# Size and Power of Multivariate Outlier Detection Rules

**Andrea Cerioli, Marco Riani, and Francesca Torti**

**Abstract** Multivariate outliers are usually identified by means of robust distances. A statistically principled method for accurate outlier detection requires both availability of a good approximation to the finite-sample distribution of the robust distances and correction for the multiplicity implied by repeated testing of all the observations for outlyingness. These principles are not always met by the currently available methods. The goal of this paper is thus to provide data analysts with useful information about the practical behaviour of some popular competing techniques. Our conclusion is that the additional information provided by a data-driven level of trimming is an important bonus which ensures an often considerable gain in power.

## 1 Introduction

Obtaining reliable information on the quality of the available data is often the first of the challenges facing the statistician. It is thus not surprising that the systematic study of methods for detecting outliers and immunizing against their effect has a long history in the statistical literature. See, e.g., Cerioli et al. (2011a), Hadi et al. (2009), Hubert et al. (2008) and Morgenthaler (2006) for recent reviews on this topic. We quote from Morgenthaler (2006, p. 271) that "Robustness of statistical methods in the sense of insensitivity to grossly wrong measurements is probably as old as the experimental approach to science". Perhaps less known is the fact that

A. Cerioli (✉) · M. Riani
Dipartimento di Economia, Università di Parma, Parma, Italy
e-mail: andrea.cerioli@unipr.it; mriani@unipr.it

F. Torti
Dipartimento di Economia, Università di Parma, Parma, Italy

Joint Research Centre, European Commission, Ispra (VA), Italy

B. Lausen et al. (eds.), *Algorithms from and for Nature and Life*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-00035-0_1,

similar concerns were also present in the Ancient Greece more than 2,400 years ago, as reported by Thucydides in his History of The Peloponnesian War (III 20): "The Plataeans, who were still besieged by the Peloponnesians and Boeotians, ... made ladders equal in length to the height of the enemy's wall, which they calculated by the help of the layers of bricks on the side facing the town ... A great many counted at once, and, although some might make mistakes, the calculation would be oftener right than wrong; for they repeated the process again and again ... In this manner they ascertained the proper length of the ladders".[1]

With multivariate data outliers are usually identified by means of robust distances. A statistically principled rule for accurate multivariate outlier detection requires:

(a) An accurate approximation to the finite-sample distribution of the robust distances under the postulated model for the "good" part of the data;
(b) Correction for the multiplicity implied by repeated testing of all the observations for outlyingness.

These principles are not always met by the currently available methods. The goal of this paper is to provide data analysts with useful information about the practical behaviour of popular competing techniques. We focus on methods based on alternative high-breakdown estimators of multivariate location and scatter, and compare them to the results from a rule adopting a more flexible level of trimming, for different data dimensions. The present thus extends that of (Cerioli et al. 2011b), where only low dimensional data are considered. Our conclusion is that the additional information provided by a data-driven approach to trimming is an important bonus often ensuring a considerable gain in power. This gain may be even larger when the number of variables increases.

## 2 Distances for Multivariate Outlier Detection

### *2.1 Mahalanobis Distances and the Wilks' Rule*

Let $y_1, \ldots, y_n$ be a sample of $v$-dimensional observations from a population with mean vector $\mu$ and covariance matrix $\Sigma$. The basic population model for which most of the results described in this paper were obtained is that

$$y_i \sim N(\mu, \Sigma) \qquad i = 1, \ldots, n. \tag{1}$$

[1]The Authors are grateful to Dr. Spyros Arsenis and Dr. Domenico Perrotta for pointing out this historical reference.

The sample mean is denoted by $\hat{\mu}$ and $\hat{\Sigma}$ is the unbiased sample estimate of $\Sigma$. The Mahalanobis distance of observation $y_i$ is

$$d_i^2 = (y_i - \hat{\mu})' \hat{\Sigma}^{-1} (y_i - \hat{\mu}). \tag{2}$$

For simplicity, we omit the fact that $d_i^2$ is squared and we call it a distance.

Wilks (1963) showed in a seminal paper that, under the multivariate normal model (1), the Mahalanobis distances follow a scaled Beta distribution:

$$d_i^2 \sim \frac{(n-1)^2}{n} \text{Beta}\left(\frac{v}{2}, \frac{n-v-1}{2}\right) \qquad i = 1, \ldots, n. \tag{3}$$

Wilks also conjectured that a Bonferroni bound could be used to test outlyingness of the most remote observation without losing too much power. Therefore, for a nominal test size $\alpha$, Wilk's rule for multivariate outlier identification takes the largest Mahalanobis distance among $d_1^2, \ldots, d_n^2$, and compares it to the $1 - \alpha/n$ quantile of the scaled Beta distribution (3). This gives an outlier test of nominal test size $\leq \alpha$.

Wilks' rule, adhering to the basic statistical principles (a) and (b) of Sect. 1, provides an accurate and powerful test for detecting a single outlier even in small and moderate samples, as many simulation studies later confirmed. However, it can break down very easily in presence of more than one outlier, due to the effect of masking. Masking occurs when a group of extreme outliers modifies $\hat{\mu}$ and $\hat{\Sigma}$ in such a way that the corresponding distances become negligible.

## 2.2 Robust Distances

One effective way to avoid masking is to replace $\hat{\mu}$ and $\hat{\Sigma}$ in (2) with high-breakdown estimators. A robust distance is then defined as

$$\tilde{d}_i^2 = (y_i - \tilde{\mu})' \tilde{\Sigma}^{-1} (y_i - \tilde{\mu}), \tag{4}$$

where $\tilde{\mu}$ and $\tilde{\Sigma}$ denote the chosen robust estimators of location and scatter. We can expect multivariate outliers to be highlighted by large values of $\tilde{d}_i^2$, even if masked in the corresponding Mahalanobis distances (2), because now $\tilde{\mu}$ and $\tilde{\Sigma}$ are not affected by the outliers.

One popular choice of $\tilde{\mu}$ and $\tilde{\Sigma}$ is related to the Minimum Covariance Determinant (MCD) criterion (Rousseeuw and Van Driessen 1999). In the first stage, we fix a coverage $\lfloor n/2 \rfloor \leq h < n$ and we define the MCD subset to be the subsample of $h$ observations whose covariance matrix has the smallest determinant. The MCD estimator of $\mu$, say $\tilde{\mu}_{(\text{MCD})}$, is the average of the MCD subset, whereas the MCD estimator of $\Sigma$, say $\tilde{\Sigma}_{(\text{MCD})}$, is proportional to the dispersion matrix of this

subset (Pison et al. 2002). A second stage is then added with the aim of increasing efficiency, while preserving the high-breakdown properties of $\tilde{\mu}_{(\mathrm{MCD})}$ and $\tilde{\Sigma}_{(\mathrm{MCD})}$. Therefore, a one-step reweighting scheme is applied by giving weight $w_i = 0$ to observations whose first-stage robust distance exceeds a threshold value. Otherwise the weight is $w_i = 1$. We consider the Reweighted MCD (RMCD) estimator of $\mu$ and $\Sigma$, which is defined as

$$\tilde{\mu}_{\mathrm{RMCD}} = \frac{\sum_{i=1}^{n} w_i y_i}{w}, \quad \tilde{\Sigma}_{\mathrm{RMCD}} = \frac{\kappa \sum_{i=1}^{n} w_i (y_i - \tilde{\mu}_{(\mathrm{RMCD})})(y_i - \tilde{\mu}_{(\mathrm{RMCD})})'}{w - 1},$$

where $w = \sum_{i=1}^{n} w_i$ and the scaling $\kappa$, depending on the values of $m$, $n$ and $v$, serves the purpose of ensuring consistency at the normal model. The resulting robust distances for multivariate outlier detection are then

$$\tilde{d}^2_{i(\mathrm{RMCD})} = (y_i - \tilde{\mu}_{\mathrm{RMCD}})' \tilde{\Sigma}^{-1}_{\mathrm{RMCD}} (y_i - \tilde{\mu}_{\mathrm{RMCD}}) \qquad i = 1, \ldots, n. \tag{5}$$

Multivariate S estimators are another common option for $\tilde{\mu}$ and $\tilde{\Sigma}$. For $\tilde{\mu} \in \Re^v$ and $\tilde{\Sigma}$ a positive definite symmetric $v \times v$ matrix, they are defined to be the solution of the minimization problem $|\tilde{\Sigma}| = \min$ under the constraint

$$\frac{1}{n} \sum_{i=1}^{n} \rho(\tilde{d}_i^2) = \zeta, \tag{6}$$

where $\tilde{d}_i^2$ is given in (4), $\rho(x)$ is a smooth function satisfying suitable regularity and robustness properties, and $\zeta = E\{\rho(z'z)\}$ for a $v$-dimensional vector $z \sim N(0, I)$. The $\rho$ function in (6) rules the weight given to each observation to achieve robustness. Different specifications of $\rho(x)$ lead to numerically and statistically different S estimators. In this paper we deal with two such specifications. The first one is the popular Tukey's Biweight function

$$\rho(x) = \begin{cases} \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4} & \text{if } |x| \le c \\ \frac{c^2}{6} & \text{if } |x| > c, \end{cases} \tag{7}$$

where $c > 0$ is a tuning constant which controls the breakdown point of S estimators; see Rousseeuw and Leroy (1987, pp.135–143) and Riani et al. (2012) for details. The second alternative that we consider is the slightly more complex Rocke's Biflat function, described, e.g., by Maronna et al. (2006, p. 190). This function assigns weights similar to (7) to distance values close to the median, but null weights outside a user-defined interval. Specifically, let

$$\eta = \min \left( \frac{\chi^2_{v,(1-\gamma)}}{v} - 1, 1 \right), \tag{8}$$

where $\chi^2_{v,(1-\gamma)}$ is the $1-\gamma$ quantile of $\chi^2_v$. Then, the weight under Rocke's Biflat function is 0 whenever a normalized version of the robust distance $\tilde{d}_i^2$ is outside the interval $[1-\eta, 1+\eta]$. This definition ensures better performance of S estimators when $v$ is large. Indeed, it can be proved (Maronna et al. 2006, p. 221) that the weights assigned by Tukey's Biweight function (7) become almost constant as $v \to \infty$. Therefore, robustness of multivariate S estimators is lost in many practical situations where $v$ is large. Examples of this behaviour will be seen in Sect. 3.2 even for $v$ as small as 10.

Given the robust, but potentially inefficient, S estimators of $\mu$ and $\Sigma$, an improvement in efficiency is sometimes advocated by computing refined location and shape estimators which satisfy a more efficient version of (6) (Salibian-Barrera et al. 2006). These estimators, called MM estimators, are defined as the minimizers of

$$\frac{1}{n}\sum_{i=1}^{n}\rho_*(\tilde{\tilde{d}}_i^2), \tag{9}$$

where

$$\tilde{\tilde{d}}_i^2 = (y_i - \tilde{\tilde{\mu}})'\tilde{\tilde{\Sigma}}^{-1}(y_i - \tilde{\tilde{\mu}}) \tag{10}$$

and the function $\rho_*(x)$ provides higher efficiency than $\rho(x)$ at the null model (1). Minimization of (9) is performed over all $\tilde{\tilde{\mu}} \in \Re^v$ and all $\tilde{\tilde{\Sigma}}$ belonging to the set of positive definite symmetric $v \times v$ matrices with $|\tilde{\tilde{\Sigma}}| = 1$. The MM estimator of $\mu$ is then $\tilde{\tilde{\mu}}$, while the estimator of $\Sigma$ is a rescaled version of $\tilde{\tilde{\Sigma}}$. Practical implementation of MM estimators is available using Tukey's Biweight function only (Todorov and Filzmoser 2009). Therefore, we follow the same convention in the performance comparison to be described in Sect. 3.

## 2.3 The Forward Search

The idea behind the Forward Search (FS) is to apply a flexible and data-driven trimming strategy to combine protection against outliers and high efficiency of estimators. For this purpose, the FS divides the data into a good portion that agrees with the postulated model and a set of outliers, if any (Atkinson et al. 2004). The method starts from a small, robustly chosen, subset of the data and then fits subsets of increasing size, in such a way that outliers and other observations not following the general structure are revealed by diagnostic monitoring. Let $m_0$ be the size of the starting subset. Usually $m_0 = v + 1$ or slightly larger. Let $S^{(m)}$ be the subset of data fitted by the FS at step $m$ ($m = m_0, \ldots, n$), yielding estimates $\hat{\mu}(m)$, $\hat{\Sigma}(m)$ and distances

$$\hat{d}_i^2(m) = \{y_i - \hat{\mu}(m)\}'\hat{\Sigma}(m)^{-1}\{y_i - \hat{\mu}(m)\} \qquad i = 1, \ldots, n.$$

These distances are ordered to obtain the fitting subset at step $m+1$. Whilst $S^{(m)}$ remains outlier free, they will not suffer from masking.

The main diagnostic quantity computed by the FS at step $m$ is

$$\hat{d}^2_{i_{\min}}(m): \qquad i_{\min} = \arg\ \min \hat{d}^2_i(m) \ \text{ for } i \notin S^{(m)}, \tag{11}$$

i.e. the distance of the closest observation to $S^{(m)}$, among those not belonging to this subset. The rationale is that the robust distance of the observation entering the fitting subset at step $m+1$ will be large if this observation is an outlier. Its peculiarity will then be revealed by a peak in the forward plot of $d^2_{i_{\min}}(m)$.

All the FS routines, as well as the algorithms for computing most of the commonly adopted estimators for regression and multivariate analysis, are contained in the FSDA toolbox for MATLAB and are freely downloadable from http://www.riani.it/MATLAB or from the web site of the Joint Research Centre of the European Commission. This toolbox also contains a series of dynamic tools which enable the user to link the information present in the different plots produced by the FS, such as the index or forward plot of robust Mahalanobis distances $\hat{d}^2_i(m)$ and the scatter plot matrix; see Perrotta et al. (2009) for details.

## 3 Comparison of Alternative Outlier Detection Rules

Precise outlier identification requires cut-off values for the robust distances when model (1) is true. If $\tilde{\mu} = \tilde{\mu}_{\text{RMCD}}$ and $\tilde{\Sigma} = \tilde{\Sigma}_{\text{RMCD}}$, Cerioli et al. (2009) show that the usually trusted asymptotic approximation based on the $\chi^2_v$ distribution can be largely unsatisfactory. Instead, Cerioli (2010) proposes a much more accurate approximation based on the distributional rules

$$\tilde{d}^2_{i(\text{RMCD})} \sim \frac{(w-1)^2}{w}\text{Beta}\left(\frac{v}{2}, \frac{w-v-1}{2}\right) \qquad \text{if} \qquad w_i = 1 \tag{12}$$

$$\sim \frac{w+1}{w}\frac{(w-1)v}{w-v}F_{v,w-v} \qquad \text{if} \qquad w_i = 0, \tag{13}$$

where $w_i$ and $w$ are defined as in Sect. 2.2. Cerioli and Farcomeni (2011) show how the same distributional results can be applied to deal with multiplicity of tests to increase power and to provide control of alternative error rates in the outlier detection process.

In the context of the Forward Search, Riani et al. (2009) propose a formal outlier test based on the sequence $\hat{d}^2_{i_{\min}}(m)$, $m = m_0, \ldots, n-1$, obtained from (11). In this test, the values of $\hat{d}^2_{i_{\min}}(m)$ are compared to the FS envelope

$$V^2_{m,\alpha}/\sigma_T(m)^2,$$

where $V^2_{m,\alpha}$ is the $100\alpha\,\%$ cut-off point of the $(m+1)$th order statistic from the scaled $F$ distribution

$$\frac{(m^2-1)v}{m(m-v)}F_{v,m-v}, \tag{14}$$

and the factor

$$\sigma_T(m)^2 = \frac{P(X^2_{v+2} < \chi^2_{v,m/n})}{m/n} \tag{15}$$

allows for trimming of the $n-m$ largest distances. In (15), $X^2_{v+2} \sim \chi^2_{v+2}$ and $\chi^2_{v,m/n}$ is the $m/n$ quantile of $\chi^2_v$.

The flexible trimming strategy enjoyed by the FS ensures a balance between the two enemy brothers of robust statistics: robustness against contamination and efficiency under the postulated multivariate normal model. This makes the Forward Search a valuable benchmark against which alternative competitors should be compared. On the other hand, very little is known about the finite sample behaviour of the outlier detection rules which are obtained from the multivariate S and MM estimators summarized in Sect. 2.2. In the rest of this section, we thus explore the performance of the alternative rules with both "good" and contaminated data, under different settings of the required user-defined tuning constants. We also provide comparison with power results obtained with the robust RMCD distances (5) and with the flexible trimming approach given by the FS.

## *3.1 Size*

Size estimation is performed by Monte Carlo simulation of data sets generated from the $v$-variate normal distribution $N(0, I)$, due to affine invariance of the robust distances (4). The estimated size of each outlier detection rule is defined to be the proportion of simulated data sets for which the null hypothesis of no outliers, i.e. the hypothesis that all $n$ observations follow model (1), is wrongly rejected. For S and MM estimation, the finite sample null distribution of the robust distances $\tilde{d}^2_i$ is unknown, even to a good approximation. Therefore, these distances are compared to the $1-\alpha/n$ quantile of their asymptotic distribution, which is $\chi^2_v$. As in the Wilks' rule of Sect. 2.1, the Bonferroni correction ensures that the actual size of the test of no outliers will be bounded by the specified value of $\alpha$ if the $\chi^2_v$ approximation is adequate.

In our investigation we also evaluate the effect on empirical test sizes of each of some user-defined tuning constants required for practical computation of multivariate S and MM estimators. See, e.g., Todorov and Filzmoser (2009) for details. Specifically, we consider:

- `bdp`: breakdown point of the S estimators, which is inherited by the MM estimators as well (the default value is 0.5);
- `eff`: efficiency of the MM estimators (the default value is 0.95);

- `effshape`[2]: dummy variable setting whether efficiency of the MM estimators is defined with respect to shape (`effshape` = 1) or to location (`effshape` = 0, the default value);
- `nsamp`: number of sub-samples of dimension $(p+1)$ in the resampling algorithm for fast computation of S estimators (our default value is 100);
- `refsteps`: maximum number of iterations in the Iterative Reweighted Least Squares algorithm for computing MM estimators (our default value is 20);
- `gamma`: tail probability in (8) for Rocke's Biflat function (the default value is 0.1).

Tables 1 and 2 report the results for $n = 200$, $v = 5$ and $v = 10$, when $\alpha = 0.01$ is the nominal size for testing the null hypothesis of no outliers and 5,000 independent data sets are generated for each of the selected combinations of parameter values. The outlier detection rule based on S estimators with Tukey's Biweight function (7) is denoted by ST. Similarly, SR is the S rule under Rocke's Biflat function. It is seen that the outlier detection rules based on the robust S and MM distances with Tukey's Biweight function can be moderately liberal, but with estimated sizes often not too far from the nominal target. As expected, liberality is an increasing function of dimension and of the breakdown point, both for S and MM estimators. Efficiency of the MM estimators (`eff`) is the only tuning constant which seems to have a major impact on the null behaviour of these detection rules. On the other hand, SR has the worst behaviour under model (1) and its size can become unacceptably high, especially when $v$ grows. As a possible explanation, we note that a number of observations having positive weight under ST receive null weight with SR (Maronna et al. 2006, p. 192). This fact introduces a form of trimming in the corresponding estimator of scatter, which is not adequately taken into account. The same result also suggests that better finite-sample approximations to the null distribution of the robust distances $\tilde{d}_i^2$ with Rocke's Biflat function are certainly worth considering.

### *3.2 Power*

We now evaluate the power of ST, SR and MM multivariate outlier detection rules. We also include in our comparison the FS test of Riani et al. (2009), using (14), and the finite-sample RMCD technique of Cerioli (2010), relying on (12) and (13). These additional rules have very good control of the size of the test of no outliers even for sample sizes considerably smaller than $n = 200$, thanks to their accurate cut-off values. Therefore, we can expect a positive bias in the estimated power of all the procedures considered in Sect. 3.1, and especially so in that of SR.

[2]In the RRCOV packege of the R software this option is called `eff.shape`

**Table 1** Estimated size of the test of the hypothesis of no outliers for $n = 200$ and nominal test size $\alpha = 0.01$. ST is the outlier detection rule based on S estimators with Tukey's Biweight function (7); MM is the rule based on MM estimators, using again Tukey's Biweight function (7). Five thousand independent data sets are generated for each of the selected combinations of parameter values

| | | | bdp | | eff | | effshape | nsamp | | refsteps | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | all parameters at default value | 0.15 | 0.25 | 0.8 | 0.98 | 0 | 10 | 500 | 10 | 500 |
| $v = 5$ | ST | 0.023 | 0.010 | 0.014 | 0.023 | 0.023 | 0.023 | 0.026 | 0.024 | 0.023 | 0.023 |
| | MM | 0.021 | 0.019 | 0.020 | 0.023 | 0.015 | 0.023 | 0.021 | 0.020 | 0.022 | 0.023 |
| $v = 10$ | ST | 0.033 | 0.005 | 0.007 | 0.033 | 0.033 | 0.033 | 0.031 | 0.036 | 0.033 | 0.033 |
| | MM | 0.038 | 0.035 | 0.028 | 0.068 | 0.019 | 0.038 | 0.029 | 0.030 | 0.034 | 0.036 |

**Table 2** Estimated size of the test of the hypothesis of no outliers for $n = 200$ and nominal test size $\alpha = 0.01$, using S estimators with Rocke's Biflat function (SR), for different values of $\gamma$ in (8). Five thousand independent data sets are generated for each of the selected combinations of parameter values

| | gamma | | | | | |
|---|---|---|---|---|---|---|
| | 0.15 | 0.10 | 0.05 | 0.025 | 0.01 | 0.001 |
| $v = 5$ | 0.066 | 0.057 | 0.055 | 0.056 | 0.056 | 0.061 |
| $v = 10$ | 0.089 | 0.080 | 0.079 | 0.078 | 0.077 | 0.081 |

Average power of an outlier detection rule is defined to be the proportion of contaminated observations rightly named to be outliers. We estimate it by simulation, in the case $n = 200$ and for $v = 5$ and $v = 10$. For this purpose, we generate $v$-variate observations from the location-shift contamination model

$$y_i \sim (1-\delta)N(0, I) + \delta N(0 + \lambda e, I), \qquad i = 1, \ldots, n, \tag{16}$$

where $0 < \delta < 0.5$ is the contamination rate, $\lambda$ is a positive scalar and $e$ is a column vector of ones. The $0.01/n$ quantile of the reference distribution is our cut-off value for outlier detection. We only consider the default choices for the tuning constants in Tables 1 and 2, given that their effect under the null has been seen to be minor. We base our estimate of average power on 5,000 independent data sets for each of the selected combinations of parameter values.

It is worth noting that standard clustering algorithms, like $g$-means, are likely to fail to separate the two populations in (16), even in the ideal situation where there is a priori knowledge that $g = 2$. For instance, we have run a small benchmark study with $n = 200$, $v = 5$ and two overlapping populations by setting $\lambda = 2$ and $\delta = 0.05$ in model (16). We have found that the misclassification rate of $g$-means can be as high as 25 % even in this idyllic scenario where the true value of $g$ is known and the covariance matrices are spherical. The situation obviously becomes much worse when $g$ is unknown and must be inferred from the data. Furthermore, clustering algorithms based on Euclidean distances, like $g$-means, are not affine invariant and would thus provide different results on unstandardized data.

Tables 3–5 show the performance of the outlier detection rules under study for different values of $\delta$ and $\lambda$ in model (16). If the contamination rate is small, it is seen that the four methods behave somewhat similarly, with FS often ranking first and MM always ranking last as $\lambda$ varies. However, when the contamination rate increases, the advantage of the FS detection rule becomes paramount. In that situation both ST and MM estimators are ineffective for the purpose of identifying multivariate outliers. As expected, SR improves considerably over ST when $v = 10$ and $\delta = 0.15$, but remains ineffective when $\delta = 0.3$. Furthermore, it must be recalled that the actual size of SR is considerably larger, and thus power is somewhat biased.

**Table 3** Estimated average power for different shifts $\lambda$ in the contamination model (16), in the case $n = 200$, $v = 5$ and $v = 10$, when the contamination rate $\delta = 0.05$. Five thousand independent data sets are generated for each of the selected combinations of parameter values

| | | Mean shift $\lambda$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2 | 2.2 | 2.4 | 2.6 | 2.8 | 3 |
| $v = 5$ | ST | 0.344 | 0.525 | 0.696 | 0.827 | 0.912 | 0.963 |
| | SR | 0.387 | 0.549 | 0.698 | 0.820 | 0.908 | 0.957 |
| | MM | 0.148 | 0.280 | 0.466 | 0.672 | 0.836 | 0.935 |
| | RMCD | 0.227 | 0.390 | 0.574 | 0.732 | 0.856 | 0.936 |
| | FS | 0.359 | 0.567 | 0.730 | 0.840 | 0.909 | 0.953 |
| $v = 10$ | ST | 0.758 | 0.919 | 0.978 | 0.995 | 0.999 | 1 |
| | SR | 0.856 | 0.946 | 0.986 | 0.997 | 0.999 | 1 |
| | MM | 0.479 | 0.782 | 0.942 | 0.990 | 0.998 | 1 |
| | RMCD | 0.684 | 0.839 | 0.956 | 0.987 | 0.997 | 1 |
| | FS | 0.808 | 0.911 | 0.968 | 0.991 | 0.998 | 1 |

**Table 4** Quantities as in Table 3, but now for $\delta = 0.15$

| | | Mean shift $\lambda$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 2 | 2.4 | 2.6 | 2.8 | 3 | 3.4 |
| $v = 5$ | ST | 0.073 | 0.532 | 0.772 | 0.901 | 0.960 | 0.996 |
| | SR | 0.275 | 0.433 | 0.594 | 0.742 | 0.854 | 0.925 |
| | MM | 0.006 | 0.010 | 0.012 | 0.016 | 0.026 | 0.397 |
| | RMCD | 0.096 | 0.428 | 0.652 | 0.815 | 0.913 | 0.988 |
| | FS | 0.580 | 0.803 | 0.878 | 0.935 | 0.965 | 0.993 |
| $v = 10$ | ST | 0.006 | 0.007 | 0.008 | 0.01 | 0.013 | 0.041 |
| | SR | 0.696 | 0.825 | 0.895 | 0.923 | 0.931 | 0.946 |
| | MM | 0.001 | 0.001 | 0.001 | 0.001 | 0.003 | 0.030 |
| | RMCD | 0.530 | 0.938 | 0.959 | 0.993 | 1 | 1 |
| | FS | 0.887 | 0.938 | 0.974 | 0.991 | 0.998 | 1 |

A qualitative explanation for the failure of multivariate MM estimators is shown in Fig. 1 in the simple case $v = 2$. The four plots display bivariate ellipses corresponding to 0.95 probability contours at different iterations of the algorithm for computing MM estimators, for a data set simulated from the contamination model (16) with $n = 200$, $\delta = 0.15$ and $\lambda = 3$. The data can be reproduced using function `randn(200,2)` of MATLAB and putting the random number seed to 2. The contaminated units are shown with symbol $\circ$ and the two lines which intersect the estimate of the robust centroid are plotted using a dash-dot symbol. The upper left-hand panel corresponds to the first iteration (i1), where the location estimate is $\tilde{\mu} = (0.19, 0.18)'$ and the value of the robust correlation $r$ derived from $\tilde{\Sigma}$ is 0.26. In this case the robust estimates are not too far from the true parameter values $\mu = (0, 0)'$ and $\Sigma = I$, and the corresponding outlier detection rule (i.e., the ST

**Table 5** Quantities as in Table 3, but now for $\delta = 0.30$

| | | Mean shift $\lambda$ | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 2 | 2.4 | 2.6 | 2.8 | 3 | 4 | 6 |
| $v = 5$ | ST | 0.003 | 0.005 | 0.006 | 0.007 | 0.009 | 0.016 | 0.092 |
| | SR | 0.006 | 0.033 | 0.286 | 0.372 | 0.458 | 0.557 | 1 |
| | MM | 0.002 | 0.003 | 0.004 | 0.005 | 0.006 | 0.012 | 0.085 |
| | RMCD | 0.010 | 0.159 | 0.381 | 0.637 | 0.839 | 1 | 1 |
| | FS | 0.627 | 0.915 | 0.920 | 0.941 | 0.967 | 1 | 1 |
| $v = 10$ | ST | 0.002 | 0.002 | 0.003 | 0.003 | 0.003 | 0.004 | 0.011 |
| | SR | 0.002 | 0.005 | 0.004 | 0.005 | 0.009 | 0.011 | 0.039 |
| | MM | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 |
| | RMCD | 0.207 | 0.842 | 0.969 | 0.994 | 0.999 | 1 | 1 |
| | FS | 0.904 | 0.929 | 0.961 | 0.980 | 0.989 | 0.995 | 1 |

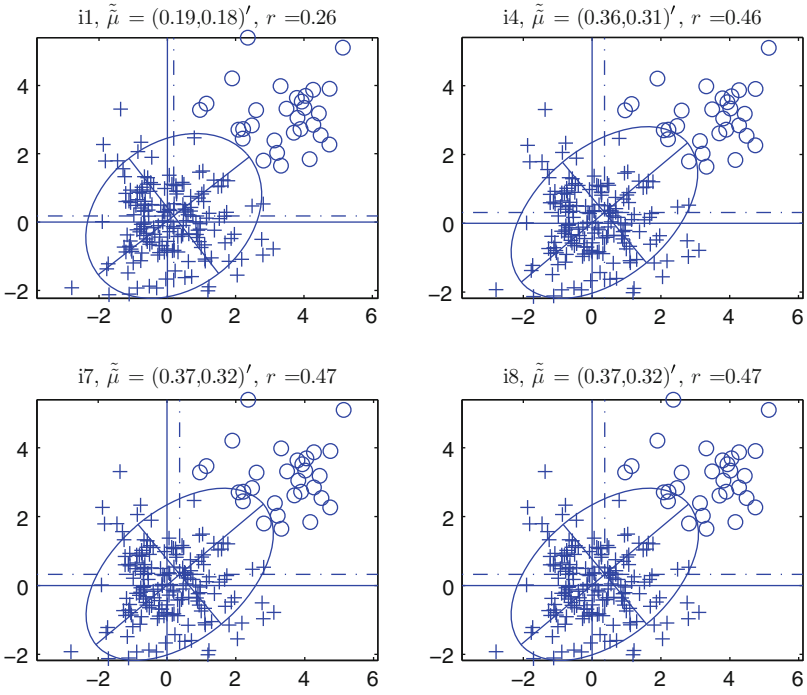


**Fig. 1** Ellipses corresponding to 0.95 probability contours at different iterations of the algorithm for computing multivariate MM estimators, for a data set simulated from the contamination model (16) with $n = 200$, $v = 2$, $\delta = 0.15$ and $\lambda = 3$

rule in Tables 3–5) can be expected to perform reasonably well. On the contrary, as the algorithm proceeds, the ellipse moves its center far from the origin and the variables artificially become more correlated. The value of $r$ in the final iteration (i8) is 0.47 and the final centroid $\tilde{\tilde{\mu}}$ is $(0.37; 0.32)'$. These features increase the bias

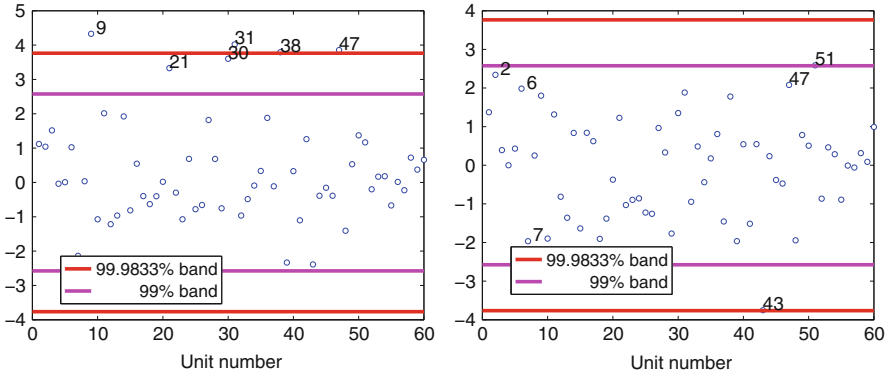


**Fig. 2** Index plots of robust scale residuals obtained using MM estimation with a preliminary $S$-estimate of scale based on a 50 % breakdown point. *Left-hand panel*: 90 % nominal efficiency; *right-hand panel*: 95 % nominal efficiency. The *horizontal lines* correspond to the 99 % individual and simultaneous bands using the standard normal

of the parameter estimates and can contribute to masking in the supposedly robust distances (10).

A similar effect can also be observed with univariate ($v = 1$) data. For instance, Atkinson and Riani (2000, pp. 5–9) and Riani et al. (2011) give an example of a regression dataset with 60 observations on three explanatory variables where there are six masked outliers (labelled 9, 21 30, 31, 38 47) that cannot be detected using ordinary diagnostic techniques. The scatter plot of the response against the three explanatory variables and the traditional plot of residuals against fitted values, as well as the *qq* plot of OLS residuals, do not reveal observations far from the bulk of the data. Figure 2 shows the index plots of the scaled MM residuals. In the left-hand panel we use a preliminary $S$ estimate of scale with Tukey's Biweight function (7) and 50 % breakdown point, and 90 % efficiency in the MM step under the same $\rho$ function. In the right-hand panel we use the same preliminary scale estimate as before, but the efficiency is 95 %. As the reader can see, these two figures produce a very different output. While the plot on the right (which is similar to the masked index plot of OLS residuals) highlights the presence of a unit (number 43) which is on the boundary of the simultaneous confidence band, only the plot on the left (based on a smaller efficiency) suggests that there may be six atypical units (9, 21 30, 31, 38 47), which are indeed the masked outliers.

## 4 Conclusions

In this paper we have provided a critical review of some popular rules for identifying multivariate outliers and we have studied their behaviour both under the null hypothesis of no outliers and under different contamination schemes. Our results

show that the actual size of the outlier tests based on multivariate S and MM estimators using Tukey's Biweight function and relying on the $\chi^2_v$ distribution is larger than the nominal value, but the extent of the difference is often not dramatic. The effect of the many tuning constants required for their computation is also seen to be minor, except perhaps efficiency in the case of MM estimators. Therefore, when applied to uncontaminated data, these rules can be considered as a viable alternative to multivariate detection methods based on trimming and requiring more sophisticated distributional approximations.

However, smoothness of Tukey's Biweight function becomes a trouble when power is concerned, especially if the contamination rate is large and the number of dimensions grows. In such instances our simulations clearly show the advantages of trimming over S and MM estimators. In particular, the flexible trimming approach ensured by the Forward Search is seen to greatly outperform the competitors, even the most liberal ones, in almost all our simulation scenarios and is thus to be recommended.

**Acknowledgements** The authors thank the financial support of the project MIUR PRIN MISURA - Multivariate models for risk assessment.

## References

Atkinson, A. C., & Riani, M. (2000). *Robust diagnostic regression analysis*. New-York: Springer.

Atkinson, A. C., Riani, M., & Cerioli, A. (2004). *Exploring multivariate data with the forward search*. New York: Springer.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association, 105*, 147–156.

Cerioli, A., & Farcomeni, A. (2011). Error rates for multivariate outlier detection. *Computational Statistics and Data Analysis, 55*, 544–553.

Cerioli, A., Riani, M., & Atkinson, A. C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statistics and Computing, 19*, 341–353

Cerioli, A., Atkinson, A. C., & Riani, M. (2011a). Some perspectives on multivariate outlier detection. In S. Ingrassia, R. Rocci, & M. Vichi (Eds.), *New perspectives in statistical modeling and data analysis* (pp. 231–238). Berlin/Heidelberg: Springer.

Cerioli, A., Riani, M., & Torti, F. (2011b). Accurate and powerful multivariate outlier detection. *58th congress of ISI*, Dublin.

Hadi, A. S., Rahmatullah Imon, A. H. M., & Werner, M. (2009). Detection of outliers. *WIREs Computational Statistics, 1*, 57–70.

Hubert, M., Rousseeuw, P. J., & Van aelst, S. (2008). High-breakdown robust multivariate methods. *Statistical Science, 23*, 92–119.

Maronna, R. A., Martin, D. G., & Yohai, V. J. (2006). *Robust statistics*. New York: Wiley.

Morgenthaler, S. (2006). A survey of robust statistics. *Statistical Methods and Applications, 15*, 271–293 (Erratum 16, 171–172).

Perrotta, D., Riani, M., & Torti, F. (2009). New robust dynamic plots for regression mixture detection. *Advances in Data Analysis and Classification, 3*, 263–279.

Pison, G., Van aelst, S., & Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika, 55*, 111–123.

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society B, 71*, 447–466.

Riani, M., Torti, F., & Zani, S. (2011). Outliers and robustness for ordinal data. In R. S. Kennet & S. Salini (Eds.), *Modern analysis of customer satisfaction surveys: with applications using R*. Chichester: Wiley.

Riani, M., Cerioli, A., & Torti, F. (2012). A new look at consistency factors and efficiency of robust scale estimators. Submitted.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rousseeuw, P. J. & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*, 212–223.

Salibian-barrera, M., Van Aelst, S., & Willems, G. (2006). Principal components analysis based on multivariate mm estimators with fast and robust bootstrap. *Journal of the American Statistical Association, 101*, 1198–1211.

Todorov, V., & Filzmoser, P. (2009). An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software, 32*, 1–47.

Wilks, S. S. (1963). Multivariate statistical outliers. *Sankhya A, 25*, 407–426.