Error rates for multivariate outlier detection

Andrea Cerioli[a,*], Alessio Farcomeni[b]

[a] *University of Parma Via Kennedy 6; 43100 Parma, Italy.*
[b] *Sapienza - University of Rome, piazzale Aldo Moro, 5; 00186 Roma, Italy*

Abstract

Multivariate outlier identification requires the choice of reliable cut-off points for the robust distances that measure the discrepancy from the fit provided by high-breakdown estimators of location and scatter. Multiplicity issues affect the identification of the appropriate cut-off points. It is described how careful choice of the error rate which is controlled during the outlier detection process can yield a good compromise between high power and low swamping, when alternatives to the Family Wise Error Rate are considered. Correspondingly, multivariate outlier detection rules based on the False Discovery Rate and the False Discovery Exceedance criteria are proposed. The properties of these rules are evaluated through simulation. The rules are then applied to real data examples. The conclusion is that the proposed approach provides a sensible strategy in many situations of practical interest.

Keywords: false discovery rate, false discovery exceedance, multiple outliers, reweighted MCD, masking and swamping

1. Introduction

With multivariate data, multiple outliers are revealed by their large distances from the robust fit provided by high-breakdown estimators of location and scatter (Hubert *et al.*, 2008). An important issue is the occurrence of multiplicity problems when outlier detection is set up in a statistical testing framework. Multiplicity arises because the candidate outliers are not known in advance and all the observations are tested in sequence starting from the most remote one. Different error rates may be of interest when performing multiple tests. The multiplicity problem has not been considered thoroughly in the literature about outlier detection, even if there are notable exceptions like Davies and Gather (1993) and Becker and Gather (1999), who define outward testing procedures and use Sidak correction to guarantee that the level of swamping is below a threshold.

The goal of this paper is to show how carefully choosing the error rate to be controlled in multiple outlier detection can provide a reasonable compromise between good performance under the null hypothesis of no outliers and high power under contamination. In particular, we focus on multivariate outlier detection rules based on the False Discovery Rate (FDR) of Benjamini and Hochberg (1995), on the False Discovery eXceedance (FDX) of Lehmann and Romano (2005) and van der Laan *et al.* (2004), and we compare the power of the resulting outlier tests with those of alternative procedures attaining the same nominal size. We also evaluate the positive FDR (pFDR) of the procedures (Storey, 2002, 2003). We conclude that controlling these error rates, especially the FDR, can be a sensible strategy for outlier identification in many situations of practical interest.

The rest of the paper is as follows: in the remainder of this section we briefly review the error rates which are of interest in multiple testing. In Section 2 we set out our strategies for FDR and FDX control, and for pFDR estimation, in multivariate outlier identification. The merits of these strategies are illustrated with a simulation study in Section 3 and on two motivating examples in Section 4.

1.1. Multiplicity control and outlier detection

Let y_i be a v-variate observation with mean vector μ and covariance matrix Σ. Our basic model explaining the genesis of y_i is a two-components mixture model of the kind: $y_i|z_i \sim F_{z_i}$ for some unobserved $z_i \in \{0, 1\}$. The clean observations arise from $F_0 \sim N(\mu, \Sigma)$, while the contaminated observations are those for which $z_i = 1$, with F_1 arbitrary. Outlier detection is stated in terms of testing n null hypotheses

$$H_{0i} : y_i \sim N(\mu, \Sigma), \qquad i = 1, \ldots, n. \tag{1}$$

Each test is performed by computing the squared robust distance

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

where $\tilde{\mu}$ and $\tilde{\Sigma}$ are high-breakdown estimators of μ and Σ. In this paper we take $\tilde{\mu}$ and $\tilde{\Sigma}$ to be the reweighted MCD (RMCD) estimators of Rousseeuw and Van Driessen (1999).

Suppose that there are M_0 clean observations and M_1 contaminated ones. R is the number of observations declared to be outliers, i.e. those for which (1) is rejected. Table 1 summarizes the outcome of the outlier detection process. The values of $N_{0|1}$ and $N_{1|0}$ determine the effects of *masking* and

Table 1: Outcome in testing n observations for outlyingness

	Null Hypotheses (1)		
	Not Rejected	**Rejected**	**Total**
True	$N_{0\|0}$	$N_{1\|0}$	M_0
False	$N_{0\|1}$	$N_{1\|1}$	M_1
Total	$n-R$	R	n

swamping, respectively. Furthermore, the quantities in Table 1 are used to define error rates, which are deemed to be under control when they are bound, before the experiment, to be below a threshold α.

Traditional methods in multiple testing involve control of the *Family Wise Error Rate* (FWER), defined as the probability of making one or more false rejections. There is a plethora of methods for FWER control, the simplest being Bonferroni correction, which consists in performing each individual test at level α/n. Another simple, but slightly more powerful, one-step procedure is Sidak correction, where each test is performed at level

$$\gamma = 1 - (1-\alpha)^{1/n}. \tag{3}$$

The observations selected after control of the FWER are all trusted to be outliers. The main drawback of FWER control is its low power. The consequences of FWER control may thus be close to those of masking.

A different approach is proposed by Benjamini and Hochberg (1995), who define the *False Discovery Rate* (FDR):

$$FDR = E\left[\frac{N_{1|0}}{R}|R>0\right]\Pr(R>0). \tag{4}$$

The FDR is the expected proportion of erroneously rejected hypotheses, if any. The method developed by Benjamini and Hochberg (1995) (BH) is a

stepwise procedure which proceeds by rejecting all tests corresponding to p-values below $\rho_i\alpha/n$, where ρ_i is the rank of the i-th p-value. A very similar error rate, the positive FDR (pFDR), is defined by Storey (2002, 2003) as

$$pFDR = E\left[\frac{N_{1|0}}{R}|R > 0\right], \tag{5}$$

thus restricting to the cases in which there is at least one rejection. The pFDR has a nice Bayesian interpretation (Storey, 2003). It can be directly controlled or, as we do in this paper, it can be estimated to further evaluate the performance of any testing procedure. The pFDR is estimated by

$$\widehat{pFDR} = \frac{\hat{a}p_{(r)}}{r(1-(1-p_{(r)})^n)}, \tag{6}$$

where $r > 0$ denotes the observed value of R, $p_{(r)}$ is the largest p-value associated with rejected tests and $\hat{a}$ is an estimator of the number of true null hypotheses. In this paper we use the Schweder and Spjøtvoll (1982) estimator, and set $\hat{a} = 2(n - \tau_{0.5})$, where $\tau_{0.5}$ denotes the count of p-values smaller than or equal to 0.5.

Both (4) and (5) are based on an expectation, whereas the actual proportion of false discoveries may be larger than α. Therefore, Lehmann and Romano (2005) and van der Laan *et al.* (2004) independently define the False Discovery eXceedance (FDX) as the probability of the false discovery proportion being above a threshold, that is,

$$FDX = \Pr\left(\frac{N_{1|0}}{R \vee 1} > c\right), \tag{7}$$

where typically, and also in this paper, $c = 0.1$. We control the FDX using the Lehmann and Romano (2005) (LR) procedure, which rejects all tests

corresponding to p-values below $(\lfloor \rho_i c \rfloor + 1)\alpha/(n + \lfloor r_i c \rfloor + 1 - r_i)$, where $\lfloor \cdot \rfloor$ is the integer part.

There is a plethora of other available methods, and error rates, for a review of which we refer to Farcomeni (2008). An important feature for our purposes is that Bonferroni and Sidak corrections provide strong control of the FWER, which is bounded no matter the number and the configuration of outliers. Instead, the procedures based on the FDR and FDX ensure weak control of the FWER, which is then bounded only under the complete null hypothesis of no outliers

$$H_0 : \cap_{i=1}^{n} H_{0i}. \tag{8}$$

The main consequence for outlier detection is that FDR (or FDX) control provides a balance between ignoring multiplicity, as in Hardin and Rocke (2005) or in Hubert *et al.* (2008), and strictly correcting for multiplicity through FWER control, as in Becker and Gather (1999) or in Cerioli *et al.* (2009). The improvement obtained by controlling (4) or (7) may be particularly advantageous when n is high, or when many samples of moderate size need to be analyzed in sequence. In such instances the total number of hypotheses (1) to be tested will be large and the loss of power induced by strong FWER control will become more relevant.

2. FDR and FDX rules for multivariate outlier detection

The performance of any outlier detection method with well-behaved data sets is ruled by two basic elements:

a) availability of a good approximation to the unknown finite-sample null distribution of the squared robust distances (2);

b) correction for the multiplicity implied by repeated testing of the n individual hypotheses (1).

Avoiding b) leads to identify a proportion α of false outliers in *any good data set*, a situation that can have negative consequences in many applications like the examples described in Section 4. Swamping effects are magnified when component a) is also absent. Evidence that asymptotic distributions may fit poorly even with sample sizes in the order of the hundreds is now well documented and extends from the multivariate setting to regression: see, e.g., Cerioli *et al.* (2009), Hardin and Rocke (2005), Maronna and Yohai (2010) and Riani *et al.* (2009).

The Finite Sample RMCD detection rule (FSRMCD) developed by Cerioli (2010a) addresses both a) and b). This methodology is able to control the nominal size of the test of no outliers even when the ratio n/v is very unfavourable (e.g. when $n/v \approx 5$) and asymptotic approximations fail. Being based on Sidak correction, the FSRMCD procedure aims at providing strong control of the FWER. As remarked in §1.1, the main drawback of this approach is its low power. A less stringent and more powerful requirement is weak control of the FWER, one example of which is the Iterated RMCD (IRMCD) procedure, also proposed by Cerioli (2010a). The IRMCD methodology consists in adding a further iteration step without multiplicity adjustment when FSRMCD suggests the presence of at least one outlier. The associated degree of swamping is controlled when there is no evidence of contamination. If at least one outlier is found, however, the proportion of false outliers will rise up to almost α.

Control of the FDR can provide a better balance, since it puts a bound on the expected proportion of false outliers. The key issue when the FDR is controlled in the outlier detection process is that the acceptable degree of swamping is allowed to depend on the number of selected outliers, not just on evidence of contamination. This is particularly useful when some swamping can be permitted provided that the number of outliers is high: the focus is not on each single outlier, but on the set of selected outliers. When the FDX is controlled, the probability that the proportion of false discoveries exceeds c is bounded, guaranteeing that this proportion is low with high probability.

This view suggests the adoption of multiplicity-adjusted procedures for multivariate outlier detection that focus on the FDR and on the FDX criteria. Our finite-sample proposals based on the RMCD estimators, FDR-RMCD and FDX-RMCD for short, are given below. They assume that the clean part of the data comes from $N(\mu, \Sigma)$, as in (1). This assumption is not uncommon: see, e.g., Hardin and Rocke (2005), Morgenthaler (2007), Pison *et al.* (2002) and Todorov and Filzmoser (2010). Furthermore, Cerioli (2010b) shows how the normality assumption can be robustly checked using the same distributional results used in Step 4 below.

The FDR-RMCD and FDX-RMCD rules are made of five steps.

Step 1. Set the coverage h of the MCD subset. Typical choices are

$$h = \left\lfloor \frac{n+v+1}{2} \right\rfloor \approx 0.5n \tag{9}$$

and

$$h = \lfloor 2\lfloor \frac{n+v+1}{2} \rfloor - n + 1.5(n - \lfloor \frac{n+v+1}{2} \rfloor) \rfloor \approx 0.75n. \tag{10}$$

Compute the raw MCD estimators of location and scatter, using the algorithm of Rousseeuw and Van Driessen (1999). Apply to the estimator of

scatter both the consistency and the finite sample corrections described by Pison *et al.* (2002).

Step 2. Compute the weights w_i, $i = 1, \ldots, n$, with $w_i = 1$ if the MCD squared distances are below the 0.975 quantile of the scaled F distribution of Hardin and Rocke (2005), and $w_i = 0$ otherwise. Let $m = \sum_{i=1}^{n} w_i$.

Step 3. Compute the RMCD estimators:

$$\tilde{\mu}_{(\mathrm{RMCD})} = \frac{1}{m} \sum_{i=1}^{n} w_i y_i;$$

$$\tilde{\Sigma}_{(\mathrm{RMCD})} = \frac{0.975}{\Pr(\chi^2_{v+2} < \chi^2_{v,0.975})} \sum_{i=1}^{n} \frac{w_i (y_i - \tilde{\mu}_{(\mathrm{RMCD})})(y_i - \tilde{\mu}_{(\mathrm{RMCD})})'}{m-1},$$

where the first term of $\tilde{\Sigma}_{(\mathrm{RMCD})}$ is a consistency factor corresponding to a nominal trimming of 2.5% in Step 2.

Step 4. Compute the squared reweighted distances

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

and assume for each of them the distribution:

$$d^2_{i(\mathrm{RMCD})} \sim \frac{(m-1)^2}{m} \mathrm{Beta}(\frac{v}{2}, \frac{m-v-1}{2}) \qquad \text{if} \qquad w_i = 1 \qquad (11)$$

$$\sim \frac{m+1}{m} \frac{(m-1)v}{m-v} F_{v,m-v} \qquad \text{if} \qquad w_i = 0. \qquad (12)$$

Step 5. Compute p-values for the squared reweighted distances $d^2_{i(\mathrm{RMCD})}$, using their reference distribution in Step 4. Control the FDR using the BH procedure *or* the FDX using the LR procedure. The rejected hypotheses correspond to the outliers. If desired, estimate the pFDR using (6).

Step 1 is also implemented in function *covMcd* of the R package *robustbase*. Steps 2–4 mimic those of the FSRMCD procedure of Cerioli

(2010a) and ensure an accurate approximation to the unknown null distribution of the squared RMCD distances. The robust outlier detection problem then allows for use of the standard FDR or FDX controlling procedures in Step 5. To our knowledge this is a substantially new proposal for the purpose of multiple outlier detection, whose merits are evaluated in the next sections. For the moment, we note that a crucial requirement for proper error rate control in real life situations is that the p-values actually follow the uniform distribution. Therefore, the applicability of Step 5 is intimately related to the existence of a good approximation to the null distribution of the squared robust distances $d^2_{i(\mathrm{RMCD})}$, which is indeed provided by Steps 2–4.

A final issue regards the potential problem of dependence among the n squared distances $d^2_{i(\mathrm{RMCD})}$, which are based on the same estimates $\tilde{\mu}_{(\mathrm{RMCD})}$ and $\tilde{\Sigma}_{(\mathrm{RMCD})}$. In general, it is known that Bonferroni procedure is robust with respect to arbitrary dependence of the test statistics, Sidak procedure with respect to positive dependence, BH procedure both with respect to positive (Benjamini and Yekutieli, 2001) and weak (Farcomeni, 2007) dependence, and that LR procedure is also robust with respect to certain forms of dependence (Lehmann and Romano, 2005). We now show a simulation to provide empirical confirmation of the minor impact of dependence on the multiple testing corrections.

For a given test, let α be the FWER under the complete null, that is,

$$\alpha = \Pr(N_{1|0} \geq 1 | M_0 = n), \tag{13}$$

in the notation of Table 1. Let $\hat{\alpha}_1(\gamma)$ be the Monte Carlo estimate of the right-hand side of (13) when each hypothesis (1) is tested at nominal size γ.

If the test statistics are independent, we also obtain $\alpha = 1 - (1 - \gamma)^n$ from Sidak relationship (3). Therefore, under independence, we can compute an alternative estimate of α as

$$\hat{\alpha}_2(\gamma) = 1 - (1 - \hat{\gamma})^n,$$

where $\hat{\gamma}$ is the Monte Carlo estimate of $E[N_{1|0}]/n$ when each H_{0i} is tested at nominal size γ. If the n tests are independent, $\hat{\alpha}_1(\gamma)$ and $\hat{\alpha}_2(\gamma)$ should be approximately equal. We then compute the ratio

$$\tilde{\alpha}(\gamma) = \frac{\hat{\alpha}_1(\gamma)}{\hat{\alpha}_2(\gamma)}$$

as our measure of the effect of dependence when each individual test of (1) is performed at nominal size γ. Table 2 reports the values of $\tilde{\alpha}(\gamma)$ obtained from 5000 simulations run under (8) with different values of n and v. Due to affine invariance of the RMCD distances, we take $\mu = 0$ and $\Sigma = I$. We choose γ to give a nominal $\alpha = 0.05$ in equation (3). The effect of dependence among the squared robust distances $d^2_{i(\mathrm{RMCD})}$ is seen to be minor, even in moderately small samples.

We also computed Monte Carlo estimates of the average rank correlation, a resistant measure of dependence (Croux and Dehon, 2010), between $d^2_{i(\mathrm{RMCD})}$ and $d^2_{i'(\mathrm{RMCD})}$, for $i \neq i'$. Our simulations show this measure to be generally smaller in magnitude than $-1/(n-1)$, the theoretical correlation between two squared classical Mahalanobis distances. We thus conclude that dependence is not a major issue for multiple outlier detection, unless perhaps when n is very small. In that case, the BH and LR procedures adopted in Step 5 of our proposed rules could be replaced by alternative techniques, like

Table 2: Dependence measure $\tilde{\alpha}(\gamma)$, for $\gamma = 0.05$. A value of $\tilde{\alpha}(\gamma)$ close to 1 indicates that the effect of dependence among the squared robust distances $d^2_{i(\mathrm{RMCD})}$ is negligible.

	$n = 90$	$n = 125$	$n = 200$	$n = 400$
$v = 5$	0.96	0.97	0.97	1.00
$v = 10$	0.97	0.98	1.01	0.99
$v = 15$	0.96	0.99	1.01	1.01

those of Benjamini and Yekutieli (2001) and Lehmann and Romano (2005), that are able to control the FDR and the FDX under arbitrary dependency.

3. Enemy brothers: power and swamping

We now show the results of a simulation experiment run under the location-shift contamination model $N(\mu + \lambda e, \Sigma)$, where λ is a positive scalar and e is a column-vector of ones. In our study, a proportion ω of observations come from the location-shift contamination model, while the remaining $n(1 - \omega)$ observations come from the null $N(\mu, \Sigma)$ model. We call ω the contamination rate. We also define power to be the proportion of contaminated observations correctly labelled as outliers. Without loss of generality, we assume that $\mu = 0$ and $\Sigma = I$.

In order to show the behaviour of the methods in absence of outliers, Table 3 provides the results for $\lambda = 0$. Each entry in this table is the estimated size of the test of no outliers, given in (13), for a nominal $\alpha = 0.05$. Both FDR-RMCD and FDX-RMCD display good null performance. They can thus be compared to the available FSRMCD and IRMCD procedures from the point of view of power.

Table 3: Estimated sizes of the test of no outliers for a nominal $\alpha = 0.05$. Simulation settings as in Figures 1 and 2.

	FDR-RMCD	FDX-RMCD	FSRMCD	IRMCD
$n = 200$ and $v = 10$	0.044	0.044	0.048	0.048
$n = 2000$ and $v = 50$	0.045	0.045	0.045	0.045

We first examine the power performance of FDR-RMCD and FDX-RMCD in the case $n = 200$ and $v = 10$, a setting similar to the structure of the glass data of Section 4.1. In situations where the expected proportion of contaminants is not very high, it is often suggested to go for a compromise between robustness and efficiency. Therefore, in our experiment we choose h as in (10) if $\omega \leq 0.10$, and h equal to (9) otherwise. We base our power estimation on 500 simulations.

Figure 1 and Table 4 display the results obtained for different shifts λ and contamination rates ω. In this framework, where n is not high, the gain in power provided by FDR control with respect to FWER control is moderate for $\omega = 0.02$, but grows for larger contamination rates. The number of false detections for FDR-RMCD generally increases with ω and λ, but it is still comparable to that of FSRMCD. On the other hand, the degree of swamping provided by IRMCD approaches 5% when λ grows. For small values of ω, FDX-RMCD is practically indistinguishable from FSRMCD. When ω is larger, FDX-RMCD provides a gain in power with respect to FSRMCD, still anyway quite below FDR-RMCD.

Figure 2 compares power when there is a large data structure. In particular, here $n = 2000$, $v = 50$ and power is estimated on 200 simulations

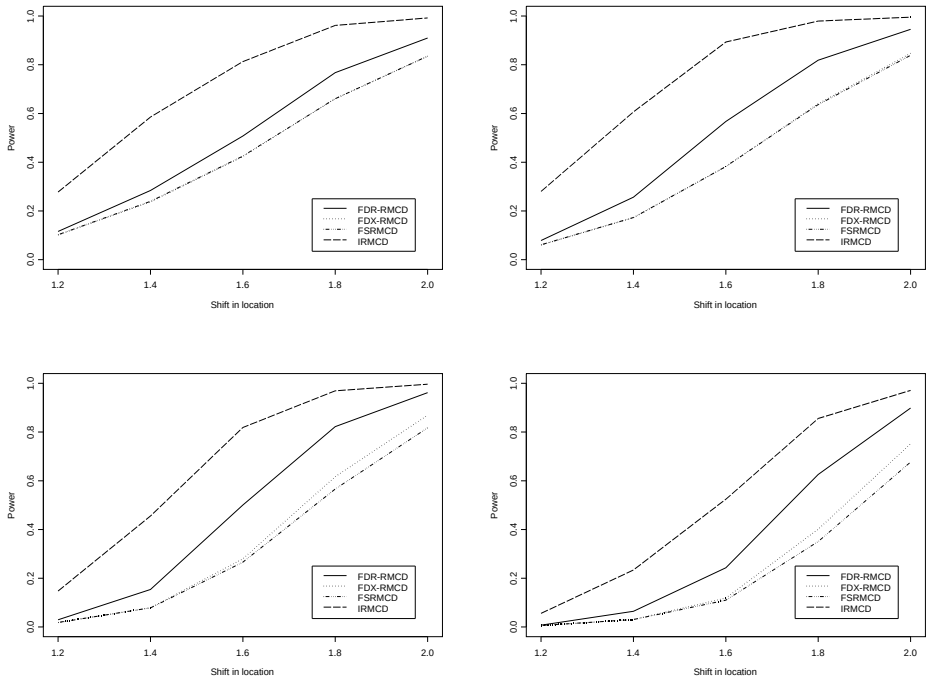

Figure 1: Power of FDR-RMCD, FDX-RMCD, FSRMCD and IRMCD under a multivariate location-shift contamination model, for $n = 200$ and $v = 10$. Nominal $\alpha = 0.05$. Upper panels: $\omega = 0.02$ (left) and $\omega = 0.05$ (right). Lower panels: $\omega = 0.10$ (left) and $\omega = 0.15$ (right). Power is estimated on 500 simulations for each ω and λ.

for each value of λ and ω. Again, the good compromise behaviour of FDR-RMCD stands out clearly among the alternative procedures. As expected, the gain in power provided by FDR and FDX control with respect to FSRMCD is higher for larger n.

The counts reported in Table 4 for the case $n = 2000$ make clear the rationale of FDR control, which allows a larger number of false detections when the number of outliers increases. Note that the number of false discoveries that must be tolerated by using IRMCD in this setting can be

Table 4: Estimated number of non-contaminated observations wrongly declared to be outliers. Simulation settings as in Figures 1 and 2. The first row of each procedure refers to the case $\omega = 0.02$ for $n = 200$ and to the case $\omega = 0.01$ for $n = 2000$. The other rows refer to $\omega = 0.05$ and $\omega = 0.10$ for both sample sizes.

	$n = 200$, $v = 10$			$n = 2000$, $v = 50$		
	$\lambda = 1.2$	$\lambda = 1.6$	$\lambda = 2.0$	$\lambda = 0.8$	$\lambda = 1.0$	$\lambda = 1.2$
FDR-RMCD	0.07	0.18	0.21	0.12	0.64	0.97
	0.09	0.32	0.47	0.11	3.17	4.37
	0.07	0.45	0.87	0.05	4.62	8.38
FDX-RMCD	0.05	0.07	0.04	0.06	0.07	0.08
	0.04	0.03	0.07	0.03	0.17	0.44
	0.04	0.05	0.12	0.05	0.28	0.87
FSRMCD	0.05	0.07	0.04	0.06	0.05	0.04
	0.04	0.03	0.04	0.03	0.06	0.03
	0.04	0.04	0.05	0.05	0.06	0.07
IRMCD	3.59	8.33	9.42	67.3	94.1	94.5
	3.68	8.62	9.00	47.6	87.9	89.0
	2.11	6.99	8.19	6.37	79.2	82.9

in the order of one hundred. In applications like the example of Section 4.2, the consideration of such a high degree of swamping may supersede the appreciation of the resulting gain in power. The same conclusion holds for other outlier detection methods that do not take multiplicity into account with large data sets; see, e.g., Filzmoser *et al.* (2008).

Finally, Table 5 gives the error rates (4), (6) and (7) for the proposed rules FDR-RMCD and FDX-RMCD. Note that the reported FDR and FDX are Monte Carlo estimates, not available in applications, while $\widehat{pFDR}$ is the

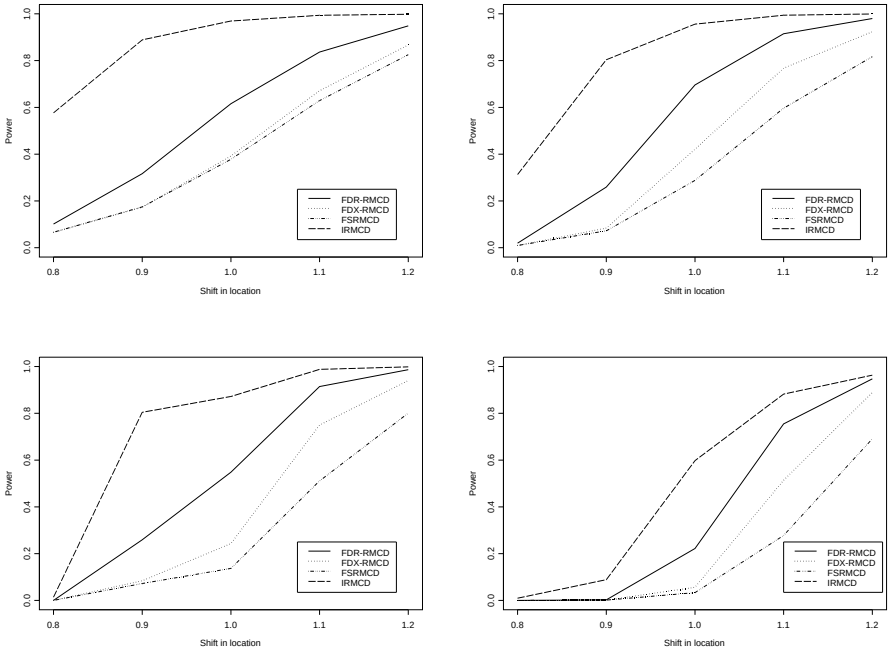

Figure 2: Power of FDR-RMCD, FDX-RMCD, FSRMCD and IRMCD under a multivariate location-shift contamination model, for $n = 2000$ and $v = 50$. Nominal $\alpha = 0.05$. Upper panels: $\omega = 0.01$ (left) and $\omega = 0.05$ (right). Lower panels: $\omega = 0.10$ (left) and $\omega = 0.15$ (right). Power is estimated on 200 simulations for each ω and λ.

average estimate of the pFDR, which is then available on real data. We have two remarks. The first regards FDX-RMCD, which often controls the FDX at a level well below the nominal 0.05. This drawback reflects the behaviour of the LR procedure implemented in Step 5 and could be overcome with less conservative corrections, like those of Farcomeni (2009) and Guo and Romano (2007). Our second remark concerns $\widehat{pFDR}$: when n and ω are small, this estimate is often quite large, indicating that when some outliers are identified, a large proportion of them is likely false. The FDR is still controlled since in

many cases there actually are no rejections. On the other hand, when n or ω are larger, the estimated pFDR drops. It finally equates the FDR when, due to the large number of observations and/or to the high contamination, we almost always reject the complete null (8). In applications, the pFDR can for instance be used as a second level of control: if some outliers are detected but $\widehat{pFDR}$ is large, one could better use FSRMCD. Note that the estimated pFDR is larger for FDX-RMCD than for FDR-RMCD when $n = 200$.

Table 5: Estimated error rates for FDR-RMCD and FDX-RMCD. Simulation settings as in Figures 1 and 2. The reported contamination rates are the same as in Table 4. First entry in each cell: FDR; second entry: pFDR; third entry: FDX

	$n = 200$, $v = 10$		$n = 2000$, $v = 50$	
	$\lambda = 1.6$	$\lambda = 2.0$	$\lambda = 1.0$	$\lambda = 1.2$
FDR-RMCD	0.05; 0.51; 0.14	0.04; 0.28; 0.19	0.04; 0.11; 0.16	0.05; 0.07; 0.11
	0.04; 0.25; 0.19	0.04; 0.11; 0.12	0.04; 0.08; 0.03	0.04; 0.04; 0.00
	0.03; 0.21; 0.12	0.04; 0.07; 0.08	0.04; 0.13; 0.00	0.04; 0.05; 0.00
FDX-RMCD	0.03; 0.62; 0.07	0.01; 0.33; 0.04	0.01; 0.15; 0.03	0.01; 0.06; 0.00
	0.01; 0.36; 0.03	0.01; 0.12; 0.01	0.01; 0.03; 0.00	0.01; 0.01; 0.00
	0.01; 0.31; 0.02	0.01; 0.06; 0.01	0.01; 0.05; 0.00	0.01; 0.01; 0.00

4. Data Analysis

We outline two real data examples, on which we demonstrate the usefulness of multiplicity corrections. Both examples might also be seen as classification situations, for which alternative solutions are available. The outlier detection framework, with respect to many classifiers, has the

disadvantage that we must use assumptions on the sample. Nevertheless, we believe that our approach is worthwhile in these examples for several reasons.

First, with statistical classifiers it may be hard to probabilistically control rates of false positives and false negatives. Usually one builds classifiers so that the estimated false positive and/or false negative rate is below a threshold, but this does not guarantee that the expected rate will be below that threshold. On the contrary, if each object is tested and a Type I error rate controlled, the rate of false positives will be probabilistically controlled on the real data. For instance, the power of classification methods reported for the data of our second example is usually higher than that given by FDR-RMCD (Hastie *et al.*, 2009, p. 301), but at the price of a proportion of false positives of about 4%, and out of control before the experiment. Furthermore, although $\tilde{\mu}_{(\mathrm{RMCD})}$ and $\tilde{\Sigma}_{(\mathrm{RMCD})}$ can resist to the presence of nearly $n/2$ outliers, in many applications the group of contaminants will be much smaller than that of clean observations. Classifiers often perform poorly in presence of imbalance, for instance ending up by assigning all observations to the larger group (Owen, 2007). We also note that in certain cases outlier detection goes beyond the classification task. In the first example we propose, contamination by glass ceramic is the only registered and the most important one, but there may be other unknown potential sources of contamination under which classification methods could break down.

4.1. Quality control: recycling

In production factories batches are often scanned for defective or alien items. Outliers must be identified individually so that they can be substituted. Hence, a high degree of swamping results in unnecessary

discarding of items, or even of entire batches, with considerable loss of money. FDR and FDX controlled outlier detection rules then arise as natural candidates in order to guarantee the required compromise between high power and acceptable swamping.

We illustrate this problem on a data set with $n = 112$ observations discussed from a different perspective in Farcomeni *et al.* (2008). Our example concerns recycling of glass. It is common that glass collected for recycling is still polluted from ceramic glass fragments, which are practically undistinguishable for many automatic sorting devices. The presence of ceramic glass can significantly affect the quality of the recycled glass, since the contaminant has a higher melting point. We thus apply multivariate outlier detection methods with the goal of separating ceramic glass (and other contaminants) from the bulk of "good" observations referring to true glass. The input data consist of the log of $v = 11$ spectral measures recorded for each fragment, after adjusting for multicollinearity. Of the 112 observations in our sample, $n_1 = 109$ are glass fragments and $n_2 = 3$ are contaminated ceramic glass fragments. The contamination rate is close to the expected proportion of outliers in a recycling plant, after manual sorting.

Table 6 summarizes our results when $\alpha = 0.05$ and h is given by (10). The FSRMCD and FDX-RMCD procedures fail to detect 2/3 of the contaminants. On the other hand, all the ceramic glass fragments are identified by IRMCD and FDR-RMCD. FDR-RMCD anyway provides a more careful balance between power and swamping, since the number of falsely rejected glass fragments is much lower than that caused by IRMCD (4 instead of 12). Note also that FDR-RMCD has the lowest estimated

pFDR. Figure 3 shows a boxplot of the squared robust distances $d^2_{i(\mathrm{RMCD})}$, with those of the three ceramic glass fragments marked by a cross. A line indicates the cut-off value set by each method (FDX-RMCD and FSRMCD yield the same cut-off). It is clear that a minimum of 3 false detections is needed in order to identify all the outliers. FDR control leads to almost the minimal swamping, while IRMCD succeeds in detecting all the contaminated objects but at the price of setting the distance threshold too low.

The empirical evidence of this example is that FDR-RMCD may improve over FSRMCD even when n is relatively small. In fact, FDR control scales with respect to the number of rejections, not to the number of tests. On the other hand, FSRMCD would be able to detect all the three ceramic glass fragments only if $\alpha \approx 0.1$, an error rate which seems hard to justify before seeing the data. Larger, and thus even less justifiable, values of α will be required by FSRMCD with larger sample sizes.

Table 6: Results of the FDR-RMCD, FDX-RMCD, FSRMCD and IRMCD procedures for the glass data. Nominal $\alpha = 0.05$.

	Detected Contaminated Fragments	False Detections	$\widehat{pFDR}$
FDR-RMCD	3/3	4/109	0.177
FDX-RMCD	1/3	3/109	0.246
IRMCD	3/3	12/109	0.613
FSRMCD	1/3	3/109	0.246

4.2. Detection: Spam

We illustrate a large sample application on an example of spam detection. We pre-processed a data set publicly available on the UCI Machine Learning

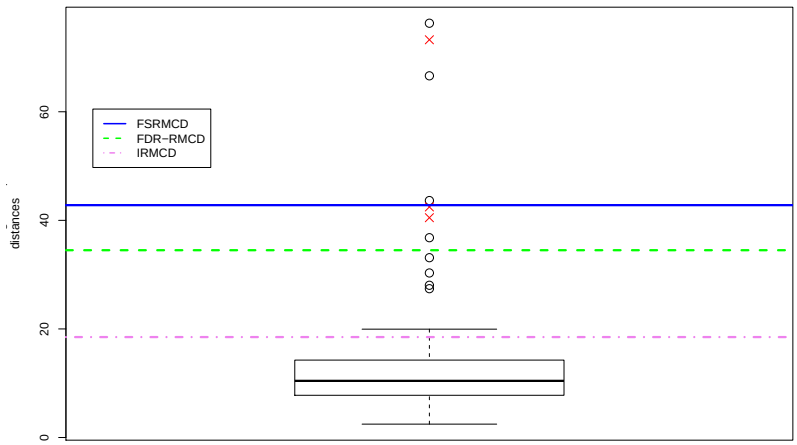

Figure 3: Boxplot of the squared distances $d^2_{i(\mathrm{RMCD})}$ for the glass example, with those of the three contaminated units marked by a cross. The horizontal lines are the cut-off values set by FSRMCD, IRMCD, FDX-RMCD and FDR-RMCD for $\alpha = 0.05$.

Repository. We have a total of $n = 2486$ emails, thirty of which are spam, and $v = 20$ variables. The variables measure characteristics of the email which could indicate its being a clean email or a spam, and include counts of occurrences of particular strings. In order to symmetrize the data, we add one unit to each count and we log-transform. The processed data set can be downloaded from `http://afarcome.interfree.it/spam20.dat`; the first thirty rows correspond to spam emails.

Similarly to the glass data of Section 4.1, a small to moderate fraction of contaminated data can be expected in almost all "batches" of email. Since swamping is a strong concern in spam detection, we now set $\alpha = 0.01$. We then apply our methods, and report the results in Table 7. FSRMCD leads to no swamping, but the number of detected spam messages is very low. This

Table 7: Results of the FDR-RMCD, FDX-RMCD, FSRMCD and IRMCD procedures for the spam data. Nominal $\alpha = 0.01$.

	Detected Spam Messages	False Detections	$\widehat{pFDR}$
FDR-RMCD	13/30	27/2456	0.032
FDX-RMCD	5/30	0/2456	0.183
IRMCD	20/30	202/2456	0.207
FSRMCD	5/30	0/2456	0.183

feature of FSRMCD is expected to become more and more visible as the number of available emails increases. FDX-RMCD yields the same results as FSRMCD. As before, FDR-RMCD and IRMCD seem to work relatively well in detecting contaminated objects, with IRMCD performing slightly better than FDR-RMCD. On the other hand, the number of messages wrongly marked to be spam by IRMCD is not acceptable, while FDR-RMCD succeeds in keeping a good balance between power and swamping. Once again, the estimated pFDR is lowest for FDR-RMCD.

Since in this application the acceptable level of swamping may depend on the personal preferences of the user, we can repeat the analysis by decreasing α. We note that using $\alpha = 0.001$ leads also the FDR-RMCD (but not the IRMCD) to zero swamping. Correspondingly, the proportion of detected spam messages obviously decreases, but FDR-RMCD still dominates FSRMCD in terms of power. With $\alpha = 0.001$, FDX-RMCD leads to the same results as FDR-RMCD.

We conclude that in this application, and in general when the number of units is in the order of the thousands and swamping is a major concern, FDR

control is to be recommended for multiple outlier detection. A very small level α may make also FDX-RMCD suitable.

5. Conclusions

In this paper we have explored alternative ways to reconcile the two opposite goals of multivariate outlier detection: achieving high power under contamination and ensuring low swamping with well behaved data. We have shown that the choice among the alternative methodologies mainly depends on the user attitude towards swamping. The FSRMCD and IRMCD procedures proposed by Cerioli (2010a) have opposite performances. With FSRMCD the level of swamping is kept under control for any number and configuration of contaminated observations, but at the expense of a potentially considerable loss of power. On the contrary, the powerful IRMCD method becomes the best choice if having 1% or 5% of false outliers is an acceptable price to pay under a contamination model.

The main proposals of this paper are the FDX-RMCD and FDR-RMCD detection rules, based on control of recently developed error rates for multiplicity correction. We have investigated their power and swamping properties in different settings. Our conclusion is that they stand out as a sensible compromise between FSRMCD and IRMCD. They can thus provide an appealing strategy in many situations of practical interest, especially for what concerns FDR-RMCD.

Acknowledgments

The authors are grateful to two anonymous reviewers for many useful comments that helped to sharpen the focus of the article. The authors also thank Anthony C. Atkinson and Marco Riani for helpful discussions on previous drafts of this work. Research of the first author was partially supported by the grant "Nuovi metodi multivariati robusti per la valutazione delle politiche sull'e-government e la società dell'informazione" of Ministero dell'Università e della Ricerca – PRIN 2008.

References

Becker, C. and Gather, U. (1999) The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association* 94, 947–955.

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.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29, 1165–1188.

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

Cerioli, A. (2010) Diagnostic checking of multivariate normality under contamination. Preprint.

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.

Croux, C. and Dehon, C. (2010) Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods and Applications*, in press.

Davies, L. and Gather, U. (1993) The identification of multiple outliers. *Journal of the American Statistical Association* 88, 782–792.

Farcomeni, A. (2007) Some results on the control of the false discovery rate under dependence. *Scandinavian Journal of Statistics* 34, 275–297.

Farcomeni, A. (2008) A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research* 17, 347–388.

Farcomeni, A., Serranti, S. and Bonifazi, G. (2008) Nonparametric analysis of infrared spectra for recognition of glass and ceramic glass fragments in recycling plants. *Waste Management* 28, 557–564.

Farcomeni, A. (2009) Generalized augmentation to control the False Discovery Exceedance in multiple testing. *Scandinavian Journal of Statistics* 36, 501–517.

Filzmoser, P., Maronna, R. and Werner, M. (2008) Outlier identification in high dimensions. *Computational Statistics and Data Analysis* 52, 1694–1711.

Guo, W. and Romano, J. (2007) A generalized Sidak-Holm procedure and control of generalized error rates under independence. *Statistical Applications in Genetics and Molecular Biology* 6, Article 3.

Hardin, J. and Rocke, D. M. (2005) The distribution of robust distances. *Journal of Computational and Graphical Statistics* 14, 910–927.

T. Hastie, R. Tibshirani, and J. Friedman (2009) *The Elements of Statistical Learning*, 2nd Edition. New York: Springer.

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

Lehmann, E.L. and Romano, J.P. (2005) Generalizations of the Familywise Error Rate. *Annals of Statistics* 33, 1138–1154.

Maronna, R.A. and Yohai, V.J. (2010) Correcting MM estimates for "fat" data sets. *Computational Statistics and Data Analysis*, doi:10.1016/j.csda.2009.09.015.

Morgenthaler, S. (2007) A survey of robust statistics. *Statistical Methods and Applications* 15, 271–293. Erratum: 16, 171–172.

Owen, A.B. (2007) Infinitely imbalanced logistic regression. *Journal of Machine Learning Research* 8, 761–773.

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

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.

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

Schweder, T. and Spjøtvoll, E. (1982) Plots of p-values to evaluate many hypotheses simultaneously. *Biometrika* 69, 493–502.

Storey, J.D. (2002) A direct approach to false discovery rates. *Journal of the Royal Statistical Society (Ser. B)* 64, 479–498.

Storey, J.D. (2003) The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* 31, 2013–2035.

Todorov, V. and Filzmoser, P. (2010) Robust statistic for the one-way MANOVA. *Computational Statistics and Data Analysis* 54, 37–48.

van der Laan, M.J., Dudoit, S. and Pollard, K.S. (2004) Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statistical Applications in Genetics and Molecular Biology* 3, Article 1.