A Multiple-Case Deletion Approach for Detecting Influential Points in High-Dimensional Regression

Tao Wang[a,b], Qun Li[c], Qingpei Zang[a] and Zhonghua Li[b,*]

[a]School of Mathematical Sciences, Huaiyin Normal University
Huaian City, 223300, P. R. China
[b]Institute of Statistics and LPMC, Nankai University
Tianjin City, 300071, P. R. China
[c]Department of Physiology, University of Szeged
Szeged, 6720, Hungary

Abstract

In high-dimensional regression, the presence of influential observations may lead to inaccurate analysis results so that it is a prime and important issue to detect these unusual points before statistical regression analysis. Most of the traditional approaches are, however, based on single-case diagnostics, and they may fail due to the presence of multiple influential observations that suffer from masking effects. In this paper, an adaptive multiple-case deletion approach is proposed for detecting multiple influential observations in the presence of masking effects in high-dimensional regression. The procedure contains two stages. Firstly, we propose a multiple-case deletion technique, and obtain an approximate clean subset of the data that is presumably free of influential observations. To enhance efficiency, in the second stage, we refine the detection rule. Monte Carlo simulation studies and a real-life data analysis investigate the effective performance of the proposed procedure.

Keywords: Influential points; Masking; Regression diagnostics; High-dimensional regression.

1 Introduction

An observation is flagged as influential if some important features of the regression analysis are noticeably changed after this observation is removed in the regression model, (see Cook, 1977, Belsley et al., 1980). The presence of these abnormal observations would possibly lead to erroneous influential results, therefore, it is necessary to detect such observations and remove them in regression analysis. As Davies and Gather (1993) pointed out, although the detection of outliers in a univariate sample has been investigated extensively in the statistical literature, the word "outlier" has never been given a precise definition.

In the past few decades, many statisticians have concentrated on the problem of detecting influential observations in various regression models, and many effective approaches have been proposed. The common approaches are by utilizing single-case deletion measures to detect the observations that unduly affect the ordinary least square (OLS) estimate, for example, Cook's distance, DFBETAS, and DFFITS, (see reference, Cook, 1977, 1979, Chatterjee et al., 1988, Cook and Weisberg, 1982, etc.). These single-case deletion methods can work well for detecting the problematic observations when the data set contain a single or few influential points.

Address correspondence to Zhonghua Li, Institute of Statistics, Nankai University, China; E-mail: zli@nankai.edu.cn

However, when there are a certain proportion of influential points in a data set, the single-case deletion methods may fail to identify all the problematic observations. This is likely due to the existence of multiple influential observations that may suffer from the so-called "masking effect", implying that when one or more influential points were deleted from the data set, another observation may emerge as extremely influential, which was not visible at first, (see reference Rousseeuw and Leroy, 1987). In other words, the real influential observations may not look like influential, therefore, multiple influential observations may mask each other and go undetected. The opposite effect is known as swamping (see Barnett and Lewis, 1994, Hadi and Simonoff, 1993), which makes non-influential observations appear influential.

To reduce the impact of masking or swamping effects by single-case deletion approaches, an effective way is by use of multiple-case deletion approach, which was proposed first by Belsley et al. (1980). It is a direct approach, and the main idea is deleting more than a single observation once, and constructing measures with and without these deleted observations. Due to the development of computing tools, many related studies have emerged, see, for example, Pena and Yohai (1999), Becker and Gather (2001), Pena (2005), Imon (2005), Nurunnabi et al. (2014), etc. Roberts et al. (2015) built upon these previous work and proposed an adaptive, automatic multiple-case deletion procedure (ADAP, in short) to detect influential points in linear regression model, and simulation studies and three real-data examples showed that the ADAP procedure is an effective method.

In high dimensional regression, data sets usually consist of a few influential points inevitably, as the dimensionality of the data increases, both the chance of an observation being influential and its potential impact on the analysis results may be amplified. Zhao et al. (2013) defined the high-dimensional influential points, and proposed high-dimensional influence measure (HIM in short) which captures the influence on the marginal correlations for high-dimensional linear model and demonstrated that it is particularly useful in downstream analysis. However, as Zhao et al. (2013)'s work utilizes single-case deletion method, it may suffer from masking effects, and the power rate of detecting the problematic points may substantially decrease when the data set consists multiple influential points, (see, Hawkins 1980; Barnett and Lewis 1984). Zhao et al. (2016) further studied the problem of multiple influential point detection in high dimensional spaces by a new group deletion procedure referred to as MIP, and they introduced two novel quantities named Max and Min statistics to overcome the masking effect and the swamping effect, respectively. However, the MIP method can not attain the targeted Type I errors and thus pays too much price on reducing swamping effect, then it requires a large amount of computation, and as a result its power rate would be largely compromised.

In this paper, with the aim to identify the true influential points as accurately as possible, we proposed an adaptive approach to detecting multiple influential points that suffer from masking effects in high-dimensional regression. Firstly, we sample many subsets from the original data set, so that, after the multiple-case deletion, these influential points in certain subsets that were masked in the original data set may be detected as influential. In a similar sprit to the single-case deletion idea of the Cook's distance, we defined a measure based on the marginal correlations (or distance correlations) with and without a fixed observation among the sampled subsets. For each observation, we calculated and obtained the maximum values of the influence measure among the sampled subsets, studied the asymptotic distribution of the influence measure and developed a

cutoff value to judge the influence of a given observation. By this procedure, we could detect and remove those problematical observations, and obtain a reliable non-influential observations subset. In the second step, to enhance efficiency, we refined the detection rule to determine whether the deleted observations are really influential or not.

The rest of this paper is organized as follows. In Section 2, we develop a multiple-case deletion approach and propose a detection algorithm for detecting multiple influential observations in high-dimensional regression model. Furthermore, a confirmatory procedure is proposed to augment the multiple-case deletion algorithm. In Section 3, we examine the performance of the procedure via Monte Carlo simulation studies. In Section 4, a real-data example is analyzed by the proposed approach. Section 5 concludes the paper.

2 Methods and Properties

2.1 Problems and existing works

Suppose we have a data set that contains a total of n observations, and there are n^* ($n^* < [n/2]$) influential observations, and $(n - n^*)$ normal observations, that is, we assume the n observations are collected from the following model

$$\begin{cases} y_i = f_0(\mathrm{x}_i) + \varepsilon_i, & \text{if} \quad i \in \mathcal{N} \setminus \mathcal{N}^*, \\ \tilde{y}_l = f_l(\tilde{\mathrm{x}}_l) + \tilde{\varepsilon}_l, & \text{if} \quad l \in \mathcal{N}^*, \end{cases} \tag{1}$$

where $\mathcal{N}^*$ ($|\mathcal{N}^*| = n^*$) is the influential points index set (a subset of $\mathcal{N} = \{1, \ldots, n\}$), $f_0(\cdot)$ and $f_l(\cdot)$, $l \in \mathcal{N}^*$ are two unknown regression functions. That is, $\{(y_i, \mathrm{x}_i),\ i \in \mathcal{N} \setminus \mathcal{N}^*\}$ denotes a data set which consist of $n - n^*$ normal observations, and $\{(\tilde{y}_l, \tilde{\mathrm{x}}_l),\ l \in \mathcal{N}^*\}$ denotes a data set which consist of n^* influential observations. Any observation with $f_i(\cdot) \neq f_0(\cdot)$ is considered as influential. Here, y_i is the i-th response variable, $\mathrm{x}_i = (x_{i1}, \ldots, x_{ip})$ denotes the i-th explanatory variable with p-dimensional vectors, analogously, $\tilde{y}_l$ and $\tilde{\mathrm{x}}_l = (\tilde{x}_{l1}, \ldots, \tilde{x}_{lp})$ denote the l-th response and explanatory variable of the influential points, the error term ε_is or $\tilde{\varepsilon}_l$s are independently and identically distributed (i.i.d.) as $N(0, \sigma^2)$.

Thus, when building a regression model as (1), the data set $\{(y_i, \mathrm{x}_i), i \in \mathcal{N}\}$ would probably be contaminated by an individual or multiple influential observations $\{(\tilde{y}_l, \tilde{\mathrm{x}}_l),\ l \in \mathcal{N}^*\}$. These influential points are usually modeled by an abnormal change occurred as the following three forms: (i), in the response variable; (ii), in the covariates; (iii), or in both the response and the explanatory variables. In this paper, we considered these three perturbation models only for generating influential points. Beyond that, it is necessary to assume that only a small proportion (far less than 50%) of observations are influential, which is reasonable in practice.

In particular, if the relationship between the response and covariates is linear, the model can be expressed as follows

$$\begin{cases} y_i = \mathrm{x}_i\beta + \varepsilon_i, & \text{if} \quad i \in \mathcal{N} \setminus \mathcal{N}^*, \\ \tilde{y}_l = \tilde{\mathrm{x}}_l\tilde{\beta} + \tilde{\varepsilon}_l, & \text{if} \quad l \in \mathcal{N}^*, \end{cases} \tag{2}$$

where $\mathcal{N}^*$, $\mathcal{N}$, $\{(y_i, \mathrm{x}_i),\ i \in \mathcal{N} \setminus \mathcal{N}^*\}$, $\{(\tilde{y}_l, \tilde{\mathrm{x}}_l),\ l \in \mathcal{N}^*\}$, ε_is and $\tilde{\varepsilon}_l$s have the same definitions as those in model (1). Here, $\beta = (\beta_1, \ldots, \beta_p)$ and $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^\top$ are two different p-vector regression coefficients. In low dimension setting, the ordinary least

square (OLS) estimate of regression coefficient was chosen as the feature with substantial change to define influential observations. Roberts et al. (2015) proposed the ADAP to detect multiple influential points with the masking effects in regression. To measure the influence of an observation K in the data set $\{(y_i, \mathrm{x}_i), i \in \mathcal{N}\}$, they utilized a "Cook's distance"-like statistic

$$C_K^I = \frac{(\hat{\beta}_{\{I,K\}} - \hat{\beta}_{\{I\}})^\top \mathrm{X}^\top \mathrm{X} (\hat{\beta}_{\{I,K\}} - \hat{\beta}_{\{I\}})}{ps^2}, \tag{3}$$

where $I = \{i_1, \ldots, i_h\}$ denotes an index set with size h to be deleted for a particular subset, X denotes an $n \times p$ design matrix with $\mathrm{X} = (\mathrm{x}_1, \ldots, \mathrm{x}_n)^\top$, $\hat{\beta}_{\{I\}}$ is the OLS estimator of coefficients based on the $n - h$ retained observations, $\hat{\beta}_{\{I,K\}}$ is the OLS estimate of coefficients when an observation K from the retained $n - h$ observations is further deleted, and s^2 is the sample estimate of σ^2 obtained from model (1) by use of all the n observations. They repeated M times to delete h observations from the whole data set, there being $n - h$ values of C_K^I for a given subset with size h, then calculated these corresponding $n - h$ values of C_K^I respectively. On the other hand, for a fixed observation K, there would be M_K ($M_K \leq M$) values of C_K^I produced by the procedure, and they chose the maximum values of C_K^I. The observation K is flagged as influential if $\max\{C_K^I\} > 1$, where the cutoff value 1 is used in their paper.

In high-dimensional settings, where the dimension p is larger than the sample size n, the ADAP method may fail to detect multiple influential observations in linear regression model, as the gram matrix is not invertible, and the OLS estimate of the coefficient is unstable in high dimensional linear regression model, hence we can not get the "Cook's distance"-like statistic. On the other hand, we are not sure whether the data comes from a high-dimensional linear model, then, we we should consider influence detection in more general models. As a result, we consider using the marginal Pearson correlations or distance correlations between the response and the predictors to define the influential observations, and an observation is flagged as influential if the corresponding correlation is noticeably changed after this observation is removed in high dimensional linear regression model. Then, in the following sections, we will give a more detailed description of these correlation-based methods.

2.2 Pearson correlation based methodology for linear model

Here we follow the idea of ADAP in Roberts et al. (2015), consider the problem of detecting multiple influential observations with masking effects in high-dimension regression model, and propose a multiple influence detection procedure. The improvement is that we choose the marginal Pearson correlation estimate instead of the OLS estimate of the coefficients, and it can deal with high dimensional problem effectively.

Firstly, we delete h observations randomly from the original n observations. Let $I = \{i_1, \ldots, i_h\}$ be the set of the indices of the deleting subset, and $R = \mathcal{N} \backslash I$ be the set of the indices of the remaining subset. Then $\mathrm{X}_{\{R\}}$ is the corresponding "after-deletion" design matrix, and $\mathrm{y}_{\{R\}}$ is the "after-deletion" response variable. The correlations between the response $\mathrm{y}_{\{R\}}$ and the j-th predictor variables $X_{j\{R\}}$ based on the $(n - h)$ retained observations are denoted by

$$\rho_{j\{R\}} = \mathrm{Cor}(X_{j\{R\}}, \mathrm{y}_{\{R\}}), \quad \text{for} \quad j = 1, \ldots, p.$$

Then, we further delete an observation K from the remaining data set $\{(y_i, \mathrm{x}_i), i \in R\}$, the corresponding "after-deletion" response is $\mathrm{y}_{\{R,-K\}}$, and design matrix is denoted as $\mathrm{X}_{\{R,-K\}}$. The correlations between $\mathrm{y}_{\{R,-K\}}$ and the j-th predictor variables $X_{j\{R,-K\}}$ based on the $(n-h-1)$ retained observations are denoted as

$$\rho_{j\{R,-K\}} = \mathrm{Cor}(X_{j\{R,-K\}}, \mathrm{y}_{\{R,-K\}}), \quad \text{for} \quad j = 1, \ldots, p.$$

Denote

$$\hat{\rho}_{j\{R\}} = \sum_{i\in R}(x_{ij} - \bar{X}_j)(y_i - \bar{Y}) / \sqrt{\sum_{i\in R}(x_{ij} - \bar{X}_j)^2 \sum_{i\in R}(y_i - \bar{Y})^2}$$

as the consistent estimate of $\rho_{j\{R\}}$, with $\bar{X}_j = \frac{1}{n-h}\sum_{i\in R} x_{ij}$, $\bar{Y} = \frac{1}{n-h}\sum_{i\in R} y_i$. Similarly, denote

$$\hat{\rho}_{j\{R,-K\}} = \sum_{i\in\{R,-K\}}(x_{ij} - \bar{X}_j)(y_i - \bar{Y}) / \sqrt{\sum_{i\in\{R,-K\}}(x_{ij} - \bar{X}_j)^2 \sum_{i\in\{R,-K\}}(y_i - \bar{Y})^2},$$

as the consistent estimate of $\rho_{j\{R,-K\}}$, with $\bar{X}_j = \frac{1}{n-h-1}\sum_{i\in\{R,-K\}} x_{ij}$, $\bar{Y} = \frac{1}{n-h-1}\sum_{i\in\{R,-K\}} y_i$.

Finally, we determine the influential of the observation K by the discrepancy between the two correlation estimates, and propose the following "HIM"-like statistic to measure the influence of an observation K in the retained data,

$$\mathcal{C}_{\{R,-K\}} = \frac{1}{p}\sum_{j=1}^{p}(\hat{\rho}_{j\{R\}} - \hat{\rho}_{j\{R,-K\}})^2. \tag{4}$$

Similar to the idea of ADAP, when there are multiple influential observations in the data set, a true influential observation K may suffer from masking effect by other influential ones. Therefore, we repeat deleting h observations from the whole data set M times, and the majority of the observations that mask observation K are deleted, then the observation K may become unmasked. For a given "after-deletion" data set $\{(\mathrm{y}_i, \mathrm{x}_i), i \in R\}$, there will be $n-h$ values of $\mathcal{C}_{\{R,-K\}}$, calculating these corresponding $n-h$ values of $\mathcal{C}_{\{R,-K\}}$ respectively. For any observation $K \in \mathcal{N}$, there will be M_K $(M_K \leq M)$ values of C_K^I produced by the procedure. As a result, we consider utilizing the maximum value of $\mathcal{C}_{\{R,-K\}}$ over all subsets to determine whether an observation K is influential when other observations are deleted, because if the observation K is a sole influential point in the entire data set, the corresponding value of $\mathcal{C}_{\{R,-K\}}$ is expected to be large.

2.3 The multiple-case deletion algorithm

The influential points detection rule can be formulated as the problem of hypothesis testing with the null hypothesis

$$H_{0i} : (y_i, \mathrm{x}_i)\text{is not an influential point for}i = 1, \ldots, n.$$

After calculating the maximum value of $\mathcal{C}_{\{R,-K\}}$ for each observation, we should find a cut-off value to determine whether an observation is influential or not. Therefore, it is possible to determine the asymptotic distribution of $\mathcal{C}_{\{R,-K\}}$ under the null hypothesis.

Here, we suppose that all the observed data $\{(y_i, \mathrm{x}_i)_{i=1}^n\}$ are generated from model (2), and under the assumptions (C1)-(C3) as those in Zhao et al. (2013), we derive a similar conclusion.

Proposition 1 Under assumptions (C1)-(C3), when there are no influential observations, and $\min(n-h,\, p) \to \infty$, then

$$(n-h)^2 \mathcal{C}_{\{R,-K\}} \to \chi^2(1),$$

where $\chi^2(1)$ is the chi-square distribution with one degree of freedom. Here, the assumption (C3) is a normality assumption on X and the error term are mainly for convenience, and can be relaxed to the distributions with sub-Gaussian tails, but it will require more lengthy proofs.

In consideration of the asymptotic distribution of $\mathcal{C}_{\{R,-K\}}$ as both $n-h$ and p go to infinity, we are ready to develop a cut-off value to determine whether an observation is unusual. For any observation K, if the maximum value of $\mathcal{C}_{\{R,-K\}}$ is greater than a certain predetermined critical value, we deem K is an influential point. That is, for a given significance level α, the i-th observation K is identified as influential if

$$\max\{\mathcal{C}_{\{R,-K\}}\} > \chi^2_{1-\alpha}(1)/(n-h)^2,$$

where $\chi^2_{1-\alpha}(1)$ is the upper α-th quantile of $\chi^2(1)$.

Under these settings, we give the following algorithm for detecting multiple influential points in high dimensional regression model.

Algorithm 1: Multiple-Case Deletion Algorithm (MDA in short).

1. Delete h observations without replacement from the original data set $\{(y_i, \mathrm{x}_i), i \in \mathcal{N}\}$, and obtain the remaining indices subset R.
2. Calculate the correlations estimate $\hat{\rho}_{j\{R\}}$ and $\hat{\rho}_{j\{R,-K\}}$ respectively, and calculate the values of $\mathcal{C}_{\{R,-K\}}$ for each observation $K \in R$.
3. Repeat Steps 1 and 2 M times, and M_K ($\leq M$) values of $\mathcal{C}_{\{R,-K\}}$ are obtained for each observation $K \in \mathcal{N}$.
4. Calculate the maximum of the M_K values of $\mathcal{C}_{\{R,-K\}}$ for each $K \in \mathcal{N}$.
5. The observation K is flagged as influential if $(n-h)^2 \max\{\mathcal{C}_{\{R,-K\}}\} > \chi^2_{1-\alpha}(1)$.

After running Steps 1-5 of the MDA algorithm, we can obtain a relatively reliable noninfluential indices subset C and a suspicious subset $D = \mathcal{N} \backslash C$. To check a particular observation $K \in D$ is really influential or not, we can test the influential of K relative to the clean subset C. Concretely, we refine the detection procedure by adding the observation K back into the noninfluential subset C and calculate the "add-back" HIM-like statistics:

$$\mathcal{C}_{\{C,+K\}} = \frac{1}{p} \sum_{j=1}^{p} (\hat{\rho}_{j\{C\}} - \hat{\rho}_{j\{C,+K\}})^2, \tag{5}$$

where $\hat{\rho}_{j\{C\}}$ is the correlation coefficients estimate based on the observations in set C only, and $\hat{\rho}_{j\{C,+K\}}$ is the correlation coefficient estimate when we add the observation K back into the clean set C.

Similar to Proposition 1, under certain conditions, when there are no influential observations, and $\min\{c+1,\, p\} \to \infty$, we have

$$(c+1)^2 \mathcal{C}_{\{C,+K\}} \to \chi^2(1),$$

where $c = |C|$ is the cardinality of the clean set C.

As a result, we can use $\chi^2_{1-\delta}(1)$ as the cut-off value to determine whether K is a really influential point. Here, δ is appropriately chosen, such as $\delta = \alpha/2$. The refined procedure augments MDA as follows:

Algorithm 2: Refined algorithm

6. Based on the result of MDA, create a clean set C and suspicious set D.

7. Calculate the correlations estimate $\hat{\rho}_{j\{C\}}$ and $\hat{\rho}_{j\{C,+K\}}$ based on the current clean data set C, and calculate $\mathcal{C}_{\{C,+K\}}$ for each of the $n-c$ observations in the suspicious data set D.

8. For a given level δ, if $c^2 \mathcal{C}_{\{C,+K\}} > \chi^2_{1-\delta}(1)$, observation K remains flagged as suspicious, otherwise observation K is added back to the clean set.

As yet, we have completed our multiple influential observation detection procedure. It is worth mentioning that the Steps 6 to 8 can be iterated many times until no further observations are flagged as influential. Considering the computational cost, we only apply a single iteration of Steps 6 to 8 in this paper. Combining Algorithms 1 and 2, we denote the refined MDA procedure as R-MDA in this paper.

2.4 Distance correlation based methodology for complex models

In practice, we are not sure the data comes from which type of model in regression analysis, in other words, specifying a correct linear model as (2) for high-dimensional data may be challenging. Therefore, we extend the proposed influence detection procedure to a broad of more complex underlying models, that is, suppose the observations (y_i, x_i) are collected from the general regression model (1). As the function $f_0(\cdot)$ may denote nonlinear relationship, the influential measure based on Pearson correlation coefficient is inappropriate. On the other hand, we recognize that distance correlation (Székely et al. 2007) is an appropriate alternative tool to metric the statistical dependence between two variables. Li et al. (2012) proposed a sure independence screening procedure based on the distance correlation, (DC-SIS, for short), and the sure screening property is valid for the DC-SIS under more general model settings, as it does not require model specification for response or predictors.

Motivated by this property, under more general model settings, we can measure the influence of the i-th observation by defining a DC-based influence statistic,

$$\mathcal{C}^*_{\{R,-i\}} = \frac{1}{p} \sum_{j=1}^{p} (\widehat{\mathrm{dcor}}_{j\{R\}} - \widehat{\mathrm{dcor}}_{j\{R,-i\}})^2, \tag{6}$$

which is expected to be more effective than the Pearson correlation-based method. Here, $\widehat{\mathrm{dcor}}_{j\{R\}}$ denotes the sample estimate of the distance correlation between the response

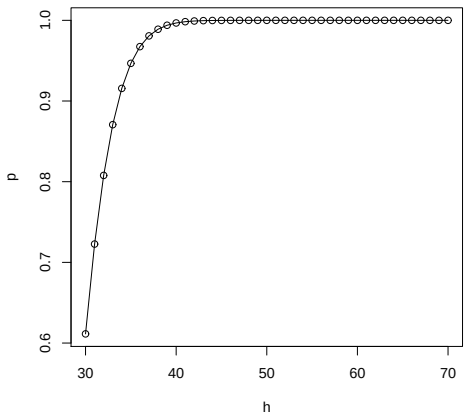

Figure 1: The scatter plot of the right side of inequality (8) and h.

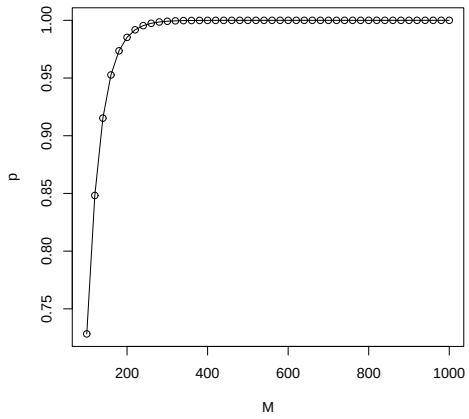

Figure 2: The scatter plot of the right side of inequality (8) and M.

and the j-th predictor. As the distribution of the statistic $\mathcal{C}^*_{\{R,-i\}}$ is still unknown in the literature and so complicated that one can sort the values of $\{\max(\mathcal{C}^*_{\{R,-i\}}), i = 1, \ldots, n\}$, and remove those observations associated with large values of $\max(\mathcal{C}^*_{\{R,-i\}})$ in practice.

The detailed derivation of the asymptotic distribution of the influential statistics $\mathcal{C}^*_{\{R,-i\}}$, however, is beyond the scope of this paper, we will focus on these exciting topics in further study. Of course, as an alternative, we can use bootstrap method to find the upper α-th quantile $F_{1-\alpha}$ of the cumulative distribution function (CDF) as the cut-off value. The observation K is flagged as influential if $(n-h)^2\max\{\mathcal{C}_{\{R,-K\}}\} > F_{1-\alpha}$. Note in practice, this bootstrap approach will cost much calculation.

2.5 The choice of M and h

It should be noted that, for the multiple deletion algorithm in Subsection 2.3, reasonable values of M and h should be considered. Assume there are n^* influential observations in the original n observations, and let A denote the event that a particular influential observation is isolated from the other $n^* - 1$ influential observations, then the probability is

$$P(A) = 1 - (1 - C^{(h-n^*+1)}_{(n-n^*)}/C^h_n)^M. \tag{7}$$

Let event B denote all these n^* influential observations being isolated from the others at least once. The probability of event B is very complicated to calculate, and we utilize the Bonferroni bound to obtain the approximate results, that is

$$P(B) \geq n^* \times (1 - (1 - C^{(h-n^*+1)}_{(n-n^*)}/C^h_n)^M) - n^* + 1. \tag{8}$$

As we can see, the right side of inequality (8) may be negative, but it provides an useful guidance as with an appropriate choice of M and h, it tends to 1. For example, when $n = 100$, $n^* = 5$, if we fix $M = 500$, as h increases, the right side of inequality (8)

tends to 1. On the other hand, if we fix $h = 49$, as M increases to 500, the right side of inequality (8) tends to 1. See Figures 1 and 2.

Then, here we just give a guidance to choose the number M and h to gain a large probability. However, we should not only consider the probability of a particular influential observation being isolated from the other $n^* - 1$ ones, as the original objective in this paper is to detect all the influential points in high dimensional regression model. As M and h increase, the chance of missing truly influential observations decreases, then we can detect mostly of true influential points. But also, the chance of incorrectly detecting a normal observation increases as the increase of M and h, so we supplement a refined step to augment the multiple-case deletion algorithm.

3 Simulation Studies

In this section, we conduct simulation studies to assess the effectiveness of R-MDA method, and compare its performance with some competitive methods, including the HIM in Zhao et al. (2013), ADAP in Roberts et al. (2015), and MIP in Zhao et al. (2016). To make fair comparisons, we follow similar model settings as those in Zhao et al. (2013). All the simulation studies were conducted by using Matlab codes.

3.1 Simulation models

In this subsection, we give two simulation examples to verify that the proposed method is reasonable.

Example 1: (High-dimensional linear model). The simulated data were generated from a "true" model which comprised a response and p explanatory variables, the relationship is formulated as

$$y_i = \mathrm{x}_i\beta + \varepsilon_i, \text{ for } i = 1, \ldots, n, \tag{9}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a p-vector of regression coefficient. We set $\beta_j = 1, (1 \le j \le 5)$ and $\beta_j = 0, (j > 5)$. The random error term ε_i is independent of the predictors, and is generated from three different distributions: the standard normal distribution, the exponential distribution with rate 0.1 and the standard t distribution with three degrees of freedom. The covariates $\mathrm{x}_i = (x_{i1}, \ldots, x_{ip})$ are generated from a multivariate normal distribution $N_p(0, \Sigma)$ with entries of $\Sigma = (\rho^{|j-l|})_{p\times p}$ for $j,\ l = 1, \ldots, p$ and $\rho = 0, 0.5, 0.9$ are chosen in the simulation study.

After generating data via model (9) in various cases mentioned above, we then remove n^* ones from the original n observations, and replaced with n^* masked influential observations $\{(\tilde{y}_l, \tilde{\mathrm{x}}_l), l \in \mathcal{N}^*\}$ from the following model

$$\tilde{y}_l = \tilde{\mathrm{x}}_l\tilde{\beta} + \varepsilon_l, \text{ for } l = 1, \ldots, n^*, \tag{10}$$

where $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^\top$ is a p-vector of regression coefficient, and ε_l has the same distribution as that in model (9). In particular, we consider the n^* influential points are generated as one of the following three cases.

Model 1: Response perturbation model. Let $\tilde{\mathrm{x}}_l = \mathrm{x}_l$, for $l = 1, \ldots, n^*$, and $\tilde{\beta}_j = (1, 1, 1, 1, 1, \kappa, \ldots, \kappa)^\top$, that is $\tilde{\beta}_j = 1, (1 \le j \le 5)$ and $\tilde{\beta}_j = \kappa, (5 < j \le p)$, κ is the

perturbation parameter that indicates the magnitude of the influence. In this case, the perturbation can be written as $\tilde{y}_l = \mathrm{x}_l\beta + \kappa \mathrm{x}_l\gamma + \varepsilon_l$, and $\gamma = (0,0,0,0,0,1,\ldots,1)^\top$, that is, the responses are contaminated by a random perturbation term $\kappa \mathrm{x}_l\gamma$.

Model 2: Explanatory variable perturbation model. Let $\tilde{y}_l = y_l$ remain unchanged, $\tilde{x}_{lj} = x_{lj} + 30\kappa I_{\{j\in S\}}$, for $l = 1,\ldots,n^*$, and $S(\subset \{1,\ldots,p\})$ is an index subset of predictors. In this case, the influence was occurred on the predictors and keep the response invariant.

Model 3: Perturbation in both the response and explanatory variables. Let the regression coefficient $\tilde{\beta} = (1,1,1,1,1,\kappa,\ldots,\kappa)^\top$, and $\tilde{x}_{lj} = x_{lj} + 30\kappa I_{\{j\in S\}}$, for $l = 1,\ldots,n^*$. Therefore, it is a combination of the above two perturbation models.

Example 2: (Nonparametric additive model) Consider the following nonlinear model

$$y_i = 2f_1(x_{i1}) + 6f_2(x_{i2}) + 4f_3(x_{i3}) + f_4(x_{i4}) + \varepsilon_i, \tag{11}$$

where $f_1(x) = x$, $f_2(x) = (2x-1)^2$, $f_3(x) = \sin(2\pi x)/(2-\sin(2\pi x))$, $f_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3$. Suppose all the observations (y_i, x_i) are generated from this model, and then we reset the first n^* observations generated from another perturbation model, e.g. Model 2 referred in Example 1.

The placement of the n^* inserted observations was such that the HIM-like value for each of the n^* observations is less than $\chi^2_{1-\alpha}(1)/n^2$ based on all n observations, but when we deleted the other n^*-1 inserted observations, the HIM-like value is greater than $\chi^2_{1-\alpha}(1)/n^2$. As a result, the data set contains $n-n^*$ normal observations and n^* influential observations. In this experiment, we set the perturbation parameter $\kappa = 0, 0.4, 0.8, 1.2, 1.6$, and $S = \{1,\ldots,[p/2]\}$. We choose $n = 100, 200, 500$, $p = 10, 100, 500, 1000$ with various values of $n^* = 5, 10, 15, 20$ in our simulation study.

3.2 Performance comparisons

In consider of all possible values of n, p, n^*, κ, h and M, it is instructive to evaluate the average values of type-I error rate and power rate. Suppose among the $n-n^*$ non-influential observations, n_1 observations are incorrectly identified as influential, and among the n^* true influential observations, n_2 observations are correctly identified. Then, the type-I error rate is defined as $n_1/(n-n^*)$, i.e., the proportion of normal observations that are incorrectly classified as influential. The power rate is defined as n_2/n^*, i.e., the proportion of contaminated observations that are correctly labeled as influential ones. For clear comparisons, we also list the standard deviations of the type-I rates and power rates in parentheses. The nominal significant level α is chosen to be 0.01, 0.05 or 0.1. As for Example 2, we perform similar algorithms as MDA proposed in Section 2 and compute the DC-based influential measure $\mathcal{C}^*_{\{R,-i\}}$. Since the asymptotic distribution of $\mathcal{C}^*_{\{R,-i\}}$ is complicated, we flag the n^* observations with the largest values of $\max(\mathcal{C}^*_{\{R,-i\}})$ as influential. The empirical power and type-I error rate values of the simulation study are presented in Tables 1-8. All the results of Example 1 and 2 are obtained with 500 replications.

Table 1 reports the average empirical type I error values as well as the standard deviations by our proposed R-MDA procedure on simulated data for various combinations of n^*, p, and α in three different models of Example 1. In this simulation study, we set $h = [n/2]-1$, $M = 500$ and the perturbation parameters $\kappa = 1.6$. It is easy to observe

Table 1: Average percentage of type I errors (size) values (%) of R-MDA under various values of n^*, p and α when $n = 100$, $\kappa = 1.6$; The standard deviations (%) are given in parentheses.

		n^*=5			n^*=10		
Model	p	α=1%	α=5%	α=10%	α=1%	α=5%	α=10%
(1)	10	0.8(0.3)	5.4(2.1)	9.7(3.0)	0.9(0.7)	4.8(2.1)	10.3(3.0)
	100	0.9(0.5)	5.3(2.2)	11.8(2.9)	0.7(0.8)	5.3(2.2)	9.6(3.2)
	500	0.9(0.6)	6.1(2.7)	12.2(3.0)	0.7(0.6)	4.4(2.1)	9.7(2.9)
	1000	0.9(0.6)	5.8(2.6)	13.0(2.6)	0.7(0.5)	3.8(1.8)	9.5(3.3)
(2)	10	0.7(0.5)	4.8(2.3)	8.2(3.2)	0.8(0.6)	4.6(2.1)	9.4(3.4)
	100	0.9(0.4)	5.7(2.3)	12.9(3.1)	0.6(0.5)	3.8(2.3)	9.4(3.3)
	500	0.9(0.3)	6.1(2.7)	13.0(2.6)	0.8(0.6)	3.9(2.7)	10.2(3.3)
	1000	0.9(0.3)	6.3(2.6)	13.2(3.1)	0.8(0.7)	3.9(2.4)	9.8(3.4)
(3)	10	0.8(0.4)	4.2(2.3)	8.8(3.1)	0.7(0.8)	4.9(2.0)	9.0(3.1)
	100	0.8(0.4)	4.6(2.4)	9.0(3.1)	0.7(0.5)	4.9(1.5)	8.4(2.1)
	500	0.9(0.3)	4.2(2.7)	8.7(2.9)	0.5(0.4)	4.6(1.2)	8.0(2.1)
	1000	0.9(0.3)	4.0(2.8)	9.2(3.1)	0.5(0.4)	3.7(1.2)	7.0(2.3)

Table 2: Performance comparison of HIM and MIP, MDA and R-MDA; Average percentage of type I errors and power values (%) for various values of p when $n = 200$, $n^* = 10$, $\alpha = 0.05$, $\kappa = 1.2$; The standard deviations (%) are given in parentheses.

		HIM		MIP		MDA		R-MDA	
Model	p	size	power	size	power	size	power	size	power
	10	0.2(0.2)	16.5(10.4)	2.0(1.1)	34.1(11.9)	9.6(2.7)	38.8(12.8)	4.7(2.2)	37.9(12.1)
(1)	100	0.0(0.0)	46.5(12.8)	1.0(1.2)	77.5(14.2)	10.2(2.5)	78.8(12.3)	5.1(2.2)	76.6(11.1)
	500	0.0(0.0)	52.4(14.4)	0.8(0.5)	89.7(6.8)	10.2(2.3)	93.8(8.7)	5.2(1.9)	92.7(8.6)
	10	0.2(0.2)	36.5(16.0)	1.9(1.8)	63.1(17.1)	6.2(2.5)	83.9(10.8)	3.4(2.1)	81.6(16.3)
(2)	100	0.0(0.0)	39.4(14,6)	1.0(0.6)	78.5(22.0)	8.8(2.0)	95.2(9.7)	4.5(2.1)	93.6(16.4)
	500	0.0(0.0)	37.1(15.1)	0.8(0.7)	83.3(18.9)	9.4(2.1)	97.9(10.2)	4.5(2.1)	96.1(17.2)
	10	0.1(0.1)	63.5(14.5)	1.6(1.3)	86.0(21.1)	8.9(2.0)	96.8(1.9)	4.9(2.0)	96.2(3.0)
(3)	100	0.0(0.0)	82.4(11.3)	0.6(0.5)	100.0(0.0)	9.0(1.6)	100.0(0.0)	5.4(1.5)	100.0(0.0)
	500	0.0(0.0)	85.8(10.7)	0.4(0.3)	100.0(0.0)	9.2(1.4)	100.0(0.0)	5.3(1.2)	100.0(0.0)

Table 3: Performance comparison of different methods in terms of type I errors and power values (%) for various perturbation parameter κ when $n = 100$, $n^* = 10$, $p = 1000$, $\alpha = 0.05$.

		HIM		MIP		MDA		R-MDA	
Model	κ	size	power	size	power	size	power	size	power
	0.0	0.2	-	3.0	-	7.4	-	5.3	-
	0.4	0.0	36.5	0.9	67.1	7.6	84.8	5.3	83.9
(1)	0.8	0.0	39.5	1.0	81.5	6.2	90.0	4.7	87.6
	1.2	0.0	44.4	0.9	88.7	8.2	87.8	6.2	89.7
	1.6	0.0	49.5	0.6	91.8	6.5	93.0	4.8	92.9
	0.0	0.2	-	3.9	-	11.5	-	6.8	-
	0.4	0.0	17.5	1.0	73.1	6.2	83.9	3.4	79.6
(2)	0.8	0.0	19.7	0.9	84.5	7.1	89.2	3.5	86.4
	1.2	0.0	20.0	0.9	85.3	7.4	90.9	3.5	88.1
	1.6	0.0	21.7	0.8	84.4	6.9	90.5	3.9	90.1
	0.0	0.3	-	3.1	-	12.4	-	6.4	-
	0.4	0.0	69.5	0.6	98.0	3.9	99.8	3.1	99.2
(3)	0.8	0.0	71.4	0.9	100.0	4.5	100.0	3.6	100.0
	1.2	0.0	73.8	0.8	100.0	3.6	100.0	3.4	100.0
	1.6	0.0	76.7	0.8	100.0	3.9	100.0	3.1	100.0

that the realized empirical type I errors of R-MDA are nearly consistent to the nominal ones for these models considered.

Table 2 reports the average empirical type I error values, power values, and the standard deviations of applying various methods for detecting $n^* = 10$ influential points. Here, we set $h = [n/2] - 1$, $M = 1000$, the perturbation parameters $\kappa = 1.2$, and $p = 10, 100, 500$. Note that, the realized type I errors of R-MDA are near to the nominal one 5% in most cases. In contrast, the type I errors of HIM tends to be extremely small, therefore, the ability of detecting influential observations would be compromised. Note the HIM method utilizes single-case deletion, it may suffer from masking effects, and the power rate of detecting the problematic points may substantially decrease when the data set consists multiple influential points. As the MIP method introduced two novel quantities named Max and Min statistics to overcome the masking effect and the swamping effect, respectively, we can see that the MIP method can not attain the targeted Type I errors and thus pays too much price on reducing swamping effect, and as a result its power rate would be largely compromised. However, masking phenomenon is usually more serious than swamping phenomenon in the influential observation detection problems, as the former can cause gross distortions, whereas the latter is often just a matter of lost efficiency (Zou, et al. 2014). In most cases, the proposed R-MDA method approximately achieves the designed type-I error rate (thus swamping effect is not serious) and has a much better power rate values (thus alleviates masking effect) in most cases.

Table 3 summarizes the results of type I error and power rate values by use of HIM, MIP, MDA and R-MDA approaches respectively. Note that in Table 3, when $\kappa = 0$, the calculation of power value is not applicable ("-") because there are no influential

Table 4: Average percentage of type I errors and power values (%) of different methods under various of n, p and n^* when $\kappa = 1.6$, $\alpha = 0.05$; The simulation data are generated from Model 2 of Example 1.

		ADAP		HIM		MIP		R-MDA	
$(n,\ p)$	n^*	size	power	size	power	size	power	size	power
$(100, 10)$	2	0.0	100.0	1.8	69.8	2.0	100.0	4.3	100.0
$(100, 10)$	5	0.0	100.0	0.9	51.1	3.2	100.0	4.4	100.0
$(100, 10)$	10	0.0	50.0	0.5	29.1	1.8	78.0	3.7	84.6
$(100, 10)$	20	0.0	10.6	0.2	17.7	1.2	37.8	2.9	69.7
$(100, 15)$	2	0.0	100.0	0.7	66.7	1.5	100.0	4.8	100.0
$(100, 15)$	5	0.0	100.0	0.6	39.1	3.7	100.0	4.2	100.0
$(100, 15)$	10	0.0	58.4	0.2	27.5	2.2	63.8	3.4	82.7
$(100, 15)$	20	0.0	7.8	0.1	12.1	2.6	58.8	3.3	73.8
$(100, 20)$	2	0.0	100.0	0.8	75.1	2.5	100.0	4.8	100.0
$(100, 20)$	5	0.0	100.0	0.0	57.14	2.2	100.0	4.2	100.0
$(100, 20)$	10	0.0	20.6	0.0	20.8	2.0	76.4	3.9	82.6
$(100, 20)$	20	0.0	0.0	0.0	11.9	1.9	49.4	3.7	67.6
$(200, 50)$	2	0.0	100.0	0.1	99.0	0.4	100.0	5.5	100.0
$(200, 50)$	5	0.0	100.0	0.0	67.1	2.6	100.0	4.7	100.0
$(200, 50)$	10	0.0	25.5	0.0	44.7	3.4	87.5	4.3	91.6
$(200, 50)$	20	0.0	10.0	0.0	26.8	1.4	67.8	3.3	78.7
$(200, 500)$	2	0.0	0.0	0.0	79.3	0.9	100.0	5.9	100.0
$(200, 500)$	5	0.0	0.0	0.0	59.1	1.2	100.0	4.8	100.0
$(200, 500)$	10	0.0	0.0	0.0	43.8	3.2	77.9	3.1	94.2
$(200, 500)$	20	0.0	0.0	0.0	26.7	1.5	68.4	2.7	80.9
$(200, 500)$	30	0.0	0.0	0.0	18.5	0.9	29.7	2.3	67.6
$(200, 500)$	40	0.0	0.0	0.0	11.1	0.7	21.4	1.7	63.3

Table 5: Average percentage of type I errors and power values (%) of R-MDA under various of when $n = 100$, $p = 1000$, $\alpha = 0.05$; The data are generated from Model 2 in Example 1.

ρ	n^*/n	κ=0.4		κ=0.8		κ=1.2		κ=1.6	
		size	power	size	power	size	power	size	power
	0.02	8.7	100.0	8.6	100.0	8.4	100.0	8.1	100.0
	0.05	6.8	100.0	6.3	100.0	6.2	100.0	5.9	100.0
0.0	0.10	4.8	87.5	3.6	90.3	3.2	79.8	3.1	88.6
	0.15	4.3	72.5	3.6	77.4	3.2	80.8	2.2	84.7
	0.20	3.6	59.5	3.0	67.1	2.3	69.0	2.1	72.8
	0.02	8.7	100.0	8.6	100.0	8.6	100.0	8.3	100.0
	0.05	6.4	100.0	6.3	100.0	6.3	100.0	6.2	100.0
0.5	0.10	5.6	80.3	5.2	87.2	4.8	89.2	4.3	90.4
	0.15	3.1	76.9	2.2	79.9	2.1	84.3	2.4	86.1
	0.20	2.7	56.7	2.1	60.9	2.2	70.1	1.4	76.2
	0.02	8.9	100.0	8.7	100.0	8.6	100.0	8.4	100.0
	0.05	6.9	100.0	6.7	100.0	6.3	100.0	6.2	100.0
0.9	0.10	4.3	81.6	4.2	83.7	3.9	85.9	3.4	88.0
	0.15	3.5	77.2	2.8	78.7	2.7	81.5	2.4	87.0
	0.20	3.4	63.2	2.9	65.4	2.1	67.0	1.6	69.4

Table 6: Average percentage of type I errors and power values (%) for various values of p when $n = 100$, $n^* = 10$, $\kappa = 1.6$, $\alpha = 0.05$; The errors are generated from three different distribution respectively.

	$\varepsilon_i \sim$	$N(0,1)$		$Exp(0.1)$		$t(3)$	
Model	p	size	power	size	power	size	power
(1)	100	5.4	78.0	5.5	78.5	6.1	77.4
	500	4.7	90.4	3.5	88.4	4.7	89.6
	1000	3.6	93.4	3.7	92.6	3.8	91.2
(2)	100	4.7	93.6	4.8	90.6	4.7	92.8
	500	4.6	93.2	4.9	90.0	5.2	96.4
	1000	4.9	96.6	5.1	92.4	5.7	97.4
(3)	100	5.1	99.2	4.9	92.8	4.7	95.8
	500	4.6	99.6	4.9	97.0	5.2	96.8
	1000	4.9	99.9	5.1	97.0	5.7	98.7

Table 7: Average percentage of type I errors and power values (%) for measuring the impact of the choice of h and M, when $n = 100$, $n^* = 10$, $p = 1000$, $\kappa = 1.6$, $\alpha = 0.05$.

		$h = 30$		$h = 40$		$h = 50$		$h = 60$	
Model	M	size	power	size	power	size	power	size	power
(1)	200	3.5	80.3	3.6	83.5	4.6	86.8	4.9	89.3
	500	3.3	86.4	4.3	90.8	5.1	93.0	6.0	93.3
	1000	3.1	88.3	3.0	92.3	2.9	93.8	2.4	96.8
(2)	200	3.0	68.3	3.9	71.0	3.1	84.8	4.5	87.3
	500	3.6	70.7	4.0	83.9	3.7	88.1	4.6	89.5
	1000	4.1	76.8	4.7	85.8	4.3	89.9	4.8	90.6

Table 8: Average percentage of type I errors and power values (%) of HIM, R-MDA and MDA-DC under various values of κ and n^*, when $n = 100$, $p = 500$, $\alpha = 0.05$.

		κ=0.4		κ=0.8		κ=1.2		κ=1.6	
Method	n^*/n	size	power	size	power	size	power	size	power
HIM	0.05	0.9	13.7	0.8	19.4	0.8	19.8	0.7	22.1
	0.10	0.9	9.5	0.9	11.1	0.9	12.0	0.9	12.6
	0.15	0.9	9.5	1.0	9.7	0.9	10.8	1.1	11.3
R-MDA	0.05	4.7	91.6	3.8	92.2	3.9	94.3	4.4	96.6
	0.10	4.1	77.6	4.2	80.3	4.8	84.9	3.6	86.8
	0.15	2.3	45.0	2.9	49.7	3.9	60.3	4.4	66.4
MDA-DC	0.05	2.4	96.3	1.1	97.9	0.0	100.0	0.0	100.0
	0.10	3.6	86.4	1.9	92.7	1.8	93.6	1.4	95.1
	0.15	3.4	81.1	3.2	84.2	2.2	85.3	1.8	92.2

points. As κ increases from 0 to 1.6, MDA and R-MDA almost can detect all the true influential points. HIM can barely detect all influential observations due to masking effect even if when $\kappa = 1.6$. MIP performs well in detecting influential observations, but compared with R-MDA, it is still not effective. Both MDA and R-MDA perform well at detecting the n^* influential observations. Although in terms of type I errors, these two approaches often erroneously flagged normal observations as influential, but compared with HIM, these two approaches have advantages of detecting all of influential points, and can control well for a given nominal significant level.

In most cases, the MDA approach can detect most of the true influential points, but there are also too many normal observations being incorrectly regarded as influential. From Tables 2-3, we observed that owing to the confirmation procedure, the R-MDA can detect most of the true influential points, and reduce the number of incorrect choices. Therefore, a confirmation step is necessary to ensure that the normal observations are not incorrectly as influential ones.

Table 4 gives the results of average type I error and power rate values by four different methods ADAP, HIM, MIP and R-MDA respectively on simulated data for various combines n, p and n^*. The simulation results show that when $p < n$, and there are few influential points in the data set, ADAP method has a good performance, this result is consistent with the conclusion of ADAP in Roberts et al. (2015). But when $p > n$, this method can not work well, as the OLS estimate of regression coefficient is not solvable. In comparison, the other three methods can be used to detect influential points in high-dimensional situations, and also in the low-dimensional case. The only downside is that the power of the HIM method is low, and this perhaps due to the masking effect. The MIP and R-MDA methods perform well to detect influential points in most cases, because they can detect influential points with higher power rate, and by contrast, the performance of R-MDA is slightly better than that of MIP.

Table 5 takes into account the effect of different correlations between the predictor variables on the results, and $\rho = 0, 0.5, 0.9$ are chosen respectively. We can observe that either the variables are independently or dependently, as the perturbation parameter κ increases from 0.4 to 1.6, and the proportion of influential points increased from 2% to 20%, the R-MDA method is effective to detect multiple influential points in most cases.

In Table 6, the simulation results show that no matter the errors are generated from which model (normal, skewed or heavy-tailed), the proposed R-MDA method can work well, as the type I errors can be well controlled near 0.05, and all the power values are close to 1. This is reasonable because when the condition (C3) of Proposition 1 is relaxed to more general cases, and the theoretical results are still feasible. Although in Table 6, we did not record the type I error values in the first step of R-MDA, the confirmation step of R-MDA is effective to control the type I error values, that is, fewer normal observations are incorrectly deemed as influential ones. But, after careful investigation, we discovered that when the errors are generated from the Exponential distribution, or the standard t distribution, the corresponding type I error values are a bit higher than those of normal error case. To our delight, we can further see that, when the dimension p increases from 100 to 1000, the power values also increased with the dimension, which once again shows that our method can be used to deal with high dimensional data.

To explore the impact of M and h in our proposed R-MDA algorithm, in Table 7, we compared the simulation results through various choices of M and h. The results are consistent with the plotted curves in Figure 1 and Figure 2. For a fixed value of M, as

h increases from $h = 30$ to $h = 70$, the power values increase, on the other hand, when we fixed the value of h, and the power values increase as M increases from 200 to 1000. Meanwhile, the type I error values increase too. As a result, an appropriate choice of M and h should be chosen, as too large or too small is inappropriate. The simulation results confirm that $h = [n/2]$ and $M \geq 1000$ is a reasonable choice.

The simulation results of Table 8 show that when κ changes from 0.4 to 1.6, and the proportion of the influential points increase from 0.05 to 0.15, the DC-based influential measure (denoted by MDA-DC for short) performs reasonably well for a complex model. By comparison, both of HIM, the single-case deletion method based on Pearson correlation, and the R-MDA, the multiple-case deletion method based on Pearson correlation can not work well when the simulated data are from a complex nonparametric additive model in Example 2. This is probably due to the fact that the distance correlation can regress the nonlinear relationship reasonably. Meanwhile, the disadvantage is that the method based DC consumes a lot of computational cost, and the theoretical work is challenging, especially in the high-dimensional setting. Therefore, the topic of detecting influential points under more general model settings deserves further study.

4 A Real Data Example

As an application illustration, we applied the proposed R-MDA approach to the NCI-60 data developed by National Cancer Institute, and found that the analysis results are substantially different when the detected influential observations are removed.

The microarray and proteomic datasets consist of data on 60 human cancer cell lines, and can be downloaded from the CellMiner program package (http://discover.nci.nih.gov/cellminer/). For the gene expression data, we used the Affymetrix HG-U133(A_ B)[a] chip (Affy) that had been normalized by the gcRMA method, resulting in a set of 43,524 predictors. For the protein expression data, 162 proteins expression values were acquired via reverse-phase protein lysate arrays and log2 transformed. As one observation named LC:NCI_H23 of the cell lines was missed in the gene expression data, 59 human cancer cell lines were used in the analysis. More details on how the data were obtained can be found in Shankavaram et al. (2007).

Similar to Alfons et al. (2013), we first order the protein expression variables according to their MAD (median absolute deviation), and show the protein expressions based on the KRT18 antibody which constitutes the variable with the largest MAD. Hence, we chose the protein KRT 18 as the response variable, which is known to be persistently expressed in carcinomas (Oshima et al., 1996). Due to the dimensionality greatly exceeds the sample size, the sparsity assumption is reasonable, that is, only a small number of predictors are relevant to the response (Fan and Lv, 2008). Then, we followed Fan and Lv (2008) to retain the top 1000 gene expression data that are mostly correlated with the protein KRT 18 in the data analysis. Thus, the resulting analysis has $p = 1000$ predictors and a sample size $n = 59$. That is, the data set consists of an response variable $\mathrm{y} = (y_1, \ldots, y_n)^\top$, and an $n \times p$ matrix of gene expression values $\mathrm{X} = (x_{ij})_{n\times p}$, for $i = 1, \ldots, n(= 59)$, $j = 1, \ldots, p(= 1000)$, where y_i is the i-th response variable of protein KRT 18, and x_{ij} denotes the expression level of the j-th gene for the i-th human cancer cell line.

Suppose that the processed gene expression data are from a high-dimensional linear regression model. Further, we believe, the $n = 59$ human cancer cell lines may be

contaminated with a few of influential points. Based on our previous simulation studies, the presence of influential points may significantly affect the accuracy of data analysis.

Then, we apply the proposed R-MDA method to the processed gene expression data, with $\alpha = 0.05$, $M = 500$ and $h = 30$. In the first step, the MDA identified 11 problematic observations, their labels are BR:BT_549, CNS:SNB_19, CNS:SNB_75, LE:CCRF_CEM, LE:HL_60, LE:MOLT_4, LE:SR, ME:MALME_3M, ME:SK_MEL_28, ME:SK_MEL_5 and LC:NCI_H460. Next, we apply the refine algorithm, and remove the observation CNS:SNB_19 back to the clean data set, therefore, we identified a total of 10 observations by the R-MDA method. Similarly, we apply the proposed MDA-DC method directly to the processed gene expression data without considering the linear relationship, and screen out the top 10 observations with larger DC-based measures as influential points. The corresponding labels are BR:BT_549, CNS:SNB_19, CNS:SNB_75, LE:CCRF_CEM, LE:MOLT_4, LE:SR, ME:MALME_3M, ME:SK_MEL_28, ME:SK_MEL_5, LC:NCI_H460. The results are consistent with the results by R-MDA except an observation CNS:SNB_19. Therefore, we may consider all the 11 problematic observations in practical applications.

As a comparison, we apply the HIM method to the processed NCI-60 data sets, and identified only 6 problematic observations, LE:CCRF_CEM, LE:MOLT_4, LE:SR, ME:MALME_3M, ME:SK_MEL_5 and LC:NCI_H460. These results show that the R-MDA method can detect more problematic observations than that of HIM method, as the multiple influential observations may be suffered from masking effects. It illustrated the single-case deletion methods may fail to detect all multiple problematic observations, and the multiple-case deletion approach is an effective method.

In addition, we apply the sparse least trimmed squares regression method (Sparse LTS in short) of Alfons et al. (2013) for analyzing the processed NCI-60 data sets, and there are 13 observations deemed as influential, which contains 11 influential points identified by MDA method above, and two other observations. Thus, we believe that the observations detected by our MDA method are not probably "false positives".

To further assess the influence of the identified observations, we compared the Lasso estimate with and without those points, and compare the results of R-MDA with that of HIM. Similar to Zhao et al. (2013), in our analysis, we used ten-fold cross-validation to select the tuning parameter and every run is random, so we repeated this analysis 100 times and report the average results.

We summarize the difference of the coefficient estimates in three aspects: the sparsity; the norm difference; and the angle between the two estimates.

Firstly, we remove the identified influential observations by the R-MDA method, the resulting Lasso estimate is considerably more sparse, the average Lasso model size with the full data is 55. Then, we remove the potential influential points identified by R-MDA, the average Lasso model size is 29. By contrast, we remove the potential influential points identified by HIM, and the average Lasso model size is 36. This shows that the existence of the potential influential points clearly shows a significant effect on the model size, and the R-MDA method is more effective than the HIM method.

Secondly, we denote $d_0 = ||\hat{\beta}_{full}||_2$ is the Lasso estimate using all the observations, $d_1 = ||\hat{\beta}_{redu}||_2$ is the estimate after removing the potential influential points identified by R-MDA, and $d_1^* = ||\hat{\beta}^*_{redu}||_2$ is the estimate after removing the potential influential points identified by HIM. To assess the difference of coefficient estimate with and without the potential influential points, we denote $d_2 = ||\hat{\beta}_{full} - \hat{\beta}_{redu}||_2$, $d_2^* = ||\hat{\beta}_{full} - \hat{\beta}^*_{redu}||_2$.

We observe that the average of $(d_0 - d_1)/d_0$ is 0.272, and that of d_2/d_0 is 0.798. By contrast, the average of $(d_0 - d_1^*)/d_0$ is 0.202, and that of d_2^*/d_0 is 0.713. Therefore, the results by the R-MDA are a little better than that of HIM, as the corresponding l_2 norm difference is slightly larger. Moreover, both methods show that the estimates with and without the potential influential points are quite different in terms of the l_2 norm.

Thirdly, to again indication how the estimates change substantially after removing the influential points, we calculate the angle between $\hat{\beta}_{full}$ and $\hat{\beta}_{redu}$, and the angle between $\hat{\beta}_{full}$ and $\hat{\beta}^*_{redu}$, which are defined as $\hat{\beta}^\top_{full}\hat{\beta}_{redu}/d_0d_1$, $\hat{\beta}^\top_{full}\hat{\beta}^*_{redu}/d_0d_1^*$ respectively. The simulated results are 0.63 and 0.66 respectively averaged over 100 times.

In summary, whether including those 10 influential observations or not may affect the results of Lasso estimate. Hence, the influential observations should be identified firstly in high dimensional data analysis, that is, leaving out those 10 observations may gain more reliable results for the majority of the cancer cell lines.

5 Conclusions

ADAP is an effective method to detect influential points in practice. Nevertheless, it cannot be applied to high-dimensional data with $p > n$. In this paper, we put forward the R-MDA method to detect multiple influential observations which suffered from masking effects in high dimensional regression model. The proposed method enjoys several appealing properties: (1) the influence measure can be calculated viably, and the calculation process is relatively simple and effective; (2) it can detect almost all multiple influence points which suffer from masking effects; (3) the method can be applied to high-dimensional regression data with relatively large values of p. Furthermore, we considered the proposed method R-MDA extended to a broad of more complex underlying models and proposed MDA-DC. Simulation studies show that the proposed methods have better performances as they can alleviate masking effects to a large extent. By comparing with the traditional single-case deletion method, HIM, and other methods, ADAP, MIP, the proposed methods are superior as they can achieve the designed type I errors and have higher power values in most cases. As mentioned, the detailed derivation of the asymptotic distribution of the influential statistics for MDA-DC warrants further study.

Acknowledgements

The authors are grateful to the editor and anonymous referees for their comments that have greatly improved this paper. This paper was supported by the National Natural Science Foundation of China (grants 11571191 and 11431006), and Project Funded by the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions.

References

[1] Alfons A, Gelper S. Sparse Least Trimmed Squares Regression for Analyzing High-dimensional Large Data Sets. *Annals of Applied Statistics* 2013; 7(1): 226-248.

[2] Barnett V, Lewis T. Outliers in Statistical Data (2nd ed.), *New York: Wiley*; 1994.

[3] Becker C, Gather U. The Largest Nonidentifiable Outlier: A Comparison of Multivariate Simultaneous Outlier Identification Rules. *Computational Statistics and Data Analysis* 2001; 36(1):119-127.

[4] Belsley D A, Kuh E, Welsch R E. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. *New York: Wiley*; 1980.

[5] Chatterjee S, Hadi A S. Sensitivity Analysis in Linear Regression. *New York: Wiley*; 1988.

[6] Cook R D. Detection of Influential Observation in Linear Regression. *Technometrics* 1977; 19(1): 15-18.

[7] Cook R D. Influential Observations in Linear Regression. *Journal of the American Statistical Association* 1979; 74(365): 169-174.

[8] Cook R D, Weisberg S. Residuals and Influence in Regression. *Chapman and Hall, New York-London*; 1982.

[9] Davies L, Gather U. The Identification of Multiple Outliers. *Journal of the American Statistical Association* 1993; 88(423): 782-792.

[10] Fan J., Lv J. Sure Independence Screening for Ultrahigh Dimensional Feature Space (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2008; 70(5): 849-911.

[11] Hadi A S, Simonoff J S. Procedures for the Identification of Multiple Outliers in Linear Models. *Journal of the American Statistical Association* 1993; 88(424): 1264-1272.

[12] Hawkins, D. Identification of Outliers. *Chapman and Hall,* 1980.

[13] Imon A H M R. Identifying Multiple Influential Observations in Linear Regression. *Journal of Applied Statistics* 2005; 32(9): 929-946.

[14] Li R., Zhong W., Zhu L. Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association*, 2012; 107(499): 1129-1139.

[15] Nurunnabi A A M, Hadi A S, Imon A H M R. Procedures for the Identification of Multiple Influential Observations in Linear Regression. *Journal of Applied Statistics* 2014; 41(6): 1315-1331.

[16] Oshima R G, Baribault H, Cauln C. Oncogenic Regulation and Function of Keratins 8 and 18. *Cancer and Metastasis Reviews* 1996; 15(4):445-471.

[17] Pena D. A New Statistic for Influence in Linear Regression. *Technometrics* 2005; 47(1):1-12.

[18] Pena D, Yohai V. A Fast Procedure for Outlier Diagnostics in Large Regression Problems. *Journal of the American Statistical Association* 1999; 94(446): 434-445.

[19] Roberts S, Martin M A, Zheng L. An Adaptive, Automatic Multiple-Case Deletion Technique for Detecting Influence in Regression. *Technometrics* 2015; 57(3): 408-417.

[20] Rousseeuw P J, Leroy A M. Robust Regression and Outlier Detection. *New York: Wiley*; 1987.

[21] Shankavaram, U. T., Reinhold, W. C., Nishizuka, S., Major, S., Morita, D., Chary, K. K., Reimers, M. A., Scherf, U., Kahn, A., Dolginow, D., Cossman, J., Kaldjian, E. P., Scudiero, D. A., Petricoin, E., Liotta, L., Lee, J. K., Weinstein, J. N. Transcript and Protein Expression Profiles of the NCI-60 Cancer Cell Panel: An Integromic Microarray Study. *Molecular Cancer Therapeutics,* 2007; 6(3): 820-832.

[22] Székely G J, Rizzo M L, Bakirov N K Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics.* 2007; 35(6), 2769-2794.

[23] Zhao J, Leng C, Li L, Wang H. High-dimensional Influence Measure. *Annals of Statistics* 2013; 41(5): 2639-2667.

[24] Zhao J, Liu C, Niu L, Leng C. Multiple Influential Point Detection in High-dimensional Spaces. 2016; https://arxiv.org/abs/1609.03320.

[25] Zou C, Tseng S, Wang Z. Outlier Detection in General Profiles Using Penalized Regression Method. *IIE Transactions,* 2014; 46, 106-117.