Outlier detection in high-dimensional regression model

Tao Wang[1,2] and Zhonghua Li[3,*]

1.School of Mathematical Sciences, Nankai University
Tianjin City, 300071, P. R. China
2.School of Mathematics and Statistics, Kashgar University
Kashgar City, 844006, P. R. China
3.Institute of Statistics and LPMC, Nankai University
Tianjin City, 300071, P. R. China

Abstract

An outlier is defined as an observation that is significantly different from the others in its dataset. In high-dimensional regression analysis, datasets often contain a portion of outliers. It is important to identify and eliminate the outliers for fitting a model to a dataset. In this paper, a novel outlier detection method is proposed for high-dimensional regression problems. The leave-one-out idea is utilized to construct a novel outlier detection measure based on distance correlation, and then an outlier detection procedure is proposed. The proposed method enjoys several advantages. First, the outlier detection measure can be simply calculated, and the detection procedure works efficiently even for high-dimensional regression data. Moreover, it can deal with a general regression, which does not require specification of a linear regression model. Finally, simulation studies show that the proposed method behaves well for detecting outliers in high-dimensional regression model and performs better than some other competing methods.

Keywords: High-dimensional, Outlier detection, Cook's distance, Leave-one-out, Distance correlation, Bootstrap

MSC2010: 62H20, 62J02, 62J05, 62J86.

1 Introduction

As Barnett and Lewis (1994) pointed out, an outlier is an observation that appears to be inconsistent with other observations in a set of data. In multiple regression models, data sets often contain a proportion of outliers. Chatterjee and Hadi (1988) thought that outliers can be occurred in the response variable, in the explanatory variable, or in both the response and explanatory variables. Since the presence of outliers can lead to biased estimation of the parameters, misspecification of the model and inappropriate predictions, it is of great concern to detect these observations and eliminate them from the data set.

Various outlier detection methods have been developed in the literature. Hawkins et al. (1984) studied the location of several outliers in multiple regression data by using elemental sets. Breunig et al. (2000) proposed a distance-based approach which considers how isolated a point is with respect to its k nearest neighbors. Türkan et al. (2012) proposed a diagnostic measure based on the robust estimator M to detect influential points. Outlier detection applications may also be found within some databases, such as Yoon et al. (2007). It is important to note that, all outlier detection approaches mentioned above have been developed under the assumption that the number of predictors in regression is fixed. As such, none is immediately applicable to high dimensional regression analysis, where the number of predictors p far exceeds the sample size n.

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

However, as many real data sets may contain hundreds or thousands of dimensions, detection of outliers in high dimensional data analysis is important because the increased dimensionality and complexity of the data may amplify the chance of an observation being outlying or influential as well as its potential impact on the analysis. When the dimension p of the data is greater than the sample size n, many of the aforementioned outlier detection methods do not work well. Aggarwal and Yu (2002) discussed a new technique for high dimensional outlier detection which finds the outliers by studying the behavior of projections from the data set. Ro et al. (2015) proposed an outlier detection procedure that replaces the classical minimum covariance determinant estimator with a high-breakdown minimum diagonal product estimator. Zhao et al. (2013) proposed a new high-dimensional measure (HIM) for diagnosis, which captures the influence on the marginal correlations for high-dimensional linear model and is particularly useful in downstream analysis including coefficient estimation, variable selection and screening. Actually, specifying a correct linear model for high dimensional data may be challenging, and the problem of outlier detection for a general regression model should be considered.

In this paper, we constructed a new measure and proposed a novel outlier detection procedure for high dimensional regression model. Székely et al. (2007) and Székely and Rizzo (2009) showed that the distance correlation (dCor) of two univariate normal random variables is a strictly increasing function of the absolute value of the Pearson correlation of these two normal random variables, and distance correlation provides a new approach to the problem of testing the joint independence of random vectors. This property motivates us to use the distance correlation to construct a novel measure for detecting outlying observations from a general high dimensional regression model. Cook (1977, 1979) studied the problem of detection of influence observations in linear regression, and proposed the Cook's distance, which is a difference measure between the ordinary least square (OLS) estimate of coefficient β on the full data and that on the subset of data without the observation in question. Similarly, we utilize this leave-one-out idea as classical Cook's distance and construct, based on distance correlation between the response and all predictors, a novel outlier detection measure denoted as $\mathcal{D}_i$, for $i = 1, \ldots, n$. Intuitively, the i-th observation $(Y_i, \mathbf{X}_i)$ is more likely to be marked as an outlier if the i-th measure $\mathcal{D}_i$ is large to some extent. Concretely, for the hypothesis that the i-th observation is not an outlier versus its alternative among the n observations, it is not clear what the exact distribution of the proposed measure is in high dimensional setting. We can obtain the asymptotic distribution by bootstrap method. Hence, we use simulations to develop a threshold rule to determine whether an individual observation is an outlier or not. Monte Carlo simulation studies and a real data example are conducted to examine the performance of the proposed procedure. Outlier identification performance are evaluated by the type I error rate, which is the proportion of good observations that are incorrectly deemed as outliers, and the power rate, which is the proportion of contaminated observations that are correctly labelled as good ones. The novel produce enables us to control the type I error and deliver robust outlier detection.

The rest of this paper is organized as follows. In Section 2, we define a new outlier detection measure, and propose a novel outlier detection procedure for high dimensional regression model. In Section 3, we assess the performance of the proposed procedure by Monte Carlo simulation studies. We further illustrate the proposed procedure by analyzing real-life dataset in Section 4. We conclude with a discussion in Section 5.

2 Method and Procedure

Let $Y = (Y_1, \ldots, Y_n)^\top$ be the $n \times 1$ response vector, and $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top = (X_1, \ldots, X_p)$ denotes the $n \times p$ design matrix with the i-th row being p-dimensional vector $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$. A pair of observation $(Y_i, \mathbf{X}_i)$ are assumed to be from the following regression model

$$Y = f(\mathbf{X}, \beta) + \varepsilon, \tag{1}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$, $\varepsilon_i, \ldots, \varepsilon_n$ are independent, identically distributed (i.i.d.) random errors with $\mathrm{E}(\varepsilon_i) = 0$, $\mathrm{var}(\varepsilon_i) = \sigma^2$ for $i = 1, \ldots, n$, $f(\mathbf{X}, \beta) = (f(\mathbf{X}_1, \beta), \ldots, f(\mathbf{X}_n, \beta))^\top$ is a nonlinear function, and $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a p-vector of parameters. In particular, if $f(\mathbf{X}, \beta) = \mathbf{X}\beta$, the equation (1) is a linear regression model.

As Hekimoglu et al. (2015) pointed, all the possible combinations of multiple outliers are considered as model errors. Following this point, we will consider two perturbation models for generating outliers. The first one is response perturbation model,

$$Y_i = f(\mathbf{X}_i, \beta) + \kappa_i + \varepsilon_i, \quad i = 1, \ldots, n^*, \tag{2}$$

where n^* is the number of outlier observations, the responses of the outliers are contaminated by a random perturbation term $\kappa_i = \kappa P_i$, $P_i = \mathbf{X}_i\gamma$, γ is a $p \times 1$ array vector, and κ is the parameter that dictates the magnitude of the outliers, with a larger value of κ indicating a larger abnormal of the outliers. When $\kappa = 0$, there is no outliers in the regression model. The other perturbation model is a model with scale perturbation as

$$Y_i = f(\mathbf{X}_i, \beta) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2/\omega_i), \tag{3}$$

where $\omega_i > 0$ for $i = 1, \ldots, n^*$, and $\omega_i = 1$ denotes the vector of null perturbation.

To quantify the influence of the i-th observation $(Y_i, \mathbf{X}_i)$ on linear regression under the classical setup when $p < n$, Cook (1977) employed the leave-one-out idea by studying the OLS estimate of coefficient β when the i-th observation is excluded from estimation, and proposed a discrepancy measure, i.e., the Cook's distance

$$D_i = \frac{\{\hat{\beta}^{(i)} - \hat{\beta}\}^\top \mathbf{X}^\top \mathbf{X} \{\hat{\beta}^{(i)} - \hat{\beta}\}}{(p+1)\hat{\sigma}^2},$$

where $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y$, $\hat{\beta}^{(i)} = ((\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)})^{-1}(\mathbf{X}^{(i)})^\top Y^{(i)}$, and $\hat{\sigma}^2 = (n-p)^{-1}\sum_{i=1}^n (Y_i - \mathbf{X}_i^\top\hat{\beta})^2$. Intuitively, if the i-th observation is an influential or outlying point, the difference between $\hat{\beta}$ and $\hat{\beta}^{(i)}$ is expected to be large. However, in the high-dimensional regression setting, where the number of predictors p far exceeds the sample size n, the classical Cook's distance is not directly computable, because the OLS estimator $\hat{\beta}$ becomes unstable.

Székely et al. (2007) and Székely and Rizzo (2009) showed that the distance correlation provides a new approach to the problem of testing the joint independence between two random vectors $X \in \mathbf{R}^p$ and $Y \in \mathbf{R}^q$, where p, q are the dimensions of X and Y. More precisely, let $\varphi_X(s)$, $\varphi_Y(t)$ be the respective characteristic functions of two random vectors X and Y, and $\varphi_{X,Y}(s,t)$ be the joint characteristic function of X and Y. The definition of dCor and some of its properties are shown below.

Definition 1 (Distance covariance). The distance covariance (dCov) between X and Y is given by

$$\mathrm{dCov}^2(X,Y) = \int_{R^{p+q}} \|\varphi_{X,Y}(s,t) - \varphi_X(s)\varphi_Y(t)\|^2 w(s,t) ds dt, \tag{4}$$

where $\|\psi\|^2 = \psi\bar{\psi}$ for a complex-valued function $\psi = \varphi_{X,Y}(s,t) - \varphi_X(s)\varphi_Y(t)$, $\bar{\psi}$ is the conjugate of ψ, $w(s,t) = (c_p c_q \|s\|_2^{1+p} \|t\|_2^{1+q})^{-1}$ is a positive weight function with constants $c_l = \pi^{(1+l)/2}/\Gamma((1+l)/2)$ for $l \in \mathbb{N}$ and $\|\cdot\|_2$ is the L_2 norm. Similarly, distance variance (dVar) is defined as the square root of

$$\mathrm{dCov}^2(X,X) = \int_{R^{p+p}} \|\varphi_{X,X}(s,t) - \varphi_X(s)\varphi_X(t)\|^2 w(s,t) ds dt.$$

Definition 2 (Distance correlation). The distance correlation (dCor) between X and Y with finite first moments is then naturally defined as

$$\mathrm{dCor}^2(X,Y) = \frac{\mathrm{dCov}^2(X,Y)}{\sqrt{\mathrm{dCov}^2(X,X)\mathrm{dCov}^2(Y,Y)}}, \tag{5}$$

if $\mathrm{dCov}^2(X,X)\mathrm{dCov}^2(Y,Y) > 0$ and equals 0 otherwise.

Lemma 1 (Properties of dCor). If $E(|X|_p + |Y|_q) < \infty$, then $0 \leq \mathrm{dCor}(X,Y) \leq 1$, and $\mathrm{dCor}(X,Y) = 0$ if and only if X and Y are independent.

The proof of Lemma 1 is similar to that of Theorem 3 in Székely et al. (2007), and thus it is omitted here. The definition of dCor in (5) suggests that the distance dependence measure is analogous to the corresponding product-moment correlation. By analogy, the results in Lemma 1 imply that dCor has certain properties of classical correlation definitions.

Coming back to outlier detection analysis, the distance correlation allows for arbitrary regression relationship of Y onto $\mathbf{X}$, regardless of whether it is linear or nonlinear. So, we use the leave-one-out principle as in the classical Cook's distance case, and compute the distance correlation between X_k and Y while the i-th observation is excluded or not in high-dimensional regression setting. Finally, we define a novel outlier detection measure based on the distance correlation, i.e.

$$\mathcal{D}_i = \frac{1}{p}\sum_{k=1}^{p} (\mathrm{dCor}(X_k, Y) - \mathrm{dCor}(X_k^{(i)}, Y^{(i)}))^2, \tag{6}$$

where $\mathrm{dCor}(X_k, Y)$ denotes the distance correlation between the k-th predictor X_k and the response Y. Analogously, $\mathrm{dCor}(X_k^{(i)}, Y^{(i)})$ denotes the distance correlation between the k-th predictor and the response with the i-th observation removed.

For ease of presentation, we write $d_k = \mathrm{dCor}(X_k, Y)$ and $d_k^{(i)} = \mathrm{dCor}(X_k^{(i)}, Y^{(i)})$ for $k = 1, \ldots, p$, then $\mathcal{D}_i$ defined in (6) can be rewritten as

$$\mathcal{D}_i = \frac{1}{p}\sum_{k=1}^{p} (d_k - d_k^{(i)})^2. \tag{7}$$

Let $\{(Y_i, \mathbf{X}_i); i \in \{1, \ldots, n\}\}$ be i.i.d. observations of size n from the population $(Y, \mathbf{X})$. It is natural to use its sample counterpart to estimate $\mathcal{D}_i$ as follows

$$\widehat{\mathcal{D}}_i = \frac{1}{p}\sum_{k=1}^{p} (\hat{d}_k - \hat{d}_k^{(i)})^2, \tag{8}$$

where $\hat{d}_k$ and $\hat{d}_k^{(i)}$ are the corresponding sample estimates of $\mathrm{dCor}(X_k, Y)$ and $\mathrm{dCor}(X_k^{(i)}, Y^{(i)})$, respectively.

We consider using $\widehat{\mathcal{D}}_i$ as a novel measure to detect outliers in high-dimensional regression model. Next we study the consistency of the proposed outlier detection measure built upon the distance correlation. Toward that end, we impose the following conditions.

(C.1) For any fixed $k = 1, \ldots, p$, d_k is a constant and will not change as p increases.

(C.2) The predictor X_k follows a multivariate normal distribution and the random noise ε_i follows a normal distribution.

Condition (C.1) only requires that for any fixed k, d_k is a constant independent of p. The normality assumption (C.2) on X is mainly for convenience. And the error term is assumed normal, then Y is normally distributed.

Lemma 2 Assume conditions (C1)- (C2) hold, and if $E(|X_k|) < \infty$ and $E(|Y|) < \infty$, then

$$\lim_{n\to\infty} \hat{d}_k^2 = d_k^2 \tag{9}$$

almost surely.

The proof of this lemma is similar to that of Corollary 1 of Theorem 2 in Székely et al. (2007) and that of Theorem 1 in Li et al. (2012), and thus it is omitted here. It will be shown that, by replacing d_k with their consistent sample estimates $\hat{d}_k$ in the novel measure, a robust procedure can be constructed to detect the outliers reliably.

Theorem 1 Under the conditions for Lemma 2, we have

$$\lim_{n\to\infty} \widehat{\mathcal{D}}_i = \mathcal{D}_i \tag{10}$$

almost surely.

Proof: According to the definitions of $\widehat{\mathcal{D}}_i$ and $\mathcal{D}_i$, we have $\mathcal{D}_i - \widehat{\mathcal{D}}_i = \frac{1}{p}\sum_{k=1}^{p}[(d_k - d_k^{(i)})^2 - (\hat{d}_k - \hat{d}_k^{(i)})^2]$. Lemma 2 showed that $\hat{d}_k^2 \to d_k^2$ almost surely, then we have $|\mathcal{D}_i - \widehat{\mathcal{D}}_i| = o_p(1)$ almost surely under certain conditions, that is, $\lim_{n\to\infty} \widehat{\mathcal{D}}_i = \mathcal{D}_i$ almost surely. □

Obviously, the i-th data $(Y_i, \mathbf{X}_i)$ is more likely to be an outlier observation if it has a large measure $\widehat{\mathcal{D}}_i$. A novel outlier detection approach can be formulated as n hypothesis tests with null hypothesis H_{0i}, the i-th observation is not an outlier versus its alternative. After calculating $\widehat{\mathcal{D}}_i$, we develop a threshold rule to determine whether an observation is an outlier or not. At a given significance level α, the i-th observation is identified as an outlier if $\widehat{\mathcal{D}}_i > F_\alpha$, where F_α is the upper α-th quantile of the cumulative distribution function of $\mathcal{D}_i$ under the null hypothesis. The exact distribution of the proposed outlier detection measure is complicated, hence, we obtain an asymptotic distribution by bootstrap procedure. Indeed, our goal is to approximate the distribution of $\mathcal{D}_i$ under the null hypothesis that the i-th observation is not an outlier. Consequently, we draw with replacement $\{i_{(1)}^{[b]}, \ldots, i_{(n)}^{[b]}\}$ from $\{1, 2, \ldots, n\}$ to form the bootstrap sample $\mathcal{D}_i^{[b]}$ for $b = 1$ to B, and study the asymptotic distribution of $\mathcal{D}_i^{[b]}$.

Under these considerations, the novel outlier detection procedure for high dimensional regression can be summarized by the following algorithm.

step 1. Create a sample $(Y_i; \mathbf{X}_i) = (Y_i; X_{i1}, \ldots, X_{ip})_{1\le i\le n}$;

step 2. Compute $\widehat{\mathcal{D}}_i$, an estimator of the dependence measure $\mathcal{D}_i$ for each observation $(Y_i; \mathbf{X}_i)$;

step 3. Realize B bootstrap samplings $\mathcal{D}_i^{[b]} (1 \le b \le B)$ of the sample $\mathcal{D}_i$ under H_0;

step 4. Compute the B bootstrap estimators of the measure $\mathcal{D}_i^{[b]}$, and denoted as $\widehat{\mathcal{D}}_i^{[b]}$;

step 5. Compute the bootstrapped upper α-th quantile of the cumulative distribution function of $\widehat{\mathcal{D}}_i^{[b]}$, denoted as $\widehat{F}_\alpha$;

step 6. The i-th observation is identified as an outlier if $\widehat{\mathcal{D}}_i > \widehat{F}_\alpha$, otherwise H_0 is accepted.

3 Simulation

In this section, we will numerically investigate the algorithm proposed above for outliers detection purpose, and evaluate the performance of the proposed methodology through a simulation study. The simulation studies are conducted using Matlab 2014a software.

3.1 Simulation Models

Consider a general regression problem with a response variable Y and an explanatory variable $\mathbf{X}$. Let the observations $\{(Y_i, \mathbf{X}_i); i = 1, \ldots, n\}$ be from the regression model (1) referred in Section 2, where $\mathbf{X}_i$ is multivariate normal $N(0, \Sigma)$. We consider autoregressive (AR) correlation with $\Sigma = (\rho_{jk})_{p\times p} = 0.5^{|j-k|}$ and moving average (MA) structures. The moving average model is constructed by $X_{ij} = \Sigma_{k=1}^{L}\eta_k z_{i,(j+k-1)}/(\Sigma_{k=1}^{L}\eta_k^2)^{1/2}$ $(i = 1, \ldots, n; j = 1, \ldots, p)$, where η_k and $\{z_{ik}\}$ are independent $U_n(0, 1)$ and $N(0, 1)$ variables respectively. We allow $L = [p^{1/2}]$. $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ are i.i.d. normal distribution $N(0, \sigma^2)$, and $\sigma = 1$ in the following simulation study.

We fix the sample size $n = 100$, and the number of predictors $p = 200, 500, 1000$, respectively. Each dataset is composed of $n - n^*$ observations from the regression model (1), and 10% of total observations are from the perturbation model (2) or model (3), so that the number of outlier observations $n^* = 10$. In the simulation study, we consider the following two models: one is a simple linear model, and the other is a nonlinear model.

Model I (Linear model). If $f(\mathbf{X}, \beta) = \mathbf{X}\beta$, model (1) is a linear model. The observations of null perturbation are from the following linear model:

$$Y_i = \mathbf{X}_i\beta + \varepsilon_i, \tag{11}$$

for $i = 1, \ldots, 100$, and $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)^\top$ is a $p \times 1$ array vector with the first five elements being 1, and the rest being 0. We consider two different type outliers coming from perturbation model (2) or perturbation model (3) with the size of 10. Concretely, we simulated $n = 100$ i.i.d. observations from model (11). Next we reset the $n^* = 10$ data observations coming from another perturbation model (2) with perturbation term $\kappa_i = \kappa P_i$, $P_i = \mathbf{X}_i\gamma$ where $\gamma = (0, 0, 0, 0, 0, 1, \ldots, 1)^\top$ is a $p \times 1$ array vector with the first five elements being 0, and the rest being 1. κ is the perturbation parameter, and let $\kappa = 0.2, 0.5, 0.8, 1.0, 1.5$ respectively. Similarly, we reset the $n^* = 10$ data observations coming from another perturbation model (3) with perturbation parameter $\omega = 0.01, 0.02, 0.05, 0.1, 0.15$ respectively.

Model II (Nonparametric additive model). Let $g_1(x) = x$, $g_2(x) = (2x - 1)^2$, $g_3(x) = \sin(2\pi x)/(2 - \sin(2\pi x))$, $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3$. Consider the observations from the following nonlinear model

$$Y_i = 2g_1(X_{i1}) + 6g_2(X_{i2}) + 4g_3(X_{i3}) + g_4(X_{i4}) + \varepsilon_i. \tag{12}$$

Similar to Model I, we simulated $n = 100$ i.i.d. observations from model (12). Next we reset $n^* = 10$ data observations as coming from another perturbation model (2) or model (3) corresponding to two different type outliers with the size 10.

The goal in this section is to detect those outliers by the proposed detection procedure, and remove them from the datset by simulation procedure.

3.2 Performance Evaluation

For the hypothesis that the i-th observation is not influential versus its alternative, we utilize bootstrap procedure and develop a threshold rule to determine whether an individual is an outlier.

We evaluate our proposed outlier identification procedure by the type I error rate and the power. Denote n^* as the number of outlier observations among the n observations, n_{tp} and n_{fp} as the number of the observations that are truly rejected and falsely rejected. Then the type I error rate=$n_{fp}/(n-n^*)$ and power rate =n_{tp}/n^*.

Under the same settings, the nominal significance level α is chosen to be 0.01, 0.05 or 0.1. We use simulations to find the cutoff values such that a desired type I error is achieved. Zhao et al. (2013) proposed the high dimensional influence diagnosis measure (HIM), evaluated the corresponding p-value, and applied the multiple testing procedure of Benjamini and Hochberg (1995), with the false discovery rate fixed at $\alpha = 0.05$, then obtained a reduced data set by removing those flagged influential or outlier observations. Therefore, we compare our proposed outlier detection procedure with HIM in terms of type I error rate and power rate.

3.3 The Results

Response perturbation and scale perturbation are considered under the AR and MA structure for model I and model II, respectively. To obtain the upper α-th quantile of the cumulative distribution function of $\widehat{\mathcal{D}}_i^{[b]}$, we set the bootstrap time $B = 500$ in the algorithm summarized in Section 2. The averages of a total of 200 random replications are reported in Tables 1-6.

Insert Tables 1-6 about here.

(1) Tables 1-2 report the type I error rate and power of the proposed method in Model I under various values of p. The nominal type I error $\alpha = 0.01, 0.05, 0.1$, and the perturbation parameter $\kappa = 1.5$ and $\omega = 0.02$. For AR correlation and MA structures, the simulated type I error rates are close to the nominal levels in most settings, which shows the effectiveness of the suggested detection procedure. Next, we evaluate the power of the proposed method. Tables 1-2 present power results for Model I when $p = 200, 500, 1000$. It is shown that the power rates increases as the dimension p increases, and the proposed method has better efficiency with larger p as expected. Similarly, Tables 3-4 report the type I error rates and power rates in Model II of the proposed method respectively. In most cases, the proposed method can maintain the desired type I error rate and a better power rate regardless of Model I or Model II.

(2) Next, we compare the proposed outlier detection procedure with HIM of Zhao et al. (2013) under Model I and Model II. Simulation results with nominal size $\alpha = 0.05$ and $n^* = 10$ are summarized in Tables 5-6. In contrast, the HIM method does not work well, with the power rates being less than that of our method in most cases. In fact, as κ or $1/\omega$ increases, the power for the proposed method to detect outlier observations increases. Thus, those outliers are more likely to be detected and eliminated from the data analysis by our method.

(3) To summarize, our simulation results confirm that the proposed outlier detection procedure for high dimensional regression model is useful to control the type I error and deliver robust outlier detection.

4 A Real Data Analysis

In this section, we apply the proposed procedure to discuss a real dataset analysis.

Example (Cardiomyopathy microarray dataset). The data was once used by Segal et al. (2003) to evaluate regression-based approaches to microarry analysis. Hall and Miller (2009) and Li et al. (2012) also used this cardiomyopathy microarray dataset to illustrate their proposed screening procedure in terms of ranking.

The dataset analysis is related to studying all types of human heart disease. In this dataset, the response Y is the Ro1 expression level, and the predictor X_k is gene expression level for $k = 1, \ldots, p$. The sample size $n = 30$ and the dimension $p = 6319$. We observe that $p \gg n$, so this is a high dimension data analysis problem. We aim to identify the most influential genes for over expression of a G protein-coupled receptor (Ro1) in mice.

Firstly, we fit an additive model :

$$Y_i = g_1(X_{ij}) + g_2(X_{ik}) + \varepsilon_i, \quad i = 1, \ldots, 30; \quad j, k = 1, \ldots, 6319,$$

where g_1, g_2 are unknown link functions and we fit them using the **R mgcv** package. Moreover, we apply outlier detection procedure for high dimensional regression proposed in Section 2, and detected three outlying specimens from 30 specimens. After that, we utilize the feature screening procedure based on distance correlation (DCS, Li et al., 2012) to identify the most influential genes. Our analysis is based on ranking the dCor between predictors and the response by the remaining 27 data pairs. This approach leads us to rank the two genes, labeled Msa.2134.0 and Msa.28021.0, as the top two genes. Figures 1-2 indicate the scatter plots and corresponding cubic-spline fit curves. Actually, they show clearly the existence of nonlinear patterns.

Insert Figures 1-2 about here.

Compared the performance of the proposed method (denoted by R-DCS) with the generalized correlation ranking (GC) method of Hall and Miller (2009) and the DCS method of Li et al. (2012), we can see from Table 7 that, R-DCS clearly achieves better performance than GC and DCS with better R^2 and deviance performance. Note that deviance means the proportion of the null deviance explained by the proposed model, with a larger value indicating better performance. This suggests that the proposed method helps DCS in removing three perturbation observations firstly, and then using the DCS procedure to screen the important genes, the adjusted R^2 values and the explained deviance are better than the results of other two procedures.

Insert Table 7 about here.

5 Conclusion and Discussion

In this paper, we constructed an outlier detection measure based on the distance correlations between the response and all predictors and proposed a novel outlier detection procedure for high dimensional regression model. We examined the performance of the proposed outlier detection procedure for linear model and nonlinear model via simulation studies, and illustrated the proposed methodology through a real data example. We would like to comment on the main advantages delivered in our work. Firstly, the new high dimensional influence measure is easy to compute regardless of high dimension setting. Moreover, it is model-free because its implementation does not require specification of the regression model. In addition, both the Monte Carlo simulation examples and a real-life data show that the proposed outlier detection procedure can greatly detect those outlier observations, and improve the filtering accuracy in feature screening problem.

Acknowledgements

The authors are grateful to the editor and an anonymous referee for their comments that have greatly improved this paper. This paper was supported by the National Natural Science Foundation of China (grants 11571191, 11371202, 11431006 and 11131002) and the Fund for the Doctoral Program of Kashgar University (grant 14-2498).

References

[1] Aggarwal, C.C. and Yu, P.S. (2002). Outlier detection for high dimensional data. *ACM Sigmod Record.* 30(2), 37-46.

[2] Barnett, V. and Lewis, T. (1994). Outliers in statistical data. New York, Wiley.

[3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B.* 57(1), 289-300.

[4] Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J. (2000). LOF: identifying density-based local outliers. *ACM Sigmod Record.* 29(2), 93-104.

[5] Chatterjee, S. and Hadi, A.S. (1988). Sensitivity Analysis in Linear Regression. New York, Wiley.

[6] Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics.* 19(1), 15-18.

[7] Cook, R.D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association.* 74(365), 169-174.

[8] Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics.* 18, 533-550.

[9] Hawkins, D.M., Bradu, D. and Kass, G.V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics.* 26(3), 197-208.

[10] Hekimoglu, S., Erdogan, B. and Erenoglu, R.C. (2015). A new outlier detection method considering outliers as model errors. *Experimental Techniques.* 39(1), 57-68.

[11] Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association.* 107(499), 1129-1139.

[12] Ro, K., Zou, C., Wang, Z. and Yin, G. (2015). Outlier detection for high-dimensional data. *Biometrika.* 102(3), 589-599.

[13] Segal, M.R., Dahlquist, K.D. and Conklin, B.R. (2003). Regression approach for microarray data analysis. *Journal of Computational Biology.* 10, 961-980.

[14] Székely G.J. and Rizzo M.L. (2009). Brownian distance covariance. *The Annals of Applied Statistics.* 3(4), 1236-1265.

[15] Székely G.J., Rizzo M.L. and Bakirov N.K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics.* 35(6), 2769-2794.

[16] Türkan, S., Çetin, M.C. and Toktamiş, Ö. (2012). Outlier detection by regression diagnostics based on robust parameter estimates. *Hacettepe Journal of Mathematics and Statistics.* 41, 147-155.

[17] Yoon, K.-A., Kwon, O.-S. and Bae, D.-H. (2007). An approach to outlier detection of software measurement data using the K-means clustering method. *in: Proceedings of the First International Symposium on Empirical Software Engineering and Measurement, IEEE Computer Society*, Washington. 443-445.

[18] Zhao, J., Leng, C., Li, L. and Wang, H. (2013). High-dimensional influence measure. *The Annals of Statistics.* 41(5), 2639-2667.

Table 1: Average type I errors (size) and power rates in Model I of response perturbation setting, with perturbation parameter $\kappa = 1.5$ for various values of $p = 200, 500, 1000$. α is the nominal significance level, and $n = 100$, $n^* = 10$.

Model I		α =0.01		α=0.05		α=0.1	
Correlation	p	size	power	size	power	size	power
AR	200	0.015	0.801	0.051	0.866	0.099	0.897
	500	0.014	0.886	0.050	0.926	0.101	0.951
	1000	0.014	0.925	0.051	0.939	0.100	0.944
MA	200	0.015	0.833	0.049	0.893	0.099	0.907
	500	0.014	0.904	0.050	0.932	0.100	0.942
	1000	0.013	0.919	0.050	0.938	0.101	0.948

AR, autoregressive; MA, moving average.

Table 2: Average type I errors (size) and power rates in Model I of scale perturbation setting, with perturbation parameter $\omega = 0.02$ for various values of $p = 200, 500, 1000$. α is the nominal significance level, and $n = 100$, $n^* = 10$.

Model I		α =0.01		α=0.05		α=0.1	
Correlation	p	size	power	size	power	size	power
AR	200	0.014	0.821	0.051	0.872	0.100	0.917
	500	0.015	0.848	0.051	0.892	0.100	0.924
	1000	0.014	0.855	0.052	0.893	0.101	0.917
MA	200	0.016	0.720	0.051	0.827	0.100	0.869
	500	0.015	0.769	0.051	0.834	0.101	0.883
	1000	0.014	0.773	0.051	0.838	0.100	0.888

Table 3: Average type I errors (size) and power rates in Model II of response perturbation setting, with perturbation parameter $\kappa = 1.5$ for various values of $p = 200, 500, 1000$. α is the nominal significance level, and $n = 100$, $n^* = 10$.

Model II		α =0.01		α=0.05		α=0.1	
Correlation	p	size	power	size	power	size	power
AR	200	0.014	0.340	0.049	0.542	0.096	0.725
	500	0.014	0.580	0.049	0.718	0.095	0.826
	1000	0.014	0.720	0.049	0.810	0.094	0.881
MA	200	0.015	0.593	0.048	0.765	0.094	0.844
	500	0.014	0.784	0.048	0.871	0.094	0.908
	1000	0.014	0.878	0.048	0.918	0.098	0.942

Table 4: Average type I errors (size) and power rates in Model II of scale perturbation setting, with perturbation parameter $\omega = 0.02$ for various values of $p = 200, 500, 1000$. α is the nominal significance level, and $n = 100$, $n^* = 10$.

Model II		α =0.01		α=0.05		α=0.1	
Correlation	p	size	power	size	power	size	power
AR	200	0.015	0.489	0.049	0.660	0.096	0.789
	500	0.013	0.544	0.050	0.704	0.095	0.809
	1000	0.013	0.548	0.049	0.712	0.097	0.814
MA	200	0.014	0.440	0.050	0.620	0.095	0.758
	500	0.014	0.485	0.048	0.648	0.094	0.765
	1000	0.014	0.53	0.049	0.683	0.096	0.803

Table 5: Power rate results for Model I and Model II of response perturbation setting, perturbation parameter $\kappa = 0.2, 0.5, 0.8, 1.0, 1.5$, the nominal significance level $\alpha = 0.05$, and $p = 1000$, $n = 100$, $n^* = 10$.

			κ				
	Method	Correlation	0.2	0.5	0.8	1.0	1.5
Model I	HIM	AR	0.474	0.551	0.566	0.548	0.565
	HDC		0.639	0.844	0.896	0.921	0.939
	HIM	MA	0.538	0.550	0.550	0.551	0.553
	HDC		0.762	0.901	0.934	0.938	0.938
Model II	HIM	AR	0.251	0.476	0.525	0.546	0.555
	HDC		0.159	0.491	0.667	0.724	0.810
	HIM	MA	0.480	0.535	0.566	0.547	0.544
	HDC		0.510	0.795	0.872	0.892	0.918

HIM, the method of Zhao et al. (2013),
HDC, the high dimensional outlier detection method based on dCor proposed in this paper.

Table 6: Power rate results for Model I and Model II of scale perturbation setting, with perturbation parameter $\omega = 0.01, 0.02, 0.05, 0.10, 0.15$, the nominal significance level $\alpha = 0.05$, and $p = 1000$, $n = 100$, $n^* = 10$.

			ω				
	Method	Correlation	0.01	0.02	0.05	0.1	0.15
Model I	HIM	AR	0.567	0.543	0.496	0.398	0.293
	HDC		0.934	0.893	0.722	0.505	0.333
	HIM	MA	0.571	0.545	0.477	0.342	0.225
	HDC		0.922	0.838	0.625	0.339	0.170
Model II	HIM	AR	0.552	0.528	0.397	0.226	0.133
	HDC		0.851	0.712	0.328	0.131	0.076
	HIM	MA	0.549	0.536	0.416	0.240	0.140
	HDC		0.843	0.683	0.349	0.141	0.078

Table 7: Simulation results for Example. Performance of the adjusted R^2 and the deviance under three procedures.

Produces	Top two genes	R^2	Deviance
GC	Msa.2877.0, Msa.1166.0	0.659	71.6%
DCS	Msa.2877.0, Msa.2134.0	0.653	70.5%
R-DCS	Msa.2134.0, Msa.28021.0	0.78	81.4%

GC, the generalized correlation ranking method in Hall and Miller (2009);
DCS, feature screening via dCor by Li et al. (2012);
R-DCS, removing outlying observations by outlier detection procedure in this paper, and then using DCS method.

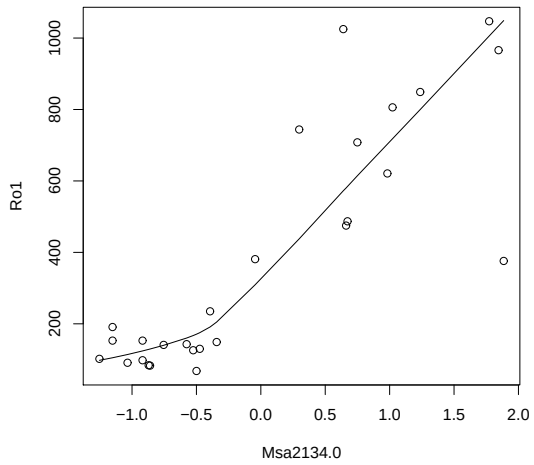

Figure 1: The scatter plot of Y versus Msa.2134.

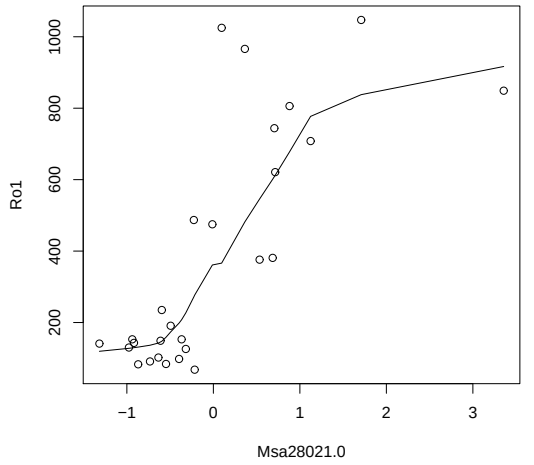

Figure 2: The scatter plot of Y versus Msa.28021.