Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis

Hengjian Cui, Runze Li & Wei Zhong

To cite this article: Hengjian Cui, Runze Li & Wei Zhong (2015) Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis, Journal of the American Statistical Association, 110:510, 630-641, DOI: 10.1080/01621459.2014.920256

To link to this article: https://doi.org/10.1080/01621459.2014.920256

View supplementary material

Accepted author version posted online: 13 May 2014.
Published online: 13 May 2014.

Submit your article to this journal

Article views: 1464

View Crossmark data

Citing articles: 37 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis

Hengjian CUI, Runze LI, and Wei ZHONG

This work is concerned with marginal sure independence feature screening for ultrahigh dimensional discriminant analysis. The response variable is categorical in discriminant analysis. This enables us to use the conditional distribution function to construct a new index for feature screening. In this article, we propose a marginal feature screening procedure based on empirical conditional distribution function. We establish the sure screening and ranking consistency properties for the proposed procedure without assuming any moment condition on the predictors. The proposed procedure enjoys several appealing merits. First, it is model-free in that its implementation does not require specification of a regression model. Second, it is robust to heavy-tailed distributions of predictors and the presence of potential outliers. Third, it allows the categorical response having a diverging number of classes in the order of $O(n^{\kappa})$ with some $\kappa \geq 0$. We assess the finite sample property of the proposed procedure by Monte Carlo simulation studies and numerical comparison. We further illustrate the proposed methodology by empirical analyses of two real-life datasets. Supplementary materials for this article are available online.

KEY WORDS: Consistency in ranking; Sure screening property; Ultrahigh dimensional data analysis

1. INTRODUCTION

Variable selection plays an important role in high-dimensional data analysis. Marginal feature screening becomes indispensable for ultrahigh dimensional data and has received much attention in the very recent literature. Various feature screening procedures have been proposed for linear models, generalized linear models, and robust linear models (Fan and Lv 2008; Fan, Samworth, and Wu 2009; Wang 2009; Li et al. 2012). These authors demonstrated that their procedures enjoy sure screening property in the terminology of Fan and Lv (2008). Feature screening procedures have been further proposed for nonparametric regression models in the literature. Fan, Feng, and Song (2011) proposed a nonparametric marginal screening procedure for additive models based on B-spline expansion. Fan, Ma, and Dai (2014) further extended the nonparametric B-spline method for varying coefficient models and proposed a marginal sure screening procedure. Liu, Li, and Wu (2014) proposed a local kernel-based marginal sure screening procedure for varying coefficient models and further established its sure screening property. The aforementioned model-based screening procedures perform well when the underlying models are correctly specified, but their performance may be quite poor in the presence of model mis-specification. Specifying a correct model for ultrahigh dimensional data may be challenging. Thus, model-free sure screening procedures are appealing and have been developed by several authors (Zhu et al. 2011; Li, Zhong, and Zhu 2012; He, Wang, and Hong 2013). Li, Zhong, and Zhu (2012) developed a sure independence screening procedure based on the distance correlation which is model-free. Its sure screening property requires subexponential tail probability conditions on predictors and response, and it is not robust to very heavy-tailed data with extreme values. Mai and Zou (2013) developed a sure feature screening procedures with ultrahigh dimensional predictors based on the Kolmogorov distance, but it is studied only for binary classification problems. Pan, Wang, and Li (2013) proposed a pairwise sure screening procedure for linear discriminant analysis with a diverging number of classes and ultrahigh dimensional predictors. However, it is based on mean difference and cannot perform well for heavy-tailed data. This work aims to develop an effective model-free and robust feature screening procedure for ultrahigh dimensional discriminant analysis with a possibly diverging number of classes.

In this article, we propose an effective sure screening procedure for discriminant analysis. We further study its theoretical properties and establish the sure screening and rank consistency properties without assuming the moment conditions on predictors under the settings of ultrahigh dimensional discriminant analysis with a diverging number of response classes. Our numerical studies show that the proposed procedure has excellent performance. It enjoys several appealing properties. It is model-free since its implementation does not require specification of the regression model. Its corresponding marginal utility may be easily evaluated without involving numerical optimization.

Due to its nature, the proposed procedure can be directly applied for continuous response with categorical predictors. This indeed is also very useful in the genomics-wide association study (GWAS), in which the phenotypes (i.e., the responses) are continuous, and the single-nucleotide polymorphisms (SNPs) as predictors are categorical. Thus, it is also of interest to develop an effective feature screening procedure for setting in which

Hengjian Cui is Professor, Department of Statistics, Capital Normal University, China (E-mail: *hjcui@bnu.edu.cn*). Runze Li is Distinguished Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111 (E-mail: *rzli@psu.edu*). Wei Zhong is Corresponding Author, Assistant Professor, Wang Yanan Institute for Studies in Economics (WISE), Department of Statistics and Fujian Key Laboratory of Statistical Science, Xiamen University, China (E-mail: *wzhong@xmu.edu.cn*).
Cui's research was supported by National Natural Science Foundation of China (NNSFC) grants 11071022, 11028103, 11231010 and Key project of Beijing Municipal Educational Commission and Beijing Center for Mathematics and Information Interdisciplinary Sciences. Li's research was supported by National Institute on Drug Abuse (NIDA) grants P50-DA10075 and P50 DA036107, and NNSFC grant 11028103. Zhong's research was supported by NNSFC grants 11301435 and 71131008 and the Fundamental Research Funds for the Central Universities. All authors equally contributed to this article. The authors are listed in alphabetical order. The authors thank the Editor, the AE, and reviewers for their constructive comments, which have greatly improved the earlier version of this article. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NNSFC or NIDA.

© 2015 American Statistical Association
Journal of the American Statistical Association
June 2015, Vol. 110, No. 510, Theory and Methods
DOI: 10.1080/01621459.2014.920256

the response is continuous, while the predictors of interest are categorical. In this article, we further extend our procedure for such settings. Some further extensions are also discussed in Section 4.

The rest of this article is organized as follows. In Section 2, we propose a new marginal utility for feature screening and further study its theoretical properties. In Section 3, we conduct Monte Carlo simulation studies to examine the finite sample performance of the proposed procedure. We further illustrate the proposed methodology by empirical analyses of real data examples. Section 4 presents some extensions of the proposed methodology. Technical proofs are given in the Appendix.

2. A NEW FEATURE SCREENING PROCEDURE

2.1 A New Index Based on Conditional Distribution Function

Let Y be a categorical response with R classes $\{y_1, y_2, \ldots, y_R\}$, and X be a continuous covariate with a support $\mathbb{R}_X$. To investigate the dependence relationship between X and Y, we naturally consider the conditional distribution function of X given Y, denoted by $F(x|Y) = \mathbb{P}(X \le x|Y)$. Denote by $F(x) = \mathbb{P}(X \le x)$ the unconditional distribution function of X and $F_r(x) = \mathbb{P}(X \le x|Y = y_r)$ the conditional distribution function of X given $Y = y_r$. If $F_r(x) = F(x)$ for any $x \in \mathbb{R}_X$ and $r = 1, 2, \ldots, R$, then X and Y are independent. This motivates us to consider the index

$$\mathrm{MV}(X|Y) = E_X[\mathrm{var}_Y(F(X|Y))] \tag{2.1}$$

to measure the dependence between X and Y. The following proposition provides the properties of the $\mathrm{MV}(X|Y)$.

Proposition 2.1. Let Y be a categorical random variable with R classes $\{y_1, y_2, \ldots, y_R\}$ and $p_r = \mathbb{P}(Y = y_r) > 0$ for all $r = 1, \ldots, R$. Let X be a continuous random variable with support $\mathbb{R}_X$. Denote $F(x) = \mathbb{P}(X \le x)$ and $F_r(x) = \mathbb{P}(X \le x|Y = y_r)$, then

(1) $\mathrm{MV}(X|Y) = \sum_{r=1}^{R} p_r \int [F_r(x) - F(x)]^2 dF(x)$.
(2) $\mathrm{MV}(X|Y) = 0$ if and only if X and Y are statistically independent.

The proof of this proposition is given in the Appendix. The results in (1) implies that the $\mathrm{MV}(X|Y)$ can be represented as the weighted average of Cramér-von Mises distances between the conditional distribution function of X given $Y = y_r$ and the unconditional distribution function of X. The second remarkable property motivates us to use the $\mathrm{MV}(X|Y)$ as a marginal utility for feature screening to characterize both linear and nonlinear relationships for ultrahigh dimensional discriminant analysis.

Let $\{(X_i, Y_i) : 1 \le i \le n\}$ be a random sample of size n from the population (X, Y). Define $\hat{p}_r = \frac{1}{n}\sum_{i=1}^{n} I\{Y_i = y_r\}$ with $I\{\cdot\}$ being the indicator function, $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x\}$, and $\hat{F}_r(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x, Y_i = y_r\}/\hat{p}_r$. It is natural to use its sample counterpart to estimate $\mathrm{MV}(X|Y)$ as follows:

$$\widehat{\mathrm{MV}}(X|Y) = \frac{1}{n}\sum_{r=1}^{R}\sum_{j=1}^{n} \hat{p}_r[\hat{F}_r(X_j) - \hat{F}(X_j)]^2. \tag{2.2}$$

To get insights into $\mathrm{MV}(X|Y)$, let us consider a simple example. Let X be a univariate standard normal random variable and generate random variables Z_k with $k = 1, 2$ by $Z_1 = cX + \varepsilon$ and $Z_2 = cX^2 + \varepsilon$, where $\varepsilon \sim N(0, 1)$ and c is a constant to control the signal-to-noise ratio. Then, we equally discretize each Z_k to a categorical variable Y_k with four classes. That is, $Y_k = I(Z_k \le q_{k1}) + 2I(q_{k1} < Z_k \le q_{k2}) + 3I(q_{k2} < Z_k \le q_{k3}) + 4I(Z_k > q_{k3}), k = 1, 2$ where $\{q_{k1}, q_{k2}, q_{k3}\}$ are the first, second, and third quartiles of Z_k, respectively. Thus, the response Y_1 depends on X through a linear term cX, while Y_2 depends on X through a quadratic term cX^2. We set sample size $n = 200$ and $c = 0, 0.5, 1$, and 2. Note that Y_k and X are independent for each $k = 1, 2$ when $c = 0$. Then, we compute the variance of conditional distribution function of X given Y_k, that is, $\mathrm{var}_{Y_k}[F(x|Y_k)]$, for $x \in [-2, 2]$ and each c. Panels (a) and (c) in Figure 1 are boxplots of $\mathrm{var}_{Y_k}[F(x|Y_k)]$ against different c values for $k = 1, 2$, respectively, where the star indicates $\widehat{\mathrm{MV}}(X|Y_k)$. Panels (b) and (d) in Figure 1 demonstrate how $\mathrm{var}_{Y_k}[F(x|Y_k)]$ with $k = 1, 2$ varies across $x \in [-2, 2]$ for different c values. It is shown that as the signal-to-noise ratio increases, $\widehat{\mathrm{MV}}(X|Y_k)$ increases. When $c = 0$, that is, X and Y_k are independent, $\widehat{\mathrm{MV}}(X|Y_k)$ are nearly close to zero; When $c > 0$, they are remarkably different above from zero. Consequently, the $\mathrm{MV}(X|Y)$ should be an effective measure to characterize and strengthen both linear and nonlinear dependence between a continuous covariate and a categorical response.

2.2 Sure Independence Screening Using MV($X|Y$)

We now propose a new model-free sure independence screening using $\mathrm{MV}(X|Y)$ for ultrahigh dimensional discriminant analysis. Let Y be the response with discrete support $\{y_1, y_2, \ldots, y_R\}$ with $R \ge 2$ and $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ be the predictor vector, where $p \gg n$ and n is the sample size. Without specifying a regression model, define the active predictor subset by

$$\mathcal{D} = \{k : F(y|\mathbf{x}) \text{ functionally depends on } X_k \text{ for some } y = y_r\},$$

and denote by $\mathcal{I} = \{1, 2, \ldots, p\} \setminus \mathcal{D}$ the inactive predictor subset.

The goal is to select a reduced model with a moderate scale which can almost fully contain $\mathcal{D}$ using an independence screening method for ultrahigh dimensional discriminant analysis. To this end, we apply the MV index for each pair (X_k, Y):

$$\omega_k = \mathrm{MV}(X_k|Y)$$

as a marginal utility to measure the importance of X_k for the response, where $k = 1, 2, \ldots, p$. Note that $\omega_k = 0$ if and only if X_k and Y are statistically independent. As a motivation, we can see that, if the partial orthogonality condition (Huang, Horowitz, and Ma 2008; Fan and Song 2010) holds, that is, $\{X_k : k \in \mathcal{D}\}$ are statistically independent of $\{X_k : k \in \mathcal{I}\}$, then ω_k is a naturally effective measure to sperate the active and inactive predictor subsets because $\omega_k > 0$ for $k \in \mathcal{D}$ and $\omega_k = 0$ for $k \in \mathcal{I}$. It also implies that the MV-based variable screening is model-free in that it is defined through conditional and unconditional distribution functions and able to characterize both linear and nonlinear relationships between the response and predictors.

For a random sample $\{(\mathbf{x}_i, Y_i) : 1 \le i \le n\}$, we can easily estimate ω_k by setting $\hat{\omega}_k = \widehat{\mathrm{MV}}(X_k|Y)$ according to Equation

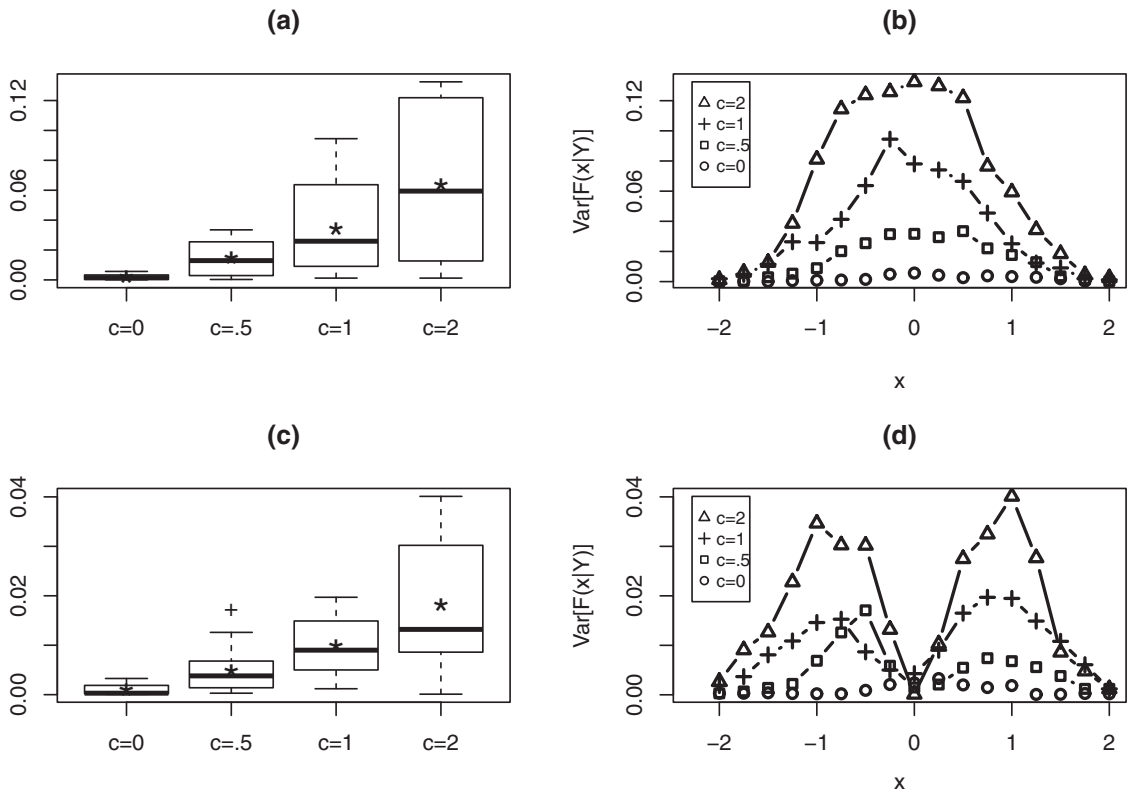

Figure 1. (a) Boxplot of $\text{var}_{Y_1}[F(x|Y_1)]$ against c with the star indicating the mean; (b) plot of $\text{var}_{Y_1}[F(x|Y_1)]$ against x for different c values; (c) boxplot of $\text{var}_{Y_2}[F(x|Y_2)]$ against c with the star indicating the mean; (d) plot of $\text{var}_{Y_2}[F(x|Y_2)]$ against x for different c values.

(2.2). Then we propose to use $\hat{\omega}_k$ to choose a submodel

$$\hat{\mathcal{D}} = \{k : \hat{\omega}_k \geq cn^{-\tau}, \text{ for } 1 \leq k \leq p\},$$

where c and τ are predetermined thresholding values defined in Condition (C2). In practice, for a given size $d < n$, one can select a reduced model:

$$\hat{\mathcal{D}}^* = \{k : \hat{\omega}_k \text{ is among the top } d \text{ largest of all}\}.$$

We refer this procedure to the MV-based sure independence screening, MV-SIS for short.

Next, we study the theoretical properties of the proposed MV-SIS. Fan and Lv (2008) and Ji and Jin (2012) demonstrated that the two-stage procedure combining independence screening and penalized estimation can outperform an one-step penalized estimation approach, such as LASSO. The effectiveness of the two-stage procedure is guaranteed by the sure screening property. That is, all active predictors can be included in the reduced model with high probability. Thus, we first establish the sure screening property for MV-SIS with assuming the following conditions:

(C1) There exist two positive constants c_1 and c_2 such that $c_1/R_n \leq \min_{1\leq r\leq R_n} p_r \leq \max_{1\leq r\leq R_n} p_r \leq c_2/R_n$. Assume that $R_n = O(n^\kappa)$ for $\kappa \geq 0$.

(C2) There exists positive constants $c > 0$ and $0 \leq \tau < 1/2$ such that $\min_{k\in\mathcal{D}} \omega_k \geq 2cn^{-\tau}$.

Condition (C1) requires that the proportion of each class of the response cannot be either too small or too large. $R_n = O(n^\kappa)$ assumed in Condition (C1) allows the diverging number of classes of the response, where the subscript n in R_n is used to emphasize R_n being allowed to be diverging with the sample size n. Condition (C2) assumes that the minimum true signal cannot be too small and it is in the order of $n^{-\tau}$ which allows the minimum true signal to vanish to zero as the sample size n approaches the infinity. Such an assumption is typical in the feature screening literature (e.g., Condition 3 in Fan and Lv (2008), Condition (C3) in Wang (2009), Condition (C2) in both Li, Zhong, and Zhu (2012), and He, Wang, and Hong (2013) etc). The following theorem presents the sure screening property of MV-SIS and its proof is provided in the Appendix.

Theorem 2.1. [Sure Screening Property] Under Condition (C1) and for any $0 \leq \kappa < 1 - 2\tau$, there exists a positive constant b depending on c, c_1, and c_2, such that

$$\mathbb{P}\left(\max_{1\leq k\leq p} |\hat{\omega}_k - \omega_k| > cn^{-\tau}\right) \leq O\left(p\exp\{-bn^{1-(2\tau+\kappa)} + (1+\kappa)\log n\}\right). \quad (2.3)$$

Under Conditions (C1) and (C2), we have that

$$\mathbb{P}(\mathcal{D} \subseteq \hat{\mathcal{D}}) \geq 1 - O\left(s_n \exp\{-bn^{1-(2\tau+\kappa)} + (1+\kappa)\log n\}\right), \quad (2.4)$$

where s_n is the cardinality of $\mathcal{D}$.

The sure screening property holds for MV-SIS under milder conditions than those for the SIS (Fan and Lv 2008) and DC-SIS (Li, Zhong, and Zhu 2012) in that we do not require the regression function of Y onto $\mathbf{x}$ to be linear and it needs little assumption on the moments of predictors. It is worth noting that MV-SIS is robust to heavy-tailed distributions of predictors

and the presence of potential outliers because $\mathrm{MV}(X_k|Y)$ inherits the robustness property of conditional distribution function. Furthermore, the sure screening property also holds for the categorical response with a diverging number of classes. Thus, the MV-SIS provides a unified alternative to existing model-based sure screening procedures for ultrahigh dimensional discriminant analysis.

According to Theorem 2.1, we know that MV-SIS can handle the NP-dimensionality $\log p = O(n^{\alpha})$, where $\alpha < 1 - 2\tau - \kappa$ with $0 \le \tau < 1/2$ and $0 \le \kappa < 1 - 2\tau$, which depends on the minimum true signal strengthen and the number of response classes. If R_n is fixed, that is, $\kappa = 0$, then the result of Theorem 2.1 is improved and its first part can be rewritten as

$$P\left\{\max_{1\le k\le p} |\hat{\omega}_k - \omega_k| > cn^{-\tau}\right\} \le O(p\exp\{-bn^{1-2\tau} + \log n\}),$$

for some constant $b > 0$. In this case, we can handle the even larger NP-dimensionality $\log p = O(n^{\alpha})$, where $\alpha < 1 - 2\tau$ with $0 \le \tau < 1/2$.

Remark. Condition (C1) can be relaxed in the way that c_1 is allowed to tend to zero in a certain rate. To be specific, we assume that $c_1 = O(n^{-\eta})$ with $0 < \eta < 2\tau + \kappa$. Under the relaxed condition, the sure screening property remains as essentially same as before, but the convergence rate becomes relatively slower. That is,

$$\mathbb{P}\left(\max_{1\le k\le p} |\hat{\omega}_k - \omega_k| > cn^{-\tau}\right) \le O\left(p\exp\{-bn^{1-(2\tau+\kappa+\eta)} + (1+\kappa)\log n\}\right).$$

Then, a smaller NP-dimensionality $\log p = O(n^{\alpha})$ with $\alpha < 1 - 2\tau - \kappa - \eta$ is allowed. For the proof, refer to Appendix A in the online supplement.

Another interesting property for independence screening is ranking consistency property in the terms of Zhu et al. (2011). To investigate the ranking consistency property of MV-SIS, we additionally assume the following condition:

(C3) $\liminf_{p\to\infty}\{\min_{k\in\mathcal{D}}\omega_k - \max_{k\in\mathcal{I}}\omega_k\} \ge c_3$, where $c_3 > 0$ is a constant.

It is easily shown that under the partial orthogonality condition (Huang, Horowitz, and Ma 2008) that $\omega_k > 0$ for $k \in \mathcal{D}$ and $\omega_k = 0$ for $k \in \mathcal{I}$, Condition (C3) naturally holds. Thus, Condition (C3) is a relatively weaker assumption than partial orthogonality condition. It requires the MV index is able to sperate active and inactive predictors well in the population level. The following theorem justifies the ranking consistency property of MV-SIS.

Theorem 2.2. [Ranking Consistency Property] If Conditions (C1) and (C3) hold for $R_n\log(n)/n = o(1)$ and $R_n\log(p)/n = o(1)$, then $\liminf_{n\to\infty}\{\min_{k\in\mathcal{D}}\hat{\omega}_k - \max_{k\in\mathcal{I}}\hat{\omega}_k\} > 0$, a.s.

Although it requires a more restrictive condition on the difference between active and inactive signals, Theorem 2.2 demonstrates a stronger theoretical result than sure screening property. That is, the sample $\mathrm{MV}(X_k|Y)$ values of active predictors are always ranked beyond those of inactive ones with high probability. Thus, with an ideal thresholding value, one might separate the active predictors and inactive predictors.

3. NUMERICAL STUDIES

In this section, we first assess the finite sample performance of the proposed MV-SIS by Monte Carlo simulation studies. Then, we conduct empirical analyses of two real data examples to illustrate the proposed MV-SIS procedure. Some additional numerical results are given in the supplementary material.

3.1 Monte Carlo Simulations

We use the minimum model size (MMS) to include all active predictors to measure the effectiveness of each screening approach. In addition, the proportion including a single active predictor X_j, denoted by $\mathcal{P}_j^s$, and the proportion including all active predictors, denoted by $\mathcal{P}_a$, are computed for a given model size $d = [n/\log n]$, where n is the sample size and $[x]$ denotes the integer part of x. All numerical studies are conducted using R code.

Example 3.1. (Ultrahigh Dimensional Linear Discriminant Analysis) In this example, we consider a linear discriminant analysis problem with ultrahigh dimensional predictors by following the similar settings in Pan, Wang, and Li (2013). For each ith observation, the categorical response Y_i is generated from two different distributions: (i) balanced, a discrete uniform distribution with R categories where $\mathbb{P}(Y_i = r) = 1/R$ with $r = 1, \ldots, R$; (ii) unbalanced, the sequence of probabilities $p_r = P(Y_i = r) = 2[1 + (r-1)/(R-1)]/3R$ is an arithmetic progression with $\max_{1\le r\le R} p_r = 2\min_{1\le r\le R} p_r$. For instance, when Y is binary, $p_1 = 1/3$ and $p_2 = 2/3$. Given $Y_i = r$, the ith predictor X_i is then generated by letting $X_i = \mu_r + \varepsilon_i$, where the mean term $\mu_r = (\mu_{r1}, \ldots, \mu_{rp}) \in \mathbb{R}^p$ is a p-dimensional vector with rth component $\mu_{rr} = 3$ but other components are all zero, and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})$ is a p-dimensional error term. Here, we consider two cases of the error term: (1) $\varepsilon_{ij} \sim N(0, 1)$; (2) $\varepsilon_{ij} \sim t(2)$ independently for each $j = 1, \ldots, p$. Note that the Case (2) makes each predictor heavy-tailed, which is designed to examine the robustness of an independence screening method. To systematically examine MV-SIS and other competitors, we will consider 2000 predictors and a binary response with $n = 40$, and a 10-categorical response with $n = 200$ for each case, respectively. That is, $(R, n, p) = (2, 40, 2000)$ and $(10, 200, 2000)$.

First, we compare the performance of MV-SIS with SIS (Fan and Lv 2008), SIRS (Zhu et al. 2011), DC-SIS (Li, Zhong, and Zhu 2012), Kolmogorov filter (Mai and Zou 2013), and PSIS (Pan, Wang, and Li 2013) for the binary response, where X_1 and X_2 are the active predictors. Table 1 summarizes the median of MMS with its associated robust estimate of the standard deviation (RSD = IQR/1.34) in the parentheses, $\mathcal{P}_j^s$ with $j = 1, 2$ and $\mathcal{P}_a$ for the given model size $d = [n/\log n]$ for each method based on 500 simulations.

Next, we consider the response with 10 categories, where $X_1, X_2, \ldots, X_{10}$ are active. Note that a value of the response Y is a nominal number, which makes SIS, SIRS, and Kolmogorov filter unapplicable. However, MV-SIS is proposed for variable screening with a multiple categorical response. To make DC-SIS applicable for this problem, we transfer the 10-categorical response to 9 dummy binary variables, which are together considered as a new multiple response. Note that Li, Zhong, and

Table 1. Simulation results for linear discriminant analysis with binary response

		Case (1): $\varepsilon_{ij} \sim N(0, 1)$				Case (2): $\varepsilon_{ij} \sim t(2)$			
p_r	Method	MMS	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_a$	MMS	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_a$
Balanced	SIS	2.0 (0.0)	1.00	1.00	1.00	2.5(9.1)	0.79	0.88	0.71
	SIRS	2.0 (0.0)	1.00	1.00	1.00	8.0(20.5)	0.71	0.76	0.55
	DC-SIS	2.0 (0.0)	1.00	1.00	1.00	2.0(0.0)	0.99	0.98	0.97
	KF	2.0 (0.0)	1.00	1.00	1.00	2.0(0.0)	0.99	0.99	0.98
	PSIS	2.0 (0.0)	1.00	1.00	1.00	2.5(9.1)	0.79	0.88	0.71
	MV-SIS	2.0 (0.0)	1.00	1.00	1.00	2.0(0.0)	1.00	0.99	0.99
Unbalanced	SIS	2.0 (0.0)	1.00	1.00	1.00	5.5(48.8)	0.75	0.75	0.55
	SIRS	2.0 (0.0)	1.00	0.99	0.99	17.0(123.3)	0.67	0.64	0.44
	DC-SIS	2.0 (0.0)	1.00	1.00	1.00	2.0(1.1)	0.95	0.96	0.92
	KF	2.0 (0.0)	1.00	1.00	1.00	2.0(0.7)	0.96	0.99	0.95
	PSIS	2.0 (0.0)	1.00	1.00	1.00	5.5(48.8)	0.75	0.75	0.55
	MV-SIS	2.0 (0.0)	1.00	1.00	1.00	2.0(0.7)	0.96	0.99	0.95

Zhu (2012) claimed that DC-SIS can be applied for the multiple response. Pan, Wang, and Li (2013) proposed a pairwise sure independence screening (PSIS) to deal with the categorical response. PSIS uses $|\hat{\mu}_{r_1 j} - \hat{\mu}_{r_2 j}|$ as the marginal signal of predictor X_j for each pair of classes (r_1, r_2) each time, where $\hat{\mu}_{rj}$ denotes the sample average of X_{ij} for $i \in \{i : Y_i = r\}$. Essentially, we consider $\max_{r_1 \neq r_2} |\hat{\mu}_{r_1 j} - \hat{\mu}_{r_2 j}|$ as the marginal signal of predictor X_j, where $r_1, r_2 = 1, 2, \ldots, 10$, denoted by PSIS*. Table 2 summarizes the median of MMS with its associated robust standard deviation in the parentheses, $\mathcal{P}_j^s$ with $j = 1, 2, \ldots, 10$ and $\mathcal{P}_a$ for the given model size $d = [n/\log n]$ based on 500 simulations.

In addition, we will compare the postscreening estimation and prediction performance between PSIS and MV-SIS for the binary response from a discrete uniform distribution. Here, we generate $p = 2000$ predictors and a binary response with different sample sizes $n = 40$ and $n = 80$. We replicate each simulation experiment a total of 500 times. For the rth simulation, we follows Pan, Wang, and Li (2013) to choose the model size using the BIC criterion, which uses the equivalence between the LDA problem and a least squares one in Mai, Zou, and Yuan (2012). Then, we define the model size (MS), percentage of correct zeros (CZ), incorrect zeros (IZ), coverage probability (CP), and the root of the sum squared error (RSSE) as follows, respectively,

$$\mathrm{MS}_r = |\mathcal{D}_r^*|, \mathrm{CZ}_r = \frac{p - |\mathcal{D} \bigcup \mathcal{D}_r^*|}{p - s_n}, \mathrm{IZ}_r = \frac{|\mathcal{D}_r^{*c} \bigcap \mathcal{D}|}{|\mathcal{D}|},$$
$$\mathrm{CP}_r = I(\mathcal{D} \subseteq \mathcal{D}_r^*), \mathrm{RSSE}_r = \|\widehat{\gamma}_r - \gamma_0\|,$$

where $\mathcal{D}_r^*$ is the selected model in the rth replication, $s_n = 2$ is the cardinality of $\mathcal{D}$, and $\gamma_0 = \mu_1 - \mu_2 = (3, -3, 0, \ldots, 0)$ is the true difference between two true means, and $\widehat{\gamma}_r = \widehat{\mu}_{1r} - \widehat{\mu}_{2r}$ is the postscreening estimator of γ_0 in the rth replication based on the selected model. Furthermore, to assess the prediction performance, an independent testing dataset is generated with the same sample size in the each simulation. The classification accuracy (CA) of the postscreening estimator is computed in each simulation. Also the classification accuracy based on the true means, denoted by CA_0, and the ratio $\mathrm{CA}/\mathrm{CA}_0$ is evaluated for comparison. We report the median of MS with its robust

Table 2. Simulation results for linear discriminant analysis with 10-categorical response

Method	MMS	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$	$\mathcal{P}_4^s$	$\mathcal{P}_4^s$	$\mathcal{P}_6^s$	$\mathcal{P}_7^s$	$\mathcal{P}_8^s$	$\mathcal{P}_9^s$	$\mathcal{P}_{10}^s$	$\mathcal{P}_a$
(i) Balanced probabilities and Case (1): $\varepsilon_{ij} \sim N(0, 1)$												
DC-SIS	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
PSIS*	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MV-SIS	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(i) Balanced probabilities and Case (2): $\varepsilon_{ij} \sim t(2)$												
DC-SIS	15.0(21.8)	0.86	0.99	0.99	0.99	0.97	0.98	0.99	0.99	0.99	0.98	0.74
PSIS*	362.5(563.6)	0.73	0.75	0.76	0.73	0.75	0.75	0.75	0.73	0.76	0.79	0.05
MV-SIS	11.0(3.7)	1.00	1.00	1.00	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.95
(ii) Unbalanced probabilities and Case (1): $\varepsilon_{ij} \sim N(0, 1)$												
DC-SIS	13.0(14.9)	0.82	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.82
PSIS*	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
MV-SIS	10.0(0.0)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
(ii) Unbalanced probabilities and Case (2): $\varepsilon_{ij} \sim t(2)$												
DC-SIS	126.5(248.3)	0.35	0.90	0.93	0.93	0.96	1.00	0.99	1.00	1.00	1.00	0.22
PSIS*	343.5(444.9)	0.68	0.66	0.56	0.58	0.64	0.63	0.60	0.73	0.61	0.67	0.05
MV-SIS	13.0(9.8)	0.93	0.98	0.98	0.98	0.98	1.00	1.00	1.00	1.00	1.00	0.85

Table 3. Simulation results for estimation and prediction performance in linear discriminant analysis with binary response with 500 simulations

n	Method	MS(RSD)	CZ(%)	IZ(%)	CP(%)	RSSE	CA(%)	CA_0(%)	RCA
				Case (1): $\varepsilon_{ij} \sim N(0, 1)$					
40	PSIS	3.0 (2.9)	99.89	0.00	100.00	1.31	95.20	98.41	96.76
	MV-SIS	3.0 (2.2)	99.91	0.00	100.00	1.16	95.34	98.41	96.90
80	PSIS	2.0 (1.5)	99.94	0.00	100.00	0.70	97.31	98.31	98.98
	MV-SIS	2.0 (0.8)	99.95	0.00	100.00	0.62	97.47	98.31	99.15
				Case (2): $\varepsilon_{ij} \sim t(2)$					
40	PSIS	6.0 (2.9)	99.76	19.50	65.00	3.65	73.42	89.91	81.81
	MV-SIS	5.0 (3.1)	99.83	3.00	94.00	2.74	78.92	89.91	87.87
80	PSIS	7.0 (4.4)	99.71	7.00	86.40	2.56	79.17	89.95	88.04
	MV-SIS	3.0 (2.9)	99.87	0.00	100.00	1.56	84.80	89.95	94.30

standard deviation in the parentheses, and the averages of other performance measures over all 500 simulations in Table 3.

Both Tables 1 and 2 indicate that the proposed MV-SIS is superior to other competitors for variable screening in the linear discriminant analysis. When the error term is heavy-tailed and the number of the response categories increases, MV-SIS has much smaller minimum model sizes (MMS) and significantly higher probabilities to include all active predictors in the selected model than other independent screenings. Thus, the robustness of our MV-SIS is an important feature, which can make it more useful in practice. The same pattern can be observed from Table 3. MV-SIS has the very close estimation and prediction performance of PSIS when the error term is normal. However, when the error deviates from a normal distribution, PSIS deteriorates while MV-SIS still performs reasonably well.

3.2 Real Data Examples

Example 3.2. Lung cancer data were previously analyzed for classification between malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung in Gordon et al. (2002) and Fan and Fan (2008). There are 12,533 genes and 181 tissue samples from two classes: 31 in class MPM and 150 in class ADCA. The training dataset contains 32 of them (16 MPM and 16 ADCA), while the remaining 149 samples (15 MPM and 134 ADCA) are used for testing.

Before classification, we first standardize the data to zero mean and unit variance. Fan and Fan (2008) showed that their features annealed independence rules (FAIR) selected 31 important genes and made no training error and 7 testing errors, while the nearest shrunken centroids (NSC) method proposed by Tibshirani et al. (2002) chose 26 genes and resulted in no training error and 11 testing errors. Then, we consider DC-SIS, PSIS, and our MV-SIS approach (denoted by MV-SIS1) following by LDA for this ultrahigh dimensional classification problem. Note that FAIR used the diagonal linear discriminant analysis after the t-test screening. To make a fair comparison, we add a procedure combining t-test screening with LDA as well, denoted by FAIR*. Furthermore, the penalized LDA method (denote by PenLDA) proposed by Witten and Tibshirani (2011) and the sparse discriminant analysis (denoted by SDA) in Clemmensen et al. (2011) are also implemented in this example for comparison. In addition, we combine our MV-SIS with SDA and consider this two-stage method as another potential approach, denoted by MV-SIS2. Similar to Example 3.1, the BIC criterion is applied to determining the model size for all competing methods in this binary classification problem. We summarize the classification results in Table 4. The MV-SIS followed by LDA (i.e., MV-SIS1) makes 0 training error and five testing errors using only five top genes, and the MV-SIS with SDA (i.e., MV-SIS2) performs even better than MV-SIS1 and SDA to achieve the smallest testing errors using only seven genes. Thus, the two-stage approaches combining MV-SIS with LDA or SDA are superior to other competitors in terms of classification errors and the selected model size for this ultrahigh dimensional lung cancer data.

To further evaluate the prediction performance, we randomly partition all 181 tissue samples into two parts: the training set including 100 samples and the testing set of the rest 81 samples. The above procedures are applied to the training data, and their performances are evaluated by the classification errors in both training and testing sets. For a fair comparison, we choose the best model sizes for all methods using the same BIC criterion. We repeat the experiment 100 times, summarize the means with associated standard deviations (in the parentheses) of the training and testing classification errors and the numbers of selected genes in Table 5, and display their distributions in Figure 2. In the result, the MV-SIS with LDA method (i.e., MV-SIS1) performs reasonably well and has both small training and testing errors using averagely around 12 genes. Among all the methods, the SDA method classifies the training samples perfectly and achieves a small testing error rate. However, SDA tends to select a considerably large number of genes and thus may lose some model interpretability. It is worth noting that

Table 4. Classification errors for lung cancer data in Example 3.2

Method	Training error	Testing error	No. of selected genes
NSC	0/32	11/149	26
FAIR	0/32	7/149	31
FAIR*	0/32	7/149	14
PenLDA	0/32	9/149	8
SDA	0/32	6/149	17
PSIS	1/32	34/149	4
DC-SIS	0/32	6/149	7
MV-SIS1	0/32	5/149	5
MV-SIS2	0/32	3/149	7

Table 5. Performance evaluation for lung cancer data in Example 3.2

Method	Training error(%)	Testing error(%)	No. of selected genes
NSC	0.87(0.90)	1.86(1.91)	17.52(11.36)
FAIR	3.07(1.32)	3.51(1.93)	13.72(7.37)
PenLDA	0.88(0.92)	1.95(1.97)	18.95(18.14)
SDA	0.00(0.00)	1.42(1.21)	39.83(2.84)
PSIS	0.06(0.24)	2.14(1.57)	26.49(6.85)
DC-SIS	0.08(0.27)	2.63(2.30)	15.54(12.53)
MV-SIS1	0.15(0.44)	1.77(1.91)	11.99(9.53)
MV-SIS2	0.20(0.40)	1.41(1.10)	11.74(6.71)

the MV-SIS with SDA (i.e., MV-SIS2) can achieve the smallest testing error rate with a much smaller number of genes. This further demonstrates the merit of the two-stage approach combining MV-SIS with SDA.

Example 3.3. This human lung carcinomas data was analyzed by using mRNA expression profiling (Bhattacharjee et al. 2001). There are 12,600 mRNA expression levels in a total of 203 snap-frozen lung tumors and normal lungs. The 203 specimens are classified into five subclasses: 139 in lung adenocarcinomas (ADEN), 21 in squamous cell lung carcinomas (SQUA), 6 in small cell lung carcinomas (SCLC), 20 in pulmonary carcinoid tumors (COID), and the remaining 17 normal lung samples (NORMAL). Before classification, we first standardize the data to zero mean and unit variance. To evaluate the prediction performance of the proposed method, we randomly select approximately $100\tau\%$ of the observations from each subclass as the training samples and the rest $100(1-\tau)\%$ observations as the testing samples, where $\tau \in (0, 1)$.

Note that the aforementioned NSC and FAIR are proposed only for binary classification problems, thus they are not applicable in this multiple classes discriminant analysis. PSIS, DC-SIS, and MV-SIS with LDA are applied to the training set and their performances are evaluated by the testing samples. For the DC-SIS and MV-SIS (denoted by MV-SIS1) with LDA procedures, the leave-one-out cross-validation is applied to choosing the optimal model size for the raining data. Besides, we also consider the penalized LDA (denoted by PenLDA) and MV-SIS followed by SDA (denoted by MV-SIS2) for comparison, and use the 10-folded cross-validation rather than the leave-one-out cross-validation to choose the best model size in order to reduce the computation time. Although SDA can be directly applied to multiple-class discriminant analysis for a given model size, searching the best model size for SDA is remarkably computational expensive for multiple-class ultrahigh dimensional data. Thus, we use MV-SIS to reduce dimensionality and then follow by SDA (i.e., MV-SIS2) instead of SDA alone in the example.

Next, we choose $\tau = 0.9, 0.8$ and repeat each experiment 100 times. Following Example 3.2, the means of the training and testing classification errors and the corresponding numbers of selected genes with their associated standard deviations (in the parentheses) are reported in Table 6. We can clearly observe that, although all methods perform reasonably well in the tumors' classification, the MV-SIS procedure with LDA or SDA are significantly better than other methods in terms of both training and testing classification errors and the number of selected genes. Specifically, the MV-SIS+SDA (i.e., MV-SIS2) procedure achieves the best performances using a small number of top genes. Furthermore, we find that the top genes selected by MV-SIS are not normally distributed and contain potential outliers. This observation explains why other methods perform relatively worse and confirms the robustness feature of the proposed MV-SIS. This example further demonstrates that the two-stage approach combing the MV-SIS method with a discriminant analysis is more favorable for ultrahigh dimensional data in practice.

4. SOME EXTENSIONS

The MV-SIS approach is proposed to screen important predictors for the ultrahigh dimensional discriminant analysis where the response is categorical, but its applications can be easily extended to some other settings. In this section, we discuss two natural extensions of MV-SIS and use simulation studies to show their excellent performances.

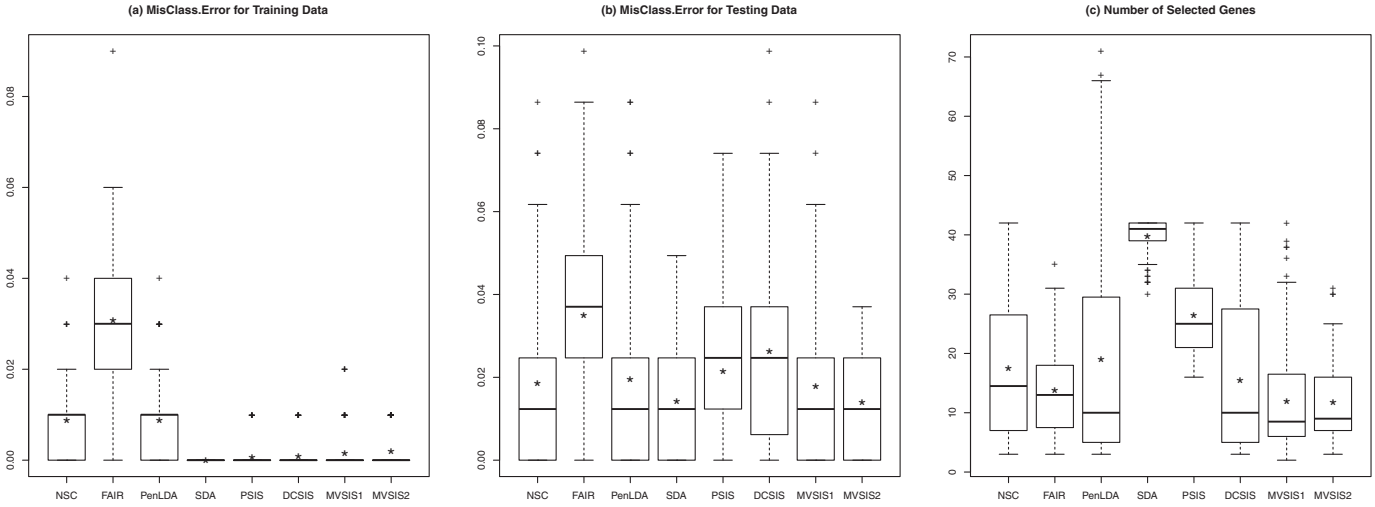

Figure 2. Lung cancer data in Example 3.2. (a) Boxplots of classification errors in the training sets over 100 random partitions of 181 samples; (b) boxplots of classification errors in the testing sets; (c) boxplots of numbers of selected genes.

Table 6. Classification errors for lung carcinomas data with five classes in Example 3.3

τ	Method	Training error(%)	Testing error(%)	No. of selected genes
	PenLDA	21.88(2.24)	21.71(3.87)	25.76(21.04)
	PSIS	3.54(0.79)	9.43(5.65)	107.54(15.71)
0.9	DC-SIS	6.85(1.35)	11.81(6.40)	32.08(3.85)
	MV-SIS1	3.65(1.15)	7.71(4.99)	20.56(8.02)
	MV-SIS2	3.65(1.15)	7.62(5.09)	31.76(10.24)
	PenLDA	22.12(2.10)	22.40(4.37)	25.04(21.81)
	PSIS	3.08(1.11)	7.90(3.89)	101.88(15.72)
0.8	DC-SIS	6.33(2.16)	13.15(5.32)	32.18(5.39)
	MV-SIS1	3.74(1.09)	8.35(4.12)	21.34(7.42)
	MV-SIS2	3.74(1.09)	6.70(4.24)	27.20(9.11)

4.1 Genome-Wide Association Studies

First, we can apply MV-SIS to ultrahigh dimensional problems with categorical predictors. In such situations, feature screening can be done by using $\mathrm{MV}(Y|X_k)$, where X_k is categorical for $k = 1, 2, \ldots, p$. Under Conditions (C1) and (C2), we can establish the sure screening property and ranking consistency property for $\omega_k = \mathrm{MV}(Y|X_k)$ with imposing Condition (C1) on each categorical SNP instead of the response. In genome-wide association studies (GWAS), modern genotyping techniques allow researchers to collect genetic data which usually contain an extremely large number of single-nucleotide polymorphisms (SNPs). In general, the SNPs as predictors are categorical with three classes, denoted by $\{AA, Aa, aa\}$. In Example 4.1, we consider applying the proposed MV-SIS for the ultrahigh dimensional GWAS problem to identify important SNPs, and compare its performance with other independence screening approaches.

Example 4.1. (Genome-Wide Association Studies) 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 $\rho_{ij} = 0.5^{|i-j|}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, 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

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

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 $Z \sim \mathcal{N}(0, 1)$, the error term ε follows $N(0, 1)$ or $t(1)$. There are five active SNPs, that is, Z_1, Z_2, Z_{10}, Z_{20}, and Z_{100}, for the response. The first four active SNPs are linearly correlated with the response Y, while the SNP Z_{100} and Y are nonlinearly correlated. It is interesting to note that the absolute value of dominant effect $|Z_{100}|$ is the corresponding additive effect in genetics. Here, we consider five different independence screening approaches: SIS, DC-SIS, SIRS, RRCS (Li et al. 2012), and MV-SIS, and set $n = 200$ and $p = 2000$ and repeat each experiment 500 times. We summarize the simulation results for $d = [n/\log(n)]$ in Table 7.

Table 7. Simulation results for Example 4.1—GWAS model

ε	Method	MMS	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_{10}^s$	$\mathcal{P}_{20}^s$	$\mathcal{P}_{100}^s$	$\mathcal{P}_a$
	SIS	1058.0(786.9)	0.96	0.97	1.00	0.99	0.02	0.02
	DCSIS	10.0(40.1)	0.96	0.95	1.00	0.99	0.79	0.72
$N(0, 1)$	SIRS	1074.0(834.8)	0.94	0.95	1.00	0.98	0.03	0.02
	RRCS	1031.0(801.6)	0.96	0.96	1.00	0.99	0.03	0.03
	MVSIS	8.0(34.3)	0.96	0.94	0.99	0.98	0.89	0.78
	SIS	1427.0(530.4)	0.26	0.28	0.42	0.42	0.02	0.00
	DC-SIS	124.0(284.8)	0.78	0.75	0.92	0.91	0.53	0.32
$t(1)$	SIRS	1050.0(672.5)	0.86	0.84	0.97	0.96	0.02	0.01
	RRCS	993.0(725.5)	0.87	0.84	0.98	0.96	0.02	0.01
	MV-SIS	46.0(139.1)	0.79	0.79	0.94	0.94	0.79	0.46

According to Table 7, when the error follows a normal distribution, all five independence screening are able to select the first four active SNPs effectively because they are linearly correlated with the response. However, only DC-SIS and MV-SIS can choose Z_{100} which nonlinearly contributed to Y. When the error is generated from $t(1)$ which is largely heavy-tailed, it is not surprising that all independence screening methods perform worse than before. However, the performance of MV-SIS is still the best one. Thus, we can conclude that MV-SIS can effectively select active categorical SNPs which are linearly or nonlinearly correlated with the response.

4.2 Nonparametric Additive Models

In this section, we further consider the application of MV-SIS for an ultrahigh dimensional nonparametric additive model to evaluate MV-SIS. Although both the response and predictors are generally continuous, we can discretize each predictor X_j into a categorical variable to make MV-SIS applicable. To be specific, we can define X_j^* using percentiles $\{\tau_1, \ldots, \tau_{K_n}\}$ of X_j by $X_{ij}^* = kI(\tau_k \le X_{ij} < \tau_{k+1})$, where $I(\cdot)$ is an indicator function, $i = 1, \ldots, n$, $j = 1, \ldots, p$, $k = 1, \ldots, K_n$ with $K_n = O(n^{1/5})$. Then, we can apply MV-SIS to the discretized predictors and use $\mathrm{MV}(Y|X_j^*)$ as the marginal screening utility to measure the importance of X_j. In practice, the sample size in each discretized class cannot be small to ensure an accurate estimation of conditional distribution function. On the other hand, the number of classes cannot be small to retain as much information of the continuous variable as possible. According to our empirical experiences, we suggest that the number of samples in each class should be greater than 20 to obtain a decent estimator of the MV index. One can also consider the number of classes as a tuning parameter and apply the cross-validation technique to choose an optimal number of classes. The following simulation example numerically examines the performance of the proposal.

Example 4.2. (Nonparametric Additive Model) Following Meier, Geer, and Bühlmann (2009), we define the following four functions

$$f_1(x) = -\sin(2x),\ f_2(x) = x^2 - 25/12,\ f_3(x) = x,$$
$$f_4(x) = e^{-x} - 2/5 \cdot \sinh(5/2).$$

Table 8. Simulation results for Example 4.2—nonparametric additive model

ε	Method	MMS	$\mathcal{P}_1^s$	$\mathcal{P}_2^s$	$\mathcal{P}_3^s$	$\mathcal{P}_4^s$	$\mathcal{P}_a$
	SIS	1084.5(690.3)	0.17	0.02	1.00	1.00	0.00
	NIS	4.0(0)	1.00	0.99	1.00	1.00	0.99
	DC-SIS	50.5(55.2)	0.47	0.79	1.00	1.00	0.37
$N(0, 1)$	SIRS	1178.0(668.6)	0.15	0.01	1.00	1.00	0.00
	QaSIS	5.0(4.5)	0.99	0.93	0.99	1.00	0.91
	RRCS	1112.5(673.9)	0.16	0.03	1.00	1.00	0.00
	MV-SIS	4.0(1.5)	0.99	0.95	1.00	1.00	0.95
	SIS	1508.0(538.1)	0.04	0.01	0.44	0.51	0.00
	NIS	1056.5(932.2)	0.25	0.15	0.22	0.37	0.08
	DC-SIS	205.0(280.1)	0.20	0.33	0.96	0.96	0.07
$t(1)$	SIRS	1222.5(645.5)	0.12	0.01	1.00	1.00	0.00
	QaSIS	16.0(37.7)	0.93	0.79	0.93	1.00	0.69
	RRCS	1212.0(688.1)	0.14	0.01	0.99	1.00	0.00
	MV-SIS	11.0(24.8)	0.93	0.81	0.99	1.00	0.75

Then we consider the following additive model

$$Y = 3f_1(X_1) + f_2(X_2) - 1.5f_3(X_3) + f_4(X_4) + \varepsilon,$$

where the predictors are generated independently from Uniform$[-2.5, 2.5]$. To examine the robustness of each independence screening approach, we consider two cases for the error term $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$: (1) $\varepsilon_i \sim N(0, 1)$; (2) $\varepsilon_i \sim t(1)$ for $i = 1, 2, \ldots, n$. In this example, besides the five approaches in Example 4.1, we further consider the nonparametric independence screening (NIS) proposed for sparse ultrahigh dimensional additive models by Fan, Feng, and Song (2011), and the quantile-adaptive sure independence screening (QaSIS) with quantile $\tau = 0.5$ proposed by He, Wang, and Hong (2013). We set $n = 200$ and $p = 2000$ and repeat each experiment 500 times for each error case. In our simulation, we discretize each predictor into a four-categorical variable using first, second, and third quartiles as knots for our MV-SIS. Simulation results are reported for the given model size $d = [n/\log(n)]$ in Table 8.

Table 8 indicates that MV-SIS performs very well after discretizing each predictor. When the error term is normal, NIS performs best followed by MV-SIS and QaSIS. Although DC-SIS may detect the nonlinearity, it occasionally misses X_1 and X_2. The probable reason is the distance correlation between Y and the first two predictors are relatively weak. When the error term follows Cauchy distribution, which makes the data heavy-tailed and generates some extreme points, NIS quickly deteriorates and yet QaSIS performs well to detect the true signals. On the other hand, MV-SIS still can effectively select the active predictors and performs even better than QaSIS, which presents its robustness merit again.

5. DISCUSSION

In this article, we have developed a new sure screening procedure for ultrahigh dimensional discriminant analysis, in which the response is allowed to have a diverging number of categories. We further established the sure screening property and the ranking consistency property of the proposed procedure without assuming any moment condition on predictors. The proposed procedure has several appealing properties. It is easily implemented, and it is robust to model specification (i.e., model-free) and robust to outliers or heavy tails of the predictors. The proposed procedure is also highly useful for analysis of data collected in GWAS, in which the phenotype may be multivariate continuous, while the predictors are categorical SNPs.

In the numerical studies, we applied linear discriminant analysis on the selected model by MV-SIS in the second-stage study. The linear discriminant analysis methods are widely used in practice and did perform reasonably well in our real data analysis. However, it is also interesting to propose a model-free and robust discriminant analysis after a model-free variable screening approach. This is out of scope of this work, but is an interesting topic for future research. Some work have been done on robust discriminant analysis. Related references include regularized discriminant analysis by Friedman (1989), robust LDA based on S-estimators by He and Fung (2000), penalized linear discriminant analysis by Witten and Tibshirani (2011), semiparametric sparse discriminant analysis by Mai and Zou (2014) and among others.

APPENDIX

Proof of Proposition 2.1. Note $F(x|Y) = \mathbb{P}(X \le x|Y)$ is a random variable of Y. □

$$\begin{aligned}
E_Y[F(x|Y)] &= \sum_{r=1}^{R} \mathbb{P}(X \le x|Y = y_r)\mathbb{P}(Y = y_r) \\
&= \sum_{r=1}^{R} \mathbb{P}(X \le x, Y = y_r) = \mathbb{P}(X \le x) = F(x), \\
\mathrm{var}_Y[F(x|Y)] &= \sum_{r=1}^{R} [\mathbb{P}(X \le x|Y = y_r) - F(x)]^2 \mathbb{P}(Y = y_r) \\
&= \sum_{r=1}^{R} p_r [F_r(x) - F(x)]^2,
\end{aligned}$$

where $p_r = \mathbb{P}(Y = y_r)$. Then

$$\begin{aligned}
\mathrm{MV}(X|Y) &= E_X[\mathrm{var}_Y(F(X|Y))] \\
&= \sum_{r=1}^{R} p_r \int [F_r(x) - F(x)]^2 dF(x).
\end{aligned}$$

The second property can be directly implied by the first one. Because the result that X and Y are statistical independent is equivalent to that $F_r(x) = F(x)$ for any $x \in \mathbb{R}_X$ and $r = 1, 2, \ldots, R$, which is also equivalent to $\sum_{r=1}^{R} p_r \int [F_r(x) - F(x)]^2 dF(x) = 0$ given $p_r > 0$ and $F(x + \delta) - F(x - \delta) > 0$ for any $\delta > 0$ and $x \in \mathbb{R}_X$. This completes the proof.

To prove Theorems 2.1 and 2.2, we need the following lemmas.

Lemma A.1. Hoeffding's Inequality Let $X_1, \ldots, X_n$ be independent random variables. Assume that $\mathbb{P}(X_i \in [a_i, b_i]) = 1$ 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

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

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

Lemma A.2. Bernstein's Inequality (van der Vaart and Wellner 1996, Lemma 2.2.9) Let $X_1, \ldots, X_n$ be independent random variables with

bounded support $[-M, M]$ and zero means, then the following inequality holds

$$\mathbb{P}(|X_1 + \cdots + X_n| > t) \le 2\exp\left\{-\frac{t^2}{2(v + Mt/3)}\right\}, \qquad (A.2)$$

for $v \ge \mathrm{var}(X_1 + \cdots + X_n)$.

We need the following notations for next lemma. Let $F_{k,r}(x) = \mathbb{P}(X_k \le x | Y = y_r)$ and $F_k(x) = \mathbb{P}(X_k \le x)$, for $1 \le k \le p$, $r = 1, \ldots, R$ and $x \in \mathbb{R}_X$. Denote

$$\begin{aligned}
f_0 &= f_0(X_k, Y) = \sum_{r=1}^{R} I\{Y = y_r\} \int [F_{k,r}(x) - F_k(x)]^2 dF_k(x);\\
\bar{f}_r &= \bar{f}_r(X_k, Y) = [F_{k,r}(X_k) - F_k(X_k)]^2;\\
f_r &= f_r(X_k, Y) = I\{Y = y_r\};\\
f_{0,x} &= f_{0,x}(X_k, Y) = I\{X_k \le x\};\\
f_{r,x} &= f_{r,x}(X_k, Y) = I\{X_k \le x, Y = y_r\}.
\end{aligned}$$

Let $\{(X_{ki}, Y_i) : 1 \le i \le n\}$ be a random sample from a population (X_k, Y). Define $\bar{f}_r^{(i)} = \bar{f}_r(X_{ik}, Y_i)$, $f_0^{(i)} = f_0(X_{ik}, Y_i)$, $f_r^{(i)} = I\{Y_i = y_r\}$, $f_{0,x}^{(i)} = I\{X_{ik} \le x\}$, $f_{r,x}^{(i)} = I\{X_{ik} \le x, Y_i = y_r\}$, for $i = 1, \ldots, n$.

Lemma A.3. For any $\epsilon \in (0, 1)$ and $1 \le r \le R$, the following inequalities are valid for univariate X_k

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} \bar{f}_r^{(i)} - E\bar{f}_r\right| \ge \epsilon\right\} \le 2\exp\left\{-2n\epsilon^2\right\}; \qquad (A.3)$$

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} f_0^{(i)} - Ef_0\right| \ge \epsilon\right\} \le 2\exp\left\{-2n\epsilon^2\right\}; \qquad (A.4)$$

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} f_r^{(i)} - Ef_r\right| \ge \epsilon\right\} \le 2\exp\left\{-\frac{n\epsilon^2}{2(p_r + \epsilon/3)}\right\}; \qquad (A.5)$$

$$\mathbb{P}\left\{\sup_{x\in\mathbb{R}_X}\left|\frac{1}{n}\sum_{i=1}^{n} f_{0,x}^{(i)} - Ef_{0,x}\right| \ge \epsilon\right\} \le 2(n+1)\exp\{-2n\epsilon^2\}; \qquad (A.6)$$

$$\begin{aligned}
&\mathbb{P}\left\{\sup_{x\in\mathbb{R}_X}\left|\frac{1}{n}\sum_{i=1}^{n} f_{r,x}^{(i)} - Ef_{r,x}\right| \ge \epsilon\right\}\\
&\quad\le 2(n+1)\exp\left\{-\frac{n\epsilon^2}{2(p_r + \epsilon/3)}\right\}, \qquad (A.7)
\end{aligned}$$

where Eh stands for $Eh(X_k, Y)$ for a function $h(X_k, Y)$ with finite expected value.

Proof. Since $|\bar{f}_r(X_k, Y)| = [F_{k,r}(X_k) - F_k(X_k)]^2 \le 1$ and $|f_0(X_k, Y)| = |\sum_{r=1}^{R} I\{Y = y_r\} \int [F_{k,r}(x) - F_k(x)]^2 dF_k(x)| \le 1$, we apply Hoeffding's inequality to obtain the inequalities (A.3) and (A.4). □

Since $f_r^{(i)} = I\{Y_i = y_r\}$ for $i = 1, \ldots, n$, then $f_r^{(i)} \sim \mathrm{Bernoulli}(p_r)$ with $Ef_r^{(i)} = p_r$ and $f_r^{(1)} + \cdots + f_r^{(n)} \sim \mathrm{Binomial}(n, p_r)$, which implies $\mathrm{var}(f_r^{(1)} + \cdots + f_r^{(n)}) = np_r(1 - p_r) \le np_r$ and $|f_r^{(i)} - p_r| \le 1$. Thus, by Bernstein's inequality, we have

$$\begin{aligned}
\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} f_r^{(i)} - Ef_r\right| \ge \epsilon\right\} &= \mathbb{P}\left\{\left|\sum_{i=1}^{n}\left(f_r^{(i)} - p_r\right)\right| \ge n\epsilon\right\}\\
&\le 2\exp\left\{-\frac{n^2\epsilon^2}{2(np_r + n\epsilon/3)}\right\}\\
&\le 2\exp\{-n\epsilon^2/(2(p_r + \epsilon/3))\}.
\end{aligned}$$

The inequality (A.5) is proved.

Note that $|f_{0,x}^{(i)} - Ef_{0,x}| = |I\{X_{ik} \le x\} - F_k(x)| \le 1$, then we apply Hoeffding's inequality and empirical process theory (Pollard 1984) to obtain (A.6). Note that $|f_{r,x}^{(i)} - Ef_{r,x}| = |I\{X_{ik} \le x, Y_i = y_r\} - F_{k,r}(x)p_r| \le 1$, then we apply Bernstein's inequality and empirical process theory (Pollard 1984) to obtain (A.7). This completes the proof of Lemma A.3.

Lemma A.4. Under Condition (C1), for any $\epsilon \in (0, 1/2)$ and $1 \le k \le p$, we have

$$\mathbb{P}\{|\hat{\omega}_k - \omega_k| \ge \epsilon\} \le O(n)R_n \exp\left\{-\frac{c_4 n}{R_n}\epsilon^2\right\} \qquad (A.8)$$

for some constant $c_4 > 0$.

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

$$\begin{aligned}
\hat{\omega}_k - \omega_k &= \frac{1}{n}\sum_{j=1}^{n}\sum_{r=1}^{R} \hat{p}_r[\hat{F}_{kr}(X_j) - \hat{F}_k(X_j)]^2\\
&\quad - \sum_{r=1}^{R} p_r \int [F_{kr}(x) - F_k(x)]^2 dF_k(x)\\
&= \sum_{r=1}^{R} \hat{p}_r\left(\int [\hat{F}_{kr}(x) - \hat{F}_k(x)]^2 d\hat{F}_k(x)\right.\\
&\quad \left. - \int [F_{kr}(x) - F_k(x)]^2 dF_k(x)\right)\\
&\quad + \sum_{r=1}^{R}(\hat{p}_r - p_r)\int [F_{kr}(x) - F_k(x)]^2 dF_k(x)\\
&= \sum_{r=1}^{R} \hat{p}_r \int \left([\hat{F}_{kr}(x) - \hat{F}_k(x)]^2 - [F_{kr}(x) - F_k(x)]^2\right) d\hat{F}_k(x)\\
&\quad + \sum_{r=1}^{R} \hat{p}_r \int [F_{kr}(x) - F_k(x)]^2 d[\hat{F}_k(x) - F_k(x)]\\
&\quad + \sum_{r=1}^{R}(\hat{p}_r - p_r)\int [F_{kr}(x) - F_k(x)]^2 dF_k(x)\\
&=: I_{k1} + I_{k2} + I_{k3}.
\end{aligned}$$

□

We first deal with the term I_{k1}.

$$\begin{aligned}
|I_{k1}| &\le 2\max_r \int |[\hat{F}_{kr}(x) - F_{kr}(x)] - [\hat{F}_k(x) - F_k(x)]| d\hat{F}_k(x)\\
&\le 2\max_r \sup_{x\in\mathbb{R}_X} (|\hat{F}_{kr}(x) - F_{kr}(x)| + |\hat{F}_k(x) - F_k(x)|)\\
&=: 2(J_{k1} + J_{k2}),
\end{aligned}$$

where the first inequality holds by $\sum_{r=1}^{R} \hat{p}_r = 1$ and $|[\hat{F}_{kr}(x) - F_{kr}(x)] + [\hat{F}_k(x) - F_k(x)]| \le |\hat{F}_{kr}(x) - F_{kr}(x)| + |\hat{F}_k(x) - F_k(x)| \le 1+1=2$, and the second inequality is implied by $\int d\hat{F}_k(x) = 1$. Then, we first deal with the term J_{k1},

$$\begin{aligned}
J_{k1} &= \max_r \sup_{x\in\mathbb{R}_X} |\hat{F}_{kr}(x) - F_{kr}(x)|\\
&= \max_r \sup_{x\in\mathbb{R}_X} \left|\frac{1}{n}\sum_{i=1}^{n} f_{r,x}^{(i)}/\hat{p}_r - Ef_{r,x}/p_r\right|\\
&\le \max_r \sup_{x\in\mathbb{R}_X} \left(\frac{\left|\frac{1}{n}\sum_{i=1}^{n} f_{r,x}^{(i)} - Ef_{r,x}\right|}{\hat{p}_r} + \frac{Ef_{r,x}\,|\hat{p}_r - p_r|}{\hat{p}_r p_r}\right)\\
&= \max_r \sup_{x\in\mathbb{R}_X} \frac{\left|\frac{1}{n}\sum_{i=1}^{n} f_{r,x}^{(i)} - Ef_{r,x}\right|}{\hat{p}_r} + \max_r \frac{\left|\frac{1}{n}\sum_{i=1}^{n} f_r^{(i)} - Ef_r\right|}{\hat{p}_r},
\end{aligned}$$

where the equality holds due to $\sup_{x\in\mathbb{R}_X} Ef_{r,x} = \sup_{x\in\mathbb{R}_X} P(X_k < x, Y = y_r) = p_r$. Thus, under Condition (C1), for any $0 < \epsilon < 1/2$,

$$
\begin{aligned}
&\mathbb{P}\{J_{k1} \geq \epsilon\} \\
&\leq \mathbb{P}\left\{\max_r \sup_{x\in\mathbb{R}_X} \frac{\left|\frac{1}{n}\sum_{i=1}^n f_{r,x}^{(i)} - Ef_{r,x}\right|}{\hat{p}_r}\right. \\
&\quad \left. + \max_r \frac{\left|\frac{1}{n}\sum_{i=1}^n f_r^{(i)} - Ef_r\right|}{\hat{p}_r} \geq \epsilon, \min_r \hat{p}_r \geq \frac{c_1}{2R_n}\right\} \\
&\quad + \mathbb{P}\{\min_r \hat{p}_r < c_1/2R_n\} \\
&\leq \mathbb{P}\left\{\max_r \sup_{x\in\mathbb{R}_X} \left|\frac{1}{n}\sum_{i=1}^n f_{r,x}^{(i)} - Ef_{r,x}\right|\right. \\
&\quad \left. + \max_r \left|\frac{1}{n}\sum_{i=1}^n f_r^{(i)} - Ef_r\right| \geq \frac{c_1\epsilon}{2R_n}\right\} \\
&\quad + \mathbb{P}\left\{\max_r \left|\frac{1}{n}\sum_{i=1}^n f_r^{(i)} - Ef_r\right| \geq \frac{c_1}{2R_n}\right\} \\
&\leq \mathbb{P}\left\{\max_r \sup_{x\in\mathbb{R}_X} \left|\frac{1}{n}\sum_{i=1}^n f_{r,x}^{(i)} - Ef_{r,x}\right| \geq \frac{c_1\epsilon}{4R_n}\right\} \\
&\quad + 2\mathbb{P}\left\{\max_r \left|\frac{1}{n}\sum_{i=1}^n f_r^{(i)} - Ef_r\right| \geq \frac{c_1\epsilon}{4R_n}\right\} \\
&\leq 2(n+1)R_n \exp\left\{-\frac{n(c_1\epsilon/4R_n)^2}{2(p_r + c_1\epsilon/12R_n)}\right\} \\
&\quad + 2R_n \exp\left\{-\frac{n(c_1\epsilon/4R_n)^2}{2(p_r + c_1\epsilon/12R_n)}\right\} \\
&\leq 2(n+3)R_n \exp\left\{-\frac{c_1^2}{32}\frac{n\epsilon^2}{R_n} \Big/ \left(c_2 + \frac{c_1\epsilon}{12}\right)\right\} \\
&\leq 2(n+3)R_n \exp\left\{-c_5\frac{n\epsilon^2}{R_n}\right\}, \qquad \text{(A.9)}
\end{aligned}
$$

for some constant $c_5 > 0$, where the second inequality holds because $\min_r \hat{p}_r < c_1/2R_n$ implies $\max_r \left|\frac{1}{n}\sum_{i=1}^n f_r^{(i)} - Ef_r\right| = \max_r |\hat{p}_r - p_r| \geq p_r - \hat{p}_r \geq c_1/R_n - c_1/2R_n = c_1/2R_n$ using $c_1/R_n \leq \min_{1\leq r\leq R_n} p_r$ in Condition (C1), the fourth inequality is due to Lemma A.3, and the fifth inequality follows that $\max_{1\leq r\leq R_n} p_r \leq c_2/R_n$ in Condition (C1). Then, we apply inequalities (A.6), (A.3), and (A.4) in Lemma A.3 to obtain the following three results, respectively,

$$
\mathbb{P}\{J_{k2} \geq \epsilon\} = \mathbb{P}\left\{\sup_{x\in\mathbb{R}_X} |\hat{F}_k(x) - F_k(x)| \geq \epsilon\right\} \leq 2(n+1)\exp\{-2n\epsilon^2\}, \qquad \text{(A.10)}
$$

$$
\begin{aligned}
\mathbb{P}\{|I_{k2}| \geq \epsilon\} &= \mathbb{P}\left\{\left|\sum_r^R \hat{p}_r\left(\frac{1}{n}\sum_{i=1}^n \bar{f}_r^{(i)} - E\bar{f}_r\right)\right| \geq \epsilon\right\} \\
&\leq \mathbb{P}\left\{\max_r \left|\frac{1}{n}\sum_{i=1}^n \bar{f}_r^{(i)} - E\bar{f}_r\right| \geq \epsilon\right\} \\
&\leq 2R_n \exp\{-2n\epsilon^2\}, \qquad \text{(A.11)}
\end{aligned}
$$

$$
\mathbb{P}\{|I_{k3}| \geq \epsilon\} = \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n f_0^{(i)} - Ef_0 \geq \epsilon\right\} \leq 2\exp\left\{-2n\epsilon^2\right\}. \qquad \text{(A.12)}
$$

Inequalities (A.9)–(A.12) together imply the result of Lemma A.4.

Proof of Theorem 2.1. For the first term of Theorem 2.1, by Lemma A.4 and $R_n = O(n^\kappa)$, we have

$$
\begin{aligned}
&\mathbb{P}\left\{\max_{1\leq k\leq p} |\hat{\omega}_k - \omega_k| \geq cn^{-\tau}\right\} \\
&\leq O(n)pR_n \exp\left\{-\frac{c_4c^2n^{1-2\tau}}{R_n}\right\} \\
&\leq O\left(pnR_n \exp\left\{-bn^{1-(2\tau+\kappa)}\right\}\right) \\
&\leq O\left(p\exp\left\{-bn^{1-(2\tau+\kappa)} + (1+\kappa)\log n\right\}\right),
\end{aligned}
$$

where $b > 0$ is a constant depending c, c_1, and c_2.

Next, we deal with the second part of Theorem 2.1. If $\mathcal{D} \nsubseteq \hat{\mathcal{D}}$, then there must exist some $k \in \mathcal{D}$ such that $\hat{\omega}_k < cn^{-\tau}$. It follows from Condition (C2) that $|\hat{\omega}_k - \omega_k| > cn^{-\tau}$ for some $k \in \mathcal{D}$, indicating that the events satisfy $\{\mathcal{D} \nsubseteq \hat{\mathcal{D}}\} \subseteq \{\hat{\omega}_k - \omega_k| > cn^{-\tau}, \text{ for some } k \in \mathcal{D}\}$, and hence $D_n = \{\max_{k\in\mathcal{D}} \hat{\omega}_k - \omega_k| \leq cn^{-\tau}\} \subseteq \{\mathcal{D} \subseteq \hat{\mathcal{D}}\}$. Consequently,

$$
\begin{aligned}
\mathbb{P}\{\mathcal{D} \subseteq \hat{\mathcal{D}}\} &\geq \mathbb{P}\{D_n\} = 1 - \mathbb{P}\left\{D_n^c\right\} = 1 - \mathbb{P}\left\{\min_{k\in\mathcal{D}} |\hat{\omega}_k - \omega_k| \geq cn^{-\tau}\right\} \\
&= 1 - s_n\mathbb{P}\{|\hat{\omega}_k - \omega_k| \geq cn^{-\tau}\} \\
&\leq 1 - O\left(s_n \exp\left\{-bn^{1-(2\tau+\kappa)} + (1+\kappa)\log n\right\}\right),
\end{aligned}
$$

where s_n is the cardinality of $\mathcal{D}$. This completes the proof of the second part.

Proof of Theorem 2.2.

$$
\begin{aligned}
&\mathbb{P}\left\{\left(\min_{k\in\mathcal{D}} \hat{\omega}_k - \max_{k\in\mathcal{I}} \hat{\omega}_k\right) < c_3/2\right\} \\
&\leq \mathbb{P}\left\{\left(\min_{k\in\mathcal{D}} \hat{\omega}_k - \max_{k\in\mathcal{I}} \hat{\omega}_k\right) - \left(\min_{k\in\mathcal{D}} \omega_k - \max_{k\in\mathcal{I}} \omega_k\right) < -c_3/2\right\} \\
&\leq \mathbb{P}\left\{\left|\left(\min_{k\in\mathcal{D}} \hat{\omega}_k - \max_{k\in\mathcal{I}} \hat{\omega}_k\right) - \left(\min_{k\in\mathcal{D}} \omega_k - \max_{k\in\mathcal{I}} \omega_k\right)\right| > c_3/2\right\} \\
&\leq \mathbb{P}\left\{2\max_{1\leq k\leq p} |\hat{\omega}_k - \omega_k| > c_3/2\right\} \\
&\leq O(n)pR_n \exp\left\{-c_6n/R_n\right\}
\end{aligned}
$$

for some constant $c_6 > 0$, where the first inequality follows Condition (C3) and the last inequality is implied by Lemma A.4. Because $R_n\log(p)/n = o(1)$ and $R_n\log(n)/n = o(1)$ imply that $p \leq \exp\{\frac{c_6}{2}n/R_n\}$, and $\frac{c_6}{2}n/R_n \geq 4\log(n)$, $\log(nR_n) \leq 2\log(n)$ for large n. Then, we have for some n_0, $\sum_{n=n_0}^{+\infty} npR_n \exp\{-c_6n/R_n\} \leq \exp\{\log(nR_n) + \frac{c_6}{2}n/R_n - c_6n/R_n\} \leq \exp\{\log(nR_n) - 4\log(n)\} \leq \sum_{n=n_0}^{+\infty} n^{-2} < +\infty$. Therefore, by Borel Contelli Lemma, we obtain that $\liminf_{n\to\infty}\{\min_{k\in\mathcal{D}}\hat{\omega}_k - \max_{k\in\mathcal{I}}\hat{\omega}_k\} \geq c_3/2 > 0$ a.s.

[Received July 2013. Revised April 2014.]

REFERENCES

Bhattacharjee, A., Richards, W., Staunton, J., Li, C., Monti, S., Vasa1, P., Ladd, C., Beheshti, J., Bueno R., Gillette, M., Loda1, M., Weber, G., Mark, E., Lander, E., Wong, W., Johnson, B., Golub, T., Sugarbaker, D., and Meyerson, M. (2001), "Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses," *PNAS*, 98, 13790–13795. [636]

Clemmensen, L., Hastie, T., Witten, D., and Ersboll, B. (2011), "Sparse Discriminant Analysis," *Technometrics*, 53, 406–415. [635]

Fan, J., and Fan, Y. (2008), "High-Dimensional Classification Using Features Annealed Independence Rules," *The Annals of Statistics*, 36, 2605–2637. [635]

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, 544–557. [630,638]

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, 1270–1284. [630]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society,* Series B, 70, 849–911. [630,632,633]

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 1829–1853. [630]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [631]

Friedman, J. (1989), "Regularized Discriminant Analysis," *Journal of the American Statistical Association*, 84, 165–175. [638]

Gordon, G., Jensen, R., Hsiao, L., Gullans, S., Blumenstock, J., Ramaswamy, S., Richards, W., Sugarbaker, D., and Bueno, R. (2002), "Translation of Microarray Data Into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios in Lung Cancer and Mesothelioma," *Cancer Research*, 62, 4963–4967. [635]

He, X., and Fung, W. K. (2000), "High Breakdown Estimation for Multiple Populations With Applications to Discriminant Analysis," *Journal of Multivariate Analysis*, 72, 151–162. [638]

He, X., Wang, L., and Hong, H. (2013), "Quantile-Adaptive Model-Free Variable Screening for High-Dimensional Heterogeneous Data," *The Annals of Statistics*, 41, 342–369. [630,632,638]

Huang, J., Horowitz, J., and Ma, S. (2008), "Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models," *The Annals of Statistics*, 36, 587–613. [631,633]

Ji, P., and Jin, J. (2012), "UPS Delivers Optimal Phase Diagram in High Dimensional Variable Selection," *The Annals of Statistics*, 40, 73–103. [632]

Li, G., Peng. H., Zhang J., and Zhu, L. (2012), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, 40, 1846–1877. [630,637]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of American Statistical Association*, 107, 1129–1139. [630,632,633]

Liu, J., Li, R., and Wu, R. (2014), "Feature Selection for Varying Coefficient Models With Ultrahigh Dimensional Covariates," *Journal of American Statistical Association*, 109, 266–274. [630]

Mai, Q., and Zou, H. (2013), "The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification," *Biometrika*, 100, 229–234. [630,633]

——— (2014), "Semiparametric Sparse Discriminant Analysis in Ultra-High Dimensions," manuscript. *2013arXiv1304.4983M*. [638]

Mai, Q., Zou, H., and Yuan, M. (2012), "A Direct Approach to Sparse Discriminant Analysis in Ultra-High Dimensions," *Biometrika*, 99, 29–42. [634]

Meier, L., Geer, V., and Bühlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [637]

Pan, R., Wang, H., and Li, R. (2013), "On the Ultrahigh Dimensional Linear Discriminant Analysis Problem With A Diverging Number of Classes," unpublished manuscript. [630,633,634]

Pollard, D. (1984), *Convergence of Stochastic Processes*, New York: Springer-Verlag Inc. [639]

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, 6567–6572. [635]

van der Vaart, A., and Wellner, J. (1996), *Weak Convergence and Empirical Processes*, New York: Springer. [638]

Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [630,632]

Witten, D., and Tibshirani, R. (2011), "Penalized Classification Using Fisher's Linear Discriminant," *Journal of the Royal Statistical Society,* Series B, 73, 753–772. [635,638]

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, 1464–1475. [630,633]