Computational Statistics and Data Analysis 65 (2013) 29–45

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust distances for outlier-free goodness-of-fit testing

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

[a] University of Parma, Via Kennedy 6, 43125 Parma, Italy

[b] Sapienza - University of Rome, piazzale Aldo Moro, 5, 00186 Roma, Italy

ARTICLE INFO

Article history:
Received 31 May 2011
Received in revised form 6 February 2012
Accepted 14 March 2012
Available online 20 March 2012

Keywords:
Beta distribution
False discovery rate
Multivariate normality
Outlier removal
Reweighted MCD
Trimming

ABSTRACT

Robust distances are mainly used for the purpose of detecting multivariate outliers. The precise definition of cut-off values for formal outlier testing assumes that the "good" part of the data comes from a multivariate normal population. Robust distances also provide valuable information on the units not declared to be outliers and, under mild regularity conditions, they can be used to test the postulated hypothesis of multivariate normality of the uncontaminated data. This approach is not influenced by nasty outliers and thus provides a robust alternative to classical tests for multivariate normality relying on Mahalanobis distances. One major advantage of the suggested procedure is that it takes into account the effect induced by trimming of outliers in several ways. First, it is shown that stochastic trimming is an important ingredient for the purpose of obtaining a reliable estimate of the number of "good" observations. Second, trimming must be allowed for in the empirical distribution of the robust distances when comparing them to their nominal distribution. Finally, alternative trimming rules can be exploited by controlling alternative error rates, such as the False Discovery Rate. Numerical evidence based on simulated and real data shows that the proposed method performs well in a variety of situations of practical interest. It is thus a valuable companion to the existing outlier detection tools for the robust analysis of complex multivariate data structures.

1. Introduction

Let y be a v-variate observation from a population with mean vector μ and covariance matrix Σ. The robust analogue of the (squared) Mahalanobis distance of y is

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

where $\tilde{\mu}$ and $\tilde{\Sigma}$ are high-breakdown estimators of μ and Σ. For simplicity, we usually omit the fact that (1) is squared and we call it a *robust distance*. Robust distances are computed for the purpose of detecting multivariate outliers. In fact, given a sample of n observations $y_1, \ldots, y_n$, the estimates $\tilde{\mu}$ and $\tilde{\Sigma}$ are not affected by the presence of nasty outliers in the sample. Therefore, the outliers themselves are revealed by their large distances (1) from the robust fit (Filzmoser et al., 2008; Hubert et al., 2008; Van Aelst et al., 2012).

In this paper our choice for $\tilde{\mu}$ and $\tilde{\Sigma}$ is the Reweighted Minimum Covariance Determinant (RMCD) estimator, for which accurate distributional results exist; see (4) and (5). These results are crucial for the practical implementation of our proposal, which extends the use of robust distances from outlier detection to goodness-of-fit testing. In particular, we exploit the fact

[*] Corresponding author. Tel.: +39 0521 902491; fax: +39 0521 902375.
E-mail addresses: andrea.cerioli@unipr.it (A. Cerioli), alessio.farcomeni@uniroma1.it (A. Farcomeni), mriani@unipr.it (M. Riani).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.03.008

that the distribution of the robust distances computed from the RMCD estimator depends on whether each observation is trimmed or not in the outlier identification process. Although we are not aware of specific research in this direction, we speculate that similar distributional results might be obtained if the RMCD is replaced by other estimators based on a "hard-trimming" approach, such as impartial trimming (Garcìa-Escudero et al., 2008), trimmed maximum likelihood (Cuesta-Albertos et al., 2008) and the Forward Search (Riani et al., 2009). Robust distances computed from high-breakdown estimators exploiting a smooth weight function, such as S and MM-estimators (Maronna et al., 2006; Alfons et al., 2011; Van Aelst and Willems, 2011), are instead compared to their asymptotic χ^2_v distribution for all the observations $y_1, \ldots, y_n$. The simulation results of Cerioli et al. (in press) show that this asymptotic approximation can provide acceptable results in the case of S-estimators for the purpose of outlier detection, when only the tail of the distribution is involved. However, the degree of accuracy of the χ^2_v approximation for the bulk of the data, which is crucial for our proposal, is unknown and should be verified if $\tilde{\mu}$ and $\tilde{\Sigma}$ in (1) are taken to be S-estimators. Given a satisfactory approximation for the distribution of the robust distances of the uncontaminated observations, our robust approach to goodness-of-fit testing would then remain valid.

The RMCD estimator is computed in two stages. In the first stage, we fix a coverage $n/2 \leq h < n$ and we define the MCD subset to be the sub-sample of h observations whose covariance matrix has the smallest determinant. The second stage aims at increasing efficiency, while preserving high-breakdown properties. For this purpose, a one-step reweighting scheme is applied by giving weight $w_i = 0$ to observations whose first-stage robust distance exceeds a threshold value. Otherwise the weight is $w_i = 1$. Let $w = \sum_{i=1}^n w_i$. The RMCD estimator of μ and Σ is then

$$\tilde{\mu}_{\mathrm{RMCD}} = \frac{1}{w}\sum_{i=1}^{n} w_i y_i, \qquad \tilde{\Sigma}_{\mathrm{RMCD}} = \frac{k_{\mathrm{RMCD}}}{w-1}\sum_{i=1}^{n} w_i (y_i - \tilde{\mu}_{(\mathrm{RMCD})})(y_i - \tilde{\mu}_{(\mathrm{RMCD})})', \tag{2}$$

where the scaling k_{RMCD} serves the purpose of ensuring consistency at the normal model (Croux and Haesbroeck, 1999). Our choice of the threshold required for computing w_i is the 0.975 quantile of the scaled F distribution proposed by Hardin and Rocke (2005). Our procedure then identifies multivariate outliers by means of the robust reweighted distances

$$d^2_{i(\mathrm{RMCD})} = (y_i - \tilde{\mu}_{(\mathrm{RMCD})})' \tilde{\Sigma}^{-1}_{(\mathrm{RMCD})} (y_i - \tilde{\mu}_{(\mathrm{RMCD})}) \quad i = 1, \ldots, n. \tag{3}$$

Precise outlier identification requires cut-off values for the distances (3). Cerioli (2010a) shows that a very accurate approximation is provided by considering

$$d^2_{i(\mathrm{RMCD})} \sim \frac{(w-1)^2}{w}\mathrm{Beta}\left(\frac{v}{2}, \frac{w-v-1}{2}\right) \quad \text{if } w_i = 1 \tag{4}$$

$$\sim \frac{w+1}{w}\frac{(w-1)v}{w-v}F_{v,w-v} \quad \text{if } w_i = 0. \tag{5}$$

Computation of these cut-off values assumes that the "good" part of the data comes from a multivariate normal population. To a large extent, the same is true for most outlier identification methods; see, e.g., Filzmoser et al. (2008), Gallegos and Ritter (2005), Garcìa-Escudero and Gordaliza (2005), Hardin and Rocke (2005), Riani et al. (2009) and Willems et al. (2009). Furthermore, a computable formula for k_{RMCD} in (2), as well as for the consistency factor of any other affine equivariant high-breakdown estimator of Σ, is only available under that hypothesis (Todorov and Filzmoser, 2009). It is not exaggerated to say that multivariate outlier detection relies heavily on the hypothesis of normality for the "good" part of the data. The same is true for many other robust multivariate techniques, whose robustness properties have been studied mainly at the normal model (Croux and Haesbroeck, 2000; Croux and Dehon, 2010; Hubert and Van Driessen, 2004; Rousseeuw et al., 2004; Van Aelst and Willems, 2011). It is thus instructive to see what happens when the normality assumption for the bulk of the data is not fulfilled.

Fig. 1 displays the output of a standard exploratory analysis for multivariate outlier identification (Maronna et al., 2006, p. 179). In the left-hand panels, the robust reweighted distances $d^2_{i(\mathrm{RMCD})}$ are computed for a sample of $n = 1000$ observations with $v = 5$. Of these observations, 930 come from the $N(0, I_v)$ distribution. The remaining 70 observations are simulated from the shifted distribution

$$N(0 + \lambda e, I_v), \tag{6}$$

where e is a (column) vector of ones and λ is a positive scalar. In this example $\lambda = 2.0$, a modest amount of contamination. The right-hand panels give the same information for a sample of $n = 1000$ observations simulated from the 5-variate t distribution on 10 degrees of freedom. The threshold displayed in the upper row is the 0.99 quantile of distribution (5), while the lower row compares the empirical quantiles of the robust distances to the theoretical values from the asymptotic χ^2_5 distribution. The conclusions reached in the two samples are similar, with about the same number of observations labelled as outliers, a few borderline units, and no clear differences in the structure of the bulk of the data. Therefore, this simple example clarifies that the knowledge usually conveyed by robust distances cannot help to discriminate between a contaminated normal model and a non-normal population. Further evidence of the need for a more sophisticated approach,

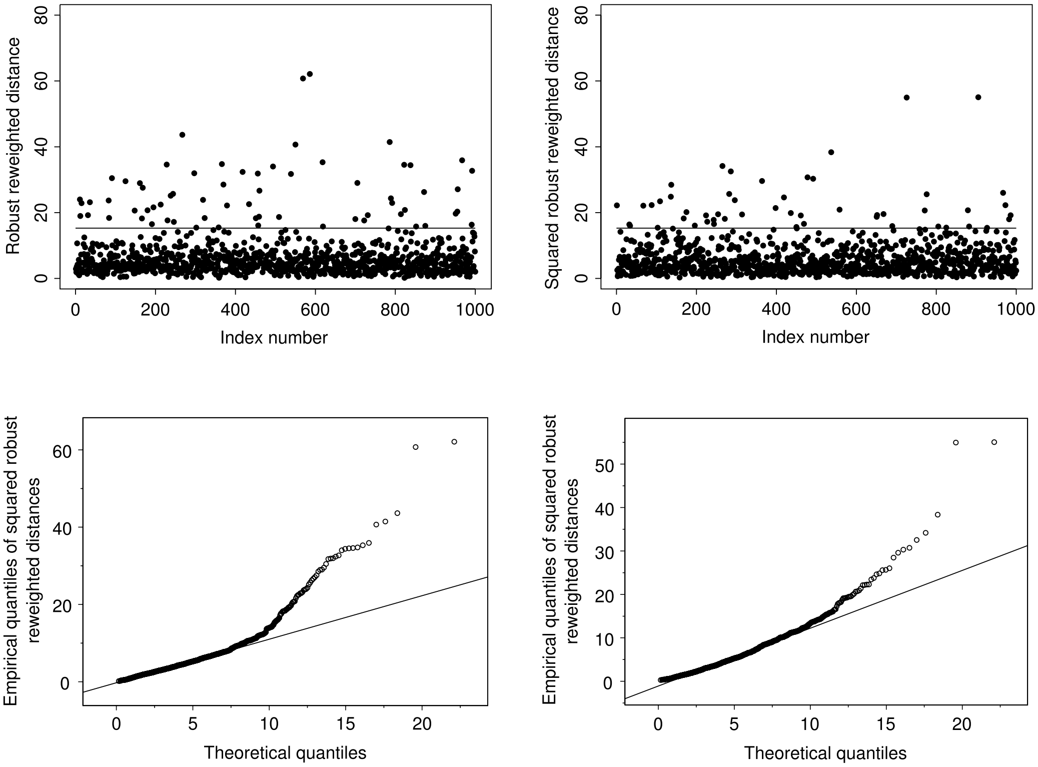

Fig. 1. Robust distances in two samples with $n = 1000$ and $v = 5$. Left (top and bottom panels): 930 observations from $N(0, I_v)$ and 70 observations from a location-shift contamination model. Right (top and bottom panels): 1000 observations from the 5-variate t distribution on 10 degrees of freedom.

which could distinguish between different generating models when outliers are present, is given in Section 7, where formal goodness-of-fit testing is applied to these and real data.

The aim of this paper is to propose a statistically sound method which is able to detect departures from the postulated normal model in the "good" part of the data, i.e. among the observations whose robust distances are below the threshold in the upper row of Fig. 1. This goal is achieved by introducing a contamination model and by analysing, under mild regularity conditions, the empirical distribution of the robust distances $d^2_{i(\text{RMCD})}$ for the units which are not likely to be contaminated. One key issue in the suggested approach is the effect of stochastic trimming. Our consideration of this effect is twofold. First, trimming is an important ingredient for the purpose of obtaining a reliable estimate of the number of "good" observations. Second, we allow for trimming in the empirical distribution of the robust reweighted distances when comparing them to their nominal distribution. Another important point of our proposal is the choice of the error rate to be controlled when removing the outliers. We elaborate on this aspect by comparing the effect of different criteria, including individual and simultaneous testing and False Discovery Rate control.

In the case $v = 1$, a related problem is studied by Alvarez-Esteban et al. (2010), who propose a goodness-of-fit test comparing a trimmed version of the empirical distribution of the data with the trimmed $N(0, 1)$ distribution. Their approach is based on the L_2-Wasserstein metric and requires the choice of a user-defined threshold which controls the degree of dissimilarity between the two (univariate) distributions. On the contrary, our procedure is simpler, relying on the classical chi-square test, and fits very naturally in the multivariate context through the robust distances (3). The null hypothesis of normality of a subset of the data coincides with the hypothesis that the selected subset arises from the normal component of the mixture, if the contamination is far away from the mean of the normal component. This assumption is made explicit in Section 3, specifically in our Lemma 1 and in the examples that follow it. Therefore, a bound on the distance between the distribution of the clean and contaminated data is not explicitly needed. Coupling the accurate finite-sample approximation (4) for the robust distances with the effect of trimming, allows our method to have good control of the size of the test of multivariate normality, also with relatively small uncontaminated samples and large dimensions. The same is not necessarily true, without Monte Carlo calibration, for other robust approaches to goodness-of-fit testing that rely on quantiles of the asymptotic χ^2_v distribution (Beirlant et al., 1999), or that neglect the importance of trimming (Singh, 1993).

The rest of this paper is organized as follows. In Section 2 we define the framework in which we work. The required regularity conditions are stated in Section 3, where we also explore their implications. In Section 4 we address the effect of stochastic trimming. Our robust chi-squared tests of multivariate normality are described in Section 5. In the same section, we investigate their performance under the null. Power comparisons under different alternatives are provided in Section 6, while numerical examples are described in Section 7. The paper ends with some final remarks in Section 8.

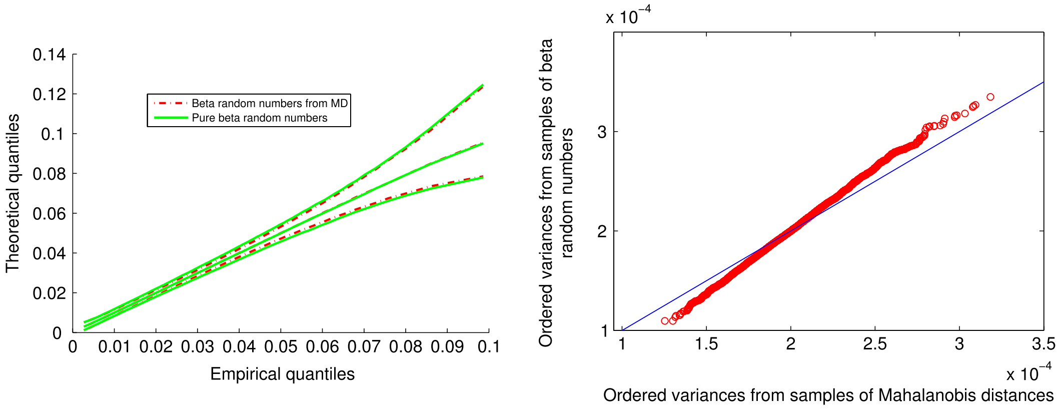

Fig. 2. Left: Q–Q plot of the curves associated with the 1%, 50% and 99% quantiles of data from Mahalanobis distances (dotted lines) and data generated directly from the Beta distribution (solid lines). Right: Q–Q plot of the distribution of the variances of numbers generated directly by the Beta distribution and the variances of numbers which come from Mahalanobis distances.

2. Multivariate normality under contamination

We formalize our approach through a contamination model. We assume that $y_1, \dots, y_n$ are independent, identically distributed v-variate observations with common distribution

$$G(y) = \gamma_{\text{sup}} G_0(y) + (1 - \gamma_{\text{sup}}) G_1(y), \tag{7}$$

where $(1 - \gamma_{\text{sup}})$ is the contamination rate and $\gamma_{\text{sup}} > 0.5$ is unknown. $G_0(y)$ and $G_1(y)$ denote the distribution functions of the "good" and of the contaminated part of the data, respectively. Let $\Phi(y, \mu, \Sigma)$ be the distribution function of the v-variate normal distribution with mean μ and covariance matrix Σ. Our null hypothesis is then

$$H_0 : G_0(y) = \Phi(y, \mu_0, \Sigma_0), \tag{8}$$

where μ_0 and Σ_0 are not known and must be estimated from the data.

As a referee noted, decomposition (7) is not unique unless we put additional assumptions on $G_0(\cdot)$ and $G_1(\cdot)$; see, e.g., Bordes et al. (2006) and Hunter et al. (2007) for such assumptions. There actually could be a set $\mathcal{G}$, with more than one element, such that for any $(G_0(\cdot), G_1(\cdot), \gamma_{\text{sup}}) \in \mathcal{G}$ the decomposition (7) holds. However, we are only interested in verifying whether, among all (possibly infinite) two-component mixture combinations leading to G, there is at least one in which one of the components is Gaussian and contains at least 50% of the data. Rejection of (8) with an appropriate test statistic would then imply that, unless a type I error has occurred, there is no mixture with Gaussian component in this family, that is, $(\Phi(y, \mu_0, \Sigma_0), G_1(\cdot), \gamma_{\text{sup}}) \notin \mathcal{G}$ for any $\gamma_{\text{sup}} > 0.5$. A decomposition similar to (7) has been adopted, with similar reasoning, also in Garcìa-Escudero and Gordaliza (2005).

One class of methods for checking multivariate normality (Mecklin and Mundfrom, 2004) is based on the comparison between the empirical Mahalanobis distances (MD)

$$d_i^2 = (y_i - \hat{\mu})' \hat{\Sigma}^{-1} (y_i - \hat{\mu}) \quad i = 1, \dots, n, \tag{9}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the classical unbiased estimators of μ and Σ, and the percentage points of their distribution at the normal model, which is $\{(n-1)^2/n\}\text{Beta}(0.5v, 0.5(n - v - 1))$ (Atkinson et al., 2004). Although a visual approach based on Q–Q plots can provide useful preliminary guidance, formal testing requires appropriate consideration of the effect of parameter estimation and of the constraints implied by the use of Mahalanobis distances. In order to show this effect we have conducted a set of 10 000 simulations each made up of datasets with $n = 200$ and $v = 6$. For each dataset we have stored the set of the n ordered Mahalanobis distances, as given in Eq. (9) and then multiplied by $200/199^2$, and a set of $n = 200$ ordered numbers generated from the Beta distribution with parameters $(3, 193/2)$. We have then considered the quantiles 1%, 50% and 99% of both sets of numbers over the collection of the 10 000 repetitions. The left-hand panel of Fig. 2 compares these sets of quantiles through a Q–Q plot. The solid lines refer to the data generated directly from the Beta distribution, while the dotted lines consider the data associated with the Mahalanobis distances. We see from the plot that the data generated from the Beta distribution have greater variability than those which come from the Mahalanobis distances. We magnify this phenomenon by comparing the ordered variances of the numbers generated directly from the Beta distribution with those which come from the Mahalanobis distances over all sets of simulations. A Q–Q plot of the two distributions, pictured in the right-hand panel of Fig. 2, clearly shows that the distribution of the variances of the Mahalanobis distances is much more concentrated.

Table 1
Estimated size of the Pearson chi-square test of H_0 comparing classical Mahalanobis distances to their scaled Beta distribution, for $n = 1000$ and $v = 5$, when the data are generated from the contamination model (7) with $G_0(y) = \Phi(y, 0, I_v)$ and $G_1(y)$ the distribution function of (6). The entries in each column refer to nominal sizes 0.10 and 0.05, respectively. 1000 simulations for each value of γ_{sup} and mean shift λ.

$\gamma_{\text{sup}} = 1$ $\lambda = 0$	$\gamma_{\text{sup}} = 0.9$ $\lambda = 2.5$	$\gamma_{\text{sup}} = 0.9$ $\lambda = 5.0$	$\gamma_{\text{sup}} = 0.8$ $\lambda = 2.5$	$\gamma_{\text{sup}} = 0.8$ $\lambda = 5.0$	$\gamma_{\text{sup}} = 0.7$ $\lambda = 2.5$	$\gamma_{\text{sup}} = 0.7$ $\lambda = 5.0$
0.071 0.041	0.922 0.869	1.000 1.000	0.144 0.085	0.271 0.166	0.243 0.149	0.385 0.278

The phenomenon which we have just seen can be explained as follows. Keeping into account that the sum of Mahalanobis distances (9) is equal to $(n-1)v$ (Atkinson et al., 2004, p. 86), we easily obtain that

$$\frac{n}{(n-1)^2}\sum_{i=1}^{n}\frac{d_i^2}{n} = \frac{v}{n-1}.$$

Notice also that if $B \sim \text{Beta}\,(0.5v, 0.5(n-v-1))$

$$E(B) = \frac{v}{n-1}.$$

This implies that when we consider a set of n Mahalanobis distances we are considering a set of n random numbers from the Beta $(0.5v, 0.5(n-v-1))$ distribution with the constraint that their average must be exactly equal to the expectation of the underlying Beta random variable. More precisely, in variability comparison, we store the two following deviances

$$\sum_{i=1}^{n}\left(\frac{n}{(n-1)^2}d_i^2 - \mu_B\right)^2 \quad \frac{n}{(n-1)^2}d_i^2 \sim B,\ \mu_B = \frac{v}{n-1}$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2 \quad y_i \sim B,\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

In the first case the mean is fixed while in the second case the mean is estimated. Therefore, the constraint on the distribution of d_i^2 due to estimation of μ and Σ may lead to empirical p-values in the goodness of fit tests which are much greater than the expected ones.

The parameter constraint can be clearly seen both in the Kolmogorov–Smirnov and in the Cramér–von Mises tests, whose asymptotic distributions are far from being standard when applied to empirical Mahalanobis distances (Beirlant et al., 1999; Koziol, 1982). On the other hand, the Pearson chi-square test is less influenced by the effect of estimating μ and Σ. Moore and Stubblebine (1981) show that the asymptotic cut-off values of this test fall between those of the χ^2_{K-2} and χ^2_{K-1} distributions, where K denotes the number of classes employed for computing the Pearson statistic. However, the chi-square test based on the Mahalanobis distances (9) is not robust and cannot be used to verify H_0. Table 1 reports its estimated size for $n = 1000$, $v = 5$ and $K = 30$. Size is estimated using 1000 independent replicates of the contamination model (7), with $G_0(y) = \Phi(y, 0, I_v)$ and $G_1(y)$ the distribution function of (6). The reported results correspond to nominal test sizes of 0.10 and 0.05, and use cut-off values from the χ^2_{K-1} distribution. Similar results have been obtained with different values of K. It is seen that the test is conservative with uncontaminated data ($\gamma_{\text{sup}} = 1$), as expected. On the contrary, the actual size of the test may explode when $\gamma_{\text{sup}} < 1$, because of the undue influence of the observations coming from G_1. The impact of the outliers is larger when the contamination rate is small and then decreases due to the effect of masking on the Mahalanobis distances. Further evidence of this behaviour is shown in Section 7.2. When the contamination rate grows, masking also causes a substantial bias in the non-robust estimates of μ and Σ, which produces a further increase in the actual test size. Our results in the following sections demonstrate how the classical Mahalanobis distances can be replaced by their robust counterpart (3) to obtain a valid test of H_0.

3. Regularity conditions on the contaminant distribution

In the rest of this paper we assume that model (7) holds. We address in this section a brief theoretical study of the regularity conditions on $G_1(y)$ needed for controlling the size of our robust goodness-of-fit test. These conditions are intimately linked to the power of the outlier identification procedure when (8) is true.

We first note that, asymptotically, the RMCD outlier identification procedure is α-level:

$$\Pr\left\{d^2_{i(\text{RMCD})} > \chi^2_{v,1-\alpha} | \gamma_{\text{sup}} = 1; H_0 \text{ is true}\right\} \to \alpha, \tag{10}$$

where $\chi^2_{v,1-\alpha}$ denotes the $(1-\alpha)$-quantile of the χ^2_v distribution. This property follows from consistency of the RMCD estimator at the normal model (Lopuhaä, 1999). We then define the probability, c_{G_1} say, that an outlier is identified by the RMCD rule under the contamination model (7):

$$c_{G_1} = \Pr\left\{d^2_{i(\text{RMCD})} > \chi^2_{v,1-\alpha} | y_i \sim G_1(y)\right\}. \tag{11}$$

This quantity is equivalent to the asymptotic power of the outlier identification procedure for a single outlying entry, and is an indirect measure of the separation between the normal and the contaminant distribution.

In what follows we use asymptotic approximations to c_{G_1}. If μ and Σ were known, the distances

$$d_i^2 = (y_i - \mu)'\Sigma^{-1}(y_i - \mu), \qquad d_{i'}^2 = (y_{i'} - \mu)'\Sigma^{-1}(y_{i'} - \mu)$$

would be independent for any y_i and $y_{i'}$, with $i \neq i'$, generated by $G_0(y)$ under H_0. Therefore, for large sample sizes $d^2_{i(\mathrm{RMCD})}$ and $d^2_{i'(\mathrm{RMCD})}$ are approximately independent under the same assumptions, due to consistency of the RMCD estimator. Let α_{GOF} be the size of a generic goodness-of-fit (GOF) test of multivariate normality when $\gamma_{\sup} = 1$ (i.e., under no contamination). Simple sufficient regularity conditions on $G_1(y)$ for this test to be of the same size α_{GOF} when $\gamma_{\sup} \leq 1$ are given in the following lemma:

Lemma 1. *Let* GOF(y) *denote any GOF statistic which leads to an α_{GOF}-level test of multivariate normality when $\gamma_{\sup} = 1$.*

For $\gamma_{\sup} \leq 1$, if $c_{G_1} = 1$ then GOF(y) *leads to an α_{GOF}-level test. If $(c_{G_1})^n \to 1$, the test based on* GOF(y) *is asymptotically α_{GOF}-level.*

Proof. We can decompose the GOF test level as

$$\begin{aligned} \Pr(R_{\mathrm{GOF}}|H_0) &= \Pr(R_{\mathrm{GOF}}|H_0, \text{all outliers removed}) \Pr(\text{all outliers removed}) \\ &\quad + \Pr(R_{\mathrm{GOF}}|H_0, \text{not all outliers removed}) \Pr(\text{not all outliers removed}) \qquad (12) \\ &\leq \alpha_{\mathrm{GOF}}(c_{G1})^{n(1-\gamma_{\sup})} + 1 - (c_{G1})^{n(1-\gamma_{\sup})}, \qquad (13) \end{aligned}$$

where R_{GOF} denotes the event that the GOF test rejects the hypothesis of normality. Now, if $c_{G_1} = 1$, then $1 - (c_{G_1})^{n\gamma_{\sup}} = 0$ and the last expression is exactly equal to α_{GOF}. If $(c_{G_1})^n \to 1$, then $1 - (c_{G_1})^{n\gamma_{\sup}} \to 0$ and the last expression is asymptotically equal to α_{GOF}. □

The lemma shows us a potential "curse of sample size" situation. In fact, as n grows, it may be harder and harder to stay close to the nominal level α_{GOF}, as it may get harder and harder to identify all outliers. The problem can be solved in practice by increasing the trimming level, thus increasing the likelihood of identifying all the observations coming from $G_1(y)$.

It is important to show under which circumstances c_{G_1} can be assumed to be so large that, even if not exactly one, the level of the goodness-of-fit test is not actually considerably inflated. First, note that by definition

$$\Pr\left\{d^2_{i(\mathrm{RMCD})} > \chi^2_{v,1-\alpha} | y_i \sim G_1(y)\right\} \to \int_{\{y:(y-\mu_0)'\Sigma_0^{-1}(y-\mu_0) > \chi^2_{v,1-\alpha}\}} dG_1. \qquad (14)$$

Explicit expressions for c_{G_1} can thus be obtained from (14) in some important special cases.

Example 1 (*Point Mass Contamination*)**.** Assume $G_1(y) = \Delta_{y_0}$, where Δ_y is a Dirac measure putting all its mass at y. Write $d^2_{0(\mathrm{RMCD})}$ for the robust distance corresponding to y_0 and let $\mathbb{I}_{[\cdot]}$ be the indicator function. Then, (11) is equal to $\mathbb{I}_{[d^2_{0(\mathrm{RMCD})} > \chi^2_{v,1-\alpha}]}$. Consequently, if we assume a point-mass contamination with outliers far enough from μ_0, the GOF test will always be asymptotically α_{GOF}-level.

Example 2 (*Mean Shift Model*)**.** Let $G_1(y)$ be the distribution function of $N(\mu_1, \Sigma_0)$. Write $\chi^2_v(nc)$ for a non-central chi-square distribution on v degrees of freedom, with non-centrality parameter nc. Then,

$$c_{G_1} = \Pr(\chi^2_v(nc) > \chi^2_{v,1-\alpha}),$$

where $nc = 0.5(\mu_0 - \mu_1)'(\mu_0 - \mu_1)$. It follows that, if the outliers arise from a Gaussian distribution, then c_{G_1} increases with the separation between the two centres of the Gaussian distributions. Furthermore the increase is at the same rate as $\|\mu_0 - \mu_1\|^2$, where $\| \ \|$ is the vector norm.

Example 3 (*Uniform Mean Shift Model*)**.** A special case of the mean shift model is provided by the distribution $N(\mu_0 + \lambda e, \Sigma_0)$, as in (6). Then,

$$c_{G_1} = \Pr(\chi^2_v(0.5v\lambda^2) > \chi^2_{v,1-\alpha}).$$

We may see this result as an instance of "blessing of dimensionality", which can be attributed to the fact that $\|e\|$ increases with the dimension: as v grows, the non-centrality parameter of the chi-square distribution increases. Consequently, c_{G_1} gets closer and closer to 1 as v increases. In general, we can assume a shift λ_j for the jth dimension. In that case the non-centrality parameter is $0.5\sum_j \lambda_j^2$, which grows to infinity as long as λ_j does not become too small as j grows. A similar phenomenon is described for the PCOut approach of Filzmoser et al. (2008), where with increasing dimensions almost all outliers are identified correctly.

Table 2
Values of c_{G_1} and $c^n_{G_1}$ in the case of a uniform mean shift model with shift λ, for different values of n, v and λ.

v	λ	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$
		c_{G_1}		$c^{200}_{G_1}$		$c^{1000}_{G_1}$	
5	2.5	0.88	0.72	0.00	0.00	0.00	0.00
5	3.75	1.00	1.00	0.80	0.25	0.32	0.00
5	5	1.00	1.00	1.00	1.00	1.00	0.99
10	2.5	1.00	1.00	0.08	0.00	0.00	0.00
10	3.75	1.00	1.00	1.00	1.00	1.00	0.99
10	5	1.00	1.00	1.00	1.00	1.00	1.00

Example 4 (*General Gaussian Contamination*). Let $G_1(y)$ be the distribution function of $N(\mu_1, \Sigma_1)$. Then,

$$c_{G_1} = \Pr\left(\sum_r \eta_r \chi^2_{h_r}(\delta^2_r) > \chi^2_{v,1-\alpha}\right),$$

where η_r are the characteristic roots of $\Sigma_1\Sigma_0$, h_r is the multiplicity of the rth root, and δ_r is a known linear combination of $\mu_0 - \mu_1$ whose exact expression can be found in Scheffé (1959, p. 418). This last probability should be computed using Monte-Carlo integration.

Table 2 shows the numerical values of c_{G_1} and $c^n_{G_1}$ in the case of a uniform mean shift model, for different values of n, v and λ, when $\alpha = 0.05$ and $\alpha = 0.01$ in (10). The parameters are chosen to represent the simulation settings of Sections 4 and 5. This table gives us an account of the degree of separation which guarantees the required null performance of the chosen goodness-of-fit test when $\gamma_{\text{sup}} < 1$. It can be seen that the magnitude of the mean shift λ needed to ensure that $c^n_{G_1}$ is large enough decreases with dimensionality. For instance, $\lambda = 3.75$ is clearly not sufficient when $v = 5$, while the same mean shift provides an almost perfect separation between $G_0(y)$ and $G_1(y)$ when $v = 10$.

A conservative upper bound on the level of the GOF test is readily obtained from Table 2. For instance, when $\lambda = 3.75$, $v = 5$, $\alpha = 0.05$, $\gamma_{\text{sup}} = 0.8$ and $n = 200$, from (13) we can compute an upper bound on the level as 0.21. In practice the level will be much smaller, as we will illustrate in our simulations of Section 5. We speculate that this happens because (12) is actually a product of two probabilities which are unlikely to be both close to 1. Indeed, the most harmful contamination instances for $\Pr(R_{\text{GOF}}|H_0)$ are those where the outliers lie sufficiently far from the bulk of the data, as shown in Table 1. But in these cases the outlier detection rule will have high power and the quantity Pr(not all outliers removed) will be small.

4. Uncovering the "good" observations

In this section we show how to identify the observations that are likely to come from $G_0(y)$ when the null hypothesis (8) is true. We also suggest a way to estimate their number after that trimming has taken place.

Let $\omega_{i0} = 1$ if $y_i \sim G_0(y)$ and $\omega_{i0} = 0$ if $y_i \sim G_1(y)$. Furthermore, let $\lfloor\ \rfloor$ denote the integer part and

$$m_0 = \sum_{i=1}^{n} \omega_{i0} = \lfloor n\gamma_{\text{sup}} \rfloor \tag{15}$$

be the total number of "good" observations. If we knew the weights ω_{i0}, the Mahalanobis distances of these observations would be

$$d^2_g = (y_g - \hat{\mu}_0)'\hat{\Sigma}_0^{-1}(y_g - \hat{\mu}_0) \quad g = 1, \ldots, m_0,$$

where

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} \omega_{i0} y_i}{m_0} \quad \text{and} \quad \hat{\Sigma}_0 = \frac{\sum_{i=1}^{n} \omega_{i0}(y_i - \hat{\mu}_0)(y_i - \hat{\mu}_0)'}{m_0 - 1}.$$

Under H_0, d^2_g is distributed as

$$\frac{(m_0 - 1)^2}{m_0}\text{Beta}\left(\frac{v}{2}, \frac{m_0 - v - 1}{2}\right) \tag{16}$$

for any m_0.

The distributional results of Cerioli (2010a) suggest that a scaled Beta distribution provides a satisfactory approximation also when the unknown weights ω_{i0} are estimated robustly by means of the reweighted MCD distances (3). Cerioli (2010b) shows empirical evidence supporting this approximation. Therefore, we base our goodness-of-fit analysis on the estimated distances

$$\hat{d}^2_g(\alpha) = \{y_g - \hat{\mu}_0(\alpha)\}'\hat{\Sigma}_0(\alpha)^{-1}\{y_g - \hat{\mu}_0(\alpha)\} \quad g = 1, \ldots, \hat{m}_0(\alpha), \tag{17}$$

Table 3
Outcome in testing n observations for outlyingness at level α.

Null hypotheses (20)			
	Not rejected	Rejected	Total
True	$N_{0\|0}$	$N_{1\|0}$	M_0
False	$N_{0\|1}$	$N_{1\|1}$	M_1
Total	$\hat{m}_0(\alpha)$	$R(\alpha)$	n

where $\hat{\mu}_0(\alpha)$ and $\hat{\Sigma}_0(\alpha)$ are the trimmed estimates of μ_0 and Σ_0 computed from the set of RMCD weights:

$$\hat{\omega}_{i0}(\alpha) = 0 \quad \text{if } d^2_{i(\text{RMCD})} > \text{cutoff}_\alpha \tag{18}$$

$$\hat{\omega}_{i0}(\alpha) = 1 \quad \text{otherwise}, \tag{19}$$

cutoff$_\alpha$ is the $(1-\alpha)$-quantile of the relevant distribution in (4) and (5), and $\hat{m}_0(\alpha) = \sum_{i=1}^n \hat{\omega}_{i0}(\alpha)$. In words, $\hat{d}^2_1(\alpha), \ldots, \hat{d}^2_{\hat{m}_0(\alpha)}(\alpha)$ are the distances of the units (suitably rearranged) that are not declared to be outliers at trimming level α by the RMCD rule using (4) and (5). Discussion about the choice of α is given later in this section and in Section 5, together with further implications of the regularity conditions stated in Lemma 1.

One crucial issue in the scaled Beta approximation to the distribution of $\hat{d}^2_g(\alpha), g = 1, \ldots, \hat{m}_0(\alpha)$, is that we need a reliable estimate of m_0 in (16). The plug-in estimator $\hat{m}_0(\alpha)$ is not adequate, since it does not take into account the effect of stochastic trimming in the outlier detection process. To solve this problem, we follow Cerioli and Farcomeni (2011) and approach outlier identification in a multiple testing framework, where the n individual hypotheses

$$H_{0i} : y_i \sim N(\mu_0, \Sigma_0), \quad i = 1, \ldots, n, \tag{20}$$

are tested in sequence by the RMCD rule using the $(1-\alpha)$-quantiles of (4) and (5). The full outcome of this process is shown in Table 3, where $R(\alpha)$ denotes the number of observations for which (20) is rejected, i.e. the number of nominated outliers, at the chosen level α. Asymptotically, this level satisfies condition (10).

We adopt an operational scheme in which observations are drawn at random from the contamination model (7). In this scheme, let V_i be the event H_{0i} `is true` and R_i be the event H_{0i} `is rejected`. If (8) holds, V_i is random with $\Pr\{V_i\} = \gamma_{\text{sup}}$. Therefore, in Table 3

$$M_0 = E\left\{\sum_{i=1}^n \mathbb{I}_{H_{0i} \text{ is true}}\right\} = n \Pr\{V_i\} = n\gamma_{\text{sup}} = m_0,$$

neglecting terms of order at most $1/n$.

If condition (11) is fulfilled, there is no asymptotic masking, i.e. we expect $N_{0|1} \approx 0$. Therefore,

$$E\{\hat{m}_0(\alpha)\} \le m_0$$

because the outlier detection rule is stochastic and $N_{1|0} \ge 0$. Specifically,

$$E\{\hat{m}_0(\alpha)\} = \lfloor m_0 - m_0\text{PCER} \rfloor = \lfloor m_0(1 - \text{PCER}) \rfloor, \tag{21}$$

where PCER is the Per-Comparison Error Rate (Farcomeni, 2008). In our framework

$$\text{PCER} = E(N_{1|0}/n) = E\left\{\sum_{i=1}^n \mathbb{I}_{[V_i \bigcap R_i]}\right\}/n = \Pr\left\{V_i \bigcap R_i\right\} = \Pr\{R_i|V_i\}\Pr\{V_i\}, \tag{22}$$

recalling that

$$E\left\{\mathbb{I}_{[V_i \bigcap R_i]}\right\} = \Pr\left\{V_i \bigcap R_i\right\}$$

is the same for all observations. The quantity

$$\Pr\{R_i|V_i\} = \alpha$$

is the (constant) probability of a Type-I error when testing each H_{0i} in (20). We also suggest to use

$$\widehat{\Pr}\{V_i\} = \frac{\hat{m}_0(\alpha)}{n},$$

as a preliminary estimate of $\Pr\{V_i\}$ in (22), thus yielding

$$\widehat{\text{PCER}} = \alpha\frac{\hat{m}_0(\alpha)}{n}.$$

Table 4
Estimates of m_0 under the null hypothesis (8) using 1000 replicates of (23) for $n = 200$, $v = 5$ (first row), $v = 10$ (second row) and different values of γ_{sup}, mean shift λ in the contaminant distribution (6) and trimming level α in the RMCD detection rule using (5) and (4). The last two values of α refer to simultaneous sizes 0.05 and 0.01 under Sidak correction.

α	$\gamma_{\text{sup}} = 1$	$\gamma_{\text{sup}} = 0.9$	$\gamma_{\text{sup}} = 0.9$	$\gamma_{\text{sup}} = 0.8$	$\gamma_{\text{sup}} = 0.8$	$\gamma_{\text{sup}} = 0.7$	$\gamma_{\text{sup}} = 0.7$
	$m_0 = 200$	$m_0 = 180$	$m_0 = 180$	$m_0 = 160$	$m_0 = 160$	$m_0 = 140$	$m_0 = 140$
	$\lambda = 0.0$	$\lambda = 2.5$	$\lambda = 5.0$	$\lambda = 2.5$	$\lambda = 5.0$	$\lambda = 2.5$	$\lambda = 5.0$
0.05	198.66	180.61	179.95	161.84	160.00	146.82	139.23
	198.91	180.06	179.87	159.89	159.80	139.54	139.16
0.01	199.42	181.34	180.40	164.60	160.71	155.84	139.94
	199.44	180.43	180.40	160.72	160.68	140.38	139.92
0.0002564	200.00	186.04	180.97	176.50	160.97	177.79	140.98
	200.00	181.16	180.95	161.52	160.96	143.39	140.98
0.0000503	200.00	188.87	181.00	182.28	161.00	185.19	141.00
	200.00	181.51	180.99	162.47	161.00	145.79	141.00

Table 5
Estimates of m_0 under the null hypothesis (8) using 1000 replicates of (23) for $n = 1000$, $v = 5$ (first row), $v = 10$ (second row) and different values of γ_{sup}, mean shift λ in the contaminant distribution (6) and trimming level α in the RMCD detection rule using (5) and (4). The last two values of α refer to simultaneous sizes 0.05 and 0.01 under Sidak correction.

α	$\gamma_{\text{sup}} = 1$	$\gamma_{\text{sup}} = 0.9$	$\gamma_{\text{sup}} = 0.9$	$\gamma_{\text{sup}} = 0.8$	$\gamma_{\text{sup}} = 0.8$	$\gamma_{\text{sup}} = 0.7$	$\gamma_{\text{sup}} = 0.7$
	$m_0 = 1000$	$m_0 = 900$	$m_0 = 900$	$m_0 = 800$	$m_0 = 800$	$m_0 = 700$	$m_0 = 700$
	$\lambda = 0.0$	$\lambda = 2.5$	$\lambda = 5.0$	$\lambda = 2.5$	$\lambda = 5.0$	$\lambda = 2.5$	$\lambda = 5.0$
0.05	996.31	898.19	897.44	798.58	796.16	701.27	694.76
	996.28	896.66	896.68	794.93	794.71	693.79	693.61
0.01	998.71	902.81	900.52	806.25	800.47	716.61	699.58
	998.81	900.47	900.28	800.11	799.93	699.39	699.32
0.0000513	1000.00	928.09	900.96	865.67	800.97	838.00	700.99
	1000.00	901.87	900.96	803.16	800.97	704.80	700.97
0.0000101	1000.00	941.01	900.99	893.60	800.99	882.00	701.00
	1000.00	903.32	900.99	806.27	800.99	709.99	701.00

From (21), our final estimator of m_0 is then

$$\hat{m}_0 = \left\lfloor \frac{\hat{m}_0(\alpha)}{1 - \widehat{\text{PCER}}} \right\rfloor + 1 = \left\lfloor \frac{n\hat{m}_0(\alpha)}{n - \alpha\hat{m}_0(\alpha)} \right\rfloor + 1. \tag{23}$$

Correspondingly,

$$\hat{\gamma}_{\text{sup}} = \frac{\widehat{\text{PCER}}}{\alpha(1 - \widehat{\text{PCER}})} = \frac{\hat{m}_0(\alpha)}{n - \alpha\hat{m}_0(\alpha)}.$$

The empirical performance of (23) is investigated in Table 4, for $n = 200$, and in Table 5, with $n = 1000$. In both cases, we report Monte Carlo estimates of $E(\hat{m}_0)$ under the null hypothesis of multivariate normality (8), with $G_0(y) = \Phi(y, 0, I_v)$ and $G_1(y)$ the distribution function of (6), for different values of v, α, γ_{sup} and λ.

It is clearly seen that the empirical performance of $\hat{m}_0$ is very satisfactory in most cases. As expected from Lemma 1, the mean of our estimator is close to the nominal target in all the simulations settings for which $c_{G1} \approx 1$. The additional good news is that performance remains satisfactory even in many situations where $c_{G1} \ll 1$ in Table 2. In those instances, the upper bound provided by (13) is clearly too conservative. Positive bias of $\hat{m}_0$ can become a problem, with strongly overlapping populations, only when the trimming probability α is small, as in the case of Sidak correction, the sample size is large and v is small. These unfavourable settings correspond to the most extreme conditions under which Lemma 1 does not hold. With no contamination ($\gamma_{\text{sup}} = 1$), or with two well separated populations, the bias may become negative if α increases, due to the asymptotic nature of the correction factor k_{RMCD} in (2). However, this effect is seen to be negligible even with α as large as 0.05.

An alternative estimate of the weights ω_{i0} is suggested by Cerioli (2010b), who uses the first-stage robust MCD distances, instead of the reweighted ones, in (18) and (19). Computational simplicity is the main motivation for that estimate, which can be implemented as a byproduct of the RMCD computations in (2). However, the results for that estimate are slightly inferior, albeit similar, to those reported in Tables 4 and 5, and we do not consider them further. The naive procedure using $\hat{m}(\alpha)$ in Table 3 as the final estimate of m_0 yields systematic underestimation of the true value if $c_{G1} \approx 1$. Also in this case we do not give the results in detail, but we postpone evidence of this effect to Section 5.2.

5. Robust chi-square tests of multivariate normality

We now describe our robust chi-square tests of multivariate normality after outlier removal. The null hypothesis H_0 is stated in (8). Our choice of the Pearson statistic is motivated by the minor impact of parameter estimation, as shown in Table 1 for $\gamma_{\text{sup}} = 1$.

Table 6

Estimates of the size of the test of H_0 using $X^2\{\mathrm{RMCD}_{0.975}(\alpha)\}$ in the case $n = 200$, $v = 5$ (first row), $v = 10$ (second row) and $K = 20$, for a nominal test size $\alpha_{\mathrm{GOF}} = 0.05$. α is the nominal size for the RMCD outlier detection rule using (5) and (4). 1000 simulations for each value of $\gamma_{\sup}$ and mean shift λ in the contaminant distribution.

α	$\gamma_{\sup} = 1$	$\gamma_{\sup} = 0.9$	$\gamma_{\sup} = 0.9$	$\gamma_{\sup} = 0.8$	$\gamma_{\sup} = 0.8$	$\gamma_{\sup} = 0.7$	$\gamma_{\sup} = 0.7$
	$m_0 = 200$	$m_0 = 180$	$m_0 = 180$	$m_0 = 160$	$m_0 = 160$	$m_0 = 140$	$m_0 = 140$
	$\lambda = 0.0$	$\lambda = 2.5$	$\lambda = 5.0$	$\lambda = 2.5$	$\lambda = 5.0$	$\lambda = 2.5$	$\lambda = 5.0$
0.05	0.033	0.039	0.044	0.029	0.034	0.034	0.047
	0.037	0.044	0.038	0.044	0.036	0.040	0.045
0.01	0.041	0.042	0.030	0.033	0.040	0.063	0.039
	0.033	0.041	0.037	0.041	0.042	0.045	0.038
0.0002564	0.041	0.058	0.029	0.076	0.034	0.083	0.040
	0.037	0.032	0.044	0.049	0.037	0.042	0.043
0.0000503	0.041	0.099	0.030	0.084	0.035	0.058	0.040
	0.033	0.038	0.043	0.050	0.037	0.056	0.042

Table 7

Estimates of the size of the test of H_0 using $X^2\{\mathrm{RMCD}_{0.975}(\alpha)\}$ in the case $n = 1000$, $v = 5$ (first row), $v = 10$ (second row) and $K = 30$, for a nominal test size $\alpha_{\mathrm{GOF}} = 0.05$. α is the nominal size for the RMCD outlier detection rule using (5) and (4). 1000 simulations for each value of $\gamma_{\sup}$ and mean shift λ in the contaminant distribution.

α	$\gamma_{\sup} = 1$	$\gamma_{\sup} = 0.9$	$\gamma_{\sup} = 0.9$	$\gamma_{\sup} = 0.8$	$\gamma_{\sup} = 0.8$	$\gamma_{\sup} = 0.7$	$\gamma_{\sup} = 0.7$
	$m_0 = 1000$	$m_0 = 900$	$m_0 = 900$	$m_0 = 800$	$m_0 = 800$	$m_0 = 700$	$m_0 = 700$
	$\lambda = 0.0$	$\lambda = 2.5$	$\lambda = 5.0$	$\lambda = 2.5$	$\lambda = 5.0$	$\lambda = 2.5$	$\lambda = 5.0$
0.05	0.046	0.041	0.035	0.041	0.034	0.046	0.042
	0.041	0.037	0.028	0.031	0.055	0.043	0.051
0.01	0.039	0.027	0.033	0.048	0.021	0.098	0.039
	0.039	0.052	0.042	0.030	0.034	0.038	0.029
0.0000513	0.041	0.425	0.038	0.553	0.034	0.254	0.029
	0.038	0.036	0.047	0.030	0.042	0.052	0.044
0.0000101	0.041	0.780	0.040	0.471	0.034	0.182	0.030
	0.037	0.035	0.046	0.058	0.043	0.099	0.043

5.1. *The* $X^2\{\mathrm{RMCD}_{0.975}(\alpha)\}$ *test*

This goodness-of-fit test compares the empirical distribution of distances (17) with the quantiles of a truncated scaled Beta distribution. The acronym recalls that, before the chi-square test, outliers are removed at level α by means of the robust reweighted distances $d^2_{i(\mathrm{RMCD})}$, after preliminary trimming at probability 0.975 in the MCD step.

The test statistic is defined as follows:

$$X^2\{\mathrm{RMCD}_{0.975}(\alpha)\} = \sum_{k=1}^{K} \frac{\left[n_k\{\hat{d}^2(\alpha)\} - n_k\{\hat{d}^2\}\right]^2}{n_k\{\hat{d}^2\}}, \tag{24}$$

where K is the number of classes in which the observations are partitioned, $n_k\{\hat{d}^2(\alpha)\}$ is the number of units, among $\hat{m}_0(\alpha)$, for which distance (17) falls within class k and $n_k\{\hat{d}^2\}$ is the predicted number of such units under the null distribution. According to (16), our estimate of the null distribution of the distances is

$$\frac{\hat{m}_0}{\hat{m}_0(\alpha)} \Pr\left\{ \frac{(\hat{m}_0 - 1)^2}{\hat{m}_0} \mathrm{Beta}\left(\frac{v}{2}, \frac{\hat{m}_0 - v - 1}{2}\right) \le \hat{d}^2 \right\},$$

where $\hat{m}_0$ is the estimate of the number of "good" units given in Eq. (23). The factor $\hat{m}_0/\hat{m}_0(\alpha)$ allows for the effect of stochastic trimming among the units for which H_0 is true, since we can only observe $\hat{m}_0(\alpha) \le m_0$ such units. The distribution of the corresponding distances $\hat{d}^2_g(\alpha)$, $g = 1, \ldots, \hat{m}_0(\alpha)$, is thus a truncated scaled Beta distribution and the probability of truncation is estimated by $\hat{m}_0(\alpha)/\hat{m}_0$. We compare our test statistic (24) to the slightly conservative χ^2_{K-1} approximation, as in Table 1. In our computations, we take equiprobable classes under H_0 and we choose values of K close to the common practical recommendation $K = 2n^{2/5}$.

Table 6 reports the estimated size of the test of H_0 using $X^2\{\mathrm{RMCD}_{0.975}(\alpha)\}$ for $n = 200$, $K = 20$, $v = 5$ and $v = 10$, and different trimming levels α, when the nominal size is $\alpha_{\mathrm{GOF}} = 0.05$. In this simulation $G_0(y) = \Phi(y, 0, I_v)$ and $G_1(y)$ is the distribution of the shifted model (6). Table 7 repeats the analysis in the case $n = 1000$. The overall structure of the results parallels the findings of Section 4, but with the additional insight provided by precise quantification of the effect of estimating m_0 on test sizes. We conclude that the null performance of our test is akin to that of the classic procedure when $\gamma_{\sup} = 1$, but much better if $\gamma_{\sup} < 1$. Our method is robust against outliers and is able to control the size of the test of

Table 8
Estimates of the size of the test of H_0 using the naive approach with trimming level $\alpha = 0.05$, for a nominal test size $\alpha_{\mathrm{GOF}} = 0.05$. For each n, the first row refers to $v = 5$, and the second row to $v = 10$. 1000 simulations for each value of $\gamma_{\sup}$ and mean shift λ in the contaminant distribution.

	$\gamma_{\sup} = 1$ $\lambda = 0.0$	$\gamma_{\sup} = 0.9$ $\lambda = 2.5$	$\gamma_{\sup} = 0.9$ $\lambda = 5.0$	$\gamma_{\sup} = 0.8$ $\lambda = 2.5$	$\gamma_{\sup} = 0.8$ $\lambda = 5.0$	$\gamma_{\sup} = 0.7$ $\lambda = 2.5$	$\gamma_{\sup} = 0.7$ $\lambda = 5.0$
$n = 200$	0.136	0.102	0.135	0.076	0.119	0.082	0.099
	0.171	0.167	0.140	0.132	0.111	0.104	0.100
$n = 1000$	1.000	0.985	0.987	0.878	0.919	0.551	0.807
	1.000	0.997	0.999	0.979	0.967	0.887	0.879

multivariate normality in most of the selected parameter settings. Contamination becomes a serious problem only if α is very small, as in the case of simultaneous testing corrections, and when n is large, as anticipated by Lemma 1. We again see evidence of the "blessing of dimensionality" effect when $v = 10$.

5.2. The effect of trimming

The naive approach to robustness, which is sometimes advocated in textbooks or in applied work (Singh, 1993; Gnanadesikan, 1997, p. 296), is to apply the standard methods to the observations that remain after outlier removal, without taking the effect of trimming into account. This reduces to computation of (24) with the theoretical frequencies $n_k\{\hat{d}^2\}$, $k = 1, \ldots, K$, obtained from the scaled Beta distribution

$$\frac{\{\hat{m}_0(\alpha) - 1\}^2}{\hat{m}_0(\alpha)} \mathrm{Beta}\left(\frac{v}{2}, \frac{\hat{m}_0(\alpha) - v - 1}{2}\right),$$

which takes $\hat{m}_0(\alpha)$ from Table 1 as the final estimate of m_0.

Table 8 shows the null performance of the naive approach, in the case of trimming level $\alpha = 0.05$, for the same simulation setting considered in Section 5.1. Comparison with Tables 6 and 7 shows that neglecting the effect of trimming can be very dangerous, especially if n grows. Remarkably, there is no "blessing of dimensionality" in this case, as performance deteriorates as v increases from 5 to 10.

5.3. FDR control in outlier detection

The theoretical result of Lemma 1 and the simulation evidence of Sections 4 and 5.1 do not support the use of strong multiplicity adjustments, like the Sidak correction, in the outlier detection step. Furthermore, the bias in the goodness-of-fit procedure possibly introduced by multiplicity adjustments is higher for large n. These findings suggest the opportunity of adopting outlier detection rules that only provide weak control of simultaneity in repeated testing of the n individual hypotheses (20). One example of weak control is the choice of the False Discovery Rate (FDR) of Benjamini and Hochberg (1995) as the relevant penalty. A potential advantage of this approach over PCER control is an increase of power due to reduced trimming when H_0 is false.

Cerioli and Farcomeni (2011) show how FDR control can be exploited for the purpose of multivariate outlier detection, providing a sensible compromise between high power and low swamping. Their procedure computes the p-values of the robust reweighted distances $d^2_{i(\mathrm{RMCD})}$, $i = 1, \ldots, n$, according to distributions (4) and (5). FDR control leads to an alternative estimate of the trimming weights (15):

$$\begin{aligned} \tilde{\omega}_{i0}(\alpha) &= 0 \quad \text{if } p_i < \rho_i \alpha / n \\ \tilde{\omega}_{i0}(\alpha) &= 1 \quad \text{otherwise,} \end{aligned} \tag{25}$$

where p_i is the p-value of $d^2_{i(\mathrm{RMCD})}$ and ρ_i is the rank of p_i. Now α denotes the chosen threshold of FDR. Let $\tilde{m}_0(\alpha) = \sum_{i=1}^{n} \tilde{\omega}_{i0}(\alpha)$. Our FDR-based estimator of m_0 is

$$\tilde{m}_0 = \left\lfloor \frac{n\tilde{m}_0(\alpha)}{n - \left[\{n - \tilde{m}_0(\alpha)\}\frac{\alpha}{n}\right]\tilde{m}_0(\alpha)} \right\rfloor + 1. \tag{26}$$

The derivation of $\tilde{m}_0$ is similar to that of (23), with PCER now estimated by

$$\widetilde{\mathrm{PCER}} = \left[\{n - \tilde{m}_0(\alpha)\}\frac{\alpha}{n}\right]\frac{\tilde{m}_0(\alpha)}{n}.$$

This PCER estimate relies on the approximation

$$\Pr\{R_{(i)} | V_{(i)}\} \approx \{n - \tilde{m}_0(\alpha)\}\frac{\alpha}{n},$$

Table 9
Estimates of the size of the test of H_0 using $X^2\{\text{FDR–RMCD}_{0.975}(\alpha)\}$, for a nominal test size $\alpha_{\text{GOF}} = 0.05$. For each n, the first row refers to $v = 5$, and the second row to $v = 10$. Here, α is the nominal value of the FDR to be controlled in the outlier detection rule using (5) and (4). 1000 simulations for each value of γ_{sup} and mean shift λ in the contaminant distribution.

α	$\gamma_{\text{sup}} = 1$	$\gamma_{\text{sup}} = 0.9$ $\lambda = 2.5$	$\gamma_{\text{sup}} = 0.9$ $\lambda = 5.0$	$\gamma_{\text{sup}} = 0.8$ $\lambda = 2.5$	$\gamma_{\text{sup}} = 0.8$ $\lambda = 5.0$	$\gamma_{\text{sup}} = 0.7$ $\lambda = 2.5$	$\gamma_{\text{sup}} = 0.7$ $\lambda = 5.0$
	$n = 200$						
0.05	0.032	0.044	0.036	0.042	0.035	0.052	0.041
	0.033	0.038	0.031	0.035	0.029	0.045	0.044
0.01	0.035	0.048	0.024	0.052	0.041	0.069	0.039
	0.035	0.041	0.043	0.037	0.047	0.030	0.053
	$n = 1000$						
0.05	0.038	0.024	0.034	0.057	0.028	0.070	0.039
	0.043	0.034	0.040	0.037	0.037	0.032	0.032
0.01	0.040	0.046	0.037	0.133	0.031	0.240	0.029
	0.038	0.034	0.040	0.024	0.036	0.037	0.043

where $\Pr\{R_{(i)}|V_{(i)}\}$ is the probability of a false rejection for the ith ordered p-value and $n-\tilde{m}_0(\alpha)$ is the number of observations declared to be outliers by (25).

The relevant distances for goodness-of-fit testing are now written as $\tilde{d}_g^2(\alpha)$, $g = 1, \ldots, \tilde{m}_0(\alpha)$, and are the analogue of (17). In obvious notation, the resulting goodness-of-fit statistic is

$$X^2\{\text{FDR–RMCD}_{0.975}(\alpha)\} = \sum_{k=1}^{K} \frac{\left[n_k\{\tilde{d}^2(\alpha)\} - n_k\{\tilde{d}^2\}\right]^2}{n_k\{\tilde{d}^2\}}, \tag{27}$$

with the estimated theoretical frequencies $n_k\{\tilde{d}^2\}$ obtained from the distribution

$$\frac{\tilde{m}_0}{\tilde{m}_0(\alpha)} \Pr\left\{\frac{(\tilde{m}_0-1)^2}{\tilde{m}_0}\text{Beta}\left(\frac{v}{2}, \frac{\tilde{m}_0 - v - 1}{2}\right) \le \tilde{d}^2\right\}.$$

The test statistic is again compared to the χ^2_{K-1} distribution.

We omit the details about the (generally good) properties of (26) as an estimator of m_0 and we directly show evidence of the null behaviour of $X^2\{\text{FDR–RMCD}_{0.975}(\alpha)\}$, in the same simulation setting as given in Section 5.1. Table 9 provides the results for a nominal test size $\alpha_{\text{GOF}} = 0.05$. It is seen that weak control of simultaneity of outlier tests, as guaranteed by the FDR trimming rule, does not alter the null properties of the goodness-of-fit procedure when $\gamma_{\text{sup}} = 1$. If the data are contaminated, liberality of $X^2\{\text{FDR–RMCD}_{0.975}(\alpha)\}$ becomes a serious problem only under extreme overlapping and with $\alpha = 0.01$. We thus argue that FDR trimming of outliers can be a sensible solution also for the purpose of robust goodness-of-fit testing, provided that α is not chosen to be too small.

5.4. Alternative contamination models

In Sections 5.1–5.3 the contamination distribution $G_1(y)$ of model (7) has been taken to be unimodal, so that we have only described the case of clustered outliers. We now show the performance of our approach when the outliers are more dispersed. Specifically, we consider two alternative contamination schemes. The first one is given by a bimodal model where $G_1(y)$ is the distribution function of

$$\pi_1 N(0 + \lambda_1 e, I_v) + \pi_2 N(0 + \lambda_2 e, I_v), \tag{28}$$

with $\lambda_2 > \lambda_1 > 0$ and mixing proportions $\pi_1 = \pi_2 = (1 - \gamma_{\text{sup}})/2$. Our second alternative contamination model is that of radial contamination, where $G_1(y)$ is the distribution function of

$$N(0, \psi I_v), \tag{29}$$

for $\psi > 1$.

Table 10 displays the results under model (28) in the case of $n = 200$, $v = 5$ and $v = 10$, for $\alpha_{\text{GOF}} = 0.05$, $\lambda_1 = 2.5$ and different values of the second group mean λ_2. Table 11 repeats the analysis under model (29) for different choices of the variance inflation factor ψ. In both instances the main conclusions remain unaltered with respect to those we have seen under the unimodal shift contamination model (6). Our test procedures have a satisfactory performance in most of the selected simulation settings, the only exceptions being when $G_0(y)$ and $G_1(y)$ strongly overlap and the amount of trimming is not sufficient to separate them. Furthermore, we see that control over $\alpha_{\text{GOF}} = 0.05$ increases with the degree of separation between $G_0(y)$ and $G_1(y)$, as predicted by Lemma 1. The dimension v again has a beneficial effect for this purpose.

Table 10
Estimates of the size of the test of H_0 using $X^2\{\text{RMCD}_{0.975}(\alpha)\}$ and $X^2\{\text{FDR–RMCD}_{0.975}(\alpha)\}$ under model (28), with $\alpha_{\text{GOF}} = 0.05$, $n = 200$ and $\lambda_1 = 2.5$, for different values of λ_2. For each α, the first row refers to $v = 5$, and the second row to $v = 10$. 1000 simulations for each value of γ_{sup} and λ_2.

α	$\gamma_{\text{sup}} = 0.9$	$\gamma_{\text{sup}} = 0.9$	$\gamma_{\text{sup}} = 0.8$	$\gamma_{\text{sup}} = 0.8$	$\gamma_{\text{sup}} = 0.7$	$\gamma_{\text{sup}} = 0.7$
	$\lambda_2 = 3.5$	$\lambda_2 = 5.0$	$\lambda_2 = 3.5$	$\lambda_2 = 5.0$	$\lambda_2 = 3.5$	$\lambda_2 = 5.0$
	$X^2\{\text{RMCD}_{0.975}(\alpha)\}$					
0.05	0.034	0.041	0.035	0.035	0.030	0.049
	0.043	0.048	0.035	0.035	0.045	0.044
0.01	0.040	0.029	0.038	0.039	0.056	0.056
	0.034	0.038	0.022	0.045	0.043	0.032
	$X^2\{\text{FDR–RMCD}_{0.975}(\alpha)\}$					
0.05	0.040	0.030	0.049	0.038	0.051	0.061
	0.035	0.031	0.026	0.039	0.035	0.038
0.01	0.035	0.033	0.047	0.052	0.064	0.054
	0.035	0.045	0.037	0.047	0.038	0.045

Table 11
Estimates of the size of the test of H_0 using $X^2\{\text{RMCD}_{0.975}(\alpha)\}$ and $X^2\{\text{FDR–RMCD}_{0.975}(\alpha)\}$ under model (29), with $\alpha_{\text{GOF}} = 0.05$ and $n = 200$, for different values of ψ. For each α, the first row refers to $v = 5$, and the second row to $v = 10$. 1000 simulations for each value of γ_{sup} and ψ.

α	$\gamma_{\text{sup}} = 0.9$	$\gamma_{\text{sup}} = 0.9$	$\gamma_{\text{sup}} = 0.8$	$\gamma_{\text{sup}} = 0.8$	$\gamma_{\text{sup}} = 0.7$	$\gamma_{\text{sup}} = 0.7$
	$\psi = 9$	$\psi = 16$	$\psi = 9$	$\psi = 16$	$\psi = 9$	$\psi = 16$
	$X^2\{\text{RMCD}_{0.975}(\alpha)\}$					
0.05	0.038	0.049	0.034	0.030	0.052	0.042
	0.039	0.034	0.031	0.029	0.042	0.027
0.01	0.041	0.039	0.038	0.038	0.104	0.037
	0.038	0.040	0.030	0.035	0.041	0.041
	$X^2\{\text{FDR–RMCD}_{0.975}(\alpha)\}$					
0.05	0.038	0.033	0.035	0.034	0.084	0.030
	0.041	0.048	0.029	0.046	0.043	0.037
0.01	0.047	0.041	0.095	0.049	0.298	0.057
	0.027	0.042	0.041	0.035	0.035	0.038

6. Power of the robust tests

We evaluate power for the robust goodness-of-fit procedures that have shown the best performance under H_0 in the different simulation settings. Therefore, we compare

- $X^2\{\text{RMCD}_{0.975}(0.05)\}$, i.e. test statistic (24) with $\alpha = 0.05$
- $X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$, i.e. test statistic (27) with $\alpha = 0.05$.

When $v = 10$, we also consider

- $X^2\{\text{RMCD}_{0.975}(0.05S)\}$, i.e. test statistic (24) with multiplicity-adjusted trimming level $\alpha = 1 - (1 - 0.05)^{1/n}$.

As a reference, we also include in our comparisons the non-robust chi-square test (X^2) described in Table 1. We estimate power by the proportion of simulations in which the null hypothesis (8) is correctly rejected. Computations are based on 1000 independent simulations for each parameter setting. We consider alternative hypotheses defined both by heavy-tailed and skew distributions.

6.1. *Multivariate t alternative*

Our first power scenario under the alternative hypothesis is run with

$$G_0(y) = G_1(y) = T_\nu,$$

where T_ν is the distribution function of the v-variate t distribution with ν degrees of freedom.

Table 12 displays the results for the case $n = 200$, with $\alpha_{\text{GOF}} = 0.05$ and $K = 20$. Table 13 repeats the analysis for $n = 1000$ and $K = 30$. The non-robust test X^2 has obviously the highest power, but it is not an eligible procedure in view of the results of Table 1. Nevertheless, it provides a useful benchmark for evaluating the loss of power which is due to trimming. It is seen that the power of all robust tests increases with n and v, as it should. The price to pay for robustness can be considerable if outlier detection is done at individual level $\alpha = 0.05$, but becomes minor for the simultaneous rules

Table 12
Estimates of the power of the test of H_0 using different statistics for $n = 200$ and $K = 20$, when all the observations are generated from a multivariate t distribution with ν degrees of freedom. 1000 simulations for each value of v and ν. The nominal test size is $\alpha_{\text{GOF}} = 0.05$.

		$\nu = 2$	$\nu = 6$	$\nu = 10$	$\nu = 16$
$v = 5$	X^2	1.000	0.985	0.672	0.259
	$X^2\{\text{RMCD}_{0.975}(0.05)\}$	0.740	0.248	0.127	0.061
	$X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$	0.925	0.624	0.410	0.197
$v = 10$	X^2	1.000	1.000	0.994	0.783
	$X^2\{\text{RMCD}_{0.975}(0.05)\}$	0.969	0.700	0.383	0.168
	$X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$	0.994	0.953	0.814	0.593
	$X^2\{\text{RMCD}_{0.975}(0.05S)\}$	1.000	0.999	0.957	0.693

Table 13
Estimates of the power of the test of H_0 using different statistics for $n = 1000$ and $K = 30$, when all the observations are generated from a multivariate t distribution with ν degrees of freedom. 1000 simulations for each value of v and ν. The nominal test size is $\alpha_{\text{GOF}} = 0.05$.

		$\nu = 2$	$\nu = 6$	$\nu = 10$	$\nu = 16$
$v = 5$	X^2	1.000	1.000	1.000	0.970
	$X^2\{\text{RMCD}_{0.975}(0.05)\}$	1.000	0.872	0.451	0.192
	$X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$	1.000	1.000	0.995	0.908
$v = 10$	X^2	1.000	1.000	1.000	1.000
	$X^2\{\text{RMCD}_{0.975}(0.05)\}$	1.000	1.000	0.986	0.807
	$X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$	1.000	1.000	1.000	1.000
	$X^2\{\text{RMCD}_{0.975}(0.05S)\}$	1.000	1.000	1.000	1.000

Table 14
Estimates of the power of the test of H_0 using different statistics for $n = 200$ and $K = 20$, when all the observations are generated from a multivariate Gaussian copula distribution with χ^2_ν univariate marginals and common correlation ρ. 1000 simulations for each value of ρ and ν. The nominal test size is $\alpha_{\text{GOF}} = 0.05$.

		$\nu = 2$	$\nu = 6$	$\nu = 10$	$\nu = 16$
$\rho = 0.1$	X^2	1.000	0.657	0.258	0.115
	$X^2\{\text{RMCD}_{0.975}(0.05)\}$	0.478	0.073	0.049	0.034
	$X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$	0.676	0.246	0.136	0.083
$\rho = 0.9$	X^2	1.000	0.999	0.830	0.409
	$X^2\{\text{RMCD}_{0.975}(0.05)\}$	0.998	0.315	0.109	0.054
	$X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$	0.998	0.496	0.270	0.172

adopted in $X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$ and $X^2\{\text{RMCD}_{0.975}(0.05S)\}$. Our simulation findings thus support the conclusion that a simultaneous approach to outlier detection in (20) enhances the power of the subsequent robust goodness-of-fit test. From this point of view, weak control of multiplicity, as in $X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$, can provide a sensible compromise between size and power requirements, especially when the number of variables is not very high.

6.2. Multivariate skew alternative

Our second power scenario is

$$G_0(y) = G_1(y) = S_\nu,$$

where S_ν is the distribution function of a v-variate random variable defined through a Gaussian copula function from univariate χ^2_ν marginals. The one-dimensional marginal distributions of S_ν are thus skew, as are the bivariate and higher-order marginals. The univariate marginal distributions are also dependent, with correlation ρ.

Our simulation setting for S_ν is similar to that for T_ν, with the additional complexity induced by the possible influence of different correlation values. Therefore, we only report the results obtained in the case $v = 5$, for which the effect of trimming is larger, for varying degrees of freedom ν and two values ρ associated to low and high correlation. Table 14 refers to the case $n = 200$, while Table 15 is for $n = 1000$. Power against this skew alternative is generally lower than power against the multivariate t distribution, not only for the robust tests, but also for X^2. However, the relative performance of the different procedures repeats almost exactly what we have already learnt in Section 6.1. The magnitude of power also increases with both n and ρ. We thus conclude that our robust goodness-of-fit procedures, and particularly the one based on FDR trimming of outliers, can be effective also for the purpose of detecting multivariate skew alternatives.

Table 15
Estimates of the power of the test of H_0 using different statistics for $n = 1000$ and $K = 30$, when all the observations are generated from a multivariate Gaussian copula distribution with χ^2_ν univariate marginals and common correlation ρ. 1000 simulations for each value of ρ and ν. The nominal test size is $\alpha_{\text{GOF}} = 0.05$.

		$\nu = 2$	$\nu = 6$	$\nu = 10$	$\nu = 16$
$\rho = 0.1$	X^2	1.000	1.000	0.979	0.719
	$X^2\{\text{RMCD}_{0.975}(0.05)\}$	0.000	0.216	0.074	0.066
	$X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$	0.000	0.844	0.671	0.427
$\rho = 0.9$	X^2	0.000	1.000	1.000	0.999
	$X^2\{\text{RMCD}_{0.975}(0.05)\}$	0.000	0.904	0.320	0.091
	$X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$	0.000	0.996	0.899	0.765

Table 16
Example 1: Empirical results of robust goodness-of-fit testing with $K = 30$ and different trimming levels. Pearson X^2 statistics with p-values in parentheses.

	Contaminated normal sample	Multivariate t sample
$X^2\{\text{RMCD}_{0.975}(0.05)\}$	33.97 (0.240)	44.89 (0.030)
$X^2\{\text{RMCD}_{0.975}(0.01)\}$	24.20 (0.719)	67.93 (<0.001)
$X^2\{\text{FDR–RMCD}_{0.975}(0.05)\}$	21.30 (0.848)	77.20 (<0.001)
$X^2\{\text{FDR–RMCD}_{0.975}(0.01)\}$	39.04 (0.101)	109.55 (<0.001)

7. Data analysis

7.1. Example 1: simulated data of Section 1

We first apply our robust test of multivariate normality to the motivating example of Section 1 based on simulated data, whose exploratory analysis was displayed in Fig. 1. The standard goodness-of-fit approach, comparing the classical Mahalanobis distances (9) to their nominal distribution

$$\{(1000-1)^2/1000\}\text{Beta}(2.5, 497),$$

produces the highly significant Pearson statistic $X^2 = 54.98$ for the contaminated normal sample, based on $K = 30$ classes. A similar conclusion is reached for the multivariate t sample, where $X^2 = 122.72$.

The goodness-of-fit results obtained through our robust approach, again with $K = 30$, are given in Table 16 for alternative trimming choices. Now the difference between the two data structures is paramount. Even for the highest level of trimming $\alpha = 0.05$, both our tests lead to the correct conclusion that the hypothesis of multivariate normality for the bulk of the data is acceptable in one instance, but has to be rejected in the other case. It is worth noting that the naive approach, which does not take the effect of trimming into account, also fails in this example, leading to a significant value of X^2 for the contaminated normal sample when the outliers are removed.

One referee has pointed out that some people could be comfortable with the decision of rejecting the normality hypothesis if a dataset is nicely generated from a normal distribution and some outliers are present, as in the contaminated normal sample of this example. Nevertheless, we emphasize that the classical test is not able to tell if the majority of the data actually follow the multivariate normal distribution. It is only through our robust procedure that it is possible to divide the goodness-of-fit problem into its main components: the behaviour of the bulk of the data and the effect of the outliers. Another major disadvantage of the classical test is masking, which may obscure the existence of a nice normal structure contaminated by outliers. The effect of masking is made clear in the following example.

7.2. Forged Swiss banknotes

Flury and Riedwyl (1988) introduce data on six variables measuring the size and other features of 200 Swiss banknotes, 100 of which are classified as genuine and 100 as forged. The notes were withdrawn from circulation and then classified by an expert. From the quality process, we may expect the sample of genuine notes to be homogeneous and normally distributed, except perhaps for a couple of possibly misclassified notes (Atkinson et al., 2004, pp. 116–137). On the other hand, the group of forged notes is known to be heterogeneous, perhaps due to the action of different forgers. Heterogeneity leads to the presence of at least 15 well identified outliers (Garcìa-Escudero and Gordaliza, 2005; Cerioli, 2010a). We thus analyse the sample of forged notes through our robust goodness-of-fit approach, with the aim of checking if the normality assumption is tenable for the bulk of forged notes as well.

Fig. 3 displays Q–Q plots of the squared Mahalanobis distances (9) and of their robust counterparts (3), against the asymptotic χ^2_6 distribution. Inspection of the robust distances clearly shows the existence of a number of outliers, which

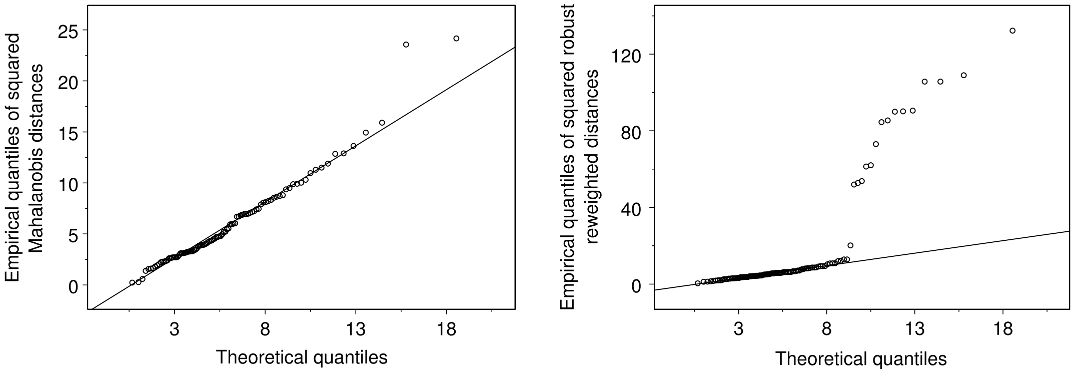

Fig. 3. Forged Swiss bank notes: Q–Q plots of Mahalanobis (left) and robust (right) squared distances.

are instead masked in the classical distances. Although the right-hand panel may suggest the idea of a normally generated sample with some outliers added, as in Example 1, the classical distances are unable to reveal this structure. Therefore, we conclude that the standard goodness-of-fit approach based on (9), yielding the non-significant value $X^2 = 5.8$ with $K = 10$ classes, is unreliable in this example. On the contrary, the Pearson statistic for the robust distances is $X^2 = 21.8$ a largely significant outcome.

Our robust test takes a different view, by separating the effect of the outliers in the right-hand panel of Fig. 3 from that of the majority of the data. The results for trimming level $\alpha = 0.05$ and $K = 10$ are (p-value in parentheses)

$$X^2\{\text{RMCD}_{0.975}(0.05)\} = 10.12\ (0.341);$$
$$X^2\{\text{FDR–RMCD}_{0.975}(0.05)\} = 3.72\ (0.929).$$

They confirm that the "best" forged notes in this sample actually follow the multivariate normal model typical of genuine notes.

8. Concluding remarks

In this paper we have developed a robust distance-based procedure for the purpose of testing multivariate normality with contaminated data. Our approach is made up of two steps, the first being accurate outlier removal through appropriate cut-off values for the robust distances, and the second being careful goodness-of-fit testing on the supposedly uncontaminated observations. We have shown that, with our approach, the first step does not alter the size of the goodness-of-fit test under regularity conditions on the contaminant distribution. Furthermore, we have provided evidence that our method has good power properties under different alternatives, including the heavy-tailed multivariate t and a class of skew multivariate distributions. Finally, we have given further support to our approach by analysing some motivating examples based on real and simulated data.

From a methodological point of view, our proposal contains two main contributions. The first one is that we develop a way to take into account the effect of stochastic trimming induced by the outlier removal process. Our consideration of this effect is twofold. First, we have shown that stochastic trimming is an important ingredient for the purpose of obtaining a reliable estimate of the number of uncontaminated observations. Then, we have allowed for trimming in the empirical distribution of the robust distances when performing the goodness-of-fit part of our procedure.

Our second contribution concerns the choice of the error rate to be controlled when removing the outliers. We have shown that the two conflicting goals of any outlier detection rule, i.e. having high power and low swamping, also affect the performance of our robust goodness-of-fit technique. With normally distributed uncontaminated observations our suggestion would be to accept a higher degree of swamping and to trim a larger number of observations, in order to achieve better separation between the "good" and the "bad" part of the data. On the other hand, when the hypothesis of multivariate normality does not hold, power increases with lower levels of trimming. Our conclusion is that control of the False Discovery Rate in the outlier removal process, as suggested by Cerioli and Farcomeni (2011), can provide a sensible balance between size and power properties in our robust goodness-of-fit approach.

Acknowledgements

The authors are grateful to an Associate Editor and two anonymous reviewers for useful remarks that helped to improve the presentation. The authors also thank Prof. Anthony C. Atkinson for helpful comments on previous drafts of this work. This research was partially supported by the grants "La valutazione statistica della qualità e dei rischi nei servizi: nuove metodologie multivariate" (A. Cerioli) and "Nuove metodologie robuste per l'analisi di dati complessi" (M. Riani) of Ministero dell'Università e della Ricerca—PRIN 2008.

References

Alfons, A., Baaske, W.E., Filzmoser, P., Mader, W., Wieser, R., 2011. Robust variable selection with application to quality of life research. Statistical Methods and Applications 20, 65–82.

Alvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A., Matrán, C., 2010. Assessing when a sample is mostly normal. Computational Statistics and Data Analysis 54, 2914–2925.

Atkinson, A.C., Riani, M., Cerioli, A., 2004. Exploring Multivariate Data with the Forward Search. Springer, New York.

Beirlant, J., Mason, D.M., Vynckier, C., 1999. Goodness-of-fit analysis for multivariate normality based on generalized quantiles. Computational Statistics and Data Analysis 30, 119–142.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society (Series B) 57, 289–300.

Bordes, L., Mottelet, S., Vandekerkhove, P., 2006. Semiparametric estimation of a two-component mixture model. The Annals of Statistics 34, 1204–1232.

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

Cerioli, A., 2010b. Diagnostic checking of multivariate normality under contamination. In: Lechevallier, Y., Saporta, G. (Eds.), Proceedings of COMPSTAT'2010. Physica-Verlag, Heidelberg, pp. 871–878.

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

Cerioli, A., Riani, M., Torti, F., 2012. Size and power of multivariate outlier detection rules. In: Lausen, B., van den Poel, D. and Ultsch, A. (Eds.), Proceedings of GfKl/IFCS 2011, Springer, Berlin (in press).

Croux, C., Dehon, C., 2010. Influence functions of the Spearman and Kendall correlation measures. Statistical Methods and Applications 19, 497–515.

Croux, C., Haesbroeck, G., 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. Journal of Multivariate Analysis 71, 161–190.

Croux, C., Haesbroeck, G., 2000. Principal components analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Biometrika 87, 603–618.

Cuesta-Albertos, J.A., Matrán, C., Mayo-Iscar, A., 2008. Trimming and likelihood: robust location and dispersion estimation in the elliptical model. The Annals of Statistics 36, 2284–2318.

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.

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

Flury, B., Riedwyl, H., 1988. Multivariate Statistics: A Practical Approach. Chapman and Hall, London.

Gallegos, M.T., Ritter, G., 2005. A robust method for cluster analysis. The Annals of Statistics 33, 347–380.

Garcìa-Escudero, L.A., Gordaliza, A., 2005. Generalized radius processes for elliptically contoured distributions. Journal of the American Statistical Association 100, 1036–1045.

Garcìa-Escudero, L.A., Gordaliza, A., Matrán, C., Mayo-Iscar, A., 2008. A general trimming approach to robust cluster analysis. The Annals of Statistics 36, 1324–1345.

Gnanadesikan, R., 1997. Methods for Statistical Data Analysis of Multivariate Observations, second ed. Wiley, New York.

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

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

Hubert, M., Van Driessen, K., 2004. Fast and robust discriminant analysis. Computational Statistics and Data Analysis 45, 301–320.

Hunter, D.R., Wang, S., Hettmansperger, T.P., 2007. Inference for mixtures of symmetric distributions. The Annals of Statistics 35, 224–251.

Koziol, J.A., 1982. A class of invariant procedures for assessing multivariate normality. Biometrika 69, 423–427.

Lopuhaä, H.P., 1999. Asymptotics of reweighted estimators of multivariate location and scatter. The Annals of Statistics 27, 1638–1665.

Maronna, R.A., Martin, R.D., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. Wiley, Chichester.

Mecklin, C.J., Mundfrom, D.J., 2004. An appraisal and bibliography of tests for multivariate normality. International Statistical Review 72, 123–138.

Moore, D.S., Stubblebine, J.B., 1981. Chi-square tests for multivariate normality with application to common stock prices. Communications in Statistics—Theory and Methods A 10, 713–738.

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

Rousseeuw, P.J., Van Aelst, S., Van Driessen, K., Agullò, J., 2004. Robust multivariate regression. Technometrics 46, 293–305.

Scheffé, H., 1959. The Analysis of Variance. Wiley, New York.

Singh, A., 1993. Omnibus robust procedures for assessment of multivariate normality and detection of multivariate outliers. In: Patil, G.P., Rao, C.R. (Eds.), Multivariate Environmental Statistics. Elsevier, pp. 445–488.

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

Van Aelst, S., Vandervieren, E., Willems, G., 2012. A Stahel–Donoho estimator based on huberized outlyingness. Computational Statistics and Data Analysis 56, 531–542.

Van Aelst, S., Willems, G., 2011. Robust and efficient one-way MANOVA tests. Journal of the American Statistical Association 106, 706–718.

Willems, G., Joe, H., Zamar, R., 2009. Diagnosing multivariate outliers detected by robust estimators. Journal of Computational and Graphical Statistics 18, 73–91.