Biometrika (2015), **xx**, x, *pp*. 1–10

Printed in Great Britain

Outlier detection for high dimensional data

BY KWANGIL RO, CHANGLIANG ZOU, ZHAOJUN WANG
Institute of Statistics, Nankai University, Tianjin, China, 300071
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 modeling and estimation. For high dimensional data where the dimension increases with sample size, classical methods based on the Mahalanobis distance are typically inapplicable. We propose an outlier detection procedure that replaces the classical minimum covariance determinant estimator with a high-breakdown minimum diagonal product estimator. The cutoff value is obtained through 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 in high dimensional data.

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

1. INTRODUCTION

Outlier detection plays a critical role in data processing, modeling, estimation, and inference. Rapid development in technology has led to emergence of high dimensional data from fields such as genomics, biomedical imaging, tomography, signal processing and finance. Conventional outlier detection methods do not work well for such data, in which the dimensionality may be very large.

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 $\mu = (\mu_1, \ldots, \mu_p)^\mathsf{T}$ 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 center of the data. The usual measure of outlyingness for an individual $Y_i = (y_{i1}, \ldots, y_{ip})^\mathsf{T}$ is the Mahalanobis distance,

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

It is critical to obtain reliable estimates of μ and Σ, as well as to determine the threshold for $d_i(\mu, \Sigma)$ to classify 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 appears 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 and 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 cutoff value for outlying points, Hardin and 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 the test size and power.

However, when the dimension of the data is higher than the sample size, the aforementioned methods are infeasible. Even for $p < n$, as p increases, traditional measures for outlier detection based on the Mahalanobis distance may become degenerate, and the contamination bias, which grows rapidly with p, would make the minimum covariance determinant unreliable for a large p (Adrover and Yohai, 2002; Alqallaf et al., 2009; Yu et al., 2012). This drawback has also been revealed by some high-dimensional location tests (Srivastava and Du, 2008; Chen and Qin, 2010). Filzmoser et al. (2008) suggested a procedure that uses the properties of principal component 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) and, as a consequence, it is not easy to find appropriate cutoff values in practice to attain a desired false alarm rate.

To overcome the difficulties involved in high-dimensional data, we modify the Mahalanobis distance so that it only contains the diagonal elements of the covariance matrix,

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

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 the 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 $\Sigma \equiv (\sigma_{jk})$ $(j, k = 1, \ldots, p)$, the diagonal matrix by $D = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{pp})$; and 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 in (2) as its 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,

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

which directly follows the Hájek–Šidák central limit theorem based on Conditions 1–2 in the Appendix. See equation (3.6) in Srivastava and Du (2008) for details.

Outlier detection can be cast 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 cutoff value of outlying distances, and also 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, which, however, may not be reliable or well-defined for high-dimensional data. Our approach searches for a subset of h observations such that the product of

the diagonal elements of the sample covariance matrix is minimal, which 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. For any $H \in \mathcal{H}$, let $\hat{\mu}(H)$ and $\hat{\Sigma}(H)$ denote the sample mean and sample covariance of $\{Y_i : i \in H\}$, respectively.

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

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

where $\mathrm{diag}\{\hat{\Sigma}(H)\}$ *is the diagonal matrix 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-dimension setting with $p = 1$, $\det[\mathrm{diag}\{\hat{\Sigma}(H)\}]$ reduces to $(h-1)^{-1}\sum_{i=1}^{h}\{y_{i1} - \hat{\mu}(H)\}^2$, and thus the minimization seeks the smallest variance that covers h observations, and accordingly $\hat{\mu}_{\mathrm{MDP}}$ is equivalent to the least trimmed squares estimator (Rousseeuw and 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, to ensure the consistency of $\hat{D}_{\mathrm{MDP}}$ for multivariate normal data. Similar to 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, $\mathrm{median}_{1\le i\le n}\, d_i^2(\hat{\mu}_{\mathrm{MDP}}, \hat{D}_{\mathrm{MDP}}) = p + o_p(1)$ as $p \to \infty$, and thus we take

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

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 and 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 $\varepsilon_n(T, \mathcal{Y}) = \min_{1\le k\le n}\left\{k/n : \sup_{\mathcal{Y}'} \parallel T(\mathcal{Y}) - T(\mathcal{Y}') \parallel = \infty\right\}$, where the supremum is taken over all possible collections of $\mathcal{Y}'$ obtained from $\mathcal{Y}$ by replacing k data points by arbitrary values. Let $m(\mathcal{Y})$ denote the cardinality of the largest subset of $\mathcal{Y}$ satisfying 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 case with $p = 1$, (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 with probability 1. This implies that $m(\mathcal{Y}) = 1$, and thus $\varepsilon_n(\hat{\mu}_{\mathrm{MDP}}, \mathcal{Y}) = \min(n - h + 1, h - 1)/n$, which does not depend on p. It follows that $h = [n/2] + 1$ yields the maximal breakdown point 50% 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 in Rousseeuw and Van Driessen (1999) to obtain the minimum diagonal product estimator. The construction in the following theorem guarantees the decrease of the objective function.

THEOREM 2. *Let H_1 be a subset of $\{1,\ldots,n\}$ with $|H_1| = h$, and let $T_1 = h^{-1}\sum_{i\in H_1} Y_i$, $S_1 = h^{-1}\sum_{i\in H_1}(Y_i - T_1)(Y_i - T_1)^{\mathsf{T}}$, 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 $d_{(1)}(T_1, D_1) \leq \cdots \leq d_{(h)}(T_1, D_1)$ are the ordered distances, and compute T_2 and D_2 based on H_2, then $\det(D_2) \leq \det(D_1)$, and the equality holds 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, by replacing the Mahalanobis distance with the modified version (2). The algorithm starts from a random subset containing $(p + 1)$ data points, while 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 and hence we 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 till convergence and obtain m diagonal product values.

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

The 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 is an outlier.

PROPOSITION 1. *Assume that Conditions 1, 3 and 4 hold. Under the null hypothesis that there is no outlier in the data,*

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

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 $d_i(\hat{\mu}_{\mathrm{MDP}}, \hat{D}_{\mathrm{MDP}})$, where $\hat{\mu}_{\mathrm{MDP}}$ and $\hat{D}_{\mathrm{MDP}}$ 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), $\mathrm{tr}(R^2)$ needs to be estimated in order to obtain the cutoff value. Let $\mathrm{tr}(R^2)_n = \mathrm{tr}(R_n^2) - p^2/n$, where R_n is the sample correlation matrix. When there is no outlier, under Conditions 1 and 3, $p^{-1}\{\mathrm{tr}(R^2)_n - \mathrm{tr}(R^2)\} \to 0$ in probability as $n, p \to \infty$ (Bai and Saranadasa, 1996). This motivates us

to use the estimator

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

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

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

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

where z_α is the upper αth quantile of the standard normal distribution and $\hat{c}_{p,n}$ is an adjustment coefficient that converges to one under Condition 3. Srivastava and Du (2008) showed by simulation that using the adjustment quantity $\hat{c}_{p,n} = 1 + \mathrm{tr}(\hat{R}^2_{\mathrm{RAW}})/p^{3/2}$ leads to 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 weight $w_i = 0$ if $d_i^2(\mu, D) > a_\delta$, and $w_i = 1$ otherwise, where a_δ is the upper δth quantile of the distribution of $d_i^2(\mu, D)$. By (3), we set $a_\delta = p + z_\delta\{2\mathrm{tr}(R^2)\}^{1/2}$.

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

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

where $(R^2)_{kk}$ *is the* k*th diagonal element of* R^2 *and* ϕ *is the standard normal probability 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 $\tilde{\mu}$ and $\tilde{D}_0$ be the sample mean and the diagonal matrix of the sample covariance $\tilde{\Sigma}$ based on those observations, respectively. Let $\tilde{D} = \tau^{-1/2}\tilde{D}_0\tau^{-1/2}$ be the refined estimators, where $\tau = \mathrm{diag}(\tau_{11}, \ldots, \tau_{pp})$, and 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

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

we have

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

In other words, we replace the p scaling factors τ_{kk} with $1 + \phi(z_\delta)\{2\mathrm{tr}(R^2)\}^{1/2}/\{p(1-\delta)\}$, which can be estimated more easily. Furthermore, $\mathrm{tr}(R^2)$ can be updated as $\mathrm{tr}(R^2)_w = \mathrm{tr}(\tilde{R}^2) - p^2/n_w$, 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 by $\tilde{w}_i$ which is defined as: $\tilde{w}_i = 0$ 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 $d_i(\hat{\mu}_{\mathrm{MDP}}, \hat{D}_{\mathrm{MDP}})$ and assign a weight to each observation according to (9) based on an appropriately chosen δ, e.g., $\delta = \alpha/2$.

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

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

The refined procedure runs fast; for instance, when $n = 100$ and $p = 400$, it only takes about two seconds to finish it, using an Intel i7-2630 CPU and FORTRAN. The R and FORTRAN codes for implementing the procedure are available in the Supplementary Material.

3. SIMULATION

In the simulation study, we fix the sample size $n = 100$. 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 a unit L_2 norm. As all the considered methods are scalar-invariant, the covariance matrix $\Sigma = R$ is used. We consider autoregressive correlation with $\rho_{jk} = 0.5^{|j-k|}$ and moving average structures. The moving average model is constructed by $y_{ij} = \sum_{k=1}^{L} \eta_k z_{i(j+k-1)}/(\sum_{k=1}^{L} \eta_k^2)^{1/2}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, where η_k and $\{z_{ik}\}$ are independent $U(0, 1)$ 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 rate of L would result in a sparse matrix R, so that Condition 1, on which the validity of asymptotic normality (3) relies, is satisfied. If we use a rate of $L = O(p)$, the corresponding correlation matrix is not sparse, and Condition 1 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 from $U(0, 1)$; and (II): b_i is a normalized p-vector in which only $p/5$ random components are from $U(0, 1)$ and all the others are zero. All the simulation results are based on 1,000 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 as 0.2 or 0.4, whereas the dimension is $p = 100$ 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 (Fritsch et al., 2011). Figure 1 presents box-plots of $\mathrm{tr}(R^2)_{\mathrm{MDP}}/\mathrm{tr}(R^2)$ and $\mathrm{tr}(R^2)_{\text{R-MCD}}/\mathrm{tr}(R^2)$ under 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^*.

The outlier identification performance is evaluated by the type I error rate, the proportion of good observations which are incorrectly classified as outliers; and the type II error rate, the proportion of contaminated observations which are incorrectly labeled as good ones. These error rates reflect the swamping probability and masking probability, respectively. Under the same settings as before, the nominal significance level α is chosen to be 0.01, 0.05 and 0.1, and $k = 10$ and $k = p^{1/2}$ are considered with the autoregressive and moving average models, respectively. Table 1 presents the type I error rate of the refined minimum diagonal product method under Case (I) for various combinations of n^* and p. The empirical type I error rates are close to the nominal levels in most cases.

We next compare the proposed outlier detection procedure with existing methods, including those in Filzmoser et al. (2008) and Fritsch et al. (2011). We also consider another alternative, the Stahel–Donoho method: first constructing the initial subset based on the Stahel–Donoho outlyingness (Maronna and Yohai, 1995; Van Aelst et al., 2012) and then applying the procedure in Fritsch et al. (2011). In both the Fritsch et al. (2011) and Stahel–Donoho methods, the size of the elemental subset for estimation is chosen

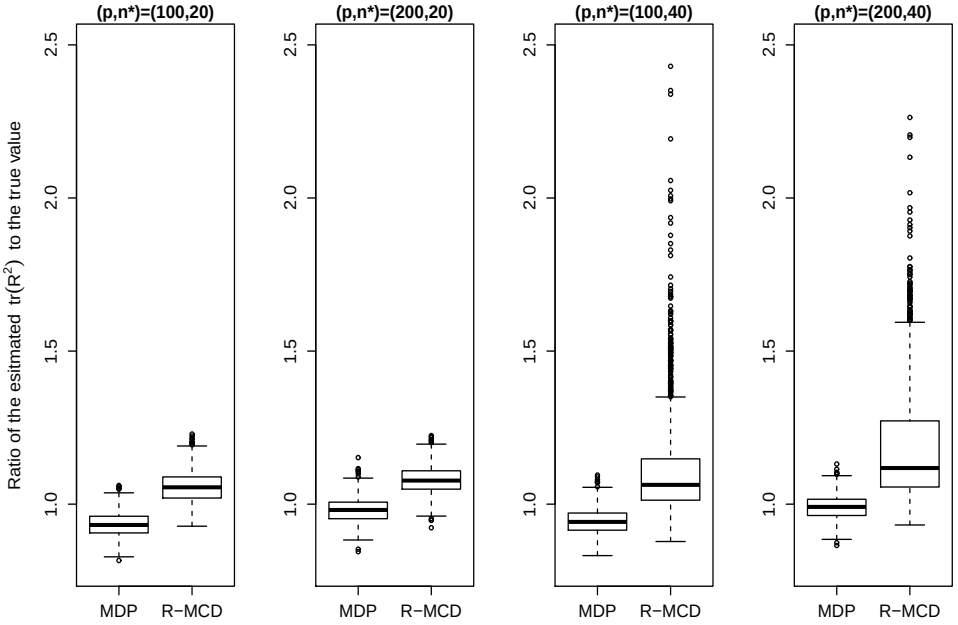

Fig. 1. Comparison of box-plots of $\mathrm{tr}(R^2)_{\mathrm{MDP}}/\mathrm{tr}(R^2)$ using the minimum diagonal product estimator and $\mathrm{tr}(R^2)_{\text{R-MCD}}/\mathrm{tr}(R^2)$ using the regularized minimum covariance determinant estimator for different paired values of (p, n^*).

Table 1. *Average type I errors (%) under Case (I) for various values of* p, n^* *and* α *when* $n = 100$

Correlation	p	$n^* = 10$			$n^* = 20$		
		$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
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 stands for autoregressive and MA for moving average.

to be the same as that in our method, i.e., $h = [n/2] + 1$. There seems to be no direct method to determine the cutoff value in the procedure of Fritsch et al. (2011), because the distribution of the regularized Mahalanobis distance in high-dimensional settings is not clear. Hence, we use the simulation to find the cutoff value so that a desired type I error is achieved by assuming Y from the univariate standard normal distribution. Although the iterated reweighted minimum covariance determinant in 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 cases. Our simulation studies, not reported here, show that when the dimension is relatively small, say $p \leq 20$, Cerioli's (2010) method outperforms the others in most cases in terms of both the type I and type II errors. The type I error rates of our method are higher

Table 2. *Average type I (α) and type II (β) errors (%) under Cases (I)-(II) for various values of p with a nominal size of $\alpha = 0.05$, when $n = 100$ and $n^* = 10$*

Case	Correlation	p	R-MDP		PCOut		R-MCD		SDM	
			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 by Filzmoser et al. (2008); R-MCD: the regularized minimum covariance determinant method by Fritsch et al. (2011); and SDM: first constructing the initial subset based on the Stahel–Donoho outlyingness and then applying R-MCD.

Table 3. *Average type I (α) and type II (β) errors (%) under Case (III) with $\psi = 2$, when $n = 100$ and $n^* = 10$*

p	R-MDP		PCOut		R-MCD		SDM	
	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

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 $\alpha = 0.05$ 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, as their type I error rates deviate greatly from the nominal level. Filzmoser et al.'s (2008) method also approximately attains a type I error rate of 0.05 and has comparable performance with our method under Case (I). However, our method performs better than Filzmoser et al.'s (2008) method under Case (II); the type II error rate of the latter increases fast when the dimension p increases.

Table 4. *Average type I (α) and type II (β) errors (%) under Case (IV)*

	Autoregressive				Moving average			
p	R-MDP		PCOut		R-MDP		PCOut	
	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

It is instructive to consider a radial contamination scheme (Cerioli, 2010), denoted as Case (III): the data are composed of $n - n^*$ observations from $N(0, R)$ and the remaining 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, and R is chosen to have an autoregressive structure. The simulation results with $\psi = 2$ are summarized in Table 3, which shows that the proposed method performs generally better than Filzmoser et al.'s (2008) method in terms of the 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 Filzmoser et al.'s (2008) method is partially due to the fact that the shift directions of those outlying observations Y_i are not the same. In such cases, dimension reduction by principal component analysis seems to be not very useful. In contrast, if $b_i = b$ for all the outliers, the information of outlyingness can be well captured by the first several components, and thus Filzmoser et al.'s (2008) method based on principal component analysis would be more powerful. To gain more insight into this, Table 4 shows the comparison results under such a scenario, 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 through $Y_i \sim N(kb, R)$, where b is a normalized p-vector consisting of p independent random variables from $U(0, 1)$. The advantage of Filzmoser et al.'s (2008) method is obvious. This 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.

Acknowledgment

The authors would like to thank the referees, Associate Editor, and Editor for comments that have resulted in significant improvements in the article. The research was supported in part by the National Natural Science Foundation of China and the Research Grants Council of Hong Kong.

Supplementary Material

The Supplementary Material contains the proofs of theoretical results and additional simulations.

APPENDIX

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

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

Condition 3. The dimension p grows with sample size n at a rate of $p = O(n^{1/\zeta})$, $1/2 < \zeta \le 1$.

Condition 4. For some $0 < \gamma < \zeta/2$, $\lim_{p\to\infty} \max_{1\le i\le p} \lambda_i/p^{\gamma} < \infty$.

Conditions 1–2 are imposed to guarantee the asymptotic convergence of the proposed distance (2). Since we apply the central limit theorem for the sum of p correlated variables, some conditions on R are inevitable. Condition

2 is used to satisfy the Hájek–Šidák condition. If all the eigenvalues of R are bounded, Condition 1 is trivially true for any p. If the correlation matrix contains many large entries, Condition 1 may not hold and neither does the asymptotic normality of (2). Thus, asymptotic normality relies on how strong the dependencies among the variables are; certain sparseness on R is needed. Stronger Conditions 2 and 4 are required to obtain Proposition 1, which is a uniform convergence result. Condition 3 includes the cases $n \leq p$, $n/p \to r,\ 0 \leq r \leq 1$ and $n > p$, but $n/p \to r,\ 1 \leq r < \infty$.

REFERENCES

ADROVER, J. & YOHAI, V. J. (2002). Projection estimates of multivariate location. *Ann. Statist.* **30**, 1760–1781.

AGULLÓ, J., CROUX, C. & VAN AELST, S. (2008). The multivariate least-trimmed squares estimator. *J. Multivar. Anal.* **99**, 311–338.

ALFONS, A., CROUX, C. & GELPER, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Ann. Appl. Statist.* **7**, 226–248.

ALQALLAF, F., VAN AELST, S., YOHAI, V. J. & ZAMAR, R. H. (2009). Propagation of outliers in multivariate data. *Ann. Statist.* **37**, 311–331.

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**, 147–156.

CERIOLI, A., RIANI, M. & ATKINSON, A. C. (2009). Controlling the size of multivariate outlier tests with the MCD estimator of scatter. *Statist. Comput.* **19**, 341–353.

CHEN, S. X. & QIN, Y. L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38**, 808–835.

CROUX, C., FILZMOSER, P. & FRITZ, H. (2013). Robust sparse principal component analysis. *Technometrics* **55**, 202–214.

DONOHO, D. L. & HUBER, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich Lehmann*, Ed. P. J. Bickel, K. A. Doksum and J. L. Hodges, pp. 157–184. Belmont: Wadsworth.

FILZMOSER, P., MARONNA, R. & WERNER, M. (2008). Outlier identification in high dimensions. *Comput. Statist. Data Anal.* **52**, 1694–1711.

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*, Ed. G. Fichtinger, A. Martel and T. Peters, pp. 264–271. Springer: Heidelberg.

HARDIN, J. & ROCKE, D. M. (2005). The distribution of robust distances. *J. Comput. Graph. Statist.* **14**, 910–927.

HÖSSJER, O. (1994). Rank-based estimates in the linear model with high breakdown point. *J. Am. Statist. Assoc.* **89**, 149–158.

HUBERT, M., ROUSSEEUW, P. J. & VERDONCK, T. (2012). A deterministic algorithm for robust location and scatter. *J . Comput. Graph. Statist.* **21**, 618–637.

MARONNA, R. A.& YOHAI, V. J. (1995). The behavior of the Stahel–Donoho robust multivariate estimator, *J. Am. Statist. Assoc.* **90**, 329–341.

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. 283–297. Dordrecht: Reidel.

ROUSSEEUW, P. J. & LEROY, A. (1987). *Robust Regression and Outlier Detection*. Wiley: New York.

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

PISON, G. VAN AELST, S. & WILLEMS, G. (2002). Small sample corrections for LTS and MCD. *Metrika* **55**, 111–123.

SRIVASTAVA, M. S. & DU, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivar. Anal.* **99**, 386–402.

VAN AELST, S., VANDERVIEREN, E. & WILLEMS, G. (2012). A Stahel–Donoho estimator based on huberized outlyingness. *Comput. Statist. Data Anal.* **56**, 531–542.

YU, G., ZOU, C. & WANG, Z. (2012). Outlier detection in the functional observations with applications to profile monitoring. *Technometrics* **54**, 308–318.