Decision Analytics Journal 10 (2024) 100424

Contents lists available at ScienceDirect

Decision Analytics Journal

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

A modified reweighted fast consistent and high-breakdown estimator for high-dimensional datasets

Ishaq A. Baba [a,c,*], Habshah Midi [a,b], Leong W. June [a,b], Gafurjan Ibragimov [a,b]

[a] Institute for Mathematical Research, Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia
[b] Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia
[c] Department of Mathematical Sciences, Faculty of Science, Taraba State University Jalingo, Taraba State, Nigeria

ARTICLE INFO

Keywords:
High-dimensional data
Outliers detection
Covariance matrix
Mahalanobis distance
Reweighted fast consistent and high breakdown point

ABSTRACT

Outlier detection and classification algorithms play a critical role in statistical analysis. The reweighted fast consistent and high breakdown point (RFCH) estimator is an outlier-resistant estimator of multivariate location and dispersion. Still, some difficulties hamper the application of the RFCH in high-dimensional settings. One main difficulty is that the RFCH cannot be applied when the dimension exceeds the sample size. We propose a modified reweighted fast consistent and high breakdown point (MRFCH) estimator to make it applicable to high-dimensional settings. The basic idea of our proposed method is to modify the Mahalanobis distance so that it uses only the diagonal elements of the scatter matrix in the computation of the RFCH algorithm. The proposed method preserves the robustness properties of the RFCH estimator. As a result, we achieve a robust and efficient high-dimensional procedure for computing location and scatter matrix estimates and a powerful outlier detection method. One of the main advantages of our proposed procedure over the existing RFCH is that it can be applied to both low and high-dimensional datasets. Based on the real-life datasets and simulation study, our proposed method showed promising results irrespective of sample size, dimensions, amount of contamination, computational time, and distance of the contamination. Thus, the new proposed algorithm can be applied to solve the problem of regression outliers in high-dimensional data (HDD) and serve as a better alternative to the minimum regularized covariance determinant (MRCD) estimator.

1. Introduction

Most real-world application datasets contain outliers, especially in biomedical and chemometrics research, where multiple features are used to monitor a dynamic or complex system. The presence of outliers hinders the use of data in modeling processes, analysis, and control [1–4]. The reliable way to tackle the problem of outliers is to allow each observation to speak for itself by considering all the data points with full measurement. Excluding data points may lead to the wrong result and a misrepresentation of the study. There are a number of statistical procedures that can be used to test for outliers, depending on the scenario.

In general, many scholars have explored the problems and solutions of outlier detection by providing different methods, methodologies, datasets, tools, and challenges, in addition to some open research questions and directions. Among them, several review papers are nonspecific and provide a comprehensive synopsis of outlier detection methods [5–8]. While some focus on low-dimensional methods and applications, others only center on high-dimensional data with specific types of techniques, application areas, and datasets. The outlier detection techniques can be categorized into unsupervised learning [9] and supervised learning algorithms [10]. The unsupervised include distance-based, density-based, and model-based procedures [11]. Closely linked to our proposal are those concerned with distance-based and high-dimensional datasets. The traditional Mahalanobis have been employed [12–16].

It is now evident that classical outlier detection techniques are not robust [17–20]. As an alternative, the minimum volume ellipsoid (MVE) [21,22] and minimum covariance determinant (MCD) [23,24] are put forward. Comparisons between the MCD and reweighted fast, consistent, and high breakdown points [25] show that the MCD is less robust and time-consuming, especially where the data points are large. Additionally, these procedures do not work well when the data is sparse and the dimension is high. [26,27] deliberated extensively on these limitations. [28,29] conducted a review of studies on outlier detection in high-dimensional settings. [30] proposed a procedure based on the transformed distance using the principal component analysis to detect outliers in high-dimensional space. [31] converted the minimum

* Corresponding author at: Department of Mathematical Sciences, Faculty of Science, Taraba State University Jalingo, Taraba State, Nigeria.
E-mail address: ishaqbaba@yahoo.com (I.A. Baba).

https://doi.org/10.1016/j.dajour.2024.100424
Received 10 May 2023; Received in revised form 9 February 2024; Accepted 12 February 2024
Available online 17 February 2024

covariance determinant estimator into the regularized minimum covariance determinant (MRCD). [32] proposed the minimum diagonal product (MDP) estimators to detect outliers in high-dimensional datasets. Producing satisfactory results in terms of Type II error compared to the Stahel–Donoho outlyingness procedure of [33], the principal component outlier detection of [30], and the regularized minimum covariance determinant of [34]. [35] combined the ideas in [31,32] to detect outliers in high-dimensional space. It is worth noting that selecting a suitable threshold is another big challenge facing both the traditional and robust distance procedures [15,36,37]. The threshold based on the quantile is the most frequently used, and this proved to be less robust against outliers, especially when the data is large and the dimension is high [35,38]. A clear shortcoming of this method is that it uses a fixed quantile for outlier detection, which reveals too many, if not all, points as outliers. [39,40] propose alternative cut-off points based on high leverage potential and Mahalanobis distances to tackle the problem. For a detailed explanation, see [41].

To address these shortcomings, a cut-off point introduced in [41] is used to classify observations as inliers or outliers. Owing to the fact that the Mahalanobis distance based on MRCD is either very expensive to compute, inconsistent, or has a low breakdown, it has become paramount to provide fast, consistent, and high breakdown multivariate estimators that can be used to detect and classify outliers in high-dimensional space.

[25] introduced a reweighted fast consistent and high breakdown (RFCH) estimator of location and scatter matrix, which proved to be faster and more resistant to outliers than the robust MCD constructed by [23]. The RFCH estimator can be used to calculate correlation estimates [42,43] and perform outlier diagnostics in multivariate space [44,45]. Its major drawback is that it is difficult to compute when $p \gg n$. In this way, there is no guarantee that the reweighting step can be achieved since it involves calculating the inverse of the covariance matrix, which leads to singularity. The curse of dimensionality poses difficulties across numerous disciplines [46,47], especially in big data analytics and high-dimensional space. Understanding these difficulties and applying appropriate procedures is critical to providing generalized and robust techniques. These drawbacks create a gap in the availability of consistent and high breakdown estimators in high-dimensional data where the number of predictor variables exceeds the number of observations. Additionally, to our knowledge, no work has extended the reweighted fast, consistent, and high breakdown estimator to make it applicable in high-dimensional settings. Inspired by these limitations, we provide a modification of the RFCH based on the Mahalanobis distances, using only the diagonal elements of the covariance matrix to make it applicable to high dimensions [32]. Our specific objectives in this paper are threefold: (i) to modify the reweighted fast, consistent, and high breakdown estimator to be used to estimate location and scatter matrix in high dimensions. (ii) to develop an outlier detection method based on the modified reweighted fast, consistent, and high breakdown estimator in high dimensions; (iii) to show using simulation and real data examples that the modified method performs and scales well for both low and high-dimensional datasets.

In the rest of this paper, we present the basic concept of multivariate Mahalanobis distance and detail explanations of the reweighted fast consistent high breakdown point estimator in Section 2. Description of the modified reweighted fast, consistent, and high breakdown point estimator algorithm and the new methodology for the outlier detection algorithm are given in Sections 3 and 4. We report the simulation study in Section 5. As for Section 6, it introduces several real-life datasets for the application of the proposed technique in high-dimensional space. Summary, conclusion, and directions for future research and applications are highlighted in Section 7.

2. Multivariate distance measure

Multivariate distance measures are widely applied to quantify the similarity and dissimilarity between two or more sets of multivariate data points. These measures are important in various fields [48,49] such as image processing [50], fraud detection [51], and fault identification in industrial process and control [52]. Some common multivariate distance measures in statistics include the Euclidean distance, the Mahalanobis distance, the correlation distance, and the Minkowski distance. Among them, Mahalanobis distance is widely used in multivariate outlier detection and classification. Therefore, this paper will focus mainly on developing Mahalanobis distance-based outlier detection in high-dimensional settings.

2.1. Robust mahalanobis distance measures

In statistics, the Mahalanobis distance is used to calculate the distance between a particular point and the center of a dataset with a mean and covariance matrix. For example, in a multivariate setting, we assume that the observations are in the $n \times p$ data matrix $X = (x_{i1}, \ldots, x_{ip})$, where $x_i = (x_1, \ldots, x_n)$ represents the ith data point, n stands for sample size, and p is the number of variables. The Mahalanobis distance (MD) can be calculated using the formula:

$$\mathbf{MD}_i = \sqrt{\left(x_i - T(X)\right)^T C(X)^{-1} \left(x_i - T(X)\right)} \tag{1}$$

where $T(X)$ is the classical mean vector and $C(X)$ represents the classical variance covariance matrix. Despite its numerous applications and popularity, the Mahalanobis distance is known to be sensitive to the presence of outliers in a dataset [49,53]. This problem is related to the efficiency and robustness of the mean and covariance matrix estimates and the nature of the cut-off point formula. It is interesting to note that outliers can be seen as a cause of bias or considered a potential data point that allows researchers to understand the process under investigation. This implies that employing a robust and systematic procedure to detect and classify outliers is critical. [54] proposed a minimum covariance determinant (MCD) algorithm for outliers to replace the $T(X)$ and $C(X)$ in Eq. (1). The MCD was developed in order to achieve a robust Mahalanobis distance-based outlier detection procedure. The idea behind MCD is to search for h data points out of n observations whose covariance has the minimum determinant. [23] have shown that Mahalanobis distance based on MCD is highly time-consuming and substituted it with the fast MCD estimator. The fast MCD that is obtained calculates a minimum covariance determinant subset with a faster runtime, making it applicable for large datasets. Thus, the robust Mahalanobis distance (RMD) based outlier detection [23] is calculated by:

$$\mathbf{RMD}_i = \sqrt{\left(x_i - \mu_{MCD}\right)^T \Sigma_{MCD}^{-1} \left(x_i - \mu_{MCD}\right)} \tag{2}$$

where μ_{MCD} is a vector of robust mean and Σ_{MCD} is the robust covariance matrix obtained from the minimum covariance determinant estimators. Several other robust Mahalanobis distance-based outlier detection algorithms have been proposed to handle the problem of outliers [41,55,56]. In another way, a fast, consistent, and high breakdown point estimator is proposed by [25] to find the location and scatter matrix as well as to compute the robust Mahalanobis distance-based outlier detection procedure. Its primary objective is to provide robustness against contaminated observations and realize high breakdown point estimates while achieving maximum computational efficiency. The Mahalanobis distance based on the reweighted fast, consistent, and high breakdown estimator is given by:

$$\mathbf{RMD}_{i,RFCH} = \sqrt{\left(x_i - \mu_{RFCH}\right)^T \Sigma_{RFCH}^{-1} \left(x_i - \mu_{RFCH}\right)} \tag{3}$$

where μ_{RFCH} is a vector of robust mean and Σ_{RFCH} is the robust covariance matrix obtained from the reweighted fast consistent and

high breakdown point estimator. The efficiency and robustness of the RFCH algorithm have been discussed [44,57,58].

Recently, [45] notes that the RFCH is fast and consistent, even faster than the MCD algorithm. Unfortunately, all the above-mentioned Mahalanobis distance-based outlier detection algorithms cannot be applied to high-dimensional datasets. To fill this gap, we proposed modifying Eq. (2) to make it suitable for high-dimensional datasets. Our new Mahalanobis distance-based outlier detection method will use the diagonal elements of the covariance matrix instead of the entire matrix (see [32]).

The chi-squared distribution criteria introduced by [59] are well known to be sensitive to outliers. Three key limitations are attached to this cut-off formula: (i) It assumes that the dimension of variable p follows a multivariate normal distribution; (ii) It takes into consideration only the dimension of the variables but does not consider the number of observations; and (iii) It is not robust against outlying observations. [60] introduced a non-parametric cut-off point using robust Mahalanobis distances. [41] applied this criteria with the diagnostic robust generalized potentials instead of the chi square criteria based on roust Mahalanobis distances. In recent years, much emphasis has been placed on applying outlier detection procedures to high-dimensional data [5,28,61,62]. [32] constructed a threshold rule-based normal distribution for outlier classification in high-dimensional space. [31] adopted the [59] cut-off point formula. [35] combined the regularized minimum covariance determinant (MRCD) estimator and minimum diagonal product (MDP) estimator to provide Mahalanobis distance-based outlier detection for high-dimensional datasets. Although a wide range of robust Mahalanobis distance-based outlier detection procedures are available (see, for example, [63,64]), this paper focuses on modifying a reweighted fast, consistent, and high breakdown estimator introduced by [25] in high-dimensional settings. This procedure was selected because of the clear intuition behind its construction. It provides estimates close to the classical mean and covariance matrix and avoids using outlying observations in its calculations. For the threshold criteria, the robust non-parametric cut-off point is applied [41,60].

2.2. RFCH estimators

Many practical outlier resistant estimators generate a sequence of k trial fits called attractors $(T_1, C_1), (T_2, C_2), \ldots, (T_k, C_k)$. Then the attractor (T_A, C_A), which minimizes some criterion, is employed to achieve the final estimator. These procedures utilize the classical estimator of mean and covariance matrix $(T_{-1,j}, C_{-1,j})$ at the beginning and compute the n Mahalanobis distances $D_i(T_{-1,j}, C_{-1,J})$. At the next iteration, the classical estimator of mean and covariance matrix (T_0, C_0) is calculated from $c_n = n/2$ cases corresponding to the smallest Mahalanobis distances. This iteration continues for k steps, resulting in the sequence of estimators $((T_{-1,j}, C_{-1,j}), (T_{0,j}, C_{0,j}), \ldots, (T_{k,j} C_{k,j}))$. Then $(T_{k,j}, C_{k,j}) = (\bar{x}, S)$ is the jth attractor for $j = 1, \ldots, k$. Note that the FCH, DGK, and MB estimators converge at $k = 5$. [25] introduced a practical outlier-resistant $\sqrt{n}$ consistent estimator that is called the Fast Consistent and High Breakdown (FCH) estimator. This procedure utilizes the [65] DGK and the median ball (MB) estimators. The DGK utilizes the classical mean and covariance matrix at the beginning point to get the final attractor $(T_{k,D}, C_{k,D})$, while the MB utilizes the classical mean and covariance calculated from the cases with $D_i(MED(X), I_p) \leq MEDD_i(MED(X), I_p)$ as an initial to get the final attractor $(T_{k,M}, C_{k,M})$, where $MED(X)$ is the coordinatewise median. If the DGK location estimator $T_{k,M}$ has a greater Euclidean distance from $MED(X)$ than half of the data, then FCH utilizes the MB attractor. The FCH utilizes the minimum determinant as the dispersion criterion to select the attractor if $\|T_{k,D}, -MED(X)\| \leq MED(D_i(MED(X), I_p))$.

Let (T_A, C_A) be the attractor used. Then the FCH estimator (T_F, C_F) takes $T_F = T_A$ and

$$C_F = \frac{MED(D_i^2(T_A, C_A))}{\chi^2_{p,0.5}} \tag{4}$$

where $\chi^2_{p,0.5}$ is the 50th percentile of the chi squared distribution with p degree of freedom. Furthermore, the RFCH utilizes two reweighting steps. Let $(\hat{\mu}_1, \tilde{\Sigma}_1)$ be the classical estimator applied to n_1 cases with $D_i^2(T_{FCH}, C_{FCH}) \leq \chi^2_{p,0.5}$ and let $\hat{\Sigma}_1 = \frac{MED(D_i^2(\hat{\mu}_1, \tilde{\Sigma}_1))}{\chi_{p,0.5}} \tilde{\Sigma}_1$. Let $T_{RFCH}, \tilde{\Sigma}_2$ be the classical estimator applied to the cases with $D_i^2(\hat{\mu}_1, \hat{\Sigma}_1) \leq \chi^2_{p,0.5}$ and defined

$$C_{RFCH} = \frac{MED(D_i^2(T_{RFCH}, \tilde{\Sigma}_2))}{\chi^2_{p,0.5}} \tilde{\Sigma}_2 \tag{5}$$

By the assumption E1 of [44], the RFCH estimator can be seen as a highly outlier-resistant $\sqrt{n}$ consistent estimator. Following [43], we summarized the modified RFCH (MRFCH) algorithm in the next section:

3. The modified RFCH estimator

Our motivation comes from the fact that applying the original RFCH algorithm involves computing the covariance matrix and Mahalanobis distances. But where the number of predictor variables exceeds the sample size, computing the inverse of the covariance matrix and its corresponding Mahalanobis distances may not be feasible. To compute the inverse of the covariance matrix and, in turn, Mahalanobis distances, we need to combine the ideas of [25,32] and then propose a modified reweighted fast consistent and high breakdown (MRFCH) point estimator that uses only the diagonal elements of the covariance matrix in its computations [32]. Following [43,45], our modified algorithm is summarized as follows:

Algorithm 1

Step 1. Modified DGK algorithm

a To calculate the MDGK location T_{MDGK} and scatter matrix C_{MDGK} estimators for high dimensions, we start by calculating the classical mean vector $\bar{x}$ and covariance matrix $cov(x)$ as initial estimators (T_0, C_0) to obtain the initial Mahalanobis distances according to the next formula:

$$MD_{i0,MDGK} = \frac{\sum_{j=1}^{p}(X_{ij} - T_{0j})^2}{\sigma_{ij}} \tag{6}$$

where $\sigma_{ij} = diag(C_0)$

b Let $C_{0,MDGK} = C_0$, where the variance covariance matrix C_0 is calculated from the original dataset. Compute the variance covariance matrix of $\tilde{X}_{1,MDGK}$ to obtain the MDGK attractor $(T_{1,MDGK}, C_{1,MDGK})$

c If the diagonal of $C_{1,MDGK} = C_0$ stops, else repeat (a-c) until convergence to obtain the final attractor $(T_{k,MDGK}, C_{k,MDGK})$, where k is the convergence step.

a Begin by calculating the initial estimator of the mean and variance covariance matrix as $(T_{0,MMB}, C_{0,MMB}) = (Med(X), I_p)$ and use the initial estimators to compute the initial Mahalanobis distances as follows:

$$MD_{i0,MMB} = \frac{\sum_{j=1}^{p}(X_{ij} - Med(X))^2}{\sigma_{ij}} \tag{7}$$

b Compute the location criterion cut-off values as $luct = Med(MD_{i0,MMB})$ where $luct \neq 0.5$ and the quartile value of $MD_{0,MB}$ can be used as the cut-off point.

c Find half the dataset such that $\tilde{X}_{1,MMB} = \{X_{ij} : MD_{i0,MMB} \leq Med(MD_{i0,MMB})\}$

d Compute the average and variance covariance matrix with respect to the $X_{1,MMB}$ dataset.

e For better concentrations, calculate the Mahalanobis distances, then repeat Steps (a-d) until convergence is achieved to get the final attractor $(T_{k,MMB}, C_{k,MMB})$ where k is the convergence step.

Step 3: The Modified RFCH algorithm

a Following [25,43,45], we determine the modified FCH attractor (T_{MFCH}, C_{MFCH}) based on the MDGK and MMB attractors obtained in steps 1 and 2 as follows:

$$T_{MFCH} = \begin{cases} T_{k,MDGK}, \text{ if } \sqrt{|C_{k,MDGK}|} < \sqrt{|C_{k,MMB}|} \\ T_{k,MMB}, \text{otherwise} \end{cases} \tag{8}$$

and

$$C_{MFCH} = \begin{cases} \frac{Med(D_i^2(T_{k,MDGK}, C_{k,MDGK}))}{\chi^2_{p,0.5}} C_{k,MDGK}, \text{ if } \sqrt{|C_{k,MDGK}|} < \sqrt{|C_{k,MMB}|} \\ \frac{Med(D_i^2(T_{k,MMB}, C_{k,MMB}))}{\chi^2_{p,0.5}} C_{k,MMB}, \text{otherwise} \end{cases} \tag{9}$$

By theorem 1 of the [43], the modified FCH (MFCH) estimator can be considered a consistent estimator.

b Use the location and scatter matrix obtained from Eqs. (8) and (9) to select observations with $D_i^2(T_{MFCH}, C_{MFCH}) \leq \chi^2_{p,0.5}$

c Compute the classical estimators of the mean and variance covariance matrix based on the selected data points in step 3 (b) to obtain the RFCH attractor:

$$C_{1,MRFCH} = \frac{MED(D_i^2(T_{MFCH}, C_{MFCH}))}{\chi_{p,0.5}} C_{MFCH} \tag{10}$$

Subsequently, find the half data so that $\tilde{X}_{1,MRFCH} = \{X_{ij} : D_i^2(T_{1,MRFCH}, C_{1,MRFCH})\} \leq \chi^2_{p,0.5}$, then compute Eq. (11):

$$C_{2,MRFCH} = \frac{MED(D_i^2(T_{1,MFCH}, C_{1,MFCH}))}{\chi_{p,0.5}} C_{1,MFCH} \tag{11}$$

d Repeat (a-c) with new Eqs. (10) and (11) until convergence is achieved to obtain the final MRFCH attractor $(T_{k,MRFCH}, C_{k,MRFCH})$, where k is the convergence step.

4. Proposed outlier detection algorithm

Our proposed Mahalanobis distance-based outlier detection procedure consists of two major parts. The first part computes the location and scatter matrix from the modified algorithm and uses them to calculate the Mahalanobis distances. In the last step, we utilize the Mahalanobis distance values and the cut-off point criteria to detect and classify outliers, which is implemented by applying the next algorithm:

Algorithm 2

Let X be a data matrix with dimensions $p \gg n$

a Perform the modified RFCH algorithm on dataset X_{ij} to get the location vector and scatter matrix $(\hat{T}_{MRFCH}, \hat{C}_{MRFCH})$

b Compute the robust Mahalanobis distances based on location and scatter matrix estimates obtained in (a) as follows:

$$RD_{i,MRFCH} = \frac{\sum_{j=1}^{p} \left(X_{ij} - \hat{T}_{j,MRFCH}\right)^2}{\sigma_{ij}} \tag{12}$$

where $\sigma_{ij} = diag\left(\hat{C}_{MRFCH}\right)$

c Following [41], determine the cut-off point as

$$CP = Med\left(RD_{i,MRFCH}\right) + 3(MAD(RD_{i,MRFCH})) \tag{13}$$

d Delare any observation having $RD_{i,MRFCH} > CP$ as an outlier, otherwise non-outlier.

In the coming section, we will assess the efficiency of the MRFCH estimate of the scatter matrix using the median mean square error formula [31]. A plot of time against sample size n and time against the number of variables p is given in Figs. 1 to 4 to demonstrate the time efficiency of our proposed method compared to existing methods. Furthermore, we evaluate the detection and classification accuracy of our proposed Mahalanobis distance-based outlier detection procedure by taking the average of the false negatives (FN) and false positives (FP), which represent outliers that were not identified or masked and non-outliers that were classified as outliers or swamped. The calculation of FN and FP requires the actual and predicted values of the instances and the use of a confusion matrix.

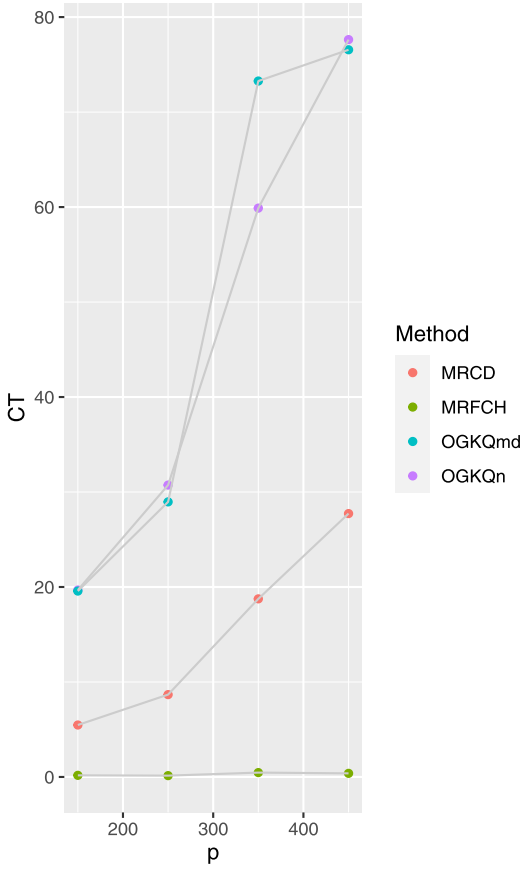

Fig. 1. Average computation time for varying p at 10% contamination with $n = 100$.

5. Simulation study

In this section, we considered two simulation examples to assess the empirical performance of our proposed methods. We designed simulation Example 1 to demonstrate the robustness of the modified algorithm in the computation of the scatter matrix. Simulation Example 2 is designed to illustrate the outlier detection and classification power of the proposed Mahalanobis distance-based outlier detection method described in Algorithm 2.

5.1. Simulation example 1

This example considered low and high-dimensional datasets, in which the datasets were generated following the simulation design of [35]. For low-dimensional data, two samples of different sizes (200 and 400) and the number of predictor variables $p = 5$ were considered. While for the high-dimensional data, we set the sample size $n = 200$ with the number of predictor variables $p =$ (400 and 800).

As per [35] X was generated from a normal distribution, that is, $X \sim N_p(0, \Sigma)$. To contaminate the data, we replaced the good observations with 10% and 20% of contaminated points. The distance k between the

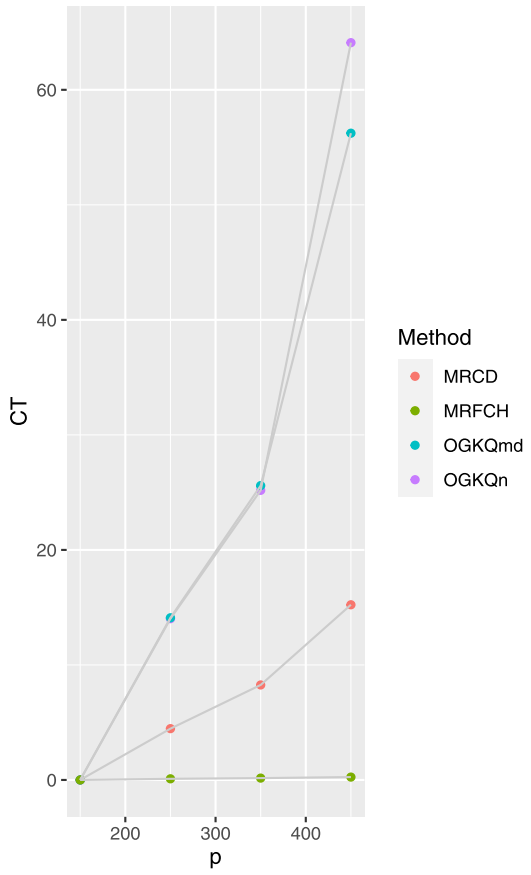

Fig. 2. Average computation time for varying p at 20% contamination with $n = 100$.

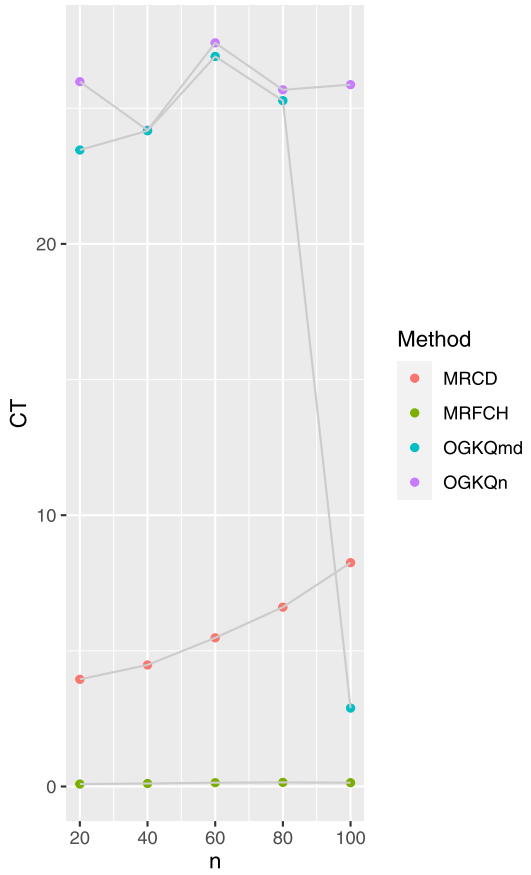

Fig. 3. Average computation time for varying n at 10% contamination with $p = 250$.

mean of the good and the contaminated observations is set at 5, 50, and 100 for small, moderate, and large contamination, respectively. For each method, we repeat the simulation 100 times ($M = 100$) to compute the median mean square error (MMSE) based on the following formulas:

$$e_{\Sigma} = MMSE = \mathbf{MED}\left(\frac{1}{p^2}\sum_{k=1}^{p}\sum_{l=1}^{p}(S_m - \Sigma_m)^2_{k,l}\right), \tag{14}$$

$m = 1,\ldots,\mathbf{M}$, where Σ_m denotes the true covariance matrix and S_m is the estimated scatter matrix obtained at each iteration. Tables 1 and 2 give the median mean squared errors. The MRCD, OGKQn, OGKmad, and MRFCH were measured for 10% and 20% contamination and different sample sizes. A good method is one that has the lowest values of MMSE. The results of the simulation for both low and high-dimensional settings show that our proposed algorithm is computationally most efficient, having the lowest MMSE values as compared to the other three methods. Furthermore, to evaluate the computational speed of our method, we generated a dataset in the same manner as described earlier. Two scenarios are considered: first, we fixed $n = 100$ while the number of variables is set to $p = (150, 250, 350,$ and $450)$, and secondly, the number of variables is set to $p = 250$, while the number of observations is set to $n = (20, 40, 60, 80$ and $100)$, each with 10% and 20% outliers and $k = 50$. The experiment is reiterated 25 times (to save time) using a computer with an Intel (Intel(R) Core(TM) i3-7020U CPU @ 2.30 GHz 2.30 GHz). It can be observed from Figs. 1 and 2 that at 10% outliers, the running time of our proposed method is much faster than the other existing methods. The OGKQn performs poorly compared to the MRCD. However, the OGKQn outperforms the MRCD and OGKmad at 20% outliers, which indicates that the speed of the OGKQn, OGKmad, and MRCD depends on the percentage of contamination as well as the number of predictor variables and sample sizes. Thus, our proposed method is the fastest and most consistent irrespective of contamination level (10% and 20%), number of predictor variables, and sample sizes compared to its competitors.

Table 1
Low dimensional scenarios with $n = 200$ and 400 and $p = 5$.

k	con	n	MMSE			
			MRCD	OGKQn	OGKmad	MRFCH
5	0.1	200	0.0229	0.0621	0.0477	0.0070
		400	0.0163	0.0737	0.0494	0.0048
	0.2	200	0.0294	0.9423	0.3811	0.0142
		400	0.0160	0.3907	0.1921	0.0107
50	0.1	200	0.0263	0.0857	0.0495	0.0085
		400	0.0139	0.0804	0.0400	0.0036
	0.2	200	0.0275	0.4429	0.2471	0.0130
		400	0.0122	0.4312	0.2678	0.0115
100	0.1	200	0.0241	0.0826	0.0309	0.0067
		400	0.0282	0.4675	0.2244	0.0129
	0.2	200	0.0128	0.0725	0.0328	0.0050
		400	0.0159	0.3426	0.2018	0.0110

5.2. Simulation example 2

In this example, we investigate the classification accuracy of our proposed method as described in Section 4. Following [32,35], for $n = 50$, and 100, we simulated clean observations from $N_p(0, \Sigma)$ and outliers are simulated from $N_p(kb_i, \Sigma)$ where b_i is a vector of p dimensional independent random variables generated from $Unif(0, 1)$ with L_2 norm, $k = p^{\frac{1}{2}}$ and Σ denoted the autoregressive correlation matrix with $\rho_{ij} = 0.5^{|i-j|}$. The outlier proportions are used as 0.1 and 0.2. The number variables p are set to 200, 400 and 800 respectively. The MRCD, MDP, and MRFCH methods were then applied to the datasets.

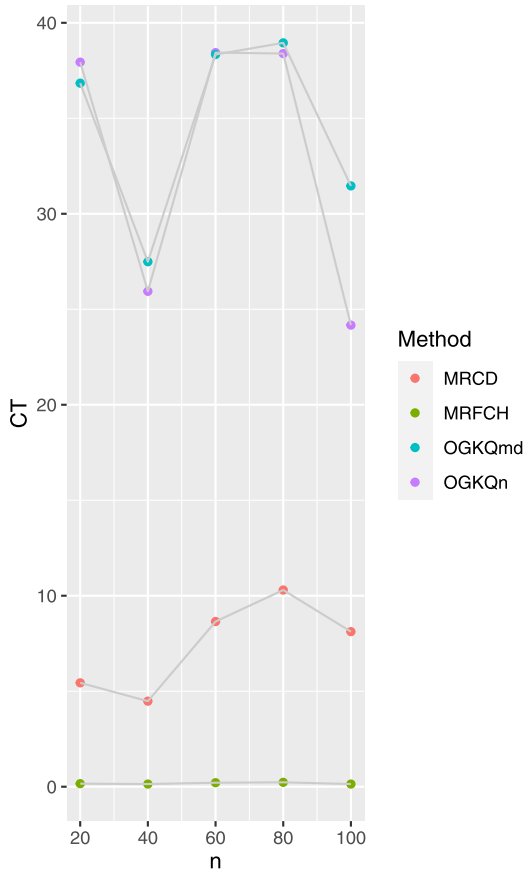

Fig. 4. Average computation time for varying n at 20% contamination with $p = 250$.

Table 2
High dimensional scenarios with $n = 200$ and $p = 400$ and 800.

k	con	p	MMSE			
			MRCD	OGKQn	OGKmad	MRFCH
5	0.1	400	0.0078	0.0199	0.0136	0.0060
		800	0.0069	0.0188	0.0139	0.0060
	0.2	400	0.0080	0.1279	0.0754	0.0067
		800	0.0078	0.0996	0.0669	0.0069
50	0.1	400	0.0075	0.0118	0.0099	0.0059
		800	0.0065	0.0121	0.0100	0.0058
	0.2	400	0.0080	0.0451	0.0318	0.0066
		800	0.0075	0.0409	0.0282	0.0067
100	0.1	400	0.0077	0.0138	0.0114	0.0059
		800	0.0079	0.0355	0.0283	0.0066
	0.2	400	0.0068	0.0128	0.0113	0.0061
		800	0.0074	0.0310	0.0240	0.0068

Table 3
False positive (FP) and false negative (FN) values of the simulation example 2.

n	con	p	MRCD		MDP		MRFCH	
			FP	FN	FP	FN	FP	FN
50	0.1	200	7	0	4.16	0	0.36	0
		400	7	0	4.60	0	0.36	0
		800	7	0	0.12	0	0.12	0
	0.2	200	2.04	0.04	4.64	0	0.44	0
		400	2	0	3.16	0	0.44	0
		800	2	0	2.88	0	0.16	0
100	0.1	200	15	0	5.24	0	1.36	0
		400	15	0	4.88	0	1.60	0
		800	15	0	4.52	0	2.68	0
	0.2	200	5.08	0.08	3.20	0	0.24	0.04
		400	5	0	3.36	0	0.24	0
		800	5	0	2.72	0	0.32	0

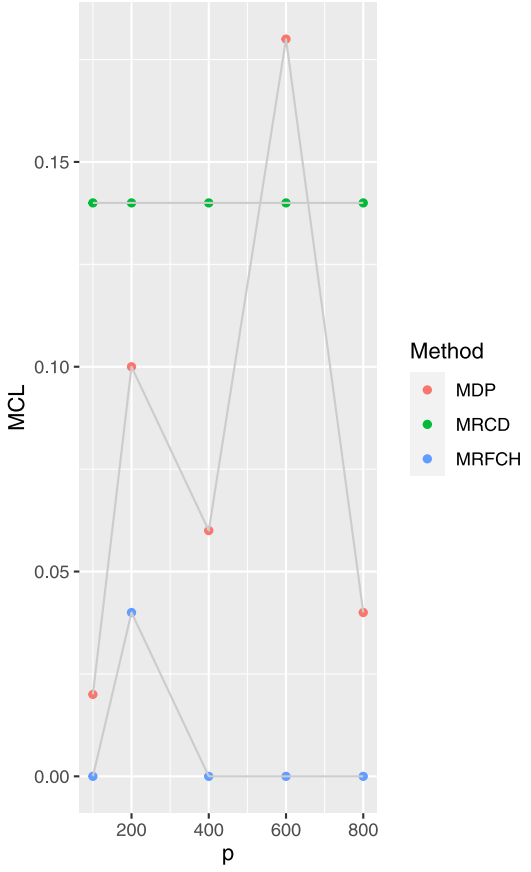

Fig. 5. Average MCL for varying p at 10% contamination with $n = 50$.

We considered MRCD and MDP because they are the most popular and most recent methods in the literature. The results in Table 3 show that our method performs excellently well, with less misclassification error. To further evaluate the performance of our proposed algorithm, Figs. 5 to Fig. 8 display the misclassification error (MCL) of the three considered methods: MRCD, MDP, and MRFCH, based on the varying number of predictors p and number of samples n. Figs. 5 and 6 clearly show that the MDP performs better than the MRCD for varying p at 10% and 20% contamination levels, except in Fig. 5 when $p = 600$. Similarly, in Figs. 7 and 8, the MRCD exhibits poor performance compared to the MDP, except in Fig. 8 when $n = 100$. This shows that the MRCD and MDP have, to a lesser extent, less detection and classification accuracy compared to our proposed procedure, which performs well over all its competitors, maintaining satisfactory misclassification errors even when the percentage of the outliers increases with an increase in sample size and dimensions. In the latter section, we will illustrate the performance of the proposed method using four high-dimensional real-life datasets.

6. Real life examples

Four high-dimensional real-life datasets are used to compare our proposed Mahalanobis distance-based outlier detection algorithm to the MRCD and MDP methods.

6.1. Octane dataset

The octane dataset contains 39 samples and two hundred and twenty-six variables, out of which $x_{25} - x_{26}$ and $x_{36} - x_{39}$ are known outliers. Several researchers, such as [26,27,31,66] have used this

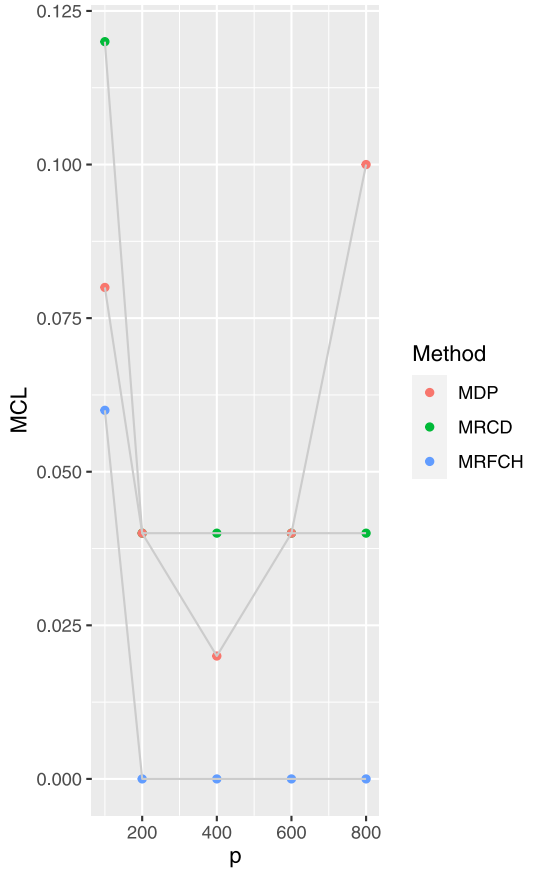

Fig. 6. Average MCL for varying p at 20% contamination with $n = 50$.

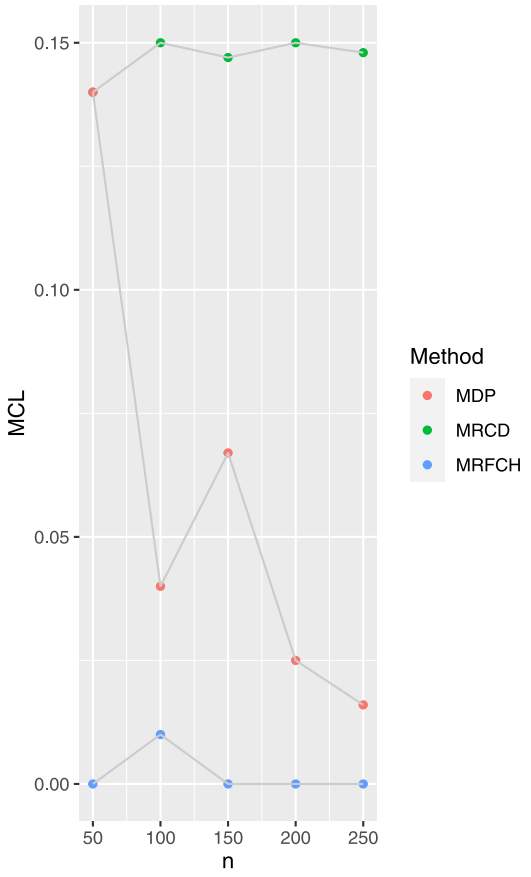

Fig. 7. Average MCL for varying n at 10% contamination with $p = 400$.

dataset to demonstrate the robustness and efficiency of their methods. We considered this dataset because of our prior knowledge about the exact number of outliers and their positions. This will ease the evaluation process of comparing our proposed method with other existing approaches. Besides, the dataset is high-dimensional, which conforms to the aim of developing our method. Fig. 9 displays the Mahalanobis distance values based on MRFCH for detecting outliers. In Fig. 9 x-axis represents the number of gasoline samples, y-axis denotes the RD_MRFCH, and the level line indicates the cut-off value, i.e., $Med(RD_{i,MRFCH}) + 3(MAD(RD_{i,MRFCH}))$. The result in Fig. 9 shows that Mahalanobis distance-based MRFCH algorithm successfully detects six (6) outliers ($x_{25}-x_{26}$ and $x_{36}-x_{39}$) as indicated in the original dataset and does not misclassify. Our results are consistent with those in [27,31,66], confirming our proposed method as a viable method.

6.2. Glass vessels dataset

The glass vessel dataset arises from a study of electron-probe X-ray microanalysis (EXPMA). Resulting in 1920 frequency measures on each of 180 glass vessels following an experiment conducted on 16th–17th-century archaeological glass vessels in a laboratory at the University of Antwerp. This experiment aimed to learn more about the production of these vessels, especially regarding their foundation and possible trade links between known producers. The number of predictor variables p in this dataset exceeds the sample size n. Several analytical methods have been applied to this dataset, which led to an accurate determination of the presence of outlying points in the data. [67] used this data to demonstrate the advantage of the partial robust M estimators over the partial least squares. Recently, [68,69] applied the dataset to a robust linear regression model for high dimensions. Although these authors acknowledged that the glass vessel data contains some outlier points, no researcher clearly pointed out which observations are outliers and which are not. [30] categorically tagged the last 38 observations as clear multivariate outliers using the distance diagnostics plot. These issues are of broad interest in the application of robust Mahalanobis distance-based outlier detection in high-dimensional datasets. However, our aim is to use this dataset to validate the results of the simulation given in Table 3, since the exact position and number of outliers in the data are known. Following [30], we removed all columns with MAD equal to zero, leaving the remaining 1905 columns of predictors for outlier investigation. We applied the MDP, MRCD, and MRFCH to the data. The results of classification performance for each method are presented as diagnostic plots in Figs. 10 to 12. It can be seen that all three methods (MDP, MRCD, and MRFCH) can detect all the last 38 observations as outliers, showing zero masking error (FN). On the other hand, all methods detect some points that are not outliers as outliers (FP), which are all evident from the results in Table 3. The MDP detects 44 points that are non-outliers as outliers, with most of the non-outliers merely at the boundary. The MRCD detects all the non-outliers as outliers without detecting even a single inlier correctly. Our proposed detection method incorrectly detects only 27 outliers with a clear cut-off line between the inliers and the outliers, except for a few points that are close to the cut-off line. Our result also confirmed that the non-parametric cut-off point criteria used in this paper are more robust than the other benchmark cut-off values. Therefore, we conclude that our proposed outlier detection algorithm achieves optimal accuracy among all the benchmark methods.

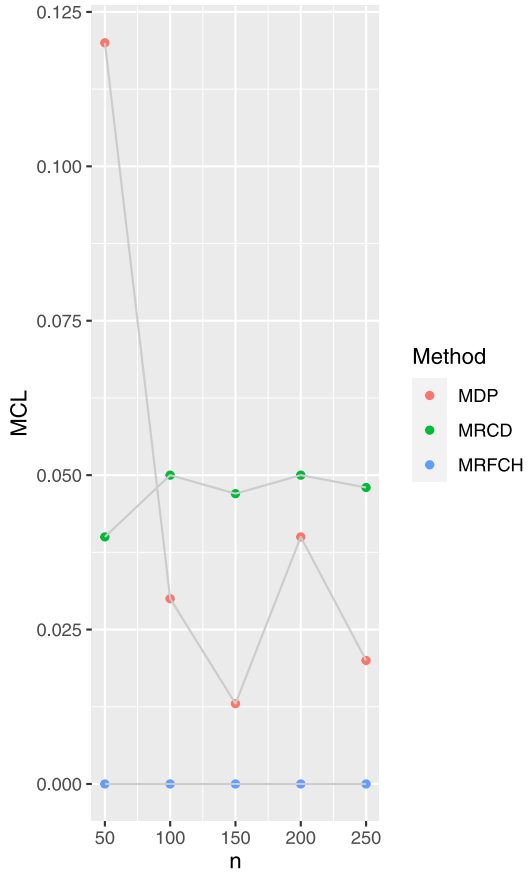

Fig. 8. Average MCL for varying n at 20% contamination with $p = 400$.

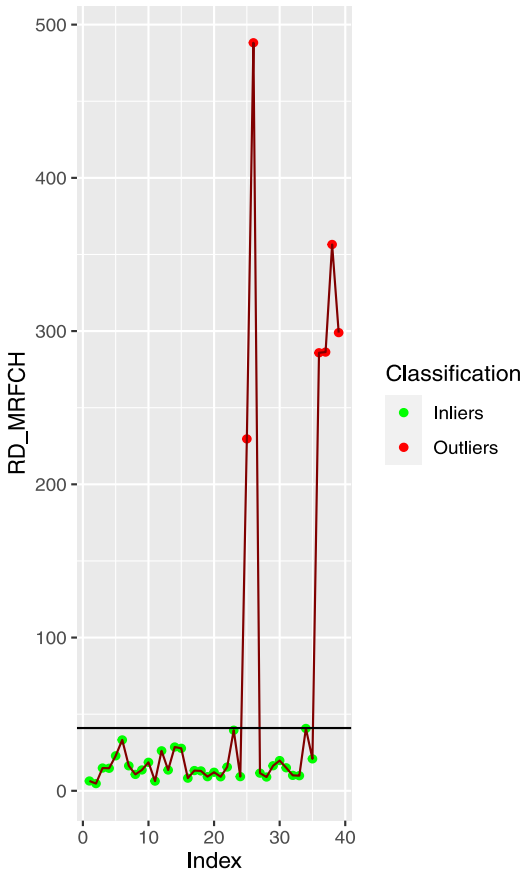

Fig. 9. Diagnostic plot for octane dataset based on the MRFCH algorithm.

6.3. NCI60 dataset

This dataset is high-dimensional, consisting of the expression levels of 6830 genes obtained from 64 cancer cell lines. The dataset was formally analyzed by [35] to illustrate behavior and an outlier detection problem. The dataset was obtained from the ISLR package in the R programming language. Following [35], we reduced the original dimension of the data to 320 by 64 to save computational time. The contaminated observations are planted by replacing 10% and 20% of the good observations by 50 to ensure that the data contains outlying points. We applied our proposed detection method and two other benchmark procedures to the modified dataset. The results in Table 4 show that our proposed method has detected outliers that are added with 100% success. The MRCD and the MDP also achieved 100% success. All at 10% and 20% contamination. The MRCD incorrectly identified 23 and 15 observation outliers for 10% and 20% contamination. MDP incorrectly detected 10 and 7 points, and our method, which achieves optimal accuracy among all the tested methods, identified only 3 and 1 outlying observations at 10% and 20%, showing the strength of the newly developed detection algorithm in solving outlier classification problems.

6.4. Brain dataset

The brain cancer dataset contains microarray expression observations for 42 brain cancer samples. The number of expressions on each array is 5597. Detailed information about this dataset can be obtained from the rda package in the R programming language. The brain dataset was previously analyzed in [35]. Similar to the NCI60 dataset, we reduced the number of observations from 5597 to 210 to save computational cost and add 10% and 20% outliers to the reduced dataset [35]. We applied all the methods used in the previous example. Table 4 displays the results for brain datasets. Also, we observed that our proposed Mahalanobis distance-based outlier detection worked well for detection and outlier classification, as it continued to maintain its position among all benchmark methods in this paper. The results of both the NCI60 and Brain datasets show consistency with the simulation results in Section 5 since the proposed algorithm and all benchmark methods detect zero (0) false negatives at 10% and 20% contamination levels.

Table 4
False Positive (FP) and False Negative (FN) for NCI60 ($n = 64$) and Brain ($n = 42$) datasets.

Data	con	MRCD		MDP		MRFCH	
		FP	FN	FP	FN	FP	FN
NCI60	0.1	23	0	10	0	3	0
	0.2	15	0	7	0	1	0
Brain	0.1	16	0	15	0	3	0
	0.2	12	0	6	0	1	0

7. Conclusion

In this paper, we developed a modified reweighted fast consistent and high break-down point estimator for a high-dimensional dataset.

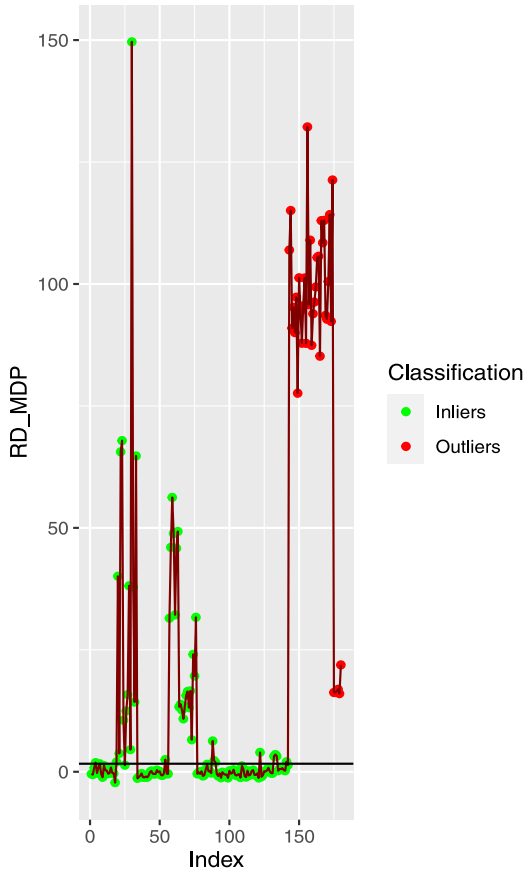

Fig. 10. Diagnostic plot for Glass vessels dataset based on the MDP algorithm.

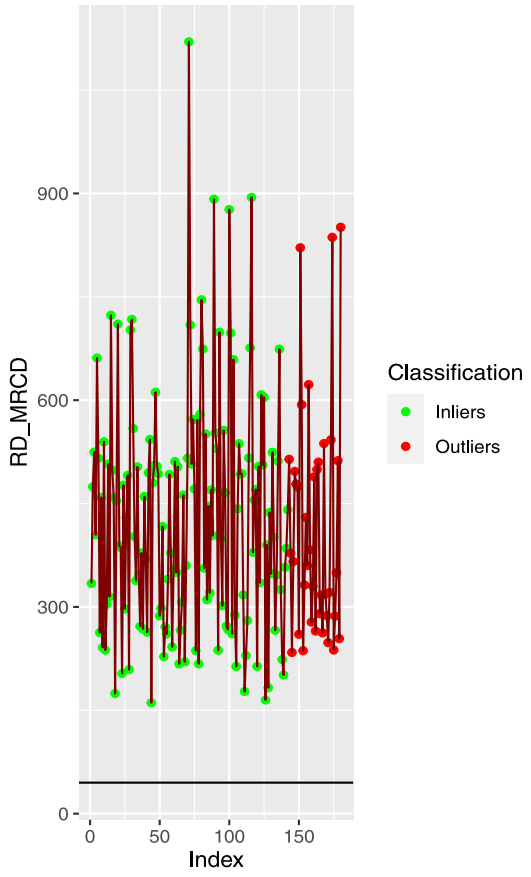

Fig. 11. Diagnostic plot for Glass vessels dataset based on MRCD algorithm.

Modeling multivariate, high-dimensional data poses numerous challenges that may lead to misleading interpretations and conclusions. Some common challenges related to high-dimensional data include computational complexity, the curse of dimensionality, and singularity, in addition to the outlier problem. Our modified algorithm attempts to address these problems. First, we apply the idea of [32] to compute the Mahalanobis distances used within the MRFCH algorithm. The final attractors from the MRFCH algorithm were used to construct a Mahalanobis distance-based outlier detection method. The Mahalanobis distance values were used to determine the threshold criteria for the identification of outliers and classification. The resulting procedure enjoys the inherent robustness properties of the original reweighted fast, consistent high break-down point estimator. The modified algorithm is tested on two simulation examples and four high-dimensional real-life datasets. The median mean squared error and misclassification error metrics were used to evaluate the performance of our proposed method. Our numerical results indicate that as the contamination level increases with increased dimension, our method consistently outperforms its competitors. This becomes more prevalent, especially for high-dimensional datasets where the number of variables is greater than the sample size. Compared to the MRCD and OGK, our modified algorithm performs relatively well in terms of computing speed. Real-life data results have shown that the new modified method is able to deal with high-dimensional datasets while achieving a small misclassification error rate. Our modified algorithm can be seen as a substantive alternative to MRCD and MDP in computing multivariate estimates of location and scatter matrix. Additionally, our method can be applied to calculating correlation coefficients and a robust Mahalanobis distance-based outlier detection technique. Specifically, several modern applications of Mahalanobi's distance-based outlier detection method exist in the literature; among them are anomaly detection, cluster analysis, image analysis, and remote sensing. Our modified algorithm is a useful tool to explore the structure of a dataset. It can also be applied to commonly used multivariate statistics such as discriminant analysis, principal component analysis, regression analysis, and correlation analysis. A sparse modified reweighted fast consistent and high breakdown point estimator and a cellwise MRFCH outlier detection-based algorithm will be nice contributions to knowledge. Robust diagnostic generalized potentials based MRFCH for the identification of high leverage points in a high-dimensional space can be proposed. Finally, robust outlier maps in a high-dimensional setting will be excellent future research topics in this area.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ishaq Abdullahi Baba reports financial support was provided by Petroleum Technology Development Fund. Habshah Midi reports financial support was provided by Universiti Putra Malaysia. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

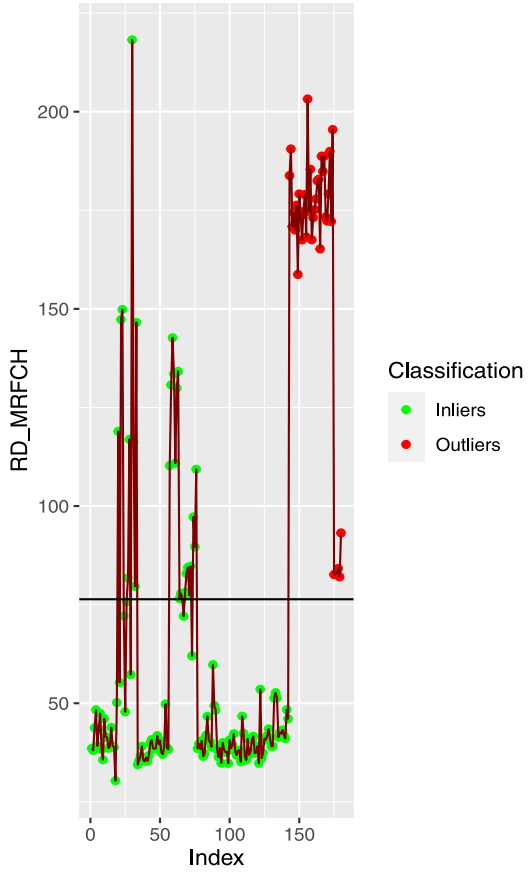

Fig. 12. Diagnostic plot for Glass vessels dataset based on the MRFCH algorithm.

References

[1] A.H. Shirdel, J.M. Böling, H.T. Toivonen, System identification in the presence of trends and outliers using sparse optimization, J. Process Control 44 (2016) 120–133, http://dx.doi.org/10.1016/j.ifacol.2017.08.1849.

[2] J. Byers, I. Popova, B. Simkins, Robust estimation of conditional risk measures using machine learning algorithm for commodity futures prices in the presence of outliers, J. Commod. Mark. 24 (2021) 100174, http://dx.doi.org/10.1016/j.jcomm.2021.100174.

[3] A.A. Amponsah, A.F. Adekoya, B.A. Weyori, A novel fraud detection and prevention method for healthcare claim processing using machine learning and blockchain technology, Decis. Anal. J. 4 (2022) 100122, http://dx.doi.org/10.1016/j.dajour.2022.100122.

[4] L. Luo, W. Wang, S. Bao, X. Peng, Y. Peng, Robust and sparse canonical correlation analysis for fault detection and diagnosis using training data with outliers, Expert Syst. Appl. 236 (2024) 121434, http://dx.doi.org/10.1016/j.eswa.2023.121434.

[5] A. Zimek, E. Schubert, H.-P. Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data, Stat. Anal. Data Min. ASA Data Sci. J. 5 (5) (2012) 363–387, http://dx.doi.org/10.1002/sam.11161.

[6] J. Zhu, Z. Ge, Z. Song, F. Gao, Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annu. Rev. Control 46 (2018) 107–133, http://dx.doi.org/10.1016/j.arcontrol.2018.09.003.

[7] A. Smiti, A critical overview of outlier detection methods, Comp. Sci. Rev. 38 (2020) 100306, http://dx.doi.org/10.1016/j.cosrev.2020.100306.

[8] M. Olteanu, F. Rossi, F. Yger, Meta-survey on outlier and anomaly detection, Neurocomputing 555 (2023) 126634, http://dx.doi.org/10.1016/j.neucom.2023.126634.

[9] G.O. Campos, A. Zimek, J. Sander, R.J. Campello, B. Micenková, E. Schubert, I. Assent, M.E. Houle, On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study, Data Min. Knowl. Discov. 30 (2016) 891–927, http://dx.doi.org/10.1007/s10618-015-0444-8.

[10] C.C. Aggarwal, C.C. Aggarwal, Supervised outlier detection, Outl. Anal. (2017) 219–248, http://dx.doi.org/10.1145/375663.375668.

[11] C.S.K. Dash, A.K. Behera, S. Dehuri, A. Ghosh, An outliers detection and elimination framework in classification task of data mining, Decis. Anal. J. 6 (2023) 100164, http://dx.doi.org/10.1016/j.dajour.2023.100164.

[12] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The mahalanobis distance, Chemometr. Intell. Lab. Syst. 50 (1) (2000) 1–18, http://dx.doi.org/10.1016/S0169-7439(99)00047-7.

[13] P. Filzmoser, R.G. Garrett, C. Reimann, Multivariate outlier detection in exploration geochemistry, Comput. Geosci. 31 (5) (2005) 579–587, http://dx.doi.org/10.1016/j.cageo.2004.11.013.

[14] X. Li, S. Deng, L. Li, Y. Jiang, Outlier detection based on robust mahalanobis distance and its application, Open J. Stat. 9 (1) (2019) 15–26, https://dpi.org/10.4236/ojs.2019.91002.

[15] E. Cabana, R.E. Lillo, H. Laniado, Multivariate outlier detection based on a robust mahalanobis distance with shrinkage estimators, Statist. Papers 62 (2021) 1583–1609, http://dx.doi.org/10.1007/s00362-019-01148-1.

[16] M. Mayrhofer, P. Filzmoser, Multivariate outlier explanations using Shapley values and Mahalanobis distances, Econom. Stat. (2023) http://dx.doi.org/10.1016/j.ecosta.2023.04.003.

[17] H.A. Lim, H. Midi, Diagnostic robust generalized potential based on index set equality (DRGP (ISE)) for the identification of high leverage points in linear model, Comput. Stat. 31 (2016) 859–877, http://dx.doi.org/10.1007/s00180-016-0662-6.

[18] P.S. Dhamale, A.S. Kashikar, Outlier detection in cylindrical data based on Mahalanobis distance, Comm. Statist. Simulation Comput. (2023) 1–11, http://dx.doi.org/10.1080/03610918.2023.2252630.

[19] J. Raymaekers, P.J. Rousseeuw, Fast robust correlation for high-dimensional data, Technometrics 63 (2) (2021) 184–198, http://dx.doi.org/10.1080/00401706.2019.1677270.

[20] J. Raymaekers, P.J. Rousseeuw, The cellwise minimum covariance determinant estimator, J. Amer. Statist. Assoc. (2023) 1–12, http://dx.doi.org/10.1080/01621459.2023.2267777.

[21] S. Van Aelst, P. Rousseeuw, Minimum volume ellipsoid, Wiley Interdiscip. Rev. Comput. Stat. 1 (1) (2009) 71–82, http://dx.doi.org/10.1002/wics.19.

[22] S. Rosa, R. Harman, Computing minimum-volume enclosing ellipsoids for large datasets, Comput. Statist. Data Anal. 171 (2022) 107452, http://dx.doi.org/10.1016/j.csda.2022.107452.

[23] P.J. Rousseeuw, K.V. Driessen, A fast algorithm for the minimum covariance determinant estimator, Technometrics 41 (3) (1999) 212–223, http://dx.doi.org/10.1080/00401706.1999.10485670.

[24] M. Hubert, M. Debruyne, P.J. Rousseeuw, Minimum covariance determinant and extensions, Wiley Interdiscip. Rev. Comput. Stat. 10 (3) (2018) e1421, http://dx.doi.org/10.1002/wics.1421.

[25] D.J. Olive, D.M. Hawkins, Robust multivariate location and dispersion, 2010, Preprint, see (www.math.siu.edu/olive/preprints.htm).

[26] R.A. Maronna, R.H. Zamar, Robust estimates of location and dispersion for high-dimensional datasets, Technometrics 44 (4) (2002) 307–317, http://dx.doi.org/10.1198/004017002188618509.

[27] M. Hubert, P.J. Rousseeuw, K. Vanden Branden, ROBPCA: a new approach to robust principal component analysis, Technometrics 47 (1) (2005) 64–79, http://dx.doi.org/10.1198/004017004000000563.

[28] I. Souiden, M.N. Omri, Z. Brahmi, A survey of outlier detection in high dimensional data streams, Comp. Sci. Rev. 44 (2022) 100463, http://dx.doi.org/10.1016/j.cosrev.2022.100463.

[29] D. Peña, V.J. Yohai, A review of outlier detection and robust estimation methods for high dimensional time series data, Econom. Stat. (2023) http://dx.doi.org/10.1016/j.ecosta.2023.02.001.

[30] P. Filzmoser, R. Maronna, M. Werner, Outlier identification in high dimensions, Comput. Statist. Data Anal. 52 (3) (2008) 1694–1711, http://dx.doi.org/10.1016/j.csda.2007.05.018.

[31] K. Boudt, P.J. Rousseeuw, S. Vanduffel, T. Verdonck, The minimum regularized covariance determinant estimator, Stat. Comput. 30 (1) (2020) 113–128, http://dx.doi.org/10.1007/s11222-019-09869-x.

[32] K. Ro, C. Zou, Z. Wang, G. Yin, Outlier detection for high-dimensional data, Biometrika 102 (3) (2015) 589–599, http://dx.doi.org/10.1093/biomet/asv021.

[33] R.A. Maronna, V.J. Yohai, The behavior of the Stahel–Donoho robust multivariate estimator, J. Amer. Statist. Assoc. 90 (429) (1995) 330–341, http://dx.doi.org/10.2307/2291158.

[34] V. Fritsch, G. Varoquaux, B. Thyreau, J.-B. Poline, B. Thirion, Detecting outlying subjects in high-dimensional neuroimaging datasets with regularized minimum covariance determinant, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2011, pp. 264–271, http://dx.doi.org/10.1007/978-3-642-23626-6_33.

[35] H. Bulut, Mahalanobis distance based on minimum regularized covariance determinant estimators for high dimensional data, Comm. Statist. Theory Methods 49 (24) (2020) 5897–5907, http://dx.doi.org/10.1080/03610926.2020.1719420.

[36] S.X. Chen, Y.-L. Qin, et al., A two-sample test for high-dimensional data with applications to gene-set testing, Ann. Statist. 38 (2) (2010) 808–835, https://www.jstor.org/stable/25662261.

[37] A.M. Baba, H. Midi, M.B. Adam, N.H.A. Rahman, Detection of influential observations in spatial regression model based on outliers and bad leverage classification, Symmetry 13 (11) (2021) 2030, http://dx.doi.org/10.3390/sym13112030.

[38] R.R. Wilcox, Introduction to Robust Estimation and Hypothesis Testing, Academic Press, 2011.

[39] A.S. Hadi, A new measure of overall potential influence in linear regression, Comput. Statist. Data Anal. 14 (1) (1992) 1–27, http://dx.doi.org/10.1016/0167-9473(92)90078-T.
[40] A.R. Imon, A.S. Hadi, Identification of multiple high leverage points in logistic regression, J. Appl. Stat. 40 (12) (2013) 2601–2616, http://dx.doi.org/10.1080/02664763.2013.822057.
[41] M. Habshah, M. Norazan, A. Rahmatullah Imon, The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression, J. Appl. Stat. 36 (5) (2009) 507–520, http://dx.doi.org/10.1080/02664760802553463.
[42] A. Alkenani, K. Yu, A comparative study for robust canonical correlation methods, J. Stat. Comput. Simul. 83 (4) (2013) 692–720, http://dx.doi.org/10.1080/00949655.2011.632775.
[43] H.S. Uraibi, H. Midi, On robust bivariate and multivariate correlation coefficient, Econ. Comput. Econ. Cybern. Stud. Res. 53 (2) (2019) http://dx.doi.org/10.24818/18423264/53.2.19.13.
[44] J. Zhang, D.J. Olive, P. Ye, Robust covariance matrix estimation with canonical correlation analysis, Int. J. Stat. Probab. 1 (2) (2012) 119, http://dx.doi.org/10.5539/ijsp.v1n2p119.
[45] O.S. Ibrahim, M.J. Mohammed, A proposed method for cleaning data from outlier values using the robust RFCH method in structural equation modeling, Int. J. Nonlinear Anal. Appl. 12 (2) (2021) 2269–2293, http://dx.doi.org/10.22075/IJNAA.2021.5374.
[46] M. Riahi-Madvar, A.A. Azirani, B. Nasersharif, B. Raahemi, A new density-based subspace selection method using mutual information for high dimensional outlier detection, Knowl.-Based Syst. 216 (2021) 106733, http://dx.doi.org/10.1016/j.knosys.2020.106733.
[47] Z. Li, L. Zhang, An ensemble outlier detection method based on information entropy-weighted subspaces for high-dimensional data, Entropy 25 (8) (2023) 1185, http://dx.doi.org/10.3390/e25081185.
[48] C.C. Aggarwal, P.S. Yu, Outlier detection with uncertain data, in: Proceedings of the 2008 SIAM International Conference on Data Mining, SIAM, 2008, pp. 483–493, http://dx.doi.org/10.1137/1.9781611972788.4.
[49] C. Leys, O. Klein, Y. Dominicy, C. Ley, Detecting multivariate outliers: Use a robust variant of the mahalanobis distance, J. Exp. Soc. Psychol. 74 (2018) 150–156, http://dx.doi.org/10.1016/j.jesp.2017.09.011.
[50] Y. Xu, Z. Wu, J. Chanussot, Z. Wei, Joint reconstruction and anomaly detection from compressive hyperspectral images using mahalanobis distance-regularized tensor RPCA, IEEE Trans. Geosci. Remote Sens. 56 (5) (2018) 2919–2930, http://dx.doi.org/10.1109/TGRS.2017.2786718.
[51] R. Domingues, F. Buonora, R. Senesi, O. Thonnard, An application of unsupervised fraud detection to passenger name records, in: 2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshop, DSN-W, IEEE, 2016, pp. 54–59, http://dx.doi.org/10.1109/DSN-W.2016.21.
[52] M. Vishwakarma, N. Kesswani, A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection, Decis. Anal. J. 7 (2023) 100233, http://dx.doi.org/10.1016/j.dajour.2023.100233.
[53] C. Fauconnier, G. Haesbroeck, Outliers detection with the minimum covariance determinant estimator in practice, Stat. Methodol. 6 (4) (2009) 363–379, http://dx.doi.org/10.1016/j.stamet.2008.12.005.
[54] P.J. Rousseeuw, Multivariate estimation with high breakdown point, Math. Stat. Appl. 8 (283–297) (1985) 37, http://dx.doi.org/10.1007/978-94-009-5438-0_20.
[55] J. Liu, Y. Qiao, Mahalanobis distance–based kernel supervised machine learning in spectral dimensionality reduction for hyperspectral imaging remote sensing, Int. J. Distrib. Sens. Netw. 16 (11) (2020) 1550147720968467, http://dx.doi.org/10.1177/1550147720968.
[56] P.O. Brown, M.C. Chiang, S. Guo, Y. Jin, C.K. Leung, E.L. Murray, A.G. Pazdor, A. Cuzzocrea, Mahalanobis distance based k-means clustering, in: International Conference on Big Data Analytics and Knowledge Discovery, Springer, 2022, pp. 256–262, http://dx.doi.org/10.1007/978-3-031-12670-3_23.
[57] H.S. Uraibi, H. Midi, S. Rana, Selective overview of forward selection in terms of robust correlations, Comm. Statist. Simulation Comput. 46 (7) (2017) 5479–5503, http://dx.doi.org/10.1080/03610918.2016.1164862.
[58] H. Midi, H.T. Hendi, J. Arasan, H. Uraibi, Fast and robust diagnostic technique for the detection of high leverage points, Pertanika J. Sci. Technol. 28 (4) (2020) http://dx.doi.org/10.47836/pjst.28.4.05, https://dpi.org/10.47836/pjst.28.4.05.
[59] P. Rousseeuw, A. Leroy, Robust Regression and Outlier Detection: Wiley Interscience, New York, 1987, http://dx.doi.org/10.1002/0471725382.
[60] A.R. Imon, M.R. Apu, Detection of high leverage points using a nonparametric cut-off point for the robust mahalanobis distance, Malays. J. Math. Sci. 10 (3) (2016) 283–295.
[61] C.C. Aggarwal, P.S. Yu, Outlier detection for high dimensional data, in: Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, 2001, pp. 37–46, http://dx.doi.org/10.1145/375663.375668.
[62] F. Kamalov, H.H. Leung, Outlier detection in high dimensional data, J. Inf. Knowl. Manag. 19 (01) (2020) 2040013, http://dx.doi.org/10.1142/S0219649220400134.
[63] P. Ampanthong, P. Suwattee, A comparative study of outlier detection procedures in multiple linear regression, in: Proceedings of the International Multiconference of Engineers and Computer Scientists, Vol. 1, 2009, https://www.researchgate.net/publication/44259638.
[64] H. Sarmadi, A. Karamodin, A novel anomaly detection method based on adaptive mahalanobis-squared distance and one-class kNN rule for structural health monitoring under environmental effects, Mech. Syst. Signal Process. 140 (2020) 106495, http://dx.doi.org/10.1016/j.ymssp.2019.106495.
[65] S.J. Devlin, R. Gnanadesikan, J.R. Kettenring, Robust estimation of dispersion matrices and principal components, J. Amer. Statist. Assoc. 76 (374) (1981) 354–362, https://www.jstor.org/stable/2287836.
[66] A. Mohammed Rashid, H. Midi, W. Dhhan, J. Arasan, Detection of outliers in high-dimensional data using nu-support vector regression, J. Appl. Stat. (2021) 1–20, http://dx.doi.org/10.1080/02664763.2021.1911965.
[67] P. Lemberge, I. De Raedt, K.H. Janssens, F. Wei, P.J. Van Espen, Quantitative analysis of 16–17th century archaeological glass vessels using PLS regression of EPXMA and μ-XRF data, J. Chemom. A J. Chemom. Soc. 14 (5–6) (2000) 751–763, http://dx.doi.org/10.1002/1099-128X(200009/12)14:5/63.0.CO;2-D.
[68] E. Smucler, V.J. Yohai, Robust and sparse estimators for linear regression models, Comput. Statist. Data Anal. 111 (2017) 116–130, http://dx.doi.org/10.1016/j.csda.2017.02.002.
[69] U. Amato, A. Antoniadis, I. De Feis, I. Gijbels, Penalised robust estimators for sparse and high-dimensional linear models, Stat. Methods Appl. 30 (1) (2021) 1–48, http://dx.doi.org/10.1007/s10260-020-00511-z.