Kalle Alaluusua

Outlier detection using robust PCA methods

School of Science

Bachelor's thesis

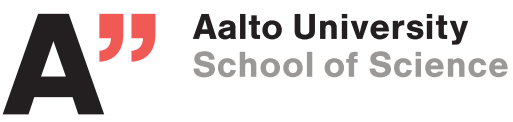
Espoo 31.8.2018

Thesis supervisor:

Asst.Prof. Pauliina Ilmonen

Thesis advisor:

D.Sc. (Tech.) Lauri Viitasaari



Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of bachelor's thesis

Author Kalle Alaluusua

Title of thesis Outlier detection using robust PCA methods

Degree programme Engineering Physics and Mathematics

Major Mathematics and Systems Analysis **Code of major** SCI3029

Supervisor Asst.Prof. Pauliina Ilmonen

Thesis advisor(s) D.Sc. (Tech.) Lauri Viitasaari

Date 15.8.2018 **Number of pages** 23+5 **Language** English

Abstract

In this thesis we apply the robust principal component analysis methods ROBPCA and its modification for skewed data to two asymmetric and non-Gaussian data sets from the field of production engineering. The outliers are identified by their large deviation from the robust center of the data, and the subspace spanned by the robust principal components. Finally, we analyze the robust principal components to gain a better understanding of the sources of variation in the data. The quality of our models is assessed by visualization methods.

As expected, the skew-adjusted algorithm proves to be more accurate in detecting the anomalous observations. The ROBPCA algorithm falsely identifies regular observations located in the tail area of skewed distributions as anomalies. We identify both univariate and multivariate outliers. The complete decomposition contribution (CDC) indices prove to be effective in describing the effect of each variable on the large deviation of the outliers.

The findings in this thesis lay the groundwork for further analysis of the data. In the wider context of improving production processes, robust logistic regression methods could be used to determine whether the measurement phenomena responsible for the outlying observations have negative connotations.

Keywords principal component analysis, multivariate statistics, robust statistics, anomaly detection, quality monitoring

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Tekniikan kandidaatintyön tiivistelmä

Tekijä Kalle Alaluusua		
Työn nimi Poikkeavien havaintojen tunnistaminen pääkomponenttianalyysin menetelmin		
Koulutusohjelma Teknillinen fysiikka ja matematiikka		
Pääaine Matematiikka ja systeemitieteet		**Pääaineen koodi** SCI3029
Vastuuopettaja Apulaisprofessori Pauliina Ilmonen		
Työn ohjaaja(t) TkT Lauri Viitasaari		
Päivämäärä 28.8.2018	**Sivumäärä** 23+5	**Kieli** Englanti

Tiivistelmä
Tässä opinnäytetyössä sovelletaan kahta vakaata pääkomponenttianalyysimenetelmää kahteen epäsymmetriseen ja ei-normaalisti jakautuneeseen havaintojoukkoon tuotantotekniikan alalta. Käytetyt menetelmät ovat ROBPCA-menetelmä ja sen epäsymmetriselle havaintojoukolle kohdennettu muunnelma. Menetelmät perustuvat pääkomponenttianalyysin, missä havaintojoukon varianssi- ja kovarianssirakennetta pyritään kuvaamaan laatimalla joukko keskenään korreloimattomia uusia muuttujia eli pääkomponentteja. Pääkomponentit ovat alkuperäisten muuttujien lineaarisia yhdistelmiä. Ensimmäinen pääkomponentti maksimoi tälle projektoidun havaintojoukon varianssin ja minimoi jäännösvirheet havaintojen ja näiden projektioiden välillä. Toinen pääkomponentti on kohtisuorassa ensimmäistä pääkomponenttia vastaan ja suunnataan jälleen siten, että tälle projektoidun havaintojoukon varianssi maksimoituu ja jäännösvirheet minimoituvat. Näin edetään, kunnes haluttu osuus alkuperäisen havaintojoukon varianssista on selitetty.

Koska klassinen pääkomponenttianalyysi perustuu oleellisesti pienimmän neliösumman menetelmään, on se herkkä poikkeaville havainnoille. Jo yksittäinen poikkeava havainto voi kasvattaa projektoidun havaintojoukon varianssin mielivaltaisen suureksi, mikä kallistaa ensimmäisen pääkomponentin virheellisesti poikkeavaa havaintoa kohti. Tällöin ensimmäisten pääkomponenttien perusteella ei voida enää tehdä johtopäätöksiä havaintojoukon enemmistön kovarianssirakenteesta. Ongelman ratkaisemiseksi on kehitetty vakaita pääkomponenttianalyysimenetelmiä, jotka luopuvat havaintojoukon normaalisuusoletuksesta ja kuvaavat tämän enemmistön ominaisuuksia poikkeavista havainnoista huolimatta. Vakaat menetelmät perustuvat pääsääntöisesti joko vakaan kovarianssimatriisin ominaisarvohajotelmaan tai vakaan hajonnan estimaatin maksimointiin. Työssä käytetyt menetelmät hyödyntävät kumpaakin keinoa.

Työssä pyritään tunnistamaan havaintojoukosta poikkeavat havainnot. Ideaalitilanteessa nämä sijaitsevat poikkeuksellisen kaukana havaintojoukon vakaasta keskipisteestä sekä vakaiden pääkomponenttien virittämästä tasosta. Lopuksi havaintojoukon merkittävimmät vaihtelun lähteet pyritään tunnistamaan tarkastelemalla vakaita pääkomponentteja. Mallin laatua arvioidaan hyödyntäen kaaviota, jossa alkuperäiset muuttujat ja mallin avulla laaditut rekonstruktiot esitetään päällekkäin.

Muunnettu menetelmä tunnistaa vieraat havainnot alkuperäistä ROBPCA-menetelmää tarkemmin, mikä on odotettavissa. ROBPCA-menetelmä luokittelee virheellisesti epäsymmetrisen jakauman ”hännän” alueelle sijoittuvat havainnot poikkeaviksi. Käytettyjen menetelmien tuottamat kaaviot mahdollistavat poikkeavien havaintojen luokittelun pääkomponenttien virittämään tasoon nähden kohtisuoriin ja yhdensuuntaisiin havaintoihin. Aineistossa havaitaan sekä yksi- että moniulotteisia poikkeavia havaintoja. Yksittäisen muuttujan vaikutuksesta havainnon poikkeavuuteen kertova CDC-indeksi moniulotteisen laadunvalvonnan alalta osoittautuu tehokkaaksi menetelmäksi näiden kuvailuun.

Tämän opinnäytetyön tulokset viitoittavat aineiston jatkoanalyysin suunnan. Laajemmassa tuotannollisessa kontekstissa voitaisiin selvittää, vaikuttavatko poikkeaviin havaintoihin johtavat ilmiöt negatiivisesti tuotantoprosessiin. Tähän voitaisiin soveltaa vakaita logistisia regressiomenetelmiä.

Avainsanat pääkomponenttianalyysi, moniulotteinen analyysi, vakaat menetelmät, poikkeavien havaintojen tunnistaminen, laadunvalvonta

Contents

1 Introduction

Principal component analysis (PCA) is a widely used technique in multivariate statistics. It aims to explain the variance-covariance structure of data through a set of new uncorrelated variables referred to as principal components (PCs). The principal components are linear combinations of the original variables, which often facilitate the interpretation of different sources of variance. PCA is often the first step of the data analysis, followed by other multivariate techniques.

In the classical approach, the first principal component is the direction in which the projected observations have the largest sample variance. The second component is orthogonal to the first component and corresponds to the direction which again maximizes the sample variance of the projected observations. Proceeding this way produces all of the principal components, which together reproduce the total system variability. However, much of this variability can often be accounted for by a small number of components. Thus, principal component analysis is an effective tool for data reduction. Though PCA is nominally a non-convex problem, it can be solved using the Lagrange multiplier method. It follows that the components can be computed as the eigenvectors of an estimate of the covariance matrix. As the estimate, the classical PCA uses the sample covariance matrix.

Like many classical estimators, sample variance and the sample covariance matrix are known to be sensitive to anomalous observations. That is to say, even a nominal portion of outlying points may prevent the leading components of classical PCA from correctly depicting the covariance structure of the data majority. Accordingly, the model may not allow to detect the anomalous observations responsible for the lack of fit and may even portray regular observations as outliers. These effects are known as *masking* and *swamping*. As a consequence, robust PCA methods have been developed to construct a subset of PCs unaffected by the outliers. The outliers are then characterized by their large deviation from the subspace spanned by the principal components. For a more extensive review on the topic, we refer to the article by Rousseeuw and Hubert (2011).

This thesis focuses on a robust PCA method referred to as ROBPCA and its modification for skewed data. The methods are applied to a data set from the field of production engineering. The results are analyzed using visualization methods as well as techniques from the field of multivariate quality control.

2 Theoretical background

2.1 Robust estimators

In statistics, estimators are rules for estimating quantities of interest based on a sample of the data. Estimators are instrumental in describing the data structure, and they often serve as the basis for higher level statistical methods. In many cases, classical estimators rely on the fundamental assumption about the normal distribution of data errors. When the data is contaminated by outliers, i.e. observations from different population than most of the data, the classical estimators no longer perform well. Thus, robust estimators have been developed to perform under wider conditions, such as the presence of outliers and non-Gaussian probability distributions.

The robustness of an estimator can be described using concepts of *breakdown point* of an estimator and its *influence function*. The breakdown point of the estimators refers to the proportion of outlying observations the estimator can withstand while still yielding meaningful estimates. The breakdown point never exceeds 50%, since with more than half of the data sampled from a contaminating distribution it is impossible to draw reliable conclusions about the underlying distribution. For instance, the mean estimator has the smallest possible breakdown point of 0%, since replacing a single observation can render the estimate arbitrarily large, while the median has the maximum breakdown point of 50%. The influence function in turn aims to describe the influence of observations upon the estimator in respect of infinitesimal perturbations. The efficiency of an estimator on the other hand describes the estimator's performance on uncontaminated data in comparison to classical estimators. For example, the efficiency of the median for a large sample size is 64% when compared to the mean estimator. For more information on robust estimators, we refer to the review paper by Daszykowski et al. (2007).

2.2 The ROBPCA method

The ROBPCA (Hubert et al. 2005) is a method for robust principal component analysis. It is resistant to outliers in the data and well suited for analysis of high-dimensional data. The ROBPCA algorithm consists of three major steps and can be described briefly as follows: Let $\mathbf{X}_{n,p}$ be the $n \times p$ input data matrix of n observations and p variables. First the data is restricted to the subspace spanned by the n observations by singular value decomposition. The dimension of the resulting subspace is at most $n - 1$ so this yields a dimensionality reduction without loss of information.

In the next step, for each observation $\mathbf{x}_i$ a measure of outlyingness (Donoho and Gasko 1992; Stahel 1981) is computed. For this the data is projected onto many univariate directions and standardized using robust univariate minimum covariance

determinant (MCD) estimators of location and scale (Rousseeuw 1984). From the resulting distances, the largest over all considered directions is the measure of outlyingness. Then a subset of $h < n$ observations with the lowest outlyingness is selected and the data is projected onto the subspace $\mathbf{V}_0$ spanned by the first k eigenvectors of the corresponding covariance matrix. The choice of k can be made by examining a *scree plot* (Jolliffe 1986) of the eigenvalues, the relative magnitude of which represent the variance explained, or by a robust PRESS algorithm (Hubert and Engelen 2007). To obtain an improved robust subspace $\mathbf{V}_1$, a subset of observations $\mathbf{x}_i$ whose orthogonal distance to the subspace $\mathbf{V}_0$ does not exceed a cutoff value $C_{OD} = (\hat{\mu} + \hat{\sigma} z_{.975})^{3/2}$ is selected. The orthogonal distance is merely the Euclidean distance between an observation and its projection in the subspace, and $z_{.975}$ is the 97.5% quantile of the Gaussian distribution. Furthermore, the estimates $\hat{\mu}$ and $\hat{\sigma}$ are obtained using the univariate MCD (Hubert et al. 2005). Respectively, $\mathbf{V}_1$ is the subspace spanned by the k dominant eigenvectors of the covariance matrix of the subset acquired.

Finally, a robust center $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$ of the projected data are computed by applying the re-weighted MCD estimator (Rousseeuw and Van Driessen 1999). The robust principal components are the k eigenvectors of this covariance matrix. The principal components span a k dimensional subspace in the original space and can be arranged into a $p \times k$ matrix $\mathbf{P}_{p,k}$ with orthogonal columns. This is referred to as a loading matrix. Centered and transformed observations $(\mathbf{X}_{n,p} - \mathbf{1}_n \hat{\boldsymbol{\mu}}^\intercal)\mathbf{P}_{p,k}$, where $\mathbf{1}_n$ is the column vector with all n components equal to 1, form a $n \times k$ matrix denoted by $\mathbf{T}_{n,k}$ and are referred to as component scores.

2.2.1 Diagnostic plot

In addition to calculating the robust loadings and component scores of the original data, the ROBPCA algorithm produces a *diagnostic plot* or *outlier map* that describes the outliers present in the data. In the context of PCA, the outliers can be divided into three categories: *good leverage points*, *orthogonal outliers* and *bad leverage points*. The good leverage points lie close to the PCA space but far from the regular observations. The orthogonal outliers on the other hand have a large orthogonal distance to the PCA space while their projection on the PCA space is inlying. Finally, the bad leverage points possess qualities of both the good leverage points and the orthogonal outliers such that they have a large orthogonal distance to the PCA space and their projection on the PCA space is distant from most of the projected data.

The outlier map plots the orthogonal distance, OD_i, from each point to its projection on the PCA space on the vertical axis. The orthogonal distance of each observation is given by

$$OD_i = \left\| \mathbf{x_i} - \hat{\boldsymbol{\mu}} - \mathbf{P}_{p,k} \mathbf{t}_i^\intercal \right\|, \tag{1}$$

where the p-variate column vector $\mathbf{x_i}$ denotes the ith observation and $\mathbf{t}_i^\intercal$ is the ith row

of $\mathbf{T}_{n,k}$. On the horizontal axis, the *robust score distance*, SD_i, of each observation is plotted. The robust score distance is defined as

$$SD_i = \sqrt{\sum_{j=1}^{k} \frac{t_{ij}^2}{l_j}}, \tag{2}$$

where t_{ij} are the robust component scores and l_j the sorted eigenvalues of the robust covariance matrix obtained in the final step of the algorithm. To distinguish the outliers from the regular observations, lines presenting cutoff values C_{OD}, introduced in Subsection 2.2, and $C_{SD} = \sqrt{\chi^2_{k,.975}}$ (Hubert et al. 2005) are drawn.

2.3 Modified ROBPCA algorithm for skewed data

Although the ROBPCA method is an effective tool for analysis of data contaminated by outliers, it assumes that the non-outlying data are approximately elliptically symmetric. When the distributions of the original variables are skewed, the algorithm tends to flag non-outlying observations as outliers. This is explained by the enhanced effect the tail of a skewed distribution has on the principal components, as its distance from the robust center can be large in relation to the range of the distribution. For this reason, skewed data could be preprocessed by a transform (for instance Box–Cox transform) before applying the algorithm. On the downside, the principal components of the transformed data might be difficult to interpret.

To eliminate the need for a transform, Hubert et al. (2009) introduced a modified ROBPCA algorithm specifically designed for skewed data. The algorithm is composed of the same steps as the ROBPCA algorithm described except for three modifications. First, Stahel–Donoho outlyingness in the second step of the algorithm is replaced by a new measure, referred to as the *adjusted outlyingness* (Hubert et al. 2009). The adjusted outlyingness is a generalization of the Stahel-Donoho outlyingness, which accounts for the skewness using *medcouple*, a robust alternative to the classical skewness coefficient (Brys et al. 2004). Medcouple, denoted by $\mathrm{MC}(\mathbf{X}_i)$, has a 25% breakdown point and a bounded influence function, which makes it attractive skewness estimator for data contaminated by outliers. The adjusted outlyingess of each point is plotted on the horizontal axis of the outlier map.

Secondly, the cutoff value C_{OD} in the second step of the algorithm is adjusted to depend adaptively on the data itself rather than on quantiles of theoretical distributions. The cutoff value used is the largest OD_i smaller than $Q_3(\{OD\}) + 1.5e^{3\mathrm{MC}(\{OD\})}\mathrm{IQR}(\{OD\})$, where $\mathrm{IQR} = Q_3 - Q_1$. If the medcouple of the orthogonal distances is negative, the cutoff value is the largest OD_i smaller than $Q_3(\{OD\}) + 1.5\mathrm{IQR}(\{OD\})$. The cutoff value C_{SD} is derived in the same way. The last modification concerns the third step of the algorithm: instead of applying the re-weighted MCD estimator, the robust principal components are calculated using the mean and covariance matrix of the h observations for which the adjusted outlyingness in the subspace $\mathbf{V}_1$ is the lowest. This is necessary, as the re-weighting

step performed in the ROBPCA method is constructed under the assumption that the regular observations are normally distributed (Hubert et al. 2005).

Both of the algorithms ROBPCA and modified ROBPCA algorithm for skewed data are well implemented in R and MATLAB. The implementations used in this thesis are a part of LIBRA: the MATLAB Library for Robust Analysis (Verboven and Hubert 2010).

3 Description of the data

In this thesis we inspect two separate production test data sets referred to as **B** and **E**. The data set **B** is a 627×20 matrix that consists of 627 observations of 20 variables, while the data set **E** is a 851×12 matrix that consists of 851 observations of 12 variables. The two data sets are the largest subsets (identified by the measurement configuration) of a larger data set, which explains the naming scheme used. The 12 variables in the data set **E** (variables 1-7, 9-11, 20 and 24) represent the same properties as identically enumerated variables in the data set **B**, but the measurement configuration differs between the data sets. Variables that have been excluded from the production process are left out from the analysis, which again justifies the unconventional naming scheme used.

Before further analysis, the variables are robustly centered and scaled. This is achieved by first subtracting the column wise medians from the data. Each column is then divided by its median absolute deviation (Andrews et al. 1972), which is a robust estimator of scale with a breakdown value of 50% and the efficiency of 37% that of the sample standard deviation. As a result, each standardized variable has a median of 0 and unit median absolute deviation, which renders the data comparable. From now on, **B** and **E** refer to the centered and scaled data sets. The robustly standardized data are presented as a whole in Figure 1. The data are plotted as a strip chart with jitter, where the univariate variables are plotted on the vertical axis, and their values on the horizontal axis. A small random vertical displacement is added to each observation to reduce the overlap. Furthermore, the observations are made transparent to emphasize the point density.

Two distributions stand out the most from the rest of distributions displayed in Figure 1 due to their distinct asymmetry. The first of them is the extremely long-tailed variable 5 distribution. The range of the distribution appears to be a factor of 10 larger than the average range of the distributions displayed in the figure. However, the tail of the distribution is light, as the median falls near the the distribution's peak, which appears black in the figure due to its high point density. The distribution is significantly left skewed in both of the data sets **B** and **E**.

The second distinctively asymmetric distribution in Figure 1 is the variable 3 distribution. The distribution is bimodal, as it has a second peak. Since the median falls just beside the left peak, its point density is significantly larger than that of the right peak. If the peaks were equally dense, the median would fall on the antimode, the least frequent value between the modes. Moreover, few observations seem to attain variable 3 values below the major mode or values above the minor mode. These remarks are valid for both of the data sets **B** and **E**.

Next we inspect the univariate distributions in greater detail by focusing on a region of interest, which includes all the peaks visible in Figure 1. The data that lies within the selected region of interest $[-5, 20]$ is displayed in the figures 2 and 3. The interval $[-5, 20]$ contains 98.8% of the standardized observations in the data set **B** and 98.4%

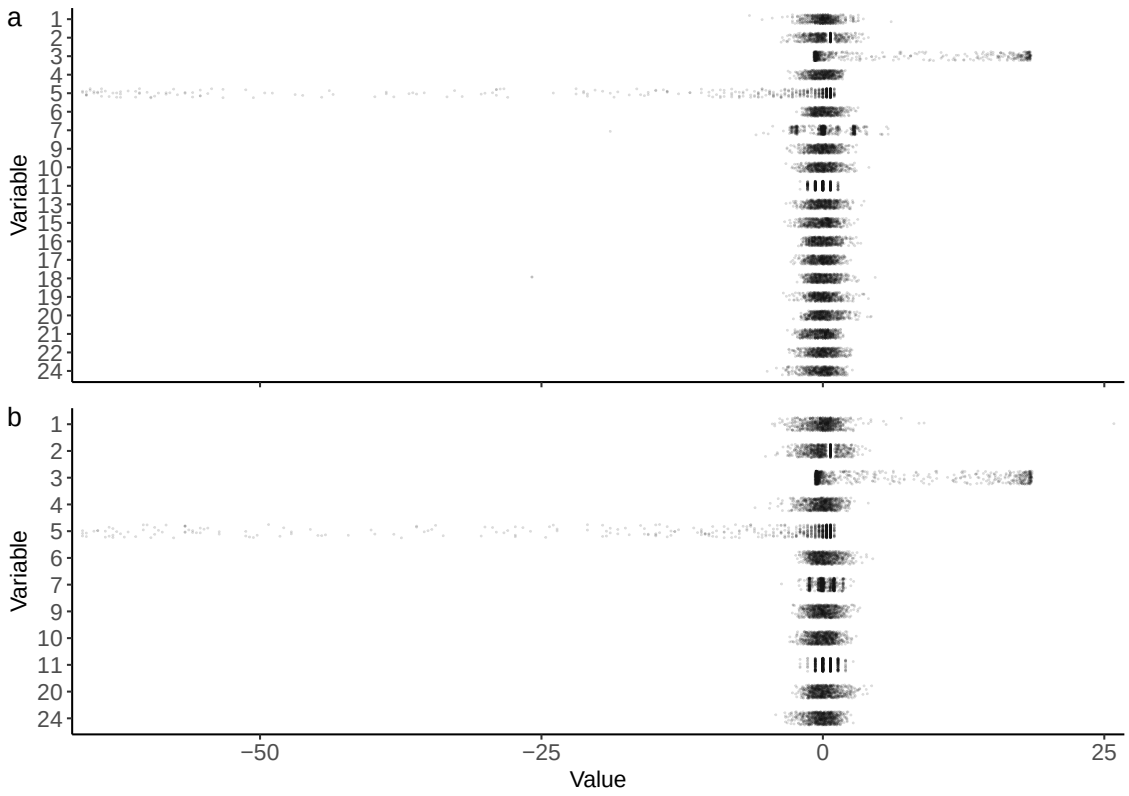


Figure 1: Jittered strip charts of the robustly standardized variables a) in the data set **B** and b) in the data set **E**.

of the standardized observations in the data set **E**. To visualize both the shapes of the distributions and the specific point locations, a combination of a density plot and a strip chart with jitter was selected. As in Figure 1, the points in the figures 2 and 3 are added a vertical displacement and transparency to further emphasize the point density. Unlike in a conventional density plot, the line is only displayed when its height is 0.1% or more relative to highest point among the ridge lines. This is done in order to highlight the possible differences in distribution tail lengths.

First we inspect the shape of the univariate distributions. The figures 2 and 3 confirm the previous observation about the bimodality of the variable 3 distribution. As suspected, the figures show that there is a sudden drop in the point density near the modes. The variable 7 distribution in the data set **B**, $\mathbf{B}_7$, has three distinct modes, although there appears to be a fourth less significant mode between the rightmost mode and the major mode. The $\mathbf{E}_7$ distribution differs from the variable $\mathbf{B}_7$ distribution, as the modes appear closer together and the troughs between the peaks in point density appear deeper. Furthermore, the strip chart below the density ridge reveals a fourth, less significant, mode on the right tail area of the distribution.

The final distribution that seemingly has multiple modes is the variable 2 distribution in both of the data sets **B** and **E**. While the distribution appears symmetric, the

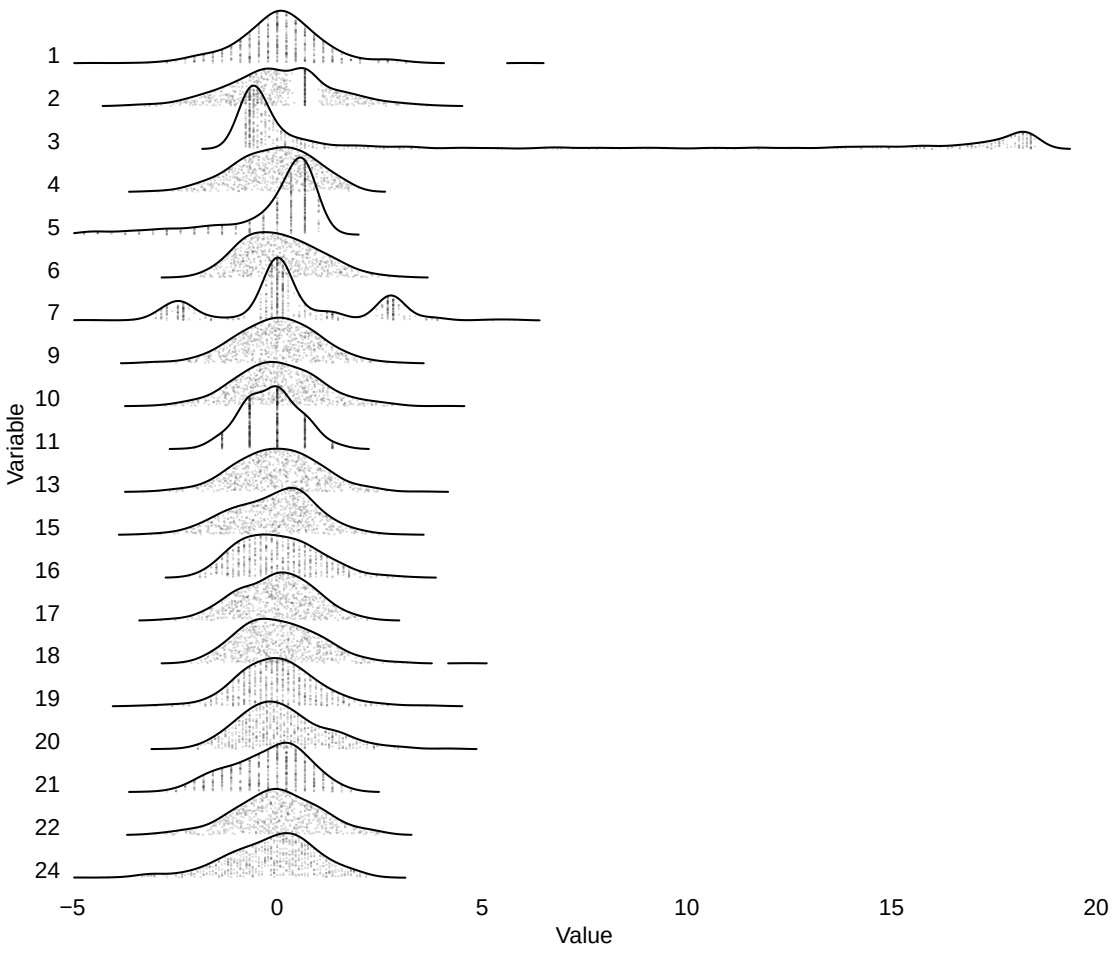


Figure 2: Density plots of the robustly standardized variables in the data set **B**, where the line is displayed when its height is 0.1% or more relative to the overall maximum.

right peak is composed of values that correspond to exactly 0 in the unstandardized distribution. In addition, no points fall within a fixed radius of the mode. Rest of the variables are unimodal, although the variables 1, 5 and especially 11 appear discrete in both of the data sets **B** and **E**. In addition, the variable $\mathbf{B}_{21}$ seems discrete. The variables 3 and 5 seem the most asymmetric in both of the distributions, while many other distributions show signs of asymmetry.

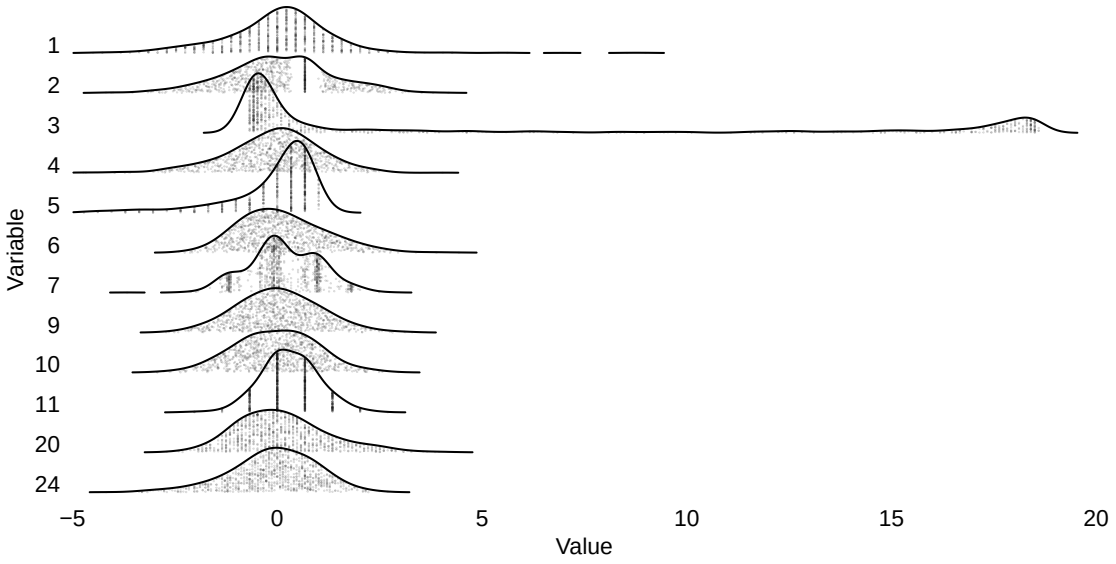


Figure 3: Density plots of the robustly standardized variables in the data set **E**, where the line is displayed when its height is 0.1% or more relative to the overall maximum.

4 Results and analysis

4.1 Preliminary analysis

Preliminary data analysis consisted of examining the pairwise scatter plots, univariate histograms and robust measures of correlation and skewness of the two data sets **B** and **E**. As a robust measure of correlation, Spearman rank correlation $\rho_S(\mathbf{X}_i, \mathbf{X}_j)$ was used. The estimator is defined as the Pearson correlation coefficient between the ranked variables (Myers and Well 2003) and it is appropriate for both continuous and discrete ordinal variables (Lehman 2005), which suits the data at hand. The preliminary analysis revealed high pairwise correlations among the variables, e.g. $\rho_S(\mathbf{E}_2, \mathbf{E}_3) = 0.98$ and $\rho_S(\mathbf{B}_{15}, \mathbf{B}_{17}) = 0.91$, which suggests that PCA-based methods are appropriate for further analysis of the data variability. The full Spearman rank correlation matrices are presented in the tables A1 and A2 in Appendix A.

As a measure of the shape and asymmetry of a distribution, medcouple was used. Based on the preliminary analysis, some of the variables appear significantly skewed. Some noteworthy medcouple values attained were $\mathrm{MC}(\mathbf{E}_3) = 0.926$, $\mathrm{MC}(\mathbf{B}_3) = 0.946$, $\mathrm{MC}(\mathbf{E}_5) = -0.714$, $\mathrm{MC}(\mathbf{B}_5) = -0.826$, $\mathrm{MC}(\mathbf{E}_7) = 0.31$ and $\mathrm{MC}(\mathbf{B}_7) = 0.543$. The complete set of medcouples acquired is presented in the tables A3 and A4 in Appendix A.

4.2 Outlier maps

Considering the findings of the preliminary analysis, both the ROBPCA method using Stahel-Donoho outlyingness (ROBPCA-SD or R-SD) and the skew-adjusted ROBPCA method using adjusted outlyingness (ROBPCA-AO or R-AO) were applied to the data sets **B** and **E**. As described in Subsection 2.2, after the initial dimensionality reduction by SVD, a set of h observations is selected and their covariance matrix is calculated. Then, as suggested by Jolliffe (1986) we examine the scree plot corresponding to the $\min(p, n-1)$ eigenvalues of the covariance matrix to select the number of components to retain. Moreover, we consider selection criterion (Hubert et al. 2005) to choose k for which

$$\sum_{i=1}^{k} \hat{l}_i \Big/ \sum_{j=1}^{r} \hat{l}_j \approx 95\%, \tag{3}$$

where $\hat{l}_i$ are the sorted eigenvalues and r is the rank of the covariance matrix calculated.

The scree plot for the ROBPCA-SD and the ROBPCA-AO methods performed on the data set **B** are presented in Figure 4. The plots feature 95% reference lines corresponding to the selection criterion 3 and only display the 8 largest eigenvalues. As Figure 4a indicates, the selection criterion holds when the number of eigenvalues

chosen in the ROBPCA-SD algorithm reaches 7, while most of the variance is explained by the first eigenvector of the covariance matrix. Since the line presenting the cumulative percentage of the variance explained is curved, the amount of variance explained by a single component added appears to diminish as k grows. As for the skew-adjusted algorithm (Figure 4b), the selection criterion holds when the number of eigenvalues chosen reaches 3, whereas most of the variance is explained by the first two components. As less components now suffice the selection criterion, the subset of the observations selected by the non skew-adjusted algorithm might represent the data structure poorly. From the third component on, the amount of variance explained appears to grow fairly linearly in relation to each component added.

To judge whether to settle for the k that satisfies the selection criterion, we turn to a rule of thumb suggested by Johnson and Wichern (2007). The rule states that only the components which, individually, explain at least a proportion of $1/r$, as in Equation 3, of the total variance should be retained. However, they add that this rule is supported by little theoretical evidence and should be applied cautiously. As for the data at hand, the rule suggests that only two leading principal components should be retained in each algorithm. The $1/r$ reference lines are presented in Figure 4. The k that suffices the ROBPCA-AO selection criterion exceeds this limit by 8% in relation to the $\min(p, n-1)$ and the k that suffices the ROBPCA-SD selection criterion by 25%. As the first few components of the ROBPCA-SD algorithm might by themselves lead to biased depiction of the data, we decide retain the number of components yielded by the selection criterion in both of the algorithms. The scree plots for both of the algorithms performed on the data set **E** appear very similar to those of the data set **B**. Thus, by the same principle we retain 4 components in the ROBPCA-SD algorithm and 3 components in the ROBPCA-AO algorithm performed on the data set **E**.

Next the score distances and the orthogonal distances of the observations in the data sets **B** and **E** are plotted as outlier maps. The outlier maps presenting the outliers detected by ROBPCA algorithm using Stahel-Donoho outlyigness and the modified ROBPCA algorithm using adjusted outlyingness are displayed in the figures 5 and 6. To highlight the differences in the results produced by the algorithms, we label up to eight outliers in the outliers maps. The R-SD algorithm flagged 250 of the observations in the data set **B** as outliers (i.e. as good leverage points, bad leverage points or orthogonal outliers), whereas the R-AO algorithm flagged only three observations. Similarly, the R-SD algorithm flagged 385 of the observations in the data set **E** as outliers, while the R-AO algorithm flagged 50 observations. As the data are very skewed, the smaller number of observations flagged by the skew-adjusted algorithm is closer to reality.

Figure 5 shows that the R-SD flagged the observations 355 and 499 in the data set **B** as bad leverage points. The observation 312 appears as a borderline case between the good and bad leverage points. R-AO on the other hand flagged the observations 355, 499 and 312 as orthogonal outliers. The rest of the observations flagged as

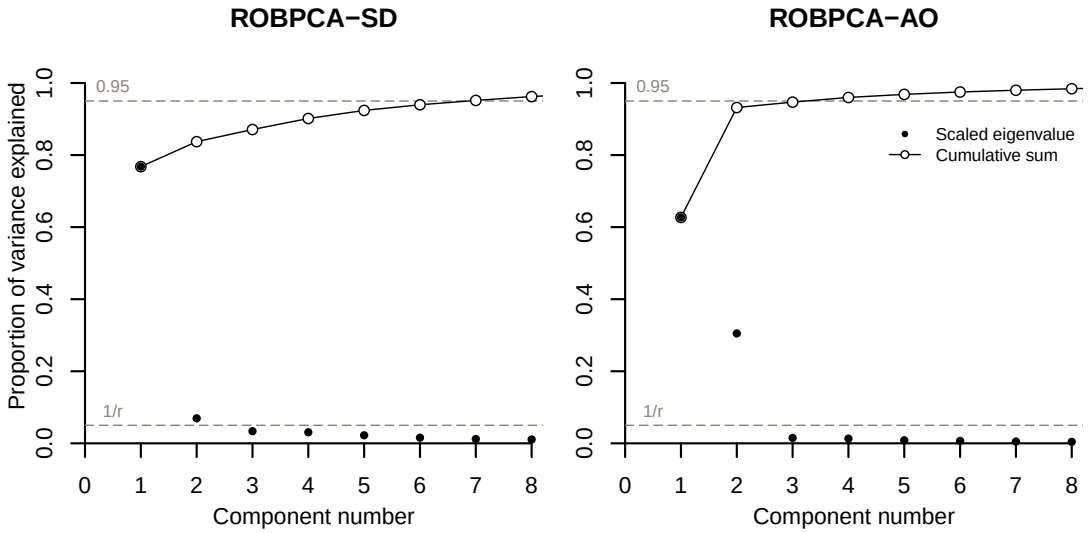


Figure 4: Scree plots for the using Stahel-Donoho outlyingness (ROBPCA-SD) and the modified ROBPCA method using adjusted outlyingness (ROBPCA-AO) performed on the data set **B** with added 0.95 and $1/r$ reference lines.

outliers in the R-SD outlier map are converted into regular observations in R-AO outlier map.

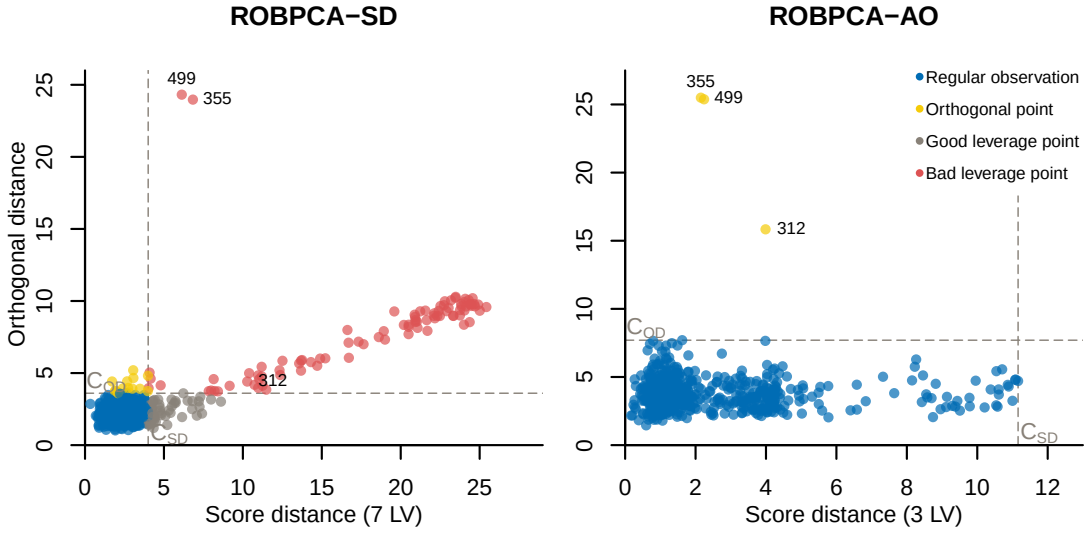


Figure 5: Outlier maps of the data set **B** produced by the ROBPCA method using Stahel-Donoho outlyingness (ROBPCA-SD) and the skew-adjusted ROBPCA method using adjusted outlyingness (ROBPCA-AO).

Very much alike the R-SD performed on the data set **B**, the R-SD applied to the

data set **E** flagged significant portion of the outliers as bad leverage points. As seen in Figure 6, some of the most extreme bad leverage points in the R-SD outlier map, observations 288, 503, 602 and 831, are converted into good leverage points by the R-AO. However, the majority of the observations flagged as bad leverage points by the R-SD are diagnosed as regular observations by the R-AO. This supports the statement in Subsection 2.3, which suggests that the tail of the skewed distribution tilts the PCA space towards it and needs to be adjusted for in order to distinguish the actual outliers. These include the observations 210, 212, 232 and 732 flagged as a bad leverage points by the R-SD, which appear as orthogonal outliers in the R-AO outlier map. From this point on, we no longer examine the results produced by the R-SD method.

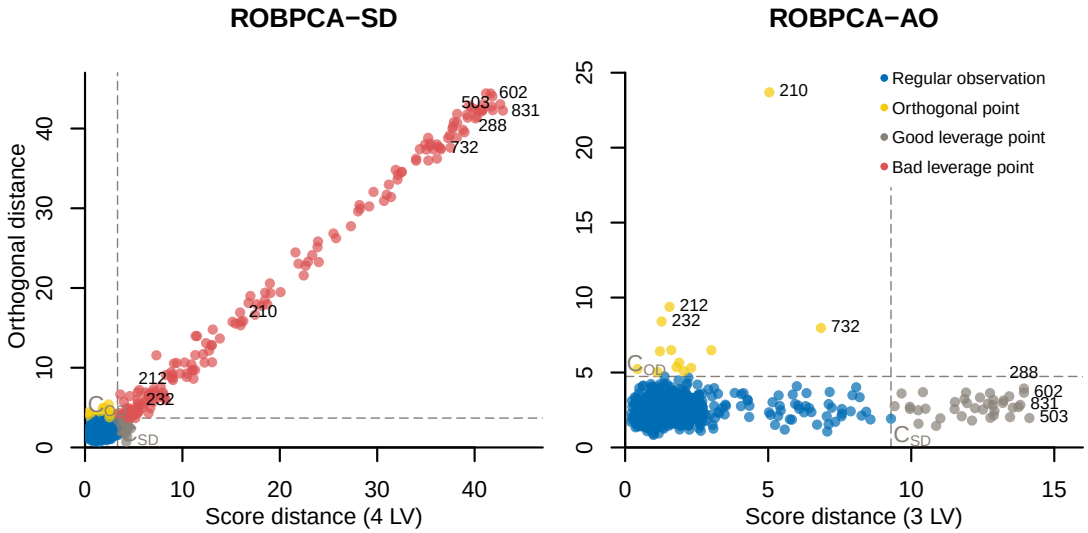


Figure 6: Outlier maps of the data set **E** produced by the ROBPCA method using Stahel-Donoho outlyingness (ROBPCA-SD) and the modified ROBPCA method using adjusted outlyingness (ROBPCA-AO).

Next we examine the univariate and multivariate nature of the outliers flagged by the ROBPCA-AO. For this, we color the outliers in the jittered strip chart introduced in Section 3. To better distinguish the outliers, the regular observations are transparent while the outliers are opaque. The results are depicted in Figure 7.

Figure 7a indicates that the observations flagged as outliers in the data set **B**, 312, 355 and 499, have atypical values for only singular variables. To be precise, the observation 312 attains an exceptionally small variable 7 value, while the observations 355 and 499 have exceptionally small variable 18 values. Thus, the observations are univariate outliers.

Figure 7b shows the observations flagged as outliers in the data set **E**. The good leverage points all lie in the remote tail area of the variable 5 distribution and attain variable 3 values near the median the bimodal distribution. Although not visible

in Figure 7b, many regular observations lie in the said tail area of the variable 5 distribution. However, those observations exclusively attain variable 3 values larger than the variable 3 antimode, situated on the positive side of the median. Furthermore, the good leverage points all have atypically small variable 4 values. This can be explained by the strong correlation of $\rho_S(\mathbf{E}_4, \mathbf{E}_5) = 0.85$ between the variable 4 distribution and the variable 5 distribution. These remarks suggest that the good leverage points are multivariate outliers. Similarly, a majority of the observations flagged as orthogonal outliers have atypically large variable 1 values and smaller than typical variable 2 values. This is not unexpected, for the two variables have a significant negative correlation of $\rho_S(\mathbf{E}_1, \mathbf{E}_2) = -0.54$.

To paint a clearer picture of how the variables contribute to each orthogonal outliers' residuals from the PCA space, we consult the complete decomposition contribution (CDC) indices (Alcala and Qin 2011). The CDC matrix for the squared prediction error index (SPE index or the squared orthogonal distances OD^2) is composed of the squared elements of the residual subspace

$$\tilde{\mathbf{X}}_{n,p} = \mathbf{X}_{n,p}(\mathbf{I}_{p,p} - \mathbf{P}_{p,k}\mathbf{P}_{p,k}^{\intercal}), \tag{4}$$

where $\mathbf{I}_{p,p}$ is the $p \times p$ identity matrix. The CDC indices for the orthogonal outliers in the data set $\mathbf{E}$ are presented in Table A5 in Appendix A. Indeed, the variables 1 and 2 attain CDC indices significantly larger than zero, which suggests that they contribute to the squared orthogonal distances of the orthogonal outliers. In addition, variables 4, 6, 7, 9, 11, 20 and 24 greatly affect some of the distances. By inspecting the rows of the CDC matrix, we see that most of the orthogonal outliers appear bivariate or multivariate outliers. Nevertheless, some observations, such as observation 212 with $\text{CDC}_1 = 67.6$ and $\text{CDC}_{i\neq 1} \leq 3.3$, are univariate. To conclude, the orthogonal outliers in the data set $\mathbf{E}$ consist of both multivariate and univariate outliers, which suggests that the measurement phenomenon responsible for the residuals varies between the observations.

4.3 Loadings

The principal components can be interpreted by inspecting the loading matrix $\mathbf{P}_{p,k}$ introduced in Subsection 2.2. The components are linear combinations of the original variables and the rows of the loading matrix depict the coefficients these variables attain in the construction. Thus, the magnitude of the coefficients measure the importance of the variables to the components with respect to the other variables. Consequently, the elements of the loading matrix are proportional to the linear correlation coefficients, $\rho(\mathbf{X}_i, \mathbf{T}_j)$, between the original data variables $\mathbf{X}_i$ and each unit scaled principal component $\mathbf{T}_j$ (Johnson and Wichern 2007). We denote the elements of the loading matrix $\mathbf{P}_{p,k}$ by $p_{i,j}$. Then, the correlation coefficients $\rho(\mathbf{X}_i, \mathbf{T}_j)$ are given by

$$\rho(\mathbf{X}_i, \mathbf{T}_j) = \frac{\text{Cov}(\mathbf{X}_i, \mathbf{T}_j)}{\sqrt{\text{Var}(\mathbf{X}_i)}\sqrt{\text{Var}(\mathbf{T}_j)}} = \frac{p_{i,j}\sqrt{\lambda_j}}{\sqrt{\sigma_{i,i}}}, \; i = 1, 2, ..., p \text{ and } j = 1, 2, ..., k, \tag{5}$$

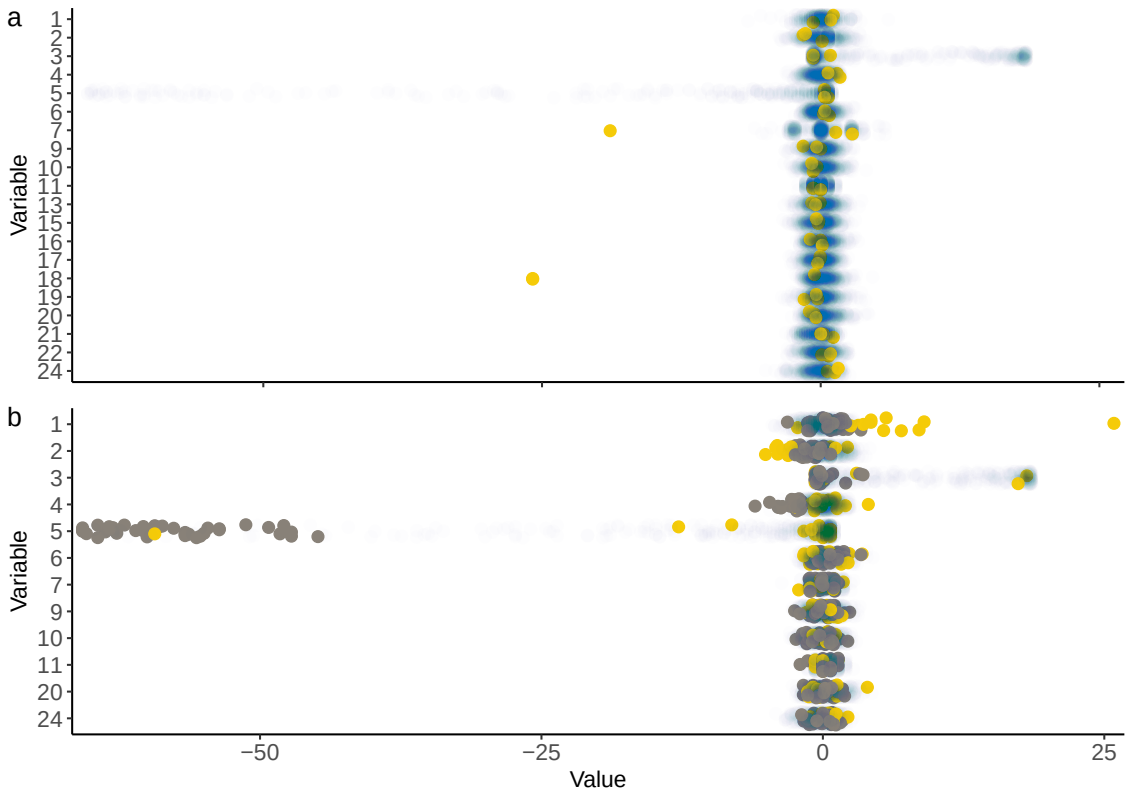


Figure 7: Jittered strip charts of the robustly standardized variables a) in the data set **B** and b) in the data set **E**, where the outliers flagged by ROBPCA-AO are colored as in the figures 5 and 6, i.e., regular observations are blue, orthogonal outliers are yellow, and good leverage points are gray.

where λ_j are the eigenvalues corresponding to the robust principal components P_j, and $\sigma_{i,i}$ are the diagonal elements of the robust covariance matrix $\hat{\mathbf{\Sigma}}$. Alternatively, $\sigma_{i,i}$ are the approximations of the $\mathrm{Var}(\mathbf{X}_i)$ without the presence of outliers. The ROBPCA-AO loading matrices for the data sets **B** and **E** are depicted in the tables A6 and A7 in Appendix A and visualized as bar plots in the figures 8 and 9. The figures 8 and 9 also depict the normalized eigenvalues corresponding to each component. These can be interpreted as the explained portion of variability in the data majority. It must be stated that these final robust eigenvalues differ from the eigenvalues used in selecting the number of components to be retained in the algorithms (Figure 4). The corresponding linear correlation coefficients $\rho(\mathbf{X}_i, \mathbf{T}_j)$ are depicted in the tables A8 and A9.

The first component of data the set **B** explains 83.0% of the variability in the data majority. Figure 8 shows that the component consists almost exclusively of the variable 5. In fact, $p_{5,1} = -1.00 = \rho(\mathbf{X}_5, \mathbf{T}_1)$ and the first component increases with decreasing variable 5 value. Similarly, the second component, which explains 15.8% of the data majority variance, increases with only one of the values, decreasing

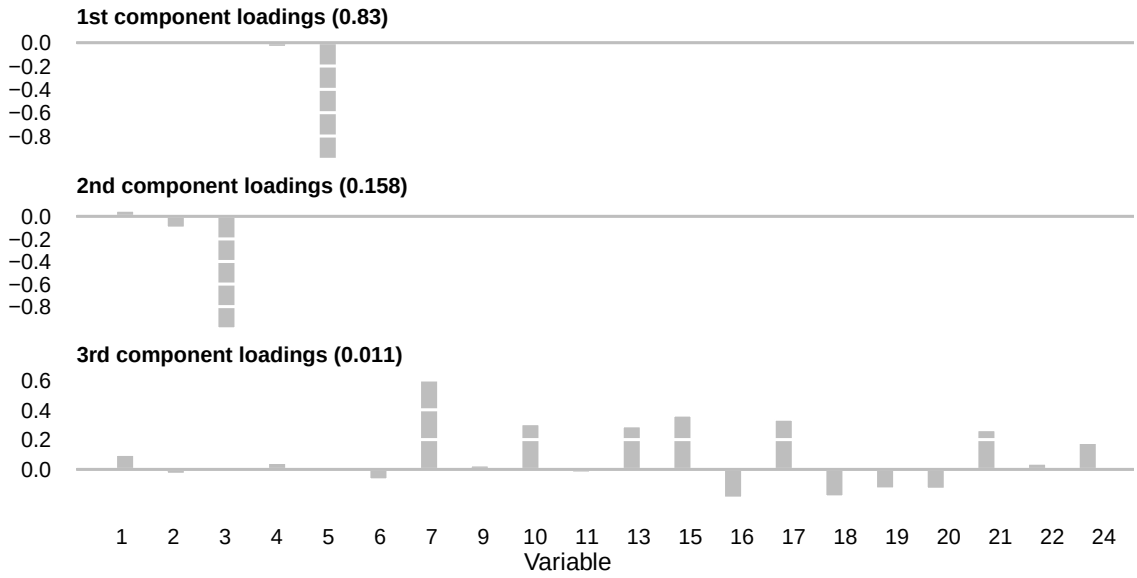


Figure 8: Component loadings of the data set **B** produced by the ROBPCA-AO, where the numbers in parentheses are the normalized eigenvalues corresponding to each component.

variable 3 value ($p_{3,2} = -0.99,\ \rho(\mathbf{X}_3, \mathbf{T}_2) = -1.00$). The third and final component retained in the construction of the robust PCA space for data set **B** explains 1.10% of the variability in the data majority and is strongly impacted by the variable 7 ($p_{7,3} = 0.60,\ \rho(\mathbf{X}_7, \mathbf{T}_3) = 1.00$). In addition, the coefficients of the variables 10, 13, 15, 17 and 21 in the third component are larger than 0.2 in magnitude. Furthermore, the correlation between the component and these variables is 0.99. This suggests that while the component increases strongest with increasing variable 7 value, these six variables vary together. In fact, the majority of the variables are strongly correlated with the third component, which may explain why the three components could replace the original 20 variables with little loss of information.

The data set **E** loadings depicted in Figure 9 tell a similar story. The first component, which explains 0.68% of the variability in the data majority, is strongly impacted by the variable 5 ($p_{5,1} = -0.98,\ \rho(\mathbf{X}_5, \mathbf{T}_1) = -0.99$). The second component, which constitutes 30.9% of the data majority variance, increases with decreasing variable 3 value ($p_{3,2} = -0.97,\ \rho(\mathbf{X}_3, \mathbf{T}_2) = -0.96$). The third component explains 1.10% of the variability in the data majority. Since $p_{1,3} = 0.45$, $p_{20,3} = 0.46$ and $p_{24,3} = -0.54$, it increases with variable 1 and variable 20 values, and the decreasing variable 24 value. The coefficients of the variables 2 and 11 in the third component are larger than 0.2 in magnitude, which implies that the five variables vary together. As before, the majority of the variables are strongly correlated with the third component.

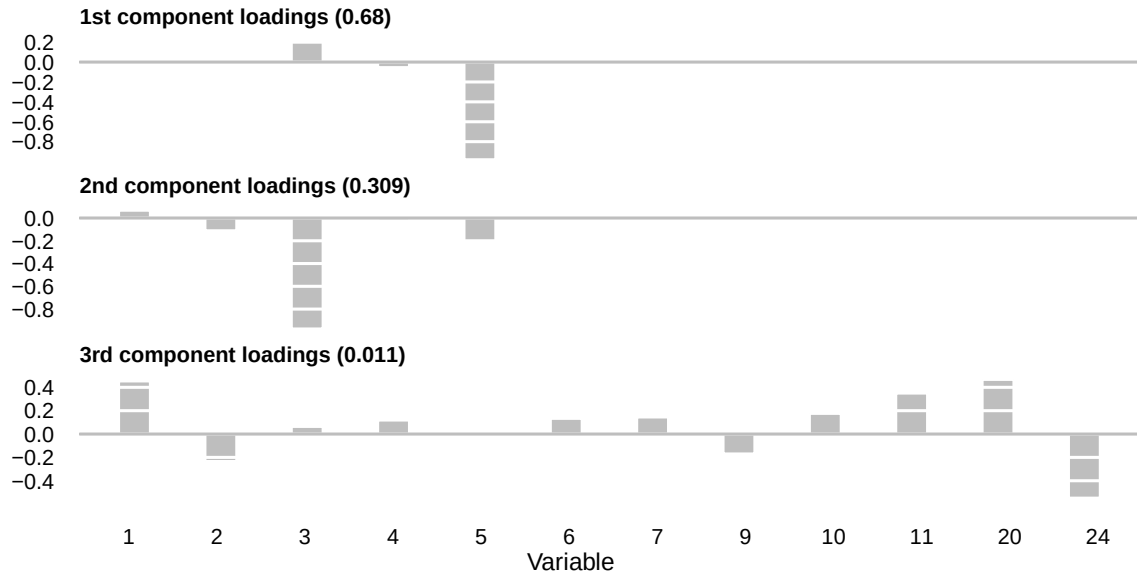


Figure 9: Component loadings of the data set $\mathbf{E}$ produced by the ROBPCA-AO, where the numbers in parentheses are the normalized eigenvalues corresponding to each component.

4.4 PCA reconstruction

Next we inspect the reconstruction matrix $\hat{\mathbf{X}}_{n,p}$ given by

$$\hat{\mathbf{X}}_{n,p} = \mathbf{T}_{n,k}\mathbf{P}_{p,k}^{\intercal}. \tag{6}$$

If $k = r$, where r is the $\min(p, n-1)$ such that all of the eigenvectors are used in the reconstruction, $\mathbf{P}_{p,k}\mathbf{P}_{p,k}^{\intercal}$ is the identity matrix. This results in

$$\hat{\mathbf{X}}_{n,p} = \mathbf{T}_{n,r}\mathbf{P}_{p,r}^{\intercal} = (\mathbf{X}_{n,p} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^{\intercal})\mathbf{P}_{p,r}\mathbf{P}_{p,r}^{\intercal} = \mathbf{X}_{n,p} - \mathbf{1}_n\hat{\boldsymbol{\mu}}^{\intercal}, \tag{7}$$

which implies that the reconstruction is perfect. On the other hand, if $k < r$, the rows of the loading matrix are no longer orthogonal. In general, $\hat{\mathbf{X}}_{n,p}$ refers to the centered observations projected onto the space spanned by the first k robust principal components in the original coordinate system.

The ROBPCA-AO outlier maps in the figures 5 and 6 show that the vast majority of the orthogonal distances fall below the orthogonal cutoff value introduced in Subsection 2.3. This suggests that few observations are inconsistent with the model and the overall reconstruction error is sufficiently small. To compare the univariate structure of the original data sets $\mathbf{B}$ and $\mathbf{E}$ with that of the reconstructed data sets $\hat{\mathbf{B}}$ and $\hat{\mathbf{E}}$, the univariate distributions are plotted one on top of other in a density plot introduced in Section 3. Before plotting, the reconstructed data $\hat{\mathbf{B}}$ and $\hat{\mathbf{E}}$ are robustly centered and scaled, as in Section 3, so that each variable has a median of 0 and unit median absolute deviation. The results are depicted in the figures 10 and 11.

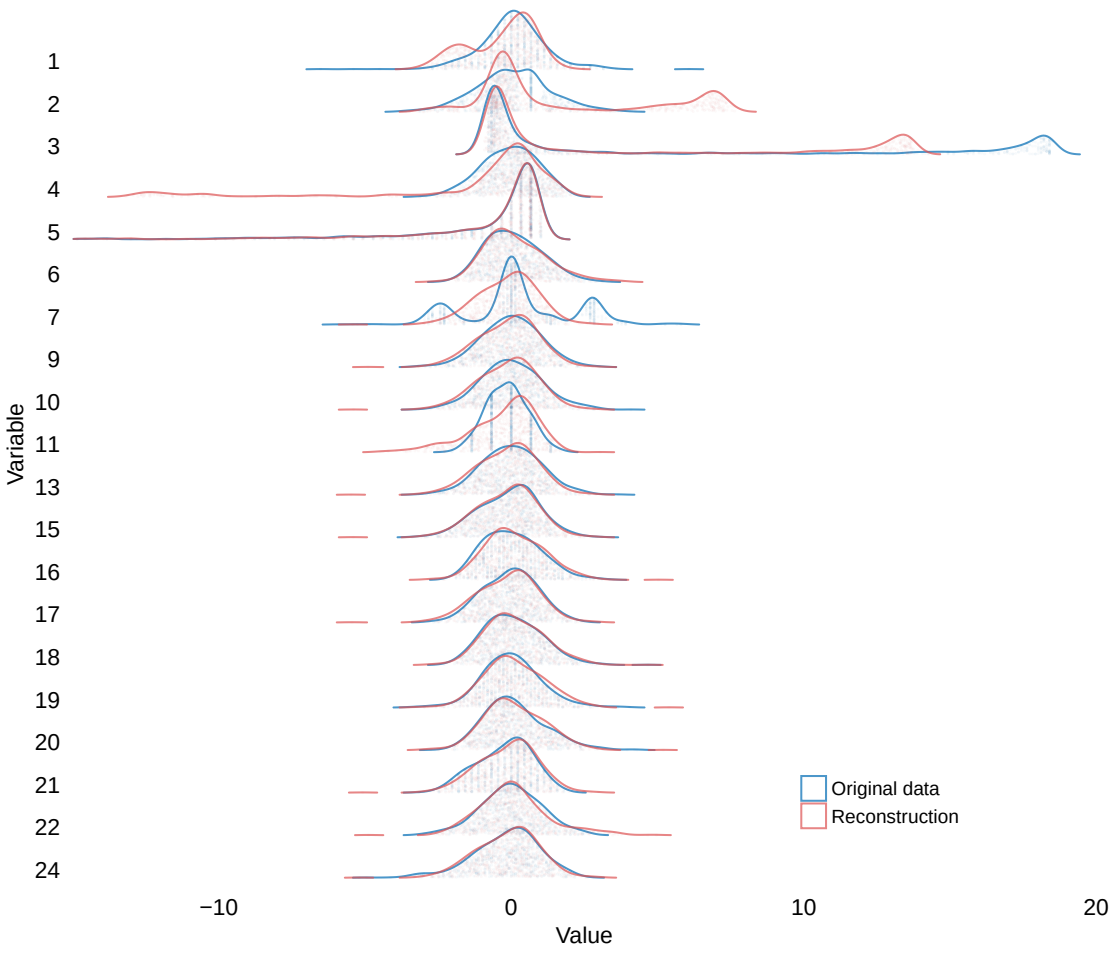


Figure 10: Density plots of the robustly standardized variables in the data set **B**, where the line is displayed when its height is 0.1% or more relative to the overall maximum.

Figure 10 shows that the the majority of the unimodal distributions in the data set **B** are reconstructed fairly accurately. On the other hand, the originally unimodal variable 1 distribution is rendered bimodal in the reconstruction. Similarly, the variable 11 reconstruction is more scattered than the original discrete distribution, and the tail of the variable 4 distribution is elongated. The bimodal variable 2 and variable 3 distributions are retained bimodal in the reconstruction matrix. The variable 3 reconstruction is the more accurate of the two reconstructions; while the reconstructed distribution is less scattered than the original distribution, the shapes of the peaks are well maintained. On the contrary, the shape of the variable 2 distribution is altered such that the right peak gets significantly spread out. Finally, the originally multimodal variable 7 distribution is rendered unimodal in the reconstruction. As seen in Figure 11, the same remarks hold true for the data set **E**.

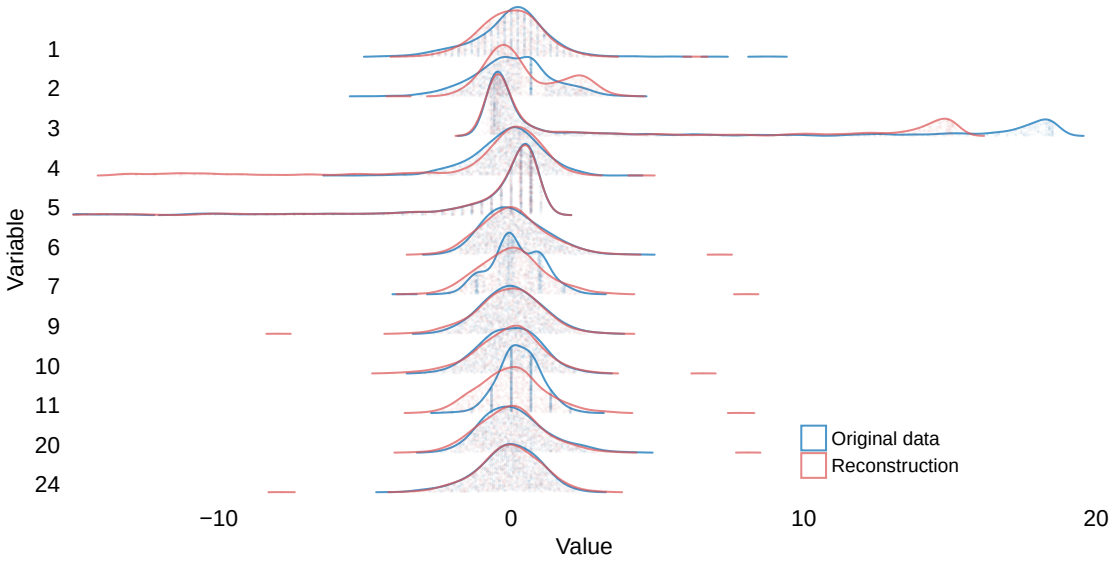


Figure 11: Density plots of the robustly standardized variables in the data set **E**, where the line is displayed when its height is 0.1% or more relative to the overall maximum.

5 Summary and conclusions

In this thesis we have applied the robust principal component analysis methods ROBPCA (Hubert et al. 2005) and its modification for skewed data (Hubert et al. 2009) to two data sets from the field of production engineering. Our aim was to detect the outlying observations in the data that were shown to be significantly asymmetric and overall non-Gaussian. The outliers are identified by their large deviation from the robust center of the data, and the subspace spanned by the robust principal components. To evaluate each individual variable's impact to this deviation, a heuristic from the field of multivariate quality control is applied. The outliers are inspected using visualization methods, including outlier maps. Finally, we analyze the robust principal components to gain a better understanding of the sources of variation in the data. The quality of our models is assessed by a specialized density plot to further ensure the validity of the results.

Confirming the hypothesis, the skew-adjusted algorithm proved to be more accurate in detecting the anomalous observations since it flagged 0.5% (**B**) and 5.9% (**E**) of the data as outliers. In contrast, the ROBPCA algorithm flagged 39.8% (**B**) and 45.2% (**E**) of the data as outliers, which suggests that the regular observations located in the tail area of skewed distributions are mistaken as anomalies. The skew-adjusted algorithm identified orthogonal outliers (observations with a large orthogonal deviation from the PCA space with inlying projections) in both of the data sets, and good leverage points (observations close to the PCA space but far from the regular observations) in one of the data sets.

Analysis of the principal components revealed that two nearly uncorrelated variables are responsible for over 90% of the variability (information) in the data. Consequently, these variables contribute almost exclusively to the large deviation of the good leverage points. We also identified two variables that frequently contribute to the orthogonal outliers' large orthogonal distances from the PCA space. However, no exhaustive proof of systematic behavior of the orthogonal outliers could be presented.

The small portion of outlying observations suggests that the vast majority of the data can be regarded as consistent with the model. Thus, the models preserve the multivariate structure of the data well. Analysis of the PCA reconstructions revealed that the characteristics of the univariate distributions (i.e. spread, symmetry and modality) were mostly preserved, although the model neglected some discrete and multimodal features of the data. The quality of the reconstructions can most likely be attributed to the third principal components, which unlike the first two components are strongly correlated with the majority of the variables in the data. On the contrary, the components only explain 1.1% of the variability in the data majority in both of the data sets.

The largest problem faced in the study concerned the number of principal components to be selected in the algorithms. The scree plots and the selection criteria

used are fairly rough approximations with no extensive theoretical evidence. A more refined technique mentioned earlier in this thesis would be based on the predicted residual error sum of squares (PRESS). Unfortunately, the MATLAB implementation of the skew-adjusted algorithm, a part of LIBRA: the MATLAB Library for Robust Analysis (Verboven and Hubert 2010), did not include the option to apply the robust PRESS algorithm (Hubert and Engelen 2007).

As mentioned in Section 1, PCA is often followed by other multivariate techniques. The findings in this thesis lay the groundwork for further analysis of the data. The multimodal nature and the underlying correlation structure of the data indicate a presence of multiple populations. To sort the observations into homogenous groups, cluster analysis methods, such as k-means algorithm, could be applied. In the wider context of improving the production process, robust logistic regression methods could be used to determine whether the measurement phenomena responsible for the outlying observations have negative connotations.

References

Alcala, C. F. and S. J. Qin (2011). "Analysis and generalization of fault diagnosis methods for process monitoring". In: *Journal of Process Control* 21.3, pp. 322–330. DOI: 10.1016/j.jprocont.2010.10.005.

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey (1972). "Robust Estimates of Location: Survey and Advances". In: *Princeton, NJ: Princeton University Press*.

Brys, G., M. Hubert, and A. Struyf (2004). "A Robust Measure of Skewness". In: *Journal of Computational and Graphical Statistics* 13.4, pp. 996–1017. DOI: 10.1198/106186004x12632.

Daszykowski, M., K. Kaczmarek, Y. V. Heyden, and B. Walczak (2007). "Robust statistics in data analysis — A review". In: *Chemometrics and Intelligent Laboratory Systems* 85.2, pp. 203–219. DOI: 10.1016/j.chemolab.2006.06.016.

Donoho, D. L. and M. Gasko (1992). "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness". In: *The Annals of Statistics* 20.4, pp. 1803–1827. DOI: 10.1214/aos/1176348890.

Hubert, M. and S. Engelen (2007). "Fast cross-validation of high-breakdown resampling methods for PCA". In: *Computational Statistics & Data Analysis* 51.10, pp. 5013–5024. DOI: 10.1016/j.csda.2006.08.031.

Hubert, M., P. J. Rousseeuw, and K. V. Branden (2005). "ROBPCA: A New Approach to Robust Principal Component Analysis". In: *Technometrics* 47.1, pp. 64–79. DOI: 10.1198/004017004000000563.

Hubert, M., P. J. Rousseeuw, and T. Verdonck (2009). "Robust PCA for skewed data and its outlier map". In: *Computational Statistics & Data Analysis* 53.6, pp. 2264–2274. DOI: 10.1016/j.csda.2008.05.027.

Johnson, R. A. and D. W. Wichern (2007). *Applied multivariate statistical analysis*. 6th ed. Prentice Hall.

Jolliffe, I. T. (1986). "Principal Component Analysis and Factor Analysis". In: *Principal Component Analysis Springer Series in Statistics*, pp. 115–128. DOI: 10.1007/978-1-4757-1904-8_7.

Lehman, A. (2005). *JMP for basic univariate and multivariate statistics: a step-by-step guide*. SAS Press, p. 123.

Myers, J. L. and A. Well (2003). *Research design and statistical analysis*. Lawrence Erlbaum Associates, p. 508.

Rousseeuw, P. J. (1984). "Least Median of Squares Regression". In: *Journal of the American Statistical Association* 79.388, p. 871. DOI: 10.2307/2288718.

Rousseeuw, P. J. and M. Hubert (2011). "Robust statistics for outlier detection". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1, pp. 73–79. DOI: 10.1002/widm.2.

Rousseeuw, P. J. and K. Van Driessen (1999). "A Fast Algorithm for the Minimum Covariance Determinant Estimator". In: *Technometrics* 41.3, p. 212. DOI: 10.2307/1270566.

Stahel, W. A. (1981). "Robust Estimation: Infinitesimal Optimality and Covariance Matrix Estimators". In: *Ph.D. thesis, ETH, Zürich*.

Verboven, S. and M. Hubert (2010). "MATLAB library LIBRA". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 2.4, pp. 509–515. DOI: 10.1002/wics.96.

A Tables

Table A1: Spearman rank correlations of the data set **E**.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_9	X_{10}	X_{11}	X_{20}	X_{24}
X_1	1	-0.54	-0.54	0.04	0.03	0.05	0.07	-0.17	0.07	0.1	-0.04	-0.14
X_2	-0.54	1	0.98	0.04	0.06	0.02	-0.06	0.01	0.06	0.03	0	-0.05
X_3	-0.54	0.98	1	0.03	0.05	0.02	-0.07	0.03	0.04	0	0	-0.02
X_4	0.04	0.04	0.03	1	0.85	0	0.06	-0.04	0.24	0.14	-0.03	-0.1
X_5	0.03	0.06	0.05	0.85	1	-0.07	0.03	-0.02	0.22	0.13	-0.01	-0.08
X_6	0.05	0.02	0.02	0	-0.07	1	-0.09	-0.12	-0.01	0.03	-0.03	-0.11
X_7	0.07	-0.06	-0.07	0.06	0.03	-0.09	1	0.04	0.1	0.1	0.12	-0.04
X_9	-0.17	0.01	0.03	-0.04	-0.02	-0.12	0.04	1	-0.05	-0.12	0.08	0.14
X_{10}	0.07	0.06	0.04	0.24	0.22	-0.01	0.1	-0.05	1	0.3	0.01	-0.11
X_{11}	0.1	0.03	0	0.14	0.13	0.03	0.1	-0.12	0.3	1	0.34	-0.56
X_{20}	-0.04	0	0	-0.03	-0.01	-0.03	0.12	0.08	0.01	0.34	1	-0.32
X_{24}	-0.14	-0.05	-0.02	-0.1	-0.08	-0.11	-0.04	0.14	-0.11	-0.56	-0.32	1

Table A2: Spearman rank correlations of the data set **B**.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_9	X_{10}	X_{11}	X_{13}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{24}
X_1	1	-0.59	-0.59	0.14	0.13	-0.06	0.09	-0.14	0.08	0.13	0.13	0.11	-0.31	0.21	-0.19	-0.02	0.08	0.2	-0.08	-0.06
X_2	-0.59	1	0.97	-0.03	-0.02	0.01	-0.03	0.02	0.05	0.01	0.04	0.03	-0.03	0.04	-0.05	0.03	0.04	0.08	0.01	-0.02
X_3	-0.59	0.97	1	-0.03	-0.02	0.02	-0.04	0.02	0.05	0	0.04	0.03	-0.03	0.04	-0.07	0.02	0.03	0.08	0.02	-0.02
X_4	0.14	-0.03	-0.03	1	0.92	-0.04	0.11	-0.04	0.14	0.06	0.17	0.12	-0.23	0.22	-0.27	0	0.07	0.18	-0.12	-0.01
X_5	0.13	-0.02	-0.02	0.92	1	-0.07	0.12	-0.06	0.16	0.06	0.17	0.13	-0.2	0.21	-0.23	-0.03	0.05	0.17	-0.08	0
X_6	-0.06	0.01	0.02	-0.04	-0.07	1	-0.01	-0.03	0	-0.01	-0.01	-0.06	0.18	-0.08	0.11	-0.06	-0.05	-0.2	0	-0.06
X_7	0.09	-0.03	-0.04	0.11	0.12	-0.01	1	-0.03	0.09	0.07	0.12	0.08	-0.21	0.16	-0.18	0.07	0.1	0.22	-0.08	0.03
X_9	-0.14	0.02	0.02	-0.04	-0.06	-0.03	-0.03	1	0.03	-0.06	0.02	0.02	0.08	-0.03	0.03	0.02	-0.02	0.01	-0.02	0.1
X_{10}	0.08	0.05	0.05	0.14	0.16	0	0.09	0.03	1	0.2	0.91	0.55	-0.09	0.49	-0.21	-0.07	0.11	0.4	0.08	0.04
X_{11}	0.13	0.01	0	0.06	0.06	-0.01	0.07	-0.06	0.2	1	0.26	-0.05	-0.18	0	-0.25	0.27	0.46	0.1	-0.27	-0.53
X_{13}	0.13	0.04	0.04	0.17	0.17	-0.01	0.12	0.02	0.91	0.26	1	0.46	-0.17	0.44	-0.29	-0.01	0.25	0.51	0.04	0
X_{15}	0.11	0.03	0.03	0.12	0.13	-0.06	0.08	0.02	0.55	-0.05	0.46	1	-0.39	0.91	-0.28	-0.34	-0.44	0.37	-0.07	0.24
X_{16}	-0.31	-0.03	-0.03	-0.23	-0.2	0.18	-0.21	0.08	-0.09	-0.18	-0.17	-0.39	1	-0.63	0.7	-0.08	-0.07	-0.37	0.34	0.01
X_{17}	0.21	0.04	0.04	0.22	0.21	-0.08	0.16	-0.03	0.49	0	0.44	0.91	-0.63	1	-0.45	-0.31	-0.35	0.44	-0.14	0.23
X_{18}	-0.19	-0.05	-0.07	-0.27	-0.23	0.11	-0.18	0.03	-0.21	-0.25	-0.29	-0.28	0.7	-0.45	1	-0.02	-0.17	-0.42	0.18	0.04
X_{19}	-0.02	0.03	0.02	0	-0.03	-0.06	0.07	0.02	-0.07	0.27	-0.01	-0.34	-0.08	-0.31	-0.02	1	0.57	-0.09	-0.46	-0.36
X_{20}	0.08	0.04	0.03	0.07	0.05	-0.05	0.1	-0.02	0.11	0.46	0.25	-0.44	-0.07	-0.35	-0.17	0.57	1	0.11	-0.15	-0.45
X_{21}	0.2	0.08	0.08	0.18	0.17	-0.2	0.22	0.01	0.4	0.1	0.51	0.37	-0.37	0.44	-0.42	-0.09	0.11	1	0.07	0.26
X_{22}	-0.08	0.01	0.02	-0.12	-0.08	0	-0.08	-0.02	0.08	-0.27	0.04	-0.07	0.34	-0.14	0.18	-0.46	-0.15	0.07	1	0.42
X_{24}	-0.06	-0.02	-0.02	-0.01	0	-0.06	0.03	0.1	0.04	-0.53	0	0.24	0.01	0.23	0.04	-0.36	-0.45	0.26	0.42	1

Table A3: Medcouples of the data set **E**.

X	medcouple
X_1	0.000
X_2	0.063
X_3	0.926
X_4	-0.089
X_5	-0.714
X_6	0.111
X_7	0.310
X_9	0.041
X_{10}	-0.055
X_{11}	0.560
X_{20}	0.111
X_{24}	-0.059

Table A4: Medcouples of data set **B**.

X	medcouple
X_1	0.200
X_2	-0.166
X_3	0.946
X_4	-0.069
X_5	-0.826
X_6	0.083
X_7	0.543
X_9	0.044
X_{10}	0.004
X_{11}	-0.333
X_{13}	-0.013
X_{15}	-0.011
X_{16}	0.067
X_{17}	0.072
X_{18}	0.097
X_{19}	-0.015
X_{20}	0.143
X_{21}	0.001
X_{22}	0.037
X_{24}	-0.158

Table A5: The complete decomposition contributions (CDC) for the residuals of the orthogonal outliers in the data set **E**.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_9	X_{10}	X_{11}	X_{20}	X_{24}
53	6.2	0.3	0	0.1	0	1.7	5.2	0.2	1.1	0.1	12.3	0.1
125	5.6	12.9	0.3	0	0	0.5	2.5	0.4	0	1.4	1.2	3.5
172	8	5.4	0.2	0	0	3.8	0.3	3.4	1	1.1	0	5.5
210	436.3	1.4	1.4	3.1	0	0.9	6.7	7.8	4.8	19.3	14	65.6
212	67.6	1.7	0.2	2.5	0	0	0.7	0.2	1.3	2.6	8	3.3
232	53.9	0.1	0.2	1	0	1.8	0.1	0.8	2	2.4	1.6	6.7
277	16.9	5.2	0.2	0.5	0	0.1	0.1	1.2	0	1.2	3.4	3
359	12.8	12.7	0.4	1.2	0	3.5	0.6	0	2	1.3	3.9	2.9
362	17.7	5.8	0.2	2.8	0	1.7	0.3	4.5	0.7	0.7	4.1	3.6
369	0.1	4.7	0	20.4	0	10.3	0.5	1.4	0.7	0.3	3	0.6
379	2.1	15.1	0.3	0.3	0	5.1	0.4	0	0.1	0	0.7	1.8
518	3.7	13.4	0.3	0.6	0	1	0.6	0.6	1.1	1	0.9	2.1
732	44.4	0.6	0.1	0.1	0	2.3	2.6	0.9	0.2	1.3	6.2	5

Table A6: Robust loadings (ROBPCA-AO) of the data set **E**.

	P_1	P_2	P_3
X_1	-0.02	0.06	0.45
X_2	0.01	-0.11	-0.23
X_3	0.21	-0.97	0.06
X_4	-0.05	-0.02	0.12
X_5	-0.98	-0.21	-0.02
X_6	0.00	0.00	0.13
X_7	0.00	0.00	0.14
X_9	0.00	-0.01	-0.17
X_{10}	-0.01	0.00	0.17
X_{11}	0.00	-0.01	0.35
X_{20}	-0.01	0.01	0.46
X_{24}	0.01	-0.01	-0.54

Table A7: Robust loadings (ROBPCA-AO) of the data set **B**.

	P_1	P_2	P_3
X_1	0.00	0.05	0.10
X_2	-0.01	-0.10	-0.03
X_3	0.00	-0.99	0.00
X_4	-0.04	-0.01	0.04
X_5	-1.00	0.00	-0.01
X_6	0.01	-0.01	-0.07
X_7	0.00	-0.01	0.60
X_9	0.00	0.00	0.03
X_{40}	0.00	-0.01	0.30
X_{11}	0.00	0.01	-0.02
X_{13}	0.00	-0.01	0.29
X_{15}	0.00	0.00	0.36
X_{16}	0.01	0.00	-0.19
X_{17}	0.00	0.00	0.33
X_{18}	0.01	0.01	-0.18
X_{19}	0.00	0.00	-0.13
X_{20}	0.00	-0.01	-0.13
X_{21}	-0.01	0.01	0.26
X_{22}	0.00	0.00	0.04
X_{24}	0.00	0.01	0.18

Table A8: Robust correlation coefficients (ROBPCA-AO) of the data set **E**.

	P_1	P_2	P_3
X_1	-0.19	0.59	0.79
X_2	0.14	-0.91	-0.38
X_3	0.29	-0.96	0.01
X_4	-0.93	-0.24	0.27
X_5	-0.99	-0.13	0.00
X_6	0.24	-0.02	0.97
X_7	-0.04	0.01	1.00
X_9	0.23	-0.27	-0.93
X_{10}	-0.43	-0.09	0.90
X_{11}	-0.04	-0.12	0.99
X_{20}	-0.12	0.05	0.99
X_{24}	0.19	-0.07	-0.98

Table A9: Robust correlation coefficients (ROBPCA-AO) of the data set **B**.

	P_1	P_2	P_3
X_1	0.15	0.87	0.47
X_2	-0.20	-0.97	-0.09
X_3	0.00	-1.00	0.00
X_4	-0.98	-0.12	0.15
X_5	-1.00	0.00	0.00
X_6	0.62	-0.23	-0.75
X_7	-0.07	-0.05	1.00
X_9	-0.24	0.26	0.94
X_{10}	-0.13	-0.09	0.99
X_{11}	-0.75	0.60	-0.28
X_{13}	-0.12	-0.06	0.99
X_{15}	-0.12	0.00	0.99
X_{16}	0.31	0.03	-0.95
X_{17}	-0.13	-0.01	0.99
X_{18}	0.42	0.13	-0.90
X_{19}	-0.24	-0.13	-0.96
X_{20}	-0.05	-0.27	-0.96
X_{21}	-0.21	0.12	0.97
X_{22}	0.83	-0.12	0.55
X_{24}	0.00	0.31	0.95