Statistics and Its Interface Volume 11 (2018) 473–489

Robust model-free feature screening based on modified Hoeffding measure for ultra-high dimensional data

Yuan Yu*, Di He†, and Yong Zhou‡,§

Sure independence screening (SIS) has become a cutting-edge dimension reduction technique to extract important features from ultrahigh-dimensional data in statistical learning. Many of the screening methods are developed to be suitable for special models that follow certain assumptions. With the availability of more data types and complicated models, a robust model-free procedure with less restrictive conditions of data is required. In this paper, we propose a modified Hoeffding measure which efficiently characterize the dependence between two random variables. The modified Hoeffding measure is between 0 and 1, and zero if and only if the two variables are independent under some mild conditions. This property enables us to propose a novel feature screening procedure based on it without specifying the regression structure. The proposed method is robust for both the predictors and response with the heavy-tailed data and outliers, and suitable for complex data including discrete and multivariate variables. In addition, it can extract important features even when the underlying model is complicated. We further establish the sure screening property and ranking consistency property even when the dimensionality is an exponential order of the sample size without assuming any moment condition on the predictors and response. Simulations and an analysis of real data demonstrate the versatility and practicability of the proposed method in comparison with other state-of-the-art approaches.

AMS 2000 subject classifications: 62E99, 62G05, 62G35, 62H20, 62P10.
Keywords and phrases: Feature screening, Hoeffding measure, Ranking consistency property, Robustness, Sure screening property, Ultrahigh-dimensional data.

*Yu's work was supported by Graduate Innovation Foundation of Shanghai University of Finance and Economics, China (2015110758).
†He's work was supported by Graduate Innovation Foundation of Shanghai University of Finance and Economics, China (CXJJ-2014-452).
‡Zhou's work was supported by the State Key Program of National Natural Science Foundation of China (71331006), the State Key Program in the Major Research Plan of National Natural Science Foundation of China (91546202).
§Corresponding author.

1. INTRODUCTION

The ultrahigh dimensional data have been frequently encountered in contemporary scientific research, where the dimension p can grow exponentially with the sample size n. [8] pointed out that analyzing such data poses simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability to modern statistical inference. To overcome the issues associated with ultra-high dimensionality, many marginal screening techniques, such as the sure independence screening (SIS) procedure [6], have been shown to filter out many uninformative variables in many scenarios. The key idea of the SIS procedure is to rank all predictors by using a utility measure between the response and each predictors and then to retain the top variables for further investigation. A desired marginal screening procedure possesses the sure screening property; that is, with probability close to one, all the important variables would survive after variable screening.

This seminal idea has motivated many methods in the recent literature. [12] used the generalized Pearson correlation as the ranking index to identify influential but not explicit features of a predictive model. [8] and [9] developed feature screening procedures for generalized linear models, which used the maximum marginal likelihood. [5] developed nonparametric SIS procedure for ultrahigh dimensional additive models. [31] proposed a screening method based on standardized marginal maximum likelihood estimator for the Cox model. [7] and [22] considered nonparametric feature screening for sparse ultrahigh dimensional varying coefficient models. By using the inverse weighted probability method, [27] proposed a robust feature screening procedure for the censored ultrahigh dimensional data for the transformation models. For the varying coefficient model about the longitudinal data, [28] developed a nonparametric feature screening.

The aforementioned feature screening procedures are all for some specified models. However, it would be very difficult to identify a correct model initially in the process of analysis for the ultrahigh dimensional data. To avoid the model misspecification problem, [33] advocated a model-free feature screening procedure using the expectation of the square of the correlation between the predictor and an indicator

function of the response. For the model-free feature screening method, there is some work done using more sophisticated dependence measures in recent years. [20] suggested to use the distance correlation as the ranking index for feature screening. [19] proposed to use the Kendall's tau to do the feature screening for generalized transformation model. [23] proposed the Kolmogorov filter for variable screening in high-dimensional binary classification. [13] proposed a quantile-adaptive model-free SIS for ultrahigh dimensional heterogeneous data. Similar to distance correlation, [26] proposed martingale difference divergence as a new measure for correlation and used it as the marginal utility for feature screening. For the ultrahigh dimensional data with categorical predictors and categorical responses, [17] proposed to use Pearson's chi-square statistic to screening variables. By proposing a mean-variance index, [2] developed a model-free feature screening for ultrahigh dimensional discriminant analysis where the categorical response can have a diverging number of classes.

Although the innovative model-free sure screening methods can be applied to more general parametric or semi-parametric models, it requires some restrictive conditions for predictor variables such as moment and linearity conditions. Moreover, [21] noted that the index used by [33] maybe zero even if there are some relationships between response and predictor variables. As for distance correlation used by [20], it is zero if and only if there no relationship between variables, but it requires moment conditions for both response and predictor variables, which is not robust for variables with heavy tails or extreme values. The rank correlation screening proposed by [19] can be robust and invariant for monotonic transformation; however, its implement is based on Kendall's tau, which can be zero even if there is an association between variables and can be applied only to measure the monotonic association. Furthermore, this method requires many restrictive conditions. Based on the above discussion, our work aims to develop a new robust model-free feature screening procedure for ultrahigh dimensional reduction problem.

In this paper, we propose a new robust model-free feature screening based on modified Hoeffding measure. Hoeffding's D measure proposed by [14] is to test the independence of the data sets by calculating the distance between the joint distribution and the product of the marginal distributions. It is a nonparametric measure of association that may identify more general types of dependence.

[10] and [3] made comparisons between the Hoeffding's measure and other dependence measures in the study of gene identification. They found that the Hoeffding's measure can effectively identify non-functional associations. More useful details of Hoeffding measure can be found in [16]. However, the Hoeffding measure has a severe drawback, that it may be zero even if there is an association between variables. Due to this reason, we introduce a new measure by modifying the Hoeffding's measure and give its some appealing properties, of which most important is that this new measure equals zero if and only if the two variables are independent under some mild condition. Based on this modified Hoeffding measure, the feature screening procedure does not require any restrictive conditions such as the moment condition for both response and predictor variables. So it is robust for both of them with the heavy-tailed data and extreme values. Also, we do not need to pre-specify a regression model to implement this procedure, which leads to the model-free properties of our method. On the other hand, our method can be applied to any random variables without knowing the distribution of the response and predictor variables in advance. Therefore, our method is also suitable for discrete random variables. Thus, our method extends the method of [2] greatly. In addition, we can also use our method to handle the grouped predictors or multivariate responses. Furthermore, we prove that the proposed method possesses both the sure screening property and ranking consistency property theoretically under weak conditions.

The remainder of this article is organized as follows: In section 2, we give some reviews of the Hoeffding measure and introduce our new modified Hoeffding measure and its properties, and then we propose our new robust model-free feature screening procedure. In section 3, we establish the theoretical properties of our proposed procedure. In section 4, we examine the finite sample performance of the proposed procedure through comprehensive Monte Carlo simulation studies and empirical analyses of real data examples. All technical proofs are relegated to the Appendix.

2. THE ROBUST MODEL-FREE FEATURE SCREENING

Most popular measures of dependence such as Pearson correlation coefficient, Kendall's tau and Spearman coefficient, may be zero even if there is an association between X and Y, so screening methods based on them may fail to identify a generalized association between the response and predictors. Although the distance correlation [29] has the required merits, i.e. $dcorr(X,Y) = 0 \Leftrightarrow X \perp Y$, [20] have shown that its implementation requires some moment conditions for both predictors and response, which will lead to its sensitivity to the non-normal data. Next, we will introduce another consistent dependence measure for our new correlation screening procedure.

2.1 Modified Hoeffding measure

Feature screening procedures intend to rank the importance of each predictor through its corresponding marginal correlation with the response, and select the predictors highly correlated with the response. Hence, SIS procedures are equivalent to implement test of independence step by step. For this reason, we attempt to use the consistent test statistics of independence to construct our feature screening index. The best known of these are those introduced by [14] and [1].

Definition 2.1. *With $F_{X,Y}(x,y)$ the joint distribution function of (X,Y), and $F_X(x)$ and $F_Y(y)$ the marginal distribution functions of X and Y, respectively, Hoeffding measure is given as*

(2.1)
$$D(X,Y) = \int [F_{X,Y}(x,y) - F_X(x)F_Y(y)]^2 dF_{X,Y}(x,y).$$

Apparently, Hoeffding's coefficient is non-negative with equality to zero under independence. [14] gave its nonparametric U-statistics estimation as below:

$$D_n = \frac{Q - 2(n-2)R + (n-2)(n-3)S}{n(n-1)(n-2)(n-3)(n-4)}, \tag{2.2}$$

where $Q = \sum_{i=1}^n (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2), R = \sum_{i=1}^n (R_i - 2)(S_i - 2)c_i, S = \sum_{i=1}^n (c_i - 1)c_i$, R_i and S_i being the respective ranks of X_i among X's and Y_i among Y's, and c_i is the number of bivariate observations (X_j, Y_j) for which $X_j \le X_i$ and $Y_j \le Y_i$. This statistics can be computed by R function hoeffd (package Hmisc).

[1] used the empirical distribution functions to estimate the Hoeffdimg coefficient as:

$$B_n = \int \left[\widehat{F}_{X,Y}(x,y) - \widehat{F}_X(x)\widehat{F}_Y(y)\right]^2 d\widehat{F}_{X,Y}(x,y), \tag{2.3}$$

where, $\widehat{F}_{X,Y}(x,y), \widehat{F}_X(x), \widehat{F}_Y(y)$ are the empirical distribution functions of $F_{X,Y}(x,y), F_X(x), F_Y(y)$ respectively.

However, the Hoeffding measure has a severe drawback, that it may be zero even if there is an association between X and Y. A counter-example is given by [14]. The reason behind this is the problem caused by the definition of Hoeffding measure itself. From its definition, even if random variables X and Y are dependent, that is $F_{X,Y}(x,y) - F_X(x)F_Y(y) \neq 0$ for some x, y, the probability density or mass function $dF_{X,Y}(x,y)$ can still be zero for these x, y. This will cause the Hoeffiding measure to be zero. Due to this reason, we modify this measure as follows.

Definition 2.2. *The modified Hoeffding measure is defined as:*

(2.4)
$$\omega(X,Y) = \int [F_{X,Y}(x,y) - F_X(x)F_Y(y)]^2 dF_X(x) dF_Y(y).$$

This new coefficient can be thought as the Cramér-von Mises distances between the joint distribution function and the product of marginal distribution functions. The following proposition provides a appealing property of the modified Hoeffding measure.

Proposition 2.1. *When (X,Y) belongs to Ω, it holds true that $\omega(X,Y) \geqslant 0$, with equality if and only if X and Y are independent. Where Ω is the class of bivariate random vector, whose marginal distribution is discrete or continuous, or a mixture of the two, that is, assume there exists probability density function $f_X(x), f_Y(y)$ and mass function $\widetilde{f_X}(x), \widetilde{f_Y}(y)$ such that $P(X < x) = \sum_{u_i < x} \widetilde{f_X}(u_i) + \int_{u<x} f_X(u)du$, $P(Y < y) = \sum_{v_i < y} \widetilde{f_Y}(v_i) + \int_{v<y} f_Y(v)dv$.*

Remark 2.1. *Although we restrict (X,Y) to the class Ω in the proposition, we conjecture that the condition can expand to more general situation, that is, for most of bivariate random vector the conclusion also holds. This remarkable property motivates us to use the modified Hoeffding measure as a marginal utility for SIS.*

Definition 2.3. *Assume $\{(X_i, Y_i), i = 1, \cdots, n\}$ is a simple random sample of size n of the random vector (X,Y). Then, the empirical distribution functions(e.d.f) of $F_X(x), F_Y(y)$ and $F_{X,Y}(x,y)$ are defined as $\widehat{F}_X(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x), \widehat{F}_Y(y) = \frac{1}{n}\sum_{i=1}^n I(Y_i \le y)$ and $\widehat{F}_{X,Y}(x,y) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x, Y_i \le y)$, respectively, where $I(\cdot)$ is the indicator function. The empirical modified Hoeffding measure ω is defined as follows:*

$$\begin{aligned}\widehat{\omega}(X,Y) &= \int \left[\widehat{F}_{X,Y}(x,y) - \widehat{F}_X(x)\widehat{F}_Y(y)\right]^2 d\widehat{F}_X(x) d\widehat{F}_Y(y) \\ &= \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \left[\widehat{F}_{X,Y}(X_i, Y_j) - \widehat{F}_X(X_i)\widehat{F}_Y(Y_j)\right]^2. \end{aligned} \tag{2.5}$$

Remark 2.2. *In the definition of $\widehat{\omega}$, we only use the empirical distribution function to estimate ω without any tuning parameter, so it is easy to compute and robust to the presence of heavy tails and extreme values for both the two variables. On the other hand, the e.d.f can estimate any distribution functions, so $\widehat{\omega}$ can be applied to estimate $\omega(X,Y)$ for most type variables. Moreover, the c.d.f and e.d.f are also suitable for multi-variate case, we can therefore use ω and $\widehat{\omega}$ to measure the dependence between multi-variables.*

Here, we use a simple simulation example to have an insight into the modified Hoeffding measure ω. Suppose (X,Y) follows a standard bivariate normal distribution with correlation coefficient ρ. For every ρ equally spaced by 0.2 in $[-1,1]$, we generate a sample of size 50 and calculate $\hat{\omega}(X,Y)$. We run the simulation 200 times. Panel (a) and (b) in Figure 1 are boxplot and average plot of $\hat{\omega}(X,Y)$ against ρ, respectively. It is shown that $\omega(X,Y)$ is a strictly increasing function of $|\rho|$ in the sample level.

Similar to the Pearson or Distance correlation coefficient, we may need to standardize the modified Hoeffding measure. However, the following proposition implies that we don't need to standardize it by using this measure to the screening.

Proposition 2.2. *For any random variable X, $\omega(X,X) = \frac{1}{90}$, which means that the modified Hoeffding measure between the same random variables is a constant.*

2.2 Screening procedure using modified Hoeffding measure

Now, we propose a new robust model-free screening procedure using modified Hoeffding measure for ultra-high di-

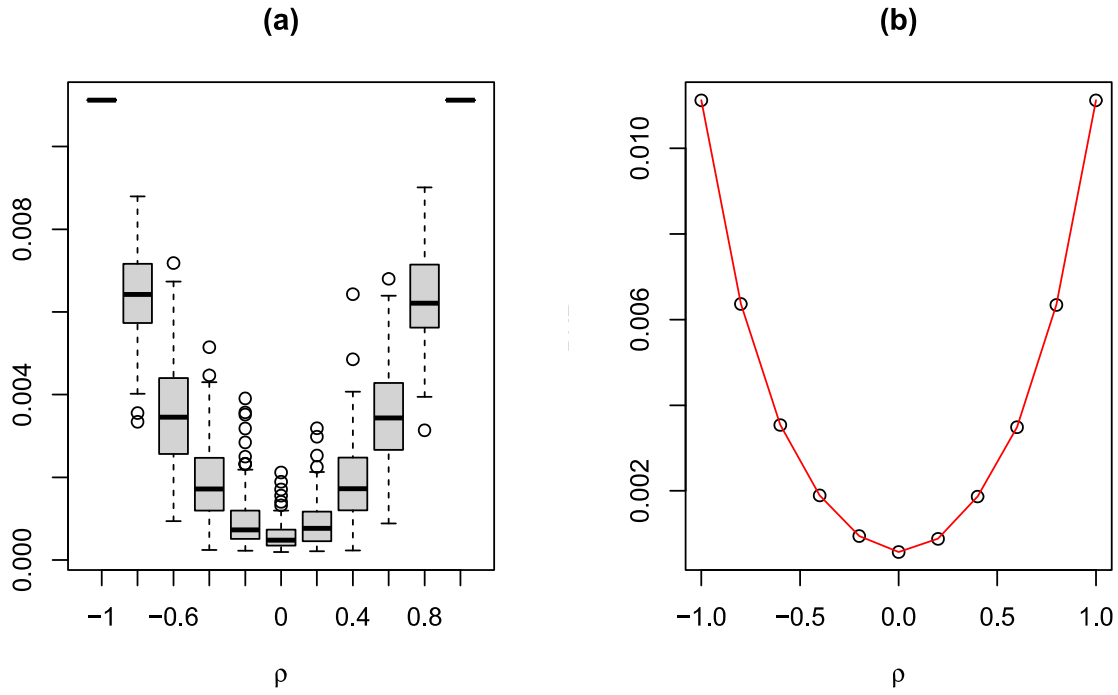

Figure 1. The plot of $\hat{\omega}$ against ρ under the bivariate normal distribution: (a) boxplot of $\hat{\omega}$ against ρ, (b) plot of average $\hat{\omega}$ against ρ.

mensional analysis. Let Y be the response variable with support Ψ_y, and $\mathbf{x} = (X_1, \cdots, X_p)^T$ be the p dimensional predictors, where, p can be much larger than the sample size n, and Y can be both discrete and continuous. Since our method can be applied to more general cases without specifying a concrete regression model, hence, we define the index set of active and inactive predictor set in the terminology of [2] as follows:

$$\begin{aligned}\mathcal{A} &= \{k : F(y|\mathbf{x}) \text{ functionally depends on } X_k \\ &\qquad \text{for some } y \in \Psi_y\}, \\ \mathcal{I} &= \{k : F(y|\mathbf{x}) \text{ does not functionally depends on } X_k \\ &\qquad \text{for any } y \in \Psi_y\}.\end{aligned}$$

We write $\mathbf{x}_{\mathcal{A}} = \{X_k : k \in \mathcal{A}\}$ and $\mathbf{x}_{\mathcal{I}} = \{X_k : k \in \mathcal{I}\}$, and refer $\mathbf{x}_{\mathcal{A}}$ to an active predictor vector and its complement $\mathbf{x}_{\mathcal{I}}$ to an inactive predictor vector. Under the sparsity condition, it implies that

$$F(y|\mathbf{x}) = F(y|\mathbf{x}_{\mathcal{A}}), \quad \text{for } y \in \Psi_y \text{ and } \mathbf{x} \in \mathbb{R}^p.$$

The definition of $\mathcal{A}$ and $\mathcal{I}$ also indicates that $\mathbf{x}$ and Y are statistically independent when $\mathbf{x}_{\mathcal{A}}$ is given, that is the inactive predictors are needless for response when the active predictors are given.

The primary goal of SIS, first proposed by [6], is to use a marginal utility to rank the importance of predictors in the ultra-high dimensional analysis, which can reduce the model to a moderate scale and the shrunken model should almost contain $\mathcal{A}$. Following the literature of SIS, we will use the modified Hoeffding measure

$$\omega_k = \omega(X_k, Y), \quad k = 1, \cdots, p$$

as a marginal utility to measure the dependence between X_k and Y for each pair (X_k, Y). In practice, if we get a random sample $\{(\mathbf{x}_i, Y_i), i = 1, \cdots, n\}$ of $(\mathbf{x}, Y)$, we can estimate ω_k as (2.3), that is

$$\hat{\omega}_k = \hat{\omega}(X_k, Y), \quad k = 1, \cdots, p.$$

Then, we propose to rank the importance of $X_k, k = 1, \cdots, p$, through the estimated value $\hat{\omega_k}$. With a prespecified threshold τ_n, we define the selected submodel as:

$$\hat{\mathcal{A}}_{\tau_n} = \{1 \leq k \leq p : \hat{\omega}_k \geq \tau_n\}.$$

In practice, this is equivalent to selecting the first $d_n \leqslant n$ predictors according to the ranked $\hat{\omega}_k$, i.e. we will select the submodel as:

$$\widehat{\mathcal{A}}^*_{d_n} = \{k : \hat{\omega}_k \text{ is among the first } d_n \text{ largest of all}\}.$$

We refer this new robust model-free feature screening as ROM-SIS thereafter. From the appealing property of modified Hoeffding measure, we expect the ROM-SIS to have the following attractive advantages in comparison to other existing model-free SIS procedures:

(A1) In the definition of ω_k and $\hat{\omega}_k$, we only use the c.d.f and e.d.f for both response and predictors, so it is robust for both of them with the heavy-tailed data and extreme values.

(A2) Similar to Hoeffding measure, the modified Hoeffding measure ω_k can also measure non-monotonic associations between X_k and Y. Hence, our method can identify most of the influential predictors for more flexible models. To implement the ROM-SIS, we do not need to pre-specify a regression model, which leads to the model free property of our method.

(A3) As we mentioned in Remark 2.2, the computation of $\hat{\omega}_k$ can be applied to any random variables without knowing the distribution of them in advance. Therefore, our method is also suitable for discrete random variables. In addition, we can also use our method to handle the grouped predictors or multivariate responses like DC-SIS proposed by [20], see simulation example 5.

3. THEORETICAL PROPERTIES

In this section, we will establish the theoretical properties of our proposed method for ultra-high dimensional feature space. We mainly expound that in two aspects: one is the sure screening property, the other is the ranking consistency property.

3.1 Sure screening property

First of all, we establish the sure screening property of the proposed ROM-SIS in the term of [6]. As our procedure is based on ω_k, which is estimated by e.d.f and their corresponding sample counterparts, therefore, $\hat{\omega}_k$ is constructed by the bounded elements so that we don't need any moment conditions for both response variables and predictors. As any other sure independent screening methods, we need the following condition:

(C1) $\min_{k\in\mathcal{A}} \omega_k \geq c_0 n^{-\kappa}$, for some $0 \leq \kappa < \frac{1}{2}$ and positive constant c_0.

Condition (C1) is the requirement for the minimal signal of the true active predictors. It supposes that the minimal true signal should reach a certain level to guarantee the effectiveness of sure independent screening. Specifically, we need its order to be $n^{-\kappa}$ which allows the minimal predictor signal to be very weak when the sample size is sufficient large. This assumption is very common in the literature of SIS such as [6] Condition 3, [9] Condition E, [5] Condition C, [20] Condition C2, etc. Now we give the sure screening property of ROM-SIS as the follows:

Theorem 3.1 (Sure screening property)**.** *For any positive constant c and $0 < \kappa < \frac{1}{2}$, there exist positive constants $c_2 < c_1 < c_3$, such that*

$$P(\max_{1\leq k\leq p} |\hat{\omega}_k - \omega_k| \geq cn^{-\kappa}) \leq p\Big\{2(n+1)\exp\{-c_1 n^{1-2\kappa}\} + 4\exp\{-c_2 n^{1-2\kappa}\} + 4\exp\{-c_3 n^{1-2\kappa}\}\Big\}.$$

In addition, if Condition (C1) holds and we set $\tau_n = c_4 n^{-\kappa}$ with $c_4 \leq \frac{c_0}{2}$, we have

$$P(\mathcal{A} \subset \widehat{\mathcal{A}}_{\tau_n}) \geq 1 - s\Big\{2(n+1)\exp\{-c_1 n^{1-2\kappa}\} + 4\exp\{-c_2 n^{1-2\kappa}\} + 4\exp\{-c_3 n^{1-2\kappa}\}\Big\},$$

where s is the cardinality of $\mathcal{A}$.

Theorem 3.1 implies that we can obtain the sure screening property for ROM-SIS only under the minimal signal condition, which is much weaker than those like SIS ([6]), SIRS ([33]), DC-SIS ([20]), and RRCS ([19]) in the sense that we do not need any assumptions on the moments of both response variables and predictors. In this respect, we can also guarantee the advantages of (A1) theoretically. Moreover, from Theorem 3.1 we can see that ROM-SIS can deal with the ultra-high dimension when $\log p = o(n^{1-2\kappa})$, which is the same order as SIS.

3.2 Ranking consistency property

In this subsection, we study the Ranking consistency property in the terminology of [2]. To obtain this property, we additionally suppose the following condition.

(C2) $\liminf_{p\to\infty} \left(\min_{k\in\mathcal{A}} \omega_k - \max_{k\in\mathcal{I}} \omega_k\right) > 0.$

Denote $\delta := \min_{k\in\mathcal{A}} \omega_k - \max_{k\in\mathcal{I}} \omega_k$, we have $\delta > 0$, under the Condition (C2). To obtain the model selection consistency, [18] proposed a sufficient condition called the partial orthogonality condition, that is, $\{X_j\}_{j\in\mathcal{A}}$ is independent of $\{X_j\}_{j\in\mathcal{I}}$, which implies $\omega_k = 0$, for $k \in \mathcal{I}$. It is easy to see that Condition (C2) is itself weaker than partial orthogonality condition. More specifically, Condition (C2) assumes that the marginal utility should have a significant separation between signal and noise variables in the population level. To successfully select the active predictors, this condition rules out some cases with strong correlation. The following Theorem 3.2 presents the ranking consistency property of our proposed screening method.

Theorem 3.2 (Ranking consistency property)**.** *Under condition (C2), there exist positive constants $c_6 < c_5 < c_7$, such that*

$$P\Big(\max_{k\in\mathcal{I}} \widehat{\omega}_k \geq \min_{k\in\mathcal{A}} \widehat{\omega}_k\Big) \leq 2p\Big\{2(n+1)\exp\{-c_5 n\} + 4\exp\{-c_6 n\} + 4\exp\{-c_7 n\}\Big\}.$$

Moreover, if we assume $\log(p) = o(n^{1-2\kappa})$, we can get

$$\liminf_{n\to\infty} \Big\{\min_{k\in\mathcal{A}} \widehat{\omega}_k - \max_{k\in\mathcal{I}} \widehat{\omega}_k\Big\} > 0, a.s..$$

Therefore, under Condition (C2), i.e. a condition on the difference between signal and noise variables, Theorem 3.2

demonstrates that the proposed ROM-SIS ensures to rank the signal variables above the noise variables almost surely in the ultra-high dimensional situations with dimensionality $p = o(n^{1-2\kappa})$. Then, $\hat{\omega}_k$ is a nature effective measure to separate the active and inactive predictor subsets. Moreover, this leads to model selection consistency.

4. NUMERICAL STUDIES

4.1 Simulations

In this subsection we assess the performance of our proposed method by several numerical experiments. We compare our proposed procedures ROM-SIS with SIS in [6], SIRS in [33], RRCS in [19], DC-SIS in [20] and MV-SIS in [2]. For MV-SIS, following [2], we rank with the marginal screening utility $MV(Y|X_j)$ when the response is continuous and predictors are discrete, and discretize each predictor into a 4-categorical variable using its 1st, 2nd and 3rd quartiles as knots when dealing with both continuous predictors and responses. We also implement the screening method based on statistics (2.2) and (2.3), and abbreviate them as Hoef and BKR respectively in the following examples. We repeat each experiment N times under the following three criteria adopted by [20].

1. $\mathcal{S}$: the minimum model size to include all active predictors. We report the 5%, 25%, 50%, 75%, and 95% quantiles of $\mathcal{S}$ out of N replications.

2. $\mathcal{P}_s$: the proportion that an individual (active) predictor is selected for a given model size d in the N replications.

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

The $\mathcal{S}$ is used to measure the model complexity of the resulting models based on the underlying screening procedures. The closer to the true model size the S is, the better the screening procedure is. The $\mathcal{P}_s$ and $\mathcal{P}_a$ allow us to examine the screening performance for an individual predictor variable and all active predictors for a given model size d.

Example 1. (Linear model): Following [9], let $\{X_k\}_{k=1}^{950}$ be iid standard normal random variables and

$$X_k = \sum_{j=1}^{s} X_j(-1)^{j+1}/5 + \sqrt{1 - \frac{s}{25}}\epsilon_k, \quad k = 951, \ldots, 1000,$$

where $\{\epsilon_k\}_{k=951}^{1000}$ are standard normally distributed. We consider the following linear model as a specific case of the additive model: $Y = \boldsymbol{\beta}^T\mathbf{X} + \epsilon$, in which $\epsilon \sim N(0, 3)$ and $\boldsymbol{\beta} = (-1, 1, -1, 1, \ldots)^T$ has s ($s = 3, 6, 12, 24$) nonzero components, taking values ± 1 alternately. We take the sample size $n = 400$ and repeat experiment $N = 200$ times.

Table 1 gives the median of minimum model size (MMS) and its associated robust estimate of the standard deviation. Note that the irrepresentable condition fails when $s > 5$ and LASSO no longer satisfies model selection consistency. It is also worth noting that SIS performs best among all the methods, particularly for $s = 24$, where MV-SIS fails badly. This is because the true model is indeed linear and the covariates are jointly normal which means marginal projection is linear as well. And our method has medium effect for this linear model among all of these methods.

Table 1. Median of minimum model size and robust estimate of standard deviations (RSD=$\frac{IQR}{1.34}$) in parentheses, where the IQR is the interquartile range. And the SNR represents signal noise ratio

Method	$s = 3$ SNR≈1.01	$s = 6$ SNR≈1.99	$s = 12$ SNR≈4.07	$s = 24$ SNR≈8.2
SIS	3(0)	56(0)	62(0.7)	126.5(56)
SIRS	3(0.7)	56(0)	62(2.2)	150(92.2)
RRCS	3(0.2)	56(0)	62(2.2)	149.5(82.6)
DC-SIS	3(0.7)	56(0)	63(2.2)	172.5(121.1)
MV-SIS	3(0.9)	56(0)	77(24.1)	410.5(231.7)
Hoef	3(0.7)	56(0)	63.5(5.2)	194(134.1)
BKR	3(0.7)	56(0)	64(4.5)	201.5(139.2)
ROM-SIS	3(0.7)	56(0)	63(4.5)	197(125.6)

Example 2. This example is adapted from [20]. We add exponential transformation models illustrated by [32] to study the performance of screening procedures on non-normal and heavy tailed data. In this example, we generate $\mathbf{x} = (X_1, X_2, \ldots, X_p)^T$ from multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$, and the error term ϵ from standard normal distribution $N(0, 1)$. We consider two structures of covariance matrices: (1) $\sigma_{ij} = 0.5^{|i-j|}$ and (2) $\sigma_{ij} = 0.8^{|i-j|}$. Other setup is: sample size $n = 200$, the number of covariates $p = 1000$ and the number of simulations $N = 500$. The response is generated from the following five models:

(4.1)
$$Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 I(X_{12} < 0) + c_4\beta_4 X_{22} + \epsilon,$$

(4.2)
$$Y = \exp\Big\{c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 I(X_{12} < 0) + c_4\beta_4 X_{22} - 5\Big\} + \epsilon,$$

(4.3)
$$Y = c_1\beta_1 X_1 X_2 + c_3\beta_2 I(X_{12} < 0)X_{22} + \epsilon,$$

(4.4)
$$Y = \exp\Big\{c_1\beta_1 X_1 X_2 + c_3\beta_2 I(X_{12} < 0)X_{22} - 5\Big\} + \epsilon,$$

(4.5)
$$Y = c_1\beta_1 X_1 + c_2\beta_2 X_2 + c_3\beta_3 I(X_{12} < 0) + \exp\{c_4|X_{22}|\}\epsilon,$$

where $I(X_{12} < 0)$ is an indicator function.

The regression functions $E(Y|\mathbf{x})$ in models (4.1)–(4.5) are all nonlinear in X_{12}. Moreover, models (4.3) and (4.4) contain interaction terms X_1X_2 and $I(X_{12} < 0)X_{22}$, models (4.2) and (4.4) possess high variance and large outliers, and

Table 2. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 2

		$\sigma_{ij} = 0.5^{\|i-j\|}$					$\sigma_{ij} = 0.8^{\|i-j\|}$				
Model	Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
(4.1)	SIS	4.0	4.0	5.0	6.0	17.0	5.0	9.0	15.0	54.5	561.4
	SIRS	4.0	4.0	5.0	7.0	18.0	5.0	9.0	15.0	64.5	504.3
	RRCS	4.0	4.0	4.0	6.0	12.0	5.0	8.0	12.0	30.2	353.9
	DC-SIS	4.0	4.0	4.0	6.0	12.0	4.0	7.0	11.0	24.0	269.8
	MV-SIS	4.0	4.0	4.0	6.0	40.0	4.0	6.0	9.0	24.0	308.1
	Hoef	4.0	4.0	4.0	6.0	15.0	4.0	7.0	11.0	22.0	218.8
	BKR	4.0	4.0	4.0	6.0	15.0	4.0	7.0	11.0	25.0	252.2
	ROM-SIS	4.0	4.0	4.0	6.0	14.0	4.0	7.0	10.0	21.0	240.2
(4.2)	SIS	91.9	272.8	488.0	766.0	954.0	117.8	347.8	554.0	758.8	958.0
	SIRS	4.0	5.0	18.0	115.2	488.6	6.0	16.0	74.0	296.8	719.1
	RRCS	4.0	5.0	16.0	113.0	582.1	5.0	12.0	62.5	293.8	789.0
	DC-SIS	46.0	199.5	409.0	641.5	891.6	71.0	262.8	491.5	705.0	907.2
	MV-SIS	4.0	5.0	29.0	172.2	632.0	5.0	10.0	57.5	238.2	715.5
	Hoef	4.0	5.0	13.0	91.2	453.4	5.0	10.0	44.0	190.5	659.2
	BKR	4.0	5.0	14.0	94.2	428.4	5.0	11.0	50.0	202.2	654.3
	ROM-SIS	4.0	5.0	12.5	82.2	399.1	5.0	10.0	38.5	184.0	640.2
(4.3)	SIS	168.9	498.2	714.5	871.0	977.2	106.8	429.5	667.0	841.2	973.1
	SIRS	108.9	392.0	644.0	872.8	978.0	64.0	305.0	598.5	838.2	979.0
	RRCS	290.8	562.0	764.0	902.0	982.0	193.5	481.0	709.5	876.0	974.0
	DC-SIS	5.0	7.0	14.0	32.0	164.2	7.0	10.0	14.0	26.0	139.0
	MV-SIS	6.0	14.0	34.0	88.2	347.4	8.0	13.0	28.0	77.0	293.0
	Hoef	9.0	18.0	32.5	70.0	202.1	8.0	12.0	19.0	37.2	149.2
	BKR	8.0	16.0	30.0	63.0	173.4	8.0	12.0	18.0	35.0	130.2
	ROM-SIS	10.0	19.0	34.5	71.0	179.1	8.0	12.8	18.0	38.0	137.0
(4.4)	SIS	198.9	506.8	723.5	884.2	982.1	248.6	580.5	759.0	884.2	983.0
	SIRS	159.9	465.5	688.0	858.0	984.0	163.5	447.8	660.0	861.5	977.0
	RRCS	155.0	462.0	704.0	874.2	979.0	138.0	438.0	706.0	880.5	982.1
	DC-SIS	168.8	472.8	701.0	870.0	974.0	219.9	559.5	736.0	881.0	971.0
	MV-SIS	17.0	122.8	363.5	694.5	926.0	24.0	156.5	366.5	655.5	921.2
	Hoef	18.0	100.5	305.5	650.2	904.1	26.0	123.8	342.5	650.0	919.0
	BKR	22.9	105.0	317.0	659.5	906.0	26.0	139.8	350.5	655.2	930.1
	ROM-SIS	17.0	89.8	283.5	657.0	926.0	24.9	113.8	340.5	648.0	932.1
(4.5)	SIS	50.9	257.0	565.0	785.8	964.0	68.9	304.2	566.5	829.2	959.1
	SIRS	21.0	174.2	411.0	726.2	947.0	30.0	181.8	429.5	716.5	960.0
	RRCS	16.0	165.8	407.5	743.0	943.6	25.0	177.2	409.0	699.5	931.0
	DC-SIS	4.0	4.0	6.0	16.2	167.1	4.0	6.0	12.5	54.2	346.0
	MV-SIS	6.0	20.0	44.5	97.0	249.3	8.0	19.8	51.5	112.2	355.2
	Hoeff	9.0	40.8	92.0	164.2	377.0	12.0	45.0	96.5	180.0	449.2
	BKR	7.0	28.8	58.0	105.2	233.3	10.0	31.8	66.0	125.0	317.0
	ROM-SIS	11.0	45.8	82.0	139.0	280.3	14.0	47.0	88.5	149.0	336.0

model (4.5) is heteroscedastic. Following Fan and Lv (2008), we choose $\beta_j = (-1)^U(a + |Z|)$ for $j = 1, 2, 3$, and 4, where $a = 4 \log n/\sqrt{n}$, $U \sim$ Bernoulli(0.4) and $Z \sim N(0, 1)$. We set$(c_1, c_2, c_3, c_4) = (2, 0.5, 3, 2)$ in this example to challenge the feature screening procedures under consideration.

Tables 2 and 3 summarize the simulation results for $\mathcal{S}$, $\mathcal{P}_s$ and $\mathcal{P}_a$. We can see that all the screening procedures are equally good for model (4.1) since it does not deviate far from linear model. However, when the true model is nonlinear in model (4.2), SIS fails as badly as DC-SIS. The reason that DC-SIS cannot identify active predictors well is because the finite exponential moment condition (C1) in [20] does not hold in the presence of extreme values in the response. In the case of model (4.3), SIS, SIRS and RRCS cannot rank the truly important predictors in the top with very high probability. This is due to the existence of interaction terms and the true model is no longer the traditional single- or multi-index model, where Y depends on the predictors $\mathbf{x}$ through a number of linear combinations $\boldsymbol{\beta}^T\mathbf{x}_{\mathcal{A}}$. In all the above cases, our proposed method performs excellently and uniformly, and surpass all the other methods with slightly better results over MV-SIS in model (4.4). Even in the heteroscedastic model (4.5), our procedures perform not too bad.

Table 3. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ in Example 2 with model size $d = [n/logn]$

		$\sigma_{ij} = 0.5^{\|i-j\|}$					$\sigma_{ij} = 0.8^{\|i-j\|}$				
		P_s				P_a	P_s				P_a
Model	Method	X1	X2	X12	X20	All	X1	X2	X12	X20	All
(4.1)	SIS	1.00	1.00	0.99	1.00	0.99	1.00	1.00	0.69	1.00	0.69
	SIRS	1.00	1.00	0.98	1.00	0.97	1.00	1.00	0.69	1.00	0.69
	RRCS	1.00	1.00	0.99	1.00	0.99	1.00	1.00	0.77	1.00	0.77
	DC-SIS	1.00	1.00	0.99	1.00	0.99	1.00	1.00	0.82	1.00	0.81
	MV-SIS	1.00	0.98	0.97	1.00	0.95	1.00	1.00	0.81	1.00	0.81
	Hoef	1.00	0.99	0.99	1.00	0.99	1.00	1.00	0.82	1.00	0.82
	BKR	1.00	1.00	0.99	1.00	0.99	1.00	1.00	0.81	1.00	0.81
	ROM-SIS	1.00	1.00	0.99	1.00	0.99	1.00	1.00	0.83	1.00	0.83
(4.2)	SIS	0.67	0.32	0.09	0.59	0.01	0.69	0.52	0.03	0.49	0.01
	SIRS	0.96	0.87	0.69	0.95	0.61	0.97	0.93	0.42	0.90	0.38
	RRCS	0.94	0.84	0.73	0.93	0.63	0.95	0.91	0.49	0.89	0.44
	DC-SIS	0.77	0.39	0.12	0.68	0.04	0.75	0.63	0.05	0.59	0.02
	MV-SIS	0.92	0.75	0.69	0.89	0.53	0.92	0.86	0.52	0.83	0.45
	Hoef	0.94	0.85	0.78	0.93	0.65	0.96	0.91	0.54	0.89	0.48
	BKR	0.94	0.82	0.78	0.92	0.64	0.94	0.90	0.54	0.89	0.47
	ROM-SIS	0.95	0.85	0.78	0.93	0.66	0.96	0.92	0.55	0.90	0.49
(4.3)	SIS	0.16	0.17	0.03	1.00	0.00	0.21	0.19	0.07	0.99	0.01
	SIRS	0.07	0.06	0.68	1.00	0.01	0.07	0.05	0.67	1.00	0.02
	RRCS	0.08	0.07	0.04	1.00	0.00	0.09	0.10	0.07	1.00	0.00
	DC-SIS	0.97	0.97	0.83	1.00	0.78	1.00	1.00	0.83	1.00	0.83
	MV-SIS	0.94	0.95	0.60	1.00	0.52	1.00	1.00	0.58	1.00	0.57
	Hoef	0.85	0.85	0.75	1.00	0.55	1.00	1.00	0.75	1.00	0.75
	BKR	0.87	0.88	0.77	1.00	0.59	1.00	1.00	0.78	1.00	0.78
	ROM-SIS	0.85	0.85	0.77	1.00	0.55	1.00	0.99	0.76	1.00	0.75
(4.4)	SIS	0.40	0.41	0.06	0.44	0.00	0.48	0.47	0.08	0.47	0.00
	SIRS	0.05	0.05	0.34	0.91	0.00	0.06	0.05	0.20	0.90	0.00
	RRCS	0.06	0.07	0.37	0.87	0.00	0.07	0.06	0.25	0.87	0.00
	DC-SIS	0.45	0.47	0.07	0.46	0.00	0.53	0.54	0.09	0.46	0.00
	MV-SIS	0.53	0.53	0.30	0.84	0.11	0.62	0.63	0.21	0.82	0.08
	Hoef	0.44	0.41	0.41	0.89	0.10	0.60	0.58	0.27	0.89	0.09
	BKR	0.44	0.41	0.36	0.89	0.10	0.60	0.59	0.24	0.89	0.08
	ROM-SIS	0.45	0.41	0.41	0.89	0.11	0.60	0.59	0.29	0.89	0.10
(4.5)	SIS	0.54	0.34	0.22	0.51	0.04	0.55	0.48	0.17	0.48	0.02
	SIRS	1.00	0.99	0.93	0.08	0.08	1.00	1.00	0.76	0.08	0.06
	RRCS	1.00	0.99	0.95	0.10	0.09	1.00	1.00	0.80	0.11	0.08
	DC-SIS	0.99	0.96	0.89	1.00	0.86	1.00	1.00	0.72	1.00	0.71
	MV-SIS	1.00	0.95	0.94	0.50	0.45	1.00	1.00	0.82	0.51	0.41
	Hoef	1.00	0.99	0.96	0.25	0.23	1.00	1.00	0.87	0.26	0.22
	BKR	1.00	0.99	0.97	0.35	0.33	1.00	1.00	0.84	0.36	0.30
	ROM-SIS	1.00	0.98	0.96	0.23	0.21	1.00	1.00	0.87	0.23	0.19

Example 3. (Transformation model): This example is adopted from [19] to study the impact of monotone transformation regression on the proposed methods. Consider the following generalized Box-Cox transformation model:

$$H(Y_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \cdots, n,$$

where the transformation functions are unknown. In the simulations, we consider the Box-Cox transformation:

$$f(x) = \begin{cases} \frac{|Y|^\lambda \text{sign}(Y) - 1}{\lambda}, & \text{when } \lambda = 0.25, 0.5, 0.75, \\ \log Y, & \text{when } \lambda = 0 \end{cases}$$

The sample $(X_1, \cdots, X_p)$ with size n is generated from a multivariate normal distribution $N(0, \Sigma)$, where $\Sigma = CS(\rho)$, and noise ϵ_i follows the standard normal distributions, $\boldsymbol{\beta} = (3, 1.5, 2, 0, \cdots, 0)^T$. The replication time is again $N = 500$, and $n = 50, p = 1000$ and $\rho = 0, 0.1, 0.5, 0.9$, respectively.

Table 4 presents the median of the minimum model size $\mathcal{S}$ and corresponding robust estimate of standard deviations in parentheses. Table 5 presents the the proportions of $\mathcal{P}_a$ with model size $d = 2n$. We can see clearly that, under every scenario, our proposed methods are comparable with the best one RRCS, which is also invariant under any strictly monotone univariate transformations.

Table 4. The median of the minimum model size $\mathcal{S}$ out of 200 replications in Example 3, and corresponding robust estimate of standard deviations (RSD = IQR/1.34, where IQR is the interquartile range) in parentheses

λ	Method	$\rho = 0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
0	SIS	383.0 (323.9)	326.0 (308.4)	474.0 (279.3)	665.0 (274.0)
	SIRS	21.0 (53.9)	19.0 (42.7)	55.0 (102.1)	214.0 (227.5)
	RRCS	22.0 (56.3)	22.0 (50.0)	54.0 (94.6)	139.5 (181.2)
	DC	296.5 (284.1)	255.0 (262.8)	423.0 (267.0)	625.5 (281.5)
	MV	112.0 (190.7)	67.5 (107.2)	100.0 (148.7)	277.0 (254.8)
	Hoef	36.0 (91.2)	35.0 (68.8)	70.0 (115.8)	168.0 (201.8)
	BKR	23.0 (51.6)	31.0 (60.6)	62.5 (111.8)	160.5 (196.1)
	ROM-SIS	31.0 (83.6)	30.0 (59.3)	73.0 (133.7)	168.0 (207.1)
0.25	SIS	104.0 (168.8)	84.5 (129.0)	239.0 (213.0)	568.5 (281.7)
	SIRS	19.0 (57.1)	20.0 (42.7)	55.0 (101.8)	208.0 (225.7)
	RRCS	20.0 (55.2)	21.0 (47.0)	52.0 (93.4)	145.5 (190.4)
	DC	59.5 (135.2)	58.0 (107.1)	138.0 (179.7)	438.0 (294.4)
	MV	95.5 (174.3)	64.5 (104.1)	102.0 (134.5)	257.5 (264.2)
	Hoef	35.0 (85.4)	28.5 (65.7)	70.0 (122.5)	171.5 (208.0)
	BKR	20.0 (50.7)	25.0 (59.7)	66.0 (115.1)	164.0 (208.6)
	ROM-SIS	29.0 (75.5)	28.0 (58.4)	72.0 (123.1)	172.0 (219.0)
0.5	SIS	24.5 (56.7)	26.0 (54.5)	61.5 (106.7)	297.5 (263.4)
	SIRS	20.0 (52.2)	20.0 (45.1)	55.5 (94.2)	198.5 (238.2)
	RRCS	19.0 (50.7)	22.0 (47.8)	55.5 (99.4)	141.5 (193.1)
	DC	24.0 (71.8)	28.0 (53.0)	49.5 (82.2)	197.0 (223.9)
	MV	97.5 (171.8)	71.5 (126.5)	92.0 (153.5)	270.0 (240.4)
	Hoef	32.5 (86.9)	33.0 (74.6)	74.0 (130.4)	164.0 (199.3)
	BKR	18.0 (47.0)	31.0 (66.6)	72.0 (123.1)	160.0 (199.4)
	ROM-SIS	31.0 (74.6)	33.0 (61.3)	73.0 (145.5)	174.0 (204.3)
0.75	SIS	18.0 (40.3)	14.0 (35.1)	34.5 (71.8)	121.5 (175.7)
	SIRS	22.5 (47.4)	18.0 (42.9)	56.5 (103.0)	207.0 (212.8)
	RRCS	24.5 (50.0)	18.0 (43.4)	51.0 (106.3)	141.5 (217.3)
	DC	24.5 (59.0)	18.0 (49.3)	37.0 (86.0)	101.5 (149.6)
	MV	105.0 (196.9)	66.0 (112.1)	100.0 (152.2)	252.0 (228.5)
	Hoef	36.5 (89.9)	28.0 (62.1)	70.5 (141.4)	172.5 (233.6)
	BKR	20.5 (51.6)	25.0 (54.6)	67.0 (134.5)	170.5 (237.9)
	ROM-SIS	34.5 (73.1)	27.0 (59.3)	73.0 (138.8)	172.0 (218.8)
1	SIS	15.0 (37.3)	13.0 (33.6)	31.0 (61.6)	80.0 (130.2)
	SIRS	21.0 (53.9)	21.0 (45.9)	48.5 (91.4)	233.0 (239.7)
	RRCS	22.0 (56.3)	20.5 (48.7)	48.0 (101.6)	156.5 (213.6)
	DC	23.0 (59.0)	19.5 (43.3)	45.0 (85.2)	110.0 (163.4)
	MV	112.0 (190.7)	66.5 (115.3)	98.5 (151.1)	285.0 (280.7)
	Hoef	36.0 (91.2)	31.0 (66.6)	66.0 (122.4)	174.5 (223.0)
	BKR	23.0 (51.6)	29.0 (58.4)	62.5 (114.6)	173.5 (224.0)
	ROM-SIS	31.0 (83.6)	27.5 (60.6)	64.0 (117.8)	175.0 (221.4)

Example 4. (A highly nonlinear situation). Here we take a look at a model with highly nonlinear structure studied by [12] with adding a small modification. Let $W_{i1}, \ldots, W_{i6}$ and $X_{i5}, \ldots, X_{i,1000}$ be independent standard normal random variables, and put

$$Y_i = 2\sin\left\{\frac{\pi}{2}(W_{i1} + 0.5W_{i2})\right\} + \sum_{j=3}^{5} W_{ij}^2 + 0.4e^{W_{i6}} + Z_{i0}$$

and $X_{i1} = W_{i1} + Z_{i1}, X_{i2} = 2W_{i2} + Z_{i2}, X_{i3} = W_{i3}W_{i4} + Z_{i3}$, and $X_{i4} = W_{i6} + Z_{i4}$, with each of the Z_{ij} being normal random variables with mean zero and standard deviation 0.1. We repeat the simulation $N = 500$ times with sample size $n = 200$.

The results are presented in Tables 6 and 7 together with computation time cost for 20 replicates. Notice that the BKR and DC-SIS are the top two best methods in this highly nonlinear situation, but BKR consumes much less time than DC-SIS due to its low computation complexity. However, SIRS and RRCS have little chance to identify the important predictors X_{i3}, and SIS and MV-SIS weakly detects X_{i3} and X_{i2} respectively.

Example 5. (Genome-Wide Association Studies) To study the influence of discrete predictors on screening procedure based on Modified Hoeffding coefficient, we adopt this example from [2]. In the genomics-wide association study

Table 5. The proportions of $\mathcal{P}_a$ in Example 3 with model size $d = 2n$

λ	Method	$\rho = 0$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
0	SIS	0.08	0.11	0.03	0.00
	SIRS	0.79	0.81	0.65	0.29
	RRCS	0.79	0.81	0.67	0.44
	DC	0.13	0.18	0.06	0.00
	MV	0.49	0.60	0.50	0.21
	Hoef	0.70	0.74	0.58	0.36
	BKR	0.79	0.78	0.61	0.36
	ROM-SIS	0.73	0.77	0.57	0.33
0.25	SIS	0.49	0.53	0.22	0.02
	SIRS	0.79	0.83	0.66	0.27
	RRCS	0.80	0.82	0.68	0.41
	DC	0.61	0.64	0.41	0.07
	MV	0.51	0.63	0.49	0.19
	Hoef	0.71	0.75	0.61	0.34
	BKR	0.79	0.78	0.64	0.34
	ROM-SIS	0.74	0.78	0.58	0.33
0.5	SIS	0.78	0.80	0.62	0.18
	SIRS	0.80	0.81	0.66	0.27
	RRCS	0.80	0.82	0.67	0.45
	DC	0.74	0.80	0.67	0.32
	MV	0.51	0.58	0.52	0.19
	Hoef	0.73	0.74	0.57	0.33
	BKR	0.81	0.75	0.59	0.34
	ROM-SIS	0.74	0.77	0.59	0.33
0.75	SIS	0.84	0.83	0.74	0.44
	SIRS	0.80	0.81	0.64	0.28
	RRCS	0.81	0.81	0.66	0.43
	DC	0.80	0.80	0.71	0.50
	MV	0.49	0.59	0.50	0.18
	Hoef	0.71	0.77	0.57	0.35
	BKR	0.79	0.79	0.60	0.36
	ROM-SIS	0.74	0.77	0.56	0.34
1	SIS	0.84	0.86	0.76	0.55
	SIRS	0.79	0.80	0.66	0.27
	RRCS	0.79	0.82	0.69	0.40
	DC	0.78	0.82	0.68	0.49
	MV	0.49	0.61	0.51	0.19
	Hoef	0.70	0.75	0.61	0.32
	BKR	0.79	0.77	0.62	0.32
	ROM-SIS	0.73	0.78	0.60	0.33

Table 6. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 4 with time cost of 20 replicates (in seconds)

	5%	25%	50%	75%	95%	Time
SIS	33.0	148.0	366.5	724.5	945.0	0.2
SIRS	63.0	285.2	570.0	828.0	980.1	0.9
RRCS	66.0	283.2	519.5	763.0	953.3	16.7
DC-SIS	6.0	20.0	87.5	277.5	754.2	220.7
MV-SIS	12.0	63.8	204.0	421.2	806.1	18.6
Hoef	7.0	26.0	99.0	268.0	743.7	14.1
BKR	6.0	21.0	71.5	205.2	672.8	19.5
ROM-SIS	7.0	24.8	93.5	256.2	763.3	203.6

Table 7. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ for Example 4 with model sizes $d_1 = [n/logn], d_2 = 2[n/logn], d_3 = 3[n/logn]$

		$\mathcal{P}_s$				$\mathcal{P}_a$
Size	Method	X1	X2	X3	X4	All
d_1	SIS	0.85	0.34	0.23	0.83	0.06
	SIRS	0.94	0.42	0.07	0.83	0.02
	RRCS	0.99	0.44	0.05	0.81	0.03
	DC-SIS	0.99	0.39	1.00	0.83	0.33
	MV-SIS	0.99	0.27	1.00	0.59	0.16
	Hoef	1.00	0.41	0.99	0.75	0.32
	BKR	1.00	0.46	0.99	0.78	0.36
	ROM-SIS	1.00	0.41	1.00	0.75	0.32
d_2	SIS	0.90	0.45	0.31	0.89	0.13
	SIRS	0.96	0.54	0.13	0.88	0.06
	RRCS	0.99	0.56	0.10	0.86	0.05
	DC-SIS	1.00	0.52	1.00	0.88	0.46
	MV-SIS	1.00	0.37	1.00	0.72	0.27
	Hoef	1.00	0.53	1.00	0.86	0.45
	BKR	1.00	0.58	1.00	0.87	0.51
	ROM-SIS	1.00	0.54	1.00	0.85	0.46
d_3	SIS	0.95	0.54	0.39	0.91	0.20
	SIRS	0.98	0.61	0.17	0.91	0.09
	RRCS	1.00	0.63	0.13	0.90	0.08
	DC-SIS	1.00	0.59	1.00	0.92	0.54
	MV-SIS	1.00	0.45	1.00	0.78	0.35
	Hoef	1.00	0.61	1.00	0.88	0.54
	BKR	1.00	0.68	1.00	0.90	0.61
	ROM-SIS	1.00	0.62	1.00	0.88	0.55

(GWAS), the response (i.e. the phenotypes) are continuous whereas the predictors (i.e. the single-nucleotide polymorphisms or SNPs) are categorical. In general, the SNPs as predictors are categorical with three classes, denoted by $\{AA, Aa, aa\}$. To mimic SNPs with equal allele frequencies, we denote Z_{ij} as the indicators of the dominant effect of the jth SNP for ith subject and generate it in the following way

$$Z_{ij} = \begin{cases} 1, & \text{if } X_{ij} < q_1 \\ 0, & \text{if } q_1 \le X_{ij} < q_3 \\ -1, & \text{if } X_{ij} \ge q_3 \end{cases}$$

where $X_i = (X_{i1}, \ldots, X_{ip}) \sim N(0, \Sigma)$, where $\Sigma = (\rho_{ij})_{p\times p}$ with $\rho_{ij} = 0.5^{|i-j|}$, and q_1 and q_3 are first and third quartiles of a standard normal distribution, respectively. Then, we generate the response (some trait or disease) by:

$$Y = \beta_1 Z_1 + \beta_2 Z_2 + 2\beta_3 Z_{10} + 2\beta_4 Z_{20} - 2\beta_5 |Z_{100}| + \epsilon,$$

where $\beta_j = (-1)^U(a + |Z|)$ for $j = 1, \ldots, 5$, where $a = 2\log n/\sqrt{n}$, $U \sim$ Bernoulli(0.4) and $Z \sim N(0, 1)$, the error term ϵ follows $N(0, 1)$ or $t(1)$. The first four active SNPs, Z_1, Z_2, Z_{10}, Z_{20} are linearly correlated with the response Y, while the SNP Z_{100} and Y are nonlinearly correlated. We set $n = 200$ and $p = 1000$ and repeat each experiment $N = 500$ times.

Table 8. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 5

	$\epsilon \sim N(0,1)$					$\epsilon \sim t(1)$				
Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS	49.9	264.0	501.5	731.0	958.2	232.0	526.0	698.0	861.5	975.0
SIRS	51.0	265.0	511.5	770.5	960.2	83.4	267.0	506.0	756.3	944.2
RRCS	150.8	644.0	1000.0	1000.0	1000.0	167.3	643.8	1000.0	1000.0	1000.0
DC-SIS	5.0	6.0	10.0	34.3	248.3	6.0	15.0	57.5	193.8	616.9
MV-SIS	5.0	6.0	8.0	29.3	265.3	5.0	8.0	21.0	82.0	485.5
Hoef	10.0	37.0	98.0	224.0	528.2	20.0	64.5	146.0	303.5	615.2
BKR	5.0	7.0	14.0	44.0	294.2	5.0	11.0	31.0	113.3	516.0
ROM-SIS	5.0	7.0	14.0	39.0	289.2	5.0	10.8	31.0	111.3	484.2

Table 9. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ for Example 5 with model sizes $d_1 = [n/logn], d_2 = 2[n/logn], d_3 = 3[n/logn]$

		$\epsilon \sim N(0,1)$						$\epsilon \sim t(1)$					
		$\mathcal{P}_s$					$\mathcal{P}_a$	$\mathcal{P}_s$					$\mathcal{P}_a$
Size	Method	X1	X2	X10	X20	X100	All	X1	X2	X10	X20	X100	All
d_1	SIS	0.97	0.96	1.00	1.00	0.04	0.04	0.33	0.30	0.44	0.45	0.03	0.01
	SIRS	0.95	0.96	1.00	1.00	0.03	0.02	0.92	0.90	0.98	0.97	0.02	0.02
	RRCS	0.39	0.39	0.40	0.40	0.05	0.01	0.37	0.37	0.40	0.40	0.03	0.01
	DC-SIS	0.95	0.95	0.99	1.00	0.83	0.77	0.83	0.81	0.93	0.92	0.64	0.40
	MV-SIS	0.91	0.92	0.99	0.99	0.92	0.79	0.87	0.85	0.98	0.96	0.85	0.61
	Hoef	0.95	0.95	1.00	1.00	0.27	0.26	0.91	0.90	0.99	0.97	0.17	0.12
	BKR	0.94	0.94	0.99	0.99	0.81	0.74	0.88	0.88	0.98	0.96	0.72	0.53
	ROM-SIS	0.93	0.95	0.99	0.99	0.82	0.74	0.88	0.88	0.98	0.96	0.72	0.53
d_2	SIS	0.99	0.98	1.00	1.00	0.07	0.07	0.39	0.36	0.49	0.50	0.07	0.01
	SIRS	0.97	0.97	1.00	1.00	0.07	0.07	0.95	0.93	0.99	0.98	0.05	0.05
	RRCS	0.40	0.40	0.40	0.40	0.09	0.02	0.40	0.39	0.40	0.40	0.07	0.02
	DC-SIS	0.97	0.97	1.00	1.00	0.90	0.86	0.88	0.86	0.96	0.95	0.73	0.55
	MV-SIS	0.95	0.96	1.00	1.00	0.94	0.86	0.91	0.90	0.99	0.97	0.91	0.73
	Hoef	0.97	0.97	1.00	1.00	0.43	0.41	0.94	0.93	0.99	0.98	0.33	0.28
	BKR	0.97	0.96	1.00	0.99	0.89	0.83	0.93	0.92	0.99	0.98	0.81	0.67
	ROM-SIS	0.96	0.97	1.00	0.99	0.89	0.83	0.92	0.92	0.99	0.98	0.81	0.67
d_3	SIS	0.99	0.99	1.00	1.00	0.11	0.11	0.44	0.41	0.52	0.54	0.11	0.02
	SIRS	0.98	0.98	1.00	1.00	0.10	0.10	0.96	0.95	0.99	0.99	0.08	0.08
	RRCS	0.40	0.40	0.40	0.40	0.12	0.03	0.40	0.39	0.41	0.41	0.12	0.03
	DC-SIS	0.98	0.98	1.00	1.00	0.93	0.89	0.90	0.88	0.96	0.96	0.78	0.62
	MV-SIS	0.97	0.97	1.00	1.00	0.95	0.89	0.93	0.92	0.99	0.98	0.94	0.80
	Hoef	0.98	0.98	1.00	1.00	0.55	0.54	0.96	0.94	0.99	0.99	0.43	0.40
	BKR	0.97	0.97	1.00	0.99	0.92	0.87	0.94	0.93	0.99	0.98	0.86	0.75
	ROM-SIS	0.97	0.97	1.00	1.00	0.92	0.87	0.94	0.93	0.99	0.98	0.87	0.75

We report the results in Tables 8 and 9. Surprisingly, RRCS fails completely. This failure highlights a drawback in RRCS that many ties occur in marginal utility vector when the predictor is finitely discrete. Since we choose the most conservative way to compute minimum model size during all the simulations, that is, we get the last index of the predictors which possess the same marginal utility with the true predictor.

When the error follows a normal distribution, all the independence screening except RRCS are able to select the first four active SNPs effectively because they are linearly correlated with the response. However, only BKR, ROM-SIS, DC-SIS and MV-SIS can choose Z_{100} which nonlinearly contributed to Y. It is interesting to notice that Hoef no longer performs as similarly as BKR and ROM-SIS in the previous examples. When the error is generated from $t(1)$ which is largely heavy-tailed, BKR and ROM-SIS perform comparably well with the best one MV-SIS.

Example 6. In this example, we consider multivariate responses data to analyze the performance of our proposed method. The example has been investigated by [20]. We compare with DC-SIS since other screening procedure cannot be directly applied for such settings. We generate $\mathbf{x} = (X_1, X_2, \ldots, X_p)^T$ from multivariate normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$, where two structures of covariance matrices: (1) $\sigma_{ij} = 0.5^{|i-j|}$ and (2) $\sigma_{ij} = 0.8^{|i-j|}$ are taken into consideration. The response

Table 10. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size $\mathcal{S}$ out of 500 replications in Example 6

		$\sigma_{ij} = 0.5^{\|i-j\|}$					$\sigma_{ij} = 0.8^{\|i-j\|}$				
Model	Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
(4.6)	DC	2.0	5.0	10.0	23.0	59.1	2.0	2.0	4.0	8.0	21.1
	BKR	2.0	6.0	20.0	67.3	279.3	2.0	3.0	6.0	24.0	116.2
	ROM-SIS	2.0	3.0	7.0	26.0	151.3	2.0	2.0	3.0	9.0	55.1
(4.7)	DC	6.0	12.0	21.0	44.3	117.2	4.0	4.0	4.0	5.0	8.0
	BKR	5.0	14.8	48.5	147.3	430.1	4.0	4.0	5.0	9.0	35.1
	ROM-SIS	4.0	7.0	19.5	65.0	303.4	4.0	4.0	4.0	6.0	13.1

Table 11. The proportions of $\mathcal{P}_s$ and $\mathcal{P}_a$ for Example 6 with model sizes $d_1 = [n/logn], d_2 = 2[n/logn], d_3 = 3[n/logn]$

		$\sigma_{ij} = 0.5^{\|i-j\|}$								$\sigma_{ij} = 0.8^{\|i-j\|}$							
		(4.6)			(4.7)					(4.6)			(4.7)				
		$\mathcal{P}_s$		$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$	$\mathcal{P}_s$		$\mathcal{P}_a$	$\mathcal{P}_s$				$\mathcal{P}_a$
Size	Method	X1	X2	All	X1	X2	X3	X4	All	X1	X2	All	X1	X2	X3	X4	All
d_1	DC	0.98	0.88	0.87	0.83	0.99	0.99	0.86	0.71	1.00	0.99	0.99	1.00	1.00	1.00	1.00	1.00
	BKR	0.83	0.70	0.62	0.70	0.90	0.88	0.70	0.43	0.89	0.85	0.80	0.99	0.99	1.00	0.97	0.96
	ROM-SIS	0.93	0.86	0.81	0.84	0.95	0.96	0.80	0.63	0.96	0.94	0.92	1.00	1.00	1.00	0.99	0.99
d_2	DC	1.00	0.98	0.98	0.93	1.00	1.00	0.95	0.88	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	BKR	0.91	0.83	0.77	0.82	0.95	0.95	0.78	0.60	0.95	0.92	0.89	0.99	1.00	1.00	0.99	0.98
	ROM-SIS	0.97	0.92	0.90	0.91	0.98	0.98	0.87	0.78	0.99	0.98	0.97	1.00	1.00	1.00	1.00	0.99
d_3	DC	1.00	0.99	0.99	0.96	1.00	1.00	0.98	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	BKR	0.95	0.89	0.85	0.87	0.96	0.97	0.83	0.69	0.97	0.96	0.94	1.00	1.00	1.00	0.99	0.99
	ROM-SIS	0.98	0.94	0.93	0.95	0.99	0.99	0.91	0.84	0.99	0.99	0.98	1.00	1.00	1.00	1.00	1.00

$\mathbf{y} = (Y_1, Y_2)^T$ are generated from normal distribution with mean zero and covariance matrix $\Sigma_{\mathbf{y}|\mathbf{x}} = (\sigma_{\mathbf{x},ij})_{2\times 2}$, where $\sigma_{\mathbf{x},11} = \sigma_{\mathbf{x},22} = 1$ and $\sigma_{\mathbf{x},12} = \sigma_{\mathbf{x},21} = \sigma(\mathbf{x})$. We look into two scenarios for the correlation function $\sigma(\mathbf{x})$:

$$\sigma(\mathbf{x}) = \sin(\beta_1^T \mathbf{x}) \tag{4.6}$$

where $\beta_1 = (0.8, 0.6, 0, \cdots, 0)^T$

$$\sigma(\mathbf{x}) = \frac{\{\exp(\beta_2^T \mathbf{x}) - 1\}}{\{\exp(\beta_2^T \mathbf{x}) + 1\}} \tag{4.7}$$

where $\beta_2 = (2 - U_1, 2 - U_2, 2 - U_3, 2 - U_4, 0, \cdots, 0)^T$ with U_i's being independent and identically distributed (iid) according to uniform distribution Uniform $[0, 1]$.

Tables 10 and 11 present the simulation results. We can see that ROM-SIS is comparable with DC-SIS and both of them performs reasonably well for the two models in terms of model complexity. Also, BKR is not that bad and it is much more computational faster than both ROM-SIS and DC-SIS. It implies that the both BKR and ROM-SIS can identify the active predictors contained in correlations between multivariate responses, which is potentially useful in gene coexpression analysis.

4.2 Leukemia data analysis

In this subsection, we apply our method to implement feature screening for the leukemia microarray data set. These data come from a study by [11] and also have been analysed in [30], [4], [6], [12]. The data are available from http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi, and the aim for analyzing this data set is to use microarray evidence to distinguish between two types of acute leukemia. There are 7129 genes and 72 samples from two classes: 47 in class ALL (acute lymphocytic leukaemia) and 25 in class AML (acute mylogenous leukaemia). Among those 72 samples, 38 (27 in class ALL and 11 in class AML) of them were set as the training sample and the remaining 34 (20 in class ALL and 14 in class AML) of them were set to be the test sample.

To examine the performance of all the screening methods mentioned previously, we follow [4] to split the 72 samples into training and test sets randomly. Specifically, we set approximately $100\gamma\%$ of the observations from class ALL and $100\gamma\%$ of the observations from class AML as standardized training samples, and the rest as test samples. We then apply screening methods to training sample to select $d = [2n/\log(n)]$ features for the classification as done in [6], where n is the size of training sample. Furthermore, we use selected features to carry out the classification in unstandardized test sample by linear discriminant analysis and calculate the test error. The above procedure is repeated 100 times for $\gamma = 0.4, 0.5$ and 0.6, respectively, and the distributions of test errors of all the screening methods are summarized in Figure 2. From these figures, we notice that ROM-SIS, Hoef and MV-SIS are best among all the screening method. MV-SIS is born to solve classification problems and our method has the same results as MV-SIS. This is not

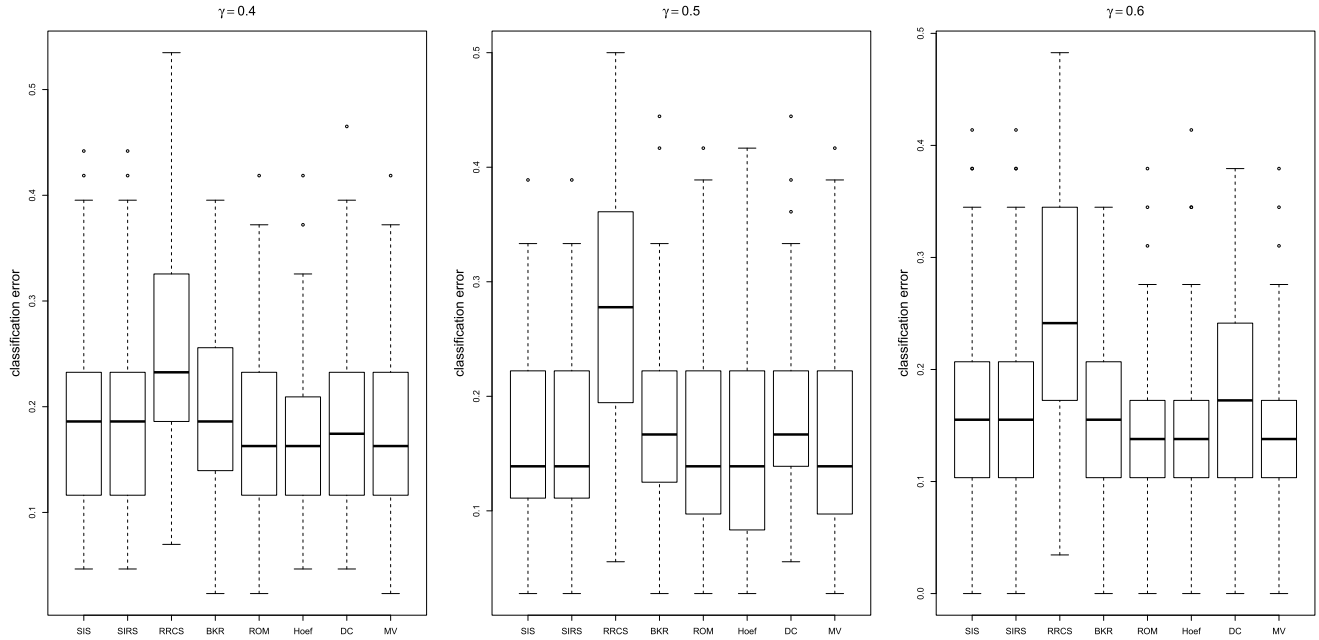

Figure 2. Boxplots of test errors of 100 random splits of 72 samples in leukemia data, where $100\gamma\%$ *of the samples from both classes are set as training samples. The three plots from left to right correspond to* $\gamma = 0.4, 0.5$ *and 0.6, respectively.*

surprising, because both the two methods are based on the Cramér-von Mises distance.

As for the influential genes, we also use the the original training sample to identify them by using the aforementioned screening methods. Surprisingly, we find that only the three screening methods based on Hoeffding's measure have some similar identification to the method proposed by [12], where the first four top ranked genes are labeled X95735_at, M27891_at, M27783_s_at and U50136_rna1_at. One reason for this result may be that both the Hoeffding's measure and generalized correlation used by [12] can character implicit relationship between variables. For the four genes, we conduct an exploratory analysis in figure 3. The histograms and boxplots of the four genes reveal that the distributions of them are highly skewed and there exists some potential outliers. So it is reasonable to use our proposed method to implement feature screening for this data set, since our method is robust to the outliers and we do not need prespecify the distributions.

APPENDIX A. APPENDIX SECTION

Proof of Proposition 2.1. The non-negativity of $\omega(X,Y)$ is obvious. We only need to prove the equivalency. If X and Y are independent, it is easy to know $\omega(X,Y) = 0$ due to $F_{X,Y}(x,y) = F_X(x)F_Y(y), \forall x,y \in \mathbb{R}$. For any random variables X and Y, we denote their support as $\mathcal{B}_X$ and $\mathcal{B}_Y$. If X and Y are dependent, the binary function $[F_{X,Y}(x,y) - F_X(x)F_Y(y)]^2 \neq 0$, for some $x \in \mathcal{B}_X, y \in \mathcal{B}_Y$. We argue that $\omega(X,Y) > 0$ as follows.

When X and Y are both continuous random variables, $\omega(X,Y) = \int_{\mathcal{B}_X}\int_{\mathcal{B}_Y}[F_{X,Y}(x,y) - F_X(x)F_Y(y)]^2 f_X(x)f_Y(y)dxdy$, where $f_X(x)$ and $f_Y(y)$ are their density functions. From the continuity of integrand, we can get $\omega(X,Y) > 0$.

When X and Y are both discrete random variables, the modified Hoeffding measure can be rewritten as: $\omega(X,Y) = \sum_{x\in\mathcal{B}_X}\sum_{y\in\mathcal{B}_Y}[F_{X,Y}(x,y) - F_X(x)F_Y(y)]^2 P(X=x)P(Y=y)$. The positivity of $\omega(X,Y)$ can be obtained by $P(X = x) > 0, \forall x \in \mathcal{B}_X$ and $P(Y = y) > 0, \forall y \in \mathcal{B}_Y$.

If X is continuous and Y is discrete, we have: $\omega(X,Y) = \sum_{y\in\mathcal{B}_Y} P(Y=y)\int_{\mathcal{B}_X}[F_{X,Y}(x,y) - F_X(x)F_Y(y)]^2 f_X(x)dx$. The positivity of $P(Y = y), y \in \mathcal{B}_Y$ and the continuity of $[F_{X,Y}(x,y) - F_X(x)F_Y(y)]^2 f_X(x)$ will lead to $\omega(X,Y) > 0$.

It is similar to get the proof for the situation where the marginal distribution is a mixture of continuous and discrete distribution. This completes the proof. □

Proof of Proposition 2.2. From the definition of distribution function for binary random vector, we have $F_{X,Y}(x,y) = P(X \leq x, Y \leq y), \forall x,y \in \mathbb{R}$. If $X = Y$, it becomes $F_{X,X}(x,y) = P(X \leq x, X \leq y) = F_X(min(x,y))$. So we can calculate $\omega(X,X)$ as follows,

$$\begin{aligned}\omega(X,X) &= \int [F_X(min(x,y)) - F_X(x)F_X(y)]^2 \, dF_X(x)dF_X(y) \\ &= 2\int_{x\leq y} [F_X(x) - F_X(x)F_X(y)]^2 \, dF_X(x)dF_X(y)\end{aligned}$$

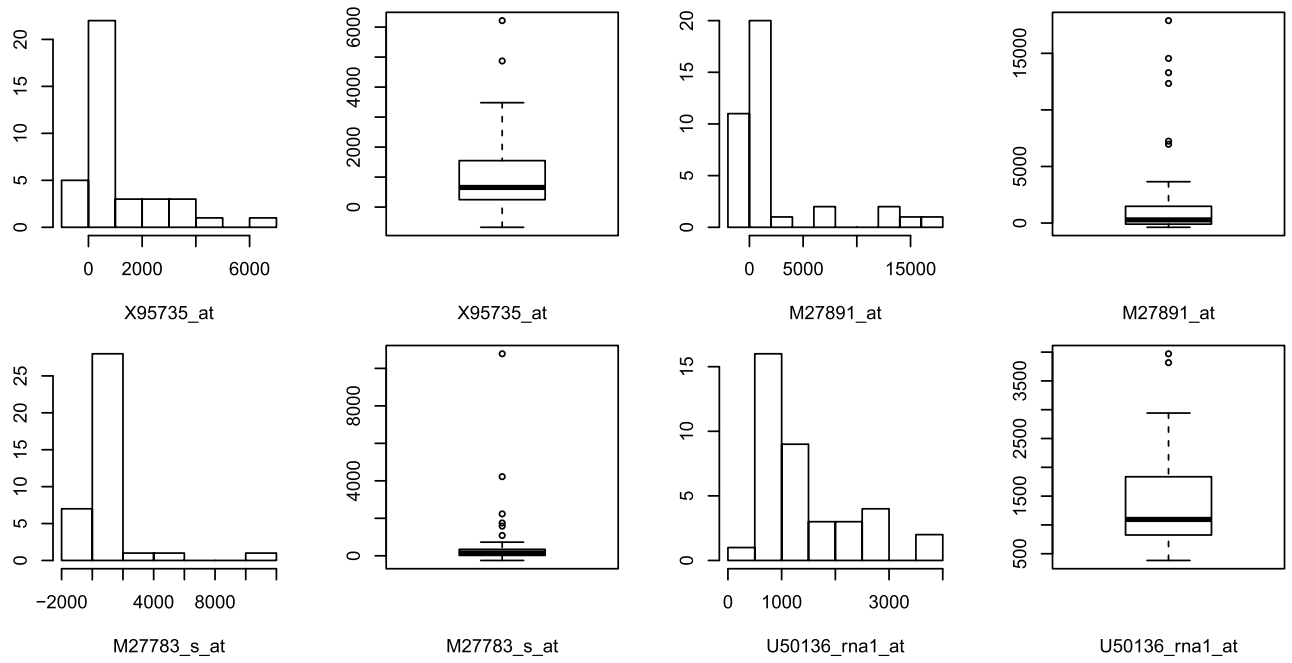

Figure 3. Histograms and boxplots for the four top ranked genes.

$$= 2\int \left\{ [1-F_X(y)]^2 \int_{-\infty}^{y} F_X^2(x) dF_x(x) \right\} dF_X(y)$$
$$= \frac{2}{3}\int F_X^3(y)[1-F_X(y)]^2 dF_X(y) = \frac{1}{90}.$$

This completes the proof. □

To prove Theorems 3.1 and Theorems 3.2, we need the following lemmas.

Lemma A.1 (Hoeffding's inequality [15]). *Let $X_1, \cdots, X_n$ be independent random variables. Assume that $P(X_i \in [a_i, b_i]) = 1$ for $1 \le i \le n$, where a_i and b_i are constants. Let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$. Then the following inequality holds*

$$P(|\bar{X} - E(\bar{X})| \ge t) \le 2exp\left\{ \frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right\}$$

where t is a positive constant and $E(\bar{X})$ is the expected value of $\bar{X}$.

Lemma A.2 (Dvoretzky-Kiefer-Wolfowitz inequality [24]). *Let $\widehat{F}_n$ denote the empirical distribution function for a sample of n i.i.d. random variables with distribution function F. Then for any $\epsilon > 0$, the following inequality holds*

$$P(\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \ge \epsilon) \le 2\exp\{-2n\epsilon^2\}.$$

We need the following notations for the next lemma to make these inequalities simple. Recall that $\widehat{F}_{k,Y}(x,y)$, $\widehat{F}_k(x), \widehat{F}_Y(y)$ are empirical functions of $F_{k,Y}(x,y), F_k(x)$, $F_Y(y)$ respectively, for $k = 1, \cdots, p, x, y \in \mathbb{R}$, where $F_{k,Y}(x,y), F_k(x), F_Y(y)$ are distribution functions of $(X_k, Y), X_k$, and Y respectively. Denote

$$\xi_Y = \frac{1}{n}\sum_{j=1}^n E_{X_k}[F_{k,Y}(X_k, Y_j) - F_k(X_k)F_Y(Y_j)]^2,$$
$$\eta_{X_k} = \frac{1}{n}\sum_{i=1}^n \left\{ \frac{1}{n}\sum_{j=1}^n [F_{k,Y}(X_{ik}, Y_j) - F_k(X_{ik})F_Y(Y_j)]^2 \right\},$$

where E_{X_k} is to compute expectations about X_k given Y_j.

Lemma A.3. *For any $\epsilon > 0$, the following inequalities are valid*

$$P\left(\left| \xi_Y - E\xi_Y \right| \ge \epsilon \right) \le 2\exp\{-2n\epsilon^2\}; \tag{A.1}$$

$$P\left(\left| \eta_{X_k} - E\eta_{X_k} \right| \ge \epsilon \right) \le 2\exp\{-2n\epsilon^2\}; \tag{A.2}$$

$$P\left(\sup_{x,y\in\mathbb{R}} \left| \widehat{F}_{k,Y}(x,y) - F_{k,Y}(x,y) \right| \ge \epsilon \right) \le 2(n+1)\exp\{-2n\epsilon^2\}; \tag{A.3}$$

$$P\left(\sup_{x,y\in\mathbb{R}} \left| \widehat{F}_k(x)\widehat{F}_Y(y) - F_k(x)F_Y(y) \right| \ge \epsilon \right) \le 4\exp\{-\frac{n\epsilon^2}{2}\}. \tag{A.4}$$

Proof. For any $x \in \mathbb{R}$, $E_{X_k}[F_{k,Y}(X_k, Y_j) - F_k(X_k)F_Y(Y_j)]^2 \le 1$, so we can apply Hoeffding in-

equality to obtain inequality (A.1). Similarly, we can get inequality (A.2).

Note that $\widehat{F}_{k,Y}(x,y)$ is the empirical distribution function of $F_{k,Y}(x,y)$, and $\left|\widehat{F}_{k,Y}(x,y)-F_{k,Y}(x,y)\right| \leq 1$, then we can also use the Hoeffding inequality and the Maximal Inequality of empirical process theory [25, p. 15] to obtain inequality (A.3).

To prove the inequality (A.4), we first get the bound of

$$\sup_{x,y\in\mathbb{R}}\left|\widehat{F}_k(x)\widehat{F}_Y(y)-F_k(x)F_Y(Y)\right|.$$

It is easy to show that

$$\begin{aligned}
&\sup_{x,y\in\mathbb{R}}\left|\widehat{F}_k(x)\widehat{F}_Y(y)-F_k(x)F_Y(Y)\right| \\
\leq\ & \sup_{x,y\in\mathbb{R}}\widehat{F}_Y(y)\left|\widehat{F}_k(x)-F_k(x)\right|+ \\
& \sup_{x,y\in\mathbb{R}}\widehat{F}_k(x)\left|\widehat{F}_Y(y)-F_Y(y)\right| \\
\leq\ & \sup_{x\in\mathbb{R}}\left|\widehat{F}_k(x)-F_k(x)\right|+\sup_{y\in\mathbb{R}}\left|\widehat{F}_Y(y)-F_Y(y)\right|,
\end{aligned}$$

where the first inequality follows from triangular inequality, and the second inequality holds by $\sup_{x\in\mathbb{R}}\widehat{F}_k(x)\leq 1$ and $\sup_{y\in\mathbb{R}}\widehat{F}_Y(y)\leq 1$.

So we can obtain:

$$\begin{aligned}
&P\left(\sup_{x,y\in\mathbb{R}}\left|\widehat{F}_k(x)\widehat{F}_Y(y)-F_k(x)F_Y(Y)\right|\geq\epsilon\right) \\
\leq\ & P\left(\sup_{x\in\mathbb{R}}\left|\widehat{F}_k(x)-F_k(x)\right|\geq\frac{\epsilon}{2}\right)+ \\
& P\left(\sup_{y\in\mathbb{R}}\left|\widehat{F}_Y(y)-F_Y(y)\right|\geq\frac{\epsilon}{2}\right) \\
\leq\ & 4exp\{-\frac{n\epsilon^2}{2}\},
\end{aligned}$$

where the last inequality follows from the Dvoretzky-Kiefer-Wolfowitz inequality. This completes the proof of Lemma A.3. □

Lemma A.4. *For any $\epsilon>0$ and $k=1,\cdots,p$, we have*

$$P(|\hat{\omega}_k-\omega_k|\geq\epsilon)\leq 2(n+1)\exp\left\{-\frac{n\epsilon^2}{72}\right\}+ 4\exp\left\{-\frac{n\epsilon^2}{288}\right\}+4\exp\left\{-\frac{2n\epsilon^2}{9}\right\}.$$

Proof. According to the definition of $\hat{\omega}_k$ and ω_k, we have

$$\begin{aligned}
&\hat{\omega}_k-\omega_k \\
&=\int\left[\widehat{F}_{k,Y}(x,y)-\hat{F}_k(x)\hat{F}_Y(Y)\right]^2 d\hat{F}_k(x)d\hat{F}_Y(y) \\
&\quad-\int\left[F_{k,Y}(x,y)-F_k(x)F_Y(Y)\right]^2 dF_k(x)dF_Y(y) \\
&=\int\left[\widehat{F}_{k,Y}(x,y)-\hat{F}_k(x)\hat{F}_Y(Y)\right]^2 d\hat{F}_k(x)d\hat{F}_Y(y) \\
&\quad-\left[F_{k,Y}(x,y)-F_k(x)F_Y(Y)\right]^2 d\hat{F}_k(x)d\hat{F}_Y(y) \\
&\quad+\int\left[F_{k,Y}(x,y)-F_k(x)F_Y(Y)\right]^2 \\
&\quad\times\left[d\hat{F}_k(x)-dF_k(x)\right]d\hat{F}_Y(y) \\
&\quad+\int\left[F_{k,Y}(x,y)-F_k(x)F_Y(Y)\right]^2 \\
&\quad\times\left[d\hat{F}_Y(y)-dF_Y(y)\right]d\hat{F}_k(x) \\
&=:\Delta_{k1}+\Delta_{k2}+\Delta_{k3}.
\end{aligned}$$

Then, we consider the part Δ_{k1} firstly,

$$\begin{aligned}
&|\Delta_{k1}| \\
=\ & \left|\int\Big(\left[\widehat{F}_{k,Y}(x,y)-\hat{F}_k(x)\hat{F}_Y(Y)\right]^2\right. \\
& \left.-\left[F_{k,Y}(x,y)-F_k(x)F_Y(Y)\right]^2\Big)d\hat{F}_k(x)d\hat{F}_Y(y)\right| \\
\leq\ & 2\int\left|\left[\hat{F}_{k,Y}(x,y)-F_{k,Y}(x,y)\right]\right. \\
& \left.-\left[\hat{F}_k(x)\hat{F}_Y(y)-F_k(x)F_Y(y)\right]\right|d\hat{F}_k(x)d\hat{F}_Y(y) \\
\leq\ & 2\sup_{x,y\in\mathbb{R}}|\hat{F}_{k,Y}(x,y)-F_{k,Y}(x,y)| \\
& +2\sup_{x,y\in\mathbb{R}}|\hat{F}_k(x)\hat{F}_Y(y)-F_k(x)F_Y(y)|
\end{aligned}$$

where the first inequality comes from

$$\begin{aligned}
&\left|\left[\hat{F}_{k,Y}(x,y)-F_{k,Y}(x,y)\right]\right. \\
&\qquad\left.+\left[\hat{F}_k(x)\hat{F}_Y(y)-F_k(x)F_Y(y)\right]\right| \\
&\leq 1+1=2,
\end{aligned}$$

and the second inequality follows from $\int d\hat{F}_k(x)d\hat{F}_Y(y)=1$. From inequality (A.3) and (A.4) of Lemma A.3, we can obtain that

$$\begin{aligned}
P(|\Delta_{k1}|\geq\epsilon)&\leq P(\sup_{x,y\in\mathbb{R}}|\hat{F}_{k,Y}(x,y)-F_{k,Y}(x,y)|\geq\frac{\epsilon}{4}) \\
&\quad+P(\sup_{x,y\in\mathbb{R}}|\hat{F}_k(x)\hat{F}_Y(y)-F_k(x)F_Y(y)|\geq\frac{\epsilon}{4}) \\
&\leq 2(n+1)\exp\{-\frac{n\epsilon^2}{8}\}+4\exp\{-\frac{n\epsilon^2}{32}\}. \qquad \text{(A.5)}
\end{aligned}$$

For the part Δ_{k2}, it follows from simple calculation that,

$$\begin{aligned}
\Delta_{k2}=&\int\left[F_{k,Y}(x,y)-F_k(x)F_Y(Y)\right]^2 d\hat{F}_k(x)d\hat{F}_Y(y) \\
&-\int\left[F_{k,Y}(x,y)-F_k(x)F_Y(Y)\right]^2 dF_k(x)d\hat{F}_Y(y)
\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{n}\sum_{j=1}^{n}[F_{k,Y}(X_{ik},Y_j) - F_k(X_{ik})F_Y(Y_j)]^2\Big\}$$
$$- E_{X_k}\Big\{\frac{1}{n}\sum_{j=1}^{n}[F_{k,Y}(X_k,Y_j) - F_k(X_k)F_Y(Y_j)]^2\Big\}$$
$$= \eta_{X_k} - E\eta_{X_k}.$$

Then we can use the inequality (A.2) of Lemma A.3 to obtain:

$$P(|\Delta_{k2}| \geq \epsilon) \leq 2\exp\{-2n\epsilon^2\}. \tag{A.6}$$

Lastly, we deal with Δ_{k3}, and we have

$$\begin{aligned}\Delta_{k3} &= \int [F_{k,Y}(x,y) - F_k(x)F_Y(Y)]^2\, dF_k(x)d\hat{F}_Y(y)\\ &\quad - \int [F_{k,Y}(x,y) - F_k(x)F_Y(Y)]^2\, dF_k(x)dF_Y(y)\\ &= \frac{1}{n}\sum_{j=1}^{n} E_{X_k}\left[F_{k,Y}(X_k,Y_j) - F_k(X_k)F_Y(Y_j)\right]^2\\ &\quad - E_Y E_{X_k}\left[F_{k,Y}(X_k,Y) - F_k(X_k)F_Y(Y)\right]^2\\ &= \xi_Y - E\xi_Y.\end{aligned}$$

So, by the inequality (A.1), we can obtain:

$$P(|\Delta_{k3}| \geq \epsilon) \leq 2\exp\{-2n\epsilon^2\}. \tag{A.7}$$

Inequalities (A.5)–(A.7) together imply that

$$\begin{aligned}&P(|\hat{\omega}_k - \omega_k| \geq \epsilon)\\ \leq\ & P\left(|\Delta_{k1}| \geq \frac{\epsilon}{3}\right) + P\left(|\Delta_{k2}| \geq \frac{\epsilon}{3}\right) + P\left(|\Delta_{k3}| \geq \frac{\epsilon}{3}\right)\\ \leq\ & 2(n+1)\exp\left\{-\frac{n\epsilon^2}{72}\right\}\\ & +4\exp\left\{-\frac{n\epsilon^2}{288}\right\} + 4\exp\left\{-\frac{2n\epsilon^2}{9}\right\}.\end{aligned}$$

This obtains the result of Lemma A.4. □

Proof of Theorem 3.1. Firstly, we prove the first part of the Theorem 3.1. By Lemma A.4, for any positive constant c, we have

$$\begin{aligned}&P(\max_{1\leq k\leq p}|\hat{\omega}_k - \omega_k| \geq cn^{-\kappa})\\ \leq\ & pP(|\hat{\omega}_k - \omega_k| \geq cn^{-\kappa})\\ \leq\ & p\Big\{2(n+1)\exp\{-c_1 n^{1-2\kappa}\}\\ & +4\exp\{-c_2 n^{1-2\kappa}\} + 4\exp\{-c_3 n^{1-2\kappa}\}\Big\},\end{aligned}$$

where $c_1 = \frac{c}{72}, c_2 = \frac{c}{288}, c_3 = \frac{2c}{9}$.

Next, we show the second part of Theorem 3.1. Denote the set

$$\Gamma_n = \Big\{\max_{k\in\mathcal{A}}|\hat{\omega}_k - \omega_k| \leq \frac{c_0 n^{-\kappa}}{2}\Big\},$$

on this set, by condition (C1): $\min_{k\in\mathcal{A}}\omega_k \geq c_0 n^{-\kappa}$, we have

$$\hat{\omega}_k \geq \omega_k - |\hat{\omega}_k - \omega_k| \geq \frac{c_0 n^{-\kappa}}{2}.$$

Therefore, by the choice of $\tau_n = c_4 n^{-\kappa}, c_4 \leq \frac{c_0}{2}$, and Lemma A.4, we have

$$\begin{aligned}&P(\mathcal{A} \subset \widehat{\mathcal{A}}_{\tau_n}) \geq P(\Gamma_n)\\ &\geq 1 - s\Big\{2(n+1)\exp\{-c_1 n^{1-2\kappa}\}\\ &\quad + 4\exp\{-c_2 n^{1-2\kappa}\} + 4\exp\{-c_3 n^{1-2\kappa}\}\Big\}.\end{aligned}$$ □

Proof of Theorem 3.2. Recall the assumption of condition (C2), we know that $\delta = \min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{I}} > 0$. Thus,

$$\begin{aligned}&P\Big(\max_{k\in\mathcal{I}}\widehat{\omega}_k \geq \min_{k\in\mathcal{A}}\widehat{\omega}_k\Big)\\ \leq\ & P\Big(|\max_{k\in\mathcal{I}}\widehat{\omega}_k - \max_{k\in\mathcal{I}}\omega_k| + |\min_{k\in\mathcal{A}}\widehat{\omega}_k - \min_{k\in\mathcal{A}}\omega_k| \geq \delta\Big)\\ \leq\ & P\Big(\max_{k\in\mathcal{I}}|\widehat{\omega}_k - \omega_k| \geq \frac{\delta}{2}\Big) + P\Big(\min_{k\in\mathcal{A}}|\widehat{\omega}_k - \omega_k| \geq \frac{\delta}{2}\Big)\\ \leq\ & 2P\Big(\max_{1\leq k\leq p}|\widehat{\omega}_k - \omega_k| \geq \frac{\delta}{2}\Big),\end{aligned}$$

by using Lemma A.4, we can obtain the first inequality of Theorem 3.2. For the second statement,

$$\begin{aligned}&P\Big(\min_{k\in\mathcal{A}}\widehat{\omega}_k - \max_{k\in\mathcal{I}}\widehat{\omega}_k < \frac{\delta}{2}\Big)\\ \leq\ & P\left(\Big[\min_{k\in\mathcal{A}}\widehat{\omega}_k - \max_{k\in\mathcal{I}}\widehat{\omega}_k\Big] - \Big[\min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{I}}\omega_k\Big] < -\frac{\delta}{2}\right)\\ \leq\ & P\left(\left|\Big[\min_{k\in\mathcal{A}}\widehat{\omega}_k - \max_{k\in\mathcal{I}}\widehat{\omega}_k\Big] - \Big[\min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{I}}\omega_k\Big]\right| > \frac{\delta}{2}\right)\\ \leq\ & P\left(2\max_{1\leq k\leq p}|\widehat{\omega}_k - \omega_k| > \frac{\delta}{2}\right)\\ \leq\ & 2\exp\{\log(p) + \log(n+1) - c_1^* n\}\\ & +4\exp\{\log(p) - c_2^* n\} + 4\exp\{\log(p) - c_3^* n\},\end{aligned}$$

where the last inequality comes from Lemma A.4, and c_1^*, c_2^*, c_3^* are some positive constants.

If $\log(p) = o(n^{1-2\kappa})$, we can obtain that $\max\Big(\log(p) + \log(n+1) - c_1^* n, log(p) - c_2^* n, \log(p) - c_3^* n\Big) \leq -2\log(n)$, for large n.

So, for some N,

$$\begin{aligned}&\sum_{n=N}^{+\infty} P\Big(\min_{k\in\mathcal{A}}\widehat{\omega}_k - \max_{k\in\mathcal{I}}\widehat{\omega}_k < \frac{\delta}{2}\Big)\\ &\leq \sum_{n=N}^{+\infty} 10\exp\{-2\log(n)\} < +\infty.\end{aligned}$$

Then, the conclusion follows from the Borel Contelli Lemma. □

ACKNOWLEDGEMENTS

The authors would like to thank the Editor, the Associate Editor and two Referees for their constructive comments which greatly improved the paper.

Received 7 December 2016

REFERENCES

[1] Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *The annals of mathematical statistics*, pages 485–498. MR0125690

[2] Cui, H., Li, R., and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510):630–641. MR3367253

[3] de Siqueira Santos, S., Takahashi, D. Y., Nakata, A., and Fujita, A. (2013). A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in bioinformatics*, 15(6):906–918.

[4] Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605.

[5] Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494).

[6] Fan, J. and 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.

[7] Fan, J., Ma, Y., and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109(507):1270–1284.

[8] Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *The Journal of Machine Learning Research*, 10:2013–2038.

[9] Fan, J., Song, R., et al. (2010). Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604.

[10] Fujita, A., Sato, J. R., Demasi, M. A. A., Sogayar, M. C., Ferreira, C. E., and Miyano, S. (2009). Comparing pearson, spearman and hoeffding's d measure for gene expression association analysis. *Journal of bioinformatics and computational biology*, 7(04):663–684.

[11] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

[12] 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(3).

[13] He, X., Wang, L., Hong, H. G., et al. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41(1):342–369. MR3059421

[14] Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*, pages 546–557.

[15] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30.

[16] Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric statistical methods*. John Wiley & Sons. MR3221959

[17] Huang, D., Li, R., and Wang, H. (2014). Feature screening for ultrahigh dimensional categorical data with applications. *Journal of Business & Economic Statistics*, 32(2):237–244.

[18] Huang, J., Horowitz, J. L., and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, pages 587–613.

[19] Li, G., Peng, H., Zhang, J., Zhu, L., et al. (2012a). Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877.

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

[21] Lin, L., Sun, J., and Zhu, L. (2013). Nonparametric feature screening. *Computational Statistics & Data Analysis*, 67:162–174.

[22] Liu, J., Li, R., and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 109(505):266–274.

[23] Mai, Q. and Zou, H. (2012). The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1):229–234. MR3034336

[24] Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283. MR1062069

[25] Pollard, D. (1984). *Convergence of stochastic processes*. Springer-Verlag New York Inc.

[26] Shao, X. and Zhang, J. (2014). Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318.

[27] Song, R., Lu, W., Ma, S., and Jessie Jeng, X. (2014a). Censored rank independence screening for high-dimensional survival data. *Biometrika*, 101(4):799–814.

[28] Song, R., Yi, F., and Zou, H. (2014b). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, 24(4):1735.

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

[30] Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

[31] Zhao, S. D. and Li, Y. (2012). Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of multivariate analysis*, 105(1):397–411.

[32] Zhong, W. (2014). Robust sure independence screening for ultrahigh dimensional non-normal data. *Acta Mathematica Sinica, English Series*, 30(11):1885–1896.

[33] Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496). MR2896849

Yuan Yu
School of Statistics
Shandong University of Finance and Economics,
Jinan 250014, China
School of Statistics and Management
Shanghai University of Finance and Economics
Shanghai 200433, China
E-mail address: yuyuan_mail@126.com

Di He
School of Statistics and Management
Shanghai University of Finance and Economics
200433 China
E-mail address: hedi8910@163.com

Yong Zhou
Institute of Statistics and Interdisciplinary Sciences
and School of Statistics
Faculty of Economics and Management
East China Normal University
Shanghai 200241, China
E-mail address: yzhou@amss.ac.cn