Focus Article

Robust statistics for outlier detection

Peter J. Rousseeuw and Mia Hubert

When analyzing data, outlying observations cause problems because they may strongly influence the result. Robust statistics aims at detecting the outliers by searching for the model fitted by the majority of the data. We present an overview of several robust methods and outlier detection tools. We discuss robust procedures for univariate, low-dimensional, and high-dimensional data such as estimation of location and scatter, linear regression, principal component analysis, and classification. *WIREs Data Mining Knowl Discov* 2011 1 73–79
DOI: 10.1002/widm.2

INTRODUCTION

In real data sets, it often happens that some observations are different from the majority. Such observations are called *outliers*. Outlying observations may be errors, or they could have been recorded under exceptional circumstances, or belong to another population. Consequently, they do not fit the model well. It is very important to be able to detect these outliers.

In practice, one often tries to detect outliers using diagnostics starting from a classical fitting method. However, classical methods can be affected by outliers so strongly that the resulting fitted model does not allow to detect the deviating observations. This is called the *masking* effect. In addition, some good data points might even appear to be outliers, which is known as *swamping*. To avoid these effects, the goal of *robust statistics* is to find a fit that is close to the fit we would have found without the outliers. We can then identify the outliers by their large deviation from that robust fit.

First, we describe some robust procedures for estimating univariate location and scale. Next, we discuss multivariate location and scatter, as well as linear regression. We also give a summary of available robust methods for principal component analysis (PCA), classification, and clustering. For a more extensive review, see Ref 1. Some full-length books on this topic are Refs 2, 3.

*Correspondence to: peter@rousseeuw.net
Renaissance Technologies, New York, and Katholieke Universiteit Leuven, Belgium

DOI: 10.1002/widm.2

ESTIMATING UNIVARIATE LOCATION AND SCALE

As an example, suppose we have five measurements of a length:

$$6.27, \quad 6.34, \quad 6.25, \quad 6.31, \quad 6.28 \qquad (1)$$

and we want to estimate its true value. For this, one usually computes the *mean* $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, which in this case equals $\bar{x} = (6.27 + 6.34 + 6.25 + 6.31 + 6.28)/5 = 6.29$. Let us now suppose that the fourth measurement has been recorded wrongly and the data become

$$6.27, \quad 6.34, \quad 6.25, \quad 63.1, \quad 6.28. \qquad (2)$$

In this case, we obtain $\bar{x} = 17.65$, which is far from the unknown true value. On the contrary, we could also compute the median of these data. To this end, we sort the observations in (2) from smallest to largest:

$$6.25 \leq 6.27 \leq 6.28 \leq 6.34 \leq 63.10.$$

The median is then the middle value, yielding 6.28, which is still reasonable. We say that the median is more robust against an outlier.

More generally, the location-scale model states that the n univariate observations x_i are independent and identically distributed (i.i.d.) with distribution function $F[(x - \mu)/\sigma]$ where F is known. Typically, F is the standard Gaussian distribution function Φ. We then want to find estimates for the center μ and the scale parameter σ.

The classical estimate of location is the mean. As we saw above, the mean is very sensitive to even 1 aberrant value out of the n observations. We say that the *breakdown value*[4,5] of the sample mean is $1/n$, so it is 0% for large n. In general, the breakdown

value is the smallest proportion of observations in the data set that need to be replaced to carry the estimate arbitrarily far away. See Ref 6 for precise definitions and extensions. The robustness of an estimator is also measured by its *influence function*,[7] which measures the effect of one outlier. The influence function of the mean is unbounded, which again illustrates that the mean is not robust.

For a general definition of the median, we denote the ith ordered observation as $x_{(i)}$. Then, the median is $x_{[(n+1)/2]}$ if n is odd, and $[x_{(n/2)} + x_{(n/2+1)}]/2$ if n is even. Its breakdown value is about 50%, meaning that the median can resist up to 50% of outliers, and its influence function is bounded. Both properties illustrate the median's robustness.

The situation for the scale parameter σ is similar. The classical estimator is the standard deviation $s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}$. Because a single outlier can already make s arbitrarily large, its breakdown value is 0%. For instance, for the clean data (1) above, we have $s = 0.035$, whereas for the data (2) with the outlier, we obtain $s = 25.41$! A robust measure of scale is the median of all absolute deviations from the median (MAD), given by the **m**edian of all **a**bsolute **d**eviations from the median:

$$\text{MAD} = 1.483 \operatorname*{median}_{i=1,\ldots,n} \left| x_i - \operatorname*{median}_{j=1,\ldots,n}(x_j) \right|. \qquad (3)$$

The constant 1.483 is a correction factor that makes the MAD unbiased at the normal distribution. The MAD of (2) is the same as that of (1), namely 0.044. We can also use the Q_n estimator,[8] defined as

$$Q_n = 2.2219\,\{|x_i - x_j|; i < j\}_{(k)}$$

with $k = \binom{h}{2} \approx \binom{n}{2}/4$ and $h = \lfloor \frac{n}{2} \rfloor + 1$. Here, $\lfloor \ldots \rfloor$ rounds down to the nearest integer. This scale estimator is thus the first quartile of all pairwise differences between two data points. The breakdown value of both the MAD and the Q_n estimator is 50%.

Also popular is the interquartile range (IQR) defined as the difference between the third and first quartiles, that is, $\text{IQR} = x_{\lceil 3n/4 \rceil} - x_{\lfloor n/4 \rfloor}$ (where $\lceil \ldots \rceil$ rounds up to the nearest integer). Its breakdown value is only 25% but it has a simple interpretation.

The robustness of the median and the MAD comes at a cost: At the normal model they are less efficient than the mean. To find a better balance between robustness and efficiency, many other robust procedures have been proposed such as M-estimators.[9] They are defined implicitly as the solution of the equation

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\theta}}{\hat{\sigma}}\right) = 0 \qquad (4)$$

for a real function ψ. The denominator $\hat{\sigma}$ is an initial robust scale estimate such as the MAD. A solution to (4) can be found by the Newton–Raphson algorithm, starting from the initial location estimate $\hat{\theta}^{(0)} = \text{median}_i(x_i)$. Popular choices for ψ are the Huber function $\psi(x) = x \min(1, c/|x|)$, and Tukey's biweight function $\psi(x) = x(1 - (x/c)^2)^2 I(|x| \le c)$. These M-estimators contain a tuning parameter c, which needs to be chosen in advance.

People often use rules to detect outliers. The classical rule is based on the z-scores of the observations given by

$$z_i = (x_i - \bar{x})/s \qquad (5)$$

where s is the standard deviation. More precisely, the rule flags x_i as outlying if $|z_i|$ exceeds 2.5, say. But in the above-mentioned example (2) with the outlier, the z-scores are

$$-0.45, \quad -0.45, \quad -0.45, \quad 1.79, \quad -0.45$$

so none of them attains 2.5. The largest value is only 1.79, which is quite similar to the largest z-score for the clean data (1), which equals 1.41. The z-score of the outlier is small because it subtracts the nonrobust mean (which was drawn toward the outlier) and because it divides by the nonrobust standard deviation (which the outlier has made much larger than in the clean data). Plugging robust estimators of location and scale into (5), such as the median and the MAD, yields the robust scores

$$\left(x_i - \operatorname*{median}_{j=1,\ldots,n}(x_j)\right)/\text{MAD} \qquad (6)$$

which are more useful; in the contaminated example (2), the robust scores are

$$-0.22, \quad 1.35, \quad -0.67, \quad 1277.5, \quad 0.0$$

in which the outlier greatly exceeds the 2.5 cutoff.

Also Tukey's boxplot is often used to pinpoint possible outliers. In this plot, a box is drawn from the first quartile $Q_1 = x_{\lfloor n/4 \rfloor}$ to the third quartile $Q_3 = x_{\lceil 3n/4 \rceil}$ of the data. Points outside the interval $[Q_1 - 1.5\ \text{IQR}, Q_3 + 1.5\ \text{IQR}]$, called the fence, are traditionally marked as outliers. Note that the boxplot assumes symmetry because we add the same amount to Q_3 as what we subtract from Q_1. At asymmetric distributions, the usual boxplot typically flags many regular data points as outlying. The skewness-adjusted boxplot corrects for this by using a robust measure of skewness in determining the fence.[10]

MULTIVARIATE LOCATION AND COVARIANCE ESTIMATION

From now on, we assume that the data are p-dimensional and are stored in an $n \times p$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ the ith observation. Classical measures of location and scatter are given by the empirical mean $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ and the empirical covariance matrix $\boldsymbol{S}_x = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T/(n-1)$. As in the univariate case, both classical estimators have a breakdown value of 0%, that is, a small fraction of outliers can completely ruin them.

Robust estimates of location and scatter can be obtained by the minimum covariance determinant (MCD) method of Rousseeuw.[11,12] The MCD looks for those h observations in the data set (where the number h is given by the user) whose classical covariance matrix has the lowest possible determinant. The MCD estimate of location $\hat{\boldsymbol{\mu}}_0$ is then the average of these h points, whereas the MCD estimate of scatter $\hat{\boldsymbol{\Sigma}}_0$ is their covariance matrix, multiplied by a consistency factor. We can then give each $\boldsymbol{x}_i$ some weight w_i, for instance, by putting $w_i = 1$ if the initial robust distance

$$\begin{aligned} \mathrm{RD}_i(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0) &= \sqrt{(x_i - \hat{\boldsymbol{\mu}}_0)^T \hat{\boldsymbol{\Sigma}}_0^{-1} (x_i - \hat{\boldsymbol{\mu}}_0)} \\ &\leq \sqrt{\chi^2_{p,0.975}} \end{aligned} \tag{7}$$

and $w_i = 0$ otherwise. The weighted mean $\hat{\boldsymbol{\mu}}_w = (\sum_{i=1}^{n} w_i \boldsymbol{x}_i)/(\sum_{i=1}^{n} w_i)$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_w = (\sum_{i=1}^{n} w_i (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_w)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_w)^T)/(\sum_{i=1}^{n} w_i - 1)$ then have a better finite-sample efficiency. The final robust distances $\mathrm{RD}(\boldsymbol{x}_i)$ are obtained by inserting $\hat{\boldsymbol{\mu}}_w$ and $\hat{\boldsymbol{\Sigma}}_w$ into (7).

The MCD estimator, as well as its weighted version, has a bounded-influence function and breakdown value $(n-h+1)/n$, hence the number h determines the robustness of the estimator. The MCD has its highest possible breakdown value when $h = \lfloor (n+p+1)/2 \rfloor$. When a large proportion of contamination is expected, h should thus be chosen close to $0.5n$. Otherwise an intermediate value for h, such as $0.75n$, is recommended to obtain a higher finite-sample efficiency. We refer to Ref 13 for an overview of the MCD estimator and its properties.

The computation of the MCD estimator is nontrivial and naively requires an exhaustive investigation of all h-subsets out of n. Rousseeuw and Van Driessen[14] constructed a much faster algorithm called FAST-MCD. It starts by randomly drawing many $p+1$ observations from the data set. On the basis of these subsets, h-subsets are constructed by so-called C-steps (see Ref 14 for details).

Many other robust estimators of location and scatter have been presented in the literature. The first such estimator was proposed by Stahel[15] and Donoho[16] (see also Ref 17). They defined the so-called Stahel–Donoho outlyingness of a data point $\boldsymbol{x}_i$ as

$$\mathrm{outl}(\boldsymbol{x}_i) = \max_{\boldsymbol{d}} \frac{\left|\boldsymbol{x}_i^T \boldsymbol{d} - \mathrm{median}_{j=1,\ldots,n}(\boldsymbol{x}_j^T \boldsymbol{d})\right|}{\mathrm{MAD}_{j=1,\ldots,n}(\boldsymbol{x}_j^T \boldsymbol{d})} \tag{8}$$

where the maximum is over all directions (i.e., all p-dimensional unit length vectors $\boldsymbol{d}$), and $\boldsymbol{x}_j^T \boldsymbol{d}$ is the projection of $\boldsymbol{x}_j$ on the direction $\boldsymbol{d}$. Next, they gave each observation a weight w_i based on $\mathrm{outl}(\boldsymbol{x}_i)$, and computed the resulting weighted mean and covariance matrix.

Multivariate M-estimators[18] have a relatively low breakdown value due to possible implosion of the estimated scatter matrix. More recently, robust estimators of multivariate location and scatter include S-estimators,[2,19] MM-estimators,[20] and the orthogonalized Gnanadesikan-Kettenring (OGK) estimator.[21]

LINEAR REGRESSION

The multiple linear regression model assumes that in addition to the p independent x-variables also a response variable y is measured, which can be explained by a linear combination of the x variables. More precisely, the model says that for all observations $(\boldsymbol{x}_i, y_i)$ it holds that

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \\ &= \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i + \epsilon_i \quad i = 1, \ldots, n \end{aligned} \tag{9}$$

where the errors ϵ_i are assumed to be independent and identically distributed with zero mean and constant variance σ^2. Applying a regression estimator to the data yields $p+1$ regression coefficients, combined as $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$. The residual r_i of case i is defined as the difference between the observed response y_i and its estimated value $\hat{y}_i$.

The classical least squares (LS) method to estimate $\boldsymbol{\theta}$ minimizes the sum of the squared residuals. It is popular because it allows to compute the regression estimates explicitly, and it is optimal if the errors are normally distributed. However, LS is extremely sensitive to regression outliers, that is, observations that do not obey the linear pattern formed by the majority of the data.

The *least trimmed squares estimator* (LTS) proposed by Rousseeuw[11] is given by

$$\text{minimize} \sum_{i=1}^{h} (r^2)_{(i)} \tag{10}$$

where $(r^2)_{(1)} \leq (r^2)_{(2)} \leq \ldots \leq (r^2)_{(n)}$ are the ordered squared residuals. (They are first squared and then ordered.) The value h plays the same role as in the MCD estimator. For $h \approx n/2$, we find a breakdown value of 50%, whereas for larger h, we obtain roughly $(n-h)/n$. A fast algorithm for the LTS estimator (FAST-LTS) has been developed.[22]

The scale of the errors σ can be estimated by $\hat{\sigma}^2_{\text{LTS}} = c^2_{h,n} \sum_{i=1}^{h} (r^2)_{(i)}/h$, where r_i are the residuals from the LTS fit and $c_{h,n}$ is a constant that makes $\hat{\sigma}$ unbiased at Gaussian error distributions. We can then identify regression outliers by their standardized LTS residuals $r_i/\hat{\sigma}_{\text{LTS}}$. We can also use the standardized LTS residuals to assign a weight to every observation. The weighted LS estimator with these LTS weights inherits the nice robustness properties of LTS, but is more efficient and yields all the usual inferential output, such as t-statistics, F-statistics, an R^2 statistic, and the corresponding p-values.

The outlier map[23] plots the standardized LTS residuals versus robust distances (7) based on (for instance) the MCD estimator, which is applied to the x-variables only. This allows to distinguish vertical outliers (with small robust distances and outlying residuals), good leverage points (with outlying robust distances but small residuals), and bad leverage points (with outlying robust distances and outlying residuals).

The earliest theory of robust regression was based on M-estimators,[24] R-estimators,[25] and L-estimators.[26] The breakdown value of all these methods is 0% because of their vulnerability to bad leverage points. Generalized M-estimators (GM-estimators)[7] were the first to attain a positive breakdown value, which unfortunately still went down to zero for increasing p.

The low finite-sample efficiency of LTS can be improved by replacing its objective function by a more efficient scale estimator applied to the residuals r_i. This approach has led to the introduction of regression S-estimators[27] and MM-estimators.[28]

FURTHER DIRECTIONS

Principal Component Analysis

PCA is a very popular dimension-reduction method. It tries to explain the covariance structure of the data by a small number of components. These components are linear combinations of the original variables and often allow for an interpretation and a better understanding of the different sources of variation. PCA is often the first step of the data analysis, followed by other multivariate techniques.

In the classical approach, the first principal component corresponds to the direction in which the projected observations have the largest variance. The second component is then orthogonal to the first and again maximizes the variance of the data points projected on it. Continuing in this way produces all the principal components, which correspond to the eigenvectors of the empirical covariance matrix. Unfortunately, both the classical variance (which is being maximized) and the classical covariance matrix (which is being decomposed) are very sensitive to anomalous observations. Consequently, the first components from classical PCA are often attracted toward outlying points and may not capture the variation of the regular observations.

A first group of robust PCA methods is obtained by replacing the classical covariance matrix by a robust covariance estimator such as the weighted MCD estimator or MM-estimators.[29,30] Unfortunately, the use of these covariance estimators is limited to small-to-moderate dimensions because they are not defined when p is larger than n.

A second approach to robust PCA uses Projection Pursuit techniques. These methods maximize a robust measure of spread to obtain consecutive directions on which the data points are projected, see Refs 31–34.

The ROBPCA[35] approach is a hybrid, which combines ideas of projection pursuit and robust covariance estimation. The projection pursuit part is used for the initial dimension reduction. Some ideas based on the MCD estimator are then applied to this lower-dimensional data space.

For outlier detection, a PCA outlier map[35] can be constructed, similar to the regression outlier map. It plots for every observation its (Euclidean) *orthogonal distance* to the PCA-subspace, against its *score distance*, which measures the robust distance of its projection to the center of all the projected observations. Doing so, four types of observations are visualized. *Regular observations* have a small orthogonal and a small score distance. Observations with a large score distance but a small orthogonal distance are called good leverage points. *Orthogonal outliers* have a large orthogonal distance but a small score distance. Bad leverage points have a large orthogonal distance and a large score distance. They lie far outside the space spanned by the robust principal

components, and after projection far from the regular data points. Typically, they have a large influence on classical PCA, as the eigenvectors will be tilted toward them.

Other proposals for robust PCA include spherical PCA,[36] which first projects the data onto a sphere with a robust center, and then applies PCA to these projected data. For a review of robust versions of principal component regression and partial Least Square, see Ref 1.

Classification

The goal of classification, also known as discriminant analysis or supervised learning, is to obtain rules that describe the separation between known groups of p-dimensional observations. This allows to classify new observations into one of the groups. We denote the number of groups by l and assume that we can describe each population π_j by its density f_j. We write p_j for the membership probability, that is, the probability for any observation to come from π_j.

For low-dimensional data, a popular classification rule results from maximizing the Bayes posterior probability. At gaussian distributions, this leads to the maximization of the quadratic discriminant scores $d_j^Q(\boldsymbol{x})$ given by

$$d_j^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j) + \ln(p_j). \quad (11)$$

When all the covariance matrices are assumed to be equal, these scores can be simplified to

$$d_j^L(\boldsymbol{x}) = \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \ln(p_j) \quad (12)$$

where $\boldsymbol{\Sigma}$ is the common covariance matrix. This leads to linear classification boundaries. Robust classification rules can be obtained by replacing the classical covariance matrices by robust alternatives such as the MCD estimator and S-estimators, as in Refs 37–40.

When the data are high dimensional, this approach cannot be applied anymore because the robust covariance estimators become uncomputable. This problem can be solved by first applying PCA to the entire set of observations. Alternatively, one can also apply a PCA method to each group separately. This is the idea behind the soft independent modeling of class analogy (SIMCA) method.[41] A robustification of SIMCA is obtained by first applying robust PCA to each group, and then constructing a classification rule for new observations based on their orthogonal distance to each subspace and their score distance within each subspace.[42]

When linear models are not appropriate, support vector machines (SVM) are powerful tools to handle nonlinear structures.[43] An SVM with an unbounded kernel (such as a linear kernel) is not robust and suffers the same problems as traditional linear classifiers. But when a bounded kernel is used, the resulting nonlinear SVM classification handles outliers quite well.

Clustering

Cluster analysis is an important technique when handling large data sets. It searches for homogeneous groups in the data, which afterward may be analyzed separately. Nonhierarchical cluster methods search for the best clustering in k groups for a user-specified k.

For spherical clusters, the most popular method is k-means, which minimizes the sum of the squared Euclidean distances of the observations to the mean of their group.[44] This method is not robust as it uses group averages. To overcome this problem, one of the first robust proposals was the Partitioning around Medoids method.[45] It searches for k observations (medoids) such that the sum of the unsquared distances of the observations to the medoid of their group is minimized.

Later on, the more robust trimmed k-means method has been proposed,[46] inspired by the trimming idea of the MCD and the LTS. It searches for the h-subset (with h as in the definition of MCD and LTS) such that the sum of the squared distances of the observations to the mean of their group is minimized. Consequently, not all observations need to be classified, as $n-h$ cases can be left unassigned. To perform the trimmed k-means clustering, an iterative algorithm has been developed, using C-steps that are similar to those in the FAST-MCD and FAST-LTS algorithms.[47] For nonspherical clusters, constrained maximum likelihood approaches[48,49] were developed.

Beyond Outlier Detection

There are other aspects to robust statistics apart from outlier detection. For instance, robust estimation can be used in automated settings such as computer vision.[50,51] Another aspect is statistical inference such as the construction of robust hypothesis tests, p-values, confidence intervals, and model selection (e.g., variable selection in regression). This aspect is studied in Refs 3 and 7, which also cite earlier work.

SOFTWARE AVAILABILITY

Matlab functions for many of the procedures mentioned in this article are part of the LIBRA toolbox,[52,53] which can be downloaded from `wis.kuleuven.be/stat/robust`.

The MCD and LTS estimators are also built into S-PLUS as well as SAS (version 11 or higher) and SAS/IML (version 7 or higher). The free software R provides many robust procedures in the packages `robustbase` (e.g., huberM, Qn, adjbox, covMcd, covOGK, ltsReg, and lmrob) and `rrcov` (many robust covariance estimators, linear and quadratic discriminant analysis, and several robust PCA methods). Robust clustering can be performed with the `cluster` and `tclust` packages.

REFERENCES

1. Hubert M, Rousseeuw PJ, Van Aelst S, High breakdown robust multivariate methods. *Statist Sci* 2008, 23:92–119.
2. Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. New York: Wiley Interscience, 1987.
3. Maronna RA, Martin DR, Yohai VJ. *Robust Statistics: Theory and Methods*. New York: Wiley, 2006.
4. Hampel FR. A general qualitative definition of robustness. *Ann Math Stat* 1971, 42:1887–1896.
5. Donoho DL, Huber PJ. The notion of breakdown point. In: Bickel P, Doksum K, and Hodges JL, eds., *A Festschrift for Erich Lehmann* Belmont, MA: Wadsworth; 1983, 157–184.
6. Hubert M, Debruyne M. Breakdown Value. *Wiley Interdiscipl Rev: Comput Stat* 2009, 1:296–302.
7. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley; 1986.
8. Rousseeuw PJ, Croux C. Alternatives to the median absolute deviation. *J Amn Statist Assoc* 1993, 88:1273–1283.
9. Huber PJ. Robust estimation of a location parameter. *Ann Math Stat* 1964, 35:73–101.
10. Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. *Comput Stat Data Anal* 2008, 52:5186–5201.
11. Rousseeuw PJ. Least median of squares regression. *J Amn Statist Assoc* 1984, 79:871–880.
12. Rousseeuw PJ. Multivariate estimation with high breakdown point. In: Grossmann W, Pflug G., Vincze I, and wertz W, eds. *Mathematical Statistics and Applications*. Vol.B. Dordrecht, The Netherlands: Reidel Publishing Company; 1985, 283–297.
13. Hubert M, Debruyne M. Minimum covariance determinant. *Wiley Interdiscipl Rev: Comput Stat* 2010, 2:36–43.
14. Rousseeuw PJ, Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999, 41:212–223.
15. Stahel WA. Robuste Schätzungen: Infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen, Ph.D. thesis, ETH Zürich, 1981.
16. Donoho DL, Gasko M. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann Statist* 1992, 20:1803–1827.
17. Maronna RA, Yohai VJ. The behavior of the Stahel-Donoho robust multivariate estimator. *J Amn Statist Assoc* 1995, 90:330–341.
18. Maronna RA. Robust *M*-estimators of multivariate location and scatter. *Ann Statist* 1976, 4:51–67.
19. Davies L. Asymptotic behavior of *S*-estimators of multivariate location parameters and dispersion matrices. *Ann Statist* 1987, 15:1269–1292.
20. Tatsuoka KS, Tyler DE. On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *Ann Statist* 2000, 28:1219–1243.
21. Maronna RA, Zamar RH. Robust estimates of location and dispersion for high-dimensional data sets. *Technometrics* 2002, 44:307–317.
22. Rousseeuw PJ, Van Driessen K. Computing LTS regression for large data sets. *Data Min Knowledge Discovery* 2006, 12:29–45.
23. Rousseeuw PJ, van Zomeren BC. Unmasking multivariate outliers and leverage points. *J Amn Statist Assoc* 1990, 85:633–651.
24. Huber PJ. *Robust Statistics*. New York: Wiley; 1981.
25. Jurecková J. Nonparametric estimate of regression coefficients. *Ann Math Stat* 1971, 42:1328–1338.
26. Koenker R, Portnoy S. *L*-estimation for linear models. *J Amn Statist Assoc* 1987, 82:851–857.
27. Rousseeuw PJ, Yohai VJ. Robust regression by means of *S*-estimators, In: Robust and nonlinear time series analysis, Franke J, Härdle W, and Martin RD eds. *Lecture Notes in Statistics* No. 26, Springer-Verlag; 1984, 256–272.
28. Yohai VJ. High breakdown point and high efficiency robust estimates for regression. *Ann Statist* 1987, 15:642–656.

29. Croux C, Haesbroeck G. Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika* 2000, 87:603–618.
30. Salibian-Barrera M, Van Aelst S, Willems G. PCA based on multivariate MM-estimators with fast and robust bootstrap. *J Amn Statist Assoc* 2006, 101:1198–1211.
31. Li G, Chen Z. Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *J Amn Statist Assoc* 1985, 80:759–766.
32. Croux C, Ruiz-Gazen A. High breakdown estimators for principal components: the projection-pursuit approach revisited. *J Multivariate Anal* 2005, 95:206–226.
33. Hubert M, Rousseeuw PJ, Verboven S. A fast robust method for principal components with applications to chemometrics. *Chemom Intell Lab Syst* 2002, 60:101–111.
34. Croux C, Filzmoser P, Oliveira MR. Algorithms for projection-pursuit robust principal component analysis. *Chemom Intell Lab Syst* 2007, 87:218–225.
35. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: A new approach to robust principal components analysis. *Technometrics* 2005, 47:64–79.
36. Locantore N, Marron JS, Simpson DG, Tripoli N, Zhang JT, Cohen KL. Robust principal component analysis for functional data. *Test* 1999, 8:1–73.
37. Hawkins DM, McLachlan GJ. High-breakdown linear discriminant analysis. *J Amn Statist Assoc* 1997, 92:136–143.
38. He X, Fung WK. High breakdown estimation for multiple populations with applications to discriminant analysis. *J Multivariate Anal* 2000, 72:151–162.
39. Croux C, Dehon C. Robust linear discriminant analysis using *S*-estimators. *Can J Statist* 2001, 29:473–492.
40. Hubert M, Van Driessen K. Fast and robust discriminant analysis. *Comput Stat Data Anal* 2004, 45:301–320.
41. Wold S. Pattern recognition by means of disjoint principal components models. *Pattern Recognit* 1976, 8:127–139.
42. Vanden Branden K, Hubert M. Robust classification in high dimensions based on the SIMCA method. *Chemom Intell Lab Syst* 2005, 79:10–21.
43. Steinwart I, Christmann A. *Support Vector Machines*. New York: Springer; 2008.
44. MacQueen JB. Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on mathematical statistics and probability*, vol. 1, University of California Press, 1967, pp. 281–297.
45. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley; 1990.
46. Cuesta-Albertos JA, Gordaliza A, Matrán C. Trimmed k-means: An attempt to robustify quantizers. *Ann Statist* 1997, 25:553–576.
47. García-Escudero LA, Gordaliza A, Matrán C. Trimming tools in exploratory data analysis. *J Comput Graph Stat* 2003, 12:434–449.
48. Gallegos MT, Ritter G. A robust method for cluster analysis. *Ann Statist* 2005, 33:347–380.
49. García-Escudero LA, Gordaliza A, Matrán C, Mayo Iscar A. A general trimming approach to robust cluster analysis. *Ann Statist* 2008, 36:1324–1345.
50. Meer P, Mintz D, Rosenfeld A, Kim DY. Robust regression methods in computer vision: A review. *Int J Comput Vision* 1991, 6:59–70.
51. Stewart CV. MINPRAN: A new robust estimator for computer vision. *IEEE Trans Pattern Anal Machine Intell* 1995, 17:925–938.
52. Verboven S, Hubert M. LIBRA: a Matlab library for robust analysis. *Chemom Intell Lab Syst* 2005, 75:127–136.
53. Verboven S, Hubert M. Matlab library LIBRA. *Wiley Interdiscipl Rev: Comput Stat* 2010, in press.

