AIMS Mathematics, 8(6): 14270–14287.
DOI: 10.3934/math.2023730
Received: 14 November 2022
Revised: 27 February 2023
Accepted: 24 March 2023
Published: 17 April 2023

http://www.aimspress.com/journal/Math

Research article

Feature screening for ultrahigh-dimensional binary classification via linear projection

Peng Lai[1,2], **Mingyue Wang**[1], **Fengli Song**[1,2,*] **and Yanqiu Zhou**[3]

[1] School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China
[2] Center for Applied Mathematics of Jiangsu Province, Nanjing University of Information Science & Technology, Nanjing 210044, China
[3] School of Science, Guangxi University of Science and Technology, Liuzhou 545006, China

* **Correspondence:** Email: songfl@nuist.edu.cn.

Abstract: Linear discriminant analysis (LDA) is one of the most widely used methods in discriminant classification and pattern recognition. However, with the rapid development of information science and technology, the dimensionality of collected data is high or ultrahigh, which causes the failure of LDA. To address this issue, a feature screening procedure based on the Fisher's linear projection and the marginal score test is proposed to deal with the ultrahigh-dimensional binary classification problem. The sure screening property is established to ensure that the important features could be retained and the irrelevant predictors could be eliminated. The finite sample properties of the proposed procedure are assessed by Monte Carlo simulation studies and a real-life data example.

Keywords: ultrahigh-dimensional data; linear projection; marginal score test; feature screening; sure screening property
Mathematics Subject Classification: 62H30, 62F07

1. Introduction

More and more attention has been paid to the analysis of ultrahigh-dimensional data with the rapid development of information science and technology. The requirement of dealing with high-dimensional data efficiently should be well satisfied. Ultrahigh-dimensional data often appear in the area of biomedical imaging, gene expression and proteomics studies and so on. The dimensionality p of the collected data is allowed to diverge at a nonpolynomial rate with the sample size n, which is $\log p = O(n^{\xi})$ for some $\xi > 0$. Hence dimension reduction seems imperative for the efficient manipulation and analysis of ultrahigh-dimensional data.

Feature reduction techniques such as principal component analysis and linear discriminant analysis (LDA) have been proposed and applied successfully in practice to reduce the dimensionality of the original features without losing too much information. LDA is one of the most popular approaches in discriminant classification and pattern recognition, and it aims to find a proper linear transformation so that each sample vector with a high dimension is projected into a low-dimension vector while preserving the original cluster structure as much as possible. However, when the dimension is ultrahigh, the classical classification methods are no longer applicable, such as LDA [1], because of too many redundant variables. For example, the ovarian cancer data which were studied by Sorace and Zhan [2] consisted of serum samples of 162 ovarian cancer patients and 91 control subjects. For each sample, 15154 distinct mass-to-charge ratios (M/Z) were available for analysis. It is interesting to identify proteomic patterns (corresponding to the M/Z value) that can distinguish ovarian cancer subjects from control subjects. The small number of samples and many redundant variables make discriminant classification unable to work effectively.

To deal with the ultrahigh-dimensional discriminant classification difficulty, many marginal feature screening procedures are proposed by statisticians to reduce the dimension rapidly, and then some classical discriminant analysis methods could be processed. Mai and Zou [3] proposed a feature screening procedure named Kolmogorov filter (KF) for binary classification based on the Kolmogorov-Smirnov statistic, which enjoyed the sure screening property under much-weakened model assumptions. Mai and Zou [4] proposed the fused Kolmogorov filter, which generalized the KF procedure to the multi-classification case. Lai et al. [5] proposed the feature screening procedure based on the expected conditional Kolmogorov filter for the ultrahigh-dimensional binary classification problem with a dependent variable. Cui et al. [6] proposed a model-free feature screening index named MV for ultrahigh-dimensional discriminant analysis based on the difference between conditional and unconditional distribution functions. Pan et al. [7] developed a pairwise sure independence screening procedure (PSIS) for LDA, but this procedure depended on the parametric modeling assumptions and may perform poorly for heavy-tailed data. Cheng et al. [8] proposed a robust ranking screening procedure based on the conditional expectation of the rank of predictor samples for the ultrahigh-dimensional discriminant analysis, which was robust against the heavy-tailed distributions, potential outliers and the sample shortage for some categories. He et al. [9] generalized the MV procedure by modifying MV with a weight function. The proposed Anderson-Darling sure independence screening procedure (AD-SIS) could be more robust against the heavy-tailed distributions. Song et al. [10] proposed a robust composite weighted quantile screening procedure based on the difference between the conditional and the unconditional quantiles of the feature. Different from the existing methods, which used the differences of means or the differences of conditional cumulative distribution functions between classes as the screening indexes. Sheng and Wang [11] proposed a new feature screening method to rank the importance of predictors based on the classification accuracy of marginal classifiers.

Although many feature screening procedures for the ultrahigh-dimensional discriminant analysis problems have been proposed, some of them are even model-free; the study for the LDA problem, one of the most popular approaches in discriminant classification and pattern recognition, is still very attractive. In order to solve the linear discrimination problem with ultrahigh-dimensional features, the dimension reduction method based on the linear projection may bring a better performance than the model-free methods. In this paper, we propose a feature screening procedure based on the Fisher's linear discriminant framework. By minimizing the linear projection of the sum of squares in the

original cluster structures and maximizing the linear projection of the sum of squares between groups, the marginal score test is constructed and combined with the linear projection optimal problem. First, the proposed method can screen out the irrelevant predictors in the linear discriminant function through the use of estimating equations. Second, the proposed procedure processes the sure screening property. Third, the simple structure of the screening index makes the calculation fast.

The rest of this paper is organized as follows. In Section 2, we construct the feature screening estimating equations, propose the feature screening procedure and further study its theoretical properties. In Section 3, we present Monte Carlo simulation studies to examine the finite sample performance of the proposed procedure. We also use the proposed procedure in a real data example. All technical details are presented in the Appendix.

2. Linear projection feature screening

Consider the class data samples $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, where Y_i is the binary class index variable that equals one of $\{1, 2\}$, and $\mathbf{X}_i \in \mathbb{R}^p$. Suppose that $G = \{G_1, G_2\} \subset \mathbb{R}^{p\times n}$, where each $G_j \subset \mathbb{R}^{p\times n_j}$ for $j = 1, 2$ represents an independent class data set $\{\mathbf{X}_{ij}\} = \{\mathbf{X}_i \text{ when } Y_i = j\}$, n_j denotes the number of the samples of the class G_j; and, $n_1 + n_2 = n$. When the dimensionality p is large, the discriminant analysis based on the data would be rather complex and inefficient. LDA aims to find a linear projection $\boldsymbol{\beta} \in \mathbb{R}^p$, so it maps each sample vector $\mathbf{X}_i$ to a new low-dimension sample $\boldsymbol{\beta}^\top \mathbf{X}_i$, $i = 1, \ldots, n$. It seeks to project the observations into a lower space such that the intergroup variance of the projected samples is large and the intragroup variance is small. A classification rule is obtained by assigning the sample to its nearest centroid in the transformed space. To find the projection direction and delete the irrelevant predictors simultaneously, we construct the linear projection feature screening procedure in the following.

2.1. Screening method

Based on the linear projection, a random sample $\mathbf{X}_i$ is projected to $\boldsymbol{\beta}^\top \mathbf{X}_i$. Define

$$
\begin{aligned}
SSE &= \sum_{j=1}^{2}\sum_{i=1}^{n_j}(\boldsymbol{\beta}^\top \mathbf{X}_{ij} - \boldsymbol{\beta}^\top \bar{\mathbf{X}}_j)^2 = \boldsymbol{\beta}^\top \sum_{j=1}^{2}\sum_{i=1}^{n_j}(\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_j)^\top \boldsymbol{\beta} := \boldsymbol{\beta}^\top \mathbf{E}\boldsymbol{\beta}, \\
SS(TR) &= \sum_{j=1}^{2} n_j(\boldsymbol{\beta}^\top \bar{\mathbf{X}}_j - \boldsymbol{\beta}^\top \bar{\mathbf{X}})^2 = \boldsymbol{\beta}^\top \sum_{j=1}^{2} n_j(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})(\bar{\mathbf{X}}_j - \bar{\mathbf{X}})^\top \boldsymbol{\beta} := \boldsymbol{\beta}^\top \mathbf{B}\boldsymbol{\beta},
\end{aligned} \tag{2.1}
$$

where $\bar{\mathbf{X}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} \mathbf{X}_{ij}$ and $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$. Thus, the linear projection procedure aims to obtain $\boldsymbol{\beta}$ by

$$
\max_{\boldsymbol{\beta}} \frac{\boldsymbol{\beta}^\top \mathbf{B}\boldsymbol{\beta}}{\boldsymbol{\beta}^\top \mathbf{E}\boldsymbol{\beta}}, \quad s.t. \quad \boldsymbol{\beta} \neq 0 \quad and \quad \|\boldsymbol{\beta}\| = 1. \tag{2.2}
$$

When the dimensionality p is ultrahigh, the traditional solutions for (2.2) fail. For example, the related eigenvectors of the eigenvalues solved from $|\mathbf{B} - \lambda\mathbf{E}| = 0$ are hard to obtain. For ultrahigh-dimensional problems, sparsity is often present, meaning that only a small number of predictors contribute significantly to the LDA process. We denote the active set and the inactive set as $\mathcal{A} = \{k : \beta_k \neq 0, 1 \leq k \leq p\}$ and $\mathcal{A}^c = \{k : \beta_k = 0, 1 \leq k \leq p\}$, respectively. Note that (2.2) is a constrained optimization problem. To avoid the constraint, without loss of generality, we assume $\boldsymbol{\beta} = (\beta_1, \beta_2 \ldots, \beta_p)^\top = (\beta_1, \boldsymbol{\beta}_{(-1)}^\top)^\top$, where $\boldsymbol{\beta}_{(-1)} = (\beta_2, \ldots, \beta_p)^\top$, $\beta_1 > 0$ and $\beta_1 = \sqrt{1 - \|\boldsymbol{\beta}_{(-1)}\|^2}$.

$\beta_1 > 0$ means X_1 is important for the linear discriminant classification. The active predictor X_1 can be identified by comparing each marginal correlation of X_k and Y for $k = 1, \ldots, p$. We propose a feature screening procedure to identify potential active predictors as follows. Assume $E(X_k) = 0$, $Var(X_k) = 1$, and redefine $X_k = X_k - E(X_k|X_1)$ for $k = 2, \ldots, p$. The transformation of X_k can clear out the linear correlation between X_k and the active predictor X_1.

By introducing $\boldsymbol{\beta}_{(-1)}$, we estimate $\boldsymbol{\beta}(\boldsymbol{\beta}_{(-1)})$ by maximizing

$$\max_{\boldsymbol{\beta}_{(-1)}} \hat{L}(\boldsymbol{\beta}_{(-1)}) = \max_{\boldsymbol{\beta}_{(-1)}} \frac{\boldsymbol{\beta}^\top \mathbf{B}\boldsymbol{\beta}}{\boldsymbol{\beta}^\top \mathbf{E}\boldsymbol{\beta}}, \tag{2.3}$$

or solving

$$\frac{\partial \hat{L}(\boldsymbol{\beta}_{(-1)})}{\partial \boldsymbol{\beta}_{(-1)}} = \frac{1}{(\boldsymbol{\beta}^\top \mathbf{E}\boldsymbol{\beta})^2} \left[2J_{\beta_{(-1)}}^\top \mathbf{B}\boldsymbol{\beta} \cdot \boldsymbol{\beta}^\top \mathbf{E}\boldsymbol{\beta} - 2\boldsymbol{\beta}^\top \mathbf{E}\boldsymbol{\beta} \cdot J_{\beta_{(-1)}}^\top \mathbf{E}\boldsymbol{\beta} \right] = 0, \tag{2.4}$$

where $J_{\beta_{(-1)}} = \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\beta}_{(-1)}} = (b_1, \ldots, b_p)^\top$ is a $p \times (p-1)$ matrix, $b_1 = -(1 - \|\boldsymbol{\beta}_{(-1)}\|^2)^{-1/2}\boldsymbol{\beta}_{(-1)}$ and $b_s = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$ with sth element equals to 1, $s = 2, \ldots, p$. Let $\hat{L}'_k(\boldsymbol{\beta})$ be the kth component of $\frac{\partial \hat{L}(\boldsymbol{\beta}_{(-1)})}{\partial \boldsymbol{\beta}_{(-1)}}$. Therefore, $(\hat{L}'_2(\boldsymbol{\beta}), \ldots, \hat{L}'_p(\boldsymbol{\beta}))^\top = 0$ are estimating equations of $\boldsymbol{\beta}_{(-1)}$. Since the sparsity property is satisfied, motivated by the score test screening procedure proposed by Zhao and Li [12], for each $k (k \neq 1)$, we consider a marginal estimating equation of $\beta_k (k = 2 \ldots, p)$ and assume that all the other covariates are unrelated to the linear discriminant classification except X_1. Denote this marginal estimating equation by $L'_k(\boldsymbol{\beta})$, and $\hat{\omega}_k(\beta_k) = \hat{L}'_k(\beta_1, 0, \ldots, 0, \beta_k, 0, \ldots, 0) = 0$. From this marginal estimating equation, if $|\hat{\omega}_k(0)|$ is bigger than 0, it means that $\beta_k = 0$ is not the solution of this estimating equation. Thus X_k is a possible active predictor. Otherwise, the coefficient $\beta_k = 0$ denotes that X_k is not important in the linear discriminant analysis. Therefore, similar to Zhao and Li [12] and Ma et al. [13], each $|\hat{\omega}_k(0)| = \hat{L}'_k(1, 0, \ldots, 0)$ is the numerator of the score statistic for a hypothesis: $\beta_k = 0 (k \geq 2)$ under the kth marginal model and therefore can be a sensible screening statistic. Here $\beta_1 = 1$ is from $\|\boldsymbol{\beta}\| = 1$. Let $\hat{\omega}_k = \hat{\omega}_k(0)$. It follows

$$\hat{\omega}_k = \hat{\omega}_k(0) = \frac{2}{E_{11}^2} \left[E_{11}B_{k1} - B_{11}E_{k1} \right], k = 2, \ldots, p, \tag{2.5}$$

where $E_{11} = \sum_{j=1}^2 \sum_{i=1}^n [X_{i1} - \bar{X}_{j1}]^2 I(Y_i = j)$, $E_{k1} = \sum_{j=1}^2 \sum_{i=1}^n [X_{ik} - \bar{X}_{jk}][X_{i1} - \bar{X}_{j1}] I(Y_i = j)$, $B_{11} = \sum_{j=1}^2 \sum_{i=1}^n [\bar{X}_{j1} - \bar{X}_1]^2 I(Y_i = j)$ and $B_{k1} = \sum_{j=1}^2 \sum_{i=1}^n [\bar{X}_{jk} - \bar{X}_k][\bar{X}_{j1} - \bar{X}_1] I(Y_i = j)$. To simplify the calculation and theoretical derivation, define

$$\hat{\omega}_k^* = \hat{\omega}_k^*(0) = \frac{1}{n^2} \left[E_{11}B_{k1} - B_{11}E_{k1} \right], k = 2, \ldots, p. \tag{2.6}$$

Note that $\hat{\omega}_k^*$ is a scaled version of $\hat{\omega}_k$. They lead to the same result of feature ranking and screening.

Define $\omega_k^* = \omega_k^*(0) = T_{11}T_{12} - T_{21}T_{22}$, where

$$T_{11} = \sum_{j=1}^2 E\left\{ \left[X_1 - \frac{E(X_1 I_j)}{E(I_j)} \right]^2 I_j \right\}, \quad T_{21} = \sum_{j=1}^2 E\left\{ \left[\frac{E(X_1 I_j)}{E(I_j)} - E(X_1) \right]^2 I_j \right\},$$

$$T_{12} = \sum_{j=1}^2 E\left\{ \left[\frac{E(X_k I_j)}{E(I_j)} - E(X_k) \right] \left[\frac{E(X_1 I_j)}{E(I_j)} - E(X_1) \right] I_j \right\},$$

$$T_{22} = \sum_{j=1}^{2} E\left\{\left[X_k - \frac{E(X_k I_j)}{E(I_j)}\right]\left[X_1 - \frac{E(X_1 I_j)}{E(I_j)}\right] I_j\right\}, \tag{2.7}$$

where $I_j = I(Y = j)$, and $I(\cdot)$ is the indicator function. From (2.6), if X_k and Y are independent, then it follows

$$\hat{\omega}_k^* \xrightarrow{P} \omega_k^* = -\sum_{j=1}^{2} E\left\{I_j \frac{E^2(X_1 I_j)}{E^2(I_j)}\right\} \cdot \sum_{j=1}^{2} E\left\{X_1 X_k I_j\right\} = 0,\ n \to \infty.$$

Therefore, $\hat{\omega}_k$ could be used as the feature screening index.

For a given threshold value c_n, the active set is estimated as

$$\hat{\mathcal{A}}_{c_n} = \{2 \le k \le p : |\hat{\omega}_k^*| \ge c_n\}. \tag{2.8}$$

Usually, the predefined c_n is not easy to be identified. Another way is to select the top d_n predictors and estimate the active set as

$$\hat{\mathcal{A}}_{d_n} = \{2 \le k \le p : |\hat{\omega}_k^*| \text{ ranks among the top } d_n\}. \tag{2.9}$$

The submodel size d_n is a predefined threshold value, e.g., $d_n = v[n/\log(n)]$, v is some positive integer, see Fan and Lv [14]. In practice, v is chosen to be bigger than 1 to enhance the probability of selecting all the relevant predictors.

2.2. Sure screening property

Next, we establish the theoretical property of the proposed feature screening method. To study the sure screening property, the following regularity conditions are assumed.

- C1. **X** satisfies the sub-exponential tail probability uniformly in p. That is, there exists a positive constant s_0 such that for all $0 < s \le 2s_0$,

$$\sup_p \max_{1 \le k \le p} E\left\{\exp\left(sX_k^2\right)\right\} < \infty.$$

- C2. There exist some constants $c > 0$ and $0 \le \kappa < \frac{1}{2}$ such that $\min_{k \in \mathcal{A}} \omega_k^* \ge 2cn^{-\kappa}$.

Theorem 1. *(Sure Screening Property) Under Condition C1, for any $0 < \gamma < \frac{1}{2} - \kappa$, there exist positive constants $c_1 > 0$ and $c_2 > 0$ such that*

$$P\left(\max_{1 \le k \le p} \left|\hat{\omega}_k^* - \omega_k^*\right| \ge cn^{-\kappa}\right) \le O\left\{2p \exp\left(-c_1 n^{1-2\gamma-2\kappa}\right) + 2np \exp\left(-c_2 n^{\gamma}\right)\right\}. \tag{2.10}$$

Further, if both conditions C1 and C2 hold, by taking $c_n = cn^{-\kappa}$ in (2.8), we have

$$P(\mathcal{A} \subset \hat{\mathcal{A}}_{c_n}) \ge 1 - O\left\{2s_n \exp\left(-c_1 n^{1-2\gamma-2\kappa}\right) + 2ns_n \exp\left(-c_2 n^{\gamma}\right)\right\}, \tag{2.11}$$

where s_n is the cardinality of $\mathcal{A}$, which is sparse and may vary with n.

Theorem 2. *(Minimum Model Size) Under conditions in Theorem 1, for any $c_n = c_3 n^{-\kappa}, c_3 > 0$, there exist positive constants c_4 and c_5, such that*

$$P\left(\left\|\hat{\mathcal{A}}_{c_n}\right\|_0 \leq O(n^{\kappa}\sum_{k=1}^{p}|\omega_k^*|)\right) \geq 1 - O\left\{2p\exp\left(-c_4 n^{1-2\gamma-2\kappa}\right) + 2np\exp\left(-c_5 n^{\gamma}\right)\right\}. \tag{2.12}$$

Here $\|\cdot\|_0$ denotes the cardinality of a set.

Remark 1. *Theorem 1 shows that the sure screening property holds for the proposed linear projection feature screening procedure. The dimensionality p is allowed to increase at an exponential rate of the sample size n, i.e., $p = o(\exp(n^{\alpha}))$. From (2.10) of Theorem 1, the left term of (2.10) tends to 0 if $0 < \alpha < 1 - 2\gamma - 2\kappa$. Furthermore, it shows that the feature screening procedure could retain all the important classification predictors with probability tending to 1, which means $P(\mathcal{A} \subset \hat{\mathcal{A}}_{c_n}) \to 1$. The screened features could be utilized in the linear discriminant analysis. If the dimensionality is still high, some penalized methods could be processed. Theorem 2 shows that as long as $\sum_{k=1}^{p}|\omega_k^*|$ is of a polynomial order of sample size, then the number of selected variables is also of polynomial order of sample size.*

3. Numerical examples

In this section, we present two simulation studies of the popular discriminant analysis models, the logistic model and the probit model, and one real data analysis to assess the finite sample performances of the proposed method (LDA-SIS). Furthermore, we compare the effectiveness of our proposed method with other existing competitive screening methods, including the T-test (Fan and Fan [1]), DC (Li et al. [15]), KF (Mai and Zou [3]), MV (Cui et al. [6]), PSIS (Pan et al. [7]) and RRS (Cheng et al. [8]).

3.1. Monte Carlo Simulations

For each simulation, we set the dimensionality p to 1000 and 2000, and the sample size n to 100 and 200, respectively. All the simulation results are based on 1000 replications. Similar to Fan and Lv [14] and Li et al. [15], the screening threshold number is set to be $d_n = [n/\log(n)]$. The following criteria are considered to evaluate the performances of all screening methods.

- MMS: The minimum model size of the submodel contains all active predictors. The five quantiles of MMS over 1000 replications are presented.
- P_k: The proportion of the kth active predictor is selected into the model with size d_n.
- P_a: The proportion that all active predictors are selected into the model.

Example 1 (Logistic Model): Consider the logistic regression model

$$logit(p_y) = \mathbf{X}^{\top}\beta, \quad p_y = P(Y = 1|\mathbf{X}).$$

The covariate $\mathbf{X} = (X_1, \mathbf{X}_{(-1)})^{\top}, \mathbf{X}_{(-1)} = (X_2, \ldots, X_p)^{\top}$ is generated from $X_1 \sim \mathcal{N}(0,1)$ and $\mathbf{X}_{(-1)} \sim \mathcal{N}_{p-1}(0,\Sigma)$, where Σ is a $(p-1)\times(p-1)$ covariance matrix with elements $\sigma_{ij} = \rho^{|i-j|}, i, j = 1, \ldots, p-1$. We consider $\rho = 0.2, 0.5$ and 0.8, respectively. Set $\beta = (1.4, 1.2, 1.0, 0.8, 0.6, \mathbf{0}_{p-5})^{\top}$ and the random

error ε which is added to $\mathbf{X}^\top\beta$ follows $N(0, 1)$. The simulation results of Example 1 are shown in Tables 1 and 2.

Table 1. The selecting rates P_a and $P_k's$ in Example 1.

		$n = 100$						$n = 200$					
	Method	P_1	P_2	P_3	P_4	P_5	P_a	P_1	P_2	P_3	P_4	P_5	P_a
$\rho = 0.2$	LDA-SIS	1	0.94	0.944	0.789	0.372	0.276	1	1	1	0.997	0.83	0.827
$p = 1000$	DC	1	0.887	0.887	0.678	0.293	0.17	1	0.996	1	0.98	0.737	0.721
	T-test	1	0.914	0.904	0.725	0.316	0.202	1	0.999	1	0.988	0.779	0.771
	RRS	1	0.896	0.899	0.688	0.318	0.187	1	0.998	1	0.985	0.751	0.74
	KF	1	0.805	0.803	0.555	0.231	0.087	1	0.99	0.994	0.934	0.631	0.577
	MV	1	0.881	0.879	0.662	0.286	0.159	1	0.996	0.999	0.976	0.722	0.702
	PSIS	1	0.914	0.904	0.725	0.316	0.202	1	0.999	1	0.988	0.779	0.771
$p = 2000$	LDA-SIS	1	0.898	0.907	0.676	0.283	0.158	1	0.998	0.999	0.989	0.75	0.74
	DC	1	0.809	0.819	0.556	0.213	0.083	1	0.997	0.999	0.96	0.643	0.621
	T-test	1	0.836	0.841	0.6	0.235	0.1	1	0.998	0.998	0.966	0.678	0.657
	RRS	1	0.812	0.82	0.572	0.223	0.092	1	0.998	0.999	0.961	0.657	0.633
	KF	1	0.679	0.699	0.465	0.164	0.035	1	0.988	0.985	0.877	0.532	0.463
	MV	1	0.787	0.796	0.546	0.203	0.075	1	0.995	0.998	0.946	0.631	0.599
	PSIS	1	0.836	0.841	0.6	0.235	0.1	1	0.998	0.998	0.966	0.678	0.657
$\rho = 0.5$	LDA-SIS	1	0.996	0.998	0.993	0.828	0.821	1	1	1	1	0.998	0.998
$p = 1000$	DC	1	0.99	0.997	0.979	0.759	0.739	1	1	1	1	0.994	0.994
	T-test	1	0.993	0.999	0.989	0.787	0.774	1	1	1	1	0.995	0.995
	RRS	1	0.991	0.995	0.982	0.764	0.743	1	1	1	1	0.994	0.994
	KF	1	0.963	0.992	0.944	0.644	0.593	1	1	1	1	0.977	0.977
	MV	1	0.986	0.997	0.972	0.737	0.715	1	1	1	1	0.992	0.992
	PSIS	1	0.993	0.999	0.989	0.787	0.774	1	1	1	1	0.995	0.995
$p = 2000$	LDA-SIS	1	0.996	0.998	0.985	0.771	0.761	1	1	1	1	0.993	0.993
	DC	1	0.982	0.99	0.971	0.681	0.65	1	1	1	1	0.972	0.972
	T-test	1	0.985	0.993	0.978	0.724	0.702	1	1	1	1	0.984	0.984
	RRS	1	0.978	0.991	0.974	0.684	0.652	1	1	1	1	0.976	0.976
	KF	1	0.927	0.97	0.913	0.55	0.464	1	1	1	0.999	0.941	0.941
	MV	1	0.969	0.988	0.963	0.653	0.612	1	1	1	1	0.964	0.964
	PSIS	1	0.985	0.993	0.978	0.724	0.702	1	1	1	1	0.984	0.984
$\rho = 0.8$	LDA-SIS	1	1	1	1	0.999	0.999	1	1	1	1	1	1
$p = 1000$	DC	1	1	1	1	0.999	0.999	1	1	1	1	1	1
	T-test	1	1	1	1	0.999	0.999	1	1	1	1	1	1
	RRS	1	1	1	1	0.999	0.999	1	1	1	1	1	1
	KF	1	1	1	0.999	0.994	0.994	1	1	1	1	1	1
	MV	1	1	1	1	0.998	0.998	1	1	1	1	1	1
	PSIS	1	1	1	1	0.999	0.999	1	1	1	1	1	1
$p = 2000$	LDA-SIS	1	1	1	1	1	1	1	1	1	1	1	1
	DC	1	1	1	1	1	1	1	1	1	1	1	1
	T-test	1	1	1	1	1	1	1	1	1	1	1	1
	RRS	1	1	1	1	0.999	0.999	1	1	1	1	1	1
	KF	1	1	0.999	0.999	0.981	0.979	1	1	1	1	1	1
	MV	1	1	1	1	0.999	0.999	1	1	1	1	1	1
	PSIS	1	1	1	1	1	1	1	1	1	1	1	1

Table 2. The different quantiles of MMS in Example 1.

		$n = 100$					$n = 200$				
	Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
$\rho = 0.2$	LDA-SIS	7	19	56	173.25	575	5	5	8	22	153.05
$p = 1000$	DC	9	34	106	279.25	736.55	5	6	12	43	275.1
	T-test	8	27	80	233.25	651.2	5	6	11	35	212.1
	RRS	9	31	95	251	726.25	5	6	12	41	236
	KF	15	56	167.5	343	770.3	5	9	25	90.25	361.4
	MV	10	38.75	115.5	288.25	764.3	5	6	14	49	293.05
	PSIS	8	27	80	233.25	651.2	5	6	11	35	212.1
$p = 2000$	LDA-SIS	10	38	124	394.5	1345.05	5	5	11	42.25	320.55
	DC	14	73	220.5	605.75	1461.05	5	7	19	95.25	689
	T-test	13	59	185.5	488.75	1404	5	6	16	72	482.3
	RRS	13	66.75	200.5	567.5	1406.1	5	7	18	84.25	540.2
	KF	29.95	113.5	338.5	741.75	1536.35	5	12	45	175	810.05
	MV	15	74	246	640.5	1455.3	5	7	22	107	695.05
	PSIS	13	59	185.5	488.75	1404	5	6	16	72	482.3
$\rho = 0.5$	LDA-SIS	5	5	6	14	89	5	5	5	5	7
$p = 1000$	DC	5	5	8	23	155	5	5	5	5	9
	T-test	5	5	7	18	123.05	5	5	5	5	7
	RRS	5	5	8	22	136.1	5	5	5	5	8.05
	KF	5	7	15	51	238.1	5	5	5	7	20
	MV	5	5	9	27.25	174.05	5	5	5	5	10
	PSIS	5	5	7	18	123.05	5	5	5	5	7
$p = 2000$	LDA-SIS	5	5	7	20	175.15	5	5	5	5	8
	DC	5	5	11	39	303	5	5	5	5	19
	T-test	5	5	10	31	241.45	5	5	5	5	16
	RRS	5	6	11	38	270.05	5	5	5	5	18.05
	KF	5	9	24	91	486	5	5	5	8.25	46
	MV	5	6	13	49	298.5	5	5	5	6	22
	PSIS	5	5	10	31	241.45	5	5	5	5	16
$\rho = 0.8$	LDA-SIS	5	5	5	5	6	5	5	5	5	5
$p = 1000$	DC	5	5	5	5	6	5	5	5	5	5
	T-test	5	5	5	5	6	5	5	5	5	5
	RRS	5	5	5	5	6	5	5	5	5	5
	KF	5	5	5	6	8	5	5	5	5	6
	MV	5	5	5	5	6	5	5	5	5	5
	PSIS	5	5	5	5	6	5	5	5	5	5
$p = 2000$	LDA-SIS	5	5	5	5	6	5	5	5	5	5
	DC	5	5	5	5	6	5	5	5	5	5
	T-test	5	5	5	5	6	5	5	5	5	5
	RRS	5	5	5	5	6	5	5	5	5	5
	KF	5	5	5	6	11	5	5	5	5	6
	MV	5	5	5	5	7	5	5	5	5	5
	PSIS	5	5	5	5	6	5	5	5	5	5

Example 2 (Probit Model): Consider the probit regression model

$$p_y = \Phi(\mathbf{X}^\top \beta), \quad p_y = P(Y = 1|\mathbf{X}),$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. Assume that the true active set $\mathcal{A} = \{1, 5, 20, 21, 100\}$, and β is the p dimensional parametric vector with $\beta_{\mathcal{A}} = (1, 1, 1, 1, 1)^\top$

and 0 otherwise. The other settings are the same to Example 1. The simulation results of Example 2 are shown in Tables 3 and 4.

Table 3. The selecting rates P_a and $P_k's$ in Example 2.

		$n = 100$						$n = 200$					
	Method	P_1	P_2	P_3	P_4	P_5	P_a	P_1	P_2	P_3	P_4	P_5	P_a
$\rho = 0.2$	LDA-SIS	1	0.779	0.933	0.934	0.76	0.499	1	0.991	0.999	1	0.992	0.982
$p = 1000$	DC	1	0.703	0.882	0.893	0.689	0.362	1	0.976	0.998	0.998	0.984	0.956
	T-test	1	0.727	0.9	0.907	0.713	0.401	1	0.983	1	0.999	0.987	0.969
	RRS	1	0.713	0.889	0.895	0.697	0.376	1	0.976	0.998	0.997	0.986	0.957
	KF	1	0.595	0.808	0.785	0.581	0.2	1	0.946	0.993	0.991	0.945	0.88
	MV	1	0.684	0.873	0.871	0.661	0.324	1	0.971	0.998	0.997	0.977	0.943
	PSIS	1	0.727	0.9	0.907	0.713	0.401	1	0.983	1	0.999	0.987	0.969
$p = 2000$	LDA-SIS	1	0.697	0.88	0.884	0.721	0.359	1	0.985	1	0.998	0.975	0.958
	DC	1	0.616	0.822	0.818	0.644	0.226	1	0.962	0.998	0.996	0.96	0.916
	T-test	1	0.649	0.852	0.848	0.673	0.275	1	0.976	1	0.998	0.967	0.941
	RRS	1	0.621	0.825	0.82	0.648	0.229	1	0.968	0.998	0.996	0.96	0.922
	KF	1	0.466	0.703	0.696	0.497	0.084	1	0.903	0.983	0.99	0.896	0.783
	MV	1	0.583	0.803	0.793	0.62	0.184	1	0.958	0.998	0.995	0.949	0.9
	PSIS	1	0.649	0.852	0.848	0.673	0.275	1	0.976	1	0.998	0.967	0.941
$\rho = 0.5$	LDA-SIS	1	0.725	0.99	0.995	0.705	0.486	1	0.985	1	1	0.982	0.968
$p = 1000$	DC	1	0.655	0.975	0.982	0.644	0.373	1	0.965	1	1	0.958	0.926
	T-test	1	0.691	0.983	0.986	0.681	0.429	1	0.972	1	1	0.97	0.944
	RRS	1	0.665	0.976	0.984	0.656	0.397	1	0.967	1	1	0.968	0.937
	KF	1	0.535	0.926	0.938	0.526	0.219	1	0.922	0.999	1	0.897	0.824
	MV	1	0.63	0.969	0.977	0.624	0.349	1	0.958	1	1	0.948	0.908
	PSIS	1	0.691	0.983	0.986	0.681	0.429	1	0.972	1	1	0.97	0.944
$p = 2000$	LDA-SIS	1	0.636	0.979	0.977	0.645	0.369	1	0.959	1	1	0.96	0.919
	DC	1	0.556	0.962	0.957	0.566	0.274	1	0.945	1	1	0.939	0.884
	T-test	1	0.589	0.975	0.969	0.617	0.326	1	0.951	1	1	0.952	0.903
	RRS	1	0.567	0.962	0.956	0.57	0.274	1	0.948	1	1	0.942	0.891
	KF	1	0.428	0.896	0.892	0.443	0.125	1	0.877	0.998	1	0.867	0.76
	MV	1	0.534	0.941	0.95	0.539	0.245	1	0.935	1	1	0.929	0.868
	PSIS	1	0.589	0.975	0.969	0.617	0.326	1	0.951	1	1	0.952	0.903
$\rho = 0.8$	LDA-SIS	1	0.641	0.999	0.999	0.57	0.343	1	0.966	1	1	0.953	0.922
$p = 1000$	DC	1	0.592	0.997	0.999	0.519	0.283	1	0.95	1	1	0.91	0.863
	T-test	1	0.625	0.999	1	0.532	0.301	1	0.962	1	1	0.932	0.895
	RRS	1	0.598	0.997	0.998	0.525	0.287	1	0.954	1	1	0.926	0.881
	KF	1	0.47	0.983	0.982	0.42	0.178	1	0.898	1	1	0.857	0.769
	MV	1	0.561	0.996	0.995	0.506	0.259	1	0.95	1	1	0.907	0.86
	PSIS	1	0.625	0.999	1	0.532	0.301	1	0.962	1	1	0.932	0.895
$p = 2000$	LDA-SIS	1	0.588	1	0.999	0.496	0.265	1	0.96	1	1	0.922	0.883
	DC	1	0.519	0.998	0.996	0.434	0.199	1	0.931	1	1	0.877	0.812
	T-test	1	0.562	0.998	0.997	0.459	0.229	1	0.951	1	1	0.895	0.847
	RRS	1	0.522	0.995	0.997	0.432	0.202	1	0.933	1	1	0.882	0.819
	KF	1	0.414	0.982	0.974	0.34	0.127	1	0.834	1	1	0.784	0.641
	MV	1	0.496	0.995	0.994	0.411	0.176	1	0.913	1	1	0.86	0.782
	PSIS	1	0.562	0.998	0.997	0.459	0.229	1	0.951	1	1	0.895	0.847

Table 4. The different quantiles of MMS in Example 2.

		n = 100					n = 200				
	Method	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
ρ = 0.2	LDA-SIS	5	10	22	55	252	5	5	5	6	17.05
p = 1000	DC	6	15	38	92	394.25	5	5	6	8	36
	T-test	6	12	30	77	288.05	5	5	5	7	23.05
	RRS	6	14	34	89	316.2	5	5	5	8	31.05
	KF	10	26.75	64.5	151	453.05	5	6	8	18	99
	MV	6	16	40	105	410.05	5	5	6	10	42.05
	PSIS	6	12	30	77	288.05	5	5	5	7	23.05
p = 2000	LDA-SIS	6	15	36	101.5	490.1	5	5	5	7	29
	DC	8	24	63	181	720.4	5	5	6	11	55.05
	T-test	7	19.75	49	135	561.55	5	5	6	9	42
	RRS	8	23	56	164.25	634	5	5	6	10	58
	KF	14	52	124	298.25	957.2	5	6	11	30	157.05
	MV	8	28	69	197.25	784.55	5	5	7	12	76.05
	PSIS	7	19.75	49	135	561.55	5	5	6	9	42
ρ = 0.5	LDA-SIS	5	9	22	71	295.05	5	5	6	9	28
p = 1000	DC	6	13	35	113	444.1	5	5	7	12	49
	T-test	6	11	29	86	351.05	5	5	6	10	41.05
	RRS	6	13	33.5	95.25	399.3	5	5	7	12	46
	KF	8	24	65	165	516.7	5	6	10	24	106.1
	MV	6	14	41	118	490.75	5	5	7	13.25	55
	PSIS	6	11	29	86	351.05	5	5	6	10	41.05
p = 2000	LDA-SIS	6	13	37	114	556.95	5	5	6	10	58.05
	DC	7	20	56	191.25	813.25	5	5	7	15	98.1
	T-test	6	16	45	155	616.7	5	5	7	12.25	74
	RRS	7	19	52.5	180.75	679	5	5	7	14.25	89.05
	KF	10.95	40	118	320.25	978.35	5	7	13	36	228.15
	MV	7	22.75	64	219.25	801.65	5	6	8	17	121.05
	PSIS	6	16	45	155	616.7	5	5	7	12.25	74
ρ = 0.8	LDA-SIS	8	16	31	86.25	334.25	7.95	11	14	19	51.05
p = 1000	DC	9	20	45	125.25	508.15	8	11	15	23	78.05
	T-test	9	18	37.5	105	399	8	11	15	21	65.05
	RRS	9	19	41	114	444.1	8	11	15	22	68
	KF	10	31	74.5	188.25	564.25	8	12	19	34	136.05
	MV	10	21	48	135.25	523.3	8	12	16	24	81.05
	PSIS	9	18	37.5	105	399	8	11	15	21	65.05
p = 2000	LDA-SIS	9	21	48	151	619.1	8	11	14	22	66.05
	DC	10	28	73	228.5	897.3	8	11	17	29	106.05
	T-test	9.95	23	60.5	172	676.45	8	11	15	26	81
	RRS	10	28	70	201.25	746.25	8	11	17	29	101
	KF	13	49	136	334.25	1113.7	8	13	24	61	211.15
	MV	10	33	82	241.25	957	8	11	18	33	121.2
	PSIS	9.95	23	60.5	172	676.45	8	11	15	26	81

From Tables 1–4, we can find that the proposed LDA-SIS procedure has better feature screening performances than the other procedures. The proportion of all active predictors selected into the screened submodel (P_a) is larger for the LDA-SIS procedure, and the minimum model size of the submodel which contains all active predictors (MMS) of the LDA-SIS procedure is smaller. From Tables 1 and 2, with the correlation parameter ρ increasing, the performances of the feature screening

procedures become better. This phenomenon shows that when the active predictors have strong relationships with each other, the proposed feature screening procedure could select the important predictors more correctly. On the other hand, from Tables 3 and 4, with the correlation parameter ρ increasing, the performances of the feature screening procedures become worse. It shows that when the active predictors have strong relationships with the inactive predictors, the feature screening accuracy would be compromised. Furthermore, with the sample size increasing, better results would be obtained.

3.2. Real data analysis

We applied the LDA-SIS feature screening procedure to ovarian cancer data previously studied by Sorace and Zhan [2], Fushiki et al. [16], Zhang et al. [17], and Zhang et al. [18]. This dataset was generated using surface-enhanced laser desorption time-of-flight mass spectrometry and comprises serum samples from 162 ovarian cancer patients and 91 control subjects. The data are available on the Clinical Proteomics Program Databank website (http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp). For each ovarian cancer sample, we analyzed 15,154 distinct mass-to-charge ratios (M/Z). As reported by Sorace and Zhan [2], the region with M/Z values below 500 was often discarded as noise, resulting in a reduction of the dimensionality of the biomarker features from 15,154 to 12,757. Our goal in this study was to identify proteomic patterns corresponding to specific M/Z values that could distinguish ovarian cancer subjects from control subjects.

We randomly split the 253 samples into the training data set and the testing data set. In particular, we sampled approximately $100\gamma\%$ of the ovarian cancer patients and $100\gamma\%$ of the control subjects as the training data set, and the rest as the testing data set. We standardized the data to zero mean and unit variance before the discriminant classification.

Different feature screening procedures are utilized to identify the important potential biomarkers in the standardized training data. In our LDA-SIS procedure, we select the variable with the largest value of the Kolmogorov-Smirnov statistics (Mai and Zou [3]) as X_1. Let $d_n = [c_0 n_{tr}/\log(n_{tr})]$, $c_0 = 0.25, 0.5, 1$, and n_{tr} is the sample size of the training data. In this case, $d_n = 8$, 17, and 35, respectively. After the first feature screening step, the kernel support vector machine (KSVM) method with Gaussian kernel function, and the penalized logistic model (PLM) with LASSO method (Tibshirani, R. [19]) are applied in the modeling step based on the screened d_n potential biomarkers, respectively. And their performances are evaluated by the testing data. The packages **e1071** and **glmnet** are used here.

The procedure is repeated 200 times with $\gamma = 0.7$ and 0.8, respectively. Three assessment criteria are introduced to investigate the classification performance of the different methods.

- Testing error: The number of errors identified on the testing set.
- TPR (sensitivity or true positive rate): The proportion of ovarian cancer patients diagnosed correctly.
- PPV (positive predictive value): The proportion of samples diagnosed with ovarian cancer who did have the disease.

Table 5 summarizes the median and robust standard deviation (RSD) in the parentheses of testing error, TPR, and PPV. The results show that our proposed LDA-SIS method outperforms the other methods based on these evaluation criteria. Furthermore, we observe that increasing the proportion of training data selected ($\gamma = 0.8$) leads to better performance across all model sizes $d_n(8, 17, 35)$.

Table 5. Classification performance of ovarian cancer data.

	d_n	Assessment Criteria	LDA-SIS	DC	T-test	RRS	KF	MV	PSIS
	8	KSVM-Testing error	3(1.49)	8(2.43)	8(2.24)	8(2.43)	8(2.24)	8(2.24)	8(2.24)
		KSVM-TPR	0.97(0.06)	0.83(0.07)	0.83(0.07)	0.85(0.08)	0.84(0.07)	0.85(0.07)	0.83(0.07)
		KSVM-PPV	0.97(0.04)	0.96(0.04)	0.96(0.04)	0.95(0.04)	0.96(0.04)	0.95(0.04)	0.96(0.04)
		PLM-Testing error	3(2.24)	5(2.24)	5(2.24)	5(2.24)	6(2.99)	5(2.24)	5(2.24)
		PLM-TPR	0.97(0.05)	0.94(0.05)	0.93(0.05)	0.94(0.04)	0.93(0.06)	0.94(0.05)	0.93(0.05)
		PLM-PPV	0.97(0.04)	0.92(0.05)	0.92(0.05)	0.93(0.07)	0.91(0.07)	0.93(0.06)	0.92(0.06)
0.8	17	KSVM-Testing error	3(2.24)	8(2.99)	8(2.99)	7(3.73)	8(2.99)	7(2.99)	8(2.99)
		KSVM-TPR	0.94(0.04)	0.83(0.06)	0.83(0.06)	0.87(0.08)	0.85(0.07)	0.87(0.08)	0.83(0.06)
		KSVM- PPV	0.97(0.04)	0.96(0.04)	0.95(0.04)	0.94(0.04)	0.94(0.04)	0.94(0.04)	0.95(0.04)
		PLM-Testing error	2(1.49)	4(2.24)	5(2.24)	3(2.24)	4(2.24)	3(2.24)	5(2.24)
		PLM-TPR	0.97(0.04)	0.94(0.05)	0.94(0.07)	0.95(0.04)	0.93(0.05)	0.95(0.04)	0.94(0.06)
		PLM-PPV	0.97(0.04)	0.95(0.04)	0.94(0.05)	0.96(0.06)	0.95(0.07)	0.96(0.06)	0.94(0.05)
	35	KSVM-Testing error	3(2.24)	8(3.73)	8(3.73)	6(2.99)	7(3.73)	6(2.99)	8(3.73)
		KSVM-TPR	0.97(0.07)	0.84(0.07)	0.85(0.07)	0.88(0.08)	0.85(0.08)	0.88(0.08)	0.85(0.07)
		KSVM- PPV	0.97(0.04)	0.95(0.04)	0.95(0.04)	0.94(0.04)	0.94(0.04)	0.94(0.04)	0.95(0.04)
		PLM-Testing error	2(1.49)	3(2.24)	3(2.24)	3(2.24)	3(2.24)	3(2.24)	3(2.24)
		PLM-TPR	0.97(0.04)	0.96(0.03)	0.96(0.04)	0.96(0.05)	0.96(0.04)	0.96(0.04)	0.96(0.04)
		PLM-PPV	0.97(0.04)	0.95(0.05)	0.96(0.06)	0.97(0.06)	0.95(0.04)	0.96(0.04)	0.96(0.06)
	8	KSVM-Testing error	4(2.24)	11(3.17)	12(3.17)	11(3.92)	11(2.99)	11(3.73)	12(3.17)
		KSVM-TPR	0.96(0.04)	0.83(0.06)	0.83(0.06)	0.85(0.07)	0.84(0.07)	0.85(0.07)	0.83(0.06)
		KSVM-PPV	0.97(0.03)	0.96(0.04)	0.96(0.04)	0.95(0.04)	0.96(0.04)	0.95(0.04)	0.96(0.04)
		PLM-Testing error	4(2.24)	9(2.99)	9(2.99)	8(2.99)	9(2.99)	8(2.99)	6(2.99)
		PLM-TPR	0.96(0.03)	0.94(0.05)	0.94(0.05)	0.95(0.05)	0.94(0.06)	0.95(0.05)	0.94(0.04)
		PLM-PPV	0.96(0.04)	0.89(0.05)	0.90(0.05)	0.91(0.05)	0.89(0.05)	0.90(0.04)	0.90(0.05)
0.7	17	KSVM-Testing error	5(2.24)	12(2.99)	12(3.73)	10(4.48)	11(4.48)	11(4.48)	12(3.73)
		KSVM-TPR	0.94(0.06)	0.84(0.07)	0.84(0.06)	0.87(0.08)	0.85(0.07)	0.87(0.07)	0.84(0.06)
		KSVM-PPV	0.98(0.04)	0.96(0.04)	0.95(0.04)	0.95(0.04)	0.95(0.04)	0.95(0.04)	0.95(0.04)
		PLM-Testing error	4(2.24)	7(2.99)	7(3.73)	6(2.99)	7(2.99)	6(2.99)	7(3.73)
		PLM-TPR	0.97(0.02)	0.95(0.04)	0.95(0.04)	0.96(0.04)	0.96(0.04)	0.96(0.04)	0.95(0.04)
		PLM-PPV	0.96(0.04)	0.92(0.05)	0.92(0.05)	0.93(0.04)	0.92(0.05)	0.93(0.04)	0.92(0.06)
	35	KSVM-Testing error	5(2.99)	12(3.73)	11(4.48)	9(4.66)	11(4.48)	10(4.48)	11(4.48)
		KSVM-TPR	0.94(0.07)	0.84(0.07)	0.85(0.07)	0.89(0.07)	0.87(0.08)	0.88(0.08)	0.85(0.07)
		KSVM-PPV	0.98(0.04)	0.95(0.04)	0.95(0.04)	0.95(0.04)	0.94(0.04)	0.94(0.04)	0.95(0.04)
		PLM-Testing error	3(1.49)	6(2.99)	6(2.24)	5(2.99)	6(2.99)	5(2.99)	5(2.24)
		PLM-TPR	0.98(0.02)	0.96(0.04)	0.96(0.03)	0.96(0.03)	0.96(0.04)	0.96(0.03)	0.96(0.03)
		PLM-PPV	0.98(0.02)	0.94(0.04)	0.94(0.04)	0.95(0.04)	0.94(0.04)	0.94(0.04)	0.94(0.04)

4. Conclusions

In this article, we employed Fisher's linear projection and the marginal score test to study the feature screening procedure for the ultrahigh-dimensional binary classification problem. Although many

feature screening procedures for the ultrahigh-dimensional discriminant analysis problems have been proposed, some of them are even model-free, the study for the LDA problem, one of the most popular approaches in discriminant classification and pattern recognition, is still very attractive. By minimizing the linear projection of the sum of squares in the original cluster structures and maximizing the linear projection of the sum of squares between groups, we constructed the marginal score test and combined it with the linear projection optimal problem to build the feature screening index. The sure screening property and the minimum model size of the procedure are studied. The sure screening property ensures that the feature screening procedure can retain all the important classification predictors with the probability tending to 1. And the minimum model size of the procedure proposed by Theorem 2 shows that as long as $\sum_{k=1}^{p} |\omega_k^*|$ is of a polynomial order of the sample size, the number of the selected variables is also a polynomial order of the sample size. The finite sample performance of the proposed procedure was illustrated by Monte Carlo studies and a real-data example. The simulation studies demonstrate that the proposed feature screening method performs well, and the simple structure of the screening index makes the calculation fast.

Acknowledgment

Peng Lai's research is supported by National Natural Science Foundation of China (11771215). Yanqiu Zhou's research is supported by Guangxi Science and Technology Base and Talent Project (2020ACI9151), and Guangxi University Young and Middle-aged Teachers Basic Research Ability Improvement Project (2021KY0343).

Conflict of interest

The authors declare no conflict of interest.

Appendix: Technical proofs

The proofs of Theorem 1 and Theorem 2 in this paper are similar to the proofs of Theorem 1 in Li et al. [15] and Theorems 1–2 in Liu et al. [20]. Similar lemmas are used here to facilitate proving the proposed theorems, these lemmas are listed in the following and the proofs of these lemmas can be found in the Appendices of Li et al. [15] and Liu et al. [20].

Lemma 1. *Let $\mu = E(X)$. If $P(a_1 \le X \le b_1) = 1$, then*

$$E\left[\exp\{s(X-\mu)\}\right] \le \exp\left\{s^2(b_1-a_1)^2/8\right\}, \text{ for any } s > 0.$$

Lemma 2. *Let $h(X_1, \ldots, X_m)$ be a kernel of the U statistics U_n, and $\theta = E\{h(X_1, \ldots, X_m)\}$. If $a \le h(X_1, \ldots, X_m) \le b$, then, for any $\epsilon > 0$ and $n > m$, we have*

$$P(U_n - \theta \ge \epsilon) \le \exp\left(-\frac{2[n/m]\epsilon^2}{(b-a)^2}\right),$$

where $[n/m]$ denotes the integer part of n/m.

Furthermore, due to the symmetry of U statistic, we also have

$$P(|U_n - \theta| \geq \epsilon) \leq 2\exp\left(-\frac{2[n/m]\epsilon^2}{(b-a)^2}\right).$$

In the following, we give the proofs of Theorem 1 and Theorem 2. For convenience, we denote M, M_i, and $c_i, i = 1, 2, \ldots$, as the generic constants depending on the context. Define $I_j = I(Y = j)$ and $I(Y_i = j) = I_{ij}$.

Proof of Theorem 1. For $\hat{\omega}_k^* - \omega_k^*$, we have

$$\hat{\omega}_k^* - \omega_k^* = \left[\hat{T}_{11}\hat{T}_{12} - \hat{T}_{21}\hat{T}_{22}\right] - \left[T_{11}T_{12} - T_{21}T_{22}\right], \tag{1}$$

where $\hat{T}_{11} = \frac{1}{n}E_{11}$, $\hat{T}_{12} = \frac{1}{n}B_{k1}$, $\hat{T}_{21} = \frac{1}{n}B_{11}$ and $\hat{T}_{22} = \frac{1}{n}E_{k1}$. We first consider $\hat{T}_{11} - T_{11}$.

Define $\tilde{T}_{11} = \frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n}[X_{i1} - \frac{E(X_1I_j)}{E(I_j)}]^2 I_{ij}$. We have

$$P(|\hat{T}_{11} - T_{11}| \geq \varepsilon) \leq P(|\hat{T}_{11} - \tilde{T}_{11}| + |\tilde{T}_{11} - T_{11}| \geq \varepsilon) \leq P(|\tilde{T}_{11} - T_{11}| \geq \varepsilon/2), \tag{2}$$

with n sufficiently large, i.e., $n \geq M_1$. It follows

$$P(|\tilde{T}_{11} - T_{11}| \geq \varepsilon/2) \leq \sum_{j=1}^{2} P\left(\left|\hat{T}_{11}^* - T_{11}^*\right| \geq \frac{\varepsilon}{4}\right),$$

where

$$\hat{T}_{11}^* = \frac{1}{n}\sum_{i=1}^{n}\left[X_{i1} - \frac{E(X_1I_j)}{E(I_j)}\right]^2 I_{ij} \quad \text{and} \quad T_{11}^* = E\left\{\left[X_1 - \frac{E(X_1I_j)}{E(I_j)}\right]^2 I_j\right\}.$$

Obviously, $\hat{T}_{11}^*$ is the U-statistic, and T_{11}^* is the kernel of the U-statistic of $\hat{T}_{11}^*$. Define $h_1(X_{i1}, Y_i) = \left[X_{i1} - \frac{E(X_1I_j)}{E(I_j)}\right]^2 I_{ij}$. Thus, we have

$$\begin{aligned}
\hat{T}_{11}^* &= \frac{1}{n}\sum_{i=1}^{n}\left[X_{i1} - \frac{E(X_1I_j)}{E(I_j)}\right]^2 I_{ij} I\left(h_1(X_{i1}, Y_i) \leq M\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left[X_{i1} - \frac{E(X_1I_j)}{E(I_j)}\right]^2 I_{ij} I\left(h_1(X_{i1}, Y_i) > M\right) \\
&:= \hat{T}_{111}^* + \hat{T}_{112}^*.
\end{aligned} \tag{3}$$

Accordingly, we decompose T_{11}^* into two parts

$$\begin{aligned}
T_{11}^* &= E\left\{\left[X_1 - \frac{E(X_1I_j)}{E(I_j)}\right]^2 I_j I\left(h_1(X_1, Y) \leq M\right)\right\} \\
&\quad + E\left\{\left[X_1 - \frac{E(X_1I_j)}{E(I_j)}\right]^2 I_j I\left(h_1(X_1, Y) > M\right)\right\} \\
&:= T_{111}^* + T_{112}^*.
\end{aligned} \tag{4}$$

Clearly, $\hat{T}^*_{111}$ and $\hat{T}^*_{112}$ are unbiased estimators of T^*_{111} and T^*_{112}.

Similar to the proof of Theorem 2 in Zhu et al.[21], with the Markov's inequality and the properties of U-statistic, for any $t > 0$, we can obtain that

$$\begin{aligned} P\left(\hat{T}^*_{111} - T^*_{111} \geq \varepsilon\right) &\leq \exp(-t\varepsilon)\exp\left(-tT^*_{111}\right) E\left\{\exp\left(t\hat{T}^*_{111}\right)\right\} \\ &\leq \exp(-t\varepsilon)E^n\left\{\exp\left(\frac{t}{n}\left[h_1(X_{i1}, Y_i)I(h_1(X_{i1}, Y_i) \leq M) - T^*_{111}\right]\right)\right\} \\ &\leq \exp\left(-t\varepsilon + \frac{M^2t^2}{8n}\right), \end{aligned}$$

where the last inequality is concluded from Lemma 1. By choosing $t = 4\varepsilon n/M^2$, we have

$$P\left(\hat{T}^*_{111} - T^*_{111} \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2 n}{M^2}\right).$$

Therefore, by the symmetry of U-statistic, we can get

$$P\left(|\hat{T}^*_{111} - T^*_{111}| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2 n}{M^2}\right). \tag{5}$$

Next, we show the consistency of $\hat{T}^*_{112}$.

With the Cauchy-Schwartz and Markov's inequalities, for any $s' > 0$,

$$(T^*_{112})^2 \leq E\left\{h_1^2(X_1, Y)\right\} E\left\{\exp(s'h_1(X_1, Y))\right\} / \exp(s'M).$$

Note that

$$h_1(X_1, Y) \leq 2X_1^2 + 2\frac{E^2(X_1I_j)}{E^2(I_j)},$$

which yields

$$E\left\{\exp(s'h_1(X_1, Y))\right\} \leq \exp\left(\frac{2s'E^2(X_1I_j)}{E^2(I_j)}\right) E\left\{\exp(2s'X_1^2)\right\}.$$

By condition C1, if we choose $M = cn^\gamma$, for $0 < \gamma < \frac{1}{2} - \kappa$, then $T^*_{112} \leq \frac{\varepsilon}{2}$ when n is sufficiently large. Consequently, similar to the proof of (B.4) in Li et al.[15], there exist some constant c_1 and some $s > 0$ such that

$$P\left(\left|\hat{T}^*_{112} - T^*_{112}\right| \geq \varepsilon\right) \leq P\left(\left|T^*_{112}\right| \geq \frac{\varepsilon}{2}\right) \leq c_1 n\exp\left(-\frac{sM}{4}\right). \tag{6}$$

Recall that $M = cn^\gamma$. Combining (3)–(6), we have

$$\begin{aligned} P\left(\left|\hat{T}^*_{11} - T^*_{11}\right| \geq \varepsilon\right) &\leq P\left(\left|\hat{T}^*_{111} - T^*_{111}\right| \geq \frac{\varepsilon}{2}\right) + P\left(\left|\hat{T}^*_{112} - T^*_{112}\right| \geq \frac{\varepsilon}{2}\right) \\ &\leq 2\exp\left(-2c_2\varepsilon^2 n^{1-2\gamma}\right) + c_1 n\exp\left(-c_3 n^\gamma\right), \end{aligned}$$

where c_2 and c_3 are some positive constants.

Similarly, for $u, v = 1, 2$, we can prove

$$P\left(\left|\hat{T}^*_{uv} - T^*_{uv}\right| \geq \varepsilon\right) \leq 2\exp\left(-2c_{2uv}\varepsilon^2 n^{1-2\gamma}\right) + c_{1uv}n\exp\left(-c_{3uv}n^{\gamma}\right),$$

where c_{1uv}, c_{2uv} and c_{3uv} are some positive constants. Therefore, it follows

$$\begin{aligned} P\left(\left|\hat{T}_{uv} - T_{uv}\right| \geq \varepsilon\right) &\leq \sum_{j=1}^{2} P\left(\left|\hat{T}^*_{uv} - T^*_{uv}\right| \geq \frac{\varepsilon}{2m}\right) \\ &\leq 4\exp\left(-\frac{c_{2uv}\varepsilon^2}{8}n^{1-2\gamma}\right) + 2c_{1uv}n\exp\left(-c_{3uv}n^{\gamma}\right), \end{aligned} \tag{7}$$

for $u, v = 1, 2$.

Therefore, similar to the proof of Lemma 4 and Lemma 5 in Liu et al.[20], by (1), (2) and (7), we get

$$\begin{aligned} P\left(\left|\hat{\omega}^*_k - \omega^*_k\right| \geq \varepsilon\right) &\leq P\left(\left|\hat{T}_{11}\hat{T}_{12} - T_{11}T_{12}\right| \geq \frac{\varepsilon}{2}\right) + P\left(\left|\hat{T}_{21}\hat{T}_{22} - T_{21}T_{22}\right| \geq \frac{\varepsilon}{2}\right) \\ &\leq 8\exp\left(-\frac{c_4\varepsilon^2}{4}n^{1-2\gamma}\right) + 2c_5 n\exp\left(-c_6 n^{\gamma}\right), \end{aligned}$$

where $c_4 - c_6$ are some positive constants. Thus,

$$P\left(\max_{1\leq k\leq p}\left|\hat{\omega}^*_k - \omega^*_k\right| \geq cn^{-\kappa}\right) \leq O\left\{2p\exp\left(-\frac{c_7}{4}n^{1-2\gamma-2\kappa}\right) + 2np\exp\left(-c_8 n^{\gamma}\right)\right\}. \tag{8}$$

Next, we prove the second part of Theorem 1 using the similar method of the proof of Theorem 1 in Li et al. [15]. If $\mathcal{A} \nsubseteq \hat{\mathcal{A}}_{c_n}$, then there must exist some $k \in \mathcal{A}$ such that $\hat{\omega}^*_k \leq cn^{-\kappa}$. Since $\min_{k\in\mathcal{A}} \omega^*_k \geq 2cn^{-\kappa}$, it indicates that

$$\left\{\mathcal{A} \nsubseteq \hat{\mathcal{A}}_{c_n}\right\} \subseteq \left\{\left|\hat{\omega}^*_k - \omega^*_k\right| > cn^{-\kappa}, \text{for some } k \in \mathcal{A}\right\}.$$

Therefore,

$$\begin{aligned} P(\mathcal{A} \subseteq \hat{\mathcal{A}}_{c_n}) &\geq 1 - P\left(\min_{k\in\mathcal{A}}\left|\hat{\omega}^*_k - \omega^*_k\right| \geq cn^{-\kappa}\right) \geq 1 - s_n P\left(\left|\hat{\omega}^*_k - \omega^*_k\right| \geq cn^{-\kappa}\right) \\ &\geq 1 - O\left\{2s_n\exp\left(-\frac{c_7}{4}n^{1-2\gamma-2\kappa}\right) + 2ns_n\exp\left(-c_8 n^{\gamma}\right)\right\}. \end{aligned}$$

This complete the proof of the second part.

Proof of Theorem 2. Note that for any $c_9 > 0$, the number of $\{k : |\omega^*_k| > \frac{c_9}{2}n^{-\kappa}\}$ is bounded by $O(n^{\kappa}\sum_{k=1}^{p}|\omega^*_k|)$. Then on the set

$$\mathcal{B} = \{\max_{1\leq k\leq p}|\hat{\omega}^*_k - \omega^*_k| \leq \frac{c_9}{2}n^{-\kappa}\},$$

the number of $\{k : |\hat{\omega}^*_k| > c_9 n^{-\kappa}\}$ can't exceed the number of $\{k : |\omega^*_k| > \frac{c_9}{2}n^{-\kappa}\}$. Therefore, we have

$$P\left(\left\|\hat{\mathcal{A}}_{c_n}\right\|_0 \leq O(n^{\kappa}\sum_{k=1}^{p}|\omega^*_k|)\right) \geq P(\mathcal{B}).$$

Then, by (8), the proof is completed.

References

1. J. Fan, Y. Fan, High dimensional classification using features annealed independence rules, *Ann. Stat.,* **36** (2008), 2605–2637. http://dx.doi.org/10.1214/07-AOS504
2. J. Sorace, M. Zhan, A data review and re-assessment of ovarian cancer serum proteomic profiling, *BMC Bioinformatics,* **4** (2003), 1–13. http://dx.doi.org/10.1186/1471-2105-4-24
3. Q. Mai, H. Zou, The Kolmogorov filter for variable screening in high-dimensional binary classification, *Biometrika,* **100** (2013), 229–234. http://dx.doi.org/10.1093/biomet/ass062
4. Q. Mai, H. Zou, The fused Kolmogorov filter: A nonparametric model-free screening method, *Ann. Stat.,* **43** (2015), 1471–1497. http://dx.doi.org/10.1214/14-AOS1303
5. P. Lai, F. Song, K. Chen, Z. Liu, Model free feature screening with dependent variable in ultrahigh dimensional binary classification, *Statist. Probab. Lett.,* **125** (2017), 141–148. https://doi.org/10.1016/j.spl.2017.02.011
6. H. Cui, R. Li, W. Zhong, Model-free feature screening for ultrahigh dimensional discriminant analysis, *J. Am. Stat. Assoc.,* **110** (2015), 630–641. http://dx.doi.org/10.1080/01621459.2014.920256
7. R. Pan, H. Wang, R. Li, Ultrahigh dimensional multi-class linear discriminant analysis by pairwise sure independence screening, *J. Am. Stat. Assoc.,* **111** (2016), 169–179. http://dx.doi.org/10.1080/01621459.2014.998760
8. G. Cheng, X. Li, P. Lai, F. Song, J. Yu, Robust rank screening for ultrahigh dimensional discriminant analysis, *Stat. Comput.,* **27** (2017), 535–545. http://dx.doi.org/10.1007/s11222-016-9637-2
9. S. He, S. Ma, W. Xu, A modified mean-variance feature-screening procedure for ultrahigh-dimensional discriminant analysis, *Comput. Stat. Data Anal.,* **137** (2019), 155–169. http://dx.doi.org/10.1016/j.csda.2019.02.003
10. F. Song, P. Lai, B. Shen, Robust composite weighted quantile screening for ultrahigh dimensional discriminant analysis, *Metrika,* **83** (2020), 799–820. https://doi.org/10.1007/s00184-019-00758-x
11. Y. Sheng, Q. Wang, Model-free feature screening for ultrahigh dimensional classification, *J. Multivar. Anal.,* **178** (2020), 104618. http://dx.doi.org/10.1016/j.jmva.2020.104618
12. S. Zhao, Y. Li, Score test variable screening, *Biometrics,* **70** (2014), 862–871. http://dx.doi.org/10.1111/biom.12209
13. Y. Ma, Y. Li, H. Lin, Concordance measure-based feature screening and variable selection, *Stat. Sinica,* **27** (2017), 1967–1985. http://dx.doi.org/10.5705/ss.202016.0024
14. J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, *J. R. Stat. Soc. Series B. Stat. Methodol.,* **70** (2008), 849–911. http://dx.doi.org/10.1111/j.1467-9868.2008.00674.x
15. R. Li, W. Zhong, L. Zhu, Feature screening via distance correlation Learning, *J. Am. Stat. Assoc.,* **107** (2012), 1129–1139. http://dx.doi.org/10.1080/01621459.2012.695654

16. T. Fushiki, H. Fujisawa, S. Eguchi, Identification of biomarkers from mass spectrometry data using a "common" peak approach, *BMC Bioinformatics,* **7** (2006), 358–366. http://dx.doi.org/10.1186/1471-2105-7-358

17. M. Zhang, W. Wang, Y. Du, ULDA-based heuristic feature selection method for proteomic profile analysis and biomarker discovery, *Chemometr. Intell. Lab. Syst.,* **102** (2010), 84–90. http://dx.doi.org/10.1016/j.chemolab.2010.04.005

18. M. Zhang, P. Tong, W. Wang, J. Geng, Y. Du, Proteomic profile analysis and biomarker discovery from mass spectra using independent component analysis combined with uncorrelated linear discriminant analysis, *Chemometr. Intell. Lab. Syst.,* **105** (2011), 207–214. http://dx.doi.org/10.1016/j.chemolab.2011.01.007

19. R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Series B. Methodol.,* **58** (1996), 267–288. http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x

20. J. Liu, R. Li, R. Wu, Feature selection for varying coefficient models with ultrahigh-dimensional covariates, *J. Am. Stat. Assoc.,* **109** (2014), 266–274. http://dx.doi.org/10.1080/01621459.2013.850086

21. L. Zhu, L. Li, R. Li, L. Zhu, Model-free feature screening for ultrahigh-dimensional data, *J. Am. Stat. Assoc.,* **106** (2011), 1464–1475. http://dx.doi.org/10.1198/jasa.2011.tm10563

AIMS Press