Mathematics

AIMS Mathematics, 9(2): 4032–4056.
DOI: 10.3934/math.2024197
Received: 21 November 2023
Revised: 23 December 2023
Accepted: 05 January 2024
Published: 12 January 2024

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

Research article

Group feature screening for ultrahigh-dimensional data missing at random

Hanji He[1], **Meini Li**[2] **and Guangming Deng**[3,4,*]

[1] School of Economics and Finance, South China University of Technology, Guangdong 510006, China
[2] School of Mathematics and Computer Science, Chongqing College of International Business and Economics, Chongqing 401520, China
[3] School of Mathematics and Statistics, Guilin University of Technology, Guangxi 541000, China
[4] Applied Statistics Institute, Guilin University of Technology, Guangxi 541000, China

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

Abstract: Statistical inference for missing data is common in data analysis, and there are still widespread cases of missing data in big data. The literature has discussed the practicability of two-stage feature screening with categorical covariates missing at random (IMCSIS). Therefore, we propose group feature screening for ultrahigh-dimensional data with categorical covariates missing at random (GIMCSIS), which can be used to effectively select important features. The proposed method expands the scope of IMCSIS and further improves the performance of classification learning when covariates are missing. Based on the adjusted Pearson chi-square statistics, a two-stage group feature screening method is modeled, and theoretical analysis proves that the proposed method conforms to the sure screening property. In a numerical simulation, GIMCSIS can achieve better finite sample performance under binary and multivariate response variables and multi-classification covariates. The empirical analysis through multiple classification results shows that GIMCSIS is superior to IMCSIS in imbalanced data classification.

Keywords: group feature screening; sure screening property; chi-square statistic; missing data
Mathematics Subject Classification: 62H30, 62R07

1. Introduction

With the development of information technology and networks, big data with "high dimensionality" as its main feature are widely appearing in medicine, economics, engineering, and other fields. In general, when the dimension p of the covariable increases exponentially with sample size n, we call such data ultrahigh-dimensional data [1]. Intuitively, the covariate dimension of ultrahigh-dimensional data far exceeds the sample size, and the traditional penalty variable screening method suffers from high computational complexity, poor statistical accuracy, and weak algorithm stability; therefore, it is urgent to develop a variable screening method that is suitable for ultrahigh-dimensional data. However, missing data is also another major problem in data analysis and it exists in ultrahigh-dimensional data. Even missing a smaller number of samples for a covariate can lead to an inability to calculate the significance of that variable, which will lead to the omission of important information and misjudgment in subsequent analysis. Therefore, determining how to directly find important variables in ultrahigh-dimensional data with randomly missing data points plays an important role in efficient data analysis.

To solve the problem of statistically modeling ultrahigh-dimensional data, J. Fan and J. Lv [2] first proposed sure independence screening (SIS), which measures the importance of each covariate according to the Pearson correlation coefficient between the response variable and a single covariate. Thus, the dimension of the covariates is reduced to a suitable range. To improve the use of SIS under more general assumptions, P. Hall and H. Miller [3] extended the generalized correlation coefficients, and G. Li et al. [4] proposed the robust rank correlation coefficient screening method to address transformed regression models. X. Y. Wang and C. L. Leng [5] proposed a feature screening method based on high-dimensional least-squares projection to further improve screening performance, considering that SIS is highly dependent on significant covariates and response variables with large marginal correlations. For the model-free assumption, which is more suitable for the era of big data, L. P. Zhu et al. [6] proposed a feature screening method based on covariance, namely, sure independent ranking screening. As there is no model assumption, this method can be used for models such as linear models, generalized linear models, partial linear models, and single index models. R. Li et al. [7] proposed a feature screening method based on the distance correlation coefficient by employing distance covariance through the use of an index to describe whether two arbitrary variables are independent. On this basis, X. Shao and J. Zhang [8] proposed the use of the martingale difference correlation coefficient screening method to measure the deviation of the correlation between two random variables. On the topic of feature screening for ultrahigh-dimensional discrete data, Q. Mai and H. Zou [9] focused on binary response variables, introduced Kolmogorov-Smirnov statistics into the feature screening framework, and proposed a variable selection method based on the Kolmogorov filter. D. Huang et al. [10] proposed a feature screening method based on Pearson chi-square statistics that can solve the problems of superhigh-dimensional discrete covariate data and continuous data under multiple response variables. L. Ni et al. [11] further considered adjusted Pearson chi-square SIS (APC-SIS) to analyze multiclass response data. P. Lai et al. [12] introduced Fisher linear projection and marginal score tests to construct a linear projection feature screening method for linear discriminant modeling. With the emergence of group variables, the above feature screening methods that apply to single variables are no longer applicable. W. C. Song and J. Xie [13] proposed a feature screening method for group data based on F statistics. Based on the assumptions of a linear model, D. Qiu and J. Ahn [14] proposed three methods: group SIS, group high-dimensional least squares projection, and group adjusted R-square screening. H. J. He and G. M. Deng [15] focused on the feature screening of discrete group data and constructed the group information entropy feature screening method based on joint information entropy. Z. Z. Wang et al. [16,17] further extended the role of information theory in group feature screening by using the information gain ratio and Gini coefficient. Y. L. Sang and X. Dang [18] used the Gini distance coefficient to measure the independence between discrete response variables and continuous covariates and constructed a group feature screening method for the Gini distance.

Missing data constitutes a widespread problem in data analysis, and ultrahigh-dimensional data are no exception. P. Lai et al. [19] applied the Kolmogorov filtering method to screen important covariates for the construction of propensity score functions, proposed a feature screening method for ultrahigh-dimensional data with response variables missing at random, and promoted inverse probability weighting technology to build marginal feature screening processes. Q. H. Wang and Y. J. Li [20] proposed missing indicator interpolation screening and feature screening methods based on a Venn diagram by using missing indicator information for response variables. X. X. Li et al. [21] identified key covariates by using the marginal Spearman rank correlation coefficient for conditional estimation. L. Y. Zou et al. [22] used the interpolation technique to process the distribution function of missing responses and adopted the distance correlation between the response distribution function and the covariate distribution function as the index for feature screening. L. Ni et al. [23] proposed two-stage feature screening with covariates missing at random (IMCSIS) and proved the theoretical properties of this method based on adjusted Pearson chi-square statistical feature screening.

The response variable and the covariates missing at random have been fully discussed in the feature screening of ultrahigh-dimensional data. Considering that group data are common, they also need to be taken into account in the feature screening framework. We attempted to extend the group feature screening method that is typically applied to complete data to the case of covariates missing at random to expand the application scope of the existing group feature screening method. In addition, the ultrahigh-dimensional data considered in this paper are discrete, and the application scenarios are mostly classification learning objectives. Therefore, we used the adjusted Pearson chi-square statistic and the two-stage screening procedure to improve the effectiveness of multi-classification problems in practical problems.

In this paper, we construct an ultrahigh-dimensional group feature screening method for randomly missing data and extend the feature screening method that is generally suitable for classification models. First, we define the indicator variables for missing covariates, for which it is assumed that any missing variables exist as a group structure. Second, a two-stage group feature screening method with covariates missing at random (GIMCSIS) was proposed by introducing adjusted Pearson chi-square statistics as the basic screening method. In this paper, the GIMCSIS satisfies the sure screening performance requirement. Furthermore, the performance of GIMCSIS is demonstrated via numerical simulation and empirical analysis. Specifically, compared with IMCSIS, GIMCSIS can be applied to group data and improve the feature screening performance of ultrahigh-dimensional data with covariates missing at random. In the empirical analysis, we focus on the classification model, and GIMCSIS is better than IMCSIS in terms of various classification indices.

This paper is organized as follows. Section 2 introduces two-stage group feature screening based on adjusted Pearson chi-square statistics. Then, we establish a group sure screening property. The simulation studies are given in Section 3. Section 4 provides a classification analysis, and the paper concludes with a discussion in Section 5.

2. Theory and method

2.1. Symbol and definition

First, we define the group structure data. When covariates are randomly missing, there are group covariates among both the full and partial observation covariates; therefore, we need to define new symbols and concepts. Suppose that Y is a multiclass response variable with R elements and that the covariate matrix X is a multivariate covariate matrix with G group covariates, each of which consists of one or more covariates. Considering random missing covariates in covariate data, the set of fully observed variables is defined as $U = (u_1, \ldots, u_{G1})^T$, and the partial set of observed variables is defined as $V = (v_1, \ldots, v_{G2})^T$. The covariate matrix X is represented by

$$X = (u_1, \ldots, u_{G1}, v_1, \ldots, v_{G2})^T,\ P = \sum_{k=1}^{G1} p_k + \sum_{l=1}^{G2} q_l,\ G = G1 + G2,$$

where $1 \leq k \leq G1$, $1 \leq l \leq G2$, and $G1$ and $G2$ are the groups of fully and partially observed covariates, respectively. P represents the total number of dimensions in the covariate, p_k represents the number of dimensions of the kth covariate in the fully observed covariate, and q_l represents the number of dimensions of the lth covariate in the group of partially observed observation covariates. For some of the observed covariates, δ_l is used to represent the missing indicator variables. The missing state of a single variable has only 1 or 0 states, while the missing state of a group of variables is more complicated. In this paper, the missing indicator variable δ_l^* is defined as a fully observed covariate group only when all covariables in the group are as follows:

$$\delta_l^* = \begin{cases} 1, & sum(\delta_l) = q_l; \\ 0, & sum(\delta_l) \neq q_l. \end{cases}$$

Second, it is assumed that all covariate components of the covariate matrix X are classified as J, J_g represents the last of the combinations between covariate classes in the gth group covariate matrix, and j_g represents the indicator variables in the combinations between covariate classes in the gth group covariate matrix. If $j_g = 1$, it is the first covariate-class combination, and so on; J_g is the p_g covariate class combination; and a certain covariate class combination has a classification vector representation, namely, $\left(j_1, \ldots, j_{p_g}\right)$.

Let the probability function of the response variable be $p_r = P(Y = r)$. A probability function with a group structure covariate can be expressed as $w_{j_k} = w_{\left(j_1,\ldots,j_{p_k}\right)} = P\left(u_{k1} = j_1, \ldots, u_{kz} = j_z, \ldots, u_{kp_k} = j_{p_k}\right)$ and $w_{j_l} = w_{\left(j_1,\ldots,j_{p_l}\right)} = P\left(v_{l1} = j_1, \ldots, v_{lz} = j_z, \ldots, v_{lp_l} = j_{p_l}\right)$ represents a joint probability function of variables within a group; w_{j_k} is a joint probability function for complete data, and w_{j_l} is a form in the partial case of covariates. Similarly, the probability functions of the response variable with group structure covariates are as follows: $p_{J_k r} = p_{\left(j_1,\ldots,j_{p_k}\right)r} = P\left(Y = r, u_{k1} = j_1, \ldots, u_{kz} = j_z, \ldots, u_{kp_k} = j_{p_k}\right)$ and $p_{J_l r} = p_{\left(j_1,\ldots,j_{p_l}\right)r} = P\left(Y = r, v_{l1} = j_1, \ldots, v_{lz} = j_z, \ldots, v_{lp_l} = j_{p_l}\right)$, $1 \leq k \leq G1$, $1 \leq l \leq G2$, $r = 1, \ldots, R$, and $z = 1, \ldots, p_k\ or\ 1, \ldots, p_l, \left\{j_1, \ldots, j_{p_g}\right\} = 1, \ldots, J$.

2.2. Adjusted Pearson chi-square statistic

Based on the two-stage feature screening process, we propose a two-stage feature screening method for group structure data [23]. Specifically, we use APC-SIS to construct the group feature screening process.

The APC-SIS method first uses the adjusted Pearson chi-square statistic as the feature screening index [11]. The adjusted Pearson chi-square statistic for univariate analysis is given as follows:

$$\Delta_k = \frac{1}{log J_k}\sum_{r=1}^{R}\sum_{j=1}^{J_k}\frac{\left(p_r w_j^{(k)} - \pi_{r,j}^{(k)}\right)^2}{p_r w_j^{(k)}} \tag{2.1}$$

where $p_r = P(Y = r)$, $w_j^{(k)} = P(X_k = j)$ and $\pi_{r,j}^{(k)} = P(Y = r, X_k = j)$, $r = 1,2,\cdots,R$, $j = 1,2,\cdots,J$ and $k = 1,2,\cdots,K$. When the response variable Y is independent of the covariate X_k, the product of the marginal probabilities is equal to the joint probability; then, $p_r w_j^{(k)} = \pi_{r,j}^{(k)}$. When the response variable Y and the covariate X_k are not independent, the product of the marginal probabilities is not equal to the joint probability; the larger the difference, the stronger the correlation between Y and X_k. Therefore, it is easy to obtain two properties of Δ_k; then, $\Delta_k \geq 0$, and $\Delta_k = 0$ if and only if Y is independent of X_k.

By applying the above definition, we can obtain the adjusted Pearson chi-square statistics of a group of covariables under the random missing mechanism. For complete covariate data, we can directly construct the adjusted Pearson chi-square statistic of the response variable Y and the complete covariate U_k:

$$APC_g(Y, U_k) = \frac{1}{log\, J_k}\sum_{r=1}^{R}\sum_{j=1}^{J_k}\frac{\left(p_{J_k r} - p_r w_{j_k}\right)^2}{p_r w_{j_k}}. \tag{2.2}$$

For partial covariate data, the adjusted Pearson chi-square statistic of the response variable Y and the complete covariate V_l is calculated:

$$APC_g(Y, V_l) = \frac{1}{log\, J_l}\sum_{r=1}^{R}\sum_{j=1}^{J_l}\frac{\left(p_{j_l r} - p_r w_{j_l}\right)^2}{p_r w_{j_l}}. \tag{2.3}$$

Because $w_{j_l} = P\left(v_{l1} = j_1, \dots, v_{lz} = j_z, \dots, v_{lp_l} = j_{p_l}\right) = \sum_{r=1}^{R} P\left(Y = r, v_{l1} = j_1, \dots, v_{lz} = j_z, \dots, v_{lp_l} = j_{p_l}\right) = p_{J_l r}$, when estimating $APC_g(Y, V_l)$, compared with Eq (2.2), we need to estimate only $p_{J_l r}$. Then,

$$\begin{aligned}
p_{J_l r} &= P\left(Y = r, v_{l1} = j_1, \dots, v_{lz} = j_z, \dots, v_{lp_l} = j_{p_l}\right) \\
&= \sum_u P\left(Y = r, U^{M_{lg}} = u\right) \cdot P\left(v_{l1} = j_1, \dots, v_{lz} = j_z, \dots, v_{lp_l} = j_{p_l} \middle| Y = r, U^{M_{lg}} = u\right) \\
&= \sum_u P\left(Y = r, U^{M_{lg}} = u\right) \cdot P\left(v_{l1} = j_1, \dots, v_{lz} = j_z, \dots, v_{lp_l} = j_{p_l} \middle| Y = r, U^{M_{lg}} = u, \delta_l^* = 1\right) \\
&= \sum_u \frac{P\left(Y = r, U^{M_{lg}} = u\right) \cdot P\left(v_{l1} = j_1, \dots, v_{lz} = j_z, \dots, v_{lp_l} = j_{p_l}, Y = r, U^{M_{lg}} = u, \delta_l^* = 1\right)}{P\left(Y = r, U^{M_{lg}} = u, \delta_l^* = 1\right)}
\end{aligned} \tag{2.4}$$

where $U^{M_{l_g}}$ is used to link the fully observed covariate with the partially observed covariate, and the important fully observed covariate M_{l_g} can be obtained by calculating $APC_g(\delta_l^*, U_k)$; that is, the fully observed covariate associated with the missing information is used to replace the partially observed covariate. Therefore, $\hat{p}_{J_l r}$ can be obtained from $\widehat{M}_{l_g}$ and Eq (2.4).

2.3. Two-stage group feature screening method

The two-stage group feature screening with covariates missing at random (GIMCSIS) uses the missing indicator variable as a bridge between the partially observed covariate and the response variable. Feature screening of the fully observed covariates and partially observed covariates is carried out, and the fully observed covariate information is used to replace the partially observed covariate information to realize group feature screening of the partially observed covariates. The screening process is divided into two steps:

Step 1: To map the partially observed variable information to the fully observed covariates, the fully observed covariates associated with the missing indicator variable are considered, and the partially observed covariates are replaced by the information of the fully observed covariates. Specifically, for each of the observed covariates that are missing indicator variables, the adjusted Pearson chi-square statistic is calculated as follows:

$$\widehat{APC}_g(\delta_l^*, U_k) = \frac{1}{log\, J_k} \sum_{r=1}^{R} \sum_{j_k}^{J_k} \frac{\left(\sum_{i=1}^{n} I\left(\delta_{i,l}^* = r, u_{k1} = j_1, \ldots, u_{kp_k} = j_{p_k}\right) - \sum_{i=1}^{n} I(\delta_{i,l}^* = r)\widehat{w}_{j_k}\right)^2}{n \sum_{i=1}^{n} I(\delta_{i,l}^* = r)\widehat{w}_{j_k}} \tag{2.5}$$

where $w_{j_k} = w_{\left(j_1, \ldots, j_{p_k}\right)} = P\left(u_{k1} = j_1, \ldots, u_{kz} = j_z, \ldots, u_{kp_k} = j_{p_k}\right)$ and $r = 0,1$. The active covariates are estimated by applying the following thresholds:

$$\widehat{M}_{l_g} = \left\{k: \widehat{APC}_g(\delta_l^*, U_k) > c_{\delta_l^*} n^{-\tau_{\delta_l^*}}, 1 \le k \le G_1\right\}$$

where $c_{\delta_l^*}$ and $\tau_{\delta_l^*}$ are predetermined constants that are defined in Condition (C4) in Section 2.4.

Step 2: On the basis of obtaining $\widehat{M}_{l_g}$, using $U^{M_{l_g}}$ to replace partially observed covariates, the adjusted Pearson chi-square statistics of the response variable and partially observed covariates are obtained via the following process:

$$\widehat{APC}_g(Y, V_l) = \frac{1}{log\, J_l} \sum_{r=1}^{R} \sum_{j=1}^{J_l} \frac{\left(\hat{p}_{j_l r} - \hat{p}_r \widehat{w}_{j_l}\right)^2}{\hat{p}_r \widehat{w}_{j_l}} \tag{2.6}$$

where

$$\hat{p}_{j_l r} = \frac{1}{n} \sum_{u} \frac{\sum_{i=1}^{n} I\left(y_i = r, u_i^{\widehat{M}_{l_g}} = u\right) \sum_{i=1}^{n} I\left(v_{l1} = j_1, \ldots, v_{lp_l} = j_{p_l}, y_i = r, u_i^{\widehat{M}_{l_g}} = u, \delta_{i,l}^* = 1\right)}{\sum_{i=1}^{n} I\left(y_i = r, u_i^{\widehat{M}_{l_g}} = u, \delta_{i,l}^* = 1\right)}$$

$\widehat{w}_{j_l} = \sum_{r=1}^{R} \hat{p}_{j_l r}$ and $\hat{p}_r = n^{-1} \sum_{i=1}^{n} I(y_i = r)$. The sum of all $\hat{p}_{j_l r}$ values is equivalent to all possible values of u in the set $U^{\widehat{M}_l}$. In practice, the summation term $\sum_{i=1}^{n} I\left(y_i = r, u_i^{\widehat{M}_{l_g}} = u, \delta_{i,l}^{*} = 1\right)$ in the case of a given u value is 0 when the number of covariates in $\widehat{M}_l$ is large enough. Thus, $log\, J_l$ is used to adjust the Pearson chi-square statistics, yielding $\sum_{r=1}^{R}\sum_{j=1}^{J_l} \hat{p}_{j_l r} = 1$.

For the fully observed covariates, we obtain the active covariate directly by using the adjusted Pearson chi-square statistic. For the partially observed covariates, we obtain the active covariate according to Steps 1 and 2. Therefore, the active covariates in the dataset can be estimated as follows:

$$(U,V)^{\widehat{D}} = \left\{U_k, V_l \colon \widehat{APC}_g(Y,U_k) > cn^{-\tau}, \widehat{APC}_g(Y,V_l) > cn^{-\tau}, 1 \le k \le p, 1 \le l \le q\right\}$$

where c and τ are predetermined constants.

In practice, we replace $\widehat{M}_{l_g}$ with

$$\widehat{M}_{l_g}^{*} = \left\{k \colon APC_g(\delta_l^{*}, U_k) \text{ is the largest } d_l \text{ of all } U_k\right\}$$

and replace $(U,V)^{\widehat{D}}$ with the following method:

$$(U,V)^{\widehat{D}^{*}} = \{U_k, V_l \colon APC_g(Y,U_k) \quad \text{or} \quad APC_g(Y,V_l) \text{ is the largest } d \text{ of all } U_k \text{ or } V_l\}$$

2.4. Sure screening property

Next, we establish the theoretical property of the proposed GIMCSIS. For feature screening, sure screening properties are essential and they were proposed by J. Fan and J. Lv [2]. After applying a feature screening procedure with a probability tending to 1, all of the important variables still survive. It is important to identify the conditions under which the sure screening property holds, i.e.,

$$P\left(\mathcal{M}_{*} \subseteq \mathcal{M}_{\gamma}\right) \to 1 \qquad as\, n \to \infty$$

where $\mathcal{M}_{\gamma}$ is the final model after feature screening and $\mathcal{M}_{*}$ is the true model.

Therefore, to explore the sure screening property of GIMCSIS, the following regularity conditions are assumed.

(1) There are two positive constants c_1 and c_2, such that $\frac{c_1}{R} \le p_r \le \frac{c_2}{R}$, $0 \le w_{j_g}^{U_k} \le \frac{c_2}{J}$ and $0 \le w_{j_g}^{V_l} \le \frac{c_2}{J}$; for $r = 1, \dots, R$, $j = 1, \dots, J_{U_k}$, $l = 1, \dots, q, k = 1, \dots, p$, and $J = max_{1 \le k \le p, 1 \le l \le q}\left\{J_{U_k}, J_{V_l}\right\}$.

(2) There are two constants $c > 0$ and $0 < \tau < \frac{1}{2}$, such that

$$\min_{U_k, V_l \in (U,V)^D} \left\{APC_g(Y,U_k), APC_g(Y,V_l)\right\} > 2cn^{-\tau}.$$

(3) There are two positive constants c_3 and c_4, such that $0 < c_3 \le P(\delta_l^{*} = r) \le c_4 < 1$; for $r = 0,1$, $l = 1, \dots, q$.

(4) For each $1 \le l \le q$, there are two constants $c_{\delta_l^{*}} > 0$ and $0 < \tau_{\delta_l^{*}} < \frac{1}{2}$, such that

$$min_{k\in M_l} APC_g\,(\delta_l^*, U_k) > \frac{3}{2} c_{\delta_l} n^{-\tau_{\delta_l^*}}$$

where $APC_g(\delta_l^*, U_k) = (log\, J_k)^{-1} \sum_{j=1}^{J_g} \sum_{r=0}^{1} \{P(\delta_l^* = r, U_k = j) - P(\delta_l^* = r)P(U_k = j)\}^2 / P(\delta_l^* = r)/P(U_k = j)$.

(5) There are two positive constants c_5 and c_6, such that $\frac{c_5}{2RJ^{\bar{m}}} \leq P\left(Y = r, U^{\bar{M}_l} = u, \delta_l^* = 1\right) \leq \frac{c_6}{2RJ^{\bar{m}}}$, where $\bar{M}_l = \left\{k: APC_g(\delta_l^*, U_k) > \frac{1}{2} c_{\delta_l^*} n^{-\tau_{\delta_l^*}}, 1 \leq k \leq G1\right\}$, $\bar{m} = max_{1\leq l\leq q} \left\|U^{\bar{M}_l}\right\|_0$, $r = 1, \dots, R$, and $l = 1, \dots, q$.

(6) $R = O\left(n^{\xi}\right)$ and $J = O(n^{\kappa})$, where $\xi \geq 0$, $\kappa \geq 0$, $1 - 2\tau - 6\xi - 18\kappa > 0$, $1 - 2\tau_\delta - 18\kappa > 0$, $1 - 2\tau - 10\xi - (18\bar{m} + 18)\kappa > 0$, and $\tau_\delta = max_{1\leq l\leq G2}\, \tau_{\delta_l^*}$.

Conditions (1)–(5) are commonly used in the study of feature screening [2,7,10,11,23]. Condition (1) ensures that the proportion of each class of response variables and covariates is not too small or too large, i.e., that the class of variables is balanced. Condition (2) requires that the smallest real signal converges to zero at the $n^{-\tau}$ rate at which the sample size reaches infinity. Condition (3) requires that the missing proportion be bounded. Condition (4) ensures the sure screening performance of the APC-SIS in the first step of screening. Condition (5) ensures that the denominator of Eq (2.4) is not 0, and the such a condition can be satisfied when the magnitude of $\bar{M}_l$ is small. Condition (6) requires that the divergence rate be much less than the growth rate of n.

Theorem 1. (Sure screening property) Under Conditions (1)–(6), if $log\, p = O(n^{\alpha})$, $log\, q = O\left(n^{\beta}\right)$, $\alpha < 1 - 2\tau - 6\xi - 18\kappa$, $\beta < 1 - 2\tau - 10\xi - (18\bar{m} + 18)\kappa$ and $\alpha + \beta < 1 - 2\tau_\delta - 18\kappa$, such that

$$P\left((U,V)^D \subseteq (U,V)^{\hat{D}}\right) \geq 1 - O\left(pexp\left(-b_1 n^{1-2\tau-6\xi-18\kappa} + (\xi + \kappa)logn\right)\right)$$

$$-O\left(pqexp\left(-b_2 n^{1-2\tau_\delta-18\kappa} + (\xi + 2\kappa)logn\right)\right)$$

$$-O\left(qexp\left(-b_3 n^{1-2\tau-10\xi\ \ (18\bar{m}+18)\kappa} + (\xi + (\bar{m} + 1)\kappa)logn\right)\right)$$

where b_1, b_2 and b_3 are constants. Therefore, GIMCSIS has the sure screening property.

Remark 1. To explore the feature screening of missing covariates in ultra-high dimensional data with group structuring, it is easier to make better use of the information of covariates missing at random. Inspired by L. Ni et al. [23], we propose group feature screening for ultrahigh-dimensional data with categorical covariates missing at random (GIMCSIS). GIMCSIS expands the scope of IMCSIS, and it further improves the performance of classification learning. Regarding the screening performance theory, compared to IMCSIS, GIMCSIS has a higher probability of screening important variables (see Theorem 1). Theorem 1 is also confirmed in Section 3.

3. Simulation studies

To verify the feature screening performance of GIMCSIS, we generated a series of simulation data for relevant experiments. Simulations 1 and 2 are compared with IMCSIS [23] from two perspectives: The binary response variable and the multiclass response variable. The computer configuration is as follows: CPU, Intel i5-3230M (2.6 GHz); memory, 16 GB; and operating system, Windows 10. The feature screening was implemented by using R version 4.2.2 programming, and the

RStudio interactive programming interface was used.

The metrics used to evaluate the performance of feature screening include the following steps:

Index	Description
MMS (minimum model size to include all active covariates)	The position of the last active covariate among all active covariates is usually represented by the 5%, 25%, 50%, 75%, and 95% quantiles of the MMS obtained from multiple experiments, which is used to illustrate that when MMS is similar to the number of active covariates, the model results are better.
CP (coverage probability)	The proportion of active covariates covered in a certain interval accounted for all active covariates, and the interval was used to calculate the coverage probability in many studies with three strong and weak intervals $[n/log(n)]$,$2\cdot[n/log(n)]$, $3\cdot[n/log(n)]$; also, the coverage probability under the three intervals was defined as CP1, CP2 and CP3 to illustrate the convergence performance of the model.
CPa (all-coverage probability)	The probability of all active variables in the interval $3\cdot[n/log(n)]$ in multiple experiments.

3.1. Simulation 1: Binary responses

On the basis of complete covariates, 40% of the covariates were defined as partially observed covariates. A simple model in which all covariables are multiclass and the response variables are binary is defined. The settings for the response variables y_i, latent variables z_i and covariables x_i are as follows:

	Response variable y_i
Balanced data	$p_r = P(y_i = 1) = P(y_i = 2) = 1/2$
Unbalanced data	$p_r = 2\left[1 + \frac{(R-r)}{(R-1)}\right]/3R$ with $max_{1\le r\le R}\, p_r = 2\, min_{1\le r\le R}\, p_r$
	Latent variable z_i, $z_{i,k} \sim N(\mu_{rk}, 1)$, $1 \le k \le p$.
$k > d_0$	$\mu_{rk} = 0$
$k \le d_0$ and$r = 1$	$\mu_{rk} = -0.5$ (active covariate)
$k \le d_0$ and$r = 2$	$\mu_{rk} = 0.5$ (active covariate)
	Covariable x_i
k is odd	$x_{i,k} = I\left(z_{i,k} > z_{(j/2)}\right) + 1$
k is even	$x_{i,k} = I\left(z_{i,k} > z_{(j/5)}\right) + 1$

where d_0 is the size of the active covariables and the first to tenth covariates are active covariates. In the IMCSIS method, the indicator of the active covariate is $d_0 = 10$. In the GIMCSIS method, the number of variables in each group is 3, and the indicator of the active covariate is $d_{0G} = 4$. When $\mu_{rk} = -0.5$ or $\mu_{rk} = 0.5$, the kth covariable is active. $z_{(\alpha)}$ is the αth standard normal distribution quantile, and it is used to discretize the covariates x_i.

Therefore, for a p-dimensional covariate, half of the covariates are categorical variables and the other half are categorical variables. The ratio of complete covariates to partial covariates is defined as 6:4; the first 60% of the covariates make up complete data, and the others constitute missing data. The random missing proportions mp were set to 10%, 25% and 40%, and the missing indicator variable $\delta_{i,l}$ was generated from the Bernoulli distribution $\delta_{i,l} \sim B(1,1-mp)$. The dimensions of the covariates were set to 1000, the full covariates to 600, the partial covariates to 400, and the sample sizes to 100, 120, and 150, (see Table 1).

Table 1. Results for binary response variables and multi-classification covariates.

mp		Method	CP1	CP2	CP3	CPa	mms.5.	mms.25.	mms.50.	mms.75.	mms.95.
n=100, P=1000, p=600, q=400, balanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.00**
		IMCSIS	0.85	0.90	0.90	0.04	55.90	73.25	99.00	119.25	140.05
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**5.55**
		IMCSIS	0.79	0.92	0.95	0.56	21.90	30.00	41.50	61.75	105.40
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.00**
		IMCSIS	0.85	0.90	0.90	0.04	55.90	73.25	99.00	119.25	140.05
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**6.10**
		IMCSIS	0.72	0.87	0.92	0.36	25.45	41.25	63.00	87.50	156.40
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.00**
		IMCSIS	0.85	0.90	0.90	0.04	55.90	73.25	99.00	119.25	140.05
	Step 2	GIMCSIS	**0.99**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**13.65**
		IMCSIS	0.62	0.78	0.85	0.10	44.80	61.00	102.50	136.00	187.75
n=120, P=1000, p=600, q=400, balanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	0.94	1.00	1.00	1.00	13.45	17.00	19.50	22.00	27.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**5.00**	**6.00**	**7.00**
		IMCSIS	0.89	0.98	0.99	0.94	15.45	18.25	24.00	29.00	52.30
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	0.94	1.00	1.00	1.00	13.45	17.00	19.50	22.00	27.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.75**	**8.55**
		IMCSIS	0.80	0.94	0.98	0.80	18.90	27.25	35.00	49.75	81.65
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	0.94	1.00	1.00	1.00	13.45	17.00	19.50	22.00	27.00
	Step 2	GIMCSIS	**0.97**	**0.99**	**0.99**	**0.96**	**4.00**	**4.00**	**6.00**	**11.75**	**29.05**
		IMCSIS	0.68	0.86	0.92	0.34	33.90	43.50	72.00	103.00	151.50
n=150, P=1000, p=600, q=400, balanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	1.00	1.00	1.00	1.00	10.00	10.00	10.50	11.00	12.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	0.99	1.00	1.00	1.00	12.00	14.00	15.50	18.00	23.55
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	1.00	1.00	1.00	1.00	10.00	10.00	10.50	11.00	12.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.55**
		IMCSIS	0.93	0.99	1.00	1.00	13.45	17.00	23.00	30.00	46.20

Continued on next page

mp		Method	CP1	CP2	CP3	CPa	mms.5.	mms.25.	mms.50.	mms.75.	mms.95.
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	1.00	1.00	1.00	1.00	10.00	10.00	10.50	11.00	12.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**9.00**
		IMCSIS	0.84	0.95	0.98	0.78	18.45	25.00	34.50	57.25	124.90
n=100, P=1000, p=600, q=400, unbalanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.55**
		IMCSIS	0.77	0.85	0.88	0.00	63.00	84.50	120.00	159.75	181.55
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**5.00**	**6.00**	**8.00**
		IMCSIS	0.71	0.89	0.94	0.52	22.90	40.25	47.00	66.00	82.50
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.55**
		IMCSIS	0.77	0.85	0.88	0.00	63.00	84.50	120.00	159.75	181.55
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.55**
		IMCSIS	0.65	0.82	0.89	0.28	29.35	46.00	68.50	94.50	210.25
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.55**
		IMCSIS	0.77	0.85	0.88	0.00	63.00	84.50	120.00	159.75	181.55
	Step 2	GIMCSIS	**0.94**	**0.98**	**1.00**	**1.00**	**5.00**	**6.25**	**9.00**	**15.50**	**33.20**
		IMCSIS	0.53	0.75	0.83	0.10	40.90	65.75	96.00	157.75	245.85
n=120, P=1000, p=600, q=400, unbalanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	0.99	1.00	1.00	1.00	11.00	13.00	14.00	15.75	19.55
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**5.00**	**6.00**	**8.00**
		IMCSIS	0.82	0.97	0.99	0.94	20.00	23.00	28.50	36.50	62.45
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	0.99	1.00	1.00	1.00	11.00	13.00	14.00	15.75	19.55
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**11.10**
		IMCSIS	0.74	0.92	0.97	0.74	21.45	29.00	38.50	54.50	77.50
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	0.99	1.00	1.00	1.00	11.00	13.00	14.00	15.75	19.55
	Step 2	GIMCSIS	**0.96**	**0.98**	**0.99**	**0.96**	**4.00**	**5.25**	**8.00**	**17.00**	**38.65**
		IMCSIS	0.61	0.81	0.89	0.24	37.70	55.50	72.00	94.00	206.60
n=150, P=1000, p=600, q=400, unbalanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.55**
		IMCSIS	1.00	1.00	1.00	1.00	10.00	10.00	10.00	11.00	12.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	0.92	0.99	1.00	0.98	17.00	20.00	24.50	33.25	50.55
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.55**
		IMCSIS	1.00	1.00	1.00	1.00	10.00	10.00	10.00	11.00	12.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	0.86	0.97	0.99	0.86	15.45	22.25	31.50	44.25	148.40
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.55**
		IMCSIS	1.00	1.00	1.00	1.00	10.00	10.00	10.00	11.00	12.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.00**
		IMCSIS	0.79	0.90	0.94	0.52	23.00	40.00	61.00	101.75	261.25

Table 1 reports the values of each index in 50 experiments simulated with a normal distribution. The active covariate size based on the GIMCSIS model is 4, and the active covariate size based on the IMCSIS model is 10. In the first stage, regardless of the missing proportions, GIMCSIS outperforms IMCSIS. In the second stage, the covariate is affected by random missing data, with the following results:

(1) Comparison of different sample sizes: As the sample size increases, the minimum model size (MMS) of GIMCSIS approaches the active covariate size $d_{0G} = 4$, and that of IMCSIS approaches the active covariate size $d_0 = 10$. However, GIMCSIS tends to converge faster than the set of active covariates, and the quantile of the MMS is almost equal to 4 when n = 150, while the quantile of the MMS for IMCSIS is significantly larger than the size of the active covariates. In both models, the four indicators covering the probability tend to be equal to 1. However, the coverage probability of the IMCSIS model is much weaker than that of the GIMCSIS model. GIMCSIS can achieve a better coverage probability when n = 100, while IMCSIS can achieve a better coverage probability when n = 150.

(2) Comparison of different response variables: The structure of the response variables considers both balanced and unbalanced data and is used to compare the anti-interference capacities of the methods. In general, the performance of finite samples under balanced data is better than that under unbalanced data. Among them, GIMCSIS with unbalanced data can achieve excellent MMS and coverage probability under the condition of n = 50, while IMCSIS has better screening performance under the condition of n = 150. Furthermore, GIMCSIS has stronger anti-interference capacities than IMCSIS.

(3) Comparison of different missing proportions: As the missing data proportion increases, the MMS quantiles of both methods decrease. The MMS variation in IMCSIS is larger than that in GIMCSIS. Moreover, the coverage probability of IMCSIS decreases significantly, while the coverage probability of GIMCSIS remains as 1. The above results indicate that the performance of IMCSIS decreases rapidly, while that of GIMCSIS is relatively stable when the missing data proportion continues to increase.

In summary, GIMCSIS has better ability to screen active covariates than IMCSIS in ultrahigh-dimensional data with binary response variables and covariates missing at random. The screening performance of IMCSIS decreases significantly in the second stage of screening, and the smaller the sample size, the worse the screening performance. The performance of GIMCSIS in the second stage of screening is similar to that in the first stage of screening, and it maintains a high coverage probability. For unbalanced data, the performance of both GIMCSIS and IMCSIS methods decreases, but GIMCSIS is significantly robust.

3.2. Simulation 2: Multiclass responses

Consider a complex model with more classes of covariates and four classes of response variables. Two y_i distributions are considered the same as those in Simulation 1.

The first to the 11th covariates are active covariates. In the IMCSIS method, the indicator of the active covariate is $d_0 = 11$; in the GIMCSIS method, when the number of variables in each group is 3, the indicator of the active covariate is $d_{0G} = 4$, which is the same as the Simulation 1 screening target, but the composition of group variables is different. In the case of y_i, Simulation 2 for the generation of latent variable data and active covariates is the same as Simulation 1.

Therefore, the p-dimensional covariates are evenly divided into two and five classes, respectively. The ratio of complete covariates to partial covariates is defined as 6:4; the first 60% of the covariates make up complete data, and the others constitute partial data. The random missing proportions mp are set to 10%, 25% and 40%, and the missing indicator variable $\delta_{i,l}$ is generated from the Bernoulli distribution $\delta_{i,l} \sim B(1, 1 - mp)$. The number of dimensions of the covariate are considered to be 1000 dimensions, where the full covariate dimension is 600 dimensions, the partial covariate dimension is 400 dimensions, and the sample numbers are 50, 80 and 100, (see Table 2).

Table 2. Results for multivariate response variables and multi-classification covariates.

mp		Method	CP1	CP2	CP3	CPa	mms.5.	mms.25.	mms.50.	mms.75.	mms.95.
n=50, P=1000, p=600, q=400, balanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	0.82	1.00	1.00	1.00	11.00	11.00	11.00	11.00	11.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.55**
		IMCSIS	0.65	0.87	0.94	0.60	13.00	17.25	26.00	35.00	64.55
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	0.82	1.00	1.00	1.00	11.00	11.00	11.00	11.00	11.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**7.55**
		IMCSIS	0.56	0.76	0.86	0.24	14.35	31.00	42.00	63.00	211.50
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	0.82	1.00	1.00	1.00	11.00	11.00	11.00	11.00	11.00
	Step 2	GIMCSIS	**0.78**	**0.90**	**0.91**	**0.66**	**4.00**	**8.25**	**12.00**	**46.00**	**103.50**
		IMCSIS	0.46	0.63	0.72	0.04	32.45	54.25	86.50	151.00	259.10
n=80, P=1000, p=600, q=400, balanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	1.00	1.00	1.00	1.00	11.00	11.00	11.00	11.00	11.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	0.98	1.00	1.00	1.00	11.00	11.00	12.00	13.00	16.55
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	1.00	1.00	1.00	1.00	11.00	11.00	11.00	11.00	11.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	0.93	0.99	1.00	1.00	11.00	11.25	13.00	19.00	27.00
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	1.00	1.00	1.00	1.00	11.00	11.00	11.00	11.00	11.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**5.00**
		IMCSIS	0.85	0.98	0.99	0.92	11.00	14.00	21.00	25.00	40.55
n=100, P=1000, p=600, q=400, balanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	1.00	1.00	1.00	1.00	11.00	11.00	11.00	11.00	11.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	1.00	1.00	1.00	1.00	11.00	11.00	11.00	12.00	12.00
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	1.00	1.00	1.00	1.00	11.00	11.00	11.00	11.00	11.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	0.99	1.00	1.00	1.00	11.00	11.00	11.00	12.00	15.55

Continued on next page

mp		Method	CP1	CP2	CP3	CPa	mms.5.	mms.25.	mms.50.	mms.75.	mms.95.
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	1.00	1.00	1.00	1.00	11.00	11.00	11.00	11.00	11.00
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**4.00**	**4.00**
		IMCSIS	0.99	1.00	1.00	1.00	11.00	12.00	13.00	15.00	20.55
n=50, P=1000, p=600, q=400, unbalanced											
10%	Step 1	GIMCSIS	**0.88**	**1.00**	**1.00**	**1.00**	**5.45**	**7.00**	**9.00**	**11.00**	**14.55**
		IMCSIS	0.35	0.53	0.60	0.00	218.90	277.25	360.00	411.50	455.75
	Step 2	GIMCSIS	**0.87**	**0.99**	**1.00**	**1.00**	**5.00**	**7.00**	**9.50**	**12.75**	**17.55**
		IMCSIS	0.34	0.41	0.47	0.00	177.75	226.50	268.00	348.75	387.95
25%	Step1	GIMCSIS	**0.88**	**1.00**	**1.00**	**1.00**	**5.45**	**7.00**	**9.00**	**11.00**	**14.55**
		IMCSIS	0.35	0.53	0.60	0.00	218.90	277.25	360.00	411.50	455.75
	Step2	GIMCSIS	**0.63**	**0.87**	**0.93**	**0.78**	**7.45**	**12.00**	**17.00**	**25.00**	**82.65**
		IMCSIS	0.29	0.36	0.42	0.00	187.00	248.00	286.00	323.50	379.85
40%	Step 1	GIMCSIS	**0.88**	**1.00**	**1.00**	**1.00**	**5.45**	**7.00**	**9.00**	**11.00**	**14.55**
		IMCSIS	0.35	0.53	0.60	0.00	218.90	277.25	360.00	411.50	455.75
	Step 2	GIMCSIS	**0.13**	**0.28**	**0.49**	**0.02**	**34.00**	**50.00**	**75.00**	**141.00**	**246.80**
		IMCSIS	0.26	0.31	0.36	0.00	153.00	254.25	300.50	359.75	394.10
n=80, P=1000, p=600, q=400, unbalanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**5.00**	**5.00**	**7.00**
		IMCSIS	0.64	0.81	0.89	0.08	37.90	63.25	86.00	111.75	183.30
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.00**
		IMCSIS	0.52	0.68	0.75	0.02	44.90	61.25	84.00	138.75	227.05
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**5.00**	**5.00**	**7.00**
		IMCSIS	0.64	0.81	0.89	0.08	37.90	63.25	86.00	111.75	183.30
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**5.00**	**7.00**	**11.55**
		IMCSIS	0.44	0.61	0.72	0.02	49.25	90.50	147.00	218.00	321.05
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**5.00**	**5.00**	**7.00**
		IMCSIS	0.64	0.81	0.89	0.08	37.90	63.25	86.00	111.75	183.30
	Step 2	GIMCSIS	**0.84**	**0.94**	**0.98**	**0.92**	**7.00**	**9.00**	**11.00**	**26.00**	**44.10**
		IMCSIS	0.38	0.50	0.60	0.00	90.45	123.50	192.50	277.25	353.35
n=100, P=1000, p=600, q=400, unbalanced											
10%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.55**
		IMCSIS	0.75	0.82	0.88	0.02	59.15	79.00	100.50	126.75	186.75
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.55**
		IMCSIS	0.68	0.85	0.93	0.36	27.45	42.00	53.50	65.00	111.10
25%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.55**
		IMCSIS	0.75	0.82	0.88	0.02	59.15	79.00	100.50	126.75	186.75
	Step 2	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**8.55**
		IMCSIS	0.59	0.77	0.86	0.14	38.80	59.25	83.50	126.00	215.95
40%	Step 1	GIMCSIS	**1.00**	**1.00**	**1.00**	**1.00**	**4.00**	**4.00**	**4.00**	**5.00**	**6.55**
		IMCSIS	0.75	0.82	0.88	0.02	59.15	79.00	100.50	126.75	186.75
	Step 2	GIMCSIS	**0.98**	**1.00**	**1.00**	**1.00**	**4.00**	**5.00**	**7.00**	**9.00**	**25.10**
		IMCSIS	0.49	0.66	0.77	0.02	59.95	85.75	137.50	198.75	328.85

Table 2 reports the values of each index in 50 experiments simulated with a normal distribution. The active covariate size based on the GIMCSIS model is 4, and the active covariate size based on the IMCSIS model is 11. In the first stage, GIMCSIS outperforms IMCSIS, regardless of the missing proportions. In the second stage, covariates are affected by random missing data, with the following results:

(1) Comparison of different sample sizes: As the sample size increases, the MMS of GIMCSIS approaches the active covariate size $d_{0G} = 4$, and the MMS of IMCSIS approaches the active covariate size $d_0 = 11$. However, GIMCSIS tends to converge faster than the set of active covariates, and the quantile of the MMS is almost equal to 4 when n = 100, while the quantile of the MMS in IMCSIS is significantly larger than the size of the active covariates. In contrast to the simulation results for binary response variables, the coverage probability of IMCSIS is much lower than that of binary response variables when the four indices all converge to 1, and the active variables that were screened out have an obvious trailing phenomenon. Although GIMCSIS also exhibited a similar phenomenon, the four coverage probability indicators are almost 1 when n = 80.

(2) Comparison of different response variables: The structure of the response variables considers both balanced and unbalanced data and is used to compare the anti-interference capacities of the methods. In general, the performance of finite samples under balanced data is better than that under unbalanced data. Among them, GIMCSIS of unbalanced data can achieve a good MMS and coverage probability when n = 80, while the coverage probability of IMCSIS is far from the qualified requirement even when n = 100. Furthermore, GIMCSIS has stronger anti-interference capacities than IMCSIS.

(3) Comparison of different missing proportions: As the missing data proportion increases, the quantile of the MMS for both methods decreases. The amplitude of the MMS variation in the IMCSIS dataset is larger than that in the GIMCSIS dataset, and the 75% and 95% quantiles of the IMCSIS dataset are too far from the range of active covariates. Moreover, the coverage probability of IMCSIS decreases significantly, while the coverage probability of GIMCSIS remains as 1. The above results indicate that the performance of IMCSIS decreases rapidly, while that of GIMCSIS is relatively stable when the missing data proportion continues to increase. Therefore, GIMCSIS can obtain stable screening results more effectively.

(4) Comparison with binary response variables: In Simulation 1, although the performance of IMCSIS is not as good as that of GIMCSIS, it can achieve better performance when there is a large sample size. However, in Simulation 2, the response variable is only increased from binary to four categories, and IMCSIS cannot achieve the screening goal. This shows that GIMCSIS has unique advantages in the case of more general multiple response variables.

In summary, GIMCSIS also exhibits better screening performance than IMCSIS under the conditions of ultrahigh-dimensional data with a multivariate response and covariates missing at random. The screening performance of GIMCSIS remains robust. Compared with IMCSIS, GIMCSIS has advantages in terms of a small sample size, an unbalanced response and a high deletion rate.

4. Empirical analysis

Section 3 illustrates the performance of GIMCSIS on simulated data. In practical application, whether important variables obtained via GIMCSIS can play a role in data analysis is an important issue that needs further verification. In this section, we describe the application of GIMCSIS to the predata analysis process of imbalanced data classification to test whether the important variables obtained can improve the effectiveness of classification problems.

The empirical data were obtained from the Arizona State University feature selection database (http://featureselection.asu.edu/), which contains colon cancer data consisting of 62 instances and 2000 covariates. Forty samples were negative for colon cancer, and the other 22 samples were positive for colon cancer, for an imbalance of 1.81:1. The 2000 covariates were based on the expression levels, and 2000 out of the 6500 genes were differentially expressed. Thus, the response is binary, and the covariates are continuous. There are group correlations between gene expression.

To evaluate the effectiveness of the various feature screening methods for classification problems, the average accuracy rate of the evaluation indicators was calculated via fivefold cross-validation. First, 62 samples were randomly divided into two groups according to the ratio of 1:4. Eighty percent of the samples were used as training data, and the rest were used as test data. The sample size of the training data was 50, the sample size of the test data was 12, and the covariate dimension of both datasets was 2000. The feature screening methods used were GIMCSIS and IMCSIS, where the number of active covariables in the univariate feature screening was $d_0 = 22$ and the number of active covariables in the group variable feature screening was $d_{0G} = 8$. Considering the effect of covariates on classification, we chose three classification models: Support vector machine (SVM) [24], decision tree (DT) [25] and k-nearest neighbor (KNN) [26].

The evaluation indices for the classification effect are derived from the confusion matrix, and the details are as follows. The confusion matrix is as below.

		Actual		
		Positive	Negative	Total
Prediction	Positive	TP	FP	TP+FP
	Negative	FN	TN	FN+TN
	Total	TP+FN	FP+TN	TP+FP+FN+TN

where TP is the true positive, FN is the false negative, FP is the false positive and TN is the true negative. Table 3 shows all of the evaluation indices based on the confusion matrix.

Table 3. Description of evaluation index.

Index	Description
$Accuracy$	$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$
$Precision$	$Precision = \frac{TP}{TP + FP}$
$Recall$	$Recall = \frac{TP}{TP + FN}$
$Specificity$	$Specificity = \frac{TN}{TN + FP}$
$G - mean$	$G - mean = \sqrt{Recall \times Specificity}$
$F - measure$	$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$

Table 4 reports the classification performance of the various feature screening methods on the training and test data. Overall, the classification effect of GIMCSIS was better than that of IMCSIS, with outstanding performance in terms of the recall, G-mean and F-measure. The classification method with the best classification effect was KNN. The average accuracy on the training set based on GIMCSIS was greater than 98%, and the average accuracy on the test set was greater than 88%. According to the test data, the G-mean and F-measure based on GIMCSIS were superior to those based on IMCSIS. Regarding the SVM, the average evaluation index for the GIMCSIS dataset was 1.64% greater than that for the IMCSIS dataset on the training data, and the average evaluation index for the GIMCSIS dataset was 5.84% greater than that for the IMCSIS dataset on the test data, indicating that the classification performance of the GIMCSIS dataset was better than that of the IMCSIS dataset. For DT, the average evaluation index for the GIMCSIS dataset was only 0.22% greater than that for the IMCSIS dataset on the training data, but the average evaluation index for the GIMCSIS dataset was 9.55% greater than that for the IMCSIS dataset on the test data, indicating that IMCSIS exacerbated the overfitting phenomenon. For KNN, the average evaluation index for GIMCSIS was 3.14% greater than that for IMCSIS on the training data, and the average evaluation index for GIMCSIS was 1.65% greater than that for IMCSIS on the test data, indicating that GIMCSIS affected the underfitting phenomenon of the KNN model on these data to a certain extent.

Table 4. Lung data analysis results.

	Screening	Response					
	method	Accuracy	Precision	Recall	Specificity	G-mean	F-measure
Classification method	SVM						
Train data	IMCSIS	0.8961	0.9339	0.9025	0.8843	0.8932	0.9178
	GIMCSIS	**0.9158**	**0.9385**	**0.9305**	**0.8885**	**0.9091**	**0.9344**
Test data	IMCSIS	0.8048	0.8883	0.7850	0.8667	0.8168	0.8217
	GIMCSIS	**0.8690**	**0.9050**	**0.8600**	**0.8917**	**0.8682**	**0.8746**
Classification method	DT						
Train data	IMCSIS	0.8670	0.8981	0.9111	0.7817	0.8339	0.8977
	GIMCSIS	**0.8629**	**0.8954**	**0.8929**	**0.8057**	**0.8466**	**0.8931**
Test data	IMCSIS	0.7282	0.8033	0.7770	0.6600	0.6956	0.7759
	GIMCSIS	**0.7910**	**0.8583**	**0.8270**	**0.7600**	**0.7821**	**0.8367**
Classification method	KNN						
Train data	IMCSIS	0.9515	0.9691	0.9558	0.9418	0.9484	0.9621
	GIMCSIS	**0.9878**	**0.9818**	**1.0000**	**0.9654**	**0.9824**	**0.9908**
Test data	IMCSIS	0.8885	0.8992	0.9214	0.8300	0.8735	0.9097
	GIMCSIS	**0.8872**	**0.9278**	**0.9083**	**0.8800**	**0.8884**	**0.9099**

5. Conclusions

The existing group feature screening methods mainly focus on continuous data, discrete response variables, discrete covariates, and other different cases, but feature screening with covariates missing at random has not been discussed. Considering the missing conditions of ultrahigh-dimensional data, this paper extends two-stage feature screening under a random missing mechanism to ultrahigh-dimensional data with a group structure and it presents a two-stage feature screening method with covariates missing at random. In the first stage, we use group feature screening based on adjusted Pearson chi-square statistics to find fully observed covariates that are dependent on missing indicator variables. In the second stage, the information of partially observed covariates is replaced by the fully observed covariates with dependence in the first stage so that the partially observed covariates with dependence on response variables can be found. Finally, the important features are selected by comparing the dependence between the fully observed covariates and the response variables. Compared with existing methods, GIMCSIS can efficiently extract important variables from ultrahigh-dimensional group data with covariates missing at random. In practice, the variables selected by GIMCSIS can improve the classification performance for imbalanced data, which plays an important role in expanding the path of imbalanced data analysis.

Specifically, GIMCSIS does not require model assumptions and satisfies certain screening performance requirements. According to our numerical simulation, the finite sample performance of GIMCSIS is better than that of IMCSIS, which is consistent with both the binary and multivariate response variables. The computational complexity of GIMCSIS is similar to that of IMCSIS, and the computer simulation times are similar for the same sample sizes. In the empirical analysis, we apply the GIMCSIS method to the classification model to improve the classification of ultrahigh-dimensional data with randomly missing data. The results show that GIMCSIS can identify more important covariates, and that GIMCSIS has better classification performance than IMCSIS.

Under different missing data mechanisms, the group feature screening of ultrahigh-dimensional data needs to be discussed. In addition, screening group features in the absence of both response variables and covariates is one of the more challenging problems. In terms of empirical research, discretizing continuous data to meet the needs of discretization feature screening is the key to popularizing the group feature screening of ultrahigh-dimensional discrete data.

Use of AI tools declaration

We have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

The authors are grateful to the editor and anonymous referee for their constructive comments that led to significant improvements in the paper.

The National Natural Science Foundation of China [grant number 71963008] funded this research.

Conflict of interest

The authors declare that there is no conflict of interest in the publication of this paper.

References

1. J. Q. Fan, R. Samwort, Y. C. Wu, Ultrahigh dimensional feature selection: Beyond the linear model, *J. Mach. Learn. Res.*, **10** (2009), 2013–2038. https://doi.org/10.1145/1577069.1755853

2. J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, *J. Roy. Stat. Soc. B*, **70** (2008), 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x
3. P. Hall, H. Miller, Using generalized correlation to effect variable selection in very high dimensional problems, *J. Comput. Graph. Stat.*, **18** (2009), 533–550. https://doi.org/10.1198/jcgs.2009.08041
4. G. Li, H. Peng, J. Zhang, L. Zhu, Robust rank correlation based screening, *Ann. Statist.*, **40** (2012), 1846–1877. https://doi.org/10.1214/12-AOS1024
5. X. Y. Wang, C. L. Leng, High dimensional ordinary least squares projection for screening variables, *J. Roy. Stat. Soc. B*, **78** (2016), 589–611. https://doi.org/10.1111/rssb.12127
6. L. P. Zhu, L. X. Li, R. Z. Li, L. X. 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
7. R. Li, W. Zhong, L. Zhu, Feature screening via distance correlation learning, *J. Am. Stat. Assoc.*, **107** (2012), 1129–1139. https://doi.org/10.1080/01621459.2012.695654
8. 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
9. 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
10. 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
11. 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
12. P. Lai, M. Y. Wang, F. L. Song, Y. Q. Zhou, Feature screening for ultrahigh-dimensional binary classification via linear projection, *AIMS Math.*, **8** (2023), 14270–14287. https://doi.org/10.3934/math.2023730
13. W. C. Song, J. Xie, Group feature screening via the F statistic, *Commun. Stat. Simul. C.*, **51** (2022), 1921–1931. https://doi.org/10.1080/03610918.2019.1691223
14. D. Qiu, J. Ahn, Grouped variable screening for ultra-high dimensional data for linear model, *Comput. Stat. Data Anal.*, **144** (2020), 106894. https://doi.org/10.1016/j.csda.2019.106894
15. H. J. He, G. M. Deng, Grouped feature screening for ultra-high dimensional data for the classification model, *J. Stat. Comput. Simul.,* **92** (2022), 974–997. https://doi.org/10.1080/00949655.2021.1981901
16. Z. Z. Wang, G. M. Deng, J. Q. Yu, Group feature screening based on information gain ratio for ultrahigh-dimensional data, *J. Math.*, 2022, 1600986. https://doi.org/10.1155/2022/1600986
17. Z. Z. Wang, G. M. Deng, H. Y. Xu, Group feature screening based on Gini impurity for ultrahigh-dimensional multi-classification, *AIMS Math.*, **8** (2023), 4342–4362. https://doi.org/10.3934/math.2023216
18. Y. L. Sang, X. Dang, Grouped feature screening for ultrahigh-dimensional classification via Gini distance correlation, 2023. https://doi.org/10.48550/arXiv.2304.08605
19. P. Lai, Y. M. Liu, Z. Liu, Y. Wan, Model free feature screening for ultrahigh dimensional data with responses missing at random, *Comput. Stat. Data Anal.*, **105** (2017), 201–216. https://doi.org/10.1016/j.csda.2016.08.008
20. Q. H. Wang, Y. J. Li, How to make model-free feature screening approaches for full data applicable to the case of missing response? *Scand. J. Stat.*, **45** (2018), 324–346. https://doi.org/10.1111/sjos.12290

21. X. X. Li, N. S. Tang, J. H. Xie, X. D. Yan, A nonparametric feature screening method for ultrahigh-dimensional missing response, *Comput. Stat. Data Anal.*, **142** (2020), 106828. https://doi.org/10.1016/j.csda.2019.106828
22. L. Y. Zou, Y. Liu, Z. H. Zhang, Adjusted feature screening for ultra-high dimensional missing response, *J. Stat. Comput. Simul.*, 2023. https://doi.org/10.1080/00949655.2023.2256926
23. L. Ni, F. Fang, J. Shao, Feature screening for ultrahigh dimensional categorical data with covariates missing at random, *Comput. Data Anal.*, **142** (2020), 106824. https://doi.org/10.1016/j.csda.2019.106824
24. J. Suykens, J. Vandewalle, Least squares support vector machine classifiers, *Neural Process. Lett.*, **9** (1999), 293–300. https://doi.org/10.1023/A:1018628609742
25. B. Lantz, *Machine learning with R*, 2 Eds., Packt Publishing, 2015.
26. R. J. Samworth, Optimal weighted nearest neighbour classifiers, *Ann. Stat.*, **40** (2012), 2733–2763. Available from: https://www.jstor.org/stable/41806553.

Supplementary

Lemma 1. Similar to the derivation of Lemma 1 in [23], for the categorical response Y and the categorical and fully observed covariate U_k, under Condition (1), we have

$$P\left(\left|\widehat{APC}_g(Y,U_k)-APC_g(Y,U_k)\right|>\varepsilon\right)\le O(RJ^3)\,exp\left\{-e_1\frac{n\varepsilon^2}{R^6J^{18}}\right\} \tag{S1}$$

where e_1 is a constant.

Corollary 1. For the missing indicator variable δ_l^* and the fully observed discrete covariate U_k, under Conditions (1) and (3), we have

$$P\left(\left|\widehat{APC}_g(\delta_l^*,U_k)-APC_g(\delta_l^*,U_k)\right|>\varepsilon\right)\le O(J^3)\,exp\left\{-e_2\frac{n\varepsilon^2}{J^{18}}\right\} \tag{S2}$$

where e_2 is a constant.

Lemma 2. Under Conditions (1), (3), (4) and (5), we have the following two inequalities

$$P\left(M_{l_g}\subseteq\widehat{M}_{l_g}\right)\ge 1-O(G1\cdot J^3)\,exp\left\{-e_3\frac{n^{1-2\tau_{\delta_l^*}}}{J^{18}}\right\} \tag{S3}$$

$$P\left(\widehat{M}_{l_g}\subseteq\bar{M}_{l_g}\right)\ge 1-O(G1\cdot J^3)\,exp\left\{-e_3\frac{n^{1-2\tau_{\delta_l^*}}}{J^{18}}\right\} \tag{S4}$$

where e_3 is a positive constant.

Proof. M_{l_g} and $\widehat{M}_{l_g}$ have been defined in Section 2.1 and Condition (5) respectively:

$$\widehat{M}_{l_g}=\left\{k{:}\,\widehat{APC}_g(\delta_l^*,U_k)>c_{\delta_l^*}n^{-\tau_{\delta_l^*}},1\le k\le G1\right\}$$

$$\bar{M}_{l_g}=\left\{k{:}\,APC_g(\delta_l^*,U_k)>\frac{1}{2}c_{\delta_l^*}n^{-\tau_{\delta_l^*}},1\le k\le G1\right\}.$$

Under Conditions (1), (3) and (4), it is easy to obtain the following:

$$\begin{aligned}
P\left(M_{l_g} \subseteq \widehat{M}_{l_g}\right) &\geq P\left(\left|\widehat{APC}_g(\delta_l^*, U_k) - APC_g(\delta_l^*, U_k)\right| \leq \frac{1}{2} c_{\delta_l^*} n^{-\tau_{\delta_l^*}}, \forall U_k \in M_{l_g}\right) \\
&\geq P\left(\left|\widehat{APC}_g(\delta_l^*, U_k) - APC_g(\delta_l^*, U_k)\right| \leq \frac{1}{2} c_{\delta_l^*} n^{-\tau_{\delta_l^*}}, 1 \leq k \leq G1\right) \\
&\geq 1 - \sum_{k=1}^{G1} P\left(\left|\widehat{APC}_g(\delta_l^*, U_k) - APC_g(\delta_l^*, U_k)\right| > \frac{1}{2} c_{\delta_l^*} n^{-\tau_{\delta_l^*}}\right) \\
&\geq 1 - O(G1 \cdot J^3)\, exp\left\{-c_3 \frac{n^{1-2\tau_{\delta_l^*}}}{J^{18}}\right\}
\end{aligned}$$

$$\begin{aligned}
P\left(\widehat{M}_{l_g} \subseteq \bar{M}_{l_g}\right) &= P\left(\widehat{M}_{l_g} \supseteq \bar{M}_{l_g}\right) \\
&\geq P\left(\left|\widehat{APC}_g(\delta_l^*, U_k) - APC_g(\delta_l^*, U_k)\right| \leq \frac{1}{2} c_{\delta_l^*} n^{-\tau_{\delta_l^*}}, \forall U_k \in M_{l_g}\right) \\
&\geq P\left(\left|\widehat{APC}_g(\delta_l^*, U_k) - APC_g(\delta_l^*, U_k)\right| \leq \frac{1}{2} c_{\delta_l^*} n^{-\tau_{\delta_l^*}}, 1 \leq k \leq G1\right) \\
&\geq 1 - \sum_{k=1}^{G1} P\left(\left|\widehat{APC}_g(\delta_l^*, U_k) - APC_g(\delta_l^*, U_k)\right| > \frac{1}{2} c_{\delta_l^*} n^{-\tau_{\delta_l^*}}\right) \\
&\geq 1 - O(G1 \cdot J^3)\, exp\left\{-c_3 \frac{n^{1-2\tau_{\delta_l^*}}}{J^{18}}\right\}.
\end{aligned}$$

Lemma 3. For the discrete response variable Y and the random missing covariate V_l, under Conditions (1), (3), (4) and (5), we have

$$P\left(\left|\hat{p}_{j_l r} - p_{j_l r}\right| > t\right) \leq O(G1 \cdot J^3)\, exp\left\{-e_3 \frac{n^{1-2\tau_{\delta_l^*}}}{J^{18}}\right\} + O(J^{3\bar{m}})\, exp\left\{-e_5 \frac{nt^2}{R^4 J^{12\bar{m}}}\right\} \qquad \text{(S5)}$$

where e_3 and e_5 are positive constants.

Proof. Section 2.2 gives the joint probability of group data with randomly missing data:

$$\hat{p}_{j_l r} = \frac{1}{n} \sum_u \frac{\sum_{i=1}^n I\left(y_i = r, u_i^{\widehat{M}_{l_g}} = u\right) \sum_{i=1}^n I\left(v_{l1} = j_1, \dots, v_{lp_l} = j_{p_l}, y_i = r, u_i^{\widehat{M}_{l_g}} = u, \delta_{i,l}^* = 1\right)}{\sum_{i=1}^n I\left(y_i = r, u_i^{\widehat{M}_{l_g}} = u, \delta_{i,l}^* = 1\right)}.$$

Then it is easy to get

$$\begin{aligned}
P\left(\left|\hat{p}_{j_l r} - p_{j_l r}\right| > t\right) \leq\ & P\left(\left|\hat{p}_{j_l r} - p_{j_l r}\right| > t \middle| M_{l_g} \subseteq \widehat{M}_{l_g} \subseteq \bar{M}_{l_g}\right) \\
& + P\left(M_{l_g} \not\subset \widehat{M}_{l_g}\right) + P\left(\widehat{M}_{l_g} \not\subset \bar{M}_{l_g}\right).
\end{aligned}$$

In Lemma 2, neither of the last two terms of the above formula is greater than

$O(G1 \cdot J^3)\, exp\left\{-e_3 \frac{n^{1-2\tau}\delta_l^*}{J^{18}}\right\}$, so we just have to worry about the first inequality. Before we do that, for ease of representation, we give the following notation:

Let $\phi_{r,u} = P\left(Y = r, U^{\widehat{M}_{l_g}} = u\right)$, $\varphi_{r,u} = P\left(Y = r, U^{\widehat{M}_{l_g}} = u, \delta_l^* = 1\right)$, and $\gamma_{j_l,r,u} = P\left(V_{l1} = j_1, \ldots, V_{lp_l} = j_{p_l}, Y = r, U^{\widehat{M}_{l_g}} = u, \delta_l^* = 1\right)$.

The corresponding estimators are as follows:

$$\hat{\phi}_{r,u} = n^{-1}\sum_{i=1}^{n} I\left(y_i = r, u_i^{\widehat{M}_{l_g}} = u\right)$$

$$\hat{\varphi}_{r,u} = n^{-1}\sum_{i=1}^{n} I\left(y_i = r, u_i^{\widehat{M}_{l_g}} = u, \delta_l^* = 1\right)$$

$$\hat{\gamma}_{j_l,r,u} = n^{-1}\textstyle\sum_{i=1}^{n} I\left(v_{l1} = j_1, \ldots, v_{lp_l} = j_{p_l}, y_i = r, u_i^{\widehat{M}_{l_g}} = u, \delta_{i,l}^* = 1\right).$$

Because $M_{l_g} \subseteq \widehat{M}_{l_g} \subseteq \bar{M}_{l_g}$, we have

$$\begin{aligned}
\hat{p}_{j_l r} - \bar{p}_{j_l r} &= \sum_u \frac{\hat{\phi}_{r,u}\hat{\gamma}_{j_l,r,u}}{\hat{\varphi}_{r,u}} - \sum_u \frac{\phi_{r,u}\gamma_{j_l,r,u}}{\varphi_{r,u}} \\
&= \sum_u \frac{\hat{\gamma}_{j_l,r,u}}{\hat{\varphi}_{r,u}}\left(\hat{\phi}_{r,u} - \phi_{r,u}\right) + \sum_u \phi_{r,u}\,\hat{\gamma}_{j_l,r,u}\left(\frac{1}{\hat{\varphi}_{r,u}} - \frac{1}{\varphi_{r,u}}\right) + \sum_u \frac{\phi_{r,u}}{\varphi_{r,u}}\left(\hat{\gamma}_{j_l,r,u} - \gamma_{j_l,r,u}\right) \\
&\le \sum_u \left(\hat{\phi}_{r,u} - \phi_{r,u}\right) + \sum_u \left(\frac{1}{\hat{\varphi}_{r,u}} - \frac{1}{\varphi_{r,u}}\right) + \sum_u \frac{2RJ^{3\bar{m}}}{e_4}\left(\hat{\gamma}_{j_l,r,u} - \gamma_{j_l,r,u}\right) \\
&=: I_{41} + I_{42} + I_{43}.
\end{aligned}$$

For ease of writing, the following formula ignores the conditional probability, such that $P_M(\cdot)$ is used instead of $P\left(\cdot|M_{l_g} \subseteq \widehat{M}_{l_g} \subseteq \bar{M}_{l_g}\right)$; hence

$$P_M\left(\left|\hat{p}_{j_l r} - p_{j_l r}\right| > t\right) \le P_M(I_{41} > t) + P_M(I_{43} > t) + P_M(I_{43} > t).$$

Considering I_{41},

$$\begin{aligned}
P_M(I_{41} > t) &\le P_M\left(\sum_u \left|\hat{\phi}_{r,u} - \phi_{r,u}\right| > t\right) \\
&\le \sum_u P_M\left(\left|\hat{\phi}_{r,u} - \phi_{r,u}\right| > \frac{t}{3J^{3\bar{m}}}\right)
\end{aligned}$$

$$\leq 2J^{3\bar{m}}\, exp\left\{-\frac{6n\left(\frac{t}{3J^{3\bar{m}}}\right)^{2}}{3+4\left(\frac{t}{3J^{3\bar{m}}}\right)}\right\}.$$

Similarly，

$$P_M(I_{42} > t) \leq 2J^{3\bar{m}}\, exp\left\{-\frac{6n\left(\frac{t}{3J^{3\bar{m}}}\right)^{2}}{3+4\left(\frac{t}{3J^{3\bar{m}}}\right)}\right\}$$

$$P_M(I_{43} > t) \leq 2J^{3\bar{m}}\, exp\left\{-\frac{6n\left(\frac{e_4^2 t}{48\ \ ^2 J^{6\bar{m}}}\right)^{2}}{3+4\left(\frac{e_4^2 t}{48R^2 J^{6\bar{m}}}\right)}\right\} + 2J^{3\bar{m}}\, exp\left\{-\frac{6n\left(\frac{e_4}{4RJ^{3\bar{m}}}\right)^{2}}{3+4\left(\frac{e_4}{4RJ^{3\bar{m}}}\right)}\right\}.$$

Hence，

$$P\left(\left|\hat{p}_{j_l r} - p_{j_l r}\right| > t\right) \leq O(G1 \cdot J^3)\, exp\left\{-e_3 \frac{n^{1-2\tau_{\delta_l^*}}}{J^{18}}\right\} + O(J^{3\bar{m}})\, exp\left\{-e_5 \frac{nt^2}{R^4 J^{12\bar{m}}}\right\}$$

where e_3 and e_5 are constants.

Corollary 2. For the discrete response variable Y and the random missing covariate V_l, under Conditions (1), (3), (4) and (5), we have

$$P\left(\left|\widehat{w}_{j_l} - w_{j_l}\right| > t\right) \leq O(G2 \cdot RJ^3)\, exp\left\{-e_3 \frac{n^{1-2\tau_{\delta_l^*}}}{J^{18}}\right\} + O(RJ^{3\bar{m}})\, exp\left\{-e_5 \frac{nt^2}{R^6 J^{12\bar{m}}}\right\} \tag{S6}$$

where e_3 and e_5 are constants.

Proof. Section 2.2 gives $\widehat{w}_{j_l} = \sum_{r=1}^{R} \hat{p}_{j_l r}$. Using the result in Lemma 3, we can see that

$$\begin{aligned}
P\left(\left|\widehat{w}_{j_l} - w_{j_l}\right| > t\right) &= P\left(\left|\sum_{r=1}^{R}\left(\hat{p}_{j_l r} - p_{j_l r}\right)\right| > t\right) \\
&\leq P\left(\sum_{r=1}^{R}\left|\hat{p}_{j_l r} - p_{j_l r}\right| > t\right) \\
&\leq \sum_{r=1}^{R} P\left(\left|\hat{p}_{j_l r} - p_{j_l r}\right| > \frac{t}{R}\right) \\
&\leq O(G2 \cdot RJ^3)\, exp\left\{-e_3 \frac{n^{1-2\tau_{\delta_l^*}}}{J^{18}}\right\} + O(RJ^{3\bar{m}})\, exp\left\{-e_5 \frac{nt^2}{R^6 J^{12\ \bar{}}}\right\}.
\end{aligned}$$

Lemma 4. For the discrete response variable Y and the random missing covariate V_l, under Conditions (1), (3), (4) and (5), we have

$$P\left(\left|\widehat{APC}_g(Y,V_l)-APC_g(Y,V_l)\right|>\varepsilon\right)$$

$$\leq O(G2\cdot RJ^6)\,exp\left\{-e_3\frac{n^{1-2\tau_{\delta_l^*}}}{J^{18}}\right\}+O\left(RJ^{3(\bar{m}+1)}\right)exp\left\{-e_5\frac{n\varepsilon^2}{R^{10}J^{18(\bar{m}+1)}}\right\} \tag{S7}$$

where e_3 and e_5 are positive constants.

Proof. The proof process is similar to Lemma 1, so it is omitted here.

Theorem 1. Under Conditions (1)–(6), we have

$$P\left((U,V)^D\subseteq(U,V)^{\widehat{D}}\right)\geq 1-O\left(pexp\left(-b_1n^{1-2\tau-6\xi-18}\ +(\xi+\kappa)logn\right)\right)$$

$$-O\left(pqexp\left(-b_2n^{1-2\tau_\delta-18\kappa}+(\xi+2\kappa)logn\right)\right)$$

$$-O\left(qexp\left(-b_3n^{1-2\tau-10\xi-(18^{-}+1\ )\kappa}+(\xi+(\overline{m}+1)\kappa)logn\right)\right) \tag{S8}$$

where b_1, b_2 and b_3 are constants. If $log\,p=O(n^\alpha)$, $log\,q=O\left(n^\beta\right)$, $\alpha<1-2\tau-6\xi-18\kappa$, $\beta<1-2\tau-10\xi-(18\bar{m}+18)\kappa$ and $\alpha+\beta<1-2\tau_\delta-18\kappa$, then GIMCSIS has the sure screening property.

Proof. Define four covariate sets as follows:

$$U^D=(U,V)^D\cup\left\{v_1,\dots,v_{q_g}\right\};\ V^D=(U,V)^D\cup\left\{v_1,\dots,v_{q_g}\right\}$$

$$U^{\widehat{D}}=(U,V)^{\widehat{D}}\cup\left\{v_1,\dots,v_{q_g}\right\};\ V^{\widehat{D}}=(U,V)^{\widehat{D}}\cup\left\{v_1,\dots,v_{q_g}\right\}.$$

It is obvious that $(U,V)^D=U^D\cap V^D$ and $(U,V)^{\widehat{D}}=U^{\widehat{D}}\cap V^{\widehat{D}}$.

According to Lemmas 1 and 4, we have

$$P\left((U,V)^D\subseteq(U,V)^{\widehat{D}}\right)=P\left(\left(U^D\cap V^D\right)\subseteq\left(U^{\widehat{D}}\cap V^{\widehat{D}}\right)\right)$$

$$\geq P\left(\left(U^D\subseteq U^{\widehat{D}}\right)\cap\left(V^D\subseteq V^{\widehat{D}}\right)\right)$$

$$\geq P\begin{pmatrix}\left\{\left|\widehat{APC}_g(Y,U_k)-APC_g(Y,U_k)\right|\leq cn^{-\tau},\forall U_k\in U^D\right\}\\ \cap\left\{\left|\widehat{APC}_g(Y,V_l)-APC_g(Y,V_l)\right|\leq cn^{-\tau},\forall V_l\in U^D\right\}\end{pmatrix}$$

$$\geq 1-\sum_{k=1}^{p}P\left(\left|\widehat{APC}_g(Y,U_k)-APC_g(Y,U_k)\right|>cn^{-\tau}\right)$$

$$
\begin{aligned}
&-\sum_{l=1}^{q} P\left(\left|\widehat{APC}_g(Y,V_l)-APC_g(Y,V_l)\right|>cn^{-\tau}\right)\\
&\geq 1-p\cdot O(RJ)exp\left\{-e_1\frac{c^2n^{1-2\tau}}{R^6J^{18}}\right\}\\
&-q\cdot O(G2\cdot RJ^6)exp\left\{-e_3\frac{n^{1-2\tau_{\delta_l^*}}}{J^{18}}\right\}\\
&-q\cdot O\left(RJ^{3(\bar{m}+1)}\right)exp\left\{-e_5\frac{c^2n^{1-2\tau}}{R^{10}J^{18(\bar{m}+1)}}\right\}\\
&\geq 1-O\left(pexp\left(-b_1n^{1-2\tau-6\xi-18}\ +(\xi+\kappa)logn\right)\right)\\
&-O\left(pqexp(-b_2n^{1-2\tau_\delta-18\kappa}+(\xi+2\kappa)logn)\right)\\
&-O\left(qexp\left(-b_3n^{1-2\tau-10\xi-(18\overline{m}+18)\kappa}+(\xi+(\overline{m}+1)\kappa)logn\right)\right)
\end{aligned}
$$

where $\tau_\delta = max_{1<l<G2}\tau_{\delta_l^*}$, b_1, b_2 and b_3 are constants.

© 2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0)