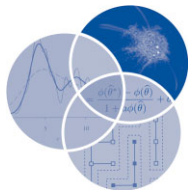
Advanced Review



Minimum covariance determinant

Mia Hubert,[1]* and Michiel Debruyne[2]

The minimum covariance determinant (MCD) estimator is a highly robust estimator of multivariate location and scatter. It can be computed efficiently with the FAST-MCD algorithm of Rousseeuw and Van Driessen. Since estimating the covariance matrix is the cornerstone of many multivariate statistical methods, the MCD has also been used to develop robust and computationally efficient multivariate techniques.

In this paper, we review the MCD estimator, along with its main properties such as affine equivariance, breakdown value, and influence function. We discuss its computation, and list applications and extensions of the MCD in theoretical and applied multivariate statistics. © 2009 John Wiley & Sons, Inc. *WIREs Comp Stat* 2010 2 36–43

The minimum covariance determinant (MCD) estimator is one of the first affine equivariant and highly robust estimators of multivariate location and scatter.[1,2] Being resistant to outlying observations, makes the MCD very helpful in outlier detection. Although already introduced in 1984, its main use has only started since the introduction of the computationally efficient FAST-MCD algorithm of Rousseeuw and Van Driessen.[3] Since then, the MCD has been applied in numerous fields such as medicine, finance, image analysis, and chemistry. Moreover, the MCD has also been used to develop many robust multivariate techniques, such as principal component analysis, factor analysis, and multiple regression.

DESCRIPTION OF THE MCD ESTIMATOR

Motivation

In the multivariate location and scatter setting we assume that the data 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. Hence n stands for the number of objects and p for the number of variables.

To illustrate we first consider a bivariate data set, hence $p = 2$. We consider the *wine* data set, available in Ref 4 and also analyzed in Ref 5. The data set contains the quantities of 13 constituents found in three types of Italian wines. We consider the first group containing 59 wines, and focus on the constituents 'malic acid' and 'proline'. A scatter plot of the data is shown in Figure 1, together with the classical and the robust 97.5% tolerance ellipse.

The classical tolerance ellipse is defined as the set of p-dimensional points $\boldsymbol{x}$ whose *Mahalanobis distance*

$$MD(\boldsymbol{x}) = \sqrt{(\boldsymbol{x} - \bar{\boldsymbol{x}})^t \boldsymbol{S}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}})} \qquad (1)$$

equals $\sqrt{\chi^2_{p,0.975}}$. We denote $\chi^2_{p,\alpha}$ as the α-quantile of the χ^2_p distribution. The Mahalanobis distance $\mathrm{MD}(\boldsymbol{x}_i)$ should tell us how far away $\boldsymbol{x}_i$ is from the center of the cloud, relative to the size of the cloud. Here $\bar{\boldsymbol{x}}$ is the sample mean and $\boldsymbol{S}$ the sample covariance matrix. We see that this tolerance ellipse tries to encompass all observations. Consequently none of the Mahalanobis distances, shown in Figure 2(a), is exceptional large and only three observations would be considered as mild outliers. On the other hand, the robust tolerance ellipse in Figure 1 which is based on the robust distances

$$RD(\boldsymbol{x}) = \sqrt{(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{MCD})^t \hat{\boldsymbol{\Sigma}}^{-1}_{MCD} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_{MCD})} \qquad (2)$$

is much smaller and encloses the regular data points. Here, $\hat{\boldsymbol{\mu}}_{MCD}$ is the MCD estimate of location, and $\hat{\boldsymbol{\Sigma}}_{MCD}$ the MCD covariance estimate. The robust distances exposed in Figure 2(b) now clearly spot eight outliers and one mild outlier (observation 3).

This illustrates the *masking effect*: classical estimates can be so highly affected by outlying

*Correspondence to: mia.hubert@wis.kuleuven.be

[1]Department of Mathematics-LStat, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium

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

DOI: 10.1002/wics.61

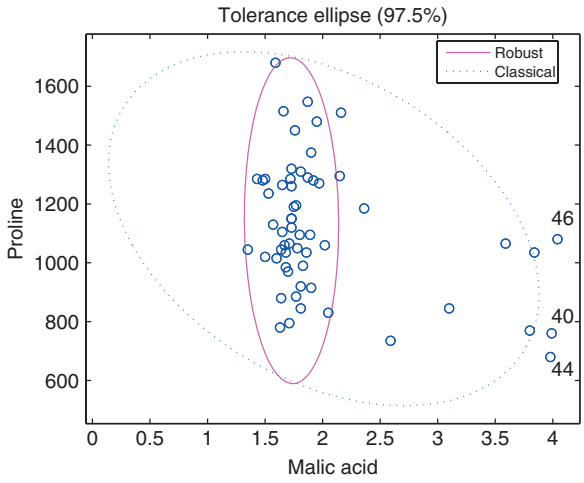


FIGURE 1 | Bivariate wine data with classical and robust tolerance ellipse.

values that diagnostic tools such as the Mahalanobis distances can no longer detect the outliers. To get a reliable analysis of these data robust estimators are required that can resist possible outliers. The MCD estimator of location and scatter is such a robust estimator.

Definition

Denote [.] the floor function. The raw MCD estimator with parameter $[(n+p+1)/2] \leq h \leq n$ defines the following location and dispersion estimates:

1. $\hat{\boldsymbol{\mu}}_0$ is the mean of the h observations for which the determinant of the sample covariance matrix is minimal.
2. $\hat{\boldsymbol{\Sigma}}_0$ is the corresponding covariance matrix multiplied by a consistency factor c_0.

Note that the MCD estimator can only be computed when $h > p$, otherwise the covariance matrix of any h-subset will be singular. Since $h \geq [(n+2)/2]$, this condition is certainly satisfied when $n \geq 2p$. To avoid the curse of dimensionality it is however recommended that $n > 5p$. To obtain consistency at the normal distribution, the consistency factor c_0 equals $\alpha / F_{\chi^2_{p+2}}(\chi^2_{p,\alpha})$ with $\alpha = \lim_{n\to\infty} h(n)/n$.[6] Also a finite-sample correction factor can be added.[7]

The MCD estimator is designed for *elliptically symmetric unimodal distributions*. A multivariate distribution with parameters $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma}$ a positive definite matrix of size p is called elliptically symmetric and unimodal if there exists a strictly decreasing function g such that the density can be written in the form

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|}}\, g((\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})). \qquad (3)$$

Consistency of the raw MCD estimator of location and scatter at elliptical models, as well as asymptotic normality of the MCD location estimator has been proved in Ref 8.

The MCD estimator is most robust by taking $h = [(n+p+1)/2]$. This corresponds at the population level with $\alpha = 0.5$. But unfortunately the MCD then suffers from low efficiency at the normal model. For example, if $\alpha = 0.5$, the asymptotic relative efficiency of the diagonal elements of the MCD scatter matrix with regard to the sample covariance matrix is only

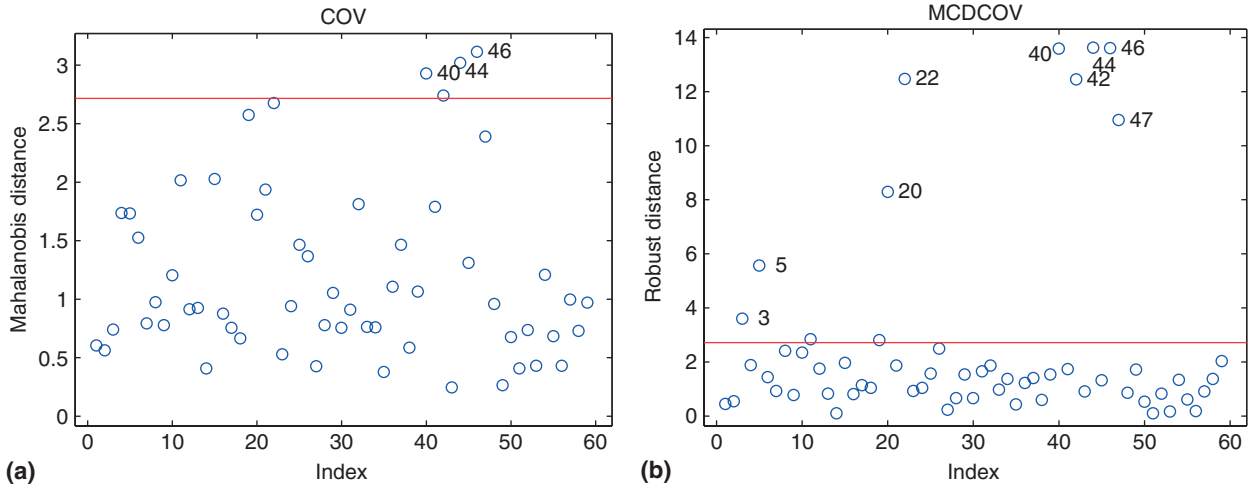


FIGURE 2 | (a) Mahalanobis distances and (b) robust distances for the bivariate wine data.

6% when $p = 2$, and 20.5% when $p = 10$. This efficiency can be increased by considering a larger α such as $\alpha = 0.75$. This yields relative efficiencies of 26.2% for $p = 2$ and 45.9% for $p = 10$.[6] On the other hand this choice of α decreases the robustness toward possible outliers.

In order to increase the efficiency while retaining high robustness, one can apply reweighted estimators.[9,10] For the MCD this yields the estimates:

$$\hat{\boldsymbol{\mu}}_{MCD} = \frac{\sum_{i=1}^{n} W(d_i^2)\boldsymbol{x}_i}{\sum_{i=1}^{n} W(d_i^2)} \tag{4}$$

$$\hat{\boldsymbol{\Sigma}}_{MCD} = c_1 \frac{1}{n} \sum_{i=1}^{n} W(d_i^2)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{MCD})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{MCD})^t \tag{5}$$

with $d_i = \sqrt{(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_0)^t \hat{\boldsymbol{\Sigma}}_0^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_0)}$ and W an appropriate weight function. The constant c_1 is again a consistency factor. A simple yet effective choice for W is $W(d^2) = I(d^2 \leq \chi^2_{p,0.975})$. This is the default choice in current implementations in S-PLUS, R, SAS, and Matlab. If we take $\alpha = 0.5$ this reweighting step increases the efficiency up to 45.5% for $p = 2$ and 82% for $p = 10$. The example of the wine data also uses the reweighted MCD estimator with $\alpha = 0.75$, but the results were similar for smaller values of α.

Remark that based on the MCD covariance matrix a robust correlation matrix can also be constructed. For all $1 \leq i \neq j \leq p$, the robust correlation between variables X_i and X_j can be estimated by

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii} s_{jj}}} \tag{6}$$

with s_{ij} the (i,j)th element of the MCD covariance estimate.

Outlier detection

As illustrated in Figure 3, the robust MCD estimator is very helpful to detect outliers in multivariate data. As the robust distances (Eq. (2)) are not sensitive to the masking effect, they can be used to flag the outliers.[11] This becomes even more useful at data sets in more than two (or three) dimensions, which become difficult to visualize.

We illustrate the outlier detection potential of the MCD on the full wine data set, with $p = 13$ variables. The distance--distance plot of Figure 3 now plots the robust distances based on the MCD versus the Mahalanobis distances.[3] From the robust analysis we see that seven observations clearly stand out, whereas the classical analysis does not flag any of the wines.

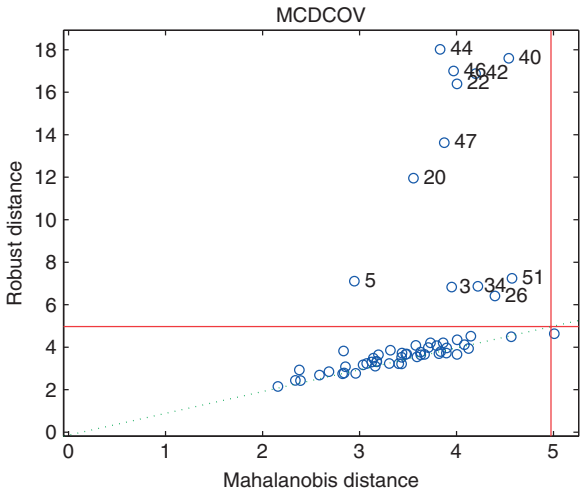


FIGURE 3 | Distance–distance plot of the full wine data.

Note that the cutoff value $\sqrt{\chi^2_{p,0.975}}$ is based on the asymptotic distribution of the robust distances, and often flags too many observations as outlying. The true distribution of the robust distances can be better approximated by an F-distribution, see Ref. 12.

PROPERTIES

Affine equivariance

The MCD estimator of location and scatter is *affine equivariant*. This means that for any nonsingular matrix $\boldsymbol{A}$ and constant vector $\boldsymbol{b} \in \mathbb{R}^p$

$$\hat{\boldsymbol{\mu}}_{MCD}(\boldsymbol{AX} + \boldsymbol{b}) = \boldsymbol{A}\hat{\boldsymbol{\mu}}_{MCD}(\boldsymbol{X}) + \boldsymbol{b} \tag{7}$$

$$\hat{\boldsymbol{\Sigma}}_{MCD}(\boldsymbol{AX} + \boldsymbol{b}) = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{MCD}(\boldsymbol{X})\boldsymbol{A}^t. \tag{8}$$

This property follows from the fact that for each subset of size h, denoted as $\boldsymbol{X}_h$, the determinant of the covariance matrix of the transformed data equals

$$|\boldsymbol{S}(\boldsymbol{AX}_h)| = |\boldsymbol{A}\boldsymbol{S}(\boldsymbol{X}_h)\boldsymbol{A}^t| = |\boldsymbol{A}|^2 |\boldsymbol{S}(\boldsymbol{X}_h)|. \tag{9}$$

Hence, the optimal h-subset (which minimizes $|\boldsymbol{S}(\boldsymbol{AX}_h)|$) remains the same as for the original data (which minimizes $|\boldsymbol{S}(\boldsymbol{X}_h)|$), and its covariance matrix is appropriately transformed. Similarly the affine equivariance of the raw MCD location estimator follows from the equivariance of the sample mean. Finally we note that the robust distances $d_i = \sqrt{(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_0)^t \hat{\boldsymbol{\Sigma}}_0^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_0)}$ are affine invariant, which implies that the reweighted estimator is again equivariant.

Affine equivariance implies that the estimator transforms well under any nonsingular reparametrization of the space of the $\boldsymbol{x}_i$. Consequently, the data might be rotated, translated, or rescaled (for example through a change of the measurement units) without affecting the outlier detection diagnostics.

The MCD is one of the first high-breakdown affine equivariant estimators of location and scatter, and was only preceded by the Stahel-Donoho estimator.[13,14] Together with the MCD, the minimum volume ellipsoid estimator was introduced[1,2] but this estimator is not asymptotically normal and it is more difficult to compute than the MCD.

Breakdown value

The breakdown value of an estimator measures the smallest fraction of observations that need to be replaced by arbitrary values to carry the estimate beyond all bounds. Denote $\boldsymbol{X}_{n,m}$ as the set obtained by replacing m data points $\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_m}$ of $\boldsymbol{X}_n$ by arbitrary values. For a multivariate location estimator T_n the breakdown value is then defined as:

$$\epsilon_n^*(T_n; \boldsymbol{X}_n) = \frac{1}{n} \min \{m \in \{1, \ldots, n\} : \sup_m ||T_n(\boldsymbol{X}_n) - T_n(\boldsymbol{X}_{n,m})|| = +\infty\}. \tag{10}$$

For a multivariate estimator of scatter we have

$$\epsilon_n^*(C_n; \boldsymbol{X}_n) = \frac{1}{n} \min \{m \in \{1, \ldots, n\} : \sup_m \max_i \{|\log(\lambda_i(C_n(\boldsymbol{X}_n))) - \log(\lambda_i(C_n(\boldsymbol{X}_{n,m})))|\}\}, \tag{11}$$

with $0 < \lambda_p(C_n) \leq \ldots \leq \lambda_1(C_n)$ the eigenvalues of C_n. This means that we consider a scatter estimator to be broken whenever any of the eigenvalues can become arbitrary large or arbitrary close to 0.

Let $k(\boldsymbol{X}_n)$ denote the maximum number of observations in the data set lying on a hyperplane of $\mathbb{R}^p$. Assume $k(\boldsymbol{X}_n) < h$, then for the raw MCD estimator of location and scatter, we have that[15]

$$\epsilon_n^*(\hat{\boldsymbol{\mu}}_0; \boldsymbol{X}_n) = \epsilon_n^*(\hat{\boldsymbol{\Sigma}}_0; \boldsymbol{X}_n) = \frac{\min(n - h + 1, h - k(\boldsymbol{X}_n))}{n}. \tag{12}$$

If the data are sampled from a continuous distribution, then almost surely $k(\boldsymbol{X}_n) = p$ which yields $\epsilon_n^*(\hat{\boldsymbol{\mu}}_0; \boldsymbol{X}_n) = \epsilon_n^*(\hat{\boldsymbol{\Sigma}}_0; \boldsymbol{X}_n) = \min(n - h + 1, \ h - p)/n$, and consequently any $[(n + p)/2] \leq h \leq [(n + p + 1)/2]$ gives the maximal breakdown value $[(n - p + 2)/2]$. This is also the highest possible breakdown value for affine equivariant scatter estimators[16] at data sets that satisfy $k(\boldsymbol{X}_n) = p$ (this is also known as *general position*). Also for affine equivariant location estimators the upper bound on the breakdown value is $[(n - p + 2)/2]$ under natural regularity conditions.[17] Note that in the limit $\lim_{n\to\infty} \epsilon_n^* = \min(1 - \alpha, \alpha)$ which is maximal for $\alpha = 0.5$.

Finally, we remark that the breakdown value of the reweighted MCD estimator $\hat{\boldsymbol{\mu}}_{MCD}$ and $\hat{\boldsymbol{\Sigma}}_{MCD}$ is not lower than the breakdown value of the raw MCD estimator, as long as the weight function W used in Eq. (4) is bounded and becomes zero for large d_i.[9]

Influence function

The influence function of an estimator measures the infinitesimal effect of point contamination on the estimator.[18] It is defined at the population level, hence it requires the functional form of the estimator T, which maps any distribution F on a value $T(F)$ in the parameter space. For multivariate location, this parameter space is $\mathbb{R}^p$, whereas for multivariate scatter estimators the parameter space corresponds with all positive definite matrices of size p. The influence function of the estimator T at the distribution F in a point $\boldsymbol{x}$ is then defined as:

$$IF(\boldsymbol{x}, T, F) = \lim_{\varepsilon \to 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} \tag{13}$$

with $F_\varepsilon = (1 - \varepsilon)F + \varepsilon \Delta_x$ a contaminated distribution with point mass in $\boldsymbol{x}$.

The influence function of the raw and the reweighted MCD has been studied in Ref. 6 and appears to be bounded. This is a desired property for robust estimators. It reflects the robustness of the estimator toward point contamination. At the standard Gaussian distribution, the influence function of the MCD location estimator becomes zero for all $\boldsymbol{x}$ with $\|\boldsymbol{x}\|^2 > \chi^2_{p,\alpha}$, hence large outliers do not influence the estimates. The same happens at the off-diagonal elements of the MCD scatter estimator. At the diagonal elements on the other hand the influence function remains constant (different from zero) when $\|\boldsymbol{x}\|^2$ is sufficiently large. This reflects that the outliers still have a bounded influence of the estimator. Moreover the influence functions are smooth, except at those $\boldsymbol{x}$ with $\|\boldsymbol{x}\|^2 = \chi^2_{p,\alpha}$. The reweighted MCD estimator has an additional jump in $\|\boldsymbol{x}\|^2 = \chi^2_{p,0.975}$ because of the discontinuity of the weight function.

Univariate MCD

For univariate data, the MCD estimates reduce to the mean and the variance of the h-subset with smallest variance. They can be computed in $O(n \log n)$ time by considering contiguous h-subsets and by computing their mean and variance recursively.[19] Their consistency and asymptotic normality is proved in Ref 2,20. For $h = [n/2] + 1$ the MCD location estimator has breakdown value $[(n+1)/2]/n$, and the MCD scale estimator $[n/2]/n$. These are the maximal values that can be attained by affine equivariant estimators.[21] The univariate MCD influence functions are also bounded, see[6] for details. The maxbias curve (which measures the maximal asymptotic bias at a certain fraction of contamination) is studied in Ref 22,23.

Note that the univariate MCD location estimator corresponds with the univariate least trimmed squares (LTS) estimator,[1] defined by

$$\min_{\mu} \sum_{i=1}^{h} (r_\mu^2)_{i:n} \tag{14}$$

where $(r_\mu^2)_{1:n} \leq (r_\mu^2)_{2:n} \leq \ldots \leq (r_\mu^2)_{n:n}$ are the ordered squared residuals $(x_i - \mu)^2$.

COMPUTATION

The exact MCD estimator is very hard to compute, as it requires the evaluation of all $\binom{n}{h}$ subsets of size h. In Ref. 3 the FAST-MCD algorithm is developed to efficiently compute the MCD. The key component of the algorithm is the C-step:

Theorem. *Take* $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ *and let* $H_1 \subset \{1, \ldots, n\}$ *be an* h*-subset, that is* $|H_1| = h$*. Put* $\hat{\boldsymbol{\mu}}_1$ *and* $\hat{\boldsymbol{\Sigma}}_1$ *the empirical mean and covariance matrix of the data in* H_1*. If* $det(\hat{\boldsymbol{\Sigma}}_1) \neq 0$ *define the relative distances*

$$d_1(i) := \sqrt{(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)^t \hat{\boldsymbol{\Sigma}}_1^{-1} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_1)} \qquad \text{for } i = 1, \ldots, n. \tag{15}$$

Now take H_2 *such that* $\{d_1(i); i \in H_2\} := \{(d_1)_{1:n}, \ldots, (d_1)_{h:n}\}$ *where* $(d_1)_{1:n} \leq (d_1)_{2:n} \leq \cdots \leq (d_1)_{n:n}$ *are the ordered distances, and compute* $\hat{\boldsymbol{\mu}}_2$ *and* $\hat{\boldsymbol{\Sigma}}_2$ *based on* H_2*. Then*

$$\det(\hat{\boldsymbol{\Sigma}}_2) \leq \det(\hat{\boldsymbol{\Sigma}}_1) \tag{16}$$

with equality if and only if $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1$ *and* $\hat{\boldsymbol{\Sigma}}_2 = \hat{\boldsymbol{\Sigma}}_1$.

If $\det(\hat{\boldsymbol{\Sigma}}_1) > 0$, the C-step thus yields a new h-subset with lower covariance determinant very easily. Note that the C stands for 'concentration' since $\hat{\boldsymbol{\Sigma}}_2$ is more concentrated (has a lower determinant) than $\hat{\boldsymbol{\Sigma}}_1$. The condition $\det(\hat{\boldsymbol{\Sigma}}_1) \neq 0$ in the C-step theorem is not a real restriction because if $\det(\hat{\boldsymbol{\Sigma}}_1) = 0$ the minimal objective value is already reached.

C-steps can be iterated until $\det(\hat{\boldsymbol{\Sigma}}_{\text{new}}) = 0$ or $\det(\hat{\boldsymbol{\Sigma}}_{\text{new}}) = \det(\hat{\boldsymbol{\Sigma}}_{\text{old}})$. The sequence of determinants obtained in this way must converge in a finite number of steps because there are only finitely many h-subsets. However, there is no guarantee that the final value $\det(\hat{\boldsymbol{\Sigma}}_{\text{new}})$ of the iteration process is the global minimum of the MCD objective function. Therefore, an approximate MCD solution can be obtained by taking many initial choices of H_1, applying C-steps to each and keeping the solution with lowest determinant.

To construct an initial subset H_1, a random $(p+1)$-subset J is drawn and $\hat{\boldsymbol{\mu}}_0 := \text{ave}(J)$ and $\hat{\boldsymbol{\Sigma}}_0 := \text{cov}(J)$ are computed [If $\det(\hat{\boldsymbol{\Sigma}}_0) = 0$ then J can be extended by adding observations until $\det(\hat{\boldsymbol{\Sigma}}_0) > 0$]. Then, for $i = 1, \ldots, n$ the distances $d_0^2(i) := (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_0)^t \hat{\boldsymbol{\Sigma}}_0^{-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_0)$ are computed and sorted. The initial H_1 subset then consists of the h observations with smallest distance d_0. This method yields better initial subsets than by drawing random h-subsets directly, because the probability of drawing an outlier-free subset is much higher when drawing $(p+1)$-subsets than with h-subsets.

The FAST-MCD algorithm contains several computational improvements. As each C-step involves the calculation of a covariance matrix, its determinant and the corresponding distances using fewer C-steps considerably improves the speed of the algorithm. It turns out that after two C-steps, many runs that will lead to the global minimum already have a considerably smaller determinant. Therefore, the number of C-steps is reduced by applying only two C-steps on each initial subset and selecting the 10 different subsets with lowest determinants. Only for these 10 subsets further C-steps are taken until convergence.

This procedure is very fast for small sample sizes n, but when n grows the computation time increases because of the n distances that need to be calculated in each C-step. For large n FAST-MCD uses a partitioning of the data set, which avoids doing all the calculations in the entire data.

Note that the FAST-MCD algorithm itself is affine equivariant. Consistency and breakdown of the approximate algorithm is discussed in Ref 24.

Implementations of the FAST-MCD algorithm are available in the package S-PLUS (as the built-in function *cov.mcd*), in R (as part of the packages *rrcov*, *robust* and *robustbase*), in SAS/IML Version 7,

and SAS Version 9 (in *PROC ROBUSTREG*). A stand-alone program can be downloaded from the website *www.agoras.ua.ac.be*, as well as a Matlab version. A Matlab function is also part of LIBRA, a Matlab Library for Robust Analysis[25] which can be downloaded from *wis.kuleuven.be/stat/robust*. Moreover it is available in the PLS_Toolbox of Eigenvector Research (*www.eigenvector.com*). Note that some functions use $\alpha = 0.5$ as default value, yielding a breakdown value of 50%, whereas other implementations use $\alpha = 0.75$.

APPLICATIONS

Many multivariate statistical methods rely on covariance estimation; hence the MCD estimator is well suited to construct robust multivariate techniques. Moreover, the trimming idea of the MCD has been generalized toward many new estimators. Here, we enumerate some applications and extensions.

The MCD analog in regression is the least trimmed squares regression estimator[1] which minimizes the sum of the h-smallest squared residuals. Equivalently, the LTS estimate corresponds with the least squares fit of the h-subset with smallest squared residuals. The FAST-LTS algorithm uses similar techniques as FAST-MCD.[26] The diagnostic plot introduced in Ref 11 exposes the regression residuals versus the robust distances of the predictors, and is very useful for outlier classification. These robust distances are, e.g., also useful for robust linear regression,[27,28] regression with continuous and categorical regressors[29] and for logistic regression.[30,31] In the multivariate regression setting (with several response variables) the MCD can be directly used to obtain MCD regression,[32] whereas MCD applied to the residuals leads to multivariate LTS estimation.[33]

Covariance estimation is also important in principal component analysis and related methods. For low-dimensional data (with $n < 5p$) the principal components can be obtained as the eigenvectors of the MCD covariance matrix,[34] whereas robust factor analysis based on the MCD has been studied in Ref 35. Robust canonical correlation is proposed in Ref 36. For high-dimensional data, projection pursuit ideas combined with the MCD results in the so-called ROBPCA method[37,38] for robust PCA. This method has led to the construction of robust principal component regression,[39] and robust partial least squares regression,[40,41] together with appropriate outlier maps. The LTS-subspace estimator[42] generalizes LTS regression to subspace estimation and orthogonal regression.

An MCD-based alternative to the Hotelling test was provided in Ref. 43. A robust bootstrap for the MCD is proposed in Ref. 44. The computation of the MCD with missing values is explored in Ref 45–47. Classification (or discriminant analysis) based on MCD is studied in Ref 48,49, whereas an alternative for high-dimensional data is developed in Ref. 50. Robust clustering is handled in Ref 51–53.

The trimming procedure of the MCD has inspired the construction of maximum trimmed likelihood estimators,[54–57] trimmed k-means,[58–60] least weighted squares regression[61] and minimum weighted covariance determinant estimation.[15] The idea of the C-step in the MCD algorithm has been extended to S-estimators.[62]

Applications of the MCD are numerous. We mention recent applications in finance and econometrics,[63,64] medicine,[65] quality control,[66] geophysics,[67] image analysis[68,69] and chemistry,[70] but this list is far from complete.

CONCLUSION

In this paper we have reviewed the MCD estimator of multivariate location and scatter. We have illustrated its resistance to outliers on an example of real data. Its main properties concerning robustness, efficiency, and equivariance were enumerated, and computational aspects were described. Finally we have provided a detailed reference list with applications and generalizations of the MCD in theoretical and applied research.

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. Rousseeuw PJ, Van Driessen K. A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics* 1999, 41:212–223.

4. Hettich S, Bay SD. *The UCI KDD Archive*. Irvine, CA: University of California, Department of Information and Computer Science; 1999.
5. Maronna RA, Martin DR, Yohai VJ. *Robust statistics: Theory and Methods*. New York: Wiley; 2006.
6. Croux C, Haesbroeck G. Influence function and efficiency of the Minimum Covariance Determinant scatter matrix estimator. *J Multivariate Anal* 1999, 71:161–190.
7. Pison G, Van Aelst S, Willems G. Small sample corrections for LTS and MCD. *Metrika* 2002, 55:111–123.
8. Butler RW, Davies PL, Jhun M. Asymptotics for the Minimum Covariance Determinant estimator. *Ann Stat* 1993, 21:1385–1400.
9. Lopuhaä HP, Rousseeuw PJ. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann Stat* 1991, 19:229–248.
10. Lopuhaä HP. Asymptotics of reweighted estimators of multivariate location and scatter. *Ann Stat* 1999, 27:1638–1665.
11. Rousseeuw PJ, van Zomeren BC. Unmasking multivariate outliers and leverage points. *J Am Stat Assoc* 1990, 85:633–651.
12. Hardin J, Rocke DM. The distribution of robust distance. *J Comput Graph Stat* 2005, 14:928–946.
13. Stahel WA, *Robuste schätzungen: infinitesimale optimalität und schätzungen von kovarianzmatrizen*. PhD thesis, ETH Zürich, 1981.
14. Donoho DL, Gasko M. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann Stat* 1992, 20:1803–1827.
15. Roelant E, Van Aelst S, Willems G. The minimum weighted covariance determinant estimator. *Metrika* 2009, 70:177–204.
16. Davies L. Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *Ann Stat* 1987, 15:1269–1292.
17. Rousseeuw PJ. Discussion on 'Breakdown and groups'. *Ann Stat* 2005, 33:1004–1009.
18. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley; 1986.
19. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. New York: Wiley-Interscience; 1987.
20. Butler RW. Nonparametric interval and point prediction using data trimmed by a Grubbs-type outlier rule. *Ann Stat* 1982, 10:197–204.
21. Croux C, Rousseeuw PJ. A class of high-breakdown scale estimators based on subranges. *Commun Stat Theory Methods* 1992, 21:1935–1951.
22. Croux C, Haesbroeck G. Maxbias curves of robust scale estimators based on subranges. *Metrika* 2001, 53:101–122.
23. Croux C, Haesbroeck G. Maxbias curves of location estimators based on subranges. *J Nonparametr Stat* 2002, 14:295–306.
24. Hawkins D, Olive D. Inconsistency of resampling algorithms for high breakdown regression estimators and a new algorithm. *J Am Stat Assoc* 2002, 97:136–148.
25. Verboven S, Hubert M. LIBRA: a Matlab library for robust analysis. *Chemometr Intell Lab Syst* 2005, 75:127–136.
26. Rousseeuw PJ, Van Driessen K. Computing LTS regression for large data sets. *Data Min Knowl Discov* 2006, 12:29–45.
27. 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.
28. Coakley CW, Hettmansperger TP. A bounded influence, high breakdown, efficient regression estimator. *J Am Stat Assoc* 1993, 88:872–880.
29. Hubert M, Rousseeuw PJ. Robust regression with both continuous and binary regressors. *J Stat Plann Infer* 1996, 57:153–163.
30. Rousseeuw PJ, Christmann A. Robustness against separation and outliers in logistic regression. *Comput Stat Data Anal* 2003, 43:315–332.
31. Croux C, Haesbroeck G. Implementing the Bianco and Yohai estimator for logistic regression. *Comput Stat Data Anal* 2003, 44:273–295.
32. Rousseeuw PJ, Van Aelst S, Van Driessen K, Agulló J. Robust multivariate regression. *Technometrics* 2004, 46:293–305.
33. Agulló J, Croux C, Van Aelst S. The multivariate least trimmed squares estimator. *J Multivariate Anal* 2008, 99:311–318.
34. 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.
35. Pison G, Rousseeuw PJ, Filzmoser P, Croux C. Robust factor analysis. *J Multivariate Anal* 2003, 84:145–172.
36. Croux C, Dehon C. Analyse canonique basée sur des estimateurs robustes de la matrice de covariance. *Rev Stat Appl* 2002, 2:5–26.
37. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: a new approach to robust principal components analysis. *Technometrics* 2005, 47:64–79.
38. Debruyne M, Hubert M. The influence function of the Stahel-Donoho covariance estimator of smallest outlyingness. *Stat Probab Lett* 2009, 79:275–282.
39. Hubert M, Verboven S. A robust PCR method for high-dimensional regressors. *J Chemometr* 2003, 17:438–452.
40. Hubert M, Vanden Branden K. Robust methods for Partial Least Squares Regression. *J Chemometr* 2003, 17:537–549.

41. Vanden Branden K, Hubert M. Robustness properties of a robust PLS regression method. *Anal Chim Acta* 2004, 515:229–241.
42. Maronna RA. Principal components and orthogonal regression based on robust scales. *Technometrics* 2005, 47:264–273.
43. Willems G, Pison G, Rousseeuw PJ, Van Aelst S. A robust Hotelling test. *Metrika* 2002, 55:125–138.
44. Willems G, Van Aelst S. A fast bootstrap method for the MCD estimator. In: Antoch J, ed. *Proceedings in Computational Statistics*. Heidelberg: Springer-Verlag; 2004, 1979–1986.
45. Cheng T-C, Victoria-Feser M. High breakdown estimation of multivariate location and scale with missing observations. *Br J Math Stat Psychol* 2002, 55:317–335.
46. Copt S, Victoria-Feser M-P. Fast algorithms for computing high breakdown covariance matrices with missing data. 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, 71–82.
47. Serneels S, Verdonck T. Principal component analysis for data containing outliers and missing elements. *Comput Stat Data Anal* 2008, 52:1712–1727.
48. Hawkins DM, McLachlan GJ. High-breakdown linear discriminant analysis. *J Am Stat Assoc* 1997, 92:136–143.
49. Hubert M, Van Driessen K. Fast and robust discriminant analysis. *Comput Stat Data Anal* 2004, 45:301–320.
50. Vanden Branden K, Hubert M. Robust classification in high dimensions based on the SIMCA method. *Chemometr Intell Lab Syst* 2005, 79:10–21.
51. Rocke DM, Woodruff DL. *A synthesis of outlier detection and cluster identification*, technical report, 1999.
52. Hardin J, Rocke DM. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Comput Stat Data Anal* 2004, 44:625–638.
53. Gallegos MT, Ritter G. A robust method for cluster analysis. *Ann Stat* 2005, 33:347–380.
54. Vandev DL, Neykov NM. About regression estimators with high breakdown point. *Statistics* 1998, 32:111–129.
55. Hadi AS, Luceño A. Maximum trimmed likelihood estimators: a unified approach, examples and algorithms. *Comput Stat Data Anal* 1997, 25:251–272.
56. Müller CH, Neykov N. Breakdown points of trimmed likelihood estimators and related estimators in generalized linear models. *J Stat Plann Infer* 2003, 116:503–519.
57. Čižek P. Robust and efficient adaptive estimation of binary-choice regression models. *J Am Stat Assoc* 2008, 103:687–696.
58. Cuesta-Albertos JA, Gordaliza A, Matrán C. Trimmed k-means: an attempt to robustify quantizers. *Ann Stat* 1997, 25:553–576.
59. Cuesta-Albertos JA, Matrán C, Mayo-Iscar A. Robust estimation in the normal mixture model based on robust clustering. *J R Stat Soc Ser B* 2008, 70:779–802.
60. García-Escudero LA, Gordaliza A, San Martín R, Van Aelst S, Zamar RH. Robust linear clustering. *J R Stat Soc B* 2009, 71:1–18.
61. Víšek JÁ. The least weighted squares I. the asymptotic linearity of normal equations. *Bull Czech Econ Soc* 2002, 9:31–58.
62. Salibian-Barrera M, Yohai VJ. A fast algorithm for S-regression estimates. *J Comput Graph Stat* 2006, 15:414–427.
63. Zaman A, Rousseeuw PJ, Orhan M. Econometric applications of high-breakdown robust regression techniques. *Econ Lett* 2001, 71:1–8.
64. Welsh R, Zhou X. Application of robust statistics to asset allocation models. *Revstat* 2007, 5:97–114.
65. Prastawa M, Bullitt E, Ho S, Gerig G. A brain tumor segmentation framework based on outlier detection. *Med Image Anal* 2004, 8:275–283.
66. Jensen WA, Birch JB, Woodal WH. High breakdown estimation methods for phase I multivariate control charts. *Qual Reliab Eng Int* 2007, 23:615–629.
67. Neykov NM, Neytchev PN, Van Gelder PHAJM, Todorov VK. Robust detection of discordant sites in regional frequency analysis. *Water Resour Res* 2007, 43.
68. Vogler C, Goldenstein S, Stolfi J, Pavlovic V, Metaxas D. Outlier rejection in high-dimensional deformable models. *Image Vis Comput* 2007, 25: 274–284.
69. Lu Y, Wang J, Kong J, Zhang B, Zhang J. An integrated algorithm for MRI brain images segmentation. *Comput Vis Approaches Med Image Anal* 2006, 4241:132–1342.
70. van Helvoort PJ, Filzmoser P, van Gaans PFM. Sequential Factor Analysis as a new approach to multivariate analysis of heterogeneous geochemical datasets: An application to a bulk chemical characterization of fluvial deposits (Rhine-Meuse delta, The Netherlands). *Appl Geochem* 2005, 20:2233–2251.

