EXPLORING HIGH-DIMENSIONAL DATA WITH ROBUST PRINCIPAL COMPONENTS

P. Filzmoser and H. Fritz
Department of Statistics and Probability Theory
Vienna University of Technology, AUSTRIA
e-mail: P.Filzmoser@tuwien.ac.at

Abstract

For high-dimensional data of low sample size it is difficult to compute principal components in a robust way. We mention an algorithm which is highly precise and fast to compute. The robust principal components are used to compute distances of the observations in the (sub-)space of the principal components and distances to this (sub-)space. Both distance measures retain valuable information about the multivariate data structure. Plotting the magnitudes of the distance measures helps to reveal important multivariate data information.

1 Introduction

Principal component analysis (PCA) is frequently used in an exploratory context to gain first insight into multivariate data. The method is fast to compute, and also the geometric concept of PCA is easy to understand. Moreover, for the purpose of graphically inspecting the multivariate data the data requirements (distributional assumptions, etc.) are very low. These features make PCA attractive for researchers and practitioners working in various fields. Pairwise plots of the first few principal components (PCs) often allow to reveal the multivariate data structure in a way that relations among the observations as well as data groups and structures can be discovered.

PCA is also often used to identify data outliers. Surprisingly, in many applications this even works, although it is well known that the directions of the PCs are determined by the eigenvectors of the covariance matrix of the data. Using the empirical covariance matrix as an estimator of the covariance results in an unbounded influence function of the eigenvalues and eigenvectors (Croux and Haesbroeck, 2000) which means that the eigendecomposition can be completely misled even in case of very small amount of contamination. On the other hand, huge outliers can attract a PC, and the analyst will be able to discover such outliers by inspecting the corresponding component. In that way, PCA does not focus on revealing the main variability of the underlying data structure, but it still is able to find interesting phenomena in the data. The analyst would then usually remove this outlier and proceed with a new PCA on the reduced data to finally reveal the multivariate data structure.

Less extreme outliers or multivariate outliers are not necessarily visible as isolated points at the projections on the PCs. Even worse, they are able to spoil the PC directions, and the goal of exploring the relevant data structure by the first few PCs may completely fail (Croux and Ruiz-Gazen, 2005). This can be avoided if PCA is robustified. The easiest possibility is to estimate the covariance matrix in a robust way, using well known estimators like the MCD (Rousseeuw and Van Driessen, 1999). The eigenvectors of this robust covariance matrix will be resistant to outliers in the data and point in directions of high variability of the main data cloud.

Exploring data by PCA and robustifying PCA as mentioned above is possible for "tall" data where n, the number of observations, is (much) larger than p, the number of variables. However, there are many applications in chemometrics, marketing, biostatistics, etc. where p is much larger than n. Robust covariance estimation e.g. using the MCD does no longer work in this case and one needs to consider other alternatives. Moreover, projecting the data on the first two PCs is often still very uninformative because of the high dimensionality of the data.

In this paper we will briefly describe methods and algorithms for robust PCA in the case $p >> n$ (Section 2). At the basis of a diagnostics plot for PCA (Hubert et al., 2005) we construct a plot in Section 3 that allows to reveal multivariate data structure. As a motivating example we use a data set of 21 NIR spectra in 268 dimensions (Swierenga et al., 1999).

2 PCA for high-dimensional data

A very appealing approach for robust PCA in high dimensions is based on the projection-pursuit (PP) principle (Li and Chen, 1985). This approach uses the initial definition of PCA of finding a direction where the projected data points have maximal variance. Subsequent directions have the same goal of maximizing the variance of the projected data, but they are supposed to be orthogonal to previously found directions. Robustifying this approach is in fact very easy, because the measure of variance simply needs to be robust. The MAD or the Qn estimators have been suggested for this purpose (Croux and Ruiz-Gazen, 2005). The difficult task, however, is to develop an algorithm for solving the maximization problem.

In the case $p >> n$ the computational complexity can be reduced considerably by first performing a singular value decomposition which allows a reduction of the dimensionality from p to only n dimensions without any loss of information (see e.g. Stanimirova et al., 2004).

The algorithm of Croux and Ruiz-Gazen (2005) uses candidate directions for finding the first PC that consist of directions from the center of the data cloud to each single data point. The resulting n directions are then evaluated by computing the (robust) variance of the projected data points, and the direction corresponding to the maximum of the variance is an approximation of the direction of the first PC. The search is then continued analogously in the subspace orthogonal to the found direction. The identified directions are approximations of the eigenvectors of a covariance-based approach, and the corresponding maximal variances are approximations of the eigenvalues.

Recently it was pointed out (Croux et al., 2007) that the algorithm of Croux and Ruiz-Gazen (2005) has serious drawbacks:

(a) The estimated eigenvalues corresponding to the k-th PC with $k > n/2$ are exactly zero if the MAD, the Qn, or any other highly robust scale measure is used. This

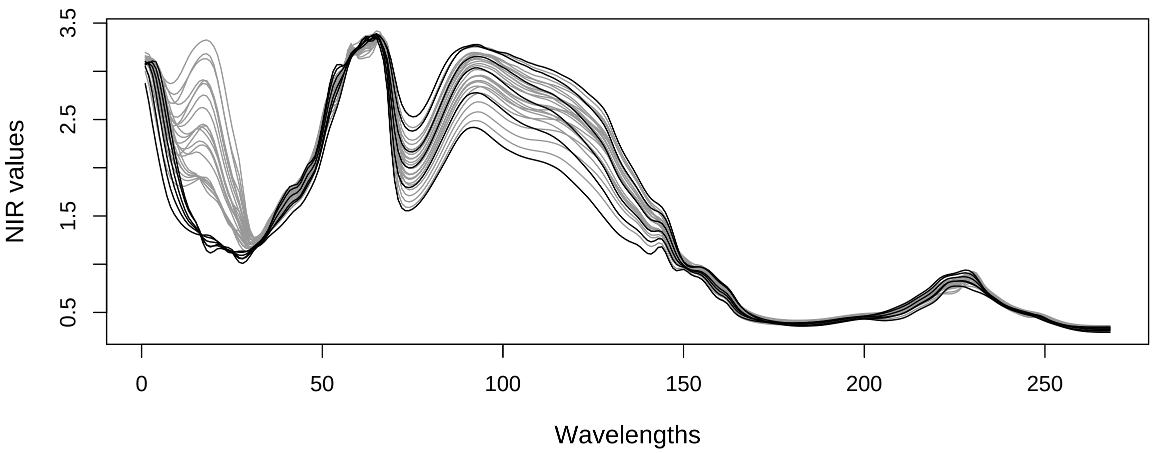

Figure 1: NIR spectra of PET yarns; the highlighted (black) spectra have deviating structure for small wavelengths.

artefact is also called *implosion* of the scale measure. Usually, this drawback has no real consequences because one is mainly interested in the first few PCs. However, if a robust covariance matrix should be estimated with the robust PCs, the result could be seriously biased.

(b) Especially for $p >> n$ the algorithm is not very precise. This can be verified by comparing the theoretical maximum which are the eigenvalues of the eigendecomposition of the sample covariance matrix with the resulting estimated eigenvalues of this algorithm using the classical variance measure.

Both disadvantages are solved by the so-called GRID algorithm (Croux et al., 2007). The idea is to search for the optimal direction only in a plane on a regular grid. The plane is first spanned by two variables, but later on also information of the other variables is used by taking linear combinations with the remaining variables. The resulting directions are highly precise: the estimated eigenvalues are in general considerably higher than for the algorithm of Croux and Ruiz-Gazen (2005), and they come very close to the true maximum. The GRID algorithm has been implemented in the library *pcaPP* of the statistical software *R*.

2.1 Example

We apply robust PCA on a data set described in Swierenga et al. (1999) of 21 NIR spectra of PET yarns, measured at 268 wavelengths. The data set is available in *R* in the package *pls* as data set *NIR*. Swierenga et al. (1999) used these data together with 28 corresponding densities to construct a robust multivariate calibration model. Here we will only consider the NIR data, and our goal is to gain insight into their multivariate data structure using robust PCA. Figure 1 shows the 28 spectra. We highlighted some spectra with somewhat abnormal behavior at small wavelengths.

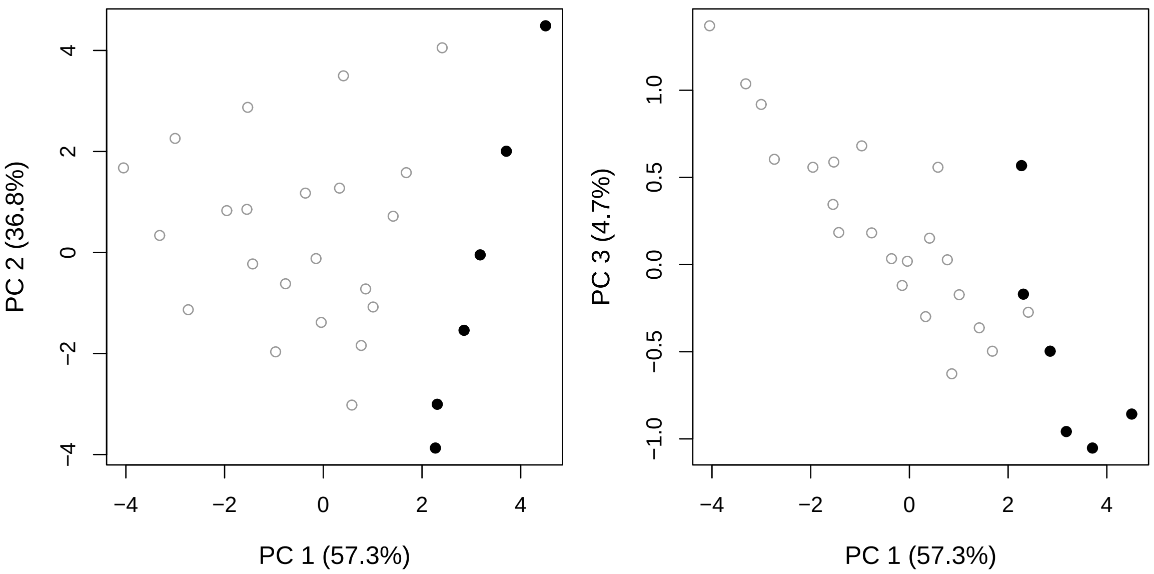

Figure 2: Robust PCA for the NIR data using the GRID algorithm with the MAD as scale measure.

Robust PCA using the GRID algorithm was applied to the NIR data set. To obtain robust components the MAD (median absolute deviation) was used as scale measure. Figure 2 shows the plot of first versus second PC (left) and first versus third PC (right). The highlighted spectra from Figure 1 are visualized as dark points. The first 3 PCs include 98.8% of the total variability. However, it is not obvious from these plots that the highlighted spectra are somehow different. It is possible that this difference is expressed in the remaining 25 PCs which, on the other hand, only contain a bit more than 1% of the total variability.

3 Orthogonal distance and score distance

Hubert et al. (2005) used the *orthogonal distance* (OD) and the *score distance* (SD) as diagnostic tools in the context of PCA. For a sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in $\mathcal{R}^p$ the OD is defined as

$$\mathrm{OD}_i^{(k)} = \|\boldsymbol{x}_i - \hat{\boldsymbol{\mu}} - \boldsymbol{P}^{(k)}\boldsymbol{t}_i^{(k)}\| \tag{1}$$

and the SD as

$$\mathrm{SD}_i^{(k)} = \left[\sum_{j=1}^{k} \frac{\|\boldsymbol{t}_i^{(k)}\|^2}{l_j}\right]^{1/2} \tag{2}$$

for $i = 1, \ldots, n$. Here, $\hat{\boldsymbol{\mu}}$ is the estimated center of the data, the matrix $\boldsymbol{P}^{(k)}$ contains the first k estimated eigenvectors in its columns, l_j are the estimated eigenvalues, and $\boldsymbol{t}_i^{(k)}$ is the ith score vector in the space of the first k principal components ($1 \le k \le r$) with r being the rank of the data. Both the OD and the SD depend on the number

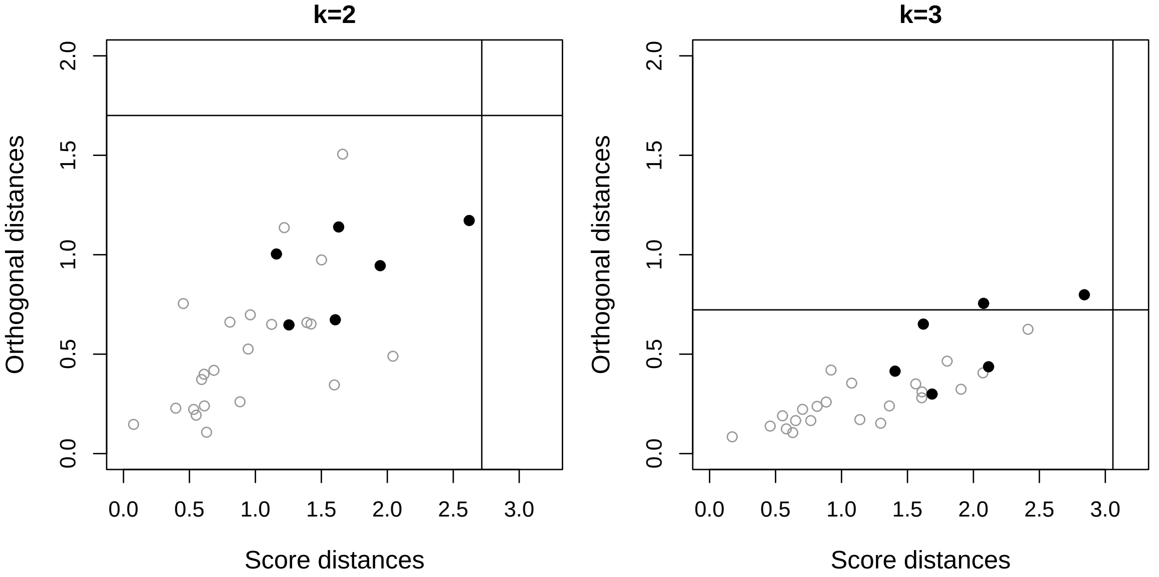

Figure 3: Diagnostic plots to the PCA from Figure 2 for 2 (left) and 3 (right) components.

of PCs considered. The OD describes the orthogonal distance of an observation to the space spanned by the first k PCs, whereas the SD is a Mahalanobis-like measure of distance of an observation within the PC space. Samples with large OD and SD can have severe leverage to a classical PCA.

Hubert et al. (2005) introduced a diagnostic plot for PCA by plotting SD versus OD. Critical thresholds for SD and OD allow to identify outlying observations. Figure 3 shows this diagnostic plot for the NIR data when $k = 2$ PCs (left) and $k = 3$ PCs are used to compute the distances. Again, the black points refer to the spectra with unusual behavior. We used similar critical values as in Hubert et al. (2005): for the SD a quantile of the chi-squared distribution with k degrees of freedom (we used $\sqrt{\chi^2_{k;0.975}}$), and for OD we also took the Wilson-Hilferty approximation for the scaled chi-squared distribution which assumes that the ODs to the power of 2/3 are approximately normally distributed. The parameters μ and σ of the normal distribution can be estimated by the median and MAD of the values $\text{OD}^{2/3}$, and the critical value can be taken as $(\hat{\mu} + \hat{\sigma} z_{0.975})^{3/2}$, with $z_{0.975}$ being the 97.5% quantile of the standard normal distribution. The critical values in Figure 3 for $k = 3$ allow to identify two observations as “outliers” because their OD is larger than the critical value. Note that our goal is not necessarily outlier detection using the PCs but rather to learn about the multivariate data structure. These plots, however, do not reveal any special phenomenon like groups of deviating data points.

It is possible, that such a diagnostic plot would reveal deviating data points in a better way if more PCs were used for computing the OD and SD. Hubert et al. (2005) suggested to take as many PCs such that about 90% of the total variability are explained. This would correspond to the right plot in Figure 3.

4 Exploring the multivariate data structure

It is easy to show that

$$\mathrm{SD}_i^{(k)} \leq \mathrm{SD}_i^{(k+1)} \qquad \text{and} \qquad \mathrm{OD}_i^{(k)} \geq \mathrm{OD}_i^{(k+1)}$$

for $1 \leq k < r$. These properties can also be observed in Figure 3. In fact, the SD is just the projection of the Mahalanobis distance on the space spanned by the first k PCs, see Equation (1). Using all PCs for computing the SD is exactly the Mahalanobis distance. Naturally, the distances in a higher dimensional space are increasing. Also for OD the above property is obvious because for increasing dimension of the PC space the orthogonal distances to the data space have to decrease, see Equation (2). On the other hand, also the critical values for OD and SD change for an increasing number k of PCs. Moreover, it is possible that an observation has large OD and small SD for k PCs, but small OD and large SD for $k+1$ PCs. This can be the case if the observation is far away from the space spanned by k PCs but very close to PC number $k+1$.

Thus, studying the OD and SD for various values of k might provide more insight into the multivariate data structure. Since it will be rather difficult to use a diagnostic plot like in Figure 3 for several or all values of k we prefer to present two separate plots for the OD and for the SD. Following the suggestions of Maronna and Zamar (2002), we multiply the SD with the scaling factor $\sqrt{\chi^2_{k;0.5}}/\mathrm{median}(SD_i^{(k)})$ which improves the chi-square approximation significantly. Afterwards, we divide the OD and the scaled SD by their critical values corresponding to the 97.5% quantiles or the respective distribution. OD and SD are thus standardized, and a comparison with the critical threshold 1 can be easily done. Figure 4 shows the resulting plots for the number of PCs ranging from 1 to 28. Each line in the plot corresponds to the standardized OD (left) or SD (right) of an observation. The black lines are the atypical observations from Figure 1. The dashed lines refer to the critical threshold 1. Most of the atypical observations are only exceeding the threshold if more than 3 PCs are taken.

The visualization of Figure 4 can be improved by observing if an observation exceeds the critical value at a given number of PCs or not. This information is shown in Figure 5 for the OD (left) and for the SD (right). A white square indicates that the observation did not exceed the critical value for given k. Light gray, dark gray and black boxes are plotted if the observations exceed the corresponding 97.5%, 99%, or 99.9% critical values, respectively. The atypical observations visible in Figure 1 refer to the indices 1, 3, 6, 10, 15, and 21, which are also atypical in the plot of the OD and/or SD in Figure 5. There are additional outstanding observations which refer to further deviating structure in some of the NIR spectra. Thus, the plot provides an impression about the multivariate data structure.

5 Summary

For high-dimensional data it is not trivial how to compute robust principal components. We propose to use the GRID algorithm (Croux et al., 2007) which is very precise and

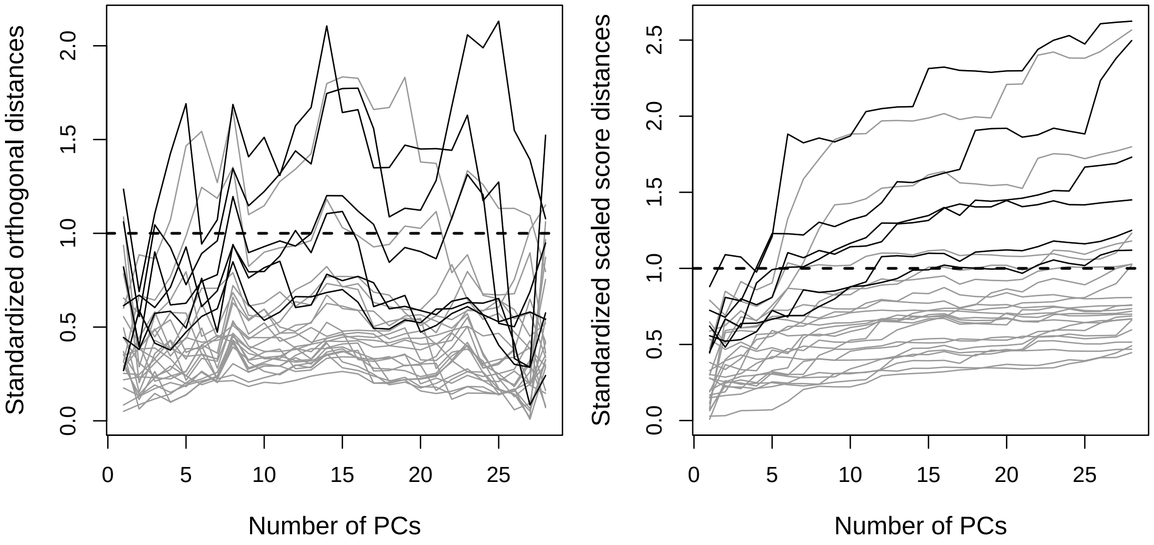

Figure 4: Standardized OD (left) and SD (right) for each observation using $k = 1, \ldots, 28$ PCs. The dashed lines indicate the 97.5% critical values.

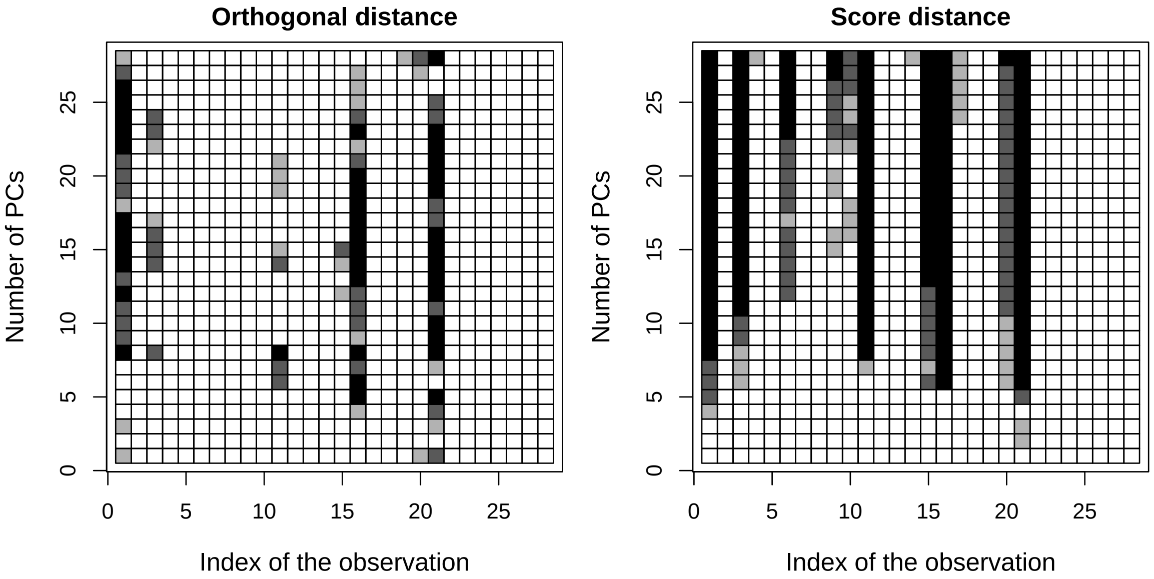

Figure 5: OD (left) and SD (right) for each observation using $k = 1, \ldots, 28$ PCs. The different gray levels indicate that the 97.5% (light gray), 99% (dark gray), or 99.9% (black) critical value was exceeded.

results in highly robust PCs. The estimated eigenvectors and eigenvalues from this approach can be used to compute orthogonal and score distances which are indications of outlyingness of the observations. We introduced a new plot of these distance measures computed for various numbers of principal components. Deviations of observations from the main data structure in a certain sub-space spanned by k PCs are indicated in the plot. Since the dimension of the sub-space changes with the number of PCs, the plot helps to reveal the multivariate data structure.

In the presentation we will also give other real data examples where this plot gives insight into the data structure. Moreover, we will demonstrate that not only the robustness of PCA is important for an informative plot, but also the precision of the PCA algorithm.

References

[1] Croux C., Filzmoser P., Oliveira M.R. 2007. Algorithms for Projection-pursuit Robust Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems*. To appear.

[2] Croux C., Haesbroeck G. (2000). Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies. *Biometrika*. Vol. **87**, pp. 603-618.

[3] Croux C., Ruiz-Gazen A. (2005). High Breakdown Estimators for Principal Components: The Projection-pursuit Approach Revisited. *Journal of Multivariate Analysis*. Vol. **95**, pp. 206-226.

[4] Hubert M., Rousseeuw P.J., Vanden Branden K. (2005). ROBCA: A New Approach to Robust Principal Component Analysis. *Technometrics*. Vol. 47, pp. 64-79.

[5] Li G., Chen Z. (1985). Projection-pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo. *Journal of the American Statistical Association*. Vol. **80**, pp. 759-766.

[6] Maronna R., Zamar R. (2002). Robust Estimates of Location and Dispersion for High-dimensional Data Sets. *Technometrics*. Vol. **44(4)**, pp. 307-317.

[7] Rousseeuw P.J., Van Driessen K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*. Vol. **41**, pp. 212-223.

[8] Stanimirova I., Walczak B., Massart D.L., Simenov V. (2004). A Comparison between two Robust PCA Algorithms. *Chemometrics and Intelligent Laboratory Systems*. Vol. **71**, pp. 83-95.

[9] Swierenga H., de Weijer A.P., van Wijk R.J., Buydens L.M.C. (1999). Strategy for Constructing Robust Multivariate Calibration Models. *Chemometrics and Intelligent Laboratory Systems*. Vol. **49(1)**, pp. 1-17.