Advanced Review

Minimum volume ellipsoid

Stefan Van Aelst[1]* and Peter Rousseeuw[2]

The minimum volume ellipsoid (MVE) estimator is based on the smallest volume ellipsoid that covers h of the n observations. It is an affine equivariant, high-breakdown robust estimator of multivariate location and scatter. The MVE can be computed by a resampling algorithm. Its low bias makes the MVE very useful for outlier detection in multivariate data, often through the use of MVE-based robust distances.

We review the basic MVE definition as well as some useful extensions such as the one-step reweighted MVE. We discuss the main properties of the MVE including its breakdown value, affine equivariance, and efficiency. We discuss the basic resampling algorithm to calculate the MVE and illustrate its use on two examples. An overview of applications is given, as well as some related classes of robust estimators of multivariate location and scatter. *WIREs Comp Stat* 2009 1 71–82

INTRODUCTION

The minimum volume ellipsoid (MVE), introduced by Rousseeuw,[1,2] was the first high-breakdown robust estimator of multivariate location and scatter that has come to be regularly used in practice. The MVE became popular thanks to its high resistance to outliers, which makes it a reliable tool for outlier detection, and the widely available, user-friendly implementations of its computational algorithm. We first review the definition of the MVE and illustrate its use on two real data examples. We then give an overview of some important properties of the MVE, which are affine equivariance, breakdown value, and efficiency. We discuss the standard resampling algorithm to calculate MVE estimates in practice and give references to alternative algorithms. We give an overview of applications of the MVE estimators of location and scatter, which often involve outlier detection in multivariate data. We also discuss some extensions of the MVE, and related estimators.

DEFINITION

We consider a multivariate dataset $X_n = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with n observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^t$; $i = 1, \ldots, n$ in p dimensions. Note that all vectors in this overview are considered to be column vectors. We focus on estimating the location and scatter of this multivariate dataset X_n. It is convenient to collect the observations of a dataset X_n in an $n \times p$ data matrix $\boldsymbol{X}$ where each row corresponds to an observation $\boldsymbol{x}_i$ of X_n.

As an example, we consider the pulp fiber data[3] which is available in the R package 'robustbase'. This dataset contains measurements of properties of pulp fibers and the paper made from them. The final aim is to investigate relations between pulp fiber properties and the resulting paper properties, see e.g., Refs 4,5 Here we focus on the pulp fiber properties. The dataset contains $n = 62$ measurements of the following four pulp fiber characteristics: arithmetic fiber length, long fiber fraction, fine fiber fraction, and zero span tensile strength. A standard approach to investigate whether this multivariate dataset forms a homogeneous group or contains aberrant points is to calculate the Mahalanobis distances of the observations, given by

$$\mathrm{MD}(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)^t \boldsymbol{S}_n^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)} \qquad i = 1, \ldots, n \tag{1}$$

where $\bar{\boldsymbol{x}}_n$ is the sample mean and $\boldsymbol{S}_n$ the sample covariance matrix of the data. It is well known that if the data follow a four-dimensional Gaussian distribution, then the squared Mahalanobis distances approximately follow a chi-square distribution with four degrees of freedom. Therefore, we compare the Mahalanobis distances to the square root of $\chi^2_{4,0.975}$,

*Correspondence to: stefan.vanaelst@ugent.be

[1]Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Ghent, Belgium.

[2]Department of Mathematics and Computer Science, University of Antwerp, Middelheimlaan 1, B-2020 Antwerp, Belgium.

DOI: 10.1002/wics.019

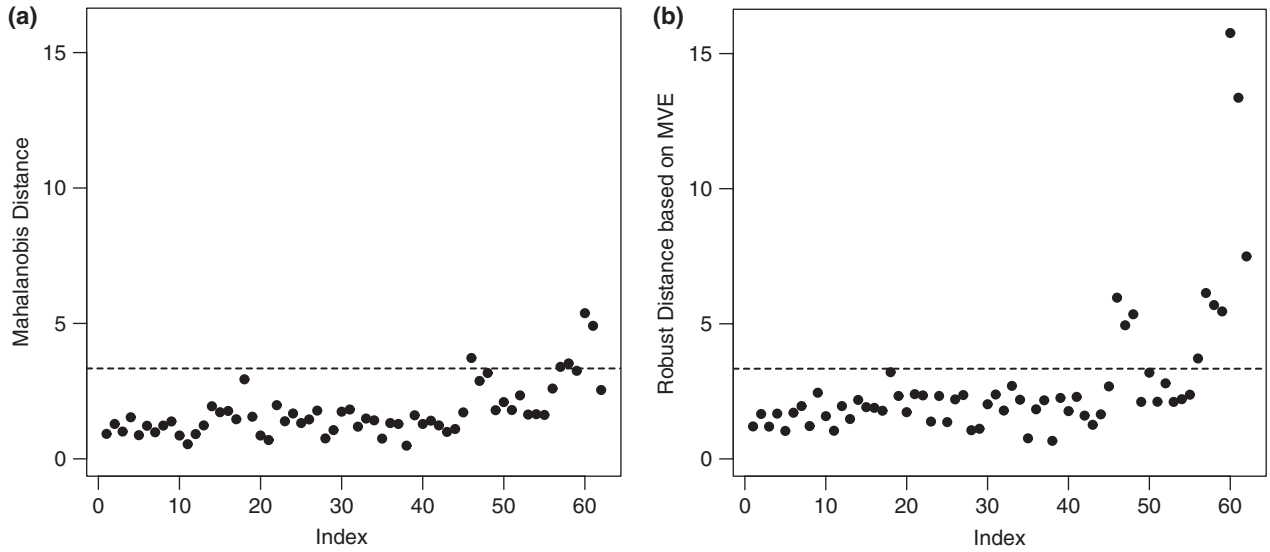

FIGURE 1 | Distances of the observations in the pulp fiber dataset based on the four pulp fiber properties: (a) Mahalanobis distances based on sample mean and sample covariance matrix; (b) Robust distances based on MVE estimates of location and scatter. The horizontal cutoff line in both panels is at $\sqrt{\chi^2_{4,0.975}} = 3.34$.

which is the 97.5% quantile of the chi-square distribution with four degrees of freedom. This cutoff value is represented by the horizontal line in Figure 1a.

If the data indeed form a homogeneous cloud, then we do not expect to find any Mahalanobis distances far above the horizontal cutoff line. Figure 1a suggests that the data are fairly homogeneous with at most two observations that deviate a little from the data cloud formed by the other observations. However, it is well known that the sample mean and sample covariance matrix can be heavily influenced by outliers in a multivariate dataset (see e.g., Refs 6–8). As a result, even if there are outliers in the dataset, they can affect the sample mean and sample covariance matrix in such a way that these outliers get small Mahalanobis distances $\mathrm{MD}(x_i)$. Hence, outliers can remain undetected in Figure 1a. This phenomenon is called the *masking effect* (see e.g., Refs 9,10). Because the dimension of the dataset in this example is fairly low, we can examine the dataset further by investigating the pairwise scatterplots in Figure 2. The ellipses shown in these scatterplots are the projections of the tolerance ellipsoid

$$\mathcal{E}(\bar{x}_n, S_n, 0.975) = \left\{x;\ \mathrm{MD}(x) \le \sqrt{\chi^2_{4,0.975}}\right\} \qquad (2)$$

on the respective coordinate planes. Hence, in a homogeneous point cloud all observations should lie within or close to the boundaries of these ellipses. From the scatterplots in Figure 2 we immediately see that although no observations lie far from the ellipses, the data do not form a homogeneous cloud. Several observations deviate from the shape of the majority of the points. These outliers have inflated the sample covariance matrix and also affected its shape, which leads to the masking effect when investigating Mahalanobis distances. Hence, to reliably estimate the center and scatter of this dataset, robust estimates of location and scatter are needed, such as the MVE estimator.

The MVE estimator of multivariate location and scatter of a dataset X_n is defined as the center and covariance structure of the ellipsoid with minimal volume that covers at least h points of X_n, where h can be chosen between $[n/2] + 1$ and n. Note that for any $x \in \mathbb{R}$, the value $[x]$ denotes the largest integer smaller than or equal to x. More formally, the MVE estimator is defined as follows.

Definition 1 *The MVE location estimator $\boldsymbol{t}_n$ and scatter estimator $\mathbf{C}_n$ minimize the determinant of C subject to the condition*

$$\#\{i;\ (\boldsymbol{x}_i - \boldsymbol{t})^t \mathbf{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{t}) \le c^2\} \ge h, \qquad (3)$$

where the minimization is over all $\boldsymbol{t} \in \mathbb{R}^p$ and $\mathbf{C} \in PDS(p)$, the class of positive definite symmetric matrices of size p.

The value c is a fixed chosen constant that determines the magnitude of $\mathbf{C}_n$. Usually, c is chosen such that $\mathbf{C}_n$ is a consistent estimator of the covariance matrix for data coming from a multivariate normal

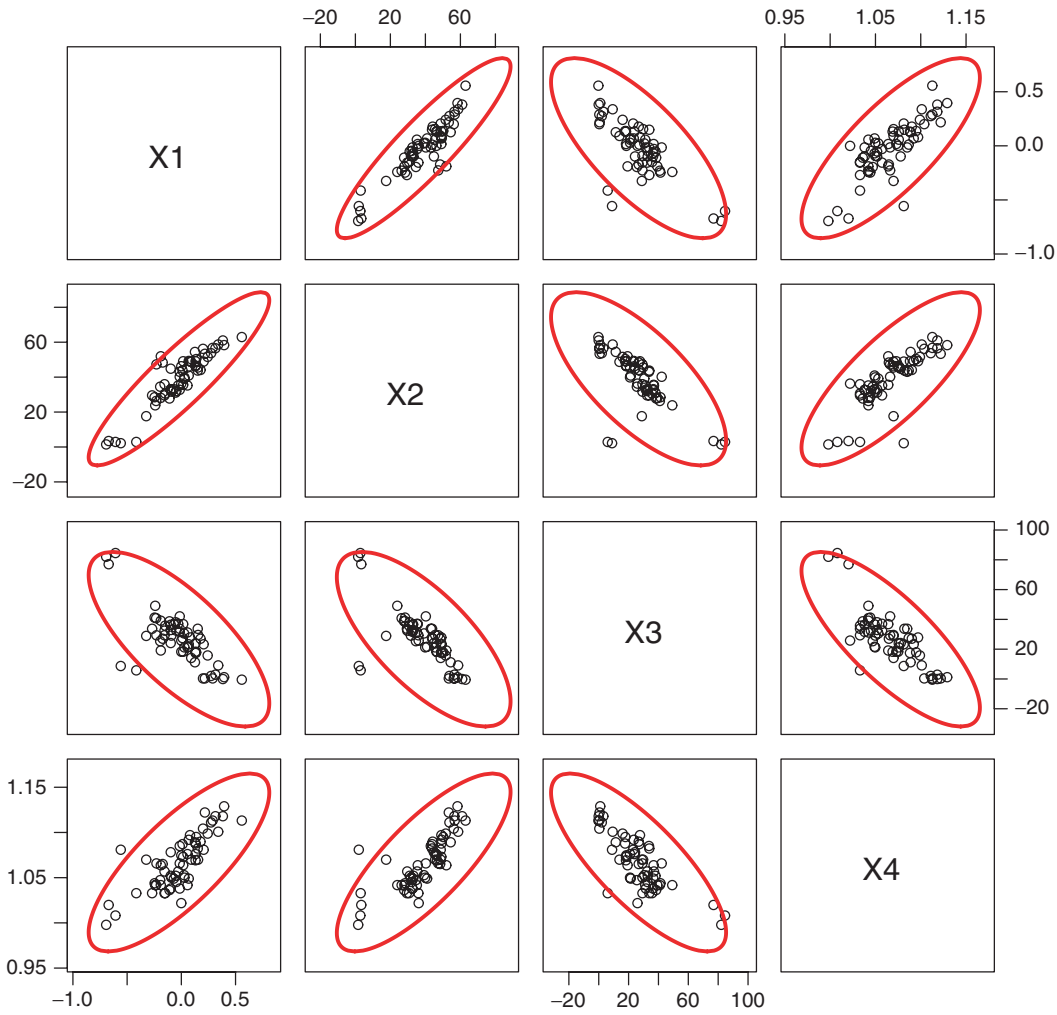

FIGURE 2 | Pairwise scatterplots of the four pulp fiber variables. The ellipses represent the 97.5% tolerance ellipsoid for the observations, based on the sample mean and sample covariance matrix.

distribution, i.e., $c = \sqrt{\chi^2_{p,\alpha}}$ where $\alpha = h/n$. From its definition it is clear that the MVE estimates the center and scatter of the h most concentrated observations in the dataset. The value of h can be chosen by the user and determines the robustness of the resulting MVE estimates. A standard choice is $h = [(n + p + 1)/2]$ because it yields the maximal breakdown value as will be explained in the next section, where we give an overview of the properties of the MVE estimator. The examples in this article all use this standard choice of h.

Let us return to the example. Figure 1b shows the robust distances of the observations based on the MVE estimates of location and scatter,[7] given by

$$\mathrm{RD}(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{t}_n)^t \boldsymbol{C}_n^{-1} (\boldsymbol{x}_i - \boldsymbol{t}_n)} \qquad i = 1, \ldots, n. \tag{4}$$

In Figure 1b we immediately see that the dataset contains two far outliers and seven less extreme outliers. Figure 3 shows the pairwise scatterplots with the MVE-based tolerance ellipsoid

$$\mathcal{E}(\boldsymbol{t}_n, \boldsymbol{C}_n, 0.975) = \left\{\boldsymbol{x};\ \mathrm{RD}(\boldsymbol{x}) \le \sqrt{\chi^2_{4,0.975}}\right\}. \tag{5}$$

These scatterplots illustrate that the MVE estimates of location and scatter indeed reflect the center and shape of the majority of the data.

As a second example, we consider an engineering problem that was first analyzed in Ref 11 Philips Mecoma (The Netherlands) produced diaphragm parts for TV sets. These are thin metal plates, molded by a press. When starting a new production line, $p = 9$ characteristics were measured for $n = 677$ parts. The aim was to gain insight in the production process and to find out whether abnormalities have

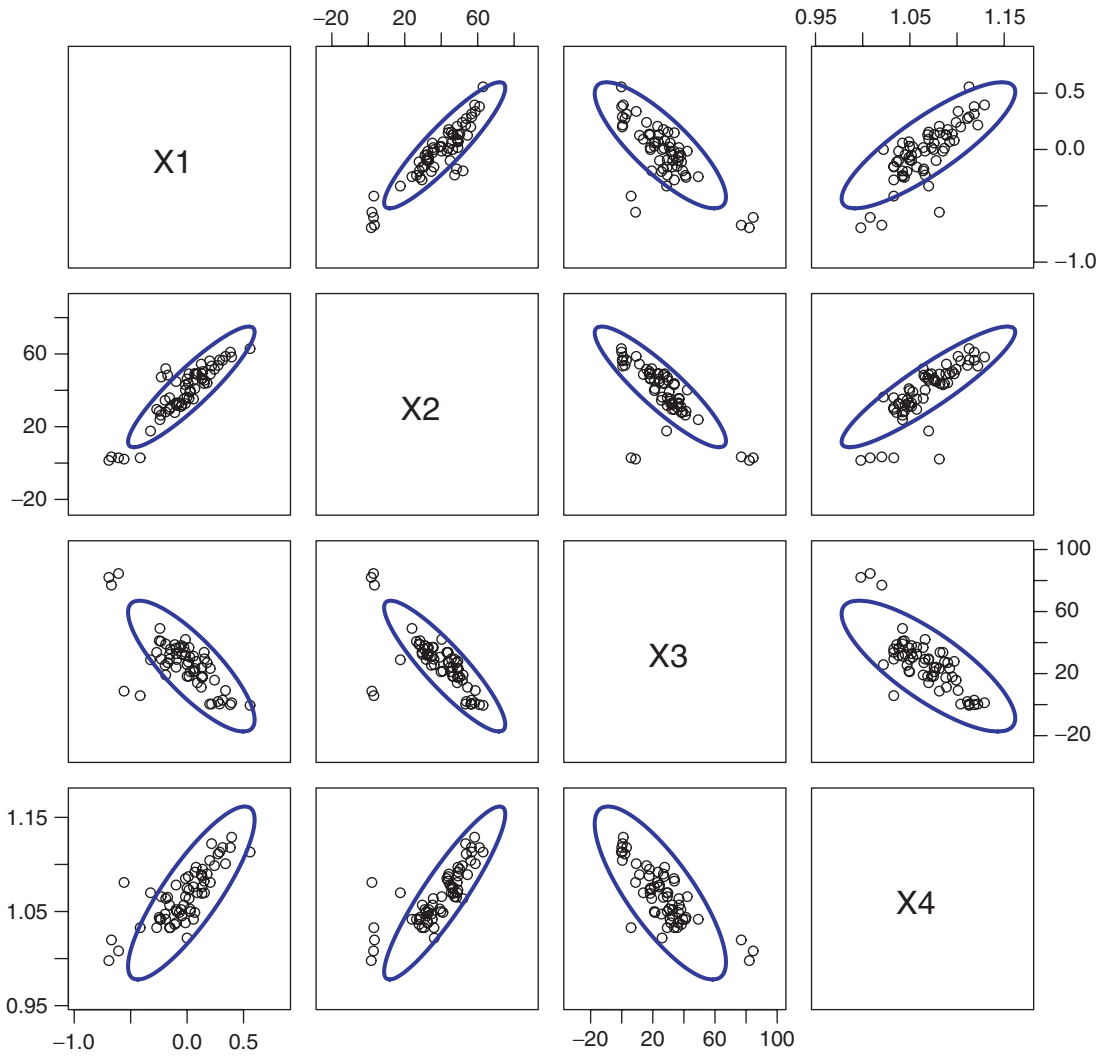

FIGURE 3 | Pairwise scatterplots of the four pulp fiber variables. The ellipses represent the 97.5% tolerance ellipsoid for the observations, based on the MVE estimates of location and scatter.

occurred and why. We can again calculate distances of the observations and check for unexpectedly large distances that indicate anomalies in the data. When classical Mahalanobis distances are used, there is no indication of severe anomalies (see Figure 1 in Ref 11), but as before this may be the consequence of the masking effect. Therefore, we now examine the MVE-based robust distances shown in Figure 4. This figure gives a much better insight into the evolution of the production process. The robust distances immediately reveal that the production line was unstable in the beginning (first 100 observations) and reveals a strongly deviating group of outliers, ranging from index 491 to index 565. Both phenomena were investigated and interpreted by engineers at Philips.

PROPERTIES

Affine equivariance

A natural property of estimators in the multivariate location and scatter model is affine equivariance, which means that the estimators behave properly under affine transformations of the data. That is, the estimators $\boldsymbol{T}$ and $\boldsymbol{C}$ of multivariate location and scatter are affine equivariant iff for any data matrix $\boldsymbol{X}$

$$\begin{aligned} \boldsymbol{T}(\boldsymbol{X}\boldsymbol{A} + \mathbf{1}_n\boldsymbol{v}^t) &= \boldsymbol{A}^t\boldsymbol{T}(\boldsymbol{X}) + \boldsymbol{v} \\ \boldsymbol{C}(\boldsymbol{X}\boldsymbol{A} + \mathbf{1}_n\boldsymbol{v}^t) &= \boldsymbol{A}^t\boldsymbol{C}(\boldsymbol{X})\boldsymbol{A} \end{aligned} \tag{6}$$

for all nonsingular $p \times p$ matrices $\boldsymbol{A}$ and $\boldsymbol{v} \in \mathbb{R}^p$. The vector $\mathbf{1}_n = (1, 1, \ldots, 1)^t \in \mathbb{R}^n$. Affine equivariance of the estimators is important because it makes the analysis independent of the measurement scale of

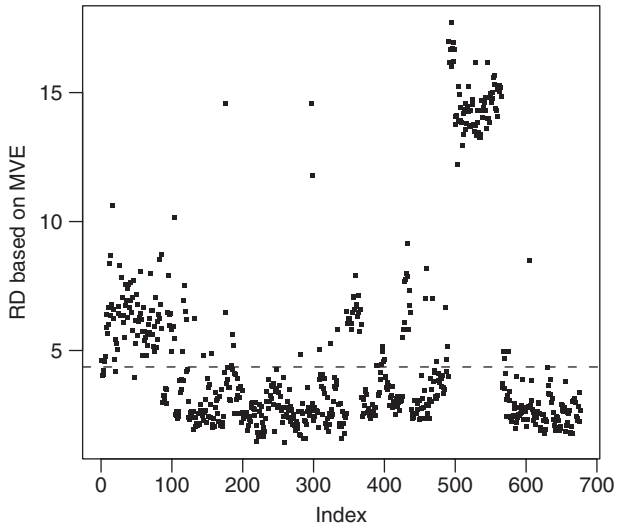

FIGURE 4 | MVE-based robust distances of the observations in the Philips dataset. The horizontal cutoff line is at $\sqrt{\chi^2_{9,0.975}} = 4.36$.

the variables as well as translations or rotations of the data.

The MVE estimates $\boldsymbol{t}_n$ and $\boldsymbol{C}_n$ of multivariate location and scatter are affine equivariant.[2,7] This follows from the fact that the nonsingular affine transformation $\boldsymbol{x} \to \boldsymbol{A}^t\boldsymbol{x} + \boldsymbol{v}$ transforms an ellipsoid with center $\boldsymbol{m}$ and scatter matrix $\boldsymbol{S}$ containing at least h points of $\boldsymbol{X}$ into an ellipsoid with center $\boldsymbol{A}^t\boldsymbol{m} + \boldsymbol{v}$ and scatter matrix $\boldsymbol{A}^t\boldsymbol{S}\boldsymbol{A}$ which contains at least h points of $\boldsymbol{X}\boldsymbol{A} + \mathbf{1}_n\boldsymbol{v}^t$. The volume of the transformed ellipsoid equals $\det(\boldsymbol{A}^t\boldsymbol{S}\boldsymbol{A})^{1/2} = |\det(\boldsymbol{A})|\det(\boldsymbol{S})^{1/2}$. Since $|\det(\boldsymbol{A})|$ is a constant, the MVE estimates of $\boldsymbol{X}\boldsymbol{A} + \mathbf{1}_n\boldsymbol{v}^t$ are indeed given by $\boldsymbol{A}^t\boldsymbol{t}_n + \boldsymbol{v}$ and $\boldsymbol{A}^t\boldsymbol{C}_n\boldsymbol{A}$ where $\boldsymbol{t}_n$, $\boldsymbol{C}_n$ are the MVE estimates of $\boldsymbol{X}$.

Breakdown value

A useful measure of the global robustness of an estimator is its breakdown value. Intuitively, the breakdown value is the smallest percentage of contamination that can have an arbitrarily large effect on the estimator (see e.g., Refs 6,7). Results for the breakdown value of the MVE estimators of location and scatter have been given in Refs 2,12,13

We discuss the finite-sample replacement breakdown value, introduced in Ref 14 For a given dataset X_n, consider all possible contaminated datasets $\tilde{X}_n$ obtained by replacing *any* m of the original observations by *arbitrary* points. Then the finite-sample breakdown value $\varepsilon^*_n(\boldsymbol{T}, X_n)$ of a location estimator $\boldsymbol{T}$ at the dataset X_n is the smallest fraction m/n of outliers that can carry the estimate over all bounds:

$$\varepsilon^*_n(\boldsymbol{T}, X_n) := \min_m \left\{ \frac{m}{n}; \sup_{\tilde{X}_n} \|\boldsymbol{T}(\tilde{X}_n) - \boldsymbol{T}(X_n)\| = \infty \right\}. \tag{7}$$

Usually $\varepsilon^*_n(\boldsymbol{T}, X_n)$ varies only slightly between samples and with the sample size n, so that we can denote its limiting value (for $n \to \infty$) by $\varepsilon^*(\boldsymbol{T})$. Similarly, the breakdown value of a covariance matrix estimator $\boldsymbol{C}$ is defined as the smallest fraction of contamination that can either take the largest eigenvalue $\lambda_1(\boldsymbol{C})$ to infinity or the smallest eigenvalue $\lambda_p(\boldsymbol{C})$ to zero. For the MVE estimators we then have the following result.

Theorem 1 *Consider a dataset $X_n \subset \mathbb{R}^p$ that is in general position, which means that no $p+1$ points lie on a hyperplane. Then the MVE estimators $(\boldsymbol{t}_n, \boldsymbol{C}_n)$ of multivariate location and scatter have finite-sample breakdown value*

$$\begin{aligned} \varepsilon^*_n(\boldsymbol{t}_n, X_n) &= \varepsilon^*_n(\boldsymbol{C}_n, X_n) \\ &= \frac{\min(n-h+1, h-p)}{n}. \end{aligned} \tag{8}$$

It follows immediately that for $n \to \infty$ the breakdown value of the MVE estimators equals $\varepsilon^*(\boldsymbol{T}) = \varepsilon^*(\boldsymbol{C}) = \min(\alpha, 1-\alpha)$ where $\alpha = h/n$ as before. From Theorem 1 it can be shown that the MVE estimates have their highest breakdown value $\varepsilon^*_n(\boldsymbol{t}_n, X_n) = \varepsilon^*_n(\boldsymbol{C}_n, X_n) = [(n-p+1)/2]/n \approx 50\%$ when $h = [(n+p+1)/2]$ (see Ref 12). One can prove that this is the maximal breakdown value for all affine equivariant estimators of scatter[15] and location.[16]

Efficiency

Davies[13] has shown that the MVE estimators of location and scatter converge at rate $n^{-1/3}$ to a non-Gaussian distribution. This low rate of convergence implies that the asymptotic efficiency of the MVE estimators is 0%. Also the finite-sample efficiency of the MVE estimates is low (see e.g., Ref 7). Therefore, one usually computes the one-step reweighted MVE estimates,[17] given by

$$\begin{aligned} \boldsymbol{t}^1_n &= \left(\sum_{i=1}^n w_i \boldsymbol{x}_i\right) \Big/ \left(\sum_{i=1}^n w_i\right) \\ \boldsymbol{C}^1_n &= \left(\sum_{i=1}^n w_i (\boldsymbol{x}_i - \boldsymbol{t}^1_n)(\boldsymbol{x}_i - \boldsymbol{t}^1_n)^t\right) \Big/ \left(\sum_{i=1}^n w_i\right) \end{aligned} \tag{9}$$

with

$$w_i = \begin{cases} 1 & \text{if } \mathrm{RD}(\boldsymbol{x}_i) \leq \sqrt{\chi^2_{p,0.975}} \\ 0 & \text{otherwise} \end{cases}$$

where $\mathrm{RD}(\boldsymbol{x}_i)$ are the robust distances of the observations based on the initial MVE estimates of location and scatter as defined in (4). These one-step reweighted MVE estimates are a weighted mean and covariance where regular observations are given weight one, but outliers (according to the initial MVE solution) are given weight zero. The one-step reweighted MVE estimators have the same breakdown value as the initial MVE estimators[12] but a much better finite-sample efficiency (see e.g., Refs 7,17). Note that many software implementations (such as the implementation in R that we used for the examples in this review) report the one-step reweighted MVE estimates by default.

Note that it has been shown more recently that the one-step reweighted MVE estimates do not improve on the convergence rate (and thus the 0% asymptotic efficiency) of the initial MVE estimator.[18] Therefore, as an alternative, a one-step M-estimator can be calculated with the MVE estimates as initial solution,[2,19] which results in an estimator with the standard $n^{-1/2}$ convergence rate to a normal asymptotic distribution. Another alternative to increase the efficiency of the MVE while retaining its robustness properties has been proposed in Ref 20.

ALGORITHM

From Definition 1 it follows that calculating the exact MVE for a dataset X_n would require examining all $\binom{n}{h}$ ellipsoids containing h observations of X_n to find the ellipsoid with smallest volume. This number of ellipsoids is usually very large, hence solving this combinatorial problem is only feasible in practice for small datasets in low dimensions.[21,22] Therefore, one resorts to approximate algorithms. The standard MVE algorithm limits its search to ellipsoids determined by subsets consisting of $(p+1)$ observations of X_n. For each subset of size $(p+1)$, indexed by $J = \{i_1, \ldots, i_{p+1}\} \subset \{1, \ldots, n\}$, its sample mean and sample covariance matrix given by

$$\bar{\boldsymbol{x}}_J = \frac{1}{p+1}\sum_{j=1}^{p+1} \boldsymbol{x}_{i_j} \quad \text{and} \quad \boldsymbol{S}_J = \frac{1}{p}\sum_{j=1}^{p+1} (\boldsymbol{x}_{i_j} - \bar{\boldsymbol{x}}_J)(\boldsymbol{x}_{i_j} - \bar{\boldsymbol{x}}_J)^t \tag{10}$$

are calculated. The covariance matrix $\boldsymbol{S}_J$ is nonsingular iff the $(p+1)$-subset is in general position. If the $(p+1)$-subset is not in general position, then observations from X_n are added until a subset with nonsingular sample covariance matrix is obtained (or a singular subsample of size h is obtained, in which case the final MVE solution is singular). The ellipsoid determined by $\bar{\boldsymbol{x}}_J$ and $\boldsymbol{S}_J$ is then inflated or deflated until it contains exactly h points: the scaling factor is given by D_J^2/c^2 with $c = \sqrt{\chi^2_{p,\alpha}}$ as before and

$$D_J^2 = [(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_J)^t (\boldsymbol{S}_J)^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_J)]_{h:n}, \tag{11}$$

where $h:n$ indicates the hth smallest squared distance among the squared distances of the n observations in X_n. The resulting ellipsoid then satisfies condition (3) and its volume is proportional to

$$[\det((D_J^2/c^2)\boldsymbol{S}_J)]^{1/2} = (D_J/c)^p \det(\boldsymbol{S}_J)^{1/2}. \tag{12}$$

The algorithm then returns the solution with smallest objective function (12) among a large number of $(p+1)$-subsets.

It has been shown that this resampling algorithm keeps the affine equivariance property of the MVE estimator. Moreover, if all $\binom{n}{p+1}$ subsets of size $(p+1)$ are considered, then the solution of the algorithm has the same breakdown value as the exact MVE.[23] However, in practice the total number of $(p+1)$-subsets is infeasibly large and only a random collection is considered. Standard implementations of the MVE algorithm use $m = 3000$ random $(p+1)$-subsets by default, to keep the computation time reasonable.[24] However, modern computers can handle many more $(p+1)$-subsets in a short period of time. For example, for the Philips data ($n = 677$, $p = 9$) that we used as an example in this review, it takes less than 9 s to calculate the approximate MVE solution based on $m = 30000$ random $(p+1)$-subsamples when using the R implementation on a standard contemporary PC.

Croux and Haesbroeck[25] proposed a modification of the standard resampling algorithm for MVE by taking an average of the solutions corresponding to several 'near-optimal' $(p+1)$-subsets instead of considering only the solution corresponding to the best $(p+1)$-subset. They showed that their average solution maintains the breakdown value and has a better finite-sample efficiency.[25,26] The standard resampling algorithm can also be improved by using location adjustment as proposed in Ref 27. An alternative improvement of the standard resampling algorithm for MVE has been proposed in Ref 8 [pp. 198–199] by updating the center and scatter estimates corresponding to the best $(p+1)$-subset, using the h observations

within its minimum volume ellipsoid. Several alternative algorithms to calculate the MVE have been proposed.[17,28–35]

The resampling algorithm to calculate the MVE estimators of multivariate location and scatter has been implemented in several software packages. In standard S-PLUS the MVE is available as the function *cov.mve*. In R, this function is part of the library *MASS*. The improved resampling algorithm proposed in Ref 8 has been implemented in the R library *rrcov* as the function *CovMve*. The MVE is also available in SAS/IML as the call *MVE* (since Version 6.12). Finally, the MVE is still available as a stand-alone FORTRAN program that can be downloaded from the website *http://www.agoras.ua.ac.be/*.

APPLICATIONS

To reliably detect outliers in multivariate data, it is not only important that the estimators of location and scatter have a high breakdown value but also that the bias of the estimators caused by a fraction of contamination below the breakdown value should be as small as possible. The maximal possible asymptotic bias, called *maxbias*, of the MVE estimates caused by a fixed fraction of contamination has been investigated in Refs 36,37 in the one-dimensional case and in Refs 38,39 in the multivariate case. It turns out that the maxbias of the MVE estimators is generally low and compares favorably to many other high-breakdown estimators of multivariate location and scatter, such as the Stahel-Donoho estimator[40–42] and the minimum covariance determinant estimator.[1,2] The good bias behavior of the MVE makes the estimator suitable for outlier detection. For this purpose, the MVE estimate of scatter is often multiplied by a finite-sample correction factor such that the resulting robust distances are appropriately scaled when the observations come from a multivariate normal distribution (see Ref 43 for details). Therefore, cutoff values of the usual χ^2_p distribution can be used for the robust distances based on the MVE.

MVE-based robust distances are often used to detect leverage points in linear regression.[17] Leverage points are outliers in the explanatory variables of the regression model and have a high influence on the standard least squares regression (see e.g., Ref 7). Detecting leverage points by examining the MVE-based robust distances of the explanatory part of the observations was first proposed in the context of least median of squares (LMS) regression.[1] LMS, or its generalization the least quantile of squares (LQS) estimator,[24] is a regression analog of the MVE. Consider a dataset $Z_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ and the multiple regression model

$$y_i = \theta_1 x_{i1} + \cdots + \theta_p x_{ip} + \epsilon_i = \boldsymbol{x}_i^t \boldsymbol{\theta} + \epsilon_i; \quad i = 1, \ldots, n \tag{13}$$

where ϵ_i are the errors centered at zero. The residuals corresponding to a fit $\boldsymbol{\theta}$ are denoted by $r_i(\boldsymbol{\theta}) = y_i - \boldsymbol{x}_i^t \boldsymbol{\theta}$. The LQS looks for the fit $\boldsymbol{\theta}_n$ that minimizes the h smallest squared residual $r_i^2(\boldsymbol{\theta})_{h:n}$ where h is usually chosen between $[n/2] + 1$ and n. From this definition it is clear that the LQS is determined by the h observations in the dataset that lie most concentrated around a hyperplane. For the choice $h = [n/2] + 1$, the LQS minimizes the median of the squared residuals which leads to the LMS. For datasets in general position, the breakdown value of the LQS is given by

$$\varepsilon_n^*(\boldsymbol{\theta}_n, Z_n) = \frac{\min(n - h + 1, h - p + 1)}{n}. \tag{14}$$

It follows immediately that for $n \to \infty$ the breakdown value of the LQS becomes $\varepsilon^*(\boldsymbol{\theta}_{LQS}) = \min(\alpha, 1 - \alpha)$ where $\alpha = h/n$ as before. Moreover, the LQS reaches its maximal breakdown value $\varepsilon_n^*(\boldsymbol{\theta}_n, Z_n) = ([(n - p)/2] + 1)/n \approx 50\%$ when $h = [(n + p + 1)/2]$ (see Ref. 24 for details).

To detect regression outliers and leverage points simultaneously, a diagnostic plot was introduced[17] which divides the observations into four categories: *regular observations, vertical outliers, good leverage points,* and *bad leverage points*. A vertical outlier is an observation whose $\boldsymbol{x}_i$ is inlying but whose $(\boldsymbol{x}_i, y_i)$ does not fit the linear trend formed by the majority of the data. A leverage point is an observation with outlying $\boldsymbol{x}_i$. It is called a good leverage point if its $(\boldsymbol{x}_i, y_i)$ fits the linear trend formed by the majority of the data, and a bad leverage point when it does not. Applications of LMS with MVE-based detection of leverage points have been given in several areas such as chemometrics,[44] management,[45] and astronomy.[46]

MVE-based robust distances were used in Refs 47–50 in the context of one-step M-estimators with high breakdown value, in Refs 51,52 in the context of high-breakdown rank regression, and in Refs 53,54 for high-breakdown estimators in heteroscedastic regression models. The MVE can also be used for estimating location with dependent data.[55]

The MVE has also been used for outlier detection in many other multivariate analysis models such as principal component analysis,[56,57] discriminant analysis,[58,59] factor analysis,[60] multiplicative factor models,[61] image segmentation,[62] and multivariate control charts.[63–65] Some textbooks also recommend

using the MVE for robust multivariate data analysis (see e.g., Ref 66 [pp. 56–61]).

EXTENSIONS

The MVE can be seen as a special case within the class of S-estimators.[67] Location-scatter S-estimators[7,15,68] are defined as follows.

Definition 2 *The S-estimators of multivariate location and scatter are the solution* $(\boldsymbol{t}_n^0, \boldsymbol{C}_n^0)$ *which minimizes the determinant of* $\boldsymbol{C}$ *under the constraint*

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\sqrt{(\boldsymbol{x}_i-\boldsymbol{t})^t\boldsymbol{C}^{-1}(\boldsymbol{x}_i-\boldsymbol{t})}\right)\leq b \qquad (15)$$

over all $\boldsymbol{t}\in\mathbb{R}^p$ *and* $\boldsymbol{C}\in \mathrm{PDS}(p)$.

Setting $b = E_F[\rho_0(\|X\|)]$ ensures consistency at the model distribution F, which usually is taken to be multivariate normal. The choice of the discontinuous function $\rho_0 = 1 - I(x \in [0, c])$ and $b = (n-h)/n$ yields the MVE estimators. It can be shown that for suitable choices of continuously differentiable loss functions ρ_0, S-estimators have a high breakdown value and are asymptotically normal.[15,68] A standard choice for the loss function ρ_0 is Tukey's biweight ρ-function, given by

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if } |x| \leq c \\ \frac{c^2}{6} & \text{if } |x| \geq c. \end{cases} \qquad (16)$$

The constant c determines the breakdown value which is given by $\varepsilon^* = 6b/c^2$, so S-estimators can be tuned to have high breakdown value. S-estimators have a positive efficiency at the multivariate normal distribution, but there exists a trade-off between efficiency and breakdown value. The efficiency of high-breakdown S-estimators can still be quite low, especially in lower dimensions, which makes them less suitable for inference. Note that S-estimators of scatter can also be based on differences of the observations, which yields a higher efficiency.[69]

MM-estimators are an extension of S-estimators that have high efficiency at the multivariate normal distribution and at the same time a high breakdown value.[70,71] Location-scatter MM-estimators are defined as follows.

Definition 3 *let* $(\boldsymbol{t}_n^0, \boldsymbol{C}_n^0)$ *be multivariate S-estimators as given by Definition 2. Denote* $s_n := det(\boldsymbol{C}_n^0)^{1/(2p)}$. *Then the multivariate MM-estimators for location and shape* $(\boldsymbol{t}_n^1, \boldsymbol{V}_n^1)$ *minimize*

$$\frac{1}{n}\sum_{i=1}^{n}\rho_1\left([(\boldsymbol{x}_i-\boldsymbol{t})^t\boldsymbol{G}^{-1}(\boldsymbol{x}_i-\boldsymbol{t})]^{\frac{1}{2}}/s_n\right)$$

among all $\boldsymbol{t}\in\mathbb{R}^p$ *and* $\boldsymbol{G}\in PDS(p)$ *for which* $det(\boldsymbol{G}) = 1$. *The MM-estimator for the scatter matrix is* $\boldsymbol{C}_n^1 := s_n^2\boldsymbol{V}_n^1$.

MM-estimators are thus two-step estimators. In the first step, a robust high-breakdown estimator s_n of the scale of the distribution is obtained. This preliminary estimate of scale is then used to calculate M-estimators of location $\boldsymbol{t}_n^1$ and shape $\boldsymbol{V}_n^1$. It can be shown that the loss function ρ_0 used to calculate the initial S-estimator determines the breakdown value of the estimators $\boldsymbol{t}_n^1, \boldsymbol{V}_n^1$, and $\boldsymbol{C}_n^1$ while the loss function ρ_1 can be tuned to obtain a high efficiency, e.g., 95% efficiency for the location estimator $\boldsymbol{t}_n^1$ when the data come from a multivariate normal distribution (see Refs 70,71 for details). Related classes of multivariate location and scatter estimators that can attain high breakdown value and high efficiency at the same time are the CM-estimators[72] and τ-estimators.[73]

Note that although these highly efficient estimators also attain a high breakdown value, there is a robustness cost in terms of the maxbias of these estimators for fractions of contamination below the breakdown value.[74] The higher bias of these estimators makes them somewhat less suitable when the main goal is outlier detection. On the other hand, their high efficiency makes them more appropriate for inference purposes. Inference can be derived from the asymptotic normal distribution of the estimators, or the bootstrap approach can be used. However, note that a standard application of the bootstrap to robust estimators poses two problems. First, the high computation time of robust estimators causes practical limitations because recalculating robust estimates a large number of times becomes very time consuming. Second, the fraction of outliers varies among bootstrap samples. Therefore, the estimator may break down in some bootstrap samples even though the fraction of outliers in the original sample does not exceed the breakdown value of the estimator. To solve both problems simultaneously, we can calculate a one-step approximation for the robust estimate in each bootstrap sample, starting from the solution in the original sample. It has been shown that when a linear correction is used, this fast bootstrap procedure is robust and consistent in the sense that the bootstrap distribution converges weakly to the distribution of the estimators (see Refs 71,75,76 for details).

The MVE estimator of multivariate location and scatter discussed here is also related to the minimum covariance determinant estimator (MCD) that was introduced in Refs 1,2. The MCD looks for the h observations whose sample covariance matrix has the smallest possible determinant. The MCD estimates of location and scatter are then the sample mean and sample covariance matrix (multiplied by a consistency factor) of that optimal subset of h observations where, as before, h is usually chosen between $[n/2]+1$ and n. The MCD estimators of location and scatter have the same breakdown value as the MVE estimators (see e.g., Refs 77). The MCD has the additional advantage that it converges to a normal distribution at the regular $n^{-1/2}$ rate.[78] Its efficiency is generally low, but can be much increased by one-step reweighting.[11,18,79] For many years, the MVE was preferred over the MCD because of its slightly better computational efficiency when using a resampling algorithm. However, in 1999 a much faster MCD algorithm was developed[11] and since then many users prefer the MCD as robust estimator of location and scatter.

Projection estimators of multivariate location and scatter can combine a high breakdown value with a small maxbias that does not depend on the dimension p.[38,80–85] However, these estimators are substantially harder to compute, especially in higher dimensions. Finally, we note that the MVE has also been extended to a class of maximum trimmed likelihood estimators.[86]

CONCLUSION

We have reviewed the MVE estimator of multivariate location and scatter. An overview of the main properties of the MVE has been given, including its affine equivariance, breakdown value, and efficiency. The finite-sample efficiency can easily be improved by reweighting the initial MVE estimator. We discussed computation of the MVE using a resampling algorithm based on $(p+1)$-subsets. Several researchers have focused on the development of efficient algorithms to calculate approximate MVE solutions. However, it seems to us that there is still room for improvement. Moreover, many of the already proposed improvements are not available in most statistical software packages, in contrast to the standard resampling algorithm. The high breakdown value and low maxbias of the MVE estimators make them very useful for outlier detection in multivariate datasets, as illustrated in this article. This property is often used in regression to detect leverage points. An overview of applications of MVE has been given as well as some extensions of MVE to larger classes of robust estimators with useful properties. Note that an extensive overview of high-breakdown robust multivariate methods has been given in Ref 87.

A challenging problem for future research is the development of robust estimators of multivariate location and scatter for very high-dimensional data, especially when the sample size is small compared to the dimension. A discussion on robustness in very high dimensions is provided in Ref 88.

REFERENCES

1. Rousseeuw PJ. Least median of squares regression. *J Am Stat Assoc* 1984, 79:871–880.
2. Rousseeuw PJ. *Multivariate estimation with high breakdown point*. In: Grossmann W, Pflug G, Vincze I, Wertz W, eds. *Mathematical Statistics and Applications*, Vol. B. Dordrecht: Reidel Publishing Company; 1985, 283–297.
3. Lee J. *Relationships between Properties of Pulp-Fibre and Paper*. PhD thesis, University of Toronto, 1992.
4. Rousseeuw PJ, Van Aelst S, Van Driessen K, Agulló J. Robust multivariate regression. *Technometrics* 2004, 46:293–305.
5. Pison G, Van Aelst S. Diagnostic plots for robust multivariate methods. *J Comput Graph Stat* 2004, 13:310–329.
6. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. *Robust Statistics: The Approach Based on Influence Functions*. New York, NY: Wiley, 1986.
7. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. New York, NY: Wiley-Interscience, 1987.
8. Maronna RA, Martin DR, Yohai VJ. *Robust Statistics: Theory and Methods*. New York: Wiley, 2006.
9. Davies L, Gather U. The identification of multiple outliers. *J Am Stat Assoc* 1993, 88:782–792.
10. Becker C, Gather U. The masking breakdown point of multivariate outlier identification rules. *J Am Stat Assoc* 1999, 94:947–955.
11. Rousseeuw PJ, Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999, 41:212–223.

12. Lopuhaä HP, Rousseeuw PJ. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann Stat* 1991, 19:229–248.

13. Davies L. The asymptotics of Rousseeuw's Minimum Volume Ellipsoid estimator. *Ann Stat* 1992, 20:1828–1843.

14. Donoho, DL, Huber, PJ. The notion of breakdown point. In: Bickel P, Doksum K, Hodges J, eds. *A Festschrift for Erich Lehmann*. Belmont, CA: Wadsworth; 1983, 157–184.

15. Davies L. Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *Ann Stat* 1987, 15:1269–1292.

16. Rousseeuw PJ. Discussion of 'Breakdown and Groups'. *Ann Stat* 2005, 33:1004–1009.

17. Rousseeuw PJ, van Zomeren BC. Unmasking multivariate outliers and leverage points. *J Am Stat Assoc* 1990, 85:633–651.

18. Lopuhaä HP. Asymptotics of reweighted estimators of multivariate location and scatter. *Ann Stat* 1999, 27:1638–1665.

19. Lopuhaä HP. Highly efficient estimators of multivariate location with high breakdown point. *Ann Stat* 1992, 20:398–413.

20. He XM, Wang G. Cross-checking using the Minimum Volume Ellipsoid estimator. *Stat Sin* 1996, 6:367–374.

21. Cook RD, Hawkins DM, Weisberg S. Exact iterative computation of the robust multivariate Minimum Volume Ellipsoid estimator. *Stat Probab Lett* 1993, 16:213–218.

22. Agulló J. Exact iterative computation of the multivariate Minimum Volume Ellipsoid estimator with a branch and bound algorithm. In: Prat A, ed. *Compstat Proceedings in Computational Statistics*. Heidelberg: Springer-Verlag, 1996, 44–45.

23. Rousseeuw PJ, Bassett G. Robustness of the p-subset algorithm for regression with high breakdown point. In: Stahel W, Weisberg S, eds. *Directions in Robust Statistics and Diagnostics, Part II*, The IMA Volumes in Mathematics and Its Applications. Springer verlag, New York, NY, 1991, 185–194.

24. Rousseeuw PJ, Hubert, M. Recent developments in progress. In: Dodge Y, ed. *L1-Statistical Procedures and Related Topics*, IMS Lecture Notes-Monograph Series, Vol. 31. Hayward, CA: Institute of Mathematical Statistics, 1997, 201–214.

25. Croux C, Haesbroeck G. An easy way to increase the finite-sample efficiency of the resampled Minimum Volume Ellipsoid estimator. *Comput Stat Data Anal* 1997, 25:125–141.

26. Croux C, Haesbroeck G. A note on finite-sample efficiencies of estimators for the Minimum Volume Ellipsoid. *J Stat Comput Simulation* 2002, 72:585–596.

27. Croux C, Haesbroeck G, Rousseeuw PJ. Location adjustment for the Minimum Volume Ellipsoid estimator. *Stat Comput* 2002, 12:191–200.

28. Hawkins DM. A feasible solution algorithm for the Minimum Volume Ellipsoid estimator in multivariate data. *Comput Stat* 1993, 8:95–107.

29. Woodruff DL, Rocke DM. Heuristic search algorithms for the Minimum Volume Ellipsoid. *J Comput Graph Stat* 1993, 2:69–95.

30. Woodruff DL, Rocke DM. Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *J Am Stat Assoc* 1994, 89:888–896.

31. Rocke DM, Woodruff DL. Robust estimation of multivariate location and shape. *J Stat Plan Inference* 1997, 57:245–255.

32. Poston WL, Wegman EJ, Priebe CE, Solka JL. A deterministic method for robust estimation of multivariate location and shape. *J Comput Graph Stat* 1997, 6:300–313.

33. Poston WL, Wegman EJ, Solka JL. D-optimal design methods for robust estimation of multivariate location and scatter. *J Stat Plan Inference* 1998, 73:205–213.

34. Hawkins DM, Olive DJ. Improved feasible solution algorithms for high breakdown estimation. *Comput Stat Data Anal* 1999, 30:1–11.

35. Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY. Quantile approximation for robust statistical estimation and kappa-enclosing problems. *Int J Comput Geometry Appl* 2000, 10:593–608.

36. Croux C, Haesbroeck G. Maxbias curves of location estimators based on subranges. *J Nonparametr Stat* 2002, 14:295–306.

37. Croux C, Haesbroeck G. Maxbias curves of robust scale estimators based on subranges. *Metrika* 2001, 53:101–122.

38. Adrover J, Yohai V. Projection estimates of multivariate location. *Ann Stat* 2002, 30:1760–1781.

39. Adrover J, Yohai V. Bias behaviour of the Minimum Volume Ellipsoid estimate. In: Hubert M, Pison G, Struyf A, Van Aelst S, eds. *Theory and Applications of Recent Robust Methods*. Basel: Statistics for Industry and Technology, Birkhäuser, 2004, 1–12.

40. Stahel WA. *Robuste Schätzungen: infinitesimale Optimalität und Schätzungen von Kovarianzmatrizen*. PhD thesis, ETH Zürich, 1981.

41. Donoho DL. *Breakdown properties of multivariate location estimators*, Qualifying paper. Boston: Harvard University, 1982.

42. Maronna RA, Yohai VJ. The behavior of the Stahel-Donoho robust multivariate estimator. *J Am Stat Assoc* 1995, 90:330–341.

43. Rousseeuw PJ, van Zomeren BC. Robust distances: simulations and cutoff values. In: Stahel W, Weisberg S, eds. *Directions in Robust Statistics and*

Diagnostics, Part II, The IMA Volumes in Mathematics and Its Applications. New York: Springer Verlag, 1991, 195–203.

44. Cho JH, Gemperline PJ. Pattern-recognition analysis of near-infrared spectra by robust distance method. *J Chemom* 1995, 9:169–178.
45. Seaver BL, Triantis KP. The impact of outliers and leverage points for technical efficiency measurement using high breakdown procedures. *Manage Sci* 1995, 41:937–956.
46. Plets H, Vynckier C. An analysis of the incidence of the vega phenomenon among main-sequence and post main-sequence stars. *Astron Astrophys* 1999, 343:496–506.
47. Simpson DG, Ruppert D, Carroll RJ. On one-step GM-estimates and stability of inferences in linear regression. *J Am Stat Assoc* 1992, 87:439–450.
48. Coakley CW, Hettmansperger TP. A bounded influence, high breakdown, efficient regression estimator. *J Am Stat Assoc* 1993, 88:872–880.
49. Du ZY, Wiens DP. Jackknifing, weighting, diagnostics and variance estimation in generalized M-estimation. *Stat Probab Lett* 2000, 46:287–299.
50. Simpson DG, Chang YI. Reweighting approximate GM estimators: asymptotics and residual-based graphics. *J Stat Plan Inference* 1997, 57:273–293.
51. Naranjo JD, Hettmansperger TP. Bounded influence rank regression. *J R Stat Soc B* 1994, 56:209–220.
52. Chang WH, McKean JW, Naranjo JD, Sheather SJ. High-breakdown rank regression. *J Am Stat Assoc* 1999, 94:205–219.
53. Bianco A, Boente G, di Rienzo J. Some results for robust GM-based estimators in heteroscedastic regression models. *J Stat Plan Inference* 2000, 89:215–242.
54. Bianco A, Boente G. On the asymptotic behavior of one-step estimates in heteroscedastic regression models. *Stat Probab Lett* 2002, 60:33–47.
55. Fried R. Robust location estimation under dependence. *J Stat Comput Simulation* 2007, 77:131–147.
56. Jackson DA, Chen Y. Robust principal component analysis and outlier detection with ecological data. *Environmetrics* 2004, 15:129–139.
57. Chen Y, Chen X, Xu L. Developing a size indicator for fish populations. *Sci Mar* 2008, 72:221–229.
58. Chork CY, Rousseeuw PJ. Integrating a high-breakdown option into discriminant analysis in exploration geochemistry. *J Geochem Explor* 1992, 43:191–203.
59. Todorov VK, Neykov NM, Neytchev PN. Robust selection of variables in the discriminant-analysis based on MVE and MCD estimators. In: Momirovic K, Mildner V, eds. *Compstat Proceedings in Computational Statistics*. Heidelberg: Physica-Verlag, 1990, 193–198.
60. Chork CY, Salminen R. Interpreting exploration geochemical data from Outokumpu, Finland - a MVE-robust factor-analysis. *J Geochem Explor* 1993, 48:1–20.
61. Pison G, Rousseeuw PJ, Filzmoser P, Croux C. Robust factor analysis. *J Multivar Anal* 2003, 84:145–172.
62. Jolion JM, Meer P, Bataouche S. Robust clustering with applications in computer vision. *IEEE Trans Pattern Anal Mach Intell* 1991, 13:791–802.
63. Vargas JA. Robust estimation in multivariate control charts for individual observations. *J Qual Technol* 2003, 35:367–376.
64. Williams JD, Woodal WH, Birch JB. Statistical monitoring of quality profiles of products and processes. *Qual Reliab Eng Int* 2007, 23:925–941.
65. Jensen WA, Birch JB, Woodall WH. High breakdown estimation methods for phase I multivariate control charts. *Qual Reliab Eng Int* 2007, 23:615–629.
66. Everitt B. *An R and S-PLUS® Companion to Multivariate Analysis*. London: Springer-Verlag, 2007.
67. Rousseeuw PJ, Yohai V. Robust regression by means of S-estimators. In: Franke J, Härdle W, Martin D, eds. *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics No. 26. Berlin: Springer Verlag, 1984, 256–272.
68. Lopuhaä HP. On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann Stat* 1989, 17:1662–1683.
69. Roelant E, Van Aelst S, Croux C. Multivariate generalized S-estimators. *J Multivar Analysis* 2009, 100:876–887.
70. Tatsuoka KS, Tyler DE. On the uniqueness of S-functionals and M-functionals under nonelliptical distributions. *Ann Stat* 2000, 28:1219–1243.
71. Salibian-Barrera M, Van Aelst S, Willems G. PCA based on multivariate MM-estimators with fast and robust bootstrap. *J Am Stat Assoc* 2006, 101:1198–1211.
72. Kent JT, Tyler DE. Constrained M-estimation for multivariate location and scatter. *Ann Stat* 1996, 24:1346–1370.
73. Lopuhaä HP. Multivariate τ-estimators for location and scatter. *Can J Stat* 1991, 19:307–321.
74. Martin RD, Zamar RH. Bias robust estimation of scale. *Ann Stat* 1993, 21:991–1017.
75. Salibian-Barrera M, Van Aelst S, Willems G. Fast and robust bootstrap. *Stat Methods Appl* 2008, 17:41–71.
76. Van Aelst S, Willems G. Multivariate regression S-estimators for robust estimation and inference. *Stat Sin* 2005, 15:981–1001.
77. Agulló J, Croux C, Van Aelst S. The multivariate Least Trimmed Squares estimator. *J Multivar Analysis* 2008, 99:311–338.

78. Butler RW, Davies PL, Jhun M. Asymptotics for the Minimum Covariance Determinant estimator. *Ann Stat* 1993, 21:1385–1400.

79. Croux C, Haesbroeck G. Influence function and efficiency of the Minimum Covariance Determinant scatter matrix estimator. *J Multivar Analysis* 1999, 71:161–190.

80. Maronna RA, Stahel WA, Yohai VJ. Bias-robust estimators of multivariate scatter based on projections. *J Multivar Analysis* 1992, 42:141–161.

81. Tyler DE. Finite-sample breakdown points of projection-based multivariate location and scatter statistics. *Ann Stat* 1994, 22:1024–1044.

82. Zuo Y. Projection-based depth functions and associated median. *Ann Stat* 2003, 31:1460–1490.

83. Zuo Y, Cui H, He X. On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *Ann Stat* 2004, 32:167–188.

84. Zuo Y, Cui H, Young D. Influence function and maximum bias of projection depth based estimators. *Ann Stat* 2004, 32:189–218.

85. Zuo Y, Cui H. Depth weighted sactter estimators. *Ann Stat* 2005, 33:381–413.

86. Hadi AS, Luceño A. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Comput Stat Data Anal* 1997, 25:251–272.

87. Hubert M, Rousseeuw PJ, Van Aelst S. High-breakdown robust multivariate methods. *Stat Sci* 2008, 23:92–119.

88. Alqallaf F, Van Aelst S, Yohai VJ, Zamar RH. Propagation of outliers in multivariate data. *Ann Stat* 2009, 37:311–331.

