Biometrika (2015), **102**, 3, *pp.* 589–599
Printed in Great Britain
doi: 10.1093/biomet/asv021
Advance Access publication 7 June 2015

Outlier detection for high-dimensional data

By KWANGIL RO, CHANGLIANG ZOU, ZHAOJUN WANG

Institute of Statistics, Nankai University, Tianjin 300071, China
rokwangil@yahoo.com.cn nk.chlzou@gmail.com zjwang@nankai.edu.cn

and GUOSHENG YIN

Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong
gyin@hku.hk

Summary

Outlier detection is an integral component of statistical modelling and estimation. For high-dimensional data, classical methods based on the Mahalanobis distance are usually not applicable. We propose an outlier detection procedure that replaces the classical minimum covariance determinant estimator with a high-breakdown minimum diagonal product estimator. The cut-off value is obtained from the asymptotic distribution of the distance, which enables us to control the Type I error and deliver robust outlier detection. Simulation studies show that the proposed method behaves well for high-dimensional data.

Some key words: Masking; Minimum covariance determinant estimator; Reweighting; Swamping.

1. Introduction

Outlier detection plays a critical role in data processing, modelling, estimation and inference. Rapid developments in technology have led to the generation of high-dimensional data in a wide range of fields, such as genomics, biomedical imaging, tomography, signal processing and finance. Conventional outlier detection methods do not work well for such data.

For multivariate data, let $\mathcal{Y} = \{Y_1, \ldots, Y_n\} \subset \mathbb{R}^p$ be independent and identically distributed p-dimensional random vectors with mean $\mu = (\mu_1, \ldots, \mu_p)^{\mathrm{T}}$ and a positive-definite covariance matrix Σ whose entries are $(\sigma_{jk})_{j,k=1,\ldots,p}$. Conventional outlier detection methods often rely on a distance measure to characterize how far a particular data point is from the centre of the data. A common measure of outlyingness for an individual $Y_i = (y_{i1}, \ldots, y_{ip})^{\mathrm{T}}$ is the Mahalanobis distance,

$$d_i^2(\mu, \Sigma) = (Y_i - \mu)^{\mathrm{T}} \Sigma^{-1} (Y_i - \mu). \tag{1}$$

It is crucial to obtain reliable estimates of μ and Σ, as well as to determine the threshold for $d_i(\mu, \Sigma)$, in order to decide whether an observation is an outlier (Cerioli et al., 2009).

In robust statistics, estimation of the multivariate location parameter μ and covariance matrix Σ is challenging, as many classical methods break down in the presence of $n/(p+1)$ outliers. One high-breakdown approach is the minimum volume ellipsoid method of Rousseeuw (1985), which searches for the ellipsoid with the smallest volume that covers h data points, with $n/2 < h < n$. However, it seems to be more advantageous to replace the minimum volume ellipsoid by the minimum covariance determinant estimator, which identifies the subset containing h

observations such that the classical covariance matrix has the lowest determinant. Furthermore, Rousseeuw & Van Driessen (1999) developed a so-called fast minimum covariance determinant algorithm, which is computationally more efficient than all existing minimum volume ellipsoid algorithms. To determine the cut-off value for outlying points, Hardin & Rocke (2005) provided a distributional result for the Mahalanobis distance in (1) based on the minimum covariance determinant estimator. Along similar lines, Cerioli (2010) developed a multivariate outlier test, which performs well in terms of both size and power.

However, when the dimension p of the data is greater than the sample size n, the aforementioned methods are infeasible. Even in the $p < n$ case, as p increases, the traditional methods for outlier detection based on the Mahalanobis distance may become degenerate, and the contamination bias, which grows rapidly with p, could make the minimum covariance determinant unreliable for large p (Adrover & Yohai, 2002; Alqallaf et al., 2009; Yu et al., 2012). This drawback has also been revealed by some high-dimensional location tests (Srivastava & Du, 2008; Chen & Qin, 2010). Filzmoser et al. (2008) proposed a procedure using the properties of principal components analysis to identify outliers in a transformed space. Fritsch et al. (2011) modified the minimum covariance determinant approach by adding a regularization term to ensure that the estimation is well-posed in high-dimensional settings. However, there is no distributional result in Fritsch et al. (2011), so in practice it is not easy to find appropriate cut-off values to achieve a desired false alarm rate.

To overcome the difficulties with high-dimensional data, we modify the Mahalanobis distance so that it involves only the diagonal elements of the covariance matrix:

$$d_i^2(\mu, D) = (Y_i - \mu)^{\mathrm{T}} D^{-1} (Y_i - \mu), \tag{2}$$

where $D = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{pp})$. We can rewrite (2) as $\sum_{j=1}^{p} (y_{ij} - \mu_j)^2 / \sigma_{jj}$, so the information on outlyingness can be extracted from each individual marginally. The modified Mahalanobis distance (2) is invariant under a group of scalar transformations. Based on (2), we propose a high-breakdown minimum diagonal product estimator and develop an algorithm and threshold rule for outlier identification.

2. Methods and properties

2·1. *Minimum diagonal product estimator*

Let $Y_1, \ldots, Y_n \sim N_p(\mu, \Sigma)$, and denote the covariance matrix by $\Sigma = (\sigma_{jk})$ $(j, k = 1, \ldots, p)$ and the diagonal matrix by $D = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{pp})$; thus the correlation matrix is $R = D^{-1/2} \Sigma D^{-1/2} \equiv (\rho_{jk})$. When μ and Σ are known, we can make an orthogonal transformation and rewrite the modified Mahalanobis distance (2) in the canonical form $d_i^2(\mu, D) = \sum_{k=1}^{p} \lambda_k \xi_k^2$, where $\{\lambda_k : k = 1, \ldots, p\}$ are the eigenvalues of the correlation matrix R and $\{\xi_k : k = 1, \ldots, p\}$ are independent standard normal variables. Given the true parameters μ and D,

$$\frac{d_i^2(\mu, D) - p}{\{2\,\mathrm{tr}(R^2)\}^{1/2}} \to N(0, 1) \tag{3}$$

as $p \to \infty$, which follows directly from the Hájek–Šidák central limit theorem based on Conditions A1 and A2 in the Appendix. See equation (3.6) in Srivastava & Du (2008) for details.

Outlier detection can be formulated as n hypothesis tests with $H_{0i} : Y_i \sim N_p(\mu, \Sigma)$ $(i = 1, \ldots, n)$. However, the least-squares estimators of μ and Σ may break down in the presence of outliers. Distance-based methods, such as (2), require robust and consistent estimation of μ

and D. If the asymptotic distribution in (3) is used, consistent estimation of $\mathrm{tr}(R^2)$ is needed to determine the cut-off value for outlying distances, and may fail when the data include outlying observations.

The minimum covariance determinant approach aims to find a subset of observations whose sample covariance matrix has the smallest determinant; this method, however, may not be reliable or well-defined for high-dimensional data. Our method searches for a subset of h observations such that the product of the diagonal elements of the sample covariance matrix is minimal, and involves only the p marginal variances. Let $\mathcal{H} = \{H \subset \{1, \ldots, n\} : |H| = h\}$ be the collection of all subsets of size h, where $|H|$ denotes the cardinality of H. For any $H \in \mathcal{H}$, let $\hat{\mu}(H)$ and $\hat{\Sigma}(H)$ denote, respectively, the sample mean and sample covariance of $\{Y_i : i \in H\}$.

DEFINITION 1. *The minimum diagonal product estimator is defined as*

$$\hat{\mu}_{\mathrm{MDP}} = \hat{\mu}(H_{\mathrm{MDP}}), \quad H_{\mathrm{MDP}} = \arg\min_{H \in \mathcal{H}} \det[\mathrm{diag}\{\hat{\Sigma}(H)\}], \tag{4}$$

where $\mathrm{diag}\{\hat{\Sigma}(H)\}$ *is the matrix of diagonal elements of* $\hat{\Sigma}(H)$.

If the minimization in (4) yields multiple solutions, we arbitrarily choose one to compute the minimum diagonal product estimator. In the one-dimensional setting, where $p = 1$, $\det[\mathrm{diag}\{\hat{\Sigma}(H)\}]$ reduces to $(h-1)^{-1}\sum_{i=1}^{h}\{y_{i1} - \hat{\mu}(H)\}^2$, so the minimization seeks the smallest variance that covers h observations, and thus $\hat{\mu}_{\mathrm{MDP}}$ is equivalent to the least-trimmed-squares estimator (Rousseeuw & Leroy, 1987).

The diagonal matrix D can be estimated by

$$\hat{D}_{\mathrm{MDP}} = c \times \mathrm{diag}\{\hat{\Sigma}(H_{\mathrm{MDP}})\}, \tag{5}$$

where c is a scale factor depending on h and n, which ensures the consistency of $\hat{D}_{\mathrm{MDP}}$ for multivariate normal data. Like its counterpart in the minimum covariance determinant (Pison et al., 2002), c can be determined as follows. We first calculate the modified Mahalanobis distances using the raw estimators of μ and D, i.e., $d_i(\hat{\mu}_{\mathrm{RAW}}, \hat{D}_{\mathrm{RAW}})$ where $\hat{\mu}_{\mathrm{RAW}} = \hat{\mu}_{\mathrm{MDP}}$ and $\hat{D}_{\mathrm{RAW}} = \mathrm{diag}\{\hat{\Sigma}(H_{\mathrm{MDP}})\}$. From Proposition 1, $\mathrm{median}_{1 \leqslant i \leqslant n}\, d_i^2(\hat{\mu}_{\mathrm{MDP}}, \hat{D}_{\mathrm{MDP}}) = p + o_{\mathrm{p}}(1)$ as $p \to \infty$, so we take

$$c = \frac{\mathrm{median}_{1 \leqslant i \leqslant n}\, d_i^2(\hat{\mu}_{\mathrm{RAW}}, \hat{D}_{\mathrm{RAW}})}{\mathrm{median}_{1 \leqslant i \leqslant n}\, d_i^2(\hat{\mu}_{\mathrm{MDP}}, \hat{D}_{\mathrm{MDP}})} \approx \frac{\mathrm{median}_{1 \leqslant i \leqslant n}\, d_i^2(\hat{\mu}_{\mathrm{RAW}}, \hat{D}_{\mathrm{RAW}})}{p}.$$

The minimum diagonal product estimator has two important properties. First, the location estimator $\hat{\mu}_{\mathrm{MDP}}$ and the diagonal matrix estimator $\hat{D}_{\mathrm{MDP}}$ are scalar equivariant but not affine equivariant. Second, to study the global robustness of the minimum diagonal product estimator, we compute its finite-sample breakdown point (Donoho & Huber, 1983). The finite-sample breakdown point ε_n of an estimator T is the smallest fraction of observations from $\mathcal{Y}$ that need to be replaced by arbitrary values to carry the estimate beyond all bounds. Formally, it is defined as $\varepsilon_n(T, \mathcal{Y}) = \min_{1 \leqslant k \leqslant n}\{k/n : \sup_{\mathcal{Y}'} \|T(\mathcal{Y}) - T(\mathcal{Y}')\| = \infty\}$, where the supremum is taken over all possible collections of $\mathcal{Y}'$ obtained from $\mathcal{Y}$ by replacing k data points with arbitrary values. Let $m(\mathcal{Y})$ denote the cardinality of the largest subset of $\mathcal{Y}$ such that all the elements are the same with respect to at least one component. It is usually required that $m(\mathcal{Y}) < h$ (Agulló et al., 2008).

THEOREM 1. *For any data* $\mathcal{Y} \subset \mathbb{R}^p$ *with* $m(\mathcal{Y}) < h$ *and* $p > 1$,

$$\varepsilon_n(\hat{\mu}_{\mathrm{MDP}}, \mathcal{Y}) = \min\{n - h + 1,\ h - m(\mathcal{Y})\}/n. \tag{6}$$

For the $p = 1$ case, (6) reduces to the breakdown point of the least-trimmed-squares estimator (Hössjer, 1994). If $\mathcal{Y}$ is continuous, then for any component of $\mathcal{Y}$ there would be no pair of values equal to each other with probability 1. This implies that $m(\mathcal{Y}) = 1$, and hence $\varepsilon_n(\hat{\mu}_{\mathrm{MDP}}, \mathcal{Y}) = \min(n - h + 1, h - 1)/n$, which does not depend on p. It follows that taking $h = [n/2] + 1$ yields the maximal breakdown point for data with $m(\mathcal{Y}) = 1$, where $[a]$ denotes the integer part of a.

2·2. *Algorithm*

We adapt the fast minimum covariance determinant algorithm of Rousseeuw & Van Driessen (1999) to obtain the minimum diagonal product estimator. The construction in the following theorem guarantees that the objective function is decreasing.

Theorem 2. *Let H_1 be a subset of $\{1, \ldots, n\}$ with $|H_1| = h$, and let $T_1 = h^{-1} \sum_{i \in H_1} Y_i$, $S_1 = h^{-1} \sum_{i \in H_1} (Y_i - T_1)(Y_i - T_1)^{\mathrm{T}}$ and $D_1 = \mathrm{diag}(S_1)$. If $\det(D_1) \neq 0$, define the distance based on T_1 and D_1, $d_i(T_1, D_1)$, for $i = 1, \ldots, n$. If we take H_2 such that $\{d_i(T_1, D_1) : i \in H_2\} = \{d_{(1)}(T_1, D_1), \ldots, d_{(h)}(T_1, D_1)\}$, where $d_{(1)}(T_1, D_1) \leqslant \cdots \leqslant d_{(h)}(T_1, D_1)$ are the ordered distances, and compute T_2 and D_2 based on H_2, then $\det(D_2) \leqslant \det(D_1)$, with equality holding if and only if $T_1 = T_2$ and $D_1 = D_2$.*

The procedures in the fast minimum covariance determinant algorithm can be applied here, upon replacing the Mahalanobis distance with the modified version (2). The algorithm starts with a random subset containing $p + 1$ data points, but such initial subsets may not be available in high-dimensional settings. In fact, the initial subsets are used to estimate the variance of each univariate variable, so we shall simply take their size to be 2. Our algorithm is described as follows.

Algorithm 1. Minimum diagonal product.

Step 1. Construct a number of, say, $m = 100$, initial subsets $H^{(0)}$ with $|H^{(0)}| = 2$.

Step 2. Apply the construction in Theorem 2 to each initial subset until convergence, and obtain m diagonal product values.

Step 3. Select the subset with the minimum diagonal product value.

This algorithm is not permutation invariant. Hubert et al. (2012) presented a deterministic algorithm without using random subsets, which is faster. Their method computes a small number of deterministic initial estimators, followed by the second step in Algorithm 1. This idea could also be adapted to the present problem and warrants further investigation.

2·3. *Minimum diagonal product distance and threshold*

After calculating $d_i(\hat{\mu}_{\mathrm{MDP}}, \hat{D}_{\mathrm{MDP}})$, we develop a threshold rule to determine whether an individual data point is an outlier.

Proposition 1. *Assume that Conditions* A1, A3 *and* A4 *in the Appendix hold, and that under the null hypothesis there is no outlier among the data. Then*

$$\max_{1 \leqslant i \leqslant n} \left| \frac{d_i^2(\hat{\mu}, \hat{D})}{\{2\,\mathrm{tr}(R^2)\}^{1/2}} - \frac{d_i^2(\mu, D)}{\{2\,\mathrm{tr}(R^2)\}^{1/2}} \right| = o_{\mathrm{p}}(1), \quad n, p \to \infty, \tag{7}$$

where $\hat{\mu}$ is the sample mean vector and $\hat{D}$ is the diagonal matrix of the sample covariance.

Although (7) presupposes that the parameters μ and D are estimated by a sample without outliers, it is also expected to be roughly valid for the distance $d_i(\hat{\mu}_{\text{MDP}}, \hat{D}_{\text{MDP}})$, where $\hat{\mu}_{\text{MDP}}$ and $\hat{D}_{\text{MDP}}$ are reliable approximations to those obtained from a clean sample. This proposition, in conjunction with (3), suggests that we could use normal distributions to construct a threshold rule.

In (3), $\text{tr}(R^2)$ needs to be estimated in order to obtain the cut-off value. Let $\text{tr}(R^2)_n = \text{tr}(R_n^2) - p^2/n$, where R_n is the sample correlation matrix. When there is no outlier, under Conditions A1 and A3 we have that $p^{-1}\{\text{tr}(R^2)_n - \text{tr}(R^2)\} \to 0$ in probability as $n, p \to \infty$ (Bai & Saranadasa, 1996). This motivates us to use the estimator

$$\text{tr}(R^2)_{\text{MDP}} = \text{tr}(\hat{R}^2_{\text{RAW}}) - p^2/h, \tag{8}$$

where $\hat{R}_{\text{RAW}}$ is the correlation matrix associated with $\hat{\Sigma}(H_{\text{MDP}})$.

At a significance level of α, using the asymptotic distribution in (3) with robust estimators instead, the ith observation is identified as an outlier if

$$d_i^2(\hat{\mu}_{\text{MDP}}, \hat{D}_{\text{MDP}}) > p + z_\alpha \{2\hat{c}_{p,n}\,\text{tr}(R^2)_{\text{MDP}}\}^{1/2}, \tag{9}$$

where z_α is the upper α-quantile of the standard normal distribution and $\hat{c}_{p,n}$ is an adjustment coefficient that converges to 1 under Condition A3. Srivastava & Du (2008) showed by simulation that using the adjustment quantity $\hat{c}_{p,n} = 1 + \text{tr}(\hat{R}^2_{\text{RAW}})/p^{3/2}$ gives faster convergence to normality.

2·4. *Refined algorithm*

To enhance efficiency, a one-step reweighting scheme is often used in practice (Cerioli, 2010). We refine the identification rule after obtaining a relatively reliable non-outlier subset based on the initial minimum diagonal product detection method. To estimate the parameters using the reweighted observations, the first and second moments of the reweighted variables are needed. Assuming the parameters μ and D to be known, we define the weights by $w_i = 0$ if $d_i^2(\mu, D) > a_\delta$ and $w_i = 1$ otherwise, where a_δ is the upper δ-quantile of the distribution of $d_i^2(\mu, D)$. By (3), we set $a_\delta = p + z_\delta\{2\,\text{tr}(R^2)\}^{1/2}$.

PROPOSITION 2. *Assume that Conditions* A1 *and* A2 *in the Appendix hold, and that under the null hypothesis there is no outlier among the data. Then* $E(y_{ik} \mid w_i = 1) = \mu_k$ *and*

$$\text{var}(y_{ik} \mid w_i = 1) = \sigma_{kk}\left[1 - \frac{2\phi(z_\delta)(R^2)_{kk}}{(1-\delta)\{2\,\text{tr}(R^2)\}^{1/2}} + o(1)\right] \equiv \sigma_{kk}\tau_{kk} \quad (k = 1, \ldots, p),$$

where $(R^2)_{kk}$ is the kth diagonal element of R^2 and ϕ is the standard normal density function.

This proposition elaborates on how to obtain approximately unbiased estimators of μ and D from the observations Y_i for which $w_i = 1$. Let $\tilde{\mu}$ be the sample mean and $\tilde{D}_0$ the diagonal matrix of the sample covariance $\tilde{\Sigma}$ based on those observations. Let $\tilde{D} = \tau^{-1/2}\tilde{D}_0\tau^{-1/2}$ be the refined estimators, where $\tau = \text{diag}(\tau_{11}, \ldots, \tau_{pp})$; accordingly, a refined distance can be constructed as $d_i(\tilde{\mu}, \tilde{D})$. However, it is not easy to obtain a consistent estimator of $(R^2)_{kk}$ in high-dimensional

settings. As it can be verified that

$$\frac{\text{median}_{1\leqslant i\leqslant n}\, d_i^2(\mu, \tau^{1/2} D \tau^{1/2})}{\text{median}_{1\leqslant i\leqslant n}\, d_i^2(\mu, D)} = \left[1 + \frac{\phi(z_\delta)\{2\,\text{tr}(R^2)\}^{1/2}}{p(1-\delta)}\right]\{1 + o(1)\}$$

as $p \to \infty$, we have

$$d_i^2(\tilde{\mu}, \tilde{D}) \approx \frac{d_i^2(\tilde{\mu}, \tilde{D}_0)}{1 + \phi(z_\delta)\{2\,\text{tr}(R^2)\}^{1/2}/\{p(1-\delta)\}}. \tag{10}$$

In other words, we can replace the p scaling factors τ_{kk} with $1 + \phi(z_\delta)\{2\,\text{tr}(R^2)\}^{1/2}/\{p(1-\delta)\}$, which can be estimated more easily. Furthermore, $\text{tr}(R^2)$ can be updated as $\text{tr}(R^2)_w = \text{tr}(\tilde{R}^2) - p^2/n_w$, where $\tilde{R}$ is the correlation matrix associated with $\tilde{\Sigma}$ and $n_w = \sum_{i=1}^n w_i$.

To derive a reliable finite-sample detection rule based on the minimum diagonal product distances, we replace w_i with $\tilde{w}_i$, which is defined by $\tilde{w}_i = 0$ if (9) holds and $\tilde{w}_i = 1$ otherwise. Finally, the refined procedure for outlier detection is summarized as follows.

Algorithm 2. Refined minimum diagonal product.

Step 1. Set a significance level α, and compute the estimators (4) and (5) with $h = [n/2] + 1$.

Step 2. Calculate the distance $d_i(\hat{\mu}_{\text{MDP}}, \hat{D}_{\text{MDP}})$, and assign a weight to each observation according to (9) based on an appropriately chosen δ, such as $\delta = \alpha/2$.

Step 3. Obtain $\tilde{\mu}$ and $\tilde{D}_0$.

Step 4. Compute the refined distance using (10), and test each observation at the significance level α with the rejection region $d_i^2(\tilde{\mu}, \tilde{D}) > p + z_\alpha\{2\tilde{c}_{p,n}\text{tr}(R^2)_w\}^{1/2}$, where $\tilde{c}_{p,n} = 1 + \text{tr}(\tilde{R}^2)/p^{3/2}$.

The refined procedure is fast; for instance, when $n = 100$ and $p = 400$, it only takes about two seconds to finish the computation in FORTRAN using an Intel i7-2630 CPU. R (R Development Core Team, 2015) and FORTRAN code for implementing the procedure are available in the Supplementary Material.

3. Simulations

In the simulation study, we fix the sample size to be $n = 100$. Each dataset is composed of $n - n^*$ observations from $N_p(0, R)$ and n^* observations from a p-variate location-shift model $Y_i \sim N_p(kb_i, R)$, where k is a constant and the b_i are p-dimensional independent random vectors with unit L_2-norm. As all the methods considered are scalar invariant, the covariance matrix $\Sigma = R$ is used. We consider autoregressive correlation with $\rho_{jk} = 0{\cdot}5^{|j-k|}$ and moving average structures. The moving average model is constructed by $y_{ij} = \sum_{k=1}^L \eta_k z_{i(j+k-1)}/(\sum_{k=1}^L \eta_k^2)^{1/2}$ $(i = 1, \ldots, n;\ j = 1, \ldots, p)$, where the η_k and $\{z_{ik}\}$ are independent Un(0, 1) and $N(0, 1)$ variables, respectively. The lag, L, determines the sparseness of R. We allow L to grow by setting $L = [p^{1/2}]$. This growth rate for L would result in a sparse matrix R so that Condition A1, on which the validity of the asymptotic normality (3) relies, is satisfied. If we use a rate of $L = O(p)$, the corresponding correlation matrix would not be sparse and Condition A1 would not hold, so our approach would not perform well, especially in terms of Type I errors. We explore two cases for the outliers: (i) b_i is a normalized p-vector consisting of p independent random variables

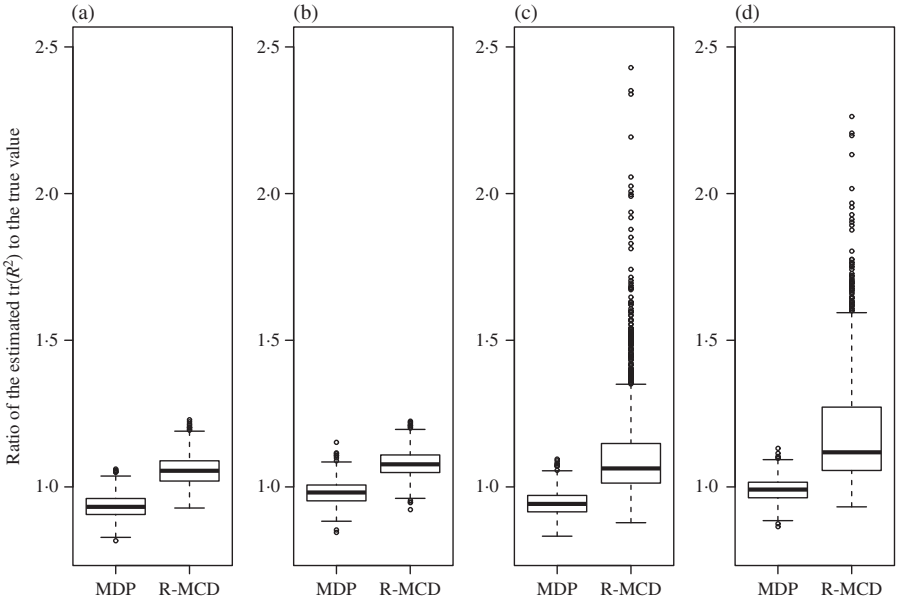

Fig. 1. Comparison of boxplots of $\mathrm{tr}(R^2)_{\mathrm{MDP}}/\mathrm{tr}(R^2)$ using the minimum diagonal product estimator and $\mathrm{tr}(R^2)_{\text{R-MCD}}/\mathrm{tr}(R^2)$ using the regularized minimum covariance determinant estimator for different pairs of p and n^* values: (a) $(p, n^*) = (100, 20)$; (b) $(p, n^*) = (200, 20)$; (c) $(p, n^*) = (100, 40)$; (d) $(p, n^*) = (200, 40)$.

from Un(0, 1); and (ii) b_i is a normalized p-vector in which only $p/5$ random components are from Un(0, 1) and the other components are all zero. All the simulation results are based on 1000 replications.

We first show that the estimator $\mathrm{tr}(R^2)_{\mathrm{MDP}}$ in (8) performs well with finite samples. The contamination rate n^*/n is set to be 0·2 or 0·4, and the dimension is taken to be $p = 100$ or 200. We compare $\mathrm{tr}(R^2)_{\mathrm{MDP}}$ and $\mathrm{tr}(R^2)_{\text{R-MCD}}$, where $\mathrm{tr}(R^2)_{\text{R-MCD}}$ is calculated based on the regularized minimum covariance determinant procedure of Fritsch et al. (2011). Figure 1 presents boxplots of $\mathrm{tr}(R^2)_{\mathrm{MDP}}/\mathrm{tr}(R^2)$ and $\mathrm{tr}(R^2)_{\text{R-MCD}}/\mathrm{tr}(R^2)$ in case (i) with the autoregressive structure and $k = 20$. Clearly, $\mathrm{tr}(R^2)_{\mathrm{MDP}}$ is accurate and generally outperforms $\mathrm{tr}(R^2)_{\text{R-MCD}}$, regardless of how large the proportion of outliers is. The advantage of $\mathrm{tr}(R^2)_{\mathrm{MDP}}$ becomes more pronounced for larger p and n^*.

Outlier identification performance is evaluated by the Type I error rate, i.e., the proportion of good observations that are incorrectly classified as outliers, and the Type II error rate, which is the proportion of contaminated observations that are incorrectly labelled as good ones. These error rates reflect the swamping probability and the masking probability, respectively. Under the same settings as before, the nominal significance level α is chosen to be 0·01, 0·05 or 0·1, and $k = 10$ and $k = p^{1/2}$ are considered with the autoregressive and moving average models, respectively. Table 1 reports the Type I error rate of the refined minimum diagonal product method in case (i) for various combinations of n^* and p. The empirical Type I error rates are close to the nominal levels in most settings.

Next, we compare the proposed outlier detection procedure with existing methods, including those of Filzmoser et al. (2008) and Fritsch et al. (2011). We also consider another alternative, the Stahel–Donoho method, which involves first constructing the initial subset based on the Stahel–Donoho outlyingness (Maronna & Yohai, 1995; Van Aelst et al., 2012) and then applying the procedure of Fritsch et al. (2011). In both the Fritsch et al. (2011) and Stahel–Donoho methods,

Table 1. *Average Type I errors* (%) *in case* (i) *for various values of* p, n^* *and* α *when* $n = 100$

		$n^* = 10$			$n^* = 20$		
Correlation	p	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
AR	50	2·4	6·9	11·7	1·7	5·3	9·2
	100	2·0	6·5	11·1	1·5	5·0	8·9
	200	1·6	6·2	10·9	1·2	4·7	8·6
	400	1·3	5·7	10·5	0·9	4·2	8·2
MA	50	2·0	6·5	11·0	1·5	4·9	8·4
	100	1·8	6·2	10·4	1·2	4·4	8·1
	200	1·6	5·5	9·7	1·2	4·3	7·7
	400	1·3	5·0	9·0	1·0	3·7	7·0

AR, autoregressive; MA, moving average.

Table 2. *Average Type I* (α) *and Type II* (β) *errors* (%) *in cases* (i) *and* (ii) *for various values of* p *with a nominal size of* $\alpha = 0{\cdot}05$, *when* $n = 100$ *and* $n^* = 10$

			R-MDP		PCOut		R-MCD		SDM	
Case	Correlation	p	α	β	α	β	α	β	α	β
(i)	AR	50	6·9	0·2	6·0	0·2	4·8	5·2	11·0	0·0
		100	6·5	1·7	5·4	1·5	12·6	3·0	7·8	0·4
		200	6·2	7·9	5·5	3·6	13·1	9·2	13·5	1·1
		400	5·7	23·4	5·5	11·6	22·7	14·0	7·1	10·7
	MA	50	6·5	31·1	8·0	25·7	11·7	19·9	32·6	3·6
		100	6·2	20·2	6·3	23·5	34·6	0·4	14·9	1·7
		200	5·5	10·1	5·5	19·2	30·8	0·0	25·7	0·3
		400	5·0	4·7	5·4	7·9	41·7	0·0	12·7	0·1
(ii)	AR	50	6·7	0·0	6·7	0·3	5·4	0·4	6·9	1·7
		100	6·4	0·4	6·4	47·2	13·4	0·0	12·2	4·0
		200	6·3	4·4	7·1	78·3	10·7	1·5	10·4	6·7
		400	5·9	19·6	7·5	86·7	17·7	5·4	8·3	12·8
	MA	50	6·7	22·4	7·4	2·1	7·8	3·4	16·0	0·3
		100	6·3	10·5	6·1	5·0	21·4	0·0	30·3	0·0
		200	5·8	2·1	5·7	27·2	26·2	0·0	15·0	0·1
		400	5·1	0·3	5·8	43·4	42·1	0·0	41·2	0·0

R-MDP, our refined minimum diagonal product method; PCOut, the principal component outlier detection procedure of Filzmoser et al. (2008); R-MCD, the regularized minimum covariance determinant method of Fritsch et al. (2011); SDM, the method of first constructing the initial subset based on the Stahel–Donoho outlyingness and then applying R-MCD.

the size of the elemental subset for estimation is chosen to be the same as that in our method, i.e., $h = [n/2] + 1$. There seems to be no direct method for determining the cut-off value in the procedure of Fritsch et al. (2011), because it is not clear what the distribution of the regularized Mahalanobis distance is in high-dimensional settings. Hence, we use simulations to find the cut-off value such that a desired Type I error is achieved by assuming Y to be from the univariate standard normal distribution. Although the iterated reweighted minimum covariance determinant of Cerioli (2010) has been shown to possess good finite-sample properties, it is not considered here as a benchmark because the method is not designed for high-dimensional settings. Our simulation studies, not reported here, show that when the dimension is relatively small, say $p \leqslant 20$, in most cases Cerioli's method outperforms the others in terms of both Type I and Type II errors.

Table 3. *Average Type I* (α) *and Type II* (β) *errors* (%) *in case* (iii) *with* $\psi = 2$, *when* $n = 100$ *and* $n^* = 10$

	R-MDP		PCOut		R-MCD		SDM	
p	α	β	α	β	α	β	α	β
50	6·9	8·9	7·7	17·2	4·6	7·3	7·6	3·1
100	6·4	0·4	6·6	48·7	16·0	0·0	9·0	0·2
200	6·1	0·0	5·5	38·9	20·5	0·0	13·9	0·0
400	5·7	0·0	4·8	9·4	18·0	0·0	8·4	0·0

Table 4. *Average Type I* (α) *and Type II* (β) *errors* (%) *in case* (iv)

	Autoregressive				Moving average			
	R-MDP		PCOut		R-MDP		PCOut	
p	α	β	α	β	α	β	α	β
50	6·7	0·3	5·2	0·0	6·7	0·0	4·8	0·0
100	6·6	1·8	4·8	0·1	6·5	0·3	4·2	0·0
200	6·2	8·4	5·0	0·2	6·2	5·1	4·5	0·1
400	5·7	23·7	5·2	1·0	5·6	22·0	4·9	0·6

The Type I error rates of our method are higher than the nominal size, because both the asymptotic distribution in (3) and the consistency of the estimator of $\mathrm{tr}(R^2)$ rely on the condition that p is sufficiently large.

Simulation results with nominal size $\alpha = 0{\cdot}05$ and $n^* = 10$ are summarized in Table 2. Again, $k = 10$ and $k = p^{1/2}$ are considered under the autoregressive and moving average models, respectively. In most cases, the proposed method can maintain the desired Type I error rate and also yield small Type II error rates. In contrast, the Fritsch et al. (2011) and Stahel–Donoho methods do not work well, with their Type I error rates deviating greatly from the nominal level. The method of Filzmoser et al. (2008) also approximately achieves a Type I error rate of 0·05 and has comparable performance to our method in case (i). However, our method performs better than that of Filzmoser et al. (2008) in case (ii): the Type II error rate of the latter increases rapidly as the dimension p increases.

It is instructive to consider a radial contamination scheme (Cerioli, 2010), which we refer to as case (iii): the data consist of $n - n^*$ observations from $N(0, R)$ and n^* from $N(0, \Sigma)$, where all the diagonal components of Σ are ψ and the off-diagonal components are the same as those of R, with R chosen to have an autoregressive structure. The simulation results with $\psi = 2$ are summarized in Table 3, which shows that the proposed method generally outperforms the method of Filzmoser et al. (2008) in terms of Type II errors as p increases. Both the Fritsch et al. (2011) and Stahel–Donoho procedures are able to identify the outliers, but their Type I errors are unsatisfactory in most cases.

The advantage of our procedure over that of Filzmoser et al. (2008) is partly due to the fact that the shift directions of the outlying observations Y_i are not the same. In such cases, dimension reduction by principal components analysis seems not to be very useful. In contrast, if $b_i = b$ for all the outliers, the information on outlyingness can be well captured by the first several components, and so the method of Filzmoser et al. (2008), based on principal components analysis, would be more powerful. To gain more insight into this case, Table 4 shows the comparison of results in such a scenario, which we refer to as case (iv). In this scenario, all the settings are the same as those in case (i) of Table 2, except that the outlying observations are generated by $Y_i \sim N(kb, R)$, where b is a normalized p-vector consisting of p independent random variables

from Un(0, 1). The advantage of the method of Filzmoser et al. (2008) is obvious in this case, which suggests that projection-based methods would be a better choice if additional information indicates that the data can be regarded as variables from a mixture distribution with only a few mixture components.

Some additional simulation results in the Supplementary Material lead to similar conclusions.

Acknowledgement

The authors would like to thank the referees, associate editor and editor for comments that have resulted in significant improvements to the article. This research was supported in part by the Foundation for the Authors of National Excellent Doctoral Dissertations, the National Natural Science Foundation of China, and the Research Grants Council of Hong Kong. Ro's work was partially completed at Kim Chaek University of Technology.

Supplementary material

Supplementary material available at *Biometrika* online contains proofs of the theoretical results and additional simulation results.

Appendix

The proposed outlier detection method requires the following four conditions.

Condition A1. For $i = 1, 2, 3, 4$, $0 < \lim_{p\to\infty} \mathrm{tr}(R^i)/p < \infty$.

Condition A2. The eigenvalues λ_i of the correlation matrix R satisfy $\lim_{p\to\infty} \max_{1\leqslant i\leqslant p} \lambda_i/p^{1/2} = 0$.

Condition A3. The dimension p grows with sample size n at a rate of $p = O(n^{1/\zeta})$ with $1/2 < \zeta \leqslant 1$.

Condition A4. For some $0 < \gamma < \zeta/2$, $\lim_{p\to\infty} \max_{1\leqslant i\leqslant p} \lambda_i/p^{\gamma} < \infty$.

Conditions A1 and A2 are imposed to guarantee the asymptotic convergence of the proposed distance (2). Since we apply the central limit theorem to the sum of p correlated variables, some conditions on R are inevitable. Condition A2 is used to satisfy the Hájek–Šidák condition. If all the eigenvalues of R are bounded, Condition A1 is trivially true for any p. If the correlation matrix contains many large entries, Condition A1 may not hold and neither does the asymptotic normality of (2). Thus, asymptotic normality relies on how strong the dependencies among the variables are; a certain degree of sparseness of R is needed. The stronger Conditions A2 and A4 are required to obtain Proposition 1, which is a uniform convergence result. Condition A3 includes the case where $n \leqslant p$ and $n/p \to r$ with $0 \leqslant r \leqslant 1$ and the case where $n > p$ but $n/p \to r$ with $1 \leqslant r < \infty$.

References

Adrover, J. & Yohai, V. J. (2002). Projection estimates of multivariate location. *Ann. Statist.* **30**, 1760–81.

Agulló, J., Croux, C. & Van Aelst, S. (2008). The multivariate least-trimmed squares estimator. *J. Mult. Anal.* **99**, 311–38.

Alqallaf, F., Van Aelst, S., Yohai, V. J. & Zamar, R. H. (2009). Propagation of outliers in multivariate data. *Ann. Statist.* **37**, 311–31.

Bai, Z. & Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6**, 311–29.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *J. Am. Statist. Assoc.* **105**, 147–56.

Cerioli, A., Riani, M. & Atkinson, A. C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statist. Comp.* **19**, 341–53.
Chen, S. X. & Qin, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38**, 808–35.
Donoho, D. L. & Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich Lehmann*, Eds. P. J. Bickel, K. A. Doksum and J. L. Hodges, pp. 157–84. Belmont: Wadsworth.
Filzmoser, P., Maronna, R. & Werner, M. (2008). Outlier identification in high dimensions. *Comp. Statist. Data Anal.* **52**, 1694–711.
Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J. B. & Thirion, B. (2011). Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant. In *Medical Image Computing and Computer Assisted Intervention, Part III*, Eds. G. Fichtinger, A. Martel and T. Peters, pp. 264–71. Heidelberg: Springer.
Hardin, J. & Rocke, D. M. (2005). The distribution of robust distances. *J. Comp. Graph. Statist.* **14**, 910–27.
Hössjer, O. (1994). Rank-based estimates in the linear model with high breakdown point. *J. Am. Statist. Assoc.* **89**, 149–58.
Hubert, M., Rousseeuw, P. J. & Verdonck, T. (2012). A deterministic algorithm for robust location and scatter. *J. Comp. Graph. Statist.* **21**, 618–37.
Maronna, R. A.& Yohai, V. J. (1995). The behavior of the Stahel–Donoho robust multivariate estimator, *J. Am. Statist. Assoc.* **90**, 329–41.
Pison, G. Van Aelst, S. & Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika* **55**, 111–23.
R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, vol. B, Ed. W. Grossmann, G. Pflug, I. Vincze and W. Werz, pp. 283–97. Dordrecht: Reidel.
Rousseeuw, P. J. & Leroy, A. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.
Rousseeuw, P. J. & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 212–23.
Srivastava, M. S. & Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Mult. Anal.* **99**, 386–402.
Van Aelst, S., Vandervieren, E. & Willems, G. (2012). A Stahel–Donoho estimator based on huberized outlyingness. *Comp. Statist. Data Anal.* **56**, 531–42.
Yu, G., Zou, C. & Wang, Z. (2012). Outlier detection in the functional observations with applications to profile monitoring. *Technometrics* **54**, 308–18.

[*Received August* 2013. *Revised March* 2015]