A Robust Variable Screening Method for High Dimensional Data

Tao Wang[1,2], Lin Zheng[1], Zhonghua Li[3,*] and Haiyang Liu[4]

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
4. Department of Aviation Material Management, Air Force Logistics College
Xuzhou City, 221000, P. R. China

Abstract

In practice, the presence of influential observations may lead to misleading results in variable screening problems. We, therefore, propose a robust variable screening procedure for high dimensional data analysis in this paper. Our method consists of two steps. The first step is to define a new high dimensional influence measure and propose a novel influence diagnostic procedure to remove those unusual observations. The second step is to utilize the sure independence screening procedure based on distance correlation to select important variables in high dimensional regression analysis. The new influence measure and diagnostic procedure that we developed are model free. To confirm the effectiveness of the proposed method, we conduct simulation studies and a real-life data analysis to illustrate the merits of the proposed approach over some competing methods. Both the simulation results and the real-life data analysis demonstrate that the proposed method can greatly control the adverse effect after detecting and removing those unusual observations, and performs better than the competing methods.

Keywords: High dimensional data analysis, Variable screening, Influential observation, Distance correlation, Bootstrap.

MSC2010: 62F40, 62J02, 62J20.

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

1 Introduction

Nowadays, modern scientific research encounters data set with tens of thousands of variables, and variable screening has received a lot of attention in high dimensional data analysis. Ever since Fan and Lv [5] proposed sure independence screening (SIS) methodology for linear regression which screens variables by ranking their marginal correlations with the response variable, various variable screening procedures for high dimensional data have been proposed and studied for different models in recent years. Fan *et al.* [7] and Fan and Song [8] further extended the SIS methodology to generalized linear models. Hall and Miller [9] studied the problem of nonlinear variable screening by generalized correlation ranking. Fan *et al.* [4] proposed a nonparametric marginal screening procedure for additive models based on B-spline expansion. Fan *et al.* [6] extended the nonparametric B-spline method to varying coefficient models and proposed a marginal sure screening procedure. Liu *et al.* [13] proposed a local kernel-based marginal sure screening procedure for varying coefficient models and established its sure screening property.

The aforementioned model-based screening procedures would perform well when the underlying models are correctly specified, but their performance may be quite poor in the presence of model mis-specification. Thus, in high dimensional data analysis, model-free sure screening procedures are appealing and have been developed in recent literatures. Zhu *et al.* [21] proposed a sure independent ranking and screening (SIRS) procedure which avoids the specification of a particular model structure. Motivated by this work, He *et al.* [10] proposed a framework called quantile-adaptive model-free screening. Li *et al.* [12] developed the sure independence screening procedure based on distance correlation (DC-SIS) procedure, which will be applied later in this paper. Mai and Zou [14] applied the fused Kolmogorov filter to deal with variable screening problems. Cui *et al.* [2] proposed a marginal variable screening procedure based on empirical conditional distribution function.

Let $\mathrm{y} = (Y_1, \ldots, Y_n)^\top$ be an n-vector of responses, and $\mathrm{X} = (\mathrm{x}_1^\top, \ldots, \mathrm{x}_n^\top)^\top = (X_1, \ldots, X_p)$ be the associated covariate vectors with sample size n and number of covariates p. The vectors $(\mathrm{x}_1, Y_1), \ldots, (\mathrm{x}_n, Y_n)$ are assumed to be independent and identically distributed (i.i.d.) realizations from a population. The i-th observation (x_i, Y_i) is flagged as influential if some important features are substantially altered after removing this observation. It is important to consider these influential observations in data analysis. When the dimension in regression is relatively low, many diagnostic procedures have been developed upon different models, Zhu *et al.* [20] gave an excellent review on the latest development in the field of influence diagnosis. Cook [1] considered the detection of influential observations in linear regression, utilizing the leave-one-out idea and ordinary least square (OLS) estimate of regression coefficients, and proposed a discrepancy measure denoted by Cook's distance. Cook's distance measures the effect of deleting a given observation on OLS parameter estimates. Those points with a large Cook's distance may be considered to be influential in the influence analysis.

For high dimensional data where the dimension increases with the sample size, the detection of influential observations is more important than for the classical regression model. Zhao *et al.* [19] proposed a high dimensional influence measure (HIM) that captures the influence on the marginal correlations for high dimensional linear model and demonstrated that it is particularly useful in downstream analysis including coefficient estimation, variable selection and screening. However, Zhao *et al.* [19]'s work only focused on high dimensional linear model or generalized linear model, and their approach may not perform so well for other general models, such as those with nonlinear relationships which are probably ignored by marginal correlation.

Székely *et al.* [18] and Székely and Rizzo [17] systematically studied the distance correlation (dCor) of two random vectors, and showed that the distance correlation equals to 0 if and only if these two random vectors are statistically independent. In this paper, we aim to detect and remove the influential observations by an effective influence diagnostic procedure, and then filter

out many noise variables and identify all important variables by DC-SIS screening procedure. We develop the diagnostic procedure by utilizing the leave-one-out idea and defining a new high dimensional influence diagnostic measure $\delta_k, k = 1, \ldots, n$ based on dCor between the response and the predictor variables. The k-th observation (x_k, Y_k) is more likely to be marked influential if its corresponding influence measure δ_k is large to some extent. As the exact distribution of the proposed diagnostic measure is complicated, we use bootstrap technique (Efron [3]) to approximate the upper α-th sample quantile of the cumulative distribution function (CDF) of δ_k. By removing those flagged influential observations, we obtain a reduced data set. Next, we utilize the screening procedure proposed by Li *et al.* [12] on the reduced data set. We conduct Monte Carlo simulation studies to numerically compare the screening results with and without those flagged influential observations. Our simulation results indicate that influential observations have strong effect on the results of variable screening. Meanwhile, our proposed diagnostic procedure could greatly help control the adverse effect after detecting and removing those influential observations in terms of variable screening results. The proposed procedure can be directly applied for a real data analysis. We use the Cardiomyopathy microarray data set to identify the most relevant genes for over expression of a G protein-coupled receptor (Ro1) in mice.

The rest of the paper is organized as follows. In Section 2, we give some preliminaries about classical Cook's distance, high dimensional influential measure and distance correlation. In Section 3, we define a novel influence measure, propose a new influence diagnostic procedure, and apply the DC-SIS variable screening procedure for high dimensional regression. In Section 4, we assess the performance of the proposed procedure by Monte Carlo simulation studies. We further illustrate the proposed procedure by analyzing a real-life data set in Section 5. Section 6 contains conclusions and further discussions.

2 Some Preliminaries

2.1 Cook's Distance

Consider the observations $\{\mathrm{x}_i, Y_i\}, i = 1, \ldots, n$, from the linear regression model

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon, \tag{1}$$

where $\mathrm{y} = (Y_1, \ldots, Y_n)^\top$ denotes the n-vector of responses, $\mathrm{X} = (\mathrm{x}_1^\top, \ldots, \mathrm{x}_n^\top)^\top$ denotes the $n \times p$ design matrix with independent and identically distributed (i.i.d.) $\mathrm{x}_1, \ldots, \mathrm{x}_n$, and X_{ij} denotes the i-th observation of the j-th variable, thus, $\mathrm{x}_i = (X_{i1}, \ldots, X_{ip})$. $\beta = (\beta_1, \ldots, \beta_p)^\top$ denotes a p-vector of regression coefficients, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is an n-vector of random errors with zero mean.

Under the classical setup of $p < n$, by minimizing the objective function $\sum_{i=1}^n (Y_i - \mathrm{x}_i\beta)^2$, the OLS estimator of regression coefficient β is $\hat{\beta} = (\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top\mathrm{y}$. To quantify the influence of the k-th observation (x_k, Y_k), $k = 1, \ldots, n$, on regression, Cook [1] defined an influence measure by employing the leave-one-out idea and finding the OLS estimator of β with and without the k-th observation, respectively. That is, let $\mathrm{y}^{(k)}$ denote an $(n-1) \times 1$ response vector with the k-th response Y_k removed, and let $\mathrm{X}^{(k)}$ denote the $(n-1) \times p$ design matrix with the k-th row $\mathbf{x}_k$ removed. Minimizing the modified objective function $\sum_{i=1, i\neq k}^n (Y_i - \mathrm{x}_i\beta)^2$, another OLS estimator $\hat{\beta}^{(k)} = ((\mathrm{X}^{(k)})^\top\mathrm{X}^{(k)})^{-1}(\mathrm{X}^{(k)})^\top\mathrm{y}^{(k)}$ is obtained, with the k-th observation (x_k, Y_k) removed.

Cook [1] asserted that the k-th observation is expected influential if the difference between

$\hat{\beta}$ and $\hat{\beta}^{(k)}$ is large, and proposed Cook's distance

$$D_k = \frac{\{\hat{\beta}^{(k)} - \hat{\beta}\}^\top \mathrm{X}^\top \mathrm{X}\{\hat{\beta}^{(k)} - \hat{\beta}\}}{(p+1)s^2}, \tag{2}$$

where $s^2 = (n-p)^{-1}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i\hat{\beta})^2$ is the mean squared error of the regression model based on all n observations.

2.2 High Dimensional Influence Measure

In high dimensional setting, where the dimension p is larger than the sample size n, the design matrix X is rectangular, having more columns than rows. The OLS estimator of β is not unique, and therefore the classical Cook's distance can not be computed directly. Then it is inappropriate to use the regression coefficient estimate to define influence measures.

To overcome this difficulty, Zhao *et al.* [19] defined a new high dimensional influence measure by choosing marginal correlation between the variables and the response, instead of regression coefficients. The marginal correlation between variables X and Y is defined as

$$\rho = E\{(X - \mu_X)(Y - \mu_Y)\}/(\sigma_X\sigma_Y),$$

where $\mu_X = E(X)$, $\mu_\mathrm{Y} = E(Y)$, σ_X^2=var(X) and σ_Y^2=var(Y).

Suppose $\{(x_i, y_i) : 1 \le i \le n\}$ is an observed random sample of size n from the joint distribution of (X, Y). Then the consistent estimator of ρ is

$$\hat{\rho} = \sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})/\sqrt{\sum_{i=1}^{n}(x_i - \bar{X})^2\sum_{i=1}^{n}(y_i - \bar{Y})^2},$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are, respectively, the sample estimates of μ_X and μ_Y.

The HIM is defined as

$$\mathcal{D}_k = \frac{1}{p}\sum_{j=1}^{p}(\hat{\rho}_j - \hat{\rho}_j^{(k)})^2, \tag{3}$$

where

$$\hat{\rho}_j = \sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(Y_i - \bar{\mathrm{y}})/\sqrt{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2\sum_{i=1}^{n}(Y_i - \bar{\mathrm{y}})^2}$$

is the estimate of correlation between response y and predictor X_j based on all observations, $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$, $\bar{\mathrm{y}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and

$$\hat{\rho}_j^{(k)} = \sum_{i=1,i\neq k}^{n}(X_{ij} - \bar{X}_j^{(k)})(Y_i - \bar{\mathrm{y}}^{(k)})/\sqrt{\sum_{i=1,i\neq k}^{n}(X_{ij} - \bar{X}_j^{(k)})^2\sum_{i=1,i\neq k}^{n}(Y_i - \bar{\mathrm{y}}^{(k)})^2}$$

is an estimate of the same correlation with k-th observation removed, $\bar{X}_j^{(k)} = \frac{1}{n-1}\sum_{i=1,i\neq k}^{n} X_{ij}$ and $\bar{\mathrm{y}}^{(k)} = \frac{1}{n-1}\sum_{i=1,i\neq k}^{n} Y_i$.

2.3 Distance Correlation

Székely *et al.* [18] introduced dCor as a measurement of dependence between two random vectors. Assume $X \in \mathbb{R}^q$, $Y \in \mathbb{R}^r$, and q, r are the dimensions of the random vectors X and Y, respectively. Let $\varphi_X(s)$, $\varphi_Y(t)$ and $\varphi_{X,Y}(s,t)$ be the characteristic function of X, the characteristic function of Y, and the joint characteristic function of (X,Y), respectively. The distance covariance (dCov) between X and Y with finite first moments is defined by

$$\mathrm{dcov}^2(X,Y) = \int_{\mathbb{R}^{q+r}} \|\varphi_{X,Y}(s,t) - \varphi_X(s)\varphi_Y(t)\|^2 w(s,t)dsdt,$$

where $w(s,t)$ is a positive weight function, $\|\varphi\|^2 = \varphi\bar{\varphi}$ for a complex-valued function φ, and $\bar{\varphi}$ is the conjugate of φ. Accordingly, the dCor between X and Y with finite first moments is defined by

$$\mathrm{dcor}^2(X,Y) = \frac{\mathrm{dcov}^2(X,Y)}{\sqrt{\mathrm{dcov}^2(X,X)\mathrm{dcov}^2(Y,Y)}},$$

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

With a properly chosen weight function $w(s,t)$, Székely *et al.* [18] stated that

$$\mathrm{dcov}^2(X,Y) = S_1 + S_2 - 2S_3,$$

where

$$\begin{aligned} S_1 &= E\{\|X - \tilde{X}\|_q \|Y - \tilde{Y}\|_r\}, \\ S_2 &= E\{\|X - \tilde{X}\|_q\} E\{\|Y - \tilde{Y}\|_r\}, \\ S_3 &= E\{E(\|X - \tilde{X}\|_q | X) E(\|Y - \tilde{Y}\|_r | Y)\}, \end{aligned}$$

$(\tilde{X}, \tilde{Y})$ is an independent copy of (X,Y), and $\|\cdot\|$ denotes the Euclidean norm. Suppose $\{(x_i, y_i) : 1 \le i \le n\}$ to be an observed random sample of size n from the joint distribution of (X,Y), then the moment estimation of S_1, S_2, S_3 can be written as

$$\begin{aligned} \hat{S}_1 &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|x_i - x_j\|_q \|y_i - y_j\|_r, \\ \hat{S}_2 &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|x_i - x_j\|_q \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \|y_i - y_j\|_r, \\ \hat{S}_3 &= \frac{1}{n^3}\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n \|x_i - x_j\|_q \|y_i - y_l\|_r. \end{aligned}$$

Then, the sample version of dCov between X and Y is given by

$$\widehat{\mathrm{dcov}^2}(X,Y) = \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3,$$

and the sample dCor between X and Y is defined by

$$\widehat{\mathrm{dcor}^2}(X,Y) = \frac{\widehat{\mathrm{dcov}^2}(X,Y)}{\sqrt{\widehat{\mathrm{dcov}^2}(X,X)\widehat{\mathrm{dcov}^2}(Y,Y)}}.$$

Székely *et al.* [18] derived the following properties of dCor. If $E(|X|_q + |Y|_r) < \infty$, then almost surely, (i). $0 \le \mathrm{dcor}(X,Y) \le 1$, and $\mathrm{dcor}(X,Y) = 0$ if and only if X and Y are independent; (ii). $\lim_{n\to\infty} \widehat{\mathrm{dcor}^2}(X,Y) = \mathrm{dcor}^2(X,Y)$; (iii). $0 \le \widehat{\mathrm{dcor}^2}(X,Y) \le 1$.

We will propose a new influence diagnostic procedure in the next section by utilizing these properties of dCor to define a new influence measure.

3 Influence Diagnosis and Variable Screening

In this section, we consider variable screening under a general model,

$$\mathrm{y} = f(\mathrm{X}) + \varepsilon, \tag{4}$$

where $f(\cdot)$ denotes an unknown functional relationship, y is an n-vector of response variable with support set ψ_y, X denotes an $n \times p$ design matrix with i.i.d. $\mathrm{x}_1, \ldots, \mathrm{x}_n$, and ε is an n-vector of random errors with zero mean.

When the dimension p is high, it is often assumed that only a small number of variables among $X_1, \ldots, X_p$ contribute to the response y. Let $F(\mathrm{y}|\mathrm{X})$ be the conditional distribution of y given X. Without specifying a regression model, we define an active variable subset as

$$\mathcal{A} = \{j : F(\mathrm{y}|\mathrm{X}) \text{ functionally depends on } X_j \text{ for some } \mathrm{y} \in \psi_y\},$$

and the complementary set $\mathcal{I} = \{1, \ldots, p\} \setminus \mathcal{A}$ is an inactive variable subset. If $j \in \mathcal{A}$, X_j is referred to as an active variable, whereas if $j \in \mathcal{I}$, X_j is referred to as an inactive variable.

The main goal of variable screening procedure is to select a reduced model with a moderate scale, which can still almost fully contain the active variable subset $\mathcal{A}$. In consideration of the arbitrary relationship between y and X, we utilized the distance correlation between the predictor variables and the response to select the important variables, and constructed a subspace $\widehat{\mathcal{A}}$ with important variables. Under the conditions (C1) and (C2) in Li *et al.* [12], the sure screening property holds for DC-SIS, i.e. $P(\mathcal{A} \subset \widehat{\mathcal{A}}) \to 1$ as $n \to \infty$. It is obvious that the constructed subspace $\widehat{\mathcal{A}}$ has larger probability of including the true model $\mathcal{A}$ under more general conditions.

The observations $\{(\mathrm{x}_i, Y_i),\ i = 1, \ldots, n\}$ from (4) may, however, contain a portion of influential observations, and the existence of these unusual observations would possibly lead to unsatisfactory results for variable screening. So, it is necessary to detect and remove those influential observations before variable screening procedure in high dimensional data analysis. As Székely *et al.* [18] pointed out, as a measurement of dependence, the distance correlation has an important property that $\mathrm{dcor}(X_j, Y) = 0$ if and only if X_j and Y are independent. Motivated by this property, we can detect influential observations by defining a dCor-based influence measure, which is expected to be more effective than marginal correlation-based HIM in the presence of nonlinear relationship between X_j and Y.

Next, we propose a new influence diagnostic procedure for high dimensional influential observations. In addition, we use the DC-SIS screening procedure on the cleaned data set to select the important variables.

3.1 A New High Dimensional Influence Measure

Firstly, we define the dCor between the response y and the j-th predictor variable X_j as

$$\mathrm{dcor}(X_j, \mathrm{y}) = \frac{\mathrm{dcov}(X_j, \mathrm{y})}{\sqrt{\mathrm{dcov}(X_j, X_j)\mathrm{dcov}(\mathrm{y}, \mathrm{y})}}.$$

We write $\gamma_j = \mathrm{dcor}(X_j, \mathrm{y})$ for simplification, and the sample estimate of γ_j can be written as

$$\widehat{\gamma}_j = \widehat{\mathrm{dcor}}(X_j, \mathrm{y}) = \frac{\widehat{\mathrm{dcov}}(X_j, \mathrm{y})}{\sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(\mathrm{y}, \mathrm{y})}}.$$

Next, utilizing the leave-one-out idea of the classical Cook's distance, when the k-th observation (x_k, Y_k) is removed, the sample estimate of the dCor between the variables and the response can be written as

$$\widehat{\gamma}_j^{(k)} = \widehat{\mathrm{dcor}}(X_j^{(k)}, \mathrm{y}^{(k)}) = \frac{\widehat{\mathrm{dcov}}(X_j^{(k)}, \mathrm{y}^{(k)})}{\sqrt{\widehat{\mathrm{dcov}}(X_j^{(k)}, X_j^{(k)})\widehat{\mathrm{dcov}}(\mathrm{y}^{(k)}, \mathrm{y}^{(k)})}}.$$

Finally, we construct a novel influence measure based on dCor as

$$\delta_k = \frac{1}{p}\sum_{j=1}^{p}(\widehat{\gamma}_j - \widehat{\gamma}_j^{(k)})^2. \tag{5}$$

It is obvious that the k-th observation (x_k, Y_k) is more likely to be marked influential, if it has a large influence measure δ_k. The new influence measure δ_k can be easily computed regardless of the variable dimensionality, and it does not require the regression function between y and X to be linear.

3.2 Influence Diagnostic Procedure

In this subsection, we propose a new influence diagnostic procedure for data from the high dimensional regression model in Equation (4).

The distribution of the proposed high-dimensional influence diagnostic measure δ_k is not known and so complicated that we consider using bootstrap to find the upper α-th quantile $F_{1-\alpha}$ of the cumulative distribution function (CDF) as the critical value. For each data pair $(\mathrm{x}_k, Y_k),\ k = 1, \ldots, n$, we compute the influence diagnostic measure δ_k and save them as $\Delta = (\delta_1, \ldots, \delta_n)$. Randomly sampling n observations from $\Delta = (\delta_1, \ldots, \delta_n)$ with replacement for B times, we obtain the B bootstrap versions of the sample estimate of the diagnostic measures $\Delta^{[1]}, \ldots, \Delta^{[B]}$. For each of the B bootstrap estimator $\Delta^{[b]} = (\delta_1^{[b]}, \ldots, \delta_n^{[b]})$, $b = 1, \ldots, B$, we compute the upper α-th sample quantile of the CDF, noted as $F_{1-\alpha}^{[b]}$. Finally, we obtain the average bootstrap upper α-th sample quantile

$$F_{1-\alpha} = \frac{1}{B}\sum_{b=1}^{B} F_{1-\alpha}^{[b]}. \tag{6}$$

For detecting and removing the influential data points in high dimensional data set $\{(\mathrm{x}_i, Y_i), i = 1, \ldots, n\}$, we formulate this problem as n hypothesis testing problems among the n observations, that is, for $k = 1, \ldots, n$,

$$\mathcal{H}_0^{(k)}:\ \text{the } k\text{−th observation is not influential } \Leftrightarrow\ \mathcal{H}_1^{(k)}:\ \text{the } k\text{−th observation is influential.}$$

Our proposed novel influence diagnosis approach can be summarized by the following algorithm.

Step 1. Compute the proposed high dimensional influence measure δ_k in Equation (5) for any data point (x_k, Y_k) in $\{(\mathrm{x}_i, Y_i), i = 1, \ldots, n\}$.

Step 2. Compare each δ_k with the corresponding $F_{1-\alpha}$ to determine which null hypothesis should be rejected.

Step 3. Those data (x_k, Y_k), whose corresponding $\delta_k > F_{1-\alpha}$, are flagged as influential. Detect and remove them, then obtain a reduced approximate clean data set H.

3.3 Variable Screening Procedure

After obtaining the reduced approximate clean data set H, we apply the DC-SIS screening procedure proposed in Li, *et al.* [12].

Step 4. Compute the sample distance correlation $\hat{\gamma}_j$ between the response y and the variables X_j for $j = 1, \ldots, p$, based on the reduced data set H.

Step 5. Sort all the $|\hat{\gamma}_j|$ with a decreasing order, and filter out those having weak distance correlations.

Toward that end, for a given size $d_n < n$, we can define a submodel

$$\hat{\mathcal{A}} = \{1 \leq j \leq p : \ |\hat{\gamma}_j| \text{ is among the first } [d_n] \text{ largest of all}\}, \tag{7}$$

where $[d_n]$ denotes the integer part of d_n, and d_n is pre-specified cutoff value related to n, such as $d_n = n/\log n$, $d_n = n/3$ and so on.

In the next section, we will evaluate the screening procedure by several criterion. We will consider the minimum model size $\mathcal{S}$, required to include all active predictors. Note that the closer to the true number of active predictor the number $\mathcal{S}$ is, the better the screening procedure. In addition, we take into account the probability that each of the true active variables is selected in a submodel $\hat{\mathcal{A}}$ of a pre-specified size.

4 Numerical Studies

In this section, by using the procedure proposed in Section 3, we assess the finite sample performance and compare it with existing competitors via Monte Carlo simulations. All numerical studies are conducted using R code.

4.1 Simulation Design

We consider the observations (x_i, Y_i) from Equation (4), that is,

$$Y_i = f(\mathrm{x}_i) + \varepsilon_i, \quad \text{for } i = 1, \ldots, n, \tag{8}$$

where Y_i is the i-th response variable, $f(\cdot)$ is a linear or nonlinear unknown link function, $\mathrm{x}_i = (X_{i1}, \ldots, X_{ip})$ is a p-dimensional vector of variables for the i-th observation, and ε_i is a random error.

In practice, the initial model (9) may be influenced by some random perturbation. More specifically, the error term ε_i is of the structure $\varepsilon_i = e_i + \omega_i$, where e_i follows the standard normal distribution and ω_i is a random perturbation variable. Under this setting, the perturbation model can be written as

$$Y_i = f(\mathrm{x}_i) + e_i + \omega_i, \quad i = 1, \ldots, n. \tag{9}$$

This shows that the response Y_i is contaminated by a perturbation variable ω_i. For example, when $\omega_1 \neq 0$, and $\omega_2 = \ldots = \omega_n = 0$, the first observation is deemed to be an influential point especially if ω_1 is related to some other variables that appear in the data set used in the screening procedure but not in the true model.

In the following Examples 4.1-4.3, we fix the sample size $n = 100$ and the dimension of variables $p = 1000$. $\mathrm{x}_i = (X_{i1}, \ldots, X_{ip})$ are generated i.i.d. from a multivariate normal distribution $N(0, \Sigma)$, where $\Sigma = (\sigma_{jl})_{p\times p}$, and $\sigma_{jl} = 0.8^{|j-l|}$ for $j, l = 1, \ldots, p$. The error term e_i is i.i.d. from standard normal distribution $N(0, 1)$. The response of the influential observations

are contaminated by a random perturbation term $\omega_i = \omega\ \mathrm{x}_i\Gamma$, where ω is the parameter that dictates the magnitude of the influential points, and Γ is a column vector of dimension p that only contains 0 and 1. For example, if $\mathrm{x}_i = (X_{i1}, \ldots, X_{ip})$ and only X_{i1}, X_{i2}, X_{i3} are true important variables, then we can set $\Gamma = (0, 0, 0, 1, 1, \ldots, 1)^\top$.

We assume 10% of the total observations as influential, so that $\tilde{n} = 10$ observations are generated from model (10) with $\omega_i \neq 0$ and the rest with $\omega_i = 0$. We set $\omega = 1.2$, which dictates the magnitude of the influential observations. Each of the experiment is repeated 1000 times in our simulation studies. In all the simulation studies, we set the bootstrap time $B = 500$, and upper quantile level $\alpha = 0.05$.

Example 4.1 High dimensional nonlinear models adapted from Li *et al.* [12].

(1.*a*) $Y_i = 3\beta_1 X_{i1} X_{i2} + \beta_2 X_{i15} + 2\beta_3 X_{i30} + e_i,\ i = 1, \ldots, n,$

(1.*b*) $Y_i = 3\beta_1 X_{i1} X_{i2} + \beta_2 \mathbf{1}(X_{i15} < 0) + 3\beta_3 X_{i30} + e_i,\ i = 1, \ldots, n,$

where $\mathbf{1}(X_{i15} < 0)$ is the indicator function, which is nonlinear in X_{15}, and e_i follows the standard normal distribution and is independent of X_{ij}. Following Fan and Lv [5], we choose $\beta_j = (-1)^U(a + |Z|)$ for $j = 1, 2, 3$, where $a = 4\log n/\sqrt{n}$, $U \sim Bernoulli(0.4)$ and $Z \sim N(0, 1)$. We first simulate $n = 100$ i.i.d. observations from this model, and then reset the first $\tilde{n} = 10$ observations from a perturbation model

$$\tilde{Y}_i = Y_i + \omega_i, \quad i = 1, \ldots, \tilde{n},$$

where ω_i is the perturbation term, that is, the first 10 responses are contaminated by a random perturbation $\omega_i = \omega\ \mathrm{x}_i\Gamma$, the $p \times 1$ vector Γ with the 1st, 2nd, 15th, and 30th elements equal to 0, and the rest elements equal to 1.

Example 4.2 Nonparametric additive model adapted from Fan *et al.* [4].

Let $g_1(x) = x$, $g_2(x) = (2x - 1)^2$, $g_3(x) = \sin(2\pi x)/(2 - \sin(2\pi x))$, $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin(2\pi x)^2 + 0.4\cos(2\pi x)^3 + 0.5\sin(2\pi x)^3$. The following model is considered

$$Y_i = 2g_1(X_{i1}) + 6g_2(X_{i2}) + 4g_3(X_{i3}) + g_4(X_{i4}) + e_i, \quad i = 1, \ldots, n.$$

Similar to Example 4.1, we first simulate $n = 100$ i.i.d. observations from this model, and then reset the first $\tilde{n} = 10$ observations from a perturbation model

$$\tilde{Y}_i = Y_i + \omega \mathrm{x}_i\Gamma, \quad i = 1, \ldots, \tilde{n},$$

where $\Gamma = (0, 0, 0, 0, 1, \ldots, 1)^\top$ is a $p \times 1$ vector.

Example 4.3 Heteroskedastic regression model.

$$Y_i = 2X_{i1} + 1.6X_{i2} + 1.2X_{i3} + 0.8X_{i4} + \exp(X_{i20} + X_{i21} + X_{i22}) \cdot e_i, \quad i = 1, \ldots, n.$$

Similar to Example 4.1, we first simulate $n = 100$ i.i.d. observations from this model, and then reset the first $\tilde{n} = 10$ observations from a perturbation model

$$\tilde{Y}_i = Y_i + \omega \mathrm{x}_i\Gamma, \quad i = 1, \ldots, \tilde{n},$$

where Γ is a $p \times 1$ vector with the 1st, 2nd, 3rd, 4th, 20th, 21st, and 22nd elements equal to 0, and the rest elements equal to 1.

4.2 Performance Evaluation

We evaluate the performance through the following three criteria.

1. $\mathcal{S}$: the minimum model size, that is, the smallest number of covariates that we need to include to ensure that all the active variables are selected.

2. $\mathcal{P}_j$: the proportion that an individual active variable X_j is selected for a given model size d_n in the 1000 replications.

3. $\mathcal{P}_a$: the proportion that all active variables are selected for a given model size d_n in the 1000 replications.

Note that we find $\mathcal{S}$ by increasing the model size by 1 at a time until all the active variables used in the data generation process are included. A threshold need not to be specified for $\mathcal{S}$, as the closer to the minimum model size the $\mathcal{S}$ is, the better the screening procedure is. We report the 5%, 25%, 50%, 75% and 95% quantiles of $\mathcal{S}$ out of the 1000 replications. When the threshold d_n is sufficiently large, the proportion $\mathcal{P}_j$s and $\mathcal{P}_a$ are all close to 1. We set the estimated model size d_n to be $d = [3n/\log n] = 65$ throughout our simulations. We compute the proportion that an individual active variable or all active variables are selected in a given model with size $d = 65$ in the 1000 replications. So, we expect that the values of $\mathcal{S}$ reasonably small, and meanwhile, the values of $\mathcal{P}_j$ and $\mathcal{P}_a$ close to 1 in our simulation studies.

For comparison, we first apply the DC-SIS to the full data, and we can get the minimum model size $\mathcal{S}$, the proportion including a single active variable $\mathcal{P}_j$, and the proportion including all active variables $\mathcal{P}_a$. Next, we utilize the influence diagnostic procedure proposed in Section 3 to detect and remove those flagged influential data (x_k, Y_k) with $\delta_k > F_{1-\alpha}$, and obtain a reduced approximate clean data set H by removing those flagged influential points. Finally, we apply the DC-SIS again, but to the reduced data set H, and obtain $\mathcal{S}$, $\mathcal{P}_j$ and $\mathcal{P}_a$, respectively.

Similarly as above, we apply the HIM diagnostic measure from subsection 2.2 to the full data, and obtain a reduced data set by removing those flagged influential points. Then, we apply the DC-SIS approach to the reduced data set and get the corresponding indicators $\mathcal{S}$, $\mathcal{P}_j$ and $\mathcal{P}_a$.

4.3 Simulation Results

The simulation results are reported in Tables 1-4 based on 1000 replications. DCS denotes that we apply the DC-SIS to the full data set; and HIM-DCS denotes that we remove influential observations by HIM procedure and then apply the DCS to the reduced data set; R-DCS denotes that we first remove influential observations by our proposed new influence diagnostic procedure, and then apply the DC-SIS to the reduced data set.

Table 1: Simulation results for Examples 4.1a-4.1b, the 5% , 25% , 50% , 75% , 95% quantiles of the minimum model size $\mathcal{S}$ out of 1000 replicates. The number of true active variables is 4.

$\mathcal{S}$	Example 4.1a					Example 4.1b				
Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
DCS	12	49	120	302	690	169	397	603	797	951
HIM-DCS	9	28	86	263	710	30	158	413	684	937
R-DCS	7	16	40	144	629	31	110	348	686	899

Table 2: Simulation results for Examples 4.2-4.3, the 5%, 25%, 50%, 75%, 95% quantiles of the minimum model size $\mathcal{S}$ out of 1000 replicates. The number of true active variables is 4 in Example 4.2 and 7 in Example 4.3.

$\mathcal{S}$	Example 4.2					Example 4.3				
Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
DCS	4	9	31	108	457	156	390	580	768	939
HIM-DCS	4	5	15	93	373	19	61	154	338	632
R-DCS	4	4	11	71	304	10	26	70	221	579

Table 3: Simulation results for Examples 4.1a-4.1b. $\mathcal{P}_j$ denotes the proportion of replicates when an individual active variable X_j was included and $\mathcal{P}_a$ denotes the proportion of iterations when all active variables were included in the submodel.

$\mathcal{P}_j,\ \mathcal{P}_a$	Example 4.1a					Example 4.1b				
Method	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_{15}$	$\mathcal{P}_{30}$	$\mathcal{P}_a$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_{15}$	$\mathcal{P}_{30}$	$\mathcal{P}_a$
DCS	0.85	0.85	0.42	0.96	0.33	0.70	0.68	0.11	0.07	0.00
HIM-DCS	0.94	0.95	0.49	0.97	0.43	0.86	0.87	0.13	0.99	0.11
R-DCS	0.99	0.98	0.62	1.00	0.61	0.93	0.92	0.15	1.00	0.13

Table 4: Simulation results for Examples 4.2-4.3. $\mathcal{P}_j$ denotes the proportion of replicates when an individual active variable X_j was included and $\mathcal{P}_a$ denotes the proportion of iterations when all active variables were included in the submodel.

$\mathcal{P}_j,\ \mathcal{P}_a$	Example 4.2					Example 4.3							
Method	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_a$	$\mathcal{P}_1$	$\mathcal{P}_2$	$\mathcal{P}_3$	$\mathcal{P}_4$	$\mathcal{P}_{20}$	$\mathcal{P}_{21}$	$\mathcal{P}_{22}$	$\mathcal{P}_a$
DCS	0.93	1.00	0.94	0.67	0.65	0.37	0.30	0.22	0.21	0.99	0.99	0.98	0.00
HIM-DCS	0.95	1.00	0.96	0.76	0.74	0.59	0.64	0.59	0.47	0.88	0.93	0.88	0.26
R-DCS	0.99	1.00	0.99	0.80	0.79	0.99	1.00	0.99	0.97	0.66	0.71	0.66	0.48

1. In Tables 1-2, the 5%, 25%, 50%, 75% and 95% percentiles of the minimum model size $\mathcal{S}$ are summarized for various models. Note that the data set are contaminated with 10% influential points. The results of the HIM-DCS and R-DCS are smaller than those obtained by DCS approach, which implies that the HIM-DCS and R-DCS approach perform better than the DCS procedure in the presence of influential points. Note also that the relationship between the response and the covariate is nonlinear, the performance of the R-DCS outperforms HIM-DCS significantly in all the related models.

2. Tables 3-4 summarizes $\mathcal{P}_j$ and $\mathcal{P}_a$ for a given model with size $d_n = 65$ in the 1000 replications. From the results of $\mathcal{P}_j$s and $\mathcal{P}_a$ in various models, we can see that the R-DCS approach outperforms DCS and HIM-DCS approach significantly in most cases, as the corresponding $\mathcal{P}_j$s and $\mathcal{P}_a$ are greater than the results from the other two approaches.

3. Tables 1-4 illustrate that the accuracy of variable screening can be affected by influential points. The R-DCS approach performs better, as the values of $\mathcal{S}$ reasonably small, and meanwhile, the values of $\mathcal{P}_j$ and $\mathcal{P}_a$ close to 1. This suggests that the proposed method improves DCS by removing the influential observations. Similarly, the HIM-DCS can perform well to reduce the effect of influential points. However, it is still outperformed by our proposed R-DCS, for most of the examples referred in this paper.

To summarize, all the simulation results indicate that the presence of influential observations will seriously affect the accuracy of variable screening. The proposed method R-DCS is able to minimize the effect of influential observations. Comparing R-DCS with HIM-DCS, the results show that the proposed method R-DCS is superior to HIM-DCS for variable screening in various models. This clearly confirm that the proposed novel influence diagnostic procedure can help control the adverse effect after detecting and removing those influential observations.

5 A Real Data Analysis

As an application illustration, we apply the proposed influence diagnosis and variable screening method to the analysis of Cardiomyopathy microarray data.

Example 5.1 Cardiomyopathy microarray data.

The cardiomyopathy microarray data are from a transgenic mouse model of dilated cardiomyopathy (Redfern *et al.* [15]). This data set has attracted considerable attention and been systematically investigated by many researchers. Examples include Segal *et al.* [16], Hall and Miller[9], Li *et al.* [12] and Li *et al.* [11].

This data set consists of an outcome measure y $= (Y_1, \ldots, Y_n)^\top$, and an $n \times p$ matrix of gene expression values X $= (X_{ij})_{n\times p}$, for $i = 1, \ldots, n(= 30)$, $j = 1, \ldots, p(= 6319)$, where Y_i is the Ro1 expression level, measured for $n = 30$ specimens, and X_{ij} denotes the expression level of the j-th gene for the i-th mouse. Note that $p \gg n$, so this is a high dimensional data analysis problem. Our aim is to determine the most relevant genes for overexpression of a G protein-coupled receptor (Ro1) in mice.

Hall and Miller [9] showed that both genes Msa.2877.0 and Msa.1166.0 are particularly important using the generalized correlation. They found Msa.2877.0 has an essentially linear relationship, and Msa.1166.0 has strong correlation of -0.75 with Msa.2877.0. Li *et al.* [12] used the DC-SIS procedure that ranks two genes, Msa.2134.0 and Msa.2877.0, at the top. They showed that their DC-SIS procedure achieves better performance in contrast to the generalized correlation ranking method of Hall and Miller [9]. Li *et al.* [11] showed that Msa.1166.0 and

Msa.7019.0 are particularly important using the robust rank correlation based screening (RRCS) procedure. Figure 1 indicates the scatter plots and corresponding cubic-spline fit curves of the relationship between the important genes (Msa.2877.0, Msa.1166.0, Msa.2134.0, Msa.7019.0) and the outcome (Ro1).

A natural question arises before we analyze this data: whether the data come from a clean set? In other words, whether the $n = 30$ specimens are contaminated with a proportion of influential data. Further, we believe, based on our previous simulation studies, that the presence of influential data may significantly affect the accuracy of variable screening. Then, we apply the proposed influential diagnosis and variable screening approach to display the important genes related to outcome (Ro1) measure Y_i.

Firstly, for each data pair (x_i, Y_i), $i = 1, \ldots, 30$, we obtain the influence diagnostic measure δ_i defined in Equation (6). We calculate the bootstrap upper 0.05 quantile of the C.D.F., denoted as $F_{0.95}$, detect and remove three flagged influential specimens named eight3054f, eight3067f, eight3081f, which $\delta_i > F_{0.95}$ from the 30 specimens, and get a reduced data set with the remaining 27 specimens. Finally, we apply the DC-SIS procedure, which ranks two genes labeled Msa.2134.0 and Msa.28021.0 at the top. We find that besides Msa.2134.0, Msa.28021.0 also seems to be an important gene related to Ro1, possibly due to the detection and removing the three influential observations.

Figure 2 indicates the scatter plots and corresponding cubic-spline fit curves for the relationship between the important genes (Msa.2134.0, Msa.28021.0) and the outcome (Ro1) based on $n = 30$ specimens (solid lines) and $n = 27$ specimens (dotted lines), respectively. The three solid circles represent the flagged influential specimens eight3054f, eight3067f and eight3081f. From Figure 2, based on $n = 30$ specimens, it can be seen that the fit curve between Msa.2134.0 and Ro1 shows an 'S' type, while the fit curve between Msa.28021.0 and Ro1 shows an 'M' type. Even if these two curves are approximated by polynomial regression, the degree will be larger than two. Therefore, these two curves exhibit clear nonlinear behaviors. Note that the distance correlation has the advantage that it can detect nonlinear relationships which are ignored by marginal correlation. Our proposed R-DCS procedure shows the advantage that it detected two important genes having nonlinear relationships with Ro1, which might be ignored by some other methods due to influential observations in the sample. Figure 2 also indicates the cubic-spline fit curves after removing the three flagged influential specimens (represented by solid circles). It can be seen that after the three flagged influential specimens are removed, the curves will be pulled downwards significantly. Therefore, the three flagged influential specimens caused higher variability in the response variable, which would explain the substantial gain of our proposed R-DCS procedure.

To assess the performance of the proposed procedure, we further fit the following additive model:

$$Y = g_1(X_j) + g_2(X_k) + e,$$

where X_j and X_k are respectively the top two genes, Msa.2134.0 and Msa.28021.0, $g_1(\cdot)$ and $g_2(\cdot)$ are two unknown link functions, e is an error term. We fit g_1 and g_2 by using the 'gam' function in the R 'mgcv' package, where 'gam' can be used to fit a generalized additive model (GAM) to data. We also measure the performance of goodness of fit by the adjusted R^2 values and the explained deviance, where deviance implies the proportion of the null deviance explained by the proposed model, with a larger value indicating better performance.

Whether removing the three influential observations eight3054f, eight3067f, eight3081f or not may affect the results of R^2 values and the deviance. The DCS method has a performance with the R^2 value of 0.736 and the deviance of 80.4% based on 30 observations. However, when we remove the three influential observations, the R-DCS method has a better performance with the R^2 value of 0.78 and the deviance of 81.4%. So, we can see that, the R-DCS procedure

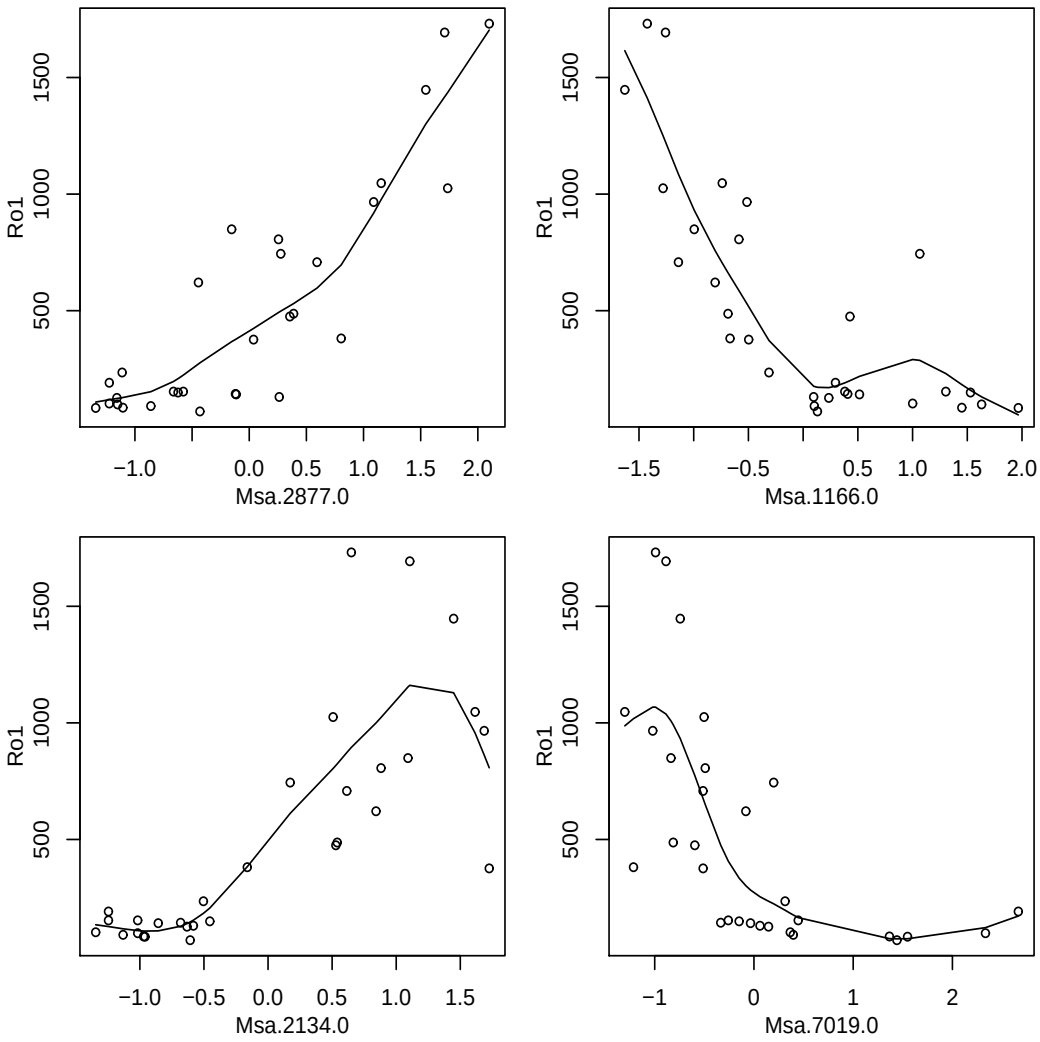

Figure 1: The scatter plots and corresponding cubic-spline fit curves of the relationship between the important genes (Msa.2877.0, Msa.1166.0, Msa.2134.0, Msa.7019.0) and the outcome (Ro1) based on $n = 30$ specimens.

clearly achieves a better performance.

6 Conclusion

In this paper, we considered the problem of influence diagnosis and variable screening in high dimensional regression. We defined a new high dimensional influence measure, and then proposed an influence diagnostic procedure. The dCor based influence measure can be more effective than the marginal correlation based influence measure in the presence of nonlinear relationship. We further utilize the DC-SIS procedure to verify the effectiveness of the proposed procedure.

The proposed procedure has several appealing properties. First, the new influence measure is easy to compute. It is motivated by the leave-one-out idea, but it is based on the dCor between the response and all predictor variables. Second, the new influence diagnosis and variable screening procedure based on dCor are more effective in the presence of nonlinear relationship between the response and all predictor variables. It is robust as 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 method can greatly reduce the adverse effect after detecting and removing those influential data points, and can improve the filtering accuracy in variable screening problem.

Acknowledgements

The authors are grateful to the editor and two 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, 11371202, 11431006 and 11131002), the Fund for the Doctoral Program of Kashgar University (grant 14-2498) and the State Scholarship Fund of China Scholarship Council.

References

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

[2] Cui H., Li R., Zhong W. (2015). Model-free Feature Screening for Ultrahigh Dimensional Discriminant Analysis. *Journal of the American Statistical Association.* **110(510)**, 630-641.

[3] Efron B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics.* **7(1)**, 1-26.

[4] Fan J., Feng Y., Song R. (2011). Nonparametric Independence Screening in Sparse Ultra-high Dimensional Additive Models. *Journal of the American Statistical Association.* **106(494)**, 544-557.

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

[6] Fan J., Ma Y., Dai W. (2014). Nonparametric Independence Screening in Sparse Ultra-high Dimensional Varying Coefficient Models. *Journal of the American Statistical Association.* **109(507)**, 1270-1284.

[7] Fan J., Samworth R., Wu Y. (2009). Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. *Journal of Machine Learning Research.* **10(5)**, 2013-2038.

[8] Fan, J., Song, R. (2010). Sure Independence Screening in Generalized Linear Models with NP-dimensionality. *Annals of Statistics.* **38(6)**, 3567-3604.

[9] Hall P., Miller H. (2009). Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems. *Journal of Computational and Graphical Statistics.* **18(3)**, 533-550.

[10] He X., Wang L., Hong H. (2013). Quantile-adaptive Model-free Variable Screening for High-dimensional Heterogeneous Data. *Annals of Statistics.* **41(1)**, 342-369.

[11] Li G., Peng H., Zhang J., Zhu L. (2012). Robust Rank Correlation Based Screening. *Annals of Statistics.* **40(3)**,1846-1877.

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

[13] Liu J., Li R., Wu R. (2014). Feature Selection for Varying Coefficient Models with Ultrahigh-Dimensional Covariates. *Journal of the American Statistical Association.* **109(505)**, 266-274.

[14] Mai Q., Zou H. (2015). The Fused Kolmogorov Filter: A Nonparametric Model-Free Screening Method. *Annals of Statistics.* **43(4)**, 1471-1497.

[15] Redfern C. H., Coward P., Degtyarev M. Y., Lee E. K., Kwa A. T., Hennighausen L., Bujard H., Fishman G. I., Conklin, B. R. (1999). Conditional Expression and Signaling of a Specifically Designed Gi-coupled Receptor in Transgenic Mice. *Nature Biotechnology.* **17(2)**, 165-169.

[16] Segal M. R., Dahlquist K. D., Conklin B. R. (2003). Regression Approaches for Microarray Data Analysis. *Journal of Computational Biology.* **10(6)**, 961-980.

[17] Székely G. J., Rizzo M. L. (2009). Brownian Distance Covariance. *Annals of Applied Statistics.* **3(4)**, 1236-1265.

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

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

[20] Zhu H., Ibrahim J. G., Cho H. (2012). Perturbation and Scaled Cook's Distance. *Annals of Statistics.* **40(2)**, 785-811.

[21] Zhu L. P., Li L., Li R., Zhu L. X. (2011). Model-Free Feature Screening for Ultrahigh Dimensional Data. *Journal of the American Statistical Association.* **106(496)**, 1464-1475.

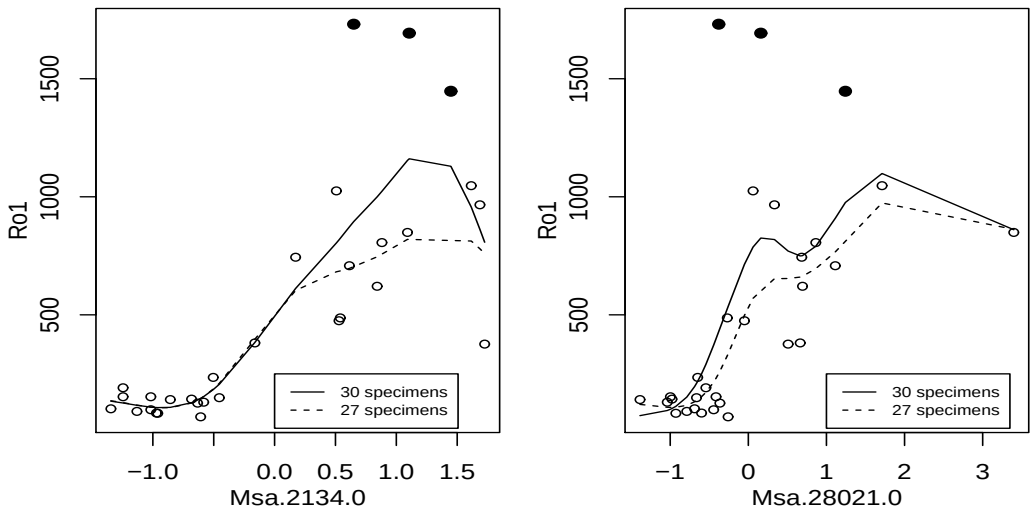

Figure 2: The scatter plots and corresponding cubic-spline fit curves for the relationship between the important genes (Msa.2134.0, Msa.28021.0) and the outcome (Ro1) based on $n = 30$ specimens (solid lines) and $n = 27$ specimens (dotted lines), respectively. The three solid circles represent the flagged influential specimens.