AIMS Mathematics, 8(2): 4342–4362.
DOI: 10.3934/math.2023216
Received: 04 September 2022
Revised: 17 November 2022
Accepted: 22 November 2022
Published: 05 December 2022

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

Research article

Group feature screening based on Gini impurity for ultrahigh-dimensional multi-classification

Zhongzheng Wang[1], **Guangming Deng**[1,2,*] **and Haiyun Xu**[3]

[1] College of science, Guilin University of Technology, Guangxi 541000, China
[2] Applied Statistics Institute, Guilin University of Technology, Guangxi 541000, China
[3] School of finance, Jiangxi University of Finance and Economics, Jiangxi 330013, China

* **Correspondence:** Email: dgm@glut.edu.cn.

Abstract: Because the majority of model-free feature screening methods concentrate on individual predictors, they are unable to consider structured predictors, such as grouped variables. In this study, we suggest a model-free and direct extension of the original sure independence screening approach for group screening using Gini impurity for a classification model. Compared to current feature screening approaches, the proposed method performs better in terms of screening efficiency and classification accuracy. It was established that the suggested group screening process exhibits sure screening properties and ranking consistency properties under specific regularity conditions. We used simulation studies to illustrate the limited sample performance of the proposed technique and real data analysis.

Keywords: ultrahigh-dimensional; group feature screening; model-free; Gini impurity; classification model
Mathematics Subject Classification: 62H30, 62R07

1. Introduction

Ultrahigh-dimensional data are commonly available for a wide range of scientific research and applications. Feature screening plays an essential role in ultrahigh-dimensional data, where Fan and Lv [5] first proposed sure independence screening (SIS) in their seminal paper. For linear regressions, they showed that the approach based on Pearson correlation learning possesses a screening property. That is, even if the number of predictors P can grow much faster than the number of observations n with $\log P = O(n^{\alpha})$ for some $\alpha \in (0, \frac{1}{2})$, all relevant predictors can be selected with a probability tending to one [6].

To address ultrahigh-dimensional feature screening in the classification problem, Mai and Zou [11] applied a Kolmogorov filter to ultrahigh-dimensional binary classification. Cui et al. [4] proposed a

screening procedure using empirical conditional distribution functions. The proposed screening methods assume that the types of data are continuous. Assume that the types of data are continuous. For categorical covariates, Huang et al. [8] constructed a model-free discrete feature screening method based on Pearson Chi-square statistics and showed its screening property fulfilling Fan et al. [6] when all covariates were binary. Ni and Fang [12] proposed a model-free feature screening procedure based on information entropy theory for multi-class classification. Ni et al. [13] further proposed a feature screening procedure based on weighting the Adjusted Pearson Chi-square for multi-class classification. Sheng and Wang [17] proposed a new model-free feature screening method based on the classification accuracy of marginal classifiers for ultrahigh-dimensional classification. However, some covariates existed in the groups, especially discrete and categorical covariates that showed microarrays, genomics, brain images and quantitative measurements. A fair number of grouped variable selection methods arise from individual variable selection and yield a sparse solution at the group level, or even at the within-group level. Refer to group LASSO [22], group SCAD [21], group MCP [2], group hierarchical LASSO [23], group bridge [9] and group exponential LASSO [1]. When the regularization parameter is set for non-sparse estimation, some grouped variable selection algorithms may fail to converge, causing non-identifiability problems and near-singularity problems. Even if the algorithm converges in the setting of a large group and small sample n, the estimated coefficients are not likely to be globally optimal solutions. Therefore, to reduce the number of groups before selecting important groups and variables within these groups, there is a need for new screening methods. For ultrahigh-dimensional data with grouping structures, Niu et al. [15] applied working independence in linear models to propose a group-screening approach. Song and Xie [18] further used F-test statistics to construct a group screening approach that improved marginal methods by reducing the burden of multiple testing and aggregating individual effects. With regard to ultrahigh-dimensional group data in the linear model, Qiu and Ahn [16] proposed group sure independence screening (gSIS), group high dimensional ordinary least-squares projector (gHOLP) and group wise adjusted R-squares screening (gAR2). He and Deng [7] applied joint information entropy to screen for important grouped covariates.

In this study, we propose a model-free group feature screening method for ultrahigh-dimensional multi-classification of categorical. Our proposed group screening method is based on the Gini impurity to evaluate the predictive power of grouped covariates. The Gini impurity is a non-purity attribute splitting index, which was proposed by Breiman et al. [3] and has been widely used in decision tree algorithms, such as CART and SPRINT. Regarding categorical covariate screening, we can apply the index of purity gain, which is the same as the information gain [12]. As in Ni and Fang [12], continuous covariates can be sliced using standard normal quantiles. The proposed grouped feature screening procedure is based on the purity gain, which is referred to as GP-SIS. Theoretically, the GP-SIS is rigorously proven to enjoy Fan and Lv [5] proposed a sure screening property that ensures that all important features can be obtained. Practically, as shown by the simulation results, compared with the existing group feature screening method and single covariate feature screening, GP-SIS has a better performance.

The remainder of this paper is organized as follows. Section 2 describes the proposed GP-SIS method in detail. In Section 3, the screening property is established. In Section 4, numerical simulations and an example of real data analysis are presented to assess the performance of the proposed method. Some concluding remarks are given in Section 5, and all proofs are provided in the

Appendix.

2. Group feature screening procedure

We first introduce the Gini impurity and purity gain and then propose a screening procedure based on purity gain.

2.1. Gini index and purity gain

Each grouped covariate can be regarded as a whole. Suppose that Y is a categorical response, and covariate matrix X is a multivariate covariate matrix of $n \times P$ dimension with G grouped covariates, which can be represented in Table 1.

Table 1. Definition of X and Y.

Y	X_{11}	$\cdots$	X_{1p_1}	$\cdots$	X_{g1}	$\cdots$	X_{gp_g}	$\cdots$	X_{G1}	$\cdots$	X_{Gp_G}
Y_1	$x_{1,11}$	$\cdots$	$x_{1,1p_1}$	$\cdots$	$x_{1,g1}$	$\cdots$	x_{1,gp_g}	$\cdots$	$x_{1,G1}$	$\cdots$	x_{1,Gp_G}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
Y_n	$x_{n,11}$	$\cdots$	$x_{n,1p_1}$	$\cdots$	$x_{n,g1}$	$\cdots$	x_{n,gp_g}	$\cdots$	$x_{n,G1}$	$\cdots$	x_{n,Gp_G}

Here, $X_g = (X_{g1}, X_{g2}, \cdots, X_{gp_g})$ represents the g-th group covariate, p_g represents the dimension of the covariates in the g-th group covariates, and $P = \sum_{g=1}^{G} p_g$. To introduce the Gini impurity and purity gain, assume that all the covariate components of the covariate matrix X are classified with J categorizes $\{1, \cdots, J\}$. The values of any element in $X_g \in \{1, \cdots, J\}$, J^{p_g} combinations were formed. J_g represents the last combinations between covariate categories in the g-th group covariate matrix, $J_g = (j_{p_g}, j_{p_g}, \cdots, j_{p_g})$. Here, $\boldsymbol{j}_g = (j_1, \cdots, j_{p_g})$ represents the indicator variable in the combination between covariate categories in the g-th group covariate matrix, and j_1 represents the first covariate category combination.

Let $p_r = P(Y = r)$ represent the probability function of a response variable, $w_{j_g} = w_{(j_1,\cdots,j_{p_g})} = P(X_{g1} = j_1, \cdots, X_{gp_g} = j_{p_g})$ represent the probability function of group covariate, and $p_{j_g r} = p_{(j_1,\cdots,j_{p_g})r} = P(Y = r|X_{g1} = j_1, \cdots, X_{gp_g} = j_{p_g})$ represent the probability function of response variables under the condition of group covariates, where $g \in \{1, \cdots, G\}, (j_1, \cdots, j_{p_g}) \in \{(1, 1, \cdots, 1), (2, 1, \cdots, 1) \cdots, (J, J, \cdots, J)\}, r \in \{1, \cdots, R\}$. The marginal Gini impurity of Y is defined as

$$Gini(Y) = 1 - \sum_{r=1}^{R} p_r^2. \tag{2.1}$$

Conditional Gini impurity is defined as

$$Gini(Y|X_g) = \sum_{\boldsymbol{j}_g = c(1,1,\cdots,1)}^{J_g} w_{j_g}(1 - \sum_{r=1}^{R} p_{j_g r}^2). \tag{2.2}$$

Similar to the information gain, the purity gain is defined as

$$\begin{aligned} GP(Y|X_g) &= Gini(Y) - Gini(Y|X_g) \\ &= 1 - \sum_{r=1}^{R} p_r^2 - \sum_{\boldsymbol{j}_g=c(1,1,\cdots,1)}^{J_g} w_{\boldsymbol{j}_g}(1 - \sum_{r=1}^{R} p_{\boldsymbol{j}_g r}^2). \end{aligned} \tag{2.3}$$

In Eq (2.1), $Gini(Y)$ is non-negative and acquires its maximum $1 - \frac{1}{R}$ if and only if $p_1 = \cdots = p_R = \frac{1}{R}$ by Jensen's inequality. The $Gini(Y|X_g)$ in Eq (2.2) is the conditional Gini impurity of Y given $X_{g1} = j_1, \cdots, X_{gp_g} = j_{p_g}$. Further support can be provided by the following proposition.

Proposition 2.1. *When X_g is a categorical covariable, we obtain $GP(Y|X_g) \geq 0$, and X_g and Y are independent if and only if $GP(Y|X_g) = 0$.*

For continuous X_g, the conditional Gini impurity cannot be directly calculated, and the purity gain by slicing X into several categories. For a fixed integer $J \geq 2$, let $q_{(j)}$ be the j/J-th percentile of X, $j = 1, \cdots, J-1$, $q_{(0)} = -\infty$ and $q_{(J)} = +\infty$. Replacing $w_{\boldsymbol{j}_g}$ and $p_{\boldsymbol{j}_g r}$ in Eq (2.3), respectively, by

$$w_{\boldsymbol{j}_g} = w_{(j_1,\cdots,j_{p_g})} = P(X_{g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \cdots, X_{gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g,(j)}]), \tag{2.4}$$

$$\begin{aligned} p_{\boldsymbol{j}_g r} &= p_{(j_1,\cdots,j_{p_g})r} \\ &= P(Y = r|X_{g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \cdots, X_{gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g,(j)}]). \end{aligned} \tag{2.5}$$

We define conditional Gini impurity based on continuous covariates:

$$Gini_J(Y|X_g) = \sum_{\boldsymbol{j}_g=c(1,1,\cdots,1)}^{J_g} w_{\boldsymbol{j}_g}(1 - \sum_{r=1}^{R} p_{\boldsymbol{j}_g r}^2), \tag{2.6}$$

$$GP_J(Y|X_g) = (1 - \sum_{r=1}^{R} p_r^2) - Gini_J(Y|X_g). \tag{2.7}$$

Proposition 2.2. *When X_g is a continuous covariable, we obtain $GP_J(Y|X_g) \geq 0$, and X_g and Y are independent if and only if $GP_J(Y|X_g) = 0$.*

2.2. Grouped feature screening procedure based on purity gain

First, we select a medium-scale simplified model that can almost fully contain D, where $D = \{g : F(Y|x)$ functionally depends on X_g for some $Y = r\}$, using an adjusted purity gain index for each pair (Y, X_g) as follows:

$$e_g = \frac{[(1 - \sum_{r=1}^{R} p_r^2) - \sum_{\boldsymbol{j}_g=c(1,1,\cdots,1)}^{J_g} w_{\boldsymbol{j}_g}(1 - \sum_{r=1}^{R} p_{\boldsymbol{j}_g r}^2)]}{\log N_g}. \tag{2.8}$$

Here, $p_r = P(Y = r)$,$w_{\boldsymbol{j}_g} = w_{(j_1,\cdots,j_{p_g})} = P(X_{g1} = j_1, \cdots, X_{gp_g} = j_{p_g})$ when X_g is a categorical group, N_g represents the number of group categories of X_g, and $p_{\boldsymbol{j}_g r} = p_{(j_1,\cdots,j_{p_g})r} = P(Y = r|X_{g1} = j_1, \cdots, X_{gp_g} = j_{p_g})$.

When X_g is defined as continuous group covariates, $p_{j_g r} = P(Y = r|X_{g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \cdots, X_{gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g,(j)}])$, where $q_{g1,j}$ represents j/J percentile of $X_{g1} \cdot J_g = J^{P_g}$, and J represents the number of slices applied to X_g. In this case, $N_g = J^{P_g}$.

There may be more categories of group covariates associated with larger purity gain in the original definition of Eq (2.3), regardless of whether the group covariates are important, especially when the number of categories involved in each group covariate is different. Therefore, Ni and Fang [12] used $\log J_k$ to construct an information gain ratio to solve this problem, where each category of X_k is the same. Similarly, when each category of X_g is the same, for Eq (2.8), we apply the $\log N_g$ to build an adjusted purity gain index to address the problem, which is also applied to continuous X_g. However, when each category of X_g is different, $1 - \sum_{j_g=c(1,1,\cdots,1)}^{J_g} w_{j_g}^2$ is defined as an adjustment factor, motivated by the split of X_g into several categories via the Decision Tree algorithm.

For group sample data $\{x_{i,g1}, \cdots, x_{i,gp_g}, y\}$, $i = 1, \cdots, n$, e_g can be easily estimated by

$$\hat{e}_g = \frac{[(1 - \sum_{r=1}^{R} \hat{p}_r^2) - \sum_{j_g=c(1,1,\cdots,1)}^{J_g} \hat{w}_{j_g}(1 - \sum_{r=1}^{R} \hat{p}_{j_g r}^2)]}{\log N_g}. \tag{2.9}$$

When X_g is categorical,

$$\hat{w}_{j_g} = \frac{1}{n} \sum_{i=1}^{n} I\{x_{i,g1} = j_1, \cdots, x_{i,gp_g} = j_{p_g}\}, \tag{2.10}$$

$$\hat{p}_{j_g r} = \frac{\sum_{i=1}^{n} I\{y_i = r, x_{i,g1} = j_1, \cdots, x_{i,gp_g} = j_{p_g}\}}{\sum_{i=1}^{n} I\{x_{i,g1} = j_1, \cdots, x_{i,gp_g} = j_{p_g}\}}. \tag{2.11}$$

When X_g is continuous,

$$\hat{w}_{j_g} = \frac{1}{n} \sum_{i=1}^{n} I\{x_{i,g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \cdots, x_{i,gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g(j)}]\}, \tag{2.12}$$

$$\hat{p}_{j_g r} = \frac{\sum_{i=1}^{n} I\{y_i = r, x_{ig1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \cdots, x_{igp_g} \in (q_{gp_g,(j-1)}, q_{gp_g(j)}]\}}{\sum_{i=1}^{n} I\{x_{ig1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \cdots, x_{igp_g} \in (q_{gp_g,(j-1)}, q_{gp_g(j)}]\}}. \tag{2.13}$$

Here, $q_{g1,j}$ is the j/Jth sample normal percentile of $\{x_{1,g1}, \cdots, x_{n,g1}\}$. In either case, $\hat{p}_r = \frac{1}{n} \sum_{i=1}^{n} I\{y_i = r\}$.

We suggest selecting a sub-model $\hat{D} = \{g : \hat{e}_g \geq cn^{-\tau}, 1 \leq g \leq G\}$, where both c and τ are predetermined thresholds established via condition (C2) in Section 3. In practice, we can choose a model $\hat{D}$= { g : $\hat{e}_g$ is among the top of d largest of all }, where $d = [n/\log n]$.

3. Group feature screening property

In this section, we establish the screening properties of the GP-SIS. Based on the sure independent screening theories proposed by Ni and Fang [12] and He and Deng [7], the following conditions are assumed:

Condition 1 (C1)**.** *There exist two positive constants c_1 and c_2 such that $c_1/R \leq p_r \leq c_2/R$,$c_1 + c_2 \leq R$, $c_1/R \leq p_{j_g r} \leq c_2/R$ and $c_1/N_g \leq w_{j_g} \leq c_2/N_g$ for every $1 \leq g \leq G$, $1 \leq r \leq R$ and $\boldsymbol{j}_g \in \{(1, 1, \cdots, 1), (2, 1, \cdots, 1), \cdots, (J, J, \cdots, J)\}$.*

Condition 2 (C2)**.** *There exists a positive constant* $c > 0$*, and* $0 \le \tau < 1/2$ *such that* $\min_{g\epsilon D} e_g \ge 2cn^{-\tau}$.

Condition 3 (C3)**.** $R = O(n^{\varepsilon}), J = \max_{1\le g\le G} N_g = O(n^{\kappa})$*, where* $\varepsilon \ge 0, \kappa \ge 0$ *and* $2\tau + 2\varepsilon + 2\kappa < 1$.

Condition 4 (C4)**.** *There exists a positive constant* c_3*, such that* $0 < f_g(x|Y = r) < c_3$ *for any* $1 \le r \le R$*, and* x *is in the domain of* X_g*, where* $f_g(x|Y = r)$ *is the Lebesgue density function of* X_g *conditional on* $Y = r$.

Condition 5 (C5)**.** *There exists a positive constant* c_4*, and* $0 \le \rho < 1/2$ *such that* $f_g(x) \ge c_4 n^{-\rho}$ *for any* $1 \le g \le G$ *and* x *in the domain of* X_g*, where* $f_g(x)$ *is the Lebesgue density function of* X_g*. Furthermore,* $f_g(x)$ *was continuous in the domain of* X_g.

Condition 6 (C6)**.** $R = O(n^{\varepsilon})$, $J = \max_{1\le g\le G} N_g = O\left(n^k\right)$*, where* $2\tau + 2\varepsilon + 2\kappa + 2\rho < 1$ *and* $\varepsilon \ge 0, \kappa \ge 0$.

Condition 7 (C7)**.** $\liminf_{p\to\infty}\{\min_{g\epsilon D} e_g - \max_{g\epsilon I} e_g\} \ge \delta$*, where* $\delta > 0$ *is a constant.*

Condition (C1) guarantees that the proportion of each class of variables cannot be either extremely small or extremely large. A similar assumption was made for conditions (C1) in Huang et al. [8] and Cui et al. [4]. According to Fan and Lv [5] and Cui et al. [4], Condition (C2) allows the minimum true signal to disappear to zero in the order of $n^{-\tau}$ as the sample size goes to infinity. According to Ni and Fang [12] and He and Deng [7], Condition (C3) provides for the covariates to diverge with a certain order and number of classes for the response, and Condition (C6) slightly modifies Condition (C3). To ensure that the sample percentiles are close to the true percentiles, Condition (C4) rules out the extreme case in which some X_g places a heavy mass in a small range. Condition (C5) requires $n^{-\rho}$ as the lower bound of the density. Cui et al. [4] and Zhu et al. [24] proposed the ranking consistency property; assuming the inactive covariate subset $I = \{1, \cdots, P\}\backslash D$, then Condition (C7) is established; a similar assumption was also made by Ni and Fang [12] and He and Deng [7].

Theorem 3.1 (Sure screening property)**.** *Under conditions (C1) to (C3), if all the covariates are categorical, we obtain:*

$$P(D \subseteq \hat{D}) \ge 1 - O(pexp - bn^{1-(2\tau+2\varepsilon+2\kappa)} + (\varepsilon + k)\log n),$$

where b denotes a positive constant. If $\log p = O(n^{\alpha})$ and $\alpha < 1 - (2\tau + 2\varepsilon + 2\kappa)$, GP-SIS exhibits a sure screening property.

Theorem 3.2 (Sure screening property)**.** *Under conditions (C4)–(C6), when the covariates comprise continuous and categorical variables, we obtain*

$$P(D \subseteq \hat{D}) \ge 1 - O(pexp - bn^{1-(2\tau+2\varepsilon+2\kappa+2\rho)} + (\varepsilon + \kappa)\log n),$$

where b denotes a positive constant. If $\log p = O(n^{\alpha})$ and $\alpha < 1 - (2\tau + 2\varepsilon + 2\kappa + 2\rho)$, GP-SIS exhibits a sure screening property.

Theorem 3.3 (Ranking consistency property)**.** *Under conditions (C1), (C4), (C5) and (C7), if* $\log \frac{RN_g}{\log n} = O(1)$ *and* $\frac{\max\{\log p, \log n\} R^4 {N_g}^4}{n^{1-2\rho}} = O(1)$*, then* $\liminf_{n\to\infty}\{min_{g\epsilon G}\hat{e}_g - max_{g\epsilon I}\hat{e}_g\} > 0, a.s.$

Theorem 3.3 testifies that the proposed screening index can effectively separate active and inactive covariates at the sample level.

4. Numerical studies

4.1. Simulation results

In this subsection, we conduct four simulation studies to demonstrate the finite sample performance of the group screen methods described in Section 2. We compared GP-SIS with IG-SIS [12] and GIG-SIS [7] in terms of performance using the following evaluation criteria: we ranked the features inside each replication in accordance with each screening criterion and noted the minimum model size (MMS) required to accommodate all of the active features. The $5, 25, 50, 75$, and 95% quantiles of the MMS over 100 replications were used to determine the screening performance. Following Shao and Zhang [20], we denote this as CPa. CPa close to 1 is evidence of sure screening for a procedure. We also consider predictor-specific inclusion proportions, which are denoted as $CP1$, $CP2$, and $CP3$, respectively. These represent the coverage probability, which is a given model size of $[n/logn]$, $2[n/logn]$, and $3[n/logn]$, including the indicators of all active covariates. This allows us to further investigate the active predictors that are easier to predict.

Model 1: categorical covariates and binary response

First, we consider the response variables of different categories. According to Ni and Fang [12] and He and Deng [7], we assume a model in which the response y_i is binary, where $R = 2$, and all covariates are categorical. We consider two distributions for y_i:

(1) Balanced, $p_r = P(y_i = r) = 1/2$;

(2) Unbalanced, $p_r = 2[1 + \frac{R-r}{R-1}]/3R$ with $\max_{1\le r\le R} p_r = 2\min_{1\le r\le R} p_r$.

The true model was defined as $D = \{1, \cdots, 9\}$ with $d_0 = 9$, and the group size was $d_{0G} = 3$. Under the condition of y_i, the latent variable is generated as $z_i = (z_{i,1}, \cdots, z_{i,P})$, where $z_{i,k}$ $N(u_{rk}, 1), 1 \le k \le P$. Subsequently, we construct the active covariates:

(1) If $k > d_0$, then $u_{rk} = 0$;

(2) If $k \le d_0$ and $r = 1$, then $u_{rk} = -0.5$;

(3) If $k \le d_0$ and $r = 2$, then $u_{rk} = 0.5$.

Next, we apply the quantile of the standard normal distribution to generate the covariates. The specific approach is as follows.

(1) When k is an odd number, that is, $x_{i,k} = I(z_{i,k} > z_{(\frac{j}{2})}) + 1$;

(2) When k is an even number, that is, $x_{i,k} = I(z_{i,k} > z_{(\frac{j}{5})}) + 1$;

where αth percentile of the standard normal distribution is $z_{(\alpha)}$.

Thus, among all P covariates, the covariates of the two categories and five categories accounted for half. In this model, we considered $P = 1500$ and $n = 80, 100, 120$.

Table 2 shows the evaluation criteria over 100 simulations for Model 1. The results argue that the proposed GP-SIS works well. When the sample size n increases, GP-SIS is close to $d_{0G} = 3$ in MMS, and both increase to 1 in coverage probability. MMS in an unbalanced response is better than in a balanced response in performance via comparing the responses of different structures. Moreover, GP-SIS is more robust than the other two methods in performance because the fluctuation range in MMS is small.

Table 2. Simulation results for example 1.

	MMS					CP			
Condition	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
	Balanced Y,n=80,p=1500								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	4.0	1.00	1.00	1.00	1.00
IG-SIS	89.8	120.0	162.5	193.0	231.1	0.72	0.79	0.80	0.00
	Balanced Y,n=100,p=1500								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
IG-SIS	17.0	22.0	26.0	31.0	40.1	0.89	0.97	1.00	1.00
	Balanced Y,n=120,p=1500								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
IG-SIS	9.0	10.0	11.0	13.0	15.0	1.00	1.00	1.00	1.00
	UnBalanced Y,n=80,p=1500								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	4.0	5.0	1.00	1.00	1.00	1.00
IG-SIS	130.9	197.0	226.0	282.5	373.4	0.66	0.79	0.84	0.00
	UnBalanced Y,n=100,p=1500								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	4.0	0.89	1.00	1.00	1.00
IG-SIS	11.0	15.0	17.0	20.0	27.0	0.90	0.99	1.00	1.00
	UnBalanced Y,n=120,p=1500								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
IG-SIS	9.0	9.0	10.0	10.0	12.0	1.00	1.00	1.00	1.00

Model 2: categorical covariates and multi-class response

We consider more covariate classifications, and the response y_i is multi-class, where $R = 10$. We consider y_i of the two distributions:

(1) Balanced, $p_r = P(y_i = r) = 1/R$;

(2) Unbalanced, $p_r = 2[1 + \frac{R-r}{R-1}]/3R$ with $\max_{1\le r\le R} p_r = 2\min_{1\le r\le R} p_r$.

The true model was defined as $D = \{1, \cdots, 9\}$ with $d_0 = 9$, and the group size was $d_{0G} = 3$. Condition on y_i, the latent variable is generated as $z_i = (z_{i,1}, \cdots, z_{i,P})$, for covariates X_k, $x_{i,k} = f_k(\varepsilon_{i,k} + \mu_{i,k})$, where $\varepsilon_{i,k} \sim t(4)$ and $f_k(\cdot)$ represents a quantile function of standard normal distribution. We then construct the active covariates by defining $u_{i,k}$:

(1) If $k > d_0$, then $u_{rk} = 0$;

(2) If $\le d_0$, then $u_{rk} = 1.5 \times (-0.9)^r$.

Next, we apply the $f_k(\cdot)$ to generate covariates and consider $P = 2000, n = 100, 150, 200$ in this model. The specific approach is as follows:

(1) For $k \le 400$, $f_k(\varepsilon_{i,k} + \mu_{i.k}) = I(z_{i,k} > z_{(\frac{j}{2})}) + 1$;

(2) For $401 \leq k \leq 800$, $f_k(\varepsilon_{i,k} + \mu_{i.k}) = I(z_{i,k} > z_{(\frac{j}{4})}) + 1$;

(3) For $801 \leq k \leq 1200$, $f_k(\varepsilon_{i,k} + \mu_{i.k}) = I(z_{i,k} > z_{(\frac{j}{6})}) + 1$;

(4) For $1201 \leq k \leq 1600$, $f_k(\varepsilon_{i,k} + \mu_{i.k}) = I(z_{i,k} > z_{(\frac{j}{8})}) + 1$;

(5) For $1601 \leq k$, $f_k(\varepsilon_{i,k} + \mu_{i.k}) = I(z_{i,k} > z_{(\frac{j}{10})}) + 1$.

Thus, among all the P covariates, the covariates of two, four, six, eight, and ten categories accounted for one-fifth each.

Table 3 shows the evaluation criteria over 100 simulations for Model 2. Two methods in performance under Model 1 are worse than Model 2. When the model is more intricate, GP-SIS in performance is better than IG-SIS. Particularly, GP-SIS and GIG-SIS have a slightly small MMS under a small sample size n. When the sample size n increases, GP-SIS is close to $d_{0G} = 3$ in MMS, and both increase to 1 in coverage probability. MMS in an unbalanced response is better than in a balanced response in performance via comparing the responses of different structures. Furthermore, GP-SIS is more robust in performance because the fluctuation range in MMS is small.

Table 3. Simulation results for example 2.

	MMS					CP			
Condition	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
	Balanced Y,n=100,p=2000								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
IG-SIS	9.0	9.0	9.0	10.0	45.1	0.99	0.99	0.99	0.95
	Balanced Y,n=150,p=2000								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
IG-SIS	9.0	9.0	9.0	9.0	9.0	1.00	1.00	1.00	1.00
	Balanced Y,n=200,p=2000								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
IG-SIS	9.0	9.0	9.0	9.0	9.0	1.00	1.00	1.00	1.00
	UnBalanced Y,n=100,p=2000								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
IG-SIS	9.0	9.0	10.0	10.0	12.0	0.66	0.79	0.84	0.00
	UnBalanced Y,n=150,p=2000								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
IG-SIS	9.0	9.0	9.0	9.0	9.0	1.00	1.00	1.00	1.00
	UnBalanced Y,n=200,p=2000								
GP-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
GIG-SIS	3.0	3.0	3.0	3.0	3.0	1.00	1.00	1.00	1.00
IG-SIS	9.0	9.0	9.0	9.0	9.0	1.00	1.00	1.00	1.00

Model 3: continuous and categorical covariates

Finally, among the covariates that are both continuous and categorical, we assume a more complex example, where response y_i is multi-class with $R = 4$. We consider y_i of the two distributions:

(1) Balanced, $p_r = P(y_i = r) = 1/R$;

(2) Unbalanced, $p_r = 2[1 + \frac{R-r}{R-1}]/3R$ with $\max_{1\le r\le R} p_r = 2\min_{1\le r\le R} p_r$.

In this model, we consider $P = 3000, n = 180, 220, 260$. The true model is defined at $D = \{1, 2, 3, 751, 752, 753, 1501, 1502, 1503, 1504, 1505, 1506\}$ with $d_0 = 12$, and the group size is 4. Under condition y_i, the latent variable is generated as $z_i = (z_{i,1}, \cdots, z_{i,p})$. For covariates $X_k, z_{i,k}$ $N(\mu_i, 1), 1 \le k \le P$, where $u_i = (u_{i1}, \ldots, u_{ip})^T$ with $u_{ik} = (-1)^r\theta_{rk}$ when $y_i = r$ and $k \in D$. According to He and Deng [7] and Ni and Fang [12], θ_{rk} is listed in Table 4. $u_{ik} = 0$ when $k \notin D$. To generate X_k:

For $k \le 750$, $x_{ik} = j, if z_{ik} \in (z_{(j-1)/4}, z_{j/4}]$;

For $750 < k \le 1500$, $x_{ik} = j, if z_{ik} \in (z_{(j-1)/10}, z_{j/10}]$;

For $1501 \le k$, $x_{ik} = z_{ik}$.

Table 4. Parameter specification of Model 3.

θ_{rk}	K											
	1	2	3	4	5	6	7	8	9	10	11	12
r=1	0.2	0.8	0.7	0.2	0.2	0.9	0.1	0.1	0.7	0.7	0.3	0.5
r=2	0.9	0.3	0.3	0.7	0.8	0.4	0.7	0.6	0.4	0.4	0.8	0.2
r=3	0.1	0.9	0.9	0.1	0.3	0.1	0.4	0.3	0.6	0.6	0.4	0.7
r=4	0.7	0.2	0.2	0.6	0.7	0.6	0.8	0.9	0.1	0.1	0.8	0.6

Thus, among all the P covariates, the covariates of four categories and ten categories accounted for one- fifth, and the other covariates were continuous. Similarly, there are three in four categories and ten in ten categories, and the active covariates are continuous, accounting for half. For continuous covariates, we applied different slices, $J = 4, 8, 10$. The corresponding approaches were defined as GP-SIS-4, IG-SIS-4, GP-SIS-8, IG-SIS-8, GP-SIS-10 and IG-SIS-10. When the numbers of covariates are grouped, He and Deng [7] proposed a grouped feature screening algorithm by using the joint information entropy to screen some important grouped covariates. We denote these as GIG-SIS-4, GIG-SIS-8 and GIG-SIS-10.

Tables 5 and 6 present the simulation results with over 100 simulations for the balanced and unbalanced cases, respectively. When the sample size n increases, GP-SIS is close to $d_{0G} = 3$ in MMS, and both increase to 1 in coverage probability. The coverage probability of GP-SIS is close to that of GIG-SIS in the five indexes. Therefore, it was proved that GP-SIS has the characteristics of group feature screening. MMS in an unbalanced response is better than in a balanced response in terms of performance by comparing the responses of different structures.

Furthermore, GP-SIS and GIG-SIS are robust in performance, because the fluctuation range in the MMS is small for the two types of responses. When different slices are applied in continuous covariates, GP-SIS and GIG-SIS are better in terms of the five indices of coverage probability and

MMS in performance by comparing the responses of different structures. Therefore, three methods are independent of the number of slices in performance.

Table 5. Simulation results for example 3: balanced Y.

	MMS					CP			
Condition	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
	Balanced Y,n=180,p=3000								
GP-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-4	15.0	22.0	32.5	52.8	92.1	0.94	0.98	0.99	0.88
GP-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-8	13.0	16.0	20.0	31.3	79.1	0.97	0.99	0.99	0.93
GP-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-10	14.0	17.0	22.0	39.3	123.1	0.96	0.99	0.99	0.87
	Balanced Y,n=220,p=3000								
GP-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-4	13.0	14.0	16.0	19.0	26.0	1.00	1.00	1.00	1.00
GP-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-8	13.0	15.0	18.0	21.0	29.0	1.00	1.00	1.00	1.00
GP-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-10	14.0	17.0	22.0	26.0	37.4	0.99	1.00	1.00	1.00
	Balanced Y,n=260,p=3000								
GP-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-4	12.0	12.0	14.0	15.0	18.1	1.00	1.00	1.00	1.00
GP-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-8	12.0	13.0	14.0	17.0	21.1	1.00	1.00	1.00	1.00
GP-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-10	12.0	14.0	16.0	19.0	28.1	1.00	1.00	1.00	1.00

Table 6. Simulation results for example 3: unbalanced Y.

	MMS					CP			
Condition	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
	UnBalanced Y,n=180,p=3000								
GP-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-4	39.0	43.0	46.0	50.3	55.1	0.84	0.97	1.00	1.00
GP-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-8	22.0	26.0	29.0	32.0	35.1	0.93	1.00	1.00	1.00
GP-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-10	21.0	23.0	26.0	29.0	33.0	0.95	1.00	1.00	1.00
	UnBalanced Y,n=220,p=3000								
GP-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-4	16.0	18.0	20.0	22.0	25.0	1.00	1.00	1.00	1.00
GP-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-8	13.0	14.0	16.0	17.0	18.0	1.00	1.00	1.00	1.00
GP-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-10	15.0	17.0	19.0	20.0	22.1	1.00	1.00	1.00	1.00
	UnBalanced Y,n=260,p=3000								
GP-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-4	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-4	12.0	12.0	13.0	13.0	14.0	1.00	1.00	1.00	1.00
GP-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-8	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-8	12.0	12.0	13.0	13.0	14.1	1.00	1.00	1.00	1.00
GP-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
GIG-SIS-10	4.0	4.0	4.0	4.0	4.0	1.00	1.00	1.00	1.00
IG-SIS-10	12.0	13.0	13.0	14.0	15.0	1.00	1.00	1.00	1.00

Model 4. Computational time complexity analysis

Similar to Model 1. However, for the distribution of y_i, we consider balanced data, that is, $P(y_i = r) = 1/2$. The true model was defined as $D = \{1, \ldots, 9\}$, with $d_0 = 9$ and $d_0G = 3$, and the group size was 3. The active and irrelevant covariates were generated in the same way as in Model 1. Similarly, half of the p-dimensional covariates are two-category, while the other half are five-category covariates. Model 4 controls for a constant sample size of 150 and considers a dimensional vector of covariates ranging from 1500 to 10500, with an equal series of 1000 equal differences. The running times of

the five methods were recorded for each experiment, and the median running time in 100 replicate experiments was recorded as the running time index of the five methods. The trends of the five methods will be compared as the dimension of covariates increases to compare the computational complexities of the methods.

The median run times in Table 7 show a linear trend for the three methods as the sample size varies linearly. It can be observed that the running time of GP-SIS is not much different from that of GIG-SIS. In ultrahigh-dimensional feature screening, both GP-SIS and GIG-SIS are robust, but the computational time of GP-SIS is shorter than that of GIG-SIS. For grouped variable feature screening, our method was superior to GIG-SIS.

Table 7. Simulation results for Model 4 (Note: Running time in seconds).

Screening Methods	P									
	1500	2500	3500	4500	5500	6500	7500	8500	9500	10500
GP-SIS	2.636	4.202	6.173	6.617	8.124	9.275	10.705	12.135	17.139	15.757
IG-SIS	3.145	5.127	7.512	7.923	9.696	10.967	12.629	14.296	19.841	18.478
GIG-SIS	3.677	6.283	8.933	9.393	11.532	13.085	15.079	17.103	24.702	22.291

4.2. Real data

In this subsection, we analyze a real data-set from the feature selection database at Arizona State University (http://featureselection.asu.edu/). The lung biological data included 203 samples and 3312 features, which were unbalanced owing the response variable. Every class is $139, 17, 21, 20$, and 6, and the covariates are not only continuous but also have group correlations. We randomly divided the data into two parts, where 90% of the data represented training data and 10% of the data represented the test data. The sample sizes of training data and test data respectively are $n = 182$ and $n = 21$. The dimensions of both the training data and test data were $P = 3312$.

We utilized a ten-fold cross-validation method to assess the performances of various classification algorithms to eliminate the model accuracy issues caused by various training data. Active covariates were chosen by GP-SIS-10, IG-SIS-10 and GIG-SIS-10 based on the training data. We classified them using a variety of techniques, including Support Vector Machine [19], Random Forest (RF), and Decision Tree (DT) [10], using the active covariates chosen based on GP-SIS, IG-SIS and GIG-SIS. G-mean and F-measure are the evaluation indices employed. The performance of feature screening for unbalanced high-dimensional data improves with higher G-means and F-measures [7]. Table 8 shows the G-mean and F-measure for the training data and test data using the three classification techniques GP-SIS-10, IG-SIS-10 and GIG-SIS-10. Among all classification methods, GP-SIS exhibited the best performance, where F-measure of GP-SIS is closer to 1 than those of the other two methods. However, the F-measure (test) of some response's class is a little small, which is close to 0 in all classification methods. In other words, the proposed GP-SIS method performed better.

Table 8. Analysis results for real data example.

	screening method	response				
		1	2	3	4	5
classification method	SVM					
	GP-SIS	0.9986	0.7995	0.8108	0.8131	0.7801
G-mean(train data)	GIG-SIS	0.9979	0.7903	0.8095	0.8111	0.7768
	IG-SIS	0.9992	0.7914	0.7865	0.8084	0.7738
	GP-SIS	0.9941	0.9790	0.9825	0.9973	0.9888
G-mean(test data)	GIG-SIS	0.9954	0.9773	0.9937	1.0000	0.9947
	IG-SIS	0.9959	0.9841	0.9758	1.0000	0.9943
	GP-SIS	0.9762	0.8080	0.8518	0.8576	0.6432
F-measure(train data)	GIG-SIS	0.9635	0.6843	0.7873	0.7942	0.5259
	IG-SIS	0.9480	0.6295	0.5900	0.7255	0.4425
	GP-SIS	0.8946	0.3352	0.4203	0.4828	0.0900
F-measure(test data)	GIG-SIS	0.9208	0.3433	0.5683	0.5446	0.2000
	IG-SIS	0.9116	0.3555	0.3422	0.5502	0.2400
classification method	DT					
	GP-SIS	0.9944	0.7861	0.7976	0.7988	0.7511
G-mean(train data)	GIG-SIS	0.9931	0.7715	0.7829	0.7914	0.7449
	IG-SIS	0.9952	0.7821	0.7777	0.8013	0.7471
	GP-SIS	0.9907	0.9896	0.9849	0.9949	0.9829
G-mean(test data)	GIG-SIS	0.9927	0.9646	0.9831	0.9891	0.9816
	IG-SIS	0.9917	0.9840	0.9751	1.0000	0.9829
	GP-SIS	0.9190	0.5290	0.5988	0.6085	0.0000
F-measure(train data)	GIG-SIS	0.89029	0.3551	0.4698	0.5202	0.0343
	IG-SIS	0.9057	0.4776	0.4432	0.5929	0.0000
	GP-SIS	0.8836	0.4005	0.4233	0.4779	0.0000
F-measure(test data)	GIG-SIS	0.8459	0.1708	0.3441	0.3847	0.0000
	IG-SIS	0.8745	0.3195	0.2819	0.4683	0.0000
classification method	RF					
	GP-SIS	1.0000	0.8099	0.8188	0.8166	0.7848
G-mean(train data)	GIG-SIS	1.0000	0.8099	0.8187	0.8166	0.7847
	IG-SIS	1.0000	0.8045	0.8098	0.8143	0.7818
	GP-SIS	0.9963	0.9819	0.9884	0.9975	0.9834
G-mean(test data)	GIG-SIS	0.9978	0.9598	0.9819	0.9913	0.9859
	IG-SIS	0.9958	0.9816	0.9761	1.0000	0.9975
	GP-SIS	1.0000	1.0000	1.0000	1.0000	1.0000
F-measure(train data)	GIG-SIS	1.0000	1.0000	1.0000	1.0000	1.0000
	IG-SIS	0.9846	0.8772	0.8925	0.9025	0.7340
	GP-SIS	0.9036	0.3533	0.5067	0.5138	0.0000
F-measure(test data)	GIG-SIS	0.8856	0.1300	0.3968	0.4885	0.0286
	IG-SIS	0.9137	0.3605	0.3722	0.5303	0.2567

5. Conclusions

In the data, there were continuous and categorical grouping covariates, and the response was categorical, which is common in practice, but the applicable screening methods are limited. We propose a GP-SIS procedure based on the Gini impurity to effectively screen grouping covariates. GP-SIS has a sure screening property and ranking consistency property, theoretically, and is model-free. When the numbers of categories of all grouping covariates are the same and different, GP-SIS is quite similar to GIG-SIS in performance, which can be shown in the simulation. Practically, as shown by the simulation results, compared with the existing group feature screening method and single covariate feature screening, GP-SIS has a better performance.

Group feature screening reports difficulties based on missing data. In the future, based on the classification model, we intend to propose a new group feature screening method for either the missing variable or response variable.

Acknowledgments

The work was supported by National Natural Science Foundation of China [grant number 71963008]. The authors are grateful to the editor and anonymous referee for their constructive comments that led to significant improvements in the paper.

Conflict of interest

All authors declare no conflicts of interest in this paper.

Availability of data

The lung biological data that support the findings of this study are available from the feature selection database of Arizona State University (http://featureselection.asu.edu/).

References

1. P. Breheny, The group exponential lasso for bi-level variable selection, *Biometrika*, **71** (2015), 731–740. https://doi.org/10.1111/biom.12300
2. P. Breheny, J. Huang, Penalized methods for bi-level variable selection, *Stat. Interface.*, **2** (2009), 369–380. https://doi.org/10.4310/SII.2009.v2.n3.a10
3. L. Breiman, J. Friedman, C. J. Stone, R. A. Olshen, *Classification and regression trees*, Belmont CA: Wadsworth International Group, 1984. https://doi.org/10.1201/9781315139470
4. H. Cui, R. Li, W. Zhong, Model-free feature screening for ultrahigh dimensional discriminant analysis, *J. Am. Stat. Assoc.*, **110** (2015), 630–641. https://doi.org/10.1080/01621459.2014.920256
5. J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, *J. R. Stat. Soc. Ser. B*, **70** (2008), 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

6. J. Fan, R. Samworth, Y. Wu, Ultrahigh dimensional feature selection: beyond the linear model, *J. Mach. Learn. Res.*, **10** (2009), 2013–2038. http://arxiv.org/abs/0812.3201

7. H. He, G. Deng, Grouped feature screening for ultra-high dimensional data for the classification model, *J. Stat. Comput. Simul.*, **92** (2022), 972–997. https://doi.org/10.1080/00949655.2021.1981901

8. D. Huang, R. Li, H. Wang, Feature screening for ultrahigh dimensional categorical data with applications, *J. Bus. Econ. Stat.,* **32** (2014), 237–244. https://doi.org/10.1080/07350015.2013.863158

9. J. Huang, S. Ma, H. Xie, C. Zhang, A group bridge approach for variable selection, *Biometrika*, **96** (2009), 339–355. https://doi.org/10.1093/biomet/asp020

10. B. Lantz, *Machine learning with R: expert techniques for predictive modeling*, 2^{ed}, Birmingha: Packt Publishing, 2019.

11. Q. Mai, H. Zou, The Kolmogorov filter for variable screening in high-dimensional binary classification, *Biometrika*, **100** (2013), 229–234. https://doi.org/10.1093/biomet/ass062

12. L. Ni, F. Fang, Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification, *J. Nonparametr. Stat.*, **28** (2016), 515–530. https://doi.org/10.1080/10485252.2016.1167206

13. L. Ni, F. Fang, F. Wan, Adjusted pearson Chi-Square feature screening for multi-classification with ultrahigh dimensional data, *Metrika*, **80** (2017), 805–828. https://doi.org/10.1007/s00184-017-0629-9

14. L. Ni, *Variable screening methods for ultra-high dimensional categorical covariates*, Shanghai: East China Normal University, 2019.

15. Y. Niu, R. Zhang, J. Liu, H. Li, Group screening for ultra-high-dimensional feature under linear model, *Stat. Theor. Relat. Field.*, **4** (2020), 43–54. https://doi.org/10.1080/24754269.2019.1633763

16. D. Qiu, J. Ahn, Grouped variable screening for ultra-high dimensional data for linear model, *Comput. Stat. Data Anal.*, **144** (2020), 1–11. https://doi.org/10.1016/j.csda.2019.106894

17. Y. Sheng, Q. Wang, Model-free feature screening for ultrahigh dimensional classification, *J. Multivar. Anal.*, **178** (2020), 1–15. https://doi.org/10.1016/j.jmva.2020.104618

18. W. Song, J. Xie, Group feature screening via the F statistic, *Commun. Stat. Simul. Comput.*, **48** (2019), 1921–1931. https://doi.org/10.1080/03610918.2019.1691223

19. J. A. K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, *Neural Process. Lett.*, **9** (1999), 293–300. https://doi.org/10.1023/A:1018628609742

20. X. Shao, J. Zhang, Martingale difference correlation and its use in high-dimensional variable screening, *J. Am. Stat. Assoc.*, **109** (2014), 1302–1318. https://doi.org/10.1080/01621459.2014.887012

21. L. Wang, G. Chen, H. Li, Group SCAD regression analysis for microarray time course gene expression data, *Bioinformatics*, **23** (2007), 1486–1494. https://doi.org/10.1093/bioinformatics/btm125

22. M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, *J. R. Stat. Soc. Ser. B*, **68** (2006), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

23.N. Zhou, J. Zhu, Group variable selection via a hierarchical lasso and its oracle property, *Stat. Interface.*, **3** (2010), 557–574. https://doi.org/10.48550/arXiv.1006.2871

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

Appendix

Proof of Proposition 2.1. To prove Proposition 2.1, we need to define $f(x) = x^2$, proved to be close Ni and Fang [12]. By Jensen's inequality,

$$\begin{aligned}\sum_{i=1}^{p_g}\sum_{j_i=1}^{J} w_{(j_1,\cdots,j_{p_g})}\sum_{r=1}^{R} p^2_{(j_1,\cdots,j_{p_g})r} &= \sum_{r=1}^{R}[\sum_{i=1}^{p_g}\sum_{j_i=1}^{J} w_{(j_1,\cdots,j_{p_g})} f(p_{(j_1,\cdots,j_{p_g})r})] \\ &\geq \sum_{r=1}^{R} f(\sum_{i=1}^{p_g}\sum_{j_i=1}^{J} w_{(j_1,\cdots,j_{p_g})} p_{(j_1,\cdots,j_{p_g})r}) \\ &= \sum_{r=1}^{R} f(\sum_{i=1}^{p_g}\sum_{j_i=1}^{J} P(X_{gi}=j_i)P(Y=r|X_{gi}=j_i)) \\ &= \sum_{r=1}^{R} p_r^2,\end{aligned}$$

and then

$$\begin{aligned}GP &= (1-\sum_{r=1}^{R} p_r^2) - \sum_{j_g}^{J_g} w_{j_g}(1-\sum_{r=1}^{R} p^2_{j_g r}) \\ &= 1-\sum_{r=1}^{R} p_r^2 - \sum_{j_g}^{J_g} w_{j_g} + \sum_{j_g}^{J_g} w_{j_g}\sum_{r=1}^{R} p^2_{j_g r} \\ &= \sum_{j_g}^{J_g} w_{j_g}\sum_{r=1}^{R} p^2_{j_g r} - \sum_{r=1}^{R} p_r^2 \\ &\geq 0.\end{aligned}$$

The above equation holds if and only if $p_{(j_1,\cdots,j_{p_g})r} = p_{(j_1,\cdots,j_{p_g})'r}$, for any $1 \leq r \leq R$, $1 \leq i \leq p_g$ and $1 \leq j_{p_g} \leq j'_{p_g} \leq J$. That is, X_g and Y are independent. □

Proof of Proposition 2.2. From the same proof as Proposition 2.1, we can get that $GP_J(Y|X_g) \geq 0$ holds if and only if $p_{(j_1,\cdots,j_{p_g})r} = p_{(j_1,\cdots,j_{p_g})'r}$. So, when X_g and Y are independent, $GP_J(Y|X_g) = 0$.

$$\begin{aligned}&P(X_{g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \cdots, X_{gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g,(j)}]|Y=r) \\ &= \frac{P(Y=r|X_{g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \cdots, X_{gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g,(j)}])/J}{P(Y=r)} \\ &= \frac{P(Y=r|X_{g1} \in (q_{g1,(j'-1)}, q_{g1,(j')}], \cdots, X_{gp_g} \in (q_{gp_g,(j'-1)}, q_{gp_g,(j')}])/J}{P(Y=r)} \\ &= P(X_{g1} \in (q_{g1,(j'-1)}, q_{g1,(j')}], \cdots, X_{gp_g} \in (q_{gp_g,(j'-1)}, q_{gp_g,(j')}]|Y=r).\end{aligned}$$

By $\sum_{j=1}^{J} P(X_{g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \cdots, X_{gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g,(j)}]) = 1$, we have $P(X_{g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \cdots, X_{gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g,(j)}]) = (1/J)^{p_g}$ and $P(X_{g1} \le q_{g1,j}, \cdots, X_{gp_g} \le q_{gp_g,j} | Y = r) = (j/J)^{p_g}$ if covariates have a similar distribution. □

Lemma 1 (Bernstein inequality)**.** *If $Z_1, \cdots, Z_n$ is an independent random variable with a mean value of 0, and bounded supporter is $[-M, M]$, then we have the inequality: $P(|\sum_{i=1}^{n} Z_i| > t) \le 2exp\{-\frac{t^2}{2(v+\frac{Mt}{3})}\}$ where $v \ge Var(\sum_{i=1}^{n} Z_i)$.*

Lemma 2. *For discrete group covariates X_g and discrete response Y, we have the following three inequalities:*

(a) $P(|\hat{p}_r - p_r| > t) \le 2\exp\{-\frac{6nt^2}{3+4t}\}$;

(b) $P(|\hat{w}_{j_g} - w_{j_g}| > t) \le 2\exp\{-\frac{6nt^2}{3+4t}\}$;

(c) $P(|\hat{p}_{j_g r} - p_{j_g r}| > t) \le 2\exp\{-\frac{6nt^2}{3+4t}\}$.

Proof of Lemma 2. Three inequalities are similar in the proofs, where inequality (a) and inequality (b), respectively, have been given at Ni [14] and He and Deng [7]. The following is the proof of inequality (c).

$$\hat{p}_{j_g r} = \frac{\sum_{i=1}^{n} I\{y_i = r, x_{i,g1} = j_1, \cdots, x_{i,gp_g} = j_{p_g}\}}{\sum_{i=1}^{n} I\{x_{i,g1} = j_1, \cdots, x_{i,gp_g} = j_{p_g}\}}.$$

The expectation of $\hat{p}_{j_g r}$ is

$$\begin{aligned} E(\hat{p}_{j_g r}) &= E(\frac{\sum_{i=1}^{n} I\{y_i = r, x_{i,g1} = j_1, \cdots, x_{i,gp_g} = j_{p_g}\}}{\sum_{i=1}^{n} I\{x_{i,g1} = j_1, \cdots, x_{i,gp_g} = j_{p_g}\}}) \\ &= E(\frac{I\{y_i = r, x_{i,g1} = j_1, \cdots, x_{i,gp_g} = j_{p_g}\}}{I\{x_{i,g1} = j_1, \cdots, x_{i,gp_g} = j_{p_g}\}}) = p_{j_g r}. \end{aligned}$$

Let $Z_i = I\{y_i = r | X_{ig1} = j_1, \cdots, X_{igp_g} = j_{p_g}\} - p_{j_g r}, Var(\sum_{i=1}^{n} Z_i) = nVar(Z_i) = np_{j_g r}(1 - p_{j_g r}) \le \frac{n}{4}$ be known, and then

$$\begin{aligned} P(|\hat{p}_{j_g r} - p_{j_g r}| > t) &= P(|n^{-1}\sum_{i=1}^{n} Z_i| > t) = P(|\sum_{i=1}^{n} Z_i| > nt) \\ &\le 2\exp\{-\frac{n^2t^2}{2\left(\frac{n}{4} + \frac{nt}{3}\right)}\} \le 2\exp\{-\frac{6nt^2}{3+4t}\}. \end{aligned}$$

According to the Bernstein inequality, the formula is held. □

Lemma 3. *With regard to discrete group covariates X_g and discrete response Y, for any $0 < \varepsilon < 1$, under condition (C1), we have $P(|\hat{e}_g - e_g| > 2\varepsilon) \le O(RJ^3)\exp\{-c_5\frac{n\varepsilon^2}{R^2J^6}\}$, where c_5 represents a positive constant.*

Proof of Lemma 3. By e_g and $\hat{e}_g$ in Section 2.2, we have

$$
\begin{aligned}
&\log N_g(\hat{e}_g - e_g) \\
&= [(1 - \sum_{r=1}^{R} \hat{p}_r^2) - \sum_{j_g}^{J_g} \hat{w}_{j_g}(1 - \sum_{r=1}^{R} \hat{p}_{j_g r}^2)] - [(1 - \sum_{r=1}^{R} p_r^2) - \sum_{j_g}^{J_g} w_{j_g}(1 - \sum_{r=1}^{R} p_{j_g r}^2)] \\
&= (\sum_{r=1}^{R} p_r^2 - \sum_{r=1}^{R} \hat{p}_r^2) + (\sum_{j_g}^{J_g} w_{j_g} - \sum_{j_g}^{J_g} \hat{w}_{j_g}) + (\sum_{j_g}^{J_g} \hat{w}_{j_g} \sum_{r=1}^{R} \hat{p}_{j_g r}^2 - \sum_{j_g}^{J_g} w_{j_g} \sum_{r=1}^{R} p_{j_g r}^2) \\
&= \sum_{r=1}^{R} (p_r^2 - \hat{p}_r^2) + \sum_{j_g}^{J_g} (w_{j_g} - \hat{w}_{j_g}) + \sum_{j_g}^{J_g} \sum_{r=1}^{R} (\hat{w}_{j_g} \hat{p}_{j_g r}^2 - w_{j_g} p_{j_g r}^2) \\
&= \sum_{r=1}^{R} (p_r - \hat{p}_r)(p_r + \hat{p}_r) + \sum_{j_g}^{J_g} (w_{j_g} - \hat{w}_{j_g}) \\
&+ \sum_{j_g}^{J_g} \sum_{r=1}^{R} [(\hat{w}_{j_g} \hat{p}_{j_g r} + w_{j_g} p_{j_g r})(\hat{p}_{j_g r} - p_{j_g r}) + \hat{p}_{j_g r} p_{j_g r} (\hat{w}_{j_g} - w_{j_g})] \\
&= I_1 + I_2 + I_3.
\end{aligned}
$$

Since $\log J \geq \log 2 \geq 0.5$, we have

$$
P(|\hat{e}_g - e_g| > \varepsilon) \leq P(|I_1| > \frac{\varepsilon}{3}) + P(|I_2| > \frac{\varepsilon}{3}) + P(|I_3| > \frac{\varepsilon}{3}).
$$

For I_1, we have

$$
\begin{aligned}
P(|I_1| > \frac{\varepsilon}{3}) &\leq \sum_{r=1}^{R} P(|(p_r - \hat{p}_r)(p_r + \hat{p}_r)| > \frac{\varepsilon}{3}) \\
&\leq \sum_{r=1}^{R} P(|(p_r - \hat{p}_r)| > \frac{c_1 \varepsilon}{3RJ^3}) \\
&\leq RJ^3 2exp\{-\frac{6n(\frac{c_1 \varepsilon}{3RJ^3})^2}{3 + 4(\frac{c_1 \varepsilon}{3RJ^3})}\}.
\end{aligned}
$$

For I_2, we have

$$
\begin{aligned}
P(|I_2| > \frac{\varepsilon}{3}) &\leq \sum_{j_g}^{J_g} P(|\hat{w}_{j_g} - w_{j_g}| > \frac{c_1 \varepsilon}{3J^3}) \\
&\leq J^3 2exp\{-\frac{6n(\frac{c_1 \varepsilon}{3J^3})^2}{3 + 4(\frac{c_1 \varepsilon}{3J^3})}\}.
\end{aligned}
$$

For I_3, we have

$$\begin{aligned} I_3 &= \sum_{j_g}^{J_g}\sum_{r=1}^{R}[(\hat{w}_{j_g}\hat{p}_{j_g r} + w_{j_g}p_{j_g r})(\hat{p}_{j_g r} - p_{j_g r}) + \hat{p}_{j_g r}p_{j_g r}(\hat{w}_{j_g} - w_{j_g})] \\ &= \sum_{j_g}^{J_g}\sum_{r=1}^{R}[(\hat{w}_{j_g}\hat{p}_{j_g r} + w_{j_g}p_{j_g r})(\hat{p}_{j_g r} - p_{j_g r})] + \sum_{j_g}^{J_g}\sum_{r=1}^{R}\hat{p}_{j_g r}p_{j_g r}(\hat{w}_{j_g} - w_{j_g}) \\ &:= I_{31} + I_{32}. \end{aligned}$$

For I_{31} and I_{32}, we have

$$P(|I_3| > \frac{\varepsilon}{3}) \le P(|I_{31}| > \frac{\varepsilon}{6}) + P(|I_{32}| > \frac{\varepsilon}{6})$$

$$\begin{aligned} P(|I_{31}| > \frac{\varepsilon}{6}) &\le \sum_{j_g}^{J_g}\sum_{r=1}^{R} P(|(\hat{w}_{j_g}\hat{p}_{j_g r} + w_{j_g}p_{j_g r})(\hat{p}_{j_g r} - p_{j_g r})| > \frac{\varepsilon}{6}) \\ &\le \sum_{j_g}^{J_g}\sum_{r=1}^{R} p(|\hat{p}_{j_g r} - p_{j_g r}| > \frac{c_1\varepsilon}{6RJ^3}) \\ &\le RJ^3 2\exp\{-\frac{6n(\frac{c_1\varepsilon}{6RJ^3})^2}{3 + 4(\frac{c_1\varepsilon}{6RJ^3})}\} \end{aligned}$$

$$\begin{aligned} P(|I_{32}| > \frac{\varepsilon}{6}) &\le \sum_{j_g}^{J_g}\sum_{r=1}^{R} P(|\hat{p}_{j_g r}p_{j_g r}(\hat{w}_{j_g} - w_{j_g})| > \frac{\varepsilon}{6}) \\ &\le \sum_{j_g}^{J_g}\sum_{r=1}^{R} P(|\hat{w}_{j_g} - w_{j_g}| > \frac{c_1\varepsilon}{6RJ^3}) \\ &\le RJ^3 2exp\{-\frac{6n(\frac{c_1\varepsilon}{6RJ^3})^2}{3 + 4(\frac{c_1\varepsilon}{6RJ^3})}\}. \end{aligned}$$

In a word, we have the inequality

$$P(|\hat{e}_g - e_g| > 2\varepsilon) \le O(RJ^3)\exp\{-c_5\frac{n\varepsilon^2}{R^2J^6}\},$$

where c_5 represents a positive constant. □

Proof of Theorem 3.1. By Conditions (C1) to (C3) and Lemma 3, we can get

$$\begin{aligned} P(D \subseteq \hat{D}) &\ge P(|\hat{e}_g - e_g| \le cn^{-\tau}, \forall g \epsilon D) \\ &\ge P(\max_{1\le g\le G}|\hat{e}_g - e_g| \le cn^{-\tau}) \\ &\ge 1 - \sum_{g=1}^{G} P(\max_{1\le g\le G}|\hat{e}_g - e_g| > cn^{-\tau}) \\ &\ge 1 - O(RJ^3)p\exp -c_5\frac{c^2n^{1-2\tau}}{R^2J^6} \\ &\ge 1 - O(pexp - bn^{1-2\tau-2\varepsilon-2\kappa} + (\varepsilon+\kappa)\log n), \end{aligned}$$

where b is a positive constant. □

Lemma 4 (Lemma A.5 [7]). *Under (C1), (C4) and (C5),mfor any* $0 < \varepsilon < 1$, *so for continuous* X_g, *we have* $P(|\hat{e}_g - e_g| > 2\varepsilon) \leq O(RN_g)\exp\{-c_6\frac{n^{1-2\rho}\varepsilon^2}{R^4N_g^4}\}$, *there exists a positive constant* c_6.

Proof of Theorem 3.2. According to Lemma 4, the proof of Theorem 3.2 is the same as Theorem 3.1 and hence is omitted. □

Proof of Theorem 3.3. According to Lemma 3 and 4 and under Conditions (C1), (C4), (C5) and (C7), we get

$$\begin{aligned}
&P(min_{g\in D}\hat{e}_g - max_{g\in I}\hat{e}_g < \frac{\delta}{2})\\
&\leq P((min_{g\in D}\hat{e}_g - max_{g\in I}\hat{e}_g) - (min_{g\in D}e_g - max_{g\in I}e_g) < -\frac{\delta}{2})\\
&\leq P(|(min_{g\in D}\hat{e}_g - max_{g\in I}\hat{e}_g) - (min_{g\in D}e_g - max_{g\in I}e_g)| > \frac{\delta}{2})\\
&\leq P(max_{1\leq g\leq G}|\hat{e}_g - e_g| > \frac{\delta}{4}) \leq O(RJ_g)p\exp\{-c_7\frac{n^{1-2\rho}}{R^4J_g^4}\}\\
&= O(\exp\{\log RN_g + \log p - c_7\frac{n^{1-2\rho}}{R^4N_g^4}\}),
\end{aligned}$$

where $c_7 = \min\{c_5, c_6\}(\frac{\delta^2}{4})$. Since $\frac{\log(RN_g)}{\log n} = O(1)$, there exists a positive constant c_8 such that $\log(RN_g) \leq c_8\log n$. Also, $\frac{\max\{\log p,\log n\}R^4{N_g}^4}{n^{1-2\rho}} = O(1)$ implies that $\log p \leq \frac{1}{2}c_7\frac{n^{1-2\rho}}{R^4N_g^4}$ and $\frac{1}{2}c_7\frac{n^{1-2\rho}}{R^4N_g^4} \geq (c_8+2)\log n$ for large n. Next, there exists a constant n_0, and we can get $\sum_{n=n_0}^{\infty}\exp\{\log RN_g + \log p - c_7\frac{n^{1-2\rho}}{R^4N_g^4}\} \leq \sum_{n=n_0}^{\infty}\exp\{c_8\log n - \frac{1}{2}c_7\frac{n^{1-2\rho}}{R^4N_g^4}\} \leq \sum_{n=n_0}^{\infty}\exp\{c_8\log n - (c_8+2)\log n\} = \sum_{n=n_0}^{\infty}n^{-2} < \infty$. According to Ni and Fang [12] and by the Borel Cantelli Lemma, we can get $\liminf_{n\to\infty}\{min_{g\epsilon G}\hat{e}_g - max_{g\epsilon I}\hat{e}_g\} \geq \frac{\delta}{2} > 0, a.s.$ □

AIMS Press