, is used to construct a robust covariance matrix. At the end, the PCA model is built for the robust covariance matrix. Based on the obtained PCA model, the remaining data objects are projected on the space defined by the robust loadings. A more detailed description of the algorithm can be found in [55] and Matlab implementation of this robust PCA method and other robust techniques [61] are available from [62].

2.7.2.6. Elliptical and spherical principal component analysis. Spherical PCA, SPCA, is another variant of robust PCA [63]. The idea behind Spherical PCA is to project all objects on the hyper sphere with radius equal 1 and center placed in the robust center of data. The robust center of data is defined as:

$$\hat{\boldsymbol{\mu}} = \min \sum_{i=1}^{n} \left|\left|\boldsymbol{x}_i - \boldsymbol{\mu}\right|\right|^k \tag{24}$$

This estimate of location, introduced by Huber [2], is the so-called "L^k M-estimate of location". In SPCA k equal to 1 is considered and $\hat{\mu}$ becomes the well-known L1-median [5]. Then, all data points are projected on the sphere with the center defined by L1-median of the data and unit radius, $\mathrm{Sph}(\mu_{\mathrm{L1}}(\mathbf{X}),1)$. The projection of data objects, $P_{\mathrm{Sph}(\mu_{\mathrm{L1}}(x),1)}\mathbf{X}$, is described by the following expression:

$$P_{\mathrm{Sph}(\mu_{\mathrm{L1}}(x),1)}\mathbf{X} = \frac{x_i - \mu_{\mathrm{L1}}(\boldsymbol{X})}{||x_i - \mu_{\mathrm{L1}}(\boldsymbol{X})||} + \mu_{\mathrm{L1}}(\mathbf{X}) \tag{25}$$

The example of data projection on a sphere is shown in Fig. 7. As shown in Fig. 7, the projected data points preserve the structure of the original data. However, the influence of outlying objects is bounded. This is done by down weighting them according to their distances from the robust data center defined by the L1-median. Now, it is possible to construct a robust PCA model based on the projected observations. By projecting original data onto robust loadings the robust scores can be obtained.

In reference [63] also another version of robust PCA, called Elliptical PCA (EPCA), is proposed. This variant of robust PCA takes into account different scales of the data variables, and instead of hypersphere the objects are projected onto a hyperellipse. The radii of hyperellipse are proportional to the σ_{Qn} scale estimator of each variable.

2.7.2.7. Robust PCA with a robust cost function. The main goal of the PCA model is to find a subspace of the f principal components such that the distances of objects to this subspace (residuals) are the smallest. Due to this condition, the PCA is a least squares model, and as all of the least squares models, is affected by outliers. Another way to robustify the PCA approach is to replace the classical least squares cost function, being

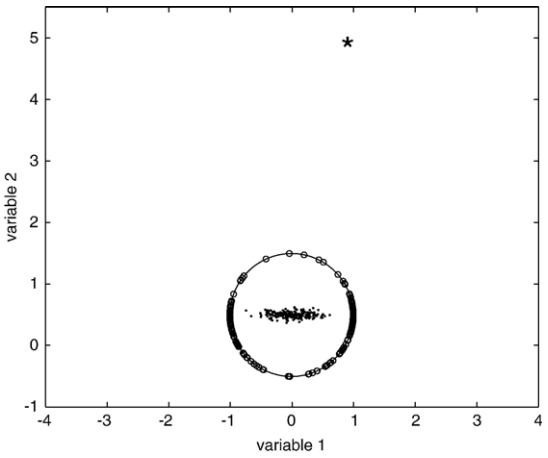

Fig. 7. Example of projected two-dimensional data (○) on a sphere. An outlier is denoted as (*).

minimized while constructing the principal components, by its robust variant. If the PCA is seen from the perspective of the data model, its residuals, i.e., the differences between original data and the data reconstructed with the PCA model with f components, can be used to derive a robust cost function that should be minimized. A simple choice to construct a robust cost function for PCA is to minimize the trimmed residuals of the PCA model with the Least Trimmed Squares estimator (LTS estimation) or M-estimator of the PCA residuals [22,64]. In order to obtain robust PCA it is necessary to find a subset of p objects for which the PCA model offers the smallest residuals for all objects. A detailed description of the algorithm for robust PCA can be found in [64].

2.7.3. Diagnostic plots based on the robust data models

The PCA and the robust PCA can be also viewed from the perspective of the data models. For the PCA model these are classical scores and classical loadings, and for the robust PCA model, robust scores and robust loadings:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \tag{26}$$

When the data is decomposed not to the full rank then, the residual term, **E**, representing the difference between the original and the predicted data with the PCA model with k principal components, occurs:

$$\mathbf{X} = \mathbf{T}_j\mathbf{P}_j^{\mathrm{T}} + \mathbf{E} \tag{27}$$

where $j=1{:}k$, **X** is the reconstructed data matrix based on the PCA model with k principal components.

The PCA model allows obtaining scores for a new object, by projecting this object onto the directions defined by the loadings. Then, score values and the loadings, can be used to calculate residuals for the new i-th object, e_i. They are obtained as a sum of squared differences between the observed x-values of the i-th object and the predicted ones with the PCA model:

$$e_i = \sum_{i=1}^{n} (x_i - \hat{x}_i)^2 \tag{28}$$

where e_i and $\hat{x}_i$ are the residual and predicted values for the i-th object based on the PCA model, respectively.

With respect to the PCA model, the data objects can be classified into four groups. The regular objects, being the first group, are relatively close to each other in the space of principal components (or robust principal components) and are located close to center of normally distributed data. Moreover, the residuals of the regular objects from the PCA model are small. The second type of objects, the so-called good leverage objects, are far from the data majority but their residuals from the PCA model are small (see Fig. 8a). The orthogonal outliers have large residuals from the PCA model but, if projected on the PCA space, they fall into the cloud of the data majority (see Fig. 8b). Therefore orthogonal outliers cannot be distinguished on the projections. The last category of data objects, bad leverages, are located far from the data majority and they have large residuals from the PCA model (see Fig. 8c).

Taking into account objects characteristics two distances are proposed for outliers detection, and namely, score and orthogonal (residual) distances. The score distance of the i-th object, SD_i, is defined as:

$$\mathrm{SD}_i = \sqrt{\sum_{j=1}^{k} \frac{\boldsymbol{t}_j^2}{v_j}} \tag{29}$$

where $j=1, 2, \ldots, k$ denotes number of robust principal components, $\boldsymbol{t}_j$ is the j-th principal component and the v_j is its eigenvalue.

The orthogonal distance of the i-th object, OD_i, is given as:

$$\mathrm{OD}_i = \sqrt{\sum_{i=1}^{n} (\boldsymbol{x}_i - \mu_{\mathrm{L1}}(\mathbf{X}) - \boldsymbol{p}_j\boldsymbol{t}_j)^2} \tag{30}$$

Based on these two distances diagnostic plots, similar to the ones presented by Rousseeuw and van Zomeren [36], can be constructed by plotting the score distances *versus* the orthogonal distances.

While constructing the robust PCA model and the diagnostic plots, two issues are of a large importance, and namely, selection of the number of robust PCs in the final model and definition of the cutoff values for the score and orthogonal distances. The selection of the number of the robust PCs, analogously as in classical PCA, is based on the robust eigenvalues. One way is to analyze a decrease of the consecutive eigenvalues in the robust PCA model [49]. An alternative to the above mentioned approach is to examine a ratio between each robust eigenvalue to the sum of all eigenvalues. The number of robust eigenvalues for which the ratio is approximately around 0.9 can be considered as optimal number of the robust PCs in the model [55].

For the score distances, the proposed cutoff value comes from the critical values of the chi-squared distribution for k components in the robust PCA model, $\sqrt{\chi^2_{k,0.975}}$ [49]. For the orthogonal distance two cutoff values were proposed. The one based on the chi-squared distribution as for the score distances [49], and the second one, the Wilson–Hilferty approximation of the chi-squared distribution [55].

What should be emphasized is that the selection of appropriate number of the robust PCs in the model is a critical and difficult task. Dependent on the number of selected robust PCs, the score and orthogonal distances as well as their corresponding cutoff values change.

An alternative to the above described cutoff values could be the use of robust z-scores of the score and orthogonal distances (standardized robust and orthogonal distances) with default cutoff value equal to 2.5 or 3. Although there are several possible choices for the robust scale used to calculate robust z-scores, the σ_{Qn} estimator of data scale is preferred due to its high efficiency and 50% breakdown point. The robust standardized z-score, ds_i, of the i-th object distance is expressed as:

$$\mathrm{ds}_i = \frac{|d_i - \mathrm{median}(\boldsymbol{d})|}{\sigma_{\mathrm{Qn}}(\boldsymbol{d})} \tag{31}$$

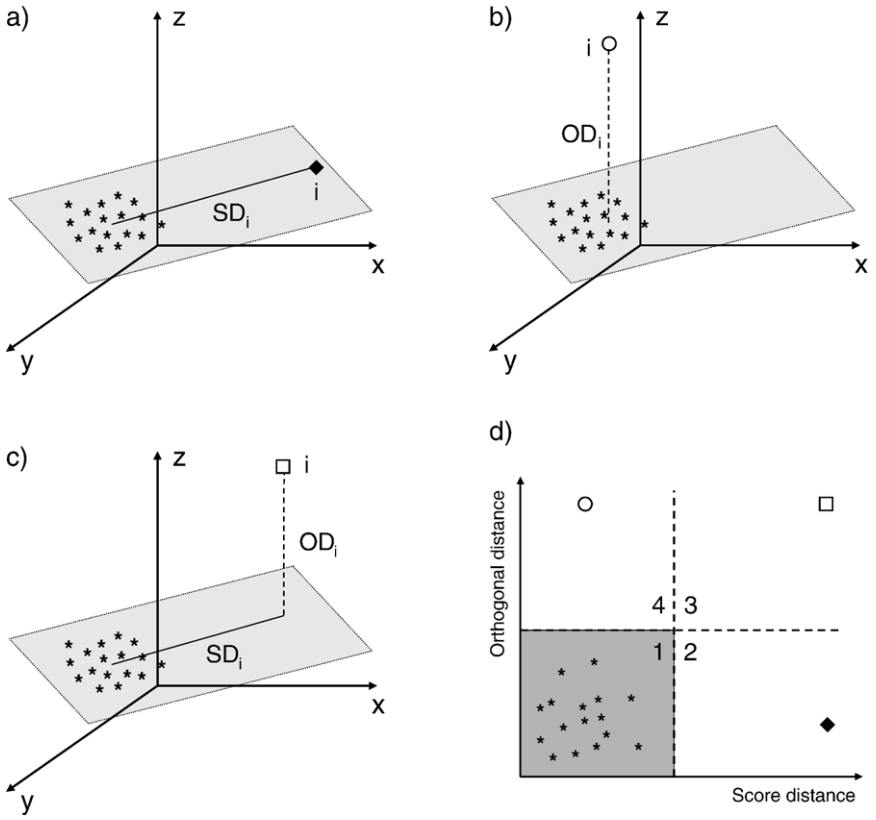

Fig. 8. Different types of outlying objects with respect to their Score and Orthogonal distances: a) good leverage objects (♦), b) orthogonal outlier (○), c) bad leverage object (□) e) distance-distance plot presenting score distances of objects versus their orthogonal distances ((*) regular observations, good leverage object (♦), orthogonal outlier (○), bad leverage object (□)).

where ***d*** is the vector, which elements represent the distances, ds_i is the robust standardized *z*-score of the *i*-th object, and σ_{Qn} is the Qn scale estimator applied to vector ***d***.

3. Conclusions

In reality, chemical data often contain outlying objects. For all techniques that employ a least square cost function, e.g. PCA or a wide range of calibration techniques including MLR, PCR, PLS, *etc*., a single outlier in the data can completely change the overall trend and make a model no longer valid for the data majority. By plugging in robust estimators to classical approaches the problem of outliers can be handled and good models describing well data majority can be constructed. A simple example of such procedure is a robust PCA obtained by replacing classical estimates of the data covariance by the robust one. The robust approaches become more popular in chemometrics. This is due to the fact that now they can be used for highly multivariate chemical data. Such method as PCA, when applied to the data allows data dimensionality reduction by replacing the original variables with orthogonal ones. Due to efficient data compression these orthogonal variables can be later used for instance in MCD or robust PCA instead of the original variables. This breaks a common problem of applying the MCD approach to the data where the number of samples is much smaller than the number of variables. Substituting original data variables by the limited number of orthogonal principal components seems to be for chemical applications already a routine, e.g. MCD, ROBPCA, *etc*.

Many of the robust approaches have the highest possible breakdown point, however a high breakdown point can lead to loss of accuracy of the estimator. When it is possible, the breakdown point should be tuned, depending on a problem at hand. This enables achieving a compromise between good robust properties of an estimator and the quality of the estimates.

There are a number of ongoing challenges concerning the use and development of robust techniques. Since some new methods for data analysis are introduced one can also expect their robust versions. For instance relatively new method in the field of chemometrics such as Support Vector Machines, SVM, has already its robust version [65]. A computational efficiency of some robust algorithms is also an important issue, and therefore their faster versions are always welcome. Somewhat less explored at the moment problem is the problem of co-existing missing elements and outliers in the studied data. Such data are typical in environmental studies. That is why robust methods, able to handle outliers and missing elements simultaneously, are strongly required. Successful attempts to derive a strategy to deal simultaneously with outliers and missing elements in the data for exploration and modeling purposes are reported in [66,67].

Acknowledgements

M. Daszykowski would like to express his sincere gratitude to the Foundation for Polish Scientists for financial support.

K. Kaczmarek's research work was supported by Bilateral Grant no BWS 03/07 by Flemish and Polish governments.

B. Walczak and Y. Vander Heyden are grateful for financial support concerning scientific activities within the Sixth Framework Programme of the European Union, project TRACE — "TRAcing food Commodities in Europe" (project no. FOOD-CT-2005-006942). The publication reflects only the author's views and the Community is not liable for any use that may be made of the information contained therein.

Appendix A.

A.1. The minimum covariance determinant algorithm

Five hundred times:

1. select at random a subset of p objects. The p value is set as a default and equal to:

$$\frac{1}{2}(m+n+1)$$

where m is the number of objects and n denotes the number of variables in the data;

2. for the selected subset compute its mean, covariance and the Mahalanobis distances; perform steps 3 and 4 twice:

3. sort the Mahalanobis distances from the smallest to the largest and select the p objects with the smallest Mahalanobis distance;

4. for the p objects compute the mean and the covariance and calculate the Mahalanobis distances for all of the object;

5. within the algorithm run collect the ten subsets of objects with the smallest determinant of the covariance matrix;

6. after 500 iterations, perform steps 3 and 4 on the best clean subsets till convergence;

7. based on the clean subset, compute the Mahalanobis distances and carry on the final outlier detection.

A.2. Robust PCA with Croux and Ruiz-Gazen algorithm

1. center data matrix, $\mathbf{X}$ (m, n) around L1-median to obtain $\mathbf{X}_c$, and determine its rank r:

$$\mathbf{X}_c = \mathbf{X} - \boldsymbol{I}^T l_{L1}(\mathbf{X})$$

2. construct the directions $\boldsymbol{p}_i$ as normalized rows of matrix $\mathbf{X}_c$;

$$\boldsymbol{p}_i = \frac{\boldsymbol{x}_{c_i}}{||\boldsymbol{x}_{c_i}||}$$

where $i=1, 2,\ldots, m$;

3. project all objects on the possible directions:

$$\boldsymbol{t}_i = \mathbf{X}_c \boldsymbol{p}_i^T$$

4. calculate robust scale of all projections and find the direction $\boldsymbol{q}$ that maximize the σ_{Qn} of projection:

$$\boldsymbol{q} = \max_i \left(\sigma_{Qn}(\boldsymbol{t}_i)\right)$$

5. project all objects on the selected direction $\boldsymbol{q}$ to obtain robust principal component:

$$\boldsymbol{t}_i = \mathbf{X}_c \boldsymbol{p}_q^T$$

6. update data matrix by its orthogonal complement:

$$\mathbf{X}_c = \mathbf{X}_c - (\boldsymbol{p}_q \boldsymbol{p}_q^T) \cdot \mathbf{X}_c$$

7. if number of extracted robust principal components is lower than f go to step 2.

A.3. Trimmed objects projection algorithm

1. center the data around coordinate-wise median:

$$\mathbf{X}_c = \mathbf{X} - \mathbf{1}^T \text{median}(\mathbf{X})$$

2. calculate the Euclidean distances between each data object and the data center as:

$$d_i = ||\boldsymbol{x}_{c_i}|| = \sum_{j=1}^{n} (x_{c_{ij}})^2,$$

where ||...|| denotes the L2-norm.

3. construct weighs w_i as:

$$w_i = \begin{cases} 1 & \text{for } d_i \le \text{median}(d_i) \\ \frac{1}{\alpha} & \text{for } d_i \rangle \text{ median}(d_i) \end{cases}$$

where α is set to 100.

4. compute cosines of the weighted projections:

$$c_{ij} = w_i \frac{\boldsymbol{x}_{c_i} \boldsymbol{x}_{c_j}^T}{||\boldsymbol{x}_{c_i}|| \cdot ||\boldsymbol{x}_{c_j}||}$$

5.find direction $\boldsymbol{q}$, which defines a direction with maximal sum of the $|c_{ij}|$ and construct a new direction as:

$$\boldsymbol{p}_i = \boldsymbol{x}_{c_q} + \sum_{j \neq q} c_{qj}^2 \cdot \text{sign}(c_{qj}) \cdot \boldsymbol{x}_{c_j}$$

6. normalize the direction $\boldsymbol{p}_i$:

$$\boldsymbol{p}_i = \frac{\boldsymbol{p}_i}{||\boldsymbol{p}_i||}$$

7. construct robust principal component by projecting $\mathbf{X}_c$ on the direction $\boldsymbol{p}_i$:

$$\boldsymbol{t}_i = \mathbf{X}_c \boldsymbol{p}_i^T$$

8. if the number of the robust PCs is smaller than f, compute residual matrix and return to step 2:

$$\mathbf{X}_c = \mathbf{X}_c - \boldsymbol{t}_i \boldsymbol{p}_i^T$$

A.4. Elliptical and spherical principal component analysis

1. find robust center of the data (L1-median, $\mu_{L1}(\mathbf{X})$);
2. in case of:

a) SPCA project all data points onto a hypersphere with radius equal to one and center in robust center found in step 1:

$$P_{\mathrm{Sph}(\mu_{L1}(\boldsymbol{x}),1)}\mathbf{X} = \frac{\boldsymbol{x}_i - \mu_{L1}(\mathbf{X})}{||\boldsymbol{x}_i - \mu_{L1}(\mathbf{X})||} + \mu_{L1}(\mathbf{X})$$

where ||...|| is the L2-norm;

b) EPCA project all data points onto hyperellipse with radii proportional to the robust standard deviation of each original variable and center in robust center found in step 1:

$$P_{\mathrm{Sph}(\mu_{L1}(\boldsymbol{x}),1)}\mathbf{X} = \frac{\frac{\boldsymbol{x}_i}{\sigma_{Qn}} - \mu_{L1}(\mathbf{X})}{||\frac{\boldsymbol{x}_i}{\sigma_{Qn}} - \mu_{L1}(\mathbf{X})||} + \mu_{L1}(\mathbf{X})$$

3. construct PCA model for projected data and obtain scores (T) and loadings (P) matrices;
4. project original data onto principal components found in step 3 to obtain robust PCA scores for original data:

$$\boldsymbol{t}_i = \mathbf{X}_c \boldsymbol{p}_i.$$

References

[1] F.R. Hampel, A general qualitative definition of robustness, Annals of Mathematical Statistics 42 (1971) 1887–1896.
[2] P.J. Huber, Robust Statistics, Wiley, New York, 1981.
[3] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: the Approach Based on Influence Functions, Wiley, New York, 1986.
[4] A. Weber, Über den standort der industrien, Tubingen. English translation by C. Friedrich (1929): Alfred Weber's Theory of Location of Industries, University of Chicago Press, 1909.
[5] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York, 1987.
[6] C.G. Small, A survey of multidimensional medians, International Statistical Review 58 (1990) 263–277.
[7] B.M. Brown, Statistical uses of the spatial median, Journal of the Royal Statistical Society. Series B 45 (1983) 25–30.
[8] C. Sârbu, H.F. Pop, Fuzzy robust estimation of central location, Talanta 54 (2001) 128–130.
[9] R. Rajkó, Treatment of model error in calibration by robust and fuzzy procedures, Analytical Letters 27 (1994) 215–228.
[10] F.R. Hampel, The influence curve and its role in robust estimation, Journal of the American Statistical Association 69 (1974) 383–393.
[11] P.J. Rousseeuw, C. Croux, Alternatives to median absolute deviation, Journal of the American Statistical Association 88 (1993) 1273–1283.
[12] C. Croux, P.J. Rousseeuw, Time-efficient algorithms for two highly robust estimators of scale, in: Y. Doge, J. Whittaker (Eds.), Computational Statistics, vol. 1, Physica-Verlag, Heidelberg, 1992, pp. 411–428.
[13] http://www.chemometria.us.edu.pl/RobustToolbox/TOMCAT.zip (last accessed on the 24th of May 2006).
[14] 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.
[15] S. de Jong, SIMPLS: an alternative approach to partial least squares regression, Chemometrics and Intelligent Laboratory Systems 18 (1993) 251–263.
[16] H. Kiers, J. Ten Berge, R. Bro, PARAFAC2. Part I. A direct fitting algorithm for the PARAFAC2 model, Journal of Chemometrics 13 (1999) 275–294.
[17] R. Bro, C. Andersson, H. Kiers, PARAFAC2. Part II. Modeling chromatographic data with retention time shifts, Journal of Chemometrics 13 (1999) 295–309.
[18] I. Stanimirova, B. Walczak, D.L. Massart, V. Simeonov, C.A. Saby, E. Di Crescenzo, STATIS, a three-way method for data analysis. Application to environmental data, Chemometrics and Intelligent Laboratory Systems 73 (2004) 219–233.
[19] B.G.M. 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, Amsterdam, 1998.
[20] S. Serneels, C. Croux, P. Filzmoser, P.J. Van Espen, Partial robust M-regression, Chemometrics and Intelligent Laboratory Systems 79 (2005) 55–64.
[21] R. Gnanadesikan, J.R. Kettenring, Robust estimates, residuals, and outlier detection with multiresponse data, Biometrics 28 (1972) 81–124.
[22] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons Inc., New York, 1987.
[23] P.J. Rousseeuw, Least median of squares regression, Journal of the American Statistical Association 79 (1984) 871–880.
[24] P.J. Rousseeuw, Multivariate estimation with high breakdown point, in: W. Grossmann, G. Pflug, I. Vinche (Eds.), Mathematical Statistics and Applications, vol. B, Reidel Dordrecht, 1985, pp. 283–297.
[25] available as Fortran code from: ftp://ftp.win.ua.ac.be/pub/software/agoras/newfiles/minvol.gz (last accessed on the 13th of September 2005).
[26] D.L. Woodruff, D.M. Rocke, Heuristic search algorithms for the minimum volume ellipsoid, Journal of Computational and Graphical Statistics 2 (1993) 69–95.
[27] R.D. Cook, D.M. Hawkins, S. Weisberg, Exact iterative computations of the robust multivariate minimum volume ellipsoid estimator, Statistics & Probability Letters 16 (1992) 213–218.
[28] J. Agulló, Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm, in: A. Prat (Ed.), Proceedings in Computational Statistics, Physica-Verlag, Heidelberg, 1996, pp. 175–180.
[29] P.J. Rousseeuw, K. Van Driessen, A fast algorithm for minimum covariance determinant estimator, Technometrics 41 (1999) 212–223.
[30] ftp://ftp.win.ua.ac.be/pub/software/agoras/newfiles/fastmcdm.gz (last accessed on the 13th of September 2005).
[31] P.J. Rousseeuw, V.J. Yohai, Robust regression by means of S-estimators, in: J. Franke, W. Härdle, D. Martin (Eds.), Robust and Nonlinear Time Series, Lecture Notes in Statistics, vol. 26, Springer, New York, 1984, pp. 256–272.
[32] V.J. Yohai, High breakdown-point and high efficiency robust estimates for regression, The Annals of Statistics 15 (1987) 642–656.
[33] C. Croux, G. Haesbrouck, Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, Journal of Multivariate Analysis 71 (1999) 161–190.
[34] K.S. Tatsuoka, D.E. Tyler, The uniqueness of S and M-functionals under non-elliptical distributions, The Annals of Statistics 28 (2000) 1219–1243.
[35] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis distance, Chemometrics and Intelligent Laboratory Systems 50 (2000) 1–18.
[36] P.J. Rousseeuw, B.C. van Zomeren, Unmasking multivariate outliers and leverage points, Journal of the American Statistical Association 85 (1990) 633–639.
[37] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and Intelligent Laboratory Systems 2 (1987) 37–52.
[38] J.S. Devlin, R. Gnanadesikan, J.R. Kettering, Robust estimation of dispersion matrix and principal components, Journal of the American Statistical Association 76 (1981) 354–362.
[39] R.A. Maronna, Robust M-estimators of multivariate scatter and location, Annals of Statistics 4 (1976) 51–67.
[40] N.A. Campbell, Robust procedures in multivariate analysis. I. Robust covariance estimation, Applied Statistics 29 (1980) 231–237.

[41] A. Singh, Outliers and robust procedures in some chemometric applications, Chemometrics and Intelligent Laboratory Systems 33 (1996) 75–100.
[42] A. Singh, J.M. Nocerino, in: J. Einax (Ed.), Chemometrics in Environmental Chemistry — Statistical Methods, vol. 2, Springer, Berlin, 1995, pp. 229–277, Part G.
[43] P.J. Huber, Projection pursuit, The Annals of Statistics 13 (1985) 435–475.
[44] A. Hyvärinen, J. Karhunen, E. Oja, Independent Component Analysis, John Wiley & Sons, Inc., New York, 2001.
[45] G.Y. Li, Z.L. Chen, Projection-pursuit approach to robust dispersion matrix and principal components: primary theory and Monte Carlo, Journal of the American Statistical Association 80 (1985) 759–766.
[46] L.P. Ammann, Robust singular value decompositions: a new approach to projection pursuit, Journal of the American Statistical Association 88 (1994) 505–514.
[47] Y.L. Xie, J.H. Wang, Y.-Z. Liang, L.X. Sun, H. Song, R.Q. Yu, Robust principal component analysis by projection pursuit, Journal of Chemometrics 7 (1993) 527–541.
[48] H. Hove, Y.-Z. Liang, O.M. Kvalheim, Trimmed objects projections: a nonparametric robust latent-structure decomposition method, Chemometrics and Intelligent Laboratory Systems 27 (1995) 33–40.
[49] M. Hubert, P.J. Rousseeuw, S. Verboven, A fast method for robust principal components with application to chemometrics, Chemometrics and Intelligent Laboratory Systems 60 (2002) 101–111.
[50] C. Croux, A. Ruiz-Gazen, A fast algorithm for robust principal components based on projection pursuit, COMPSTAT: Proceedings in Computational Statistics, Physica-Verlag, Heidelberg, 1996, pp. 211–217.
[51] C. Croux, A. Ruiz-Gazen, High breakdown estimators for principal components: the projection-pursuit approach revisited, Journal of Multivariate Analysis 95 (2005) 206–226.
[52] D. Peña, F.J. Prieto, Multivariate outlier detection and robust covariance estimation, Technometrics 41 (2001) 286–300.
[53] C. Croux, A. Ruiz-Gazen, Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies, Biometrika 87 (2000) 603–618.
[54] B. Walczak, D.L. Massart, Robust principal component regression as a detection tool for outliers, Chemometrics and Intelligent Laboratory Systems 27 (1995) 41–54.
[55] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, ROBPCA: a new approach to robust principal component analysis, Technometrics 47 (2005) 64–79.
[56] W.A. Stahel, Robuste Schätzungen, Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen. Ph.D. Thesis, ETH, Zürich.
[57] D.L. Donoho, Breakdown properties of multivariate location estimators. Ph.D. qualifying paper, Harvard University, 1982.
[58] S. Serneels, P. Filzmoser, C. Croux, P.J. Van Espen, Robust continuum regression, Chemometrics and Intelligent Laboratory Systems 76 (2005) 197–204.
[59] P. Filzmoser, S. Serneels, C. Croux, P.J. Van Espen, Robust multivariate methods: the projection pursuit approach, in: M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, W. Gaul (Eds.), From Data and Information Analysis to Knowledge Engineering, Springer Verlag, Berlin, 2006, pp. 270–277.
[60] M. Hubert, S. Engelen, Robust PCA and classification in biosciences, Bioinformatics 20 (2004) 1728–1736.
[61] S. Verboven, M. Hubert, LIBRA: a MATLAB library for robust analysis, Chemometrics and Intelligent Laboratory Systems 75 (2005) 127–136.
[62] http://wis.kuleuven.be/stat/robust/LIBRA1.html.
[63] N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, K.L. Cohen, Robust principal component analysis for functional data, Test 8 (1999) 1–7.
[64] R.A. Maronna, Principal components and orthogonal regression based on robust scales, Technometrics 47 (2005) 264–273.
[65] S. Serneels, submitted for publication. Linear robust classification by the robust least squares support vector classifier, Journal of Chemometrics.
[66] I. Stanimirova, M. Daszykowski, B. Walczak, submitted for publication. Dealing with outliers and missing elements in principal component analysis, Talanta.
[67] I. Stanimirova, S. Serneels, P. Van Espen, B. Walczak, submitted for publication. How to construct a multiple regression model for data with missing elements and outlying objects, Analytica Chimica Acta.