Multivariate Outlier Detection With High-Breakdown Estimators

Andrea CERIOLI

In this paper we develop multivariate outlier tests based on the high-breakdown Minimum Covariance Determinant estimator. The rules that we propose have good performance under the null hypothesis of no outliers in the data and also appreciable power properties for the purpose of individual outlier detection. This achievement is made possible by two orders of improvement over the currently available methodology. First, we suggest an approximation to the exact distribution of robust distances from which cut-off values can be obtained even in small samples. Our thresholds are accurate, simple to implement and result in more powerful outlier identification rules than those obtained by calibrating the asymptotic distribution of distances. The second power improvement comes from the addition of a new iteration step after one-step reweighting of the estimator. The proposed methodology is motivated by asymptotic distributional results. Its finite sample performance is evaluated through simulations and compared to that of available multivariate outlier tests.

KEY WORDS: Minimum covariance determinant estimator; Multiple outliers; Reweighting; Robust distance; Size and power.

1. INTRODUCTION

We are concerned with outlier detection in a multivariate model with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ on v dimensions. We estimate the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ based on a sample of n observations $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$. It is well known that the classical least-squares estimates can be strongly distorted and even completely break down if the sample contains outliers. For this reason highly robust estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ should be used in place of the classical ones (Rousseeuw and van Zomeren 1990; Becker and Gather 1999; Peña and Prieto 2001; Cuesta-Albertos, Matràn, and Mayo-Iscar 2008).

With high-breakdown methods robust estimation and outlier detection are two essentially equivalent tasks: given robust estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the outliers in $\mathbf{y}$ are revealed by their large distances from this robust fit. Formal identification rules are commonly obtained by comparing the squared robust distances to the χ^2_v distribution, since their exact distribution is unknown for finite samples, and by testing each observation at a fixed nominal size α not depending on n. Usually $0.01 \leq \alpha \leq 0.05$, with $\alpha = 0.025$ being perhaps the most popular choice (e.g., Hubert, Rousseeuw, and Van Aelst 2008; Willems, Joe, and Zamar 2009). However, there are two potential shortcomings in this procedure. The first one is the inadequacy of the χ^2_v approximation to obtain reliable cut-off values when the data contain no outliers. Evidence of this behavior is now well documented even in moderately large samples, especially when the number of dimensions increases (Becker and Gather 2001; Hardin and Rocke 2005; Cerioli, Riani, and Atkinson 2009; Riani, Atkinson, and Cerioli 2009). The other potential pitfall is the absence of simultaneity adjustments when comparing n distances to the relevant cut-off. Multiplicity corrections are considered in the seminal work of Wilks (1963) and in some of its extensions (e.g., Caroni and Prescott 1992), but they are not taken into account in most high-breakdown developments. A few notable exceptions are the simultaneous outlier detection rules of Becker and Gather (1999) and Cerioli, Riani, and Atkinson (2009), which, however, rely on low-power Bonferroni-type adjustments of the asymptotic χ^2_v distribution of robust distances.

The main consequence of the two shortcomings described above is a large number (often higher than $n\alpha$) of false outliers declared in any "good" dataset. Swamping may be an acceptable side effect in some situations, for example, when the user is almost certain that the data contain outliers. If this is the case, it is sensible to magnify the power of the diagnostic procedure in order to "find all the outliers that matter" (Hubert, Rousseeuw, and Van Aelst 2008, p. 92), even at the expense of an increase in the number of false discoveries. However, there are applications where even a moderate amount of swamping may have disastrous consequences. For instance, in the analysis of bivariate trade data arising in the European Union market outliers are of paramount importance because some of them may correspond to fraudulent transactions (Arsenis, Perrotta, and Torti 2005; Riani et al. 2008). Since there are hundreds of transactions to be inspected over thousands of markets, ignoring the multiplicity of tests and relying on liberal distributional assumptions would lead to a plethora of false signals for antifraud services, thus making substantial investigation of possible frauds impractical. Similar problems arise in multivariate quality control (Vargas N. 2003; Boente and Farall 2008) and in microarray data analysis (Brettschneider et al. 2008).

The goal of this paper is to develop multivariate outlier tests based on high-breakdown estimators with good performance under the null hypothesis of no outliers in the data and also with appreciable power properties for the purpose of individual outlier detection. In particular, we focus on the reweighted version of the Minimum Covariance Determinant (MCD) estimator of Hawkins and Olive (1999) and Rousseeuw and Van Driessen (1999). This estimator shares the same high-breakdown properties of the "raw" MCD estimator but has greater efficiency thanks to one-step reweighting (Croux and Haesbroeck 1999; Lopuhaä 1999). Furthermore, it benefits from the availability of fast software implemented in different languages, which makes it one of the most popular choices in applied robust statistics.

Andrea Cerioli is Professor, Dipartimento di Economia, Sezione di Statistica e Informatica, Università di Parma, Via Kennedy 6, 43100 Parma, Italy (E-mail: *andrea.cerioli@unipr.it*). The author expresses his gratitude to three anonymous reviewers for insightful comments that led to many improvements in the article. The author also thanks Marco Riani and Anthony C. Atkinson for helpful discussions on previous drafts of this work.

Journal of the American Statistical Association
March 2010, Vol. 105, No. 489, Theory and Methods
DOI: 10.1198/jasa.2009.tm09147

Our goal in this paper is achieved by two orders of improvement over the currently available methodology. First, we suggest an approximation to the exact distribution of robust distances computed from the reweighted MCD estimator, from which accurate cut-off values can be obtained even in small samples. These thresholds are virtually as accurate as those obtained by approximating the unknown distribution of the squared robust distances through Monte Carlo simulation (Cerioli, Riani, and Atkinson 2009), but they are simpler to implement and result in more powerful outlier identification rules. Our second improvement is the addition of a new iteration step after reweighting. The new step allows us to control the family-wise error rate of the n outlier tests only when it is really needed, that is, when all the data come from the null distribution, thus increasing the probability of finding contaminated observations when they are present. Like in the standard high-breakdown framework, we accept to tolerate some degree of swamping, with at most $(n-1)\alpha$ false outliers, but only when we are confident that some contamination is present in the data. The resulting procedure has power curves which are close or even superior to those of the potentially very liberal MCD-based tests.

The outline of the paper is as follows. In Section 2 we set up the problem, we motivate our approximation to the exact distribution of robust reweighted distances with asymptotic distributional results and we show simulation evidence of the good null performance of the resulting outlier test. In Section 3 we suggest the iterated reweighted MCD procedure, which ensures an increase in power when there is evidence of contamination. Simulation is used in Section 4 to assess the power properties of our procedures and to compare them with the liberal reweighted MCD test currently in use. The paper ends with an application in Section 5 and some closing remarks in Section 6.

2. MULTIVARIATE OUTLIER DETECTION WITH THE REWEIGHTED MCD

2.1 Set Up

The MCD subset is defined to be the subsample of h observations, with $n/2 \le h < n$, whose covariance matrix has the smallest determinant. Let $y_{(\mathrm{MCD})} = \{i_1, \ldots, i_h\}$ denote the indices of the observations belonging to the MCD subset. The MCD estimate of location is the average of the MCD subset,

$$\hat{\boldsymbol{\mu}}_{(\mathrm{MCD})} = \frac{1}{h} \sum_{i \in y_{(\mathrm{MCD})}} \mathbf{y}_i, \tag{1}$$

whereas the MCD estimate of scatter is proportional to the dispersion matrix of this subset:

$$\hat{\boldsymbol{\Sigma}}_{(\mathrm{MCD})} = \frac{k_{\mathrm{MCD}}(h,n,v)}{h-1} \times \sum_{i \in y_{(\mathrm{MCD})}} \big(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{(\mathrm{MCD})}\big)\big(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{(\mathrm{MCD})}\big)'. \tag{2}$$

The proportionality constant $k_{\mathrm{MCD}}(h,n,v)$, depending on the values of h, n, and v, serves the purpose of making $\hat{\boldsymbol{\Sigma}}_{(\mathrm{MCD})}$ both consistent and unbiased when each $\mathbf{y}_i \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (Croux and Haesbroeck 1999; Pison, Van Aelst, and Willems 2002). The finite-sample formula for computing $k_{\mathrm{MCD}}(h,n,v)$ is given in Equation (17). For outlier detection, we take the value of h yielding the maximum possible breakdown point, that is,

$$h = \left\lfloor \frac{n+v+1}{2} \right\rfloor \approx \frac{n}{2}, \tag{3}$$

where $\lfloor \cdot \rfloor$ denotes the integer part. Another popular choice is $h \approx 0.75n$, which yields more efficient estimates at the expense of a reduced breakdown value.

To increase efficiency, a one-step reweighting scheme is often used in practice. Reweighted estimators are computed by giving weight 0 to observations for which the squared robust distance

$$d^2_{i(\mathrm{MCD})} = \big(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{(\mathrm{MCD})}\big)' \hat{\boldsymbol{\Sigma}}^{-1}_{(\mathrm{MCD})} \big(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{(\mathrm{MCD})}\big), \quad i = 1, \ldots, n, \tag{4}$$

exceeds a threshold value. The reweighted MCD (RMCD) estimates of location and scatter are then

$$\hat{\boldsymbol{\mu}}_{(\mathrm{RMCD})} = \frac{1}{m} \sum_{i=1}^{n} w_i \mathbf{y}_i \tag{5}$$

and

$$\hat{\boldsymbol{\Sigma}}_{(\mathrm{RMCD})} = \frac{k_{\mathrm{RMCD}}(m,n,v)}{m-1} \times \sum_{i=1}^{n} w_i \big(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{(\mathrm{RMCD})}\big)\big(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{(\mathrm{RMCD})}\big)', \tag{6}$$

where $w_i = 0$ if $d^2_{i(\mathrm{MCD})} > d^2_{(\mathrm{MCD})*}$, $w_i = 1$ otherwise, and $m = \sum_{i=1}^{n} w_i$. The suggested threshold (Rousseeuw and Van Driessen 1999) is the 0.975 quantile of the χ^2_v distribution

$$d^2_{(\mathrm{MCD})*} = \chi^2_{v,0.975}.$$

The scaling $k_{\mathrm{RMCD}}(m,n,v)$ guarantees consistency of the reweighted estimator and improves its small sample behavior, as does the corresponding factor in (2). The currently adopted formula for computing $k_{\mathrm{RMCD}}(h,n,v)$ is shown in Equation (20). An alternative proposal is suggested in Step 3 of Section 2.4.

2.2 Outlier Detection and Testing

The squared robust reweighted distances

$$d^2_{i(\mathrm{RMCD})} = \big(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{(\mathrm{RMCD})}\big)' \hat{\boldsymbol{\Sigma}}^{-1}_{(\mathrm{RMCD})} \big(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_{(\mathrm{RMCD})}\big), \quad i = 1, \ldots, n, \tag{7}$$

constitute the main tool of this paper for identifying multivariate outliers. The outlier detection problem is usually phrased in terms of testing the n null hypotheses

$$H_{0i} : \mathbf{y}_i \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, n. \tag{8}$$

We say that $\mathbf{y}_i$ is a "good" observation if it satisfies H_{0i}. Otherwise, it is contaminated.

There is some debate about which error rate should be controlled when testing each H_{0i}. Let R be the number of true null hypotheses (8) which are incorrectly rejected. Following Benjamini and Hochberg (1995), the two main quantities of interest are the per-comparison error rate (PCER) $E\{R/n\}$ and the family-wise error rate (FWER) $\Pr\{R \ge 1\}$. The prevalent

approach is to test individually each hypothesis (8) at level $0.01 \le \alpha \le 0.05$ (Hubert, Rousseeuw, and Van Aelst 2008), thus focusing on the requirement that $E\{R/n\}$ be controlled at that α. Under this framework, we expect to find a proportion α of *false outliers* when all the data come from the prescribed multivariate normal distribution. Controlling the PCER increases the probability of detecting truly contaminated observations, but results in a value of the FWER that is close to 1 even in moderate samples when all the observations satisfy the null (8). The user must then be prepared to declare at least one outlier (and often many more) in most datasets of realistic size, even when contaminated observations are not present.

On the other hand, the simultaneous rules of Becker and Gather (1999) and Cerioli, Riani, and Atkinson (2009) focus on the intersection hypothesis

$$H_{0s}: \{\mathbf{y}_1 \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\} \cap \{\mathbf{y}_2 \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\} \cap \cdots \cap \{\mathbf{y}_n \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})\} \quad (9)$$

that no outlier is present in the data. Given a cut-off d^2, the size of the test of (9) is

$$P\big(\max_{i=1}^{n} d^2_{i(\mathrm{RMCD})} > d^2 | H_{0s} \text{ is true}\big), \quad (10)$$

which represents the proportion of good *datasets* that are wrongly declared to contain outliers. The Bonferroni argument can be used to choose d^2 such that (10) is kept below a specified value γ between 0.01 and 0.05 when the distribution of $d^2_{i(\mathrm{RMCD})}$ is known. However, this argument is known to lead to conservative rules ensuring that

$$P(R \ge 1) = P\big(\max_{i=1}^{n} d^2_{i(\mathrm{RMCD})} > d^2\big) \le \gamma$$

not only when H_{0s} is true, but also under any configuration of good and contaminated observations (Hochberg and Tamhane 1987). In Section 3 we consider a more powerful way of performing multiple outlier detection. Before that, we introduce a simple but effective approximation to the exact distribution of the reweighted distances that provides reliable cut-off values for our basic outlier identification rule.

2.3 The Distribution of Reweighted Distances

Our basic simultaneous outlier identification rule is motivated by two useful distributional results. These results assume that the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in (9) are known and thus hold only asymptotically for the robust reweighted distances (7). Our conjecture is that they can also provide useful guidance in finite samples, given that a reliable approximation to the distribution of the "raw" MCD distances (4) is available.

Let

$$d_i^2 = (\mathbf{y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}), \qquad i = 1, \ldots, n \quad (11)$$

be the squared population Mahalanobis distance for observation $\mathbf{y}_i$. Define the weights

$$\omega_i = \begin{cases} 0 & \text{if } d_i^2 > \chi^2_{v,1-\delta} \\ 1 & \text{otherwise,} \end{cases} \quad (12)$$

where $\chi^2_{v,1-\delta}$ is the $1-\delta$ quantile of the χ^2_v distribution.

Our first distributional result is stated in Proposition 1 and refers to the null distribution of the distances of the observations for which $\omega_i = 1$. The proof is given in the Appendix.

Proposition 1. Under the null hypothesis (9), the conditional distribution of $\mathbf{y}_i$ given $\omega_i = 1$ is a truncated multivariate normal distribution with

$$E(\mathbf{y}_i | \omega_i = 1) = \boldsymbol{\mu} \qquad \text{and} \qquad \mathrm{Var}(\mathbf{y}_i | \omega_i = 1) = \kappa_\delta^{-1} \boldsymbol{\Sigma}, \quad (13)$$

where

$$\kappa_\delta = \frac{1-\delta}{P(\chi^2_{v+2} < \chi^2_{v,1-\delta})}.$$

The usefulness of Proposition 1 is that it suggests how to obtain approximately unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the observations $\mathbf{y}_i$ for which $\omega_i = 1$. Let $\omega_\delta = \sum_{i=1}^n \omega_i$,

$$\hat{\boldsymbol{\mu}}_\delta = \frac{1}{\omega_\delta} \sum_{i=1}^{n} \omega_i \mathbf{y}_i,$$

$$\hat{\boldsymbol{\Sigma}}_\delta = \frac{\kappa_\delta}{\omega_\delta - 1} \sum_{i=1}^{n} \omega_i (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_\delta)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_\delta)',$$

and

$$d^2_{i,\delta} = (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_\delta)' \hat{\boldsymbol{\Sigma}}_\delta^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_\delta). \quad (14)$$

Note that $E(\mathbf{y}_i | \omega_i = 1) = E(\omega_i \mathbf{y}_i)/(1-\delta)$, $\mathrm{Var}(\mathbf{y}_i | \omega_i = 1) = E\{\omega_i (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_\delta)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_\delta)'\}/(1-\delta)$ and $\omega_\delta/n = (1-\delta) + O_p(\frac{1}{\sqrt{n}})$. It thus follows from Proposition 1 that

$$E(\hat{\boldsymbol{\mu}}_\delta) \to \boldsymbol{\mu} \qquad \text{and} \qquad E(\hat{\boldsymbol{\Sigma}}_\delta) \to \boldsymbol{\Sigma}$$

as $n \to \infty$. It is a standard result that the distribution of the squared Mahalanobis distances for the observations that contribute to the computation of the classical unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is a scaled Beta. Therefore, conditioning on ω_δ, we use an analogous scaled Beta distribution to approximate the distribution of the squared reweighted distances (14) for the units that contribute to the computation of the asymptotically unbiased estimators $\hat{\boldsymbol{\mu}}_\delta$ and $\hat{\boldsymbol{\Sigma}}_\delta$:

$$d^2_{i,\delta} | \omega_\delta \approx \frac{(\omega_\delta - 1)^2}{\omega_\delta} \mathrm{Beta}\left(\frac{v}{2}, \frac{\omega_\delta - v - 1}{2}\right) \quad \text{if } \omega_i = 1.$$

Our second distributional result is stated in Proposition 2 and gives a hint on the null distribution of the reweighted distances of the units for which $\omega_i = 0$. Proposition 2, which is proved in the Appendix, is given in the univariate case, where the population Mahalanobis distance (11) becomes the squared standardized residual $(y_i - \mu)^2/\sigma^2$ and $\mathcal{E}(1-\delta)$ is the symmetric central part of the $\mathrm{N}(\mu, \sigma^2)$ distribution leaving probability $\delta/2$ in each tail. A multivariate extension could follow from the asymptotic representation of affine equivariant multivariate quantiles (e.g., Chakraborty 2001).

Proposition 2. Let $v = 1$. Suppose that $y_1 \sim \mathrm{N}(\mu, \sigma^2), \ldots, y_n \sim \mathrm{N}(\mu, \sigma^2)$ with μ and σ known and define the weights ω_i, $i = 1, \ldots, n$, as in (12). Take $0 < \epsilon < \delta/2$ and let $y_{(j)}$ be the jth sample order statistic, with $j = \lfloor (1-\epsilon)n \rfloor + 1$. Then,

$$\mathrm{cov}\big(\sqrt{n} y_{(j)}; \sqrt{n} \hat{\mu}_\delta\big) \to 0 \quad \text{as } n \to \infty.$$

An analogous result also holds with $j = \lfloor \epsilon n \rfloor + 1$. Therefore, Proposition 2 shows that the observations trimmed in the reweighting step are asymptotically uncorrelated to the trimmed mean $\hat{\mu}_\delta$, when the population parameters are known and the weights (12) are used. This conclusion is remarkably different from that obtained when considering $y_{(j)}$ and the sample mean $\bar{y} = \sum_i y_i/n$ (see, e.g., DasGupta 2008, p. 95). The same

argument can be used to show that $\operatorname{cov}(y_{(j)}; \hat{\sigma}_\delta^2) = o_p(\frac{1}{\sqrt{n}})$. Furthermore, $E(\hat{\sigma}_\delta^2|\omega_\delta) \approx \sigma^2$ by Proposition 1. The usefulness of Proposition 2 is thus to suggest that, conditioning on ω_δ, the squared distance of a trimmed observation should approximately behave like that of an observation independent of the parameter estimates. This distance has a scaled F distribution. Therefore, we take

$$d_{i,\delta}^2|\omega_\delta \approx \frac{(\omega_\delta - 1)v}{\omega_\delta - v} F_{v,\omega_\delta - v} \quad \text{if } \omega_i = 0. \tag{15}$$

The distributional results given in Propositions 1 and 2 refer to the squared reweighted distances (14) which take advantage of the population weights $\omega_i, i, \ldots, n$. Consistency of high-breakdown estimators (Butler, Davies, and Jhun 1993; Lopuhaä 1999) could be used to obtain analogous asymptotic results when the true parameter values are replaced by such estimates. However, our main goal is to derive a reliable finite-sample approximation to the distribution of the squared reweighted MCD distances (7). Our conjecture is that the distributional results for $d_{i,\delta}^2$ suggested by Propositions 1 and 2 still hold approximately true for $d_{i(\mathrm{RMCD})}^2$ in finite samples, provided that the weights are defined as follows:

$$w_i = \begin{cases} 0 & \text{if } d_{i(\mathrm{MCD})}^2 > D_{v,1-\delta} \\ 1 & \text{otherwise,} \end{cases} \tag{16}$$

where $D_{v,1-\delta}$ is the $1-\delta$ quantile of the finite-sample distribution of the squared MCD distances (4). In fact, under (16), each w_i will have the same Bernoulli distribution as the corresponding population weight ω_i, and $m \sim \mathrm{Bin}(n, 1-\delta)$ as ω_δ.

Unfortunately $D_{v,1-\delta}$ is unknown, but Hardin and Rocke (2005) provide an accurate approximation to it for the values of δ useful for the purpose of outlier detection. Monte Carlo simulation is used in the next section to show the adequacy of our conjecture with this approximation.

2.4 Outlier Identification in Finite Samples

Let α be the nominal size at which each individual hypothesis (8) is tested. Given the distributional results of Section 2.3, our finite sample reweighted MCD (FSRMCD for short) rule for outlier detection is summarized as follows.

Step 1. Choose h and compute the raw MCD estimators (1) and (2) with

$$k_{\mathrm{MCD}}(h, n, v) = \frac{h/n}{P(\chi_{v+2}^2 < \chi_{v,h/n}^2)} s_{\mathrm{MCD}}(h, n, v), \tag{17}$$

where $s_{\mathrm{MCD}}(h, n, v)$ is the small sample bias-correction factor for $\hat{\mathbf{\Sigma}}_{(\mathrm{MCD})}$ obtained by Pison, Van Aelst, and Willems (2002).

Step 2. Compute the weights (16), with $D_{v,1-\delta}$ the $1-\delta$ quantile of the scaled F distribution of Hardin and Rocke (2005, p. 938) obtained using their adjusted asymptotic method. Take $\delta = 0.025$.

Step 3. Compute the RMCD estimators (5) and (6) with $k_{\mathrm{RMCD}}(h, n, v) = \kappa_\delta$ as given in Proposition 1.

Step 4. Compute the squared reweighted distances (7) and test each observation at size α, using the distribution

$$d_{i(\mathrm{RMCD})}^2 \sim \frac{(m-1)^2}{m} \mathrm{Beta}\left(\frac{v}{2}, \frac{m-v-1}{2}\right) \quad \text{if } w_i = 1 \tag{18}$$

$$\sim \frac{m+1}{m} \frac{(m-1)v}{m-v} F_{v,m-v} \quad \text{if } w_i = 0. \tag{19}$$

We introduce the factor $(m+1)/m$ in (19) to allow for estimation of $\boldsymbol{\mu}$ in the squared distance $d_{i,\delta}^2$ of Equation (15) (Atkinson, Riani, and Cerioli 2004, p. 43).

Asymptotically, the null probability of rejecting the intersection hypothesis (9) under the FSRMCD rule is $\gamma = 1-(1-\alpha)^n$, like that of the multiplicity-adjusted RMCD procedure using $\chi_{v,1-\alpha}^2$ as a cut-off for each individual outlier test (Becker and Gather 1999). We now investigate its finite sample performance through Monte Carlo simulation. We also compare our FSRMCD rule with the results of the asymptotic RMCD methodology that uses $\chi_{v,1-\alpha}^2$ as a cut-off. We focus on the case $\alpha = 1-(1-0.01)^{1/n}$, corresponding to a nominal size $\gamma = 0.01$ for testing the intersection hypothesis (9). The experience from this simulation study is that the same conclusions hold for different choices of γ.

We estimate the actual size of the test of (9) by simulating 5000 independent n-dimensional samples $\mathbf{y}$ from the v-variate $\mathrm{N}(\mathbf{0}, \mathbf{I})$ distribution, for several values of n and v. The results are valid for any $\mathbf{y}_i \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ of the same dimensions thanks to affine invariance of robust reweighted Mahalanobis distances. Aiming at the highest possible breakdown value, our first choice is to take h as in (3). We obtain the MCD subset through the Fortran algorithm *FAST-MCD* of Rousseeuw and Van Driessen (1999), which is based on repeated iteration of the concentration step (see also Hawkins and Olive 1999). To provide comparison with the available methodology, we also tune the RMCD procedure with the same corrections computed in function *covMcd* of the R package *robustbase* (Maechler 2008), where $k_{\mathrm{MCD}}(h, n, v)$ is as in (17),

$$k_{\mathrm{RMCD}}(h, n, v) = \frac{m/n}{P(\chi_{v+2}^2 < \chi_{v,1-\delta}^2)} s_{\mathrm{RMCD}}(h, n, v) \tag{20}$$

and $s_{\mathrm{RMCD}}(h, n, v)$ is a small sample calibration factor for $\hat{\mathbf{\Sigma}}_{(\mathrm{RMCD})}$ (Pison, Van Aelst, and Willems 2002). The aim of these corrections is to guarantee the best finite sample control of the size of the resulting RMCD tests. For each procedure, the size of the test of the intersection hypothesis (9) is estimated as the proportion of simulated datasets for which the null H_{0s} is falsely rejected.

Table 1 summarizes our results. Each entry in the cells of this table should be close to 0.01. It is clear that our rule greatly outperforms the standard RMCD approach across all the values of n and v. The actual size of the proposed test is virtually coincident with the nominal size when $n \geq 90$ and still acceptable for the smallest sample sizes, even if v increases. This is a remarkable result in view of the sparsity of multivariate space with few observations per dimension. On the other hand, the results obtained under the χ_v^2 approximation are exceptionally bad, with sizes up to 99% and often larger than 10%, thus rendering the standard RMCD procedure unusable for the purpose of simultaneous outlier identification in most of the experimental conditions considered in Table 1.

Our FSRMCD rule exhibits moderate liberality only when $n = 40$ and v increases. In such extreme situations Rousseeuw and van Zomeren (1990, p. 649) predict that "*any* outlier method can get into trouble" and, as a rule of thumb, they recommend applying robust multivariate methods only when $n/v > 5$. The null performance of our method can be further

Table 1. Estimated size of the proposed FSRMCD rule and of the standard RMCD procedure for testing the intersection hypothesis (9) of no outliers at nominal size $\gamma = 0.01$. The cut-off of each individual test performed by the RMCD procedure is $\chi^2_{v,1-\alpha}$, with $\alpha = 1-(1-0.01)^{1/n}$. The cardinality h of the MCD subset is given by (3). Size is estimated on 5000 simulations for each combination of n and v

		$n=40$	$n=60$	$n=90$	$n=125$	$n=200$	$n=400$
$v=5$	FSRMCD	0.017	0.017	0.015	0.013	0.011	0.010
	RMCD	0.516	0.272	0.119	0.071	0.028	0.016
$v=10$	FSRMCD	0.054	0.025	0.014	0.012	0.012	0.008
	RMCD	0.992	0.837	0.389	0.156	0.061	0.021
$v=15$	FSRMCD	0.084	0.030	0.013	0.014	0.013	0.010
	RMCD	1.000	0.999	0.884	0.478	0.126	0.040

improved by the more conservative choice $\delta = 0.01$ in Step 2 and by defining

$$k_{\mathrm{RMCD}}(h,n,v) = \frac{m/n}{P(\chi^2_{v+2} < \chi^2_{v,1-\delta})}$$

in Step 3. The first choice reduces the amount of trimming and thus the small sample bias that $k_{\mathrm{RMCD}}(h,n,v)$ is not able to allow for in Step 3. Taking m/n in the numerator of $k_{\mathrm{RMCD}}(h,n,v)$ corrects for the actual proportion of units trimmed in Step 2, which in very small samples can be substantially higher than the nominal proportion $1-\delta$. With these choices the FSRMCD entries in the first column of Table 1 become 0.006, 0.020, and 0.034, respectively. The FSRMCD simultaneous sizes for $v=10$ and $v=15$ can thus be remarkably close to the hoped-for value even with a sample size as small as $n=40$. However, since this setting will generally reduce the power of the outlier detection rule, we recommend it only when n is very small.

Increasing the cardinality of the MCD subset does not alter our findings. Table 2 summarizes the results for the less severe trimming recommendation $h \approx 0.75n$. Again, to provide comparison, the precise value of h is calculated as in the R function *covMcd*. Although the sizes in Table 2 are generally smaller than the corresponding entries in Table 1, it is clear that the standard RMCD procedure is still unusable in most situations. A small benefit is observed for FSRMCD as well, when the ratio n/v is very unfavorable.

3. THE ITERATED REWEIGHTED MCD

We increase the power of the multiple outlier test described in Section 2.4 by adding a further iteration step to the FSRMCD procedure. The iterated reweighted MCD (IRMCD for short) aims at behaving like a simultaneous outlier identification rule under the null hypothesis of no outliers and like the standard RMCD test when the data are contaminated. Furthermore, the test takes advantage of the accurate distributional results obtained in Section 2.3. The IRMCD rule is defined as the FSRMCD procedure with the additional step that follows.

Step 5. Let $q_{i,1-\alpha}$ be the $1-\alpha$ quantile of the reference distribution for $d^2_{i(\mathrm{RMCD})}$ in Step 4. If

$$\max_{i=1}^{n} I\left(d^2_{i(\mathrm{RMCD})} > q_{i,1-\alpha}\right) = 0, \tag{21}$$

then accept H_{0s}. Otherwise, test each individual hypothesis (8) using $q_{i,1-\gamma}$ as a cut-off, with $\gamma = 1-(1-\alpha)^n$, and declare $\mathbf{y}_i$ to be an outlier if

$$d^2_{i(\mathrm{RMCD})} > q_{i,1-\gamma}. \tag{22}$$

Condition (21) ensures that the IRMCD and the FSRMCD procedures have the same size for testing the null hypothesis (9) of no outliers in the data. On the other hand, (22) increases the probability of correctly identifying contaminated observations when they are present. The idea of performing n individual tests only when a simultaneous hypothesis is rejected is certainly not new and dates back at least to Fisher's least significant difference method (Hochberg and Tamhane 1987). It fits naturally in the RMCD setting as it only involves one additional iteration step. More importantly, the IRMCD rule follows the commonly accepted high-breakdown philosophy that admits a small amount of false outliers. The important contribution is that, with the IRMCD, swamping occurs only when there is substantial evidence of contamination.

Table 2. The entries represent the same quantities as in Table 1, but now the cardinality of the MCD subset is $h \approx 0.75n$

		$n=40$	$n=60$	$n=90$	$n=125$	$n=200$	$n=400$
$v=5$	FSRMCD	0.019	0.015	0.015	0.013	0.010	0.010
	RMCD	0.240	0.113	0.058	0.041	0.023	0.016
$v=10$	FSRMCD	0.035	0.018	0.015	0.012	0.013	0.010
	RMCD	0.848	0.465	0.182	0.090	0.049	0.022
$v=15$	FSRMCD	0.053	0.019	0.014	0.014	0.015	0.011
	RMCD	1.000	0.928	0.542	0.237	0.086	0.034

4. POWER OF OUTLIER IDENTIFICATION RULES

We now compare the power of the outlier detection rules suggested in this paper, FSRMCD and IRMCD, with that of other MCD-based procedures with similar size properties and also with that of the potentially very liberal RMCD approach using cut-off values from the χ^2_v distribution. The alternative finite-sample procedures based on the MCD that we consider are the Hardin and Rocke (2005) scaled-F test using the squared raw MCD distances (4) (HR), and the calibrated-χ^2_v test of Cerioli, Riani, and Atkinson (2009) based on the reweighted distances (7) (RMCDCAL). Both these procedures are adjusted for simultaneity and have good control of the size γ of the test of the intersection hypothesis (9) for the values of n and v considered in our power simulations. With regard to the standard RMCD techniques, we compute both the multiplicity-adjusted test with cut-off $\chi^2_{v,1-\alpha}$, already described in Table 1 (RMCD), and the individual RMCD test commonly adopted in practice, having $\chi^2_{v,1-\gamma}$ as a cut-off (RMCD_ind). A summary of the outlier detection rules compared in our power simulations is given in Table 3.

We generate 5000 n-dimensional samples for each combination of n and v. For a specified contamination rate $\tau < 0.5$, each dataset is composed of $n(1-\tau)$ observations from $\mathrm{N}(\mathbf{0}, \mathbf{I})$ and $n\tau$ observations from a v-variate location-shift model with constant contamination on all variables, $\mathbf{y}_i \sim \mathrm{N}(\lambda\mathbf{e}, \mathbf{I})$, where λ is a positive scalar and $\mathbf{e}$ is a column-vector of ones. Power is defined as the proportion of contaminated observations which are correctly named as outliers. We take $\gamma = 0.01$, the findings for different nominal sizes being similar. We give the results only for the maximum breakdown coverage $h = \lfloor (n+v+1)/2 \rfloor$, since the more efficient choice $h \approx 0.75n$ of Table 2 does not affect our conclusions. The only procedure that experiences a major increase in power when $h \approx 0.75n$ is the raw-MCD test HR, but the ranking of the best-performing rules remains unchanged.

Our first experimental setting is $n = 60$ and $v = 5$, a situation where the standard RMCD test should not be used for simultaneous outlier detection due to its very liberal behavior, even after adjusting for multiplicity. Figure 1 shows the power curves of the different tests under the location-shift contamination model as $\lambda > 0$ increases, respectively when $\tau = 0.05$ and $\tau = 0.20$. To provide comparison also under the null hypothesis, the power values for $\lambda = 0$ are defined to be the corresponding sizes in Table 1. Not surprisingly, the individual test RMCD_ind has size close to 1 and leads to wrongly reject the null hypothesis (9) in almost all the simulated good

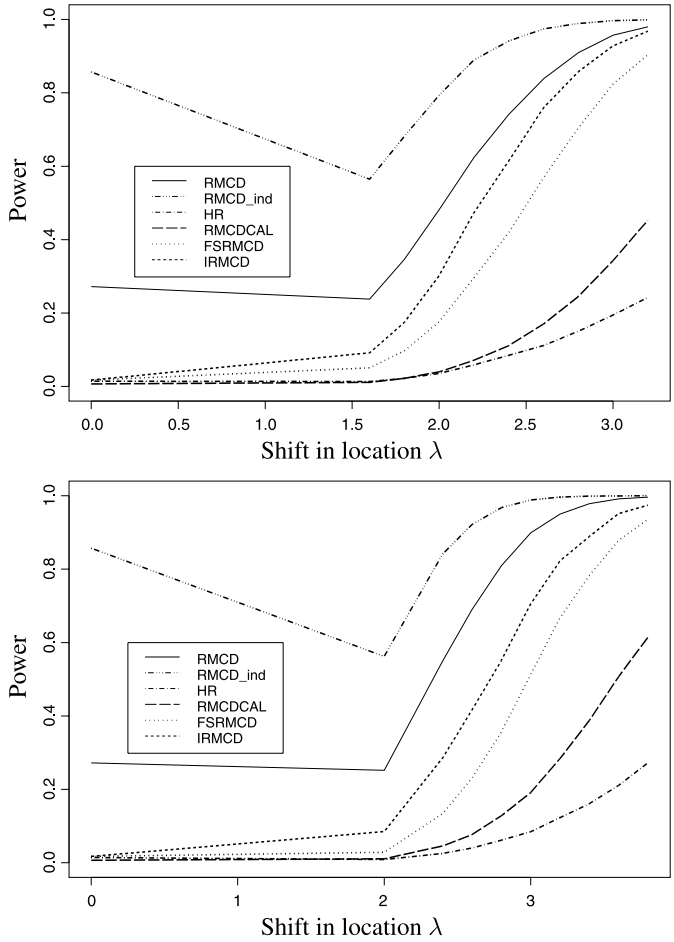

Figure 1. Power of MCD-based tests under a multivariate location-shift contamination model, for $n = 60$ and $v = 5$. The nominal size for testing the simultaneous hypothesis (9) is $\gamma = 0.01$. Upper panel: $\tau = 0.05$. Lower panel: $\tau = 0.20$. The cardinality h of the MCD subset is given by (3). Power is estimated on 5000 simulations for each value of τ and λ.

Table 3. MCD-based outlier detection rules computed for power comparison

Acronym	Synthetic description
RMCD	Squared reweighted distances $d^2_{i(\mathrm{RMCD})}$ with corrections (17) and (20); asymptotic χ^2_v distribution with simultaneity adjustment
RMCD_ind	Squared reweighted distances $d^2_{i(\mathrm{RMCD})}$ with corrections (17) and (20); asymptotic χ^2_v distribution without simultaneity adjustment
HR	Squared "raw" MCD distances $d^2_{i(\mathrm{MCD})}$ with correction (17); scaled F distribution of Hardin and Rocke (2005) with simultaneity adjustment
RMCDCAL	Squared reweighted distances $d^2_{i(\mathrm{RMCD})}$ with corrections (17) and (20); cut-off values estimated by Cerioli, Riani, and Atkinson (2009) with simultaneity adjustment
FSRMCD	Finite sample reweighted-MCD detection rule of Section 2.4; scaled Beta distribution (18) or scaled F distribution (19) with simultaneity adjustment
IRMCD	Iterated reweighted-MCD detection rule of Section 3

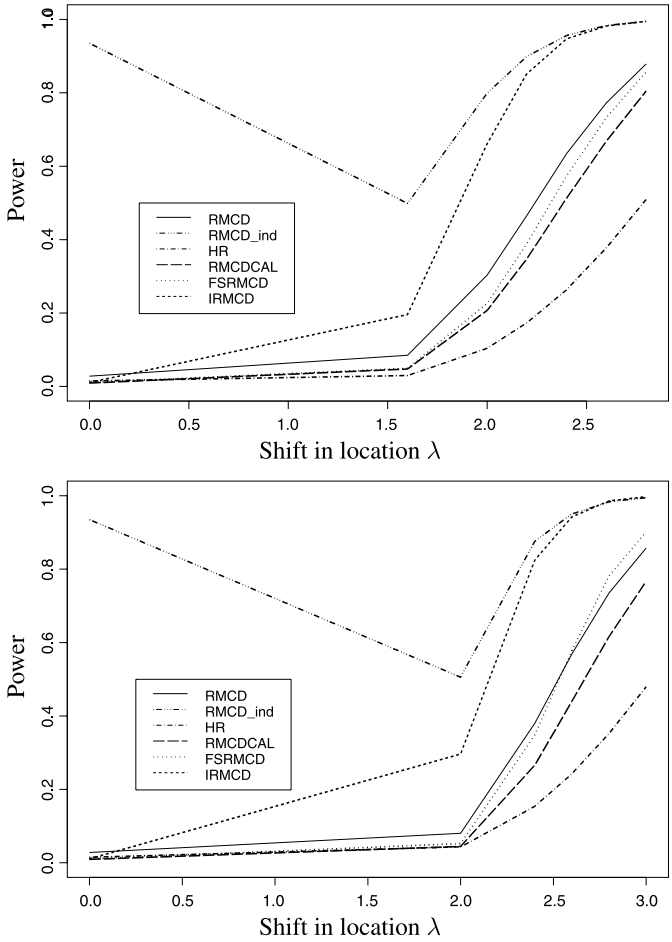

Figure 2. Power of MCD-based tests under a multivariate location-shift contamination model, for $n = 200$ and $v = 5$. The nominal size for testing the simultaneous hypothesis (9) is $\gamma = 0.01$. Upper panel: $\tau = 0.05$. Lower panel: $\tau = 0.20$. The cardinality h of the MCD subset is given by (3). Power is estimated on 5000 simulations for each value of τ and λ.

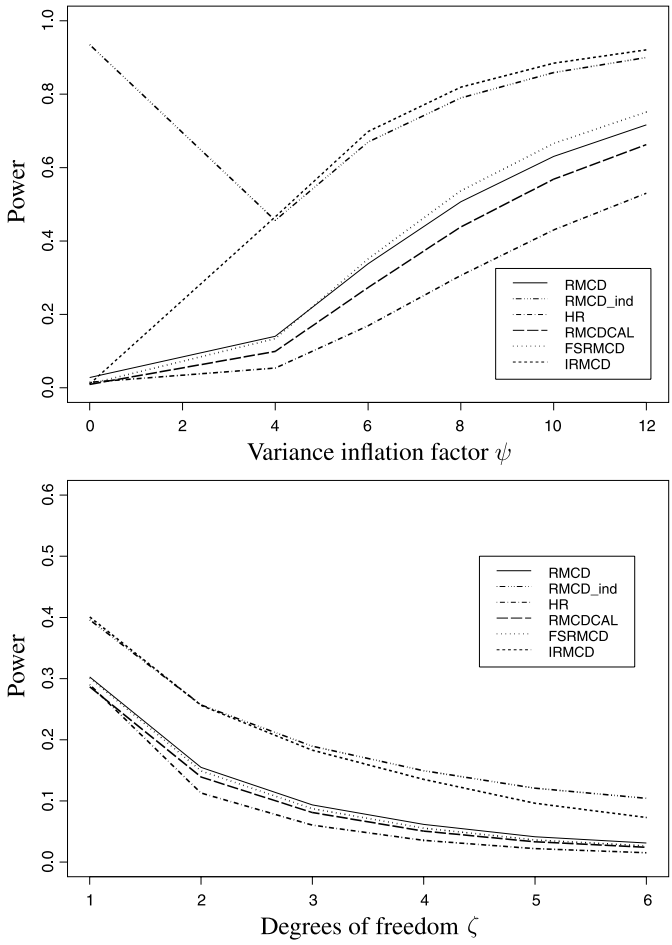

Figure 3. Power of MCD-based tests under alternative multivariate contamination models, for $n = 200$, $v = 5$, and $\tau = 0.20$. The nominal size for testing the simultaneous hypothesis (9) is $\gamma = 0.01$. Upper panel: radial contamination. Lower panel: multivariate t contamination. The cardinality h of the MCD subset is given by (3). Power is estimated on 5000 simulations for each value of ψ and ζ.

datasets. However, if λ in the location-shift model is small, the proportion of contaminated observations which are detected by RMCD_ind is much lower than 1, yielding a descent in the corresponding power curve.

It is clear from the pictures in Figure 1 that the FSRMCD procedure greatly outperforms the other two tests with similar simultaneous size (HR and RMCDCAL). The additional iteration step of the IRMCD procedure provides a further gain in power. Despite the wide gap at the origin, the power curve of IRMCD is remarkably close to those of the very liberal RMCD and RMCD_ind tests and approaches them as λ increases.

Figure 2 depicts the results for our second experimental setting with location-shift contamination, $n = 200$ and $v = 5$. Although slightly liberal, the multiplicity-adjusted RMCD test can be considered an acceptable procedure for simultaneous outlier detection in this situation and indeed its performance is very close to that of FSRMCD. Instead, the power of the IRMCD test is much larger than that of RMCD. Furthermore, we see from the plots that the power curve of IRMCD rapidly converges to that of RMCD_ind as λ increases. The proposed method, which has good control of the FWER under (9), is thus virtually as powerful as the PCER-controlling test RMCD_ind under the contamination model.

It is also instructive to look at the behavior of the different methods in contamination models exhibiting a gradation of more to less central observations. The upper panel of Figure 3 shows power under a radial contamination scheme, where $n(1 - \tau)$ observations are generated from the null distribution $N(\mathbf{0}, \mathbf{I})$ and the remaining $n\tau$ ones from $N(\mathbf{0}, \psi\mathbf{I})$, with $\psi > 1$. The lower panel displays the results when the $n\tau$ contaminated observations are generated from a v-variate t distribution with $\zeta \geq 1$ degrees of freedom. In both instances $n = 200$, $v = 5$, $\tau = 0.20$, and $\gamma = 0.01$. Contaminated observations now have the same mean as those coming from the null model. It is thus not surprising to see that power is lower than under the location shift-model. Nevertheless, the relative performance of the different methods in Figure 3 is unchanged, with the IRMCD rule behaving as the PCER-controlling test RMCD_ind. It is worth noting that there are instances where IRMCD is even more powerful than RMCD_ind. This is a bonus of the adoption of the null correction factor κ_δ in Step 3 of Section 2.4, instead of the usual RMCD correction (20).

We conclude that for multivariate outlier detection the iterated reweighted MCD rule is the one to be preferred in most

Table 4. Percentage of noncontaminated observations declared to be outliers by different procedures, in the case $n = 200$, $v = 5$, and $\tau = 0.05$, under the location-shift contamination model. The nominal size for testing the simultaneous hypothesis (9) is $\gamma = 0.01$. The cardinality h of the MCD subset is given by (3). Estimates on 5000 simulations for each value of λ

	$\lambda = 0.0$	$\lambda = 2.0$	$\lambda = 2.2$	$\lambda = 2.4$	$\lambda = 2.6$	$\lambda = 2.8$
IRMCD	0.019%	0.762%	0.868%	0.930%	0.931%	0.919%
RMCD	0.014%	0.009%	0.009%	0.009%	0.009%	0.008%
RMCD_ind	1.386%	0.984%	0.959%	0.974%	0.950%	0.941%

situations among the high-breakdown procedures considered in this paper. It has the right size under the null hypothesis of no outliers even in small samples and it shows excellent power properties under different experimental settings and contamination schemes. Therefore, it is a considerable improvement over the available MCD-based methodologies. The little price one has to pay is a small proportion of false discoveries, bounded by γ, when the simultaneous hypothesis (9) is false. For instance, Table 4 reports the percentage of noncontaminated observations wrongly declared to be outliers at $\gamma = 0.01$, in the case $n = 200$, $v = 5$, and $\tau = 0.05$, under the location-shift contamination model. The IRMCD rule induces approximately the same proportion of false discoveries as RMCD_ind if $\lambda > 0$, but the effect of swamping is negligible when (9) is true.

5. DATA ANALYSIS

Flury and Riedwyl (1988) introduce data on six variables measuring the size and other features of 200 Swiss banknotes, 100 of which are genuine and 100 forged. We expect the quality control during the production of genuine notes to be much tighter than that on forged notes, so that the null hypothesis of no outliers should be reasonable for the first sample. On the other hand, the group of forged notes is known to be heterogeneous, perhaps due to the action of different forgers. We analyze the two groups of banknotes separately and we compare the performance of outlier detection methods for each of them.

The left-hand panel of Figure 4 shows the classical squared Mahalanobis distances, computed from the unbiased estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, for the sample of 100 genuine notes. As a cut-off, we report the $1 - (1 - 0.01)^{1/100}$ percentage point of the null scaled-Beta distribution of these distances. The right-hand panel gives the corresponding robust reweighted distances (7), with h as in (3). Three cut-off values are reported for the robust distances: the scaled-F cut-off of the FSRMCD procedure with $\alpha = 1 - (1 - 0.01)^{1/100}$, the simultaneous RMCD cut-off $\chi^2_{v,1-\alpha}$ with the same value of α, and the individual RMCD cut-off $\chi^2_{v,1-0.01}$ (RMCD_ind). We do not display the threshold provided by (18), which does not show any outlier among the banknotes for which $w_i = 1$. If H_{0s} is true, classical and robust distances should roughly give the same answer. Of the robust procedures, only FSRMCD leads to accept (9) and to rightly conclude that the data form a homogeneous sample, in good agreement with the classical Mahalanobis distances. On the contrary, the chi-square RMCD rules trim too much and discard some genuine notes as forgeries. This effect, which may be particularly severe for the PCER-controlling approach, could lead to considerable waste of money if the robust distances were used in an automatic procedure of banknote scanning for antifraud purposes.

Figure 5 repeats the analysis for the group of forged banknotes. Now the Mahalanobis distances completely break down due to heavy contamination. Instead, the presence of many outliers is revealed by the FSRMCD test. The additional iteration step of the IRMCD rule suggests the existence of a further borderline unit, which is also signalled by the RMCD_ind criterion. We conclude that in this contaminated sample the IRMCD

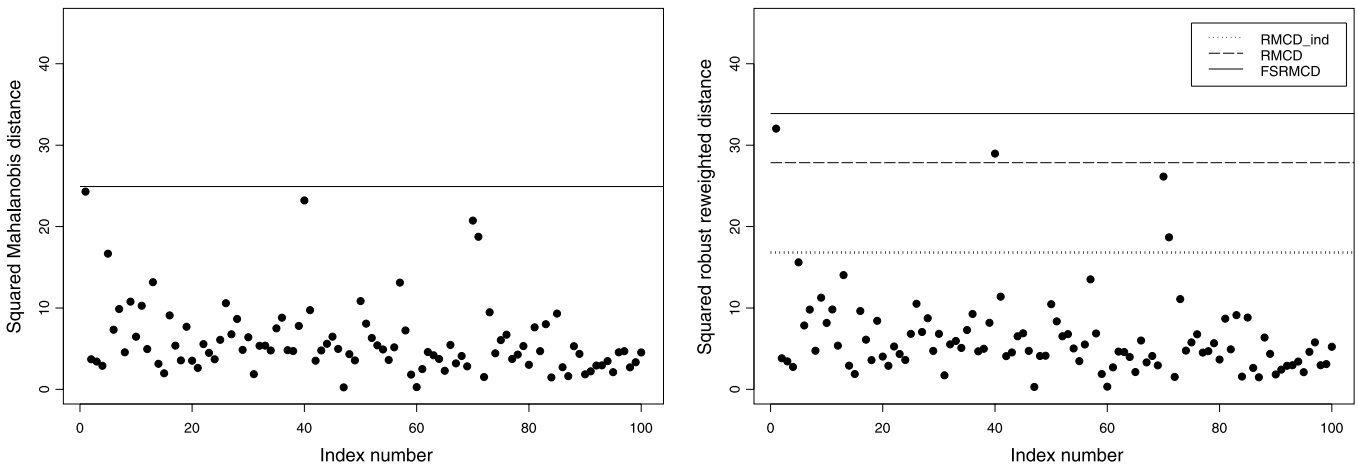

Figure 4. Genuine Swiss banknotes. Squared distances and cut-off values for multivariate outlier detection ($\gamma = 0.01$). Left-hand panel: Mahalanobis distances. Right-hand panel: robust reweighted distances with h as in (3).

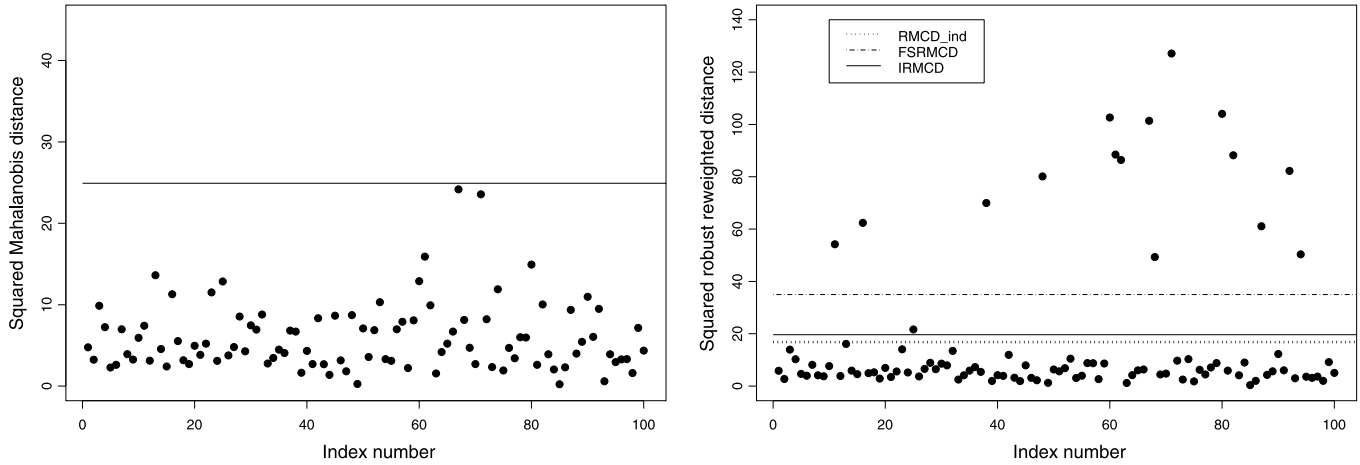

Figure 5. Forged Swiss banknotes. Squared distances and cut-off values for multivariate outlier detection ($\gamma = 0.01$). Left-hand panel: Mahalanobis distances. Right-hand panel: robust reweighted distances with h as in (3).

procedure has the same diagnostic power as repeated testing of the n individual hypotheses (8) without multiplicity adjustments. Therefore, our method proves to be as effective as the currently available high-breakdown methodology when the data contain outliers, but has a much more sensible behavior on uncontaminated datasets. Also note that Willems, Joe, and Zamar (2009) obtain an indication of 20 outliers for the group of forged notes by use of the liberal $\chi^2_{v,0.975}$ cut-off for the squared raw MCD distances (4), and reach conclusions similar to ours only after application of a number of supplementary diagnostic tools.

6. CONCLUSION

We are interested in tests for multiple outlier detection with good size properties when the data come from the multivariate normal distribution and with high power against contamination. Our results show that the currently available high-breakdown methodology (Hubert, Rousseeuw, and Van Aelst 2008), which relies on asymptotic arguments, fails to meet the first goal even in relatively large samples. We suggest two novel multivariate outlier detection rules, FSRMCD and IRMCD, that solve this problem. Our procedures are based on accurate finite-sample distributional results for the robust distances. Therefore, they are able to attain the nominal size even in otherwise problematic situations with $n < 200$. When outliers are present, both FSRMCD and IRMCD are considerably more powerful than the few available tests with similar size properties (Hardin and Rocke 2005; Cerioli, Riani, and Atkinson 2009). We conclude that, among the high-breakdown procedures considered in this paper, our proposals provide the best balance between size and power and are thus to be recommended.

The FSRMCD and IRMCD rules have the same behavior under the null hypothesis of no outliers. They differ in their attitude towards swamping. FSRMCD aims at controlling the number of false outliers for any distribution that could have generated the data. IRMCD tolerates a higher degree of swamping, but only when there is substantial evidence of contamination. The bonus of the latter approach is a considerable gain in power. Indeed, we have seen that the power of IRMCD is comparable to those of the potentially very liberal MCD-based tests currently in use. However, there may be situations where the acceptable degree of swamping could depend on the number of outliers found and not only on evidence of contamination. In these situations, our distributional results are still valid and can be used to develop outlier detection rules that focus on alternative error rates, like the false discovery rate of Benjamini and Hochberg (1995) and its extensions. Multivariate outlier tests based on the FDR criterion are the subject of ongoing research. Preliminary simulation evidence shows that these tests will have power curves intermediate between those of FSRMCD and IRMCD.

APPENDIX

Proof of Proposition 1

Under (9), the conditional distribution of $\mathbf{y}_i|\omega_i = 1$ is the distribution of a v-variate normal random vector $\mathbf{y} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ truncated outside the set

$$\mathcal{E}(1-\delta) = \{\mathbf{y} : (\mathbf{y}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y}-\boldsymbol{\mu}) \le \chi^2_{v,1-\delta}\}.$$

To this distribution we apply Tallis (1963) results on elliptical truncation and we obtain the parameters given in (13).

Proof of Proposition 2

Since μ and σ are known, we take $\mu = 0$ and $\sigma = 1$ without loss of generality. It is well known that

$$\sqrt{n}\big(y_{(j)} - F_n^{-1}(1-\epsilon)\big) \to 0$$

almost surely, where $F_n(\cdot)$ is the sample distribution function. We thus apply the Bahadur representation of $F_n^{-1}(1-\epsilon)$ to $y_{(j)}$:

$$y_{(j)} = \frac{1}{n}\sum_{l=1}^{n} z_l + o_p\left(\frac{1}{\sqrt{n}}\right),$$

where

$$z_l = z_{1-\epsilon} + \frac{(1-\epsilon) - I(y_l \le z_{1-\epsilon})}{\phi(z_{1-\epsilon})},$$

$I(\cdot)$ is the indicator function and $z_{1-\epsilon}$ is the $1-\epsilon$ quantile of $\mathrm{N}(0,1)$, with density $\phi(z_{1-\epsilon})$. Neglecting terms of order $o_p(\frac{1}{\sqrt{n}})$, the covari-

ance of interest is

$$\operatorname{cov}\left(\frac{\sqrt{n}}{n}\sum_{l=1}^{n} z_l;\ \frac{\sqrt{n}}{\omega_\delta}\sum_{i=1}^{n} y_i\omega_i\right) = \frac{n}{n\omega_\delta}\sum_{l=1}^{n}\sum_{i=1}^{n}\sigma_{il}, \tag{A.1}$$

where

$$\sigma_{il} = E\left\{\frac{(1-\epsilon) - I(y_l \le z_{1-\epsilon})}{\phi(z_{1-\epsilon})} y_i I(d_i^2 \le \chi^2_{1,1-\delta})\right\} - E\left\{\frac{(1-\epsilon) - I(y_l \le z_{1-\epsilon})}{\phi(z_{1-\epsilon})}\right\} E\{y_i I(d_i^2 \le \chi^2_{1,1-\delta})\}.$$

For $i = 1, \ldots, n$, we have that

$$E\{y_i I(d_i^2 \le \chi^2_{1,1-\delta})\} = \int_{\mathcal{E}(1-\delta)} y_i \phi(y_i)\, dy_i = (1-\delta)\mu = 0, \tag{A.2}$$

by Tallis (1963) result on elliptical truncation. Applying (A.2) we obtain

$$\sigma_{il} = E\left\{\frac{(1-\epsilon) - I(y_l \le z_{1-\epsilon})}{\phi(z_{1-\epsilon})} y_i I(d_i^2 \le \chi^2_{1,1-\delta})\right\}.$$

We need to distinguish two separate cases in the evaluation of σ_{il}.

Case (a): $i \ne l$. The observations y_i and y_l are independent. Also the functions $\varphi_1(y_i) = y_i I(d_i^2 \le \chi^2_{1,1-\delta})$ and $\varphi_2(y_l) = I(y_l \le z_{1-\epsilon})$ are thus independent. The covariance term σ_{il} becomes

$$\sigma_{il} = \frac{(1-\epsilon)(1-\delta)}{\phi(z_{1-\epsilon})}\mu - \frac{1}{\phi(z_{1-\epsilon})} E\{\varphi_1(y_i)\} E\{\varphi_2(y_l)\} = 0, \tag{A.3}$$

applying (A.2) again.

Case (b): $i = l$. The covariance term now is

$$\sigma_{ii} = \frac{(1-\epsilon)(1-\delta)}{\phi(z_{1-\epsilon})}\mu - \frac{1}{\phi(z_{1-\epsilon})} E\{y_i I(y_i \le z_{1-\epsilon}) I(d_i^2 \le \chi^2_{1,1-\delta})\}. \tag{A.4}$$

Let $\mathcal{L}(1-\epsilon) = \{y : y \le z_{1-\epsilon}\}$. The expectation in (A.4) is the expectation of $y \sim \mathrm{N}(0, 1)$ over the intersection region

$$\mathcal{I}(1-\delta; 1-\epsilon) = \{y : y^2 \le \chi^2_{1,1-\delta}\} \cap \{y : y \le z_{1-\epsilon}\} = \mathcal{E}(1-\delta) \cap \mathcal{L}(1-\epsilon).$$

Note that $\mathcal{E}(1-\delta)$ is the intersection of two sets: $\mathcal{E}(1-\delta) = \mathcal{L}(1-\delta/2) \cap \mathcal{R}(\delta/2)$, with $\mathcal{L}(1-\delta/2) = \{y : y \le z_{1-\delta/2}\}$ and $\mathcal{R}(\delta/2) = \{y : y > z_{\delta/2}\}$. Therefore,

$$\mathcal{I}(1-\delta; 1-\epsilon) = \mathcal{L}(1-\delta/2) \cap \mathcal{R}(\delta/2) \cap \mathcal{L}(1-\epsilon) = \mathcal{E}(1-\delta)$$

because $\epsilon < \delta/2$ and $\mathcal{L}(1-\delta/2) \subset \mathcal{L}(1-\epsilon)$. The covariance term is

$$\sigma_{ii} = \frac{(1-\epsilon)(1-\delta)}{\phi(z_{1-\epsilon})}\mu - \frac{1}{\phi(z_{1-\epsilon})}\int_{\mathcal{E}(1-\delta)} y_i \phi(y_i)\, dy_i \tag{A.5}$$

$$= -\frac{\epsilon(1-\delta)}{\phi(z_{1-\epsilon})}\mu = 0. \tag{A.6}$$

Finally note that

$$\frac{\omega_\delta}{n} = (1-\delta) + O_p\left(\frac{1}{\sqrt{n}}\right)$$

because $\omega_\delta \sim \mathrm{Bin}(n, 1-\delta)$. Substituting (A.3) and (A.6) into Equation (A.1) and applying Slutski's theorem to the denominator gives the result.

[Received March 2009. Revised July 2009.]

REFERENCES

Arsenis, S., Perrotta, D., and Torti, F. (2005), "Price Outliers in EU External Trade Data," internal working document on work presented at the *Enlargement and Integration Workshop 2005*, Ispra, Italy, Joint Research Centre of the European Commission, available at *http://theseus.jrc.it/index.php?id=1298*. [147]

Atkinson, A. C., Riani, M., and Cerioli, A. (2004), *Exploring Multivariate Data With the Forward Search*, New York: Springer-Verlag. [150]

Becker, C., and Gather, U. (1999), "The Masking Breakdown Point of Multivariate Outlier Identification Rules," *Journal of the American Statistical Association*, 94, 947–955. [147,149,150]

——— (2001), "The Largest Nonidentifiable Outlier: A Comparison of Multivariate Simultaneous Outlier Identification Rules," *Computational Statistics and Data Analysis*, 36, 119–127. [147]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Ser. B*, 57, 289–300. [148,155]

Boente, G., and Farall, A. (2008), "Robust Multivariate Tolerance Regions," *Technometrics*, 50, 487–500. [147]

Brettschneider, J., Bolstad, B. M., Collin, F., and Speed, T. P. (2008), "Quality Assessment for Short Oligonucleotide Microarray Data," *Technometrics*, 50, 241–264. [147]

Butler, R. W., Davies, P. L., and Jhun, M. (1993), "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics*, 21, 1385–1400. [150]

Caroni, C., and Prescott, P. (1992), "Sequential Application of Wilks's Multivariate Outlier Test," *Applied Statistics*, 41, 355–364. [147]

Cerioli, A., Riani, M., and Atkinson, A. C. (2009), "Controlling the Size of Multivariate Outlier Tests With the MCD Estimator of Scatter," *Statistics and Computing*, 19, 341–353. [147-149,152,155]

Chakraborty, B. (2001), "On Affine Equivariant Multivariate Quantiles," *Annals of the Institute of Statistical Mathematics*, 53, 380–403. [149]

Croux, C., and Haesbroeck, G. (1999), "Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator," *Journal of Multivariate Analysis*, 71, 161–190. [147,148]

Cuesta-Albertos, J. A., Matràn, C., and Mayo-Iscar, A. (2008), "Trimming and Likelihood: Robust Location and Dispersion Estimation in the Elliptical Model," *The Annals of Statistics*, 36, 2284–2318. [147]

DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, New York: Springer. [149]

Flury, B., and Riedwyl, H. (1988), *Multivariate Statistics: A Practical Approach*, London: Chapman & Hall. [154]

Hardin, J., and Rocke, D. M. (2005), "The Distribution of Robust Distances," *Journal of Computational and Graphical Statistics*, 14, 910–927. [147,150,152,155]

Hawkins, D. M., and Olive, D. J. (1999), "Improved Feasible Solution Algorithms for High Breakdown Estimation," *Computational Statistics and Data Analysis*, 30, 1–11. [147,150]

Hochberg, Y., and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, New York: Wiley. [149,151]

Hubert, M., Rousseeuw, P. J., and Van Aelst, S. (2008), "High-Breakdown Robust Multivariate Methods," *Statistical Science*, 23, 92–119. [147,149,155]

Lopuhaä, H. P. (1999), "Asymptotics of Reweighted Estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 27, 1638–1665. [147,150]

Maechler, M. (2008), "robustbase: Basic Robust Statistics," R package version 0.4-3, available at *http://cran.r-project.org/*. [150]

Peña, D., and Prieto, F. J. (2001), "Multivariate Outlier Detection and Robust Covariance Matrix Estimation," *Technometrics*, 43, 286–310. [147]

Pison, G., Van Aelst, S., and Willems, G. (2002), "Small Sample Corrections for LTS and MCD," *Metrika*, 55, 111–123. [148,150]

Riani, M., Atkinson, A. C., and Cerioli, A. (2009), "Finding an Unknown Number of Multivariate Outliers," *Journal of the Royal Statistical Society, Ser. B*, 71, 447–466. [147]

Riani, M., Cerioli, A., Atkinson, A., Perrotta, D., and Torti, F. (2008), "Fitting Mixtures of Regression Lines With the Forward Search," in *Mining Massive Data Sets for Security*, eds. F. Fogelman-Soulié, D. Perrotta, J. Piskorski, and R. Steinberger, Amsterdam: IOS Press, pp. 271–286. [147]

Rousseeuw, P. J., and Van Driessen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, 41, 212–223. [147,148,150]

Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking Multivariate Outliers and Leverage Points," *Journal of the American Statistical Association*, 85, 633–651. [147,150]

Tallis, G. M. (1963), "Elliptical and Radial Truncation in Normal Samples," *Annals of Mathematical Statistics*, 34, 940–944. [155,156]

Vargas N., J. A. (2003), "Robust Estimation in Multivariate Control Charts for Individual Observations," *Journal of Quality Technology*, 35, 367–376. [147]

Wilks, S. S. (1963), "Multivariate Statistical Outliers," *Sankhya, Ser. A*, 25, 407–426. [147]

Willems, G., Joe, H., and Zamar, R. (2009), "Diagnosing Multivariate Outliers Detected by Robust Estimators," *Journal of Computational and Graphical Statistics*, 18, 73–91. [147,155]