Hindawi
Journal of Mathematics
Volume 2022, Article ID 1600986, 15 pages
https://doi.org/10.1155/2022/1600986

Research Article

Group Feature Screening Based on Information Gain Ratio for Ultrahigh-Dimensional Data

Zhongzheng Wang,[1] **Guangming Deng**,[1,2] **and Jianqi Yu**[1]

[1]*College of Science, Guilin University of Technology, Guilin 541000, China*
[2]*Applied Statistics Institute, Guilin University of Technology, Guilin 541000, China*

Correspondence should be addressed to Guangming Deng; dgm@glut.edu.cn

Received 10 July 2022; Revised 14 September 2022; Accepted 14 October 2022; Published 2 December 2022

Academic Editor: Qiang Wu

Most model-free feature screening approaches focus on the -individual predictor; therefore, they are not able to incorporate structured predictors like grouped variables. In this article, we propose a group screening procedure via the information gain ratio for a classification model, which is a direct extension of the original sure independence screening procedure and also model-free. The proposed method yields a better screening performance and classification accuracy. It is demonstrated that the proposed group screening method possesses the sure screening property and ranking consistency properties under certain regularity conditions. Through simulation studies and real-world data analysis, we demonstrate the proposed method with the finite sample performance.

1. Introduction

Ultrahigh-dimensional data are commonly available in a wide range of scientific research and applications. Feature screening plays an essential role in the ultrahigh-dimensional data, where Fan and Lv [1] first proposed the sure independence screening (SIS) in the seminal paper. They showed that the method based on Pearson correlation learning possesses a sure screening property for linear regressions. All relevant predictors can be selected with probability tending to one 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 $\alpha \in (0, 1/2)$, as discussed by Fan et al. [2].

To address the ultrahigh-dimensional feature screening in classification problem, the Kolmogorov filter was utilized for ultrahigh-dimensional binary classification by Mai and Zou [3]. Cui et al. [4] utilized empirical conditional distribution functions in the fused mean-variance-based screening approach. The assumption that the data are continuous is made by all of the above classification feature screening approaches. Huang et al.'s study of categorical covariates [5] constructed a model-free discrete feature screening method with the help of the Pearson chi-square statistics. The method's sure screening property satisfied Fan and Lv [1] when all of the covariates were binary. For multiclass classification, Ni et al. [6] further added weighting-based adjusted Pearson chi-square feature screening. Information entropy theory was used by Ni and Fang [7] to screen model-free features for ultrahigh-dimensional multiclass classification. However, some types of covariates, particularly categorical and discrete covariates, are grouped, which are frequently represented by microarrays, genomics, quantitative measures, and brain imaging. A good number of methods for selecting grouped variables originate from selecting individual variables and produce a sparse solution at the group level or even within-group level. See, for example, group Lasso [8], group SCAD [9], group MCP [10], group hierarchical Lasso [11], group bridge [12], group exponential Lasso [13], etc. Some grouped variable selection methods may not converge when the number of groups G grows much faster than the sample size n. This is especially true when the regularization parameter is set for a non-sparse estimation, which frequently causes non-identifiability or near-singularity issues. The estimated coefficients may not be globally optimal solutions even if

the algorithm converges in the case of "large G, small n." We believe that new screening techniques that can reduce the number of groups before selecting important groups and variables within these groups are required because of these factors. Niu et al. [14] looked at data with grouping structures in ultrahigh-dimensional and suggested a group screening method based on working independence in linear models. Song and Xie's [15] group screening method made use of the F-test statistic, which improved marginal methods by reducing the burden of multiple tests and aggregating individual effects. For ultrahigh-dimensional data for a linear model, Qiu and Ahn [16] suggested group sure independence screening (gSIS), group-wise adjusted R-squares screening (gAR2), and group high dimensional ordinary least-squares projector (gHOLP). He and Deng [17] applied the joint information entropy to screen some important grouped covariates.

In this paper, we propose a new approach named Group Information Gain Ratio Sure Independence Screening (GIGR-SIS) for grouped feature screening. It is based on information gain ratio, as a direct extension of the original sure independence screening procedure, which is effective in this scenario. The proposed method selects the covariate groups whose information gain ratio are greater than a threshold value, where the threshold value is defined by a given number. Similar to Ni and Fang [7], continuous covariates are sliced using the standard normal quantile. Employing information gain ratio to assess the importance of grouped covariates, we show that GIGR-SIS possesses the sure screening property under certain regularity conditions.

This paper is organized as follows. Section 2 describes the proposed GIGR-SIS method in detail. Then we establish its sure screening property. In Section 3, we assess the performance of our method via numerical examples, including simulations and a real data application. Some concluding remarks are given in Section 4, and all the proofs are given in the Appendix.

2. Theory and Method

We first introduce entropy and information gain ratio, and then propose the screening procedure based on information gain ratio. After that, we establish the property of group feature screening.

2.1. Information Entropy and Information Gain Ratio. For group covariates, each group of covariates can be considered as a whole. Suppose Y be a categorical response with R classes $\{1, \cdots, R\}$ and covariate matrix X be a multivariate covariate matrix of dimension $n \times p$ with G-grouped covariates which can be expressed as $X = \{X_1, \ldots, X_G\}$, where $X_g = (X_{g1}, X_{g2}, \ldots, X_{gp_g})$ represents the g-th group covariate and p_g represents the dimension of the covariates belonging to the g-th group. To introduce the concept of entropy and information gain ratio, assume that all the covariate components of covariate matrix X are classified with J categories $\{1, \ldots, J\}$. The value of any element in $X_g \in \{1, \ldots, J\}$, J^{p_g} combinations are formed. J_g represents the last of the combinations between covariate categories in the g-th group covariate matrix, $J_g = (j_{p_g}, j_{p_g}, \ldots, j_{p_g})$, where $j_g = (j_1, \ldots, j_{p_g})$ represents the indicator variable in the combination between covariate categories in the g-th group covariate matrix, and j_1 is the first covariate category combination.

Let $p_r = P(Y = r)$ represent the probability function of response variable, $w_{j_g} = w_{(j_1,\ldots,j_{p_g})} = P(X_{g1} = j_1, \ldots, X_{gp_g} = j_{p_g})$ represent the probability function of group covariates, and $p_{j_g r} = p_{(j_1,\ldots,j_{p_g})r} = P(Y = r | X_{g1} = j_1, \ldots, X_{gp_g} = j_{p_g})$ represent the probability function of response variables under the condition of group covariates, where $g \in \{1, \cdots, G\}$, $(j_1, \ldots, j_{p_g}) \in \{(1, 1, \ldots, 1), (2, 1, \ldots, 1), \ldots, (J, J, \ldots, J)\}$, and $r \in \{1, \ldots, R\}$. Let $0 \times \log 0 = 0$. The marginal entropy of Y and marginal entropy of X, respectively, are defined as

$$H(Y) = -\sum_{r=1}^{R} p_r \log p_r,$$
$$H(X_g) = -\sum_{j_g = c(1,1,\ldots,1)}^{J_g} w_{j_g} \log w_{j_g}. \tag{1}$$

Conditional entropy is defined as

$$En(Y|X_g) = -\sum_{j_g = c(1,1,\ldots,1)}^{J_g} w_{j_g} \sum_{r=1}^{R} p_{j_g r} \log p_{j_g r}. \tag{2}$$

The information gain is defined as

$$\begin{aligned} IG(Y|X_g) &= H(Y) - En(Y|X_g) \\ &= \sum_{j_g = c(1,1,\ldots,1)}^{J_g} w_{j_g} \sum_{r=1}^{R} p_{j_g r} \log p_{j_g r} - \sum_{r=1}^{R} p_r \log p_r. \end{aligned} \tag{3}$$

The information gain ratio is defined as

$$\begin{aligned} IGR(Y|X_g) &= \frac{IG(Y|X_g)}{H(X_g)} = \frac{H(Y) - En(Y|X_g)}{H(X_g)} \\ &= \frac{\sum_{j_g = c(1,1,\ldots,1)}^{J_g} w_{j_g} \sum_{r=1}^{R} p_{j_g r} \log p_{j_g r} - \sum_{r=1}^{R} p_r \log p}{-\sum_{j_g = c(1,1,\ldots,1)}^{J_g} w_{j_g} \log w_{j_g}}. \end{aligned} \tag{4}$$

In equation (3), $H(Y)$ is non-negative and achieves its maximum $\log R$ if and only if $p_1 = \cdots = p_R = 1/R$ by Jensen's inequality, and the term $-\sum_{j_g = c(1,1,\ldots,1)}^{J_g} w_{j_g} \sum_{r=1}^{R} p_{j_g r} \log p_{j_g r}$ in equation (2) is the conditional entropy of Y given $X_{g1} = j_1, \ldots, X_{gp_g} = j_{p_g}$. According to Ni and Fang [7] and He and Deng [17], further support is provided by the following proposition.

Proposition 1. *We have $IG(Y|X_g) \geq 0$ when X is a categorical variable, and if and only if $IG(Y|X_g) = 0$, Y and X_g are statistically independent.*

The conditional entropy can only be determined by dividing X into multiple categories when X is continuous. For a fixed integer $J \geq 2$, let $q_{(j)}$ be the $j/\mathbf{J}$-th percentile of X, $j = 1, \ldots, J-1$, $q_{(0)} = -\infty$, and $q_{(J)} = +\infty$, replacing w_{j_g} in equation (2) by $w_{j_g} = w_{(j_1,\ldots,j_{p_g})} = P(X_{g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \ldots, X_{gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g,(j)}])$ and $p_{j_g r} = p_{(j_1,\ldots,j_{p_g})r} = P(Y = r | X_{g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \ldots, X_{gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g,(j)}])$.

Based on continuous covariates, we define conditional entropy as follows:

$$\begin{aligned} En_J\left(Y|X_g\right) &= -\sum_{j_g=c(1,1,\ldots,1)}^{J_g} w_{j_g} \sum_{r=1}^{R} p_{j_g r} \log\, p_{j_g r}, \\ IG_J\left(Y|X_g\right) &= \sum_{j_g=c(1,1,\ldots,1)}^{J_g} w_{j_g} \sum_{r=1}^{R} p_{j_g r} \log\, p_{j_g r} - \sum_{r=1}^{R} p_r \log\, p_r. \end{aligned} \tag{5}$$

Proposition 2. *We have $IG_J(Y|X_g) \geq 0$ when X is a continuous variable, and if and only if $IG_J(Y|X_g) = 0$, Y and X_g are statistically independent.*

2.2. Grouped Feature Screening Procedure Based on Information Gain Ratio. Let $X = (X_1, \ldots, X_G)^T$ be the covariate matrix and Y be a categorical response with R classes $\{1, \ldots, R\}$, $p = \sum_{g=1}^{G} p_g$, where $p \gg n$ and n is the sample size. $D = \{g: F(Y|x)$ functionally depends on X_g for some $Y = r\}$ denotes the active covariate subset.

A changed information gain ratio is used to measure the relationship between Y and X_g because we must select a simplified model with a medium scale that can almost include D. The following is the information gain ratio for each pair (Y, X_g):

$$e_g = \left|IGR\left(Y|X_g\right)\right| = \left|IG\left(Y|X_g\right)/H\left(X_g\right)\right| = \left|H(Y) - En\left(Y|X_g\right)/H\left(X_g\right)\right| = \frac{\left(\sum_{j_g=c(1,1,\ldots,1)}^{J_g} w_{j_g} \sum_{r=1}^{R} p_{j_g r} \log\, p_{j_g r} - \sum_{r=1}^{R} p_r \log\, p_r\right)}{\sum_{j_g=c(1,1,\ldots,1)}^{J_g} w_{j_g} \log\, w_{j_g}}, \tag{6}$$

where $P_r = P(Y = r)$ and $w_{j_g} = w_{(j_1,\ldots,j_{p_g})} = P(X_{g1} = j_1, \ldots, X_{gp_g} = j_{p_g})$ when X_g is a categorical group, and $P_{j_g r} = p_{(j_1,\ldots,j_{p_g})r} = P(Y = r | X_{g1} = j_1, \ldots, X_{gp_g} = j_{p_g})$.

Then $p_{j_g r} = P(Y = r | X_{g1} \in (q_{g1,(j-1)}, q_{g1,(j)}], \ldots, X_{gp_g} \in (q_{gp_g,(j-1)}, q_{gp_g,(j)}])$ when X_g is the continuous group, and $q_{g1,j}$ is the $j/\mathbf{J}$ percentile of X_{g1}, and $J_g = J^{p_g}$. In X_g, J is the number of slices that are applied to each X.

When using information gain to select features, it tends to select features with larger values. The essence of the information gain ratio is to multiply a penalty parameter based on information gain. When the number of features is large, the penalty parameter is small; when the number of features is small, the penalty parameter is large. The information gain ratio makes up for the deficiency that information gain tends to select features with large values.

With the sample group data $\left\{y, x_{i,g1}, \ldots, x_{i,gp_g}\right\}$, $i = 1, \ldots, n$, e_g can be easily estimated by

$$\widehat{e}_g = \frac{\left(\sum_{j_g=c(1,1,\ldots,1)}^{J_g} \widehat{w}_{j_g} \sum_{r=1}^{R} \widehat{p}_{j_g r} \log\, \widehat{p}_{j_g r} - \sum_{r=1}^{R} \widehat{p}_r \log\, \widehat{p}_r\right)}{\sum_{j_g=c(1,1,\ldots,1)}^{J_g} \widehat{w}_{j_g} \log\, \widehat{w}_{j_g}}. \tag{7}$$

When X_g is categorical,

$$\begin{aligned} \widehat{w}_{j_g} &= \frac{1}{n} \sum_{i=1}^{n} I\left\{x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{p_g}\right\}, \\ \widehat{p}_{j_g r} &= \frac{\sum_{i=1}^{n} I\left\{y_i = r, x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{p_g}\right\}}{\sum_{i=1}^{n} I\left\{x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{p_g}\right\}}. \end{aligned} \tag{8}$$

When X_g is continuous,

$$\begin{aligned} \widehat{w}_{j_g} &= \frac{1}{n} \sum_{i=1}^{n} I\left\{x_{i,g1} \in \left(q_{g1,(j-1)}, q_{g1,(j)}\right], \ldots, x_{i,gp_g} \in \left(q_{gp_g,(j-1)}, q_{gp_g(j)}\right]\right\}, \\ \widehat{p}_{j_g r} &= \frac{\sum_{i=1}^{n} I\left\{x_{i,g1} \in \left(q_{g1,(j-1)}, q_{g1,(j)}\right], \ldots, x_{i,gp_g} \in \left(q_{gp_g,(j-1)}, q_{gp_g(j)}\right]\right\}}{\sum_{i=1}^{n} I\left\{x_{i,g1} \in \left(q_{g1,(j-1)}, q_{g1,(j)}\right], \ldots, x_{i,gp_g} \in \left(q_{gp_g,(j-1)}, q_{gp_g(j)}\right]\right\}}, \end{aligned} \tag{9}$$

where $q_{g1,(j)}$ is the j/Jth sample normal percentile of $\{x_{1,g1}, \ldots, x_{n,g1}\}$. In either case, $\widehat{p}_r = (1/n)\sum_{i=1}^{n} I\{y_i = r\}$.

We suggest going with a submodel $\widehat{D} = \{g: \widehat{e}_g \geq cn^{-\tau}, 1 \leq g \leq G\}$, where condition (C2) in subsection 2.3 specifies predetermined thresholds for the values c and τ. In actuality, we can select a model:

$$\widehat{D} = \{g: \widehat{e}_g \text{ is among the top of } d \text{ largest of all}\} \text{ where } d = \left[\frac{n}{\log n}\right]. \tag{10}$$

2.3. Group Feature Screening Property. In this subsection, we establish the sure screening property of GIGR-SIS. Sure independence screening (SIS), which provided a statistical theoretical foundation for ultrahigh-dimensional feature screening techniques, was first proposed by Fan and Lv [1]. IG-SIS [7] and GIG-SIS [17] were testified to be satisfied with sure independence screening. We assume the following conditions based on these theories.

(C1): There are two positive constants c_1 and c_2 such that $c_1/R \leq p_r c_2/R$, $c_1 + c_2 \leq R$, $c_1/R \leq p_{j_g r} \leq c_2/R$, and $c_1/J_g w_{j_g} \leq c_2/J_g$ for every $1 \leq r \leq R$, $1 \leq g \leq G$, and $\mathbf{j}_g \in \{(1, 1, \ldots, 1), (2, 1, \ldots, 1), \ldots, (J, J, \ldots, J)\}$.

(C2): There is a positive constant $c > 0$ and $0 \leq \tau < 1/2$ such that $\min_{k \varepsilon D} e_g \geq 2cn^{-\tau}$.

(C3): $R = O(n^{\varepsilon})$, $J = \max_{1 \leq g \leq G} J_g = O(n^{\kappa})$, where $\varepsilon \geq 0$, $\kappa \geq 0$ and $2\tau + 4\varepsilon + 4\kappa < 1$.

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

(C5): There is a positive constant c_4 and $0 \leq \rho < 1/2$ such that $f_g(x) \geq c_4 n^{-\rho}$ for any $1 \leq g \leq G$ and x in the domain of X_g, where $f_g(x|Y = r)$ is the Lebesgue density function of X_g. Furthermore, $f_g(x)$ is continuous in the domain of X_g.

(C6): $R = O(n^{\varepsilon})$, $J = \max_{1 \leq g \leq G} J_g = O(n^{\kappa})$, where $2\tau + 4\varepsilon + 4\kappa + 2\rho < 1$ and $\varepsilon \geq 0$, $\kappa \geq 0$.

(C7): $\liminf_{p \longrightarrow \infty} \{\min_{g \varepsilon D} e_g - \max_{g \varepsilon I} e_g\} \geq \delta$, where $\delta > 0$ is a constant.

Condition (C1) ensures that neither a very small nor a very large proportion of any given class of variables can exist. In Huang et al.'s study [5], a similar assumption is also made for condition (C1) as well as Cui et al. [4]. When the sample size reaches infinity, condition (C2) permits the minimum true signal to disappear to zero in the order of $n^{-\tau}$. The number of classes for the response and the covariates can diverge in a particular order under conditions (C3) and (C6). Make sure that the sample percentiles are close to the actual percentiles by excluding the extreme case that some X_g places heavy mass in a narrow range under condition (C4). The density must have a lower bound of order $n^{-\rho}$ for condition (C5). According to Cui et al. [4], condition (C7) makes it possible for the active covariate subset and the inactive covariate subset $I = \{1, \ldots, p\}/D$ to be very different, as well as in Zhu et al.'s study [18].

Theorem 1. *Under conditions (C1) to (C3), if all the covariates are categorical, we have*

$$P(D \subseteq \widehat{D}) \geq 1 - O\left(p \exp\left\{-bn^{1-(2\tau+4\varepsilon+4\kappa)} + (\varepsilon + \kappa)\log n\right\}\right), \tag{11}$$

where b is a positive constant. If $\log p = O(n^{\alpha})$ and $\alpha < 1 - (2\tau + 4\varepsilon + 4\kappa)$, GIGR-SIS has a sure screening property.

Theorem 2. *Under conditions (C4) to (C6), if the covariates consist of continuous and categorical variables, we have*

$$P(D \subseteq \widehat{D}) \geq 1 - O\left(p \exp\left\{-bn^{1-(2\tau+4\varepsilon+4\kappa+2\rho)} + (\varepsilon + \kappa)\log n\right\}\right), \tag{12}$$

where b is a positive constant. If $\log p = O(n^{\alpha})$ and $\alpha < 1 - (2\tau + 4\varepsilon + 4\kappa + 2\rho)$, GIGR-SIS has a sure screening property.

The proposed screening index can effectively distinguish between active and inactive covariates at the sample level, as shown by Theorem 3.

Theorem 3. *Under conditions (C1), (C4), (C5), and (C7), if $(\log(RJ_g)/\log n) = O(1)$ and $\max\{\log p, \log n\}R^4 J_g^4/n^{1-2\rho} = O(1)$, then*

$$\liminf_{n \longrightarrow \infty} \left\{\min_{g \varepsilon G} \widehat{e}_g - \max_{g \varepsilon I} \widehat{e}_g\right\} > 0, \ a.s. \tag{13}$$

3. Numerical Studies

3.1. Simulation Results. In this subsection, we carry out five simulation studies to demonstrate the finite sample performance of our group screen methods described in section 2. We will compare the performance of GIGR-SIS with IG-SIS, GIG-SIS, gSIS, and gHOLP.

There are five indicators for assessing method performance in Models 1 through 5. All active covariates are included in the minimum model size, or MMS. Although it is superior, it is comparable to the number of active covariates. The majority of results include 5%, 25%, 50%, 75%, and 95% of MMS; CP1, coverage probability 1, the probability with which the indicators of all active covariates were included in the model size $[n/\log n]$; CP2, coverage probability 2, the probability with which all active covariate indicators were included in a model of size $2[n/\log n]$; and CP3, coverage probability 3, the probability with which the indicators of all active covariates were included in a model of size $3[n/\log n]$. The indicators of whether the selected model incorporates all active covariates are CPa or total coverage probability.

3.1.1. Model 1: Binary Response. To begin, we take into consideration a straightforward model in which all of the covariates are categorical, the number of responses is binary, and R is 2, as proposed by Ni and Fang [7] and He and Deng [17]. We take into account two y_i distributions:

(1) Balanced: $P(y_i = 1) = P(y_i = 2) = 1/2$

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

The true model is defined as $D = \{1, \ldots, 15\}$, with $d_0 = 15$ and $d_{0G} = 5$, and the group size is 5. We generated the latent variable $z_i = (z_{i,1}, \ldots, z_{i,p})$, where $z_{i,k} \sim N(u_{rk}, 1)$, $1 \le k \le p$, based on y_i. We then constructed 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 $u_{rk} = 0.5$

Finally, we generated covariates using the standard normal distribution's quantile. The particular approach is as follows:

(1) If k is an odd number, then $x_{i,k} = I(z_{i,k} > z_{(j/2)}) + 1$

(2) If k is an even number, then $x_{i,k} = I(z_{i,k} > z_{(j/5)}) + 1$

Here, $z_{(\alpha)}$ is the standard normal distribution's α-th percentile.

As a result, of all the p covariates, half are two-category, while the remaining half are five-category. Take into consideration the scenarios where the sample size n was 80, 120, or 160, and the dimension of the covariate p was 1500.

Three indicators for assessing method performance over a hundred simulations are presented in Table 1.

Evaluation of Various Sample Sizes. The MMS of GIGR-SIS, GIG-SIS, gSIS, and gHOLP all approach $d_{0G} = 5$ as sample sizes increase, and the coverage probability indexes all reach 1. However, when the sample size is 80 or 100, four IG-SIS coverage probability indexes perform worse than GIGR-SIS. The GIGR-SIS MMS is superior to the IG-SIS MMS. Therefore, in Model 1, the GIGR-SIS performs better in terms of finite sample performance than the IG-SIS.

Comparison of Various Response Structures. In the balanced response, the performance of five indexes is comparable to that of the unbalanced response. Due to the small fluctuation range in MMS, the performance of GIGR-SIS is also more robust than that of the other grouped screening methods.

3.1.2. Model 2: Multiclass Response. More covariate classification is taken into account, and the response y_i is multiclass with $R = 10$. We take into account two y_i distributions:

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

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

The true model is defined as $D = \{1, \ldots, 15\}$, with $d_0 = 15$ and $d_{0G} = 5$, and the group size is 5. For covariates X_k, $x_{i,k} = f_k(\varepsilon_{i.k} + \mu_{i.k})$, where $\varepsilon_{i.k} \sim t(4)$ and $f_k(\cdot)$ are quantile functions of the standard normal distribution, and then we generate the latent variable $z_i = (z_{i,1}, \ldots, z_{i,p})$. After that, we define $\mu_{i.k}$ to construct active covariates:

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

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

At last, we generate covariates by $f_k(\cdot)$ and take $p = 2000$ and $n = 150, 200, 250$. The particular approach is as follows:

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

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

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

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

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

One-fifth of the p covariates are quadripartite, while one-fifth of the covariates are binary. Sex partite covariates make up one-fifth of the data. The remaining covariates are decuple, while one-fifth of the covariates are octuple.

Three indicators for assessing method performance over 100 simulations for Model 2 are presented in Table 2. The performance of GIGR-SIS is significantly superior to that of other screening methods in this more complex example. In particular, the GIGR-SIS MMS has a relatively small sample size.

Evaluation of Various Sample Sizes. The MMS of GIGR-SIS approaches $d_{0G} = 5$ and four indexes of coverage probability both reach 1 as sample sizes increase. However, the coverage probability indexes IG-SIS, gSIS, and gHOLP are all worse than GIGR-SIS. Compared to the MMS of other grouped screening methods, GIGR-SIS is superior. The fact that gSIS is expected to work well when there are few correlations between the predictors is the reason for its poor screening. The fact that gHOLP is expected to work well when the predictors are correlated is the reason for its poor screening. Additionally, due to the influence of model limitations, the results of gSIS and gHOLP remain the same as the sample size increases. Therefore, in Model 2, the GIGR-SIS performs better with finite samples than other grouped screening methods.

Comparison of Various Response Structures. Five indexes perform better in the unbalanced response than in the balanced response. Due to the small fluctuation range in MMS, the performance of GIGR-SIS is also more robust than that of other screening methods.

3.1.3. Model 3: Continuous and Categorical Covariates. After that, we looked into a more complicated example with both continuous and categorical covariates, and the response is a multiclass one, and $R = 4$. We took into account two y_i distributions:

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

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

Table 1: Simulation results for Model 1.

Condition	MMS					CP			
	5%	25%	50%	75%	95%	CP1	CP2	CP3	Cpa
Balanced Y, $n = 80$, $p = 1500$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	107.90	137.75	160.00	269.75	695.05	0.574	0.713	0.777	0.000
GIG-SIS	5.00	5.00	5.00	5.00	6.00	1.000	1.000	1.000	1.000
gSIS	7.00	9.75	12.00	14.00	20.00	0.902	0.996	1.000	1.000
gHOLP	8.00	11.00	16.00	21.00	35.00	0.810	0.972	0.994	0.980
Balanced Y, $n = 100$, $p = 1500$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	139.25	191.00	246.00	283.00	331.55	0.711	0.851	0.899	0.000
GIG-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
gSIS	5.00	6.00	6.00	8.00	9.05	1.000	1.000	1.000	1.000
gHOLP	5.00	6.00	8.00	9.25	14.05	0.992	1.000	1.000	1.000
Balanced Y, $n = 120$, $p = 1500$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	24.95	30.00	37.00	44.00	52.00	0.846	0.961	0.995	0.940
GIG-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
gSIS	5.00	5.00	5.00	5.00	6.00	1.000	1.000	1.000	1.000
gHOLP	5.00	5.00	5.00	6.00	7.00	1.000	1.000	1.000	1.000
Unbalanced Y,$n = 80$, $p = 1500$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	133.95	219.50	531.00	843.50	1154.05	0.523	0.681	0.747	0.000
GIG-SIS	5.00	5.00	5.00	6.00	8.00	0.998	1.000	1.000	1.000
gSIS	7.95	13.75	19.50	25.25	54.25	0.832	0.942	0.978	0.890
gHOLP	8.00	17.00	24.00	36.25	67.05	0.804	0.902	0.950	0.750
Unbalanced Y, $n = 100$, $p = 1500$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	184.90	220.00	306.00	398.00	514.60	0.692	0.835	0.873	0.000
GIG-SIS	5.00	5.00	5.00	5.00	6.00	1.000	1.000	1.000	1.000
gSIS	5.00	6.00	7.00	8.00	12.05	0.998	1.000	1.000	1.000
gHOLP	5.00	7.00	8.00	11.00	17.15	0.982	0.998	1.000	1.000
Unbalanced Y, $n = 120$, $p = 1500$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	20.00	22.00	25.00	30.25	43.15	0.807	0.991	0.999	0.990
GIG-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
gSIS	5.00	5.00	5.00	5.00	6.00	1.000	1.000	1.000	1.000
gHOLP	5.00	5.00	5.00	5.25	6.05	1.000	1.000	1.000	1.000

The true model is defined as $D = \{1, 2, 3, 751, 752, 753, 1501, 1502, 1503, 1504, 1505, 1506\}$ with $d_0 = 12$ and $d_{0G} = 4$, and the group size is 4. We take $p = 3000$, $n = 180, 200, 220$ in this model. For covariates X_k, we generated latent variable $z_i = (z_{i,1}, \ldots, z_{i,p})$, $z_{i,k} \sim N(\mu_i, 1)$, and $1 \le k \le p$, 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 Ni and Fang [7] and He and Deng [17], θ_{rk} is given in Table 3. $u_{ik} = 0$ when $k \notin D$. To generate X_k, we have those as follows:

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

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

(3) For $1501 \le k$, then $x_{ik} = z_{ik}$

Four-category covariates account for one-fifth of all p covariates. The remaining covariates are continuous, while one-fifth of the covariates are four-category. Continuous covariates make up half of the 12 active covariates, while categorical covariates with four categories and ten categories make up the remaining covariates.

When the numbers of covariates are grouped, He and Deng [17] proposed a grouped feature screening method by using the joint information entropy to screen some important grouped covariates. We denote them as GIG-SIS-4, GIG-SIS-8, and GIG-SIS-10. The simulation results for 100 replications for the balanced and unbalanced cases are presented in Tables 4 and 5, respectively.

Evaluation of Various Sample Sizes. The MMS of GIGR-SIS approaches $d_{0G} = 4$ and the four indexes of coverage probability both reach 1 as the sample sizes increase. However, when the sample size is 180, the five coverage probability indices of gSIS and gHOLP are inferior to GIGR-SIS. GIGR-SIS outperforms IG-SIS, gSIS, and gHOLP in terms of MMS. In Model 3, the GIGR-SIS outperforms the IG-SIS, gSIS, and gHOLP in terms of performance with finite samples. Additionally, five

Table 2: Simulation results for Model 2.

Condition	MMS					CP			
	5%	25%	50%	75%	95%	CP1	CP2	CP3	Cpa
Balanced Y, $n = 100$, $p = 2000$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	1283.00	1283.00	1283.00	1283.00	1283.00	0.600	0.600	0.600	0.000
GIG-SIS	7.00	8.00	8.00	9.00	11.00	0.998	1.000	1.000	1.000
gSIS	1262.00	1262.00	1262.00	1262.00	1262.00	0.600	0.600	0.600	0.000
gHOLP	1455.00	1455.00	1455.00	1455.00	1455.00	0.600	0.600	0.600	0.000
Balanced Y, $n = 150$, $p = 2000$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	1283.00	1283.00	1283.00	1283.00	1283.00	0.600	0.600	0.600	0.000
GIG-SIS	9.00	9.00	10.00	14.00	23.50	0.978	0.994	0.998	0.990
gSIS	1262.00	1262.00	1262.00	1262.00	1262.00	0.600	0.600	0.600	0.000
gHOLP	1455.00	1455.00	1455.00	1455.00	1455.00	0.600	0.600	0.600	0.000
Balanced Y, $n = 200$, $p = 2000$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	1283.00	1283.00	1283.00	1283.00	1283.00	0.600	0.600	0.600	0.000
GIG-SIS	9.00	9.00	9.00	9.00	10.05	1.000	1.000	1.000	1.000
gSIS	1262.00	1262.00	1262.00	1262.00	1262.00	0.600	0.600	0.600	0.000
gHOLP	1455.00	1455.00	1455.00	1455.00	1455.00	0.600	0.600	0.600	0.000
Unbalanced Y, $n = 100$, $p = 2000$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	1283.00	1283.00	1283.00	1283.00	1283.00	0.600	0.600	0.600	0.000
GIG-SIS	11.00	11.00	11.00	11.00	11.00	1.000	1.000	1.000	1.000
gSIS	1262.00	1262.00	1262.00	1262.00	1262.00	0.600	0.600	0.600	0.000
gHOLP	1455.00	1455.00	1455.00	1455.00	1455.00	0.600	0.600	0.600	0.000
Unbalanced Y, $n = 150$, $p = 2000$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	1283.00	1283.00	1283.00	1283.00	1283.00	0.600	0.600	0.600	0.000
GIG-SIS	10.00	10.00	10.00	10.00	10.00	1.000	1.000	1.000	1.000
gSIS	1262.00	1262.00	1262.00	1262.00	1262.00	0.600	0.600	0.600	0.000
gHOLP	1455.00	1455.00	1455.00	1455.00	1455.00	0.600	0.600	0.600	0.000
Unbalanced Y, $n = 200$, $p = 2000$									
GIGR-SIS	5.00	5.00	5.00	5.00	5.00	1.000	1.000	1.000	1.000
IG-SIS	1283.00	1283.00	1283.00	1283.00	1283.00	0.600	0.600	0.600	0.000
GIG-SIS	9.00	9.00	9.00	9.00	9.00	1.000	1.000	1.000	1.000
gSIS	1262.00	1262.00	1262.00	1262.00	1262.00	0.600	0.600	0.600	0.000
gHOLP	1455.00	1455.00	1455.00	1455.00	1455.00	0.600	0.600	0.600	0.000

Table 3: 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

GIGR-SIS coverage probability indices are comparable to GIG-SIS. As a result, it is demonstrated that the GIGR-SIS possesses group feature screening characteristics.

Comparison of Various Response Structures. Four indexes perform better in the unbalanced response than in the balanced response. Due to the influence of model limitations, the results for gSIS and gHOLP are poor. However, there are two types of responses to which the MMS of GIGR-SIS and GIG-SIS is robust.

Comparison of Various Slice Counts. In terms of MMS and five indexes of coverage probability, the performance of GIGR-SIS is superior regardless of the number of slices applied to continuous covariates. The number of slices has no effect on GIGR-SIS or GIG-SIS's performance.

3.1.4. Model 4: Continuous Covariates. The true model is defined as $D = \{1, 2, 3, 751, 752, 753, 1501, 1502, 1503, 1504, 1505, 1506\}$ with $d_0 = 12$ and $d_{0G} = 4$, and the group size is 4. We take $p = 3000$ and $n = 180, 200, 220$ in this model.

Table 4: Simulation results for Model 3: balanced Y.

Condition	MMS					CP			
	5%	25%	50%	75%	95%	CP1	CP2	CP3	Cpa
Balanced Y, $n = 180$, $p = 3000$									
GIGR-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-4	14.95	22.00	32.50	52.75	92.05	0.940	0.975	0.990	0.880
GIG-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-4	276.75	484.00	745.50	1052.50	1402.10	0.488	0.495	0.505	0.000
gHOLP-4	400.45	773.75	1095.00	1435.75	1924.95	0.438	0.485	0.493	0.000
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	13.00	16.00	20.00	31.25	79.10	0.965	0.990	0.994	0.930
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	245.00	466.25	669.50	918.00	1243.70	0.493	0.503	0.505	0.000
gHOLP-8	379.65	730.75	949.50	1208.00	1685.90	0.475	0.493	0.493	0.000
GIGR-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-10	13.95	17.00	22.00	39.25	123.10	0.961	0.985	0.989	0.870
GIG-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-10	218.85	447.50	666.00	908.00	1283.65	0.493	0.503	0.505	0.000
gHOLP-10	368.80	692.25	955.50	1284.75	1678.70	0.473	0.490	0.493	0.000
Balanced Y, $n = 200$, $p = 3000$									
GIGR-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-4	14.00	16.00	20.00	25.00	60.00	0.982	0.994	0.998	0.980
GIG-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-4	170.00	352.50	632.00	916.00	1219.05	0.500	0.508	0.515	0.010
gHOLP-4	313.75	614.50	1026.00	1316.00	1834.45	0.470	0.495	0.500	0.000
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	13.00	15.00	18.00	24.25	46.15	0.988	0.997	0.999	0.990
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	173.20	323.25	590.00	769.00	1036.20	0.503	0.508	0.515	0.010
gHOLP-8	302.50	599.75	883.50	1183.50	1662.45	0.490	0.500	0.500	0.000
GIGR-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-10	13.00	17.00	22.50	31.50	54.55	0.968	0.995	0.997	0.960
GIG-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-10	167.05	310.00	556.50	823.25	1062.30	0.500	0.508	0.515	0.010
gHOLP-10	304.00	557.75	887.50	1183.50	1638.30	0.485	0.500	0.500	0.000
Balanced Y, $n = 220$, $p = 3000$									
GIGR-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-4	12.95	14.00	16.00	19.00	26.00	0.998	0.999	1.000	1.000
GIG-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-4	107.55	309.00	449.50	700.75	1058.05	0.503	0.530	0.548	0.050
gHOLP-4	239.50	536.00	806.50	1072.75	1548.95	0.490	0.498	0.503	0.000
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	13.00	15.00	18.00	21.00	29.00	0.995	1.000	1.000	1.000
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	105.95	290.25	436.50	611.50	866.70	0.503	0.523	0.545	0.040
gHOLP-8	278.95	533.25	732.00	971.50	1286.55	0.495	0.500	0.503	0.000
GIGR-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-10	14.00	17.00	22.00	26.00	37.35	0.985	0.999	1.000	1.000
GIG-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-10	108.30	270.75	410.00	608.25	857.90	0.503	0.525	0.543	0.040
gHOLP-10	218.75	536.00	719.00	955.50	1352.70	0.495	0.500	0.503	0.000

Condition on y_i, we generate latent variable $z_i = (z_{i,1}, \ldots, z_{i,p})$. For covariates X_k, $z_{i,k} \sim N(\mu_i, 1)$ and $1 \le k \le p$, 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 Ni and Fang [7] and He and Deng [17], θ_{rk} is given in Table 3. $u_{ik} = 0$ when $k \notin D$. To generate X_k, for $1 \le k \le p$, then $x_{ik} = z_{ik}$.

The active covariates and all of the p covariates are continuous covariates. We use a slice count of $J_k = 8$ for the continuous covariates. The names GIGR-SIS-8, IG-SIS-8, GIG-SIS-8, gSIS-8, and gHOLP-8 are used to represent the respective methods.

In Table 6, we can get that the GIGR-SIS screening method proposed in this paper works well for continuous data as the sample size increases, but gSIS and gHOLP are affected by the model limitations and are poorly screened for continuous data. The MMS of GIGR-SIS approaches $d_{0G} = 4$ when the sample sizes increase, and the four indexes of coverage probability both reach 1. Additionally, five GIGR-

Table 5: Simulation results for Model 3: unbalanced Y.

Condition	MMS					CP			
	5%	25%	50%	75%	95%	CP1	CP2	CP3	Cpa
Unbalanced Y, $n = 180$, $p = 3000$									
GIGR-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-4	38.95	43.00	46.00	50.25	55.05	0.840	0.969	1.000	1.000
GIG-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-4	531.00	551.75	562.50	575.25	594.05	0.500	0.500	0.500	0.000
gHOLP-4	586.05	787.25	889.00	1049.50	1190.45	0.360	0.488	0.498	0.000
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	22.00	26.00	29.00	32.00	35.05	0.929	1.000	1.000	1.000
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	511.95	528.75	547.00	554.25	574.00	0.500	0.500	0.500	0.000
gHOLP-8	575.65	779.25	911.50	1030.75	1191.05	0.463	0.498	0.500	0.000
GIGR-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-10	21.00	23.00	26.00	29.00	33.00	0.948	1.000	1.000	1.000
GIG-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-10	509.95	528.00	545.50	556.00	574.05	0.500	0.500	0.500	0.000
gHOLP-10	599.65	758.25	892.00	1023.00	1222.75	0.460	0.500	0.500	0.000
Unbalanced Y, $n = 200$, $p = 3000$									
GIGR-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-4	17.00	21.00	23.00	25.00	29.00	0.986	1.000	1.000	1.000
GIG-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-4	395.75	412.00	425.00	438.00	461.05	0.500	0.500	0.500	0.000
gHOLP-4	467.50	593.50	746.00	911.00	1065.65	0.448	0.500	0.500	0.000
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	13.00	15.00	17.00	18.00	21.00	1.000	1.000	1.000	1.000
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	368.90	382.75	394.00	407.00	425.05	0.500	0.500	0.500	0.000
gHOLP-8	471.05	592.00	714.00	894.25	1077.10	0.485	0.500	0.500	0.000
GIGR-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-10	17.00	19.75	21.00	23.00	27.00	0.994	1.000	1.000	1.000
GIG-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-10	350.85	365.75	377.50	391.00	406.15	0.500	0.500	0.500	0.000
gHOLP-10	465.15	573.75	667.00	851.00	1006.30	0.480	0.500	0.500	0.000
Unbalanced Y, $n = 220$, $p = 3000$									
GIGR-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-4	15.95	18.00	20.00	22.00	25.00	1.000	1.000	1.000	1.000
GIG-SIS-4	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-4	408.00	439.75	450.00	467.00	491.20	0.500	0.500	0.500	0.000
gHOLP-4	557.75	717.50	840.50	957.00	1149.45	0.428	0.498	0.500	0.000
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	13.00	14.00	16.00	17.00	18.00	1.000	1.000	1.000	1.000
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	339.95	361.00	373.50	384.00	404.00	0.500	0.500	0.500	0.000
gHOLP-8	486.35	631.50	729.50	845.50	1072.75	0.485	0.500	0.500	0.000
GIGR-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-10	14.95	17.00	19.00	20.00	22.05	1.000	1.000	1.000	1.000
GIG-SIS-10	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-10	317.00	339.75	350.00	364.00	382.20	0.500	0.500	0.500	0.000
gHOLP-10	444.55	607.00	685.00	804.25	1029.25	0.478	0.500	0.500	0.000

SIS coverage probability indices are comparable to GIG-SIS. As a result, it is demonstrated that the GIGR-SIS possesses group feature screening characteristics.

3.1.5. Model 5: Computational Time Complexity Analysis. It is the same as Model 1. But for distribution of y_i, we consider balanced data, that is, $P(y_i = r) = 1/2$. The true model is defined as $D = \{1, \ldots, 15\}$, with $d_0 = 15$ and $d_{0G} = 5$, and the group size is 5. The active and irrelevant covariates are generated in the same way as in Model 1. Similar to this, half of the p-dimensional covariates are two-category, while the other half are five-category.

Model 5 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 time of the five methods will be recorded for each experiment, and the median running

Table 6: Simulation results for Model 4.

Condition	MMS					CP			
	5%	25%	50%	75%	95%	CP1	CP2	CP3	Cpa
Balanced Y, $n = 180$, $p = 3000$									
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	12.00	13.00	16.00	23.00	30.20	0.986	0.998	0.999	0.990
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	249.05	464.50	664.00	909.25	1259.00	0.495	0.508	0.530	0.000
gHOLP-8	373.10	697.25	943.00	1207.00	1707.05	0.475	0.493	0.495	0.000
Balanced Y, $n = 200$, $p = 3000$									
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	12.00	12.00	14.00	16.00	21.05	0.997	1.000	1.000	1.000
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	173.45	319.50	586.00	758.00	1049.30	0.510	0.518	0.553	0.010
gHOLP-8	309.70	603.25	868.50	1123.00	1688.30	0.485	0.500	0.505	0.000
Balanced Y, $n = 200$, $p = 2000$									
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	12.00	13.00	13.00	15.00	23.05	0.999	1.000	1.000	1.000
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	108.80	289.75	446.00	601.75	865.95	0.515	0.550	0.580	0.040
gHOLP-8	216.85	559.75	743.00	948.00	1281.45	0.498	0.505	0.513	0.000
Unbalanced Y, $n = 180$, $p = 2000$									
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	19.00	21.00	24.00	26.00	30.00	0.960	1.000	1.000	1.000
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	508.90	530.50	545.00	555.25	573.05	0.500	0.500	0.500	0.000
gHOLP-8	594.90	782.50	906.50	1056.75	1244.70	0.458	0.498	0.500	0.000
Unbalanced Y, $n = 200$, $p = 3000$									
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	12.00	12.00	12.00	13.00	14.00	1.000	1.000	1.000	1.000
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	365.85	382.75	393.00	406.25	426.10	0.500	0.500	0.645	0.000
gHOLP-8	495.55	610.25	733.00	900.00	1126.05	0.475	0.503	0.505	0.000
Unbalanced Y, $n = 200$, $p = 2000$									
GIGR-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
IG-SIS-8	12.00	12.00	13.00	14.00	15.00	1.000	1.000	1.000	1.000
GIG-SIS-8	4.00	4.00	4.00	4.00	4.00	1.000	1.000	1.000	1.000
gSIS-8	337.75	359.00	374.00	383.00	401.10	0.500	0.505	0.675	0.000
gHOLP-8	504.00	655.25	736.50	853.50	1091.95	0.485	0.500	0.505	0.000

time in 100 replicate experiments will be recorded as the running time index of the five methods, and the trend of the five methods will be compared as the dimension of covariates increases, so as to compare the computational complexity of the methods.

The median run times in Table 7 show a linear trend for the five methods as the sample size varies linearly. It can be seen that the running time of GIGR-SIS is not much different from that of GIG-SIS, while the running time of gSIS and gHOLP is half of that of GIGR-SIS. This is because gSIS and gHOLP are limited by the model setting, and although the running time is short, the screening effect is not good; see Table 1 for details.

Table 7: Simulation results for Model 5.

p	Feature screening Methods (note: running time in seconds)				
	GIGR-SIS	IG-SIS	GIG-SIS	gSIS	gHOLP
1500	3.354	2.930	3.383	1.519	1.601
2500	5.673	4.901	5.691	2.535	2.670
3500	7.974	6.952	8.029	3.551	3.754
4500	10.320	8.927	10.400	4.573	4.816
5500	12.726	10.931	12.809	5.607	5.925
6500	15.075	12.860	15.186	6.625	7.007
7500	17.636	14.915	17.761	7.662	8.093
8500	19.533	16.544	19.660	8.556	9.038
9500	21.837	18.512	21.972	9.588	10.115
10500	24.242	20.580	24.404	10.649	11.269

3.2. Real Data. In this subsection, we analyse a real dataset, mutational status of p53 in cell lines, from the gene expression studies reported in Subramanian et al.[19] and Zeng and Breheny [20]. The p53 study aims to identify pathways that correlated with the mutational status of the gene p53, which regulates gene expression in response to various signals of cellular stress [20]. The p53 data consists of 50 sample and 4301 features, where 17 of 50 cell lines are classified as normal and 33 of which carry mutations in the p53 gene. These gene sets contain a total

Table 8: P53 data screened for gene pathways.

	Gene pathway number of the screened											
GIGR-SIS-4	91	177	241	74	82	167	264	195	185	68	146	35
GIGR-SIS-8	91	177	74	241	82	68	63	167	185	264	79	195
GIGR-SIS-10	91	177	241	74	82	167	63	68	195	141	185	79
gSIS	153	152	151	150	149	148	147	146	145	144	143	142
gHOLP	153	152	151	150	149	148	147	146	145	144	143	142

of 4301 genes, in the context of biological genetics, and the genes here do not act alone but together in the same gene pathway.

According to the analysis in Subramanian et al. [19] and Zeng and Breheny [20], the 4301 gene variables were grouped into 308 gene pathways that contained varying numbers of genes. In the screening, we selected $n/\log n = 12$ group variables as significant group variables. We applied GIGR-SIS-4, GIGR-SIS-8, and GIGR-SIS-10 methods to select important gene pathways, which are group variables, respectively. Because of the number of group variables, that is, more genes in gene pathways, here we only give the final selected group number, and the specific screening group number results are shown in Table 8. We can find that the results of our screening methods are not much different after the selection of group variables.

4. Conclusions

In this article, we have proposed the grouped feature screening GIGR-SIS method based on information gain ratio for categorical covariates in the classification model. Compared with most existing literature, we deal with variables which can be naturally grouped. The GIGR-SIS feature screening method is model-free, and we establish the sure screening property for this group screening approach. Simulation results show that the performance of GIGR-SIS outperforms that of GIG-SIS, gSIS, and gHOLP. The study on p53 data also shows the advantages of our screening method.

The grouped feature screening encounters some difficulties when data are missing. For the purpose of resolving the classification model's missing covariates or responses, we would like to propose a novel feature screening approach.

Appendix

Proposition 1 and Proposition 2 have been proved in He and Deng [17]; hence, they are omitted.

Lemma A.1 (Bernstein inequality). *If $Z_1, Z_2, \ldots, Z_n$ are independent random variables with a mean value of 0 and bounded support is $[-M, M]$, then the inequality $P(|\sum_{i=1}^{n} Z_i| > t) \le 2\exp\{-(t^2/2(v + Mt/3))\}$, where $v \ge Var(\sum_{i=1}^{n} Z_i)$.*

Lemma A.2. *We have the following three inequalities for discrete group covariates X_g and discrete response Y:*

(1) $P(|\widehat{p}_r - p_r| > t) \le 2\exp\{-6nt^2/3 + 4t\}$

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

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

Proof of Lemma A.1. The proofs for the aforementioned inequality (a) and inequality (b) have been provided in Ni [21] and He et al. [17] and are comparable. The proof of inequality (c) is given here. $\widehat{p}_{j_g r} = \sum_{i=1}^{n} I\{y_i = r,\ x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{p_g}\} / \sum_{i=1}^{n} I\left\{x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{p_g}\right\}$.

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

$$E\left(\widehat{p}_{j_g r}\right) = E\left(\frac{I\left\{y_i = r, x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{p_g}\right\}}{I\left\{x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{p_g}\right\}}\right) = p_{j_g r}. \tag{A.1}$$

Let $Z_i = I\left\{y_i = r | x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{p_g}\right\} - p_{j_g r}$ and $\mathrm{Var}(\sum_{i=1}^{n} Z_i) = n\,\mathrm{Var}(Z_i) = np_{j_g r}(1 - p_{j_g r}) \le (n/4)$ be known, then $P(|\widehat{p}_{j_g r} - p_{j_g r}| > t) = P(|n^{-1}\sum_{i=1}^{n} Z_i| > t) = P(|\sum_{i=1}^{n} Z_i| > nt) \le 2\exp\{-n^2t^2/2((n/4) + (nt/3))\} \le 2\exp\{-(6nt^2/3 + 4t)\}$.

The preceding formula is established by Bernstein inequality.

Lemma A.3. *Under condition (C1), we have $P(|\widehat{e}_g - e_g| > 2\varepsilon) \le O(RJ^3)\exp\{-c_5 n\varepsilon^2/R^4J^{12}\}$, for discrete response Y and discrete group covariates X_g, where c_5 is a constant.*

Proof of lemma A.2. Section 2.2's e_g and $\widehat{e}_g$ indicate that we have

$$
\begin{aligned}
&\widehat{e}_g - e_g \\
&= \frac{\left(\sum_{j_{g=c(1,1,\ldots,1)}}^{J_g} \widehat{w} j_g \log \widehat{w} j_g\right) \times \left(\sum_{j_{g=c(1,1,\ldots,1)}}^{J_g} \widehat{w} j_g\right) \sum_{r=1}^{R} \widehat{p}_{jg^r} \log\, \widehat{p}_{jg^r} - \sum_{r=1}^{R} \widehat{p}_r \log\, \widehat{p}_r}{\left(\sum_{j_{g=c(1,1,\ldots,1)}}^{J_g} \widehat{w} j_g \log\, \widehat{w}_{jg}\right) \times \left(\sum_{j_{g=c(1,1,\ldots,1)}}^{J_g} w_{jg} \log\, w_{jg}\right)}, \\
&\quad - \frac{\left(\sum_{j_g=c(1,1,\ldots,1)}^{J_g} w_{jg} \log\, w_{jg}\right) \times \left(\sum_{j_g=c(1,1,\ldots,1)}^{J_g} w_{jg} \sum_{r=1}^{R} p_{jg^r} \log\, p_{jg^r} - \sum_{r=1}^{R} p_r \log\, p_r\right)}{\left(\sum_{j_{g=c(1,1,\ldots,1)}}^{J_g} \widehat{w} j_g \log \widehat{w} j_g\right) \times \left(\sum_{j_{g=c(1,1,\ldots,1)}}^{J_g} w_{jg} \log\, w_{jg}\right)}, \\
&\le 2\left[\left(\sum_{j_g=c(1,1,\ldots,1)}^{J_g} \widehat{w} j_g \sum_{r=1}^{R} \widehat{p} j_{g^r} \log \widehat{p} j_{g^r} - \sum_{r=1}^{R} \widehat{p}_r \log \widehat{p}_r\right) - \left(\sum_{j_g=c(1,1,\ldots,1)}^{J_g} \widehat{w} j_g\right) \sum_{r=1}^{R} p j_{g^r} \log\, p j_{g^r} - \sum_{r=1}^{R} p_r \log\, p_r\right] \\
&= 2\left(\sum_{j_g=c(1,1,\ldots,1)}^{J_g} \widehat{w} j_g \sum_{r=1}^{R} \widehat{p} j_{g^r} \log \widehat{p} j_{g^r} - \sum_{j_g=c(1,1,\ldots,1)}^{J_g} \widehat{w} j_g \sum_{r=1}^{R} \widehat{p} j_{g^r} \log \widehat{p} j_{g^r}\right) - 2\left(\sum_{r=1}^{R} \widehat{p}_r \log \widehat{p}_r - \sum_{r=1}^{R} p_r \log\, p_r\right) \\
&= 2\left\{\sum_{r=1}^{R}\left[\sum_{j_g=c(1,1,\ldots,1)}^{J_g} \frac{1}{n} \sum_{i=1}^{n} I\left\{y_i = r, x_{i,g1} = j_1, \ldots, x_{i,gp_g} = i_{pg}\right\} \log \widehat{p} j_{g^r}\right]\right. \\
&\quad \left. - \sum_{r=1}^{R}\left[\sum_{j_g=c(1,1,\ldots,1)}^{J_g} \frac{1}{n} \sum_{i=1}^{n} p\left\{y_i = r, x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{pg}\right\} \log \widehat{p} j_{g^r}\right]\right\} \\
&\quad - 2\left(\sum_{r=1}^{R} \widehat{p}_r \log\, \widehat{p}_r - \sum_{r=1}^{R} p_r \log\, p_r\right), \\
&= 2\sum_{r=1}^{R}\left[\sum_{j_g=c(1,1,\ldots,1)}^{J_g} \frac{1}{n} \sum_{i=1}^{n} I\left\{y_i = r, x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{pg}\right\}\left(\log \widehat{p} j_{g^r} - \log\, p j_{g^r}\right)\right] \\
&\quad + 2\sum_{r=1}^{R}\left[\sum_{j_g=c(1,1,\ldots,1)}^{J_g} \frac{1}{n} \sum_{i=1}^{n}\left(I\left\{y_i = r, x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{pg}\right\} - p\left(y_i = r, x_{i,g1} = j_1, \ldots, x_{i,gp_g} = j_{pg}\right)\right) \log\, p j_{g^r}\right] \\
&\quad - 2\left(\sum_{r=1}^{R} \widehat{p}_r \log\, \widehat{p}_r - \sum_{r=1}^{R} p_r \log\, p_r\right) \\
&= 2\left(I_1 + I_2 + I_3\right) P\left(\left|\widehat{e_g} - e_g\right| > 2\varepsilon\right) \le P\left(\left|I_1 > \frac{\varepsilon}{3}\right|\right) + P\left(\left|I_2 > \frac{\varepsilon}{3}\right|\right) + P\left(\left|I_3 > \frac{\varepsilon}{3}\right|\right).
\end{aligned}
\tag{A.2}
$$

For $|I_1|$, we have

$$|I_1| \le \sum_{r=1}^{R} \sum_{j_g=c(1,1,\ldots,1)}^{J_g} \left|\log \hat{p}_{j_g r} - \log p_{j_g r}\right|. \tag{A.3}$$

Then,

$$\begin{aligned}
&P\left(|I_1| > \frac{\varepsilon}{3}\right) \\
&\le \sum_{r=1}^{R} \sum_{j_g=c(1,1,\ldots,1)}^{J_g} P\left(\left|\log \hat{p}_{j_g r} - \log p_{j_g r}\right| > \frac{\varepsilon}{3RJ^3}\right) \\
&\le \sum_{r=1}^{R} \sum_{j_g=c(1,1,\ldots,1)}^{J_g} P\left(\left|\log \hat{p}_{j_g r} - \log p_{j_g r}\right| > \frac{\varepsilon}{3RJ^3}, \left|\log \hat{p}_{j_g r} - \log p_{j_g r}\right| \le \frac{c_1}{2RJ^3}\right) + P\left(\left|\log \hat{p}_{j_g r} - \log p_{j_g r}\right| > \frac{c_1}{2RJ^3}\right) \\
&\le \sum_{r=1}^{R} \sum_{j_g=c(1,1,\ldots,1)}^{J_g} P\left(\left|\log \hat{p}_{j_g r} - \log p_{j_g r}\right| > \frac{c_1\varepsilon}{6R^2J^6}\right) + P\left(\left|\log \hat{p}_{j_g r} - \log p_{j_g r}\right| > \frac{c_1}{2RJ^3}\right) \\
&\le RJ^3 \cdot 2 \exp\left\{-\frac{6n\left(c_1\varepsilon/6R^2J^6\right)^2}{3 + 4\left(c_1\varepsilon/6R^2J^6\right)}\right\} + RJ_g \cdot 2 \exp\left\{-\frac{6n\left(c_1/2RJ^3\right)^2}{3 + 4\left(c_1/2RJ^3\right)}\right\}.
\end{aligned} \tag{A.4}$$

Similarly, for $|I_2|$, we have

$$\begin{aligned}
&|I_2| \\
&= \left|\sum_{r=1}^{R} \sum_{j_g=c(1,1,\ldots,1)}^{J_g} \left(\hat{w}_{j_g} \cdot \hat{p}_{j_g r} - w_{j_g} \cdot p_{j_g r}\right) \log p_{j_g r}\right|
\end{aligned} \tag{A.5}$$

Then,

$$\begin{aligned}
&P\left(|I_2| > \frac{\varepsilon}{3}\right) \\
&\le \sum_{r=1}^{R} \sum_{j_g=c(1,1,\ldots,1)}^{J_g} P\left(\left|\left(\hat{w}_{j_g} \cdot \hat{p}_{j_g r} - w_{j_g} \cdot p_{j_g r}\right)\right| > \frac{\varepsilon}{3RJ^3\left(\log RJ^3 + |\log c_1|\right)}\right) \\
&\le RJ^3 \cdot 2 \exp\left\{-\frac{6n\left(\varepsilon/3RJ^3\left(\log RJ^3 + |\log c_1|\right)\right)^2}{3 + 4\left(\varepsilon/3RJ^3\left(\log RJ^3 + |\log c_1|\right)\right)}\right\}
\end{aligned} \tag{A.6}$$

For $|I_3|$, we have

$$\begin{aligned}|I_3| &= \left|\sum_{r=1}^{R} p_r \log\, p_r - \sum_{r=1}^{R} \widehat{p}_r \log\, \widehat{p}_r\right| = \sum_{r=1}^{R} \left|\widehat{p}_r (\log\, \widehat{p}_r - \log\, p_r) + \log\, p_r (\widehat{p}_r - p_r)\right| \le \sum_{r=1}^{R} \left|\widehat{p}_r (\log\, \widehat{p}_r - \log\, p_r)\right| \\ &+ \sum_{r=1}^{R} \left|\log\, p_r (\widehat{p}_r - p_r)\right| = |I_{31}| + |I_{32}|. \end{aligned} \tag{A.7}$$

Then,

$$\begin{aligned} &P\left(|I_3| > \frac{\varepsilon}{3}\right) \\ &P\left(|I_{31}| > \frac{\varepsilon}{6}\right) \le \sum_{r=1}^{R} P\left(\left|\log\, \widehat{p}_r - \log\, p_r\right| > \frac{\varepsilon}{6R}\right) \le \sum_{r=1}^{R} P\left(\left|\log\, \widehat{p}_r - \log\, p_r\right| > \frac{\varepsilon}{6R}, \left|\widehat{p}_r - p_r\right| \le \frac{c_2}{2R}\right) + \sum_{r=1}^{R} P\left(\left|\widehat{p}_r - p_r\right| > \frac{c_2}{2R}\right), \\ &P\left(|I_{31}| > \frac{\varepsilon}{6}\right) \le \sum_{r=1}^{R} P\left(\left|\log\, \widehat{p}_r - \log\, p_r\right| > \frac{c_2 \varepsilon}{12R^2}\right) + \sum_{r=1}^{R} P\left(\left|\widehat{p}_r - p_r\right| > \frac{c_2}{2R}\right) \\ &\le R \cdot 2\, \exp\left\{-\frac{6n\left(c_2 \varepsilon/12R^2\right)^2}{3 + 4\left(c_2 \varepsilon/12R^2\right)}\right\} + R \cdot 2\, \exp\left\{-\frac{6n\left(c_2/2R\right)^2}{3 + 4\left(c_2/2R\right)}\right\}, \\ &P\left(|I_{32}| > \frac{\varepsilon}{6}\right) \le \sum_{r=1}^{R} P\left(\left|\widehat{p}_r - p_r\right| > \frac{\varepsilon}{6R\left(\log\, R + \left|\log\, c_1\right|\right)}\right) \le R \cdot 2\, \exp\left\{-\frac{6n\left(\varepsilon/6R\left(\log\, R + \left|\log\, c_1\right|\right)\right)^2}{3 + 4\left(\varepsilon/6R\left(\log\, R + \left|\log\, c_1\right|\right)\right)}\right\}. \end{aligned} \tag{A.8}$$

In sum, we have inequality $P(|\widehat{e}_g - e_g| > 2\varepsilon) \le O(RJ^3)\exp\{-c_5 n\varepsilon^2/R^4 J^{12}\}$, where c_5 is a constant.

Proof of Theorem 1. According to Conditions (C1) to (C3) and Lemma A.3, we have

$$\begin{aligned} &P(D \subseteq \widehat{D}) \\ &\ge 1 - O\left(RJ^3\right) p\, \exp\left\{-c_5 \frac{c^2 n^{1-2\tau}}{R^4 J^{12}}\right\} \ge 1 - O\left(p\, \exp\left\{-bn^{1-2\tau-4\varepsilon-4\kappa} + (\varepsilon + \kappa)\log\, n\right\}\right), \end{aligned} \tag{A.9}$$

where b is a positive constant.

Lemma A.4 (Lemma A.2 [21]). *For any continuous covariate X_g satisfying conditions (C4) and (C5), for any $\varepsilon > 0$, $1 \le j \le J_g$, and $1 \le r \le R$, we have $P(|\widehat{F}_k(r, q_{k,(j)}) - F_k(r, q_{k,(j)})| > \varepsilon) \le c_6 \exp\{-c_7 n^{1-2p}\varepsilon^2\}$, where $c_6 = 3c_8$ and $c_7 = \min\{1/2, c_4^2/2c_3^2\}$ are two positive constants.*

Lemma A.5 (Lemma A.5 [17]). *For continuous X_g, under (C1), (C4), and (C5), we have $P(|\widehat{e}_g - e_g| > 2\varepsilon) \le O(RJ_g)\exp\{-c_9 n^{1-2p}\varepsilon^2/R^4 J_g^4\}$ for any $0 < \varepsilon < 1$, and there is a positive constant c_9.*

Proof of Theorem A.2. Based on Lemma A.5, since the Proof of Theorem A.2 is identical to that of Theorem 1, it is omitted.

Proof of Theorem A.3. Under Conditions (C1), (C4), (C5), and (C7) and by Lemma A.3 and A.5., the proof is similar to that of Ni and Fang [7]. Then, we have

$$
\begin{aligned}
&P\left(\min_{g\in D}\widehat{e}_g - \max_{g\in I}\widehat{e}_g < \frac{\delta}{2}\right)\\
&\le P\left(\left|\left(\min_{g\in D}\widehat{e}_g - \max_{g\in I}\widehat{e}_g\right) - \left(\min_{g\in D}e_g - \max_{g\in I}e_g\right)\right| > \frac{\delta}{2}\right) \le P\left(\max_{1\le g\le G}\left|\widehat{e}_g - e_g\right| > \frac{\delta}{4}\right) \qquad \text{(A.10)}\\
&\le O\left(RJ_g\right)p\,\exp\left\{-c_{11}\frac{n^{1-2\rho}}{R^4J_g^4}\right\} = O\left(\exp\left\{\log\,RJ_g + \log\,p - c_{11}\frac{n^{1-2\rho}}{R^4J_g^4}\right\}\right),
\end{aligned}
$$

where $c_{11} = \min\{c_5, c_9\}(\delta^2/4)$. Since $\log(RJ_g)/\log n = O(1)$, there exists a positive constant c_{12} such that $\log(RJ_g) \le c_{12}\log n$. Also, $(\max\{\log p, \log n\}\ R^4{J_g}^4/n^{1-2\rho}) = O(1)$ implies that $\log p \le (1/2)c_{11}(n^{1-2\rho}/R^4J_g^4)$ and $1/2c_{11}(n^{1-2\rho}/R^4J_g^4) \ge (c_{12}+2)\log n$ for large n. Then there exists an n_0 such that $\sum_{n=n_0}^{\infty}\exp\{\log RJ_g + \log p - c_{11}n^{1-2\rho}/R^4J_g^4\} \le \sum_{n=n_0}^{\infty}\exp\{c_{12}\log n - (1/2)c_{11}n^{1-2\rho}/R^4J_g^4\} \le \sum_{n=n_0}^{\infty}\exp\{c_{12}\log n - (c_{12}+2)\log n\} = \sum_{n=n_0}^{\infty}n^{-2} < \infty$. By Borel Contelli lemma, we have $\liminf_{n\longrightarrow\infty}\{\min_{g\varepsilon G}\widehat{e}_g - \max_{g\varepsilon I}\widehat{e}_g\} \ge \delta/2 > 0, a.s.$

Data Availability

The p53 in cell lines data used in the real data analysis was obtained from the R package grpregOverlap (https://github.com/YaohuiZeng/grpregOverlap.git).

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

The work was supported by National Natural Science Foundation of China (grant no. 71963008).

References

[1] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B*, vol. 70, no. 5, pp. 849–911, 2008.

[2] J. Fan, R. Samworth, and Y. Wu, "Ultrahigh dimensional feature selection: beyond the linear model," *Journal of Machine Learning Research*, vol. 10, no. 5, pp. 2013–2038, 2009.

[3] Q. Mai and H. Zou, "The kolmogorov filter for variable screening in high-dimensional binary classification," *Biometrika*, vol. 100, no. 1, pp. 229–234, 2013.

[4] H. Cui, R. Li, and W. Zhong, "Model-free feature screening for ultrahigh dimensional discriminant analysis," *Journal of the American Statistical Association*, vol. 110, no. 510, pp. 630–641, 2015.

[5] D. Huang, R. Li, and H. Wang, "Feature screening for ultrahigh dimensional categorical data with applications," *Journal of Business & Economic Statistics*, vol. 32, no. 2, pp. 237–244, 2014.

[6] L. Ni, F. Fang, and F. Wan, "Adjusted pearson chi-square feature screening for multi-classification with ultrahigh dimensional data," *Metrika*, vol. 80, no. 6–8, pp. 805–828, 2017.

[7] L. Ni and F. Fang, "Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification," *Journal of Nonparametric Statistics*, vol. 28, no. 3, pp. 515–530, 2016.

[8] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society B*, vol. 68, no. 1, pp. 49–67, 2006.

[9] L. Wang, G. Chen, and H. Li, "Group SCAD regression analysis for microarray time course gene expression data," *Bioinformatics*, vol. 23, no. 12, pp. 1486–1494, 2007.

[10] P. Breheny and J. Huang, "Penalized methods for bi-level variable selection," *Statistics and Its Interface*, vol. 2, no. 3, pp. 369–380, 2009.

[11] N. Zhou and J. Zhu, "Group variable selection via a hierarchical lasso and its oracle property," *Statistics and Its Interface*, vol. 3, no. 4, pp. 557–574, 2010.

[12] J. Huang, S. Ma, H. Xie, and C.-H. Zhang, "A group bridge approach for variable selection," *Biometrika*, vol. 96, no. 2, pp. 339–355, 2009.

[13] P. Breheny, "The group exponential lasso for bi-level variable selection: the group exponential lasso for bi-level variable selection," *Biometrics*, vol. 71, no. 3, pp. 731–740, 2015.

[14] Y. Niu, R. Zhang, J. Liu, and H. Li, "Group screening for ultra-high-dimensional feature under linear model," *Statistical Theory and Related Fields*, vol. 4, no. 1, pp. 43–54, 2020.

[15] W. C. Song and J. Xie, "Group feature screening via the *F* statistic," *Communications in Statistics—Simulation and Computation*, vol. 51, no. 4, pp. 1921–1931, 2019.

[16] D. Qiu and J. Ahn, "Grouped variable screening for ultra-high dimensional data for linear model," *Computational Statistics & Data Analysis*, vol. 144, Article ID 106894, 2020.

[17] H. He and G. Deng, "Grouped feature screening for ultra-high dimensional data for the classification model," *Journal of Statistical Computation and Simulation*, vol. 92, no. 5, pp. 974–997, 2021.

[18] L. P. Zhu, L. X. Li, R. Z. Li, and L. X. Zhu, "Model-free feature screening for ultrahigh-dimensional data," *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1464–1475, 2011.

[19] A. Subramanian, P. Tamayo, V. K. Mootha et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15545–15550, 2005.

[20] Y. H. Zeng and P. Breheny, "Overlapping group logistic regression with applications to genetic pathway selection," *Cancer Informatics*, vol. 15, pp. CIN.S40043–187, 2016.

[21] L. Ni, *Variable screening methods for ultra-high dimensional categorical covariates*, PhD Dissertation, East China Normal University, Shanghai, China, 2019.