AIMS Mathematics, 9(2): 2874–2907.
DOI: 10.3934/math.2024142
Received: 23 October 2023
Revised: 05 December 2023
Accepted: 13 December 2023
Published: 02 January 2024

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

Research article

Ultra-high-dimensional feature screening of binary categorical response data based on Jensen-Shannon divergence

Qingqing Jiang[1] **and Guangming Deng**[1,2,*]

[1] School of Mathematics and Statistics, Guilin University of Technology, Guangxi 541000, China
[2] Applied Statistics Institute, Guilin University of Technology, Guangxi 541000, China

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

Abstract: Currently, most of the ultra-high-dimensional feature screening methods for categorical data are based on the correlation between covariates and response variables, using some statistics as the screening index to screen important covariates. Thus, with the increasing number of data types and model availability limitations, there may be a potential problem with the existence of a class of unimportant covariates that are also highly correlated with the response variable due to their high correlation with the other covariates. To address this issue, in this paper, we establish a model-free feature screening procedure for binary categorical response variables from the perspective of the contribution of features to classification. The idea is to introduce the Jensen-Shannon divergence to measure the difference between the conditional probability distributions of the covariates when the response variables take on different values. The larger the value of the Jensen-Shannon divergence, the stronger the covariate's contribution to the classification of the response variable, and the more important the covariate is. We propose two kinds of model-free ultra-high-dimensional feature screening methods for binary response data. Meanwhile, the methods are suitable for continuous or categorical covariates. When the numbers of covariate categories are the same, the feature screening is based on traditional Jensen-Shannon divergence. When the numbers of covariate categories are different, the Jensen-Shannon divergence is adjusted using the logarithmic factor of the number of categories. We theoretically prove that the proposed methods have sure screening and ranking consistency properties, and through simulations and real data analysis, we demonstrate that, in feature screening, the approaches proposed in this paper have the advantages of effectiveness, stability, and less computing time compared with an existing method.

Keywords: ultra-high-dimensional; binary categorical; Jensen-Shannon divergence; model-free; feature screening
Mathematics Subject Classification: 62H30, 62R07

1. Introduction

Due to the rapid growth of science and technology, ultra-high-dimensional data are more prevalent in various scientific study fields, including genomics, bio-imaging, and tumor classification. In ultra-high-dimensional data, the dimensionality of the variables is substantially larger than the sample size, and there are frequently very few variables among these variables that significantly influence the response variable. Therefore, it is crucial to screen a set of real covariates for this type of ultra-high-dimensional data problem. Fan and Lv [1] first proposed the method of ultra-high-dimensional feature screening and put forward the theory of sure screening, which lays the theoretical foundation for the ultra-high-dimensional feature screening method. Subsequently, a great deal of research has been developed for ultra-high-dimensional feature screening.

From the perspective of the model, current ultra-high-dimensional feature screening techniques are divided into three main categories: Based on parametric modeling assumptions, based on nonparametric and semiparametric modeling assumptions, and based on model-free assumptions. Ultra-high-dimensional feature screening based on parametric modeling assumptions: Fan and Lv [1] first proposed a marginal screening method (SIS) based on Pearson's correlation coefficient under linear modeling assumptions, where the magnitude of the absolute value of Pearson's correlation coefficient is used to measure the importance of the covariates. Given that the Pearson correlation coefficient is used to describe the degree of linear correlation between random variables, specific transformations can be applied to the covariates to account for nonlinear correlations. Therefore, Hall and Miller [2] proposed the generalized correlation coefficient to describe nonlinear relationships. Li et al. [3] proposed a robust rank correlation coefficient screening method by applying certain transformations to the response variables. Relaxing the linear model assumption to generalized linear models, Fan and Song [4] proposed a screening method based on maximum marginal likelihood estimation (MMLE-SIS). When there is less a priori knowledge about the model, nonparametric models are more adaptable than parametric models. Ultra-high-dimensional feature screening based on nonparametric and semiparametric modeling assumptions: Fan et al. [5] initially developed a marginal nonparametric screening (NIS) method for variables under the presumption of additive modeling. Liu et al. [6] proposed a conditional correlation coefficient screening method in the framework of variable coefficient modeling. In addition to the additive and variable coefficient models, Liang et al. [7] proposed a profile forward regression (PFR) screening method based on a partially linear model. It is vital to create model-free hypothetical screening methods with broad applicability when information about the model is absent. Ultra-high-dimensional feature screening based on model-free assumptions: Zhu et al. [8] first proposed a ranking screening approach (SIRS) based on covariance. The distance correlation coefficient (DC)-based screening approach was subsequently proposed by Li et al. [9]. He et al. [10] proposed the quartile adaptive screening method (QaSIS) by fitting marginal quantile regression. The correlation between two random vectors can be effectively measured by the Ball correlation, and based on this property, it can be used to rank predictor vectors. Then, Pan et al. [11] proposed a generic model-free sure independence screening procedure based on ball correlation, called BCor-SIS. Since many problems in practice cannot be accurately described by a single model, model-free screening methods can be applied more widely. Hence, studying a model-free feature screening procedure for ultra-high-dimensional data is the first focus of this work.

From a data type standpoint, the majority of existing ultra-high-dimensional feature screening

methods implicitly assume that the response variable is a continuous variable. Yet, ultra-high-dimensional data with discrete response variables is also frequently found in many areas of scientific research. For example, in medicine, identifying which genes are correlated with certain types of tumors is of interest. When the response variable is discrete, Fan and Fan [12] proposed a marginal t-test screening statistic based on a normal distribution, but its performance is poor for heavy-tailed distributions or outlier data. For this reason, Mai and Zou [13] proposed the screening method with a response variable that is a binary based on the Kolmogorov-Smirnov test statistic, which they later extended to situations where the response variable is multi-categorical. Cui et al. [14] proposed a new test based on the mean-variance index for testing the independence between a categorical random variable Y and a continuous random variable X. When all covariates are categorical variables, Huang et al. [15] developed a screening approach based on Pearson's cardinality statistic (PC-SIS). It can be seen that most of the ultra-high-dimensional variable screening methods construct the corresponding statistical indexes based on the correlation between the covariates and the response variable. In recent years, some scholars have further searched for new indexes to measure the relationship between random variables or random vectors. Ni and Fang [16], from the perspective of the amount of information, proposed a model-less feature screening method for ultra-high-dimensional variable selection based on information gain (IG-SIS). In information theory, in addition to information gain, divergence has been widely developed as a useful tool for measuring differences between information in many fields, such as in the Dempster-Shafer evidence theory: Xiao [17] proposed a new Belief Jensen-Shannon divergence to measure the discrepancy and conflict degree between the evidence. A novel reinforced belief divergence measure, known as RB, was created by Xiao [18] to measure the discrepancy between basic belief assignments in the context of the Dempster-Shafer evidence theory. Xiao [19] developed a novel generalized evidential Jensen-Shannon divergence that measures the conflict and discrepancy across several sources of evidence. To measure the disparity and discrepancy between basic belief assignments in Dempster-Shafer theory, Xiao et al. [20] suggested and examined a number of generalized evidential divergences. And, as the convergence of information theory and statistics develops, it is attractive to generalize the divergence to ultra-high-dimensional feature screening.

This paper primarily examines the feature screening procedure for the binary categorical response variable according to the state of the current feature screening for ultra-high-dimensional data. Since most current methods directly measure the specific degree of correlation between the covariates and the response variable, there may be a potential problem with the existence of a class of unimportant covariates that are also highly correlated with the response variable due to their high correlation with the other covariates. And furthermore, in real classification problems, screening out important features is not the ultimate goal but rather using the features to make classification predictions. Therefore, we do not directly measure the correlation between the response variable and the covariates, but we start from the perspective of the contribution of the features to the classification, which is the second focus of this paper's work, by introducing the Jensen-Shannon divergence to measure the difference between the conditional probability distributions of the covariates when the response variables take on different values, thus reflecting the contribution of the covariates to the classification of the response variables. The larger the value of the Jensen-Shannon divergence, the stronger the covariate's contribution to the classification of the response variable, i.e., the more important the covariate is considered.

In this study, Jensen-Shannon divergence is referred to as JS divergence for readability. The main

contributions of this paper are as follows:

(1) From the point of view of the contribution of the features to the classification, we examine a model-free feature screening procedure for binary categorical response variables, which implies less restrictive assumptions about the data and highlights the importance of features for classification prediction.

(2) The JS divergence is widely used, which is different from the I, J, and K divergences. JS divergence does not need the condition of absolute continuity of the probability distributions involved, and it has the advantages of symmetry, non-negativity, and boundedness [21], so it is very effective to use JS divergence to measure the differences between probability distributions.

(3) We propose two kinds of feature screening methods for binary response variables in different cases. When the number of covariate categories is the same, the screening method based on traditional JS divergence is used. Additionally, when the number of covariate categories is different, we propose a method to use the logarithmic factor of the number of categories to adjust the JS divergence and use it for screening variables, defined as AJS-SIS.

The suggested methods's sure screening and ranking consistency properties are further shown theoretically and through simulated studies. Furthermore, simulation experiments and real data analysis show the effectiveness, availability, and practicality of the methods proposed in this paper in terms of feature screening.

The rest of the paper is organized as follows: Section 2 describes the proposed method; Section 3 demonstrates the screening properties of the proposed method under certain conditions; Section 4 carries out simulation experiments to study the proposed method in comparison with an existing method; Section 5 is real data analysis; and Section 6 gives the conclusion.

2. Feature screening method

2.1. Basic assumption

Suppose $X = \left(x_{i1}, x_{i2}, \ldots, x_{ij}\right)$ is an $n \times p$-dimensional covariate matrix, where X obeys the assumption of independent identical distribution, and $Y = (y_1, y_2, \ldots, y_i)$ is an $n \times 1$-dimensional binary categorical response variable, where $j = 1, 2, \ldots, p, i = 1, 2, \ldots, n$. Let $\mathbf{x}_j = \left\{x_{1j}, x_{2j}, \ldots, x_{ij}\right\}, i = 1, 2, \ldots, n$. The probability function of X is denoted by $p_{j,l}$, the probability function of Y is denoted by p_r, the conditional probability function of Y given $\mathbf{x}_j$ is denoted by $p_{j,lr}$, and the conditional probability function of $\mathbf{x}_j$ given Y is denoted by $p_{lr,j}$.

The expression for p_r is as follows:

$$p_r = \Pr\left(Y = r\right), r = 1, 2,$$

$$\hat{p}_r = \frac{\sum_{i=1}^{n} I\left\{y_i = r\right\}}{n}.$$

When the covariate $\mathbf{x}_j$ is a categorical variable, let $\mathbf{x}_j$ have L categories, $L = \{1, 2, \ldots, L\}$:

$$p_{j,l} = \Pr\left(\mathbf{x}_j = l\right),$$

$$\hat{p}_{j,l} = \Pr\left(\mathbf{x}_j = l\right) = \frac{\sum_{i=1}^{n} I\left\{x_{ij} = l\right\}}{n},$$

$$p_{j,lr} = \Pr\left(\mathbf{x}_j = l \mid Y = r\right),$$

$$\hat{p}_{j,lr} = \frac{\sum_{i=1}^{n} I\left\{x_{ij} = l, y_i = r\right\}}{\sum_{i=1}^{n} I\left\{y_i = r\right\}},$$

$$p_{lr,j} = \Pr\left(Y = r \mid \mathbf{x}_j = l\right),$$

$$\hat{p}_{lr,j} = \frac{\sum_{i=1}^{n} I\left\{x_{ij} = l, y_i = r\right\}}{\sum_{i=1}^{n} I\left\{x_{ij} = l\right\}}.$$

When the covariate $\mathbf{x}_j$ is a continuous variable, reference is made to Ni and Fang [16] to cut $\mathbf{x}_j$ into categorical data using standard normal distribution quantiles:

$$p_{j,l} = \Pr\left(\mathbf{x}_j \in (q_{(J-1)}, q_{(J)}]\right),$$

$$\hat{p}_{j,l} = \frac{\sum_{i=1}^{n} I\left\{x_{ij} \in (q_{(J-1)}, q_{(J)}]\right\}}{n},$$

$$p_{j,lr} = \Pr\left(\mathbf{x}_j \in (q_{(J-1)}, q_{(J)}] \mid Y = r\right),$$

$$\hat{p}_{j,lr} = \frac{\sum_{i=1}^{n} I\left\{x_{ij} \in (q_{(J-1)}, q_{(J)}]\right\}}{\sum_{i=1}^{n} I\left\{y_i = r\right\}},$$

$$p_{lr,j} = \Pr\left(Y = r \mid \mathbf{x}_j \in (q_{(J-1)}, q_{(J)}]\right),$$

$$\hat{p}_{lr,j} = \frac{\sum_{i=1}^{n} I\left\{y_i = r\right\}}{\sum_{i=1}^{n} I\left\{x_{ij} \in (q_{(J-1)}, q_{(J)}]\right\}}.$$

Where $q_{(J)}$ is the J/J_k quantile, and $J = 1, 2, ..., J_k$, $q_{(0)} = -\infty$, $q_{(J_K)} = +\infty$.

Define two index sets: D is the set of significant covariates, D^c is the set of non-significant covariates, and $|D| = d_0$ is the number of variables in the set of significant covariates, which is expressed in set form as

$$D = \left\{j : \text{for some } Y = y,\ F\left(\mathbf{x}_j \mid y\right) \text{ is related to } Y\right\},$$

$$D^c = \{1, 2, \ldots, p\} \setminus D.$$

2.2. Information entropy

Information entropy is the information theory Shannon borrowed from the concept of thermodynamics, where, in 1948, he proposed a measure of the size of the information index and also gave the mathematical formula [22]. Taking the covariate as a categorical discrete variable $\mathbf{x}_j \in \{1, 2, \ldots, L\}$ as an example, the information entropy of the covariates $\mathbf{x}_j$ and Y are given by

$$H\left(\mathbf{x}_j\right) = -\sum_{l=1}^{L} p_{j,l} \log p_{j,l},$$

$$H\left(Y\right) = -\sum_{r=1}^{R} p_r \log p_r.$$

Where $0 \times \log 0 = 0$, and the logarithmic base is 2.

Having understood the definition of information entropy, the conditional information entropy of the covariate $\mathbf{x}_j$ given the response variable Y is defined as

$$H\left(\mathbf{x}_j \mid Y\right) = -\sum_{l=1}^{L} p_{j,lr} \log p_{j,lr},$$

$$H\left(Y \mid \mathbf{x}_j\right) = -\sum_{r=1}^{R} p_{lr,j} \log p_{lr,j}.$$

2.3. IG-SIS

The information gain is derived from information entropy, which can represent the strength of the correlation between the covariates and the response variable, and the expression for the information gain between Y and $\mathbf{x}_j$ is

$$\begin{aligned} \mathrm{IG}(Y, \mathbf{x}_j) &= \frac{1}{\log J_k}\left(H(Y) - H(Y \mid \mathbf{x}_j)\right) \\ &= \frac{1}{\log J_k}\left(\sum_{r=1}^{R}\sum_{J=1}^{J_k} p_{lr,j} \log p_{lr,j} - \sum_{r=1}^{R} p_r \log p_r - \sum_{J=1}^{J_k} p_{j,l} \log p_{j,l}\right). \end{aligned}$$

$\mathrm{IG}(Y, \mathbf{x}_j)$ represents the difference of the response variable Y between the information entropy and the conditional information entropy of the given covariate $\mathbf{x}_j$. If $\mathbf{x}_j$ is a significant variable, Y will be significantly impacted by $\mathbf{x}_j$, and thus the value of $\mathrm{IG}(Y, \mathbf{x}_j)$ is larger. Based on this, Ni and Fang [16] proposed the IG-SIS feature screening method.

The estimate of the information gain about Y and $\mathbf{x}_j$ is

$$\hat{\mathrm{IG}}(Y, \mathbf{x}_j) = \frac{1}{\log J_k}\left(\sum_{r=1}^{R}\sum_{J=1}^{J_k} \hat{p}_{lr,j} \log \hat{p}_{lr,j} - \sum_{r=1}^{R} \hat{p}_r \log \hat{p}_r - \sum_{J=1}^{J_k} \hat{p}_{j,l} \log \hat{p}_{j,l}\right). \tag{2.1}$$

After obtaining the IG values of each covariate and response variable, sort and filter all the variables by importance and select the top d_0 variables to be selected into the set of important variables.

The set of important variables is

$$\hat{D} = \left\{\mathbf{x}_j : \text{The first } d_0 \text{ descending } \hat{\mathrm{IG}}(Y, \mathbf{x}_j)\right\}.$$

2.4. JS-SIS

JS divergence (Jensen-Shannon divergence, abbreviated JSD) is a statistical measure based on the KL divergence (relative entropy). Assuming that there are two probability distributions $G = \Pr\left(\mathbf{x}_j = l \mid Y = 1\right)$ and $Q = \Pr\left(\mathbf{x}_j = l \mid Y = 2\right)$ for the same random variable $\mathbf{x}_j$ in space, JS divergence can measure the degree of difference between these two distributions, and larger JS divergence implies that the covariates are more important. According to Lin [21], the value of JS divergence is non-negative, equal to 0 when $G = Q$, and it is upper bound by 1.

JS divergence is actually a variant form of KL divergence (relative entropy), and KL divergence can be computed in the following way:

$$\begin{aligned} D_{\mathrm{KL}}(G \parallel Q) &= \sum_{j=1}^{p} G \log \frac{G}{Q} \\ &= \sum_{j=1}^{p} G \log G - \sum_{j=1}^{p} G \log Q. \end{aligned}$$

Since KL divergence is asymmetric, it cannot accurately measure the real difference between G and Q. JS divergence solves this problem by constructing the average probability distribution of G and Q.

Assume that $M = \frac{1}{2}(G + Q)$ is the average probability distribution of G and Q. The JS divergence of G and Q is defined as

$$\begin{aligned} e_j =& \mathrm{JS}(G \parallel Q) \\ =& \frac{1}{2} D_{\mathrm{KL}}(G \parallel M) + \frac{1}{2} D_{\mathrm{KL}}(Q \parallel M) \\ =& \frac{1}{2} \sum_{j=1}^{p} G \log\left(\frac{G}{M}\right) + \frac{1}{2} \sum_{j=1}^{p} Q \log\left(\frac{Q}{M}\right) \\ =& \frac{1}{2} \sum_{j=1}^{p} G \log(G) - \frac{1}{2} \sum_{j=1}^{p} G \log(M) + \frac{1}{2} \sum_{j=1}^{p} Q \log(Q) - \frac{1}{2} \sum_{j=1}^{p} Q \log(M) \\ =& \frac{1}{2} (H(G, M) - H(G)) + \frac{1}{2} (H(Q, M) - H(Q)). \end{aligned}$$

The estimate of the JS divergence between G and Q is

$$\begin{aligned} \hat{e}_j =& \mathrm{JS}\left(\hat{G} \parallel \hat{Q}\right) \\ =& \frac{1}{2}\left(H\left(\hat{G}, \hat{M}\right) - H(\hat{G})\right) + \frac{1}{2}\left(H\left(\hat{Q}, \hat{M}\right) - H(\hat{Q})\right). \end{aligned} \tag{2.2}$$

The probability distributions G and Q based on continuous covariates are defined as $G = \Pr\left(\mathbf{x}_j \in (q_{(J-1)}, q_{(J)}] \mid Y = 1\right)$, $Q = \Pr\left(\mathbf{x}_j \in (q_{(J-1)}, q_{(J)}] \mid Y = 2\right)$.

2.5. AJS-SIS

The definition of Eq (2.2) may lead to the incorrect selection of non-significant covariates with a large number of categories because covariates with more categories may have larger calculated JS

divergence values, especially when the number of categories involved in each covariate varies. To address this issue, this paper refers to Ni and Fang [16] applying $(\log J_k)^{-1}$ to construct an adjusted JS divergence to measure the importance of $\mathbf{x}_j$:

$$\begin{aligned} w_j =& e_j / \log J_k \\ =& \left[\frac{1}{2}\left(H\left(G, M\right) - H(G)\right) + \frac{1}{2}\left(H\left(Q, M\right) - H(Q)\right)\right] / \log J_k. \end{aligned} \tag{2.3}$$

The estimate of the adjusted JS divergence between G and Q is

$$\begin{aligned} \hat{w}_j =& \hat{e}_j / \log J_k \\ =& \left[\frac{1}{2}\left(H\left(\hat{G}, M\right) - H(\hat{G})\right) + \frac{1}{2}\left(H\left(\hat{Q}, \hat{M}\right) - H(\hat{Q})\right)\right] / \log J_k. \end{aligned} \tag{2.4}$$

When $\mathbf{x}_j$ is a categorical variable, J_k equals the number of categories L of $\mathbf{x}_j$, and when $\mathbf{x}_j$ is a continuous variable, J_k represents the number of categories into which $\mathbf{x}_j$ is cut by the standard normal distribution quantile.

3. Theoretical properties

Fan and Lv [1] mentioned that a method is meaningful if a feature screening method has a sure screening property. It is the basis of feature screening, which means that the probability of all significant covariates being selected tends to be 1. Therefore, subsequent feature screening methods that extend the SIS method demonstrate this property, such as those in the articles by Li et al. [9], Cui and Zhong [14], and Ni and Fang [16].

In addition to the sure screening property, a feature screening method should also have a ranking consistency property. It means that the feature screening approach is consistent and can guarantee that the values of all important covariate indexes are ranked before all other unimportant covariates.

These two properties eventually guarantee the usefulness and effectiveness of feature screening methods.

This subsection will illustrate the theoretical properties of the methods proposed in this paper under certain conditions, which are as follows:

(C1) $p = o(\exp(n^{\delta})), \delta \in (0, 1)$*, which means the variable dimension p is an exponential multiple of the sample capacity n.*

(C2) *There exist positive numbers* c_1*,* c_2*, such that* $0 < c_1 \leq p_{j,lr} \leq c_2 < 1$*,* $\forall l \in \{1, J_k\}$*,* $\forall r \in \{1, 2\}$*, and* $\forall j = 1, 2, \ldots, p$*.*

(C3) *There exist positive* $c > 0$ *and* $0 \leq \tau < 1/2$*, such that* $\min_{j \in D} e_j \geq 2cn^{-\tau}$*.*

(C4) *There exists a constant* c_3 *for* $\forall 1 \leq r \leq R$ *such that* $0 < f_k(x \mid Y = r) < c_3$*, and x is in the domain of definition of* X_k*, where* $f_k(x \mid Y = r)$ *is the Lebesgue density function of* X_k *under the condition* $Y = r$*.*

(C5) *There exists a constant* c_4 *and* $\forall 1 \leq \rho \leq 1/2$ *such that* $f_k(x) \geq c_4 n^{-\rho}$*, and x is in the domain of definition of* X_k *for* $\forall 1 \leq k \leq \rho$*, where* $f_k(x)$ *is the Lebesgue density function of* X_k*, and* $f_k(x)$ *is continuous in the domain of definition of* X_k*.*

(C6) $J = \max_{1 \leq j \leq p} J_k = O(n^{\kappa})$*,* $\kappa > 0$*,* $\forall 1 \leq \tau \leq 1/2$ *and* $\forall 1 \leq \rho \leq 1/2$ *with* $2\tau + 2\rho < 1$*.*

The above six conditions are frequently found in the literature on ultra-high-dimensional feature screening methods, such as Fan and Lv [1], Li et al. [9], Cui et al. [14], and Ni and Fang [16]. Condition (C1) indicates that it is a feature screening method used in ultra-high-dimensional problems; Condition (C2) indicates that the marginal probabilities of the response variable and the covariate are bounded by an upper and a lower limit to avoid the extreme case of the failure of the screening method; and Condition (C3) indicates that the values of the indexes belonging to the really important variables are bounded by a lower value. Condition (C4) ensures that the sample percentile is near to the true percentile by excluding an extreme case in which some X_k places a huge mass in a tiny range. Condition (C5) requires the density to have a lower bound of order $n^{-\rho}$. Condition (C6) ensures that the number of categories of covariates diverges at a certain rate.

Under these six conditions, we give the theoretical properties of the feature screening method JS-SIS when the response is a binary categorical variable.

Since $w_j = e_j/\log J_k$, $\hat{w}_j = \hat{e}_j/\log J_k$, and $logJ_k \geq \log 2 \geq 1/2$, it follows that $\Pr\left(\left|w_j - \hat{w}_j\right| > \epsilon\right) = \Pr\left(\left|e_j - \hat{e}_j\right| > \epsilon/2\right)$. Therefore, this paper gives the properties of sure screening and ranking consistency for feature screening using the index e_j of the JS divergence and a detailed theoretical proof.

3.1. Sure screening property

To distinguish between types of covariates, discrete covariates are subscripted with j and continuous covariates are subscripted with k. If the covariate is categorical, L is the number of categories of the covariate, and if the covariate is continuous, J_k is the number of categories of the covariate.

Theorem 3.1. *When the covariates are categorical, in conditions (C1) and (C2),* $0 \leq \tau < 1/2$*, and there exists a positive number c such that*

$$\Pr\left(\max_{1\leq j\leq p}\left|e_j - \hat{e}_j\right| > cn^{-\tau}\right) \leq 8pL\exp\left\{-c^2n^{1-2\tau}\Big/2L^2\right\}$$

and when $0 < \delta < 1 - 2\tau$, $\Pr(\max_{1\leq j\leq p}\left|e_j - \hat{e}_j\right| > cn^{-\tau}) \to 0, n \to \infty$. *Under conditions (C1)–(C3), when* $n \to \infty$*, there exists a positive number c such that*

$$\Pr(D \subseteq \hat{D}) \geq 1 - 8d_0L\exp\left\{-c^2n^{1-2\tau}/2L^2\right\} \to 1.$$

Theorem 3.1 states that the probability that the set of true covariates D is contained in the set of simplified covariates $\hat{D}$ as $n \to \infty$ converges to 1, which means that as the sample size n increases, eventually all the true covariates can theoretically be filtered out.

Theorem 3.2. *When the covariates are continuous variables, there exist positive constants* c_5*,* c_6*,* c_7 *under the conditions (C1), (C4)–(C6), and we have*

$$\Pr\left(\max_{1\leq j\leq p}|e_k - \hat{e}_k| > c_5n^{-\tau}\right) \leq 4c_6pJ_k\exp\left\{\frac{-c_7c_5^2n^{1-2\rho-2\tau}}{4J_k^2}\right\} \tag{3.1}$$

and when $n \to \infty$

$$\Pr(D \subseteq \hat{D}) \geq 1 - 4c_6d_0J_k\exp\left\{\frac{-c_7c_5^2n^{1-2\rho-2\tau}}{4J_k^2}\right\} \to 1.$$

Theorem 3.3. *When the covariates are continuous and categorical covariates coexist, there exists a positive constant c_9 under the conditions (C1), (C2), (C4)–(C6), and we have*

$$\begin{aligned} &\Pr\left(\max_{1\le j\le p}\left(\left|e_j-\hat{e}_j\right|+|e_k-\hat{e}_k|\right)>c_9n^{-\tau}\right)\\ &\le 8p_1L\exp\left\{-c_9^2n^{1-2\tau}/8L^2\right\}+4c_6p_2J_k\exp\left\{\frac{-c_7c_9^2n^{1-2\rho-2\tau}}{16J_k^2}\right\} \end{aligned} \tag{3.2}$$

and when $n\to\infty$,

$$\Pr(D\subseteq\hat{D})\ge 1-8d_1L\exp\left\{-c_9^2n^{1-2\tau}/8L^2\right\}-4c_6d_2J_k\exp\left\{\frac{-c_7c_9^2n^{1-2\rho-2\tau}}{16J_k^2}\right\}\to 1$$

where $p_1+p_2=p$, $d_1+d_2=d_0$.

3.2. Ranking consistency property

Theorem 3.4. *When the covariates are categorical, under conditions (C1)–(C3), assume that $\min_{j\subseteq D}\hat{e}_j-\max_{j\subseteq D^c}\hat{e}_j>0$, then we have*

$$\Pr\left\{\liminf_{n\to\infty}(\min_{j\subseteq D}\hat{e}_j-\max_{j\subseteq D^c}\hat{e}_j)>0\right\}=1.$$

Theorem 3.5. *When the covariates are continuous variables, under conditions (C1), (C3)–(C6), assume that $\min_{k\subseteq D}\hat{e}_k-\max_{k\subseteq D^c}\hat{e}_k>0$, then we have*

$$\Pr\left\{\liminf_{n\to\infty}(\min_{k\subseteq D}\hat{e}_k-\max_{k\subseteq D^c}\hat{e}_k)>0\right\}=1.$$

Theorem 3.6. *When the covariates are continuous and categorical covariates coexist, under conditions (C1)–(C6), assume that $\min_{j\subseteq D}\hat{e}_j-\max_{j\subseteq D^c}\hat{e}_j>0$ and $\min_{k\subseteq D}\hat{e}_k-\max_{k\subseteq D^c}\hat{e}_k>0$, then we have*

$$\Pr\left\{\liminf_{n\to\infty}\left((\min_{j\subseteq D}\hat{e}_j-\max_{j\subseteq D^c}\hat{e}_j)+(\min_{k\subseteq D}\hat{e}_k-\max_{k\subseteq D^c}\hat{e}_k)\right)>0\right\}=1.$$

A detailed proof of the theoretical part is in the Appendix.

4. Numerical simulation

In this section, we conduct simulation experiments to investigate the variable screening performance of our proposed methods, in which we analyze the simulation in terms of two main aspects: The type of distribution of the response variable and the type of the covariate. The methods proposed in this study are limited to binary response variables and make no assumptions regarding the data types of the covariates. In practice, four types of covariates are generally encountered: All the covariates are categorical with the same categories; all the covariates are categorical with different categories; all the covariates are continuous; both continuous and categorical covariates appear in the data, where the categories of the categorical variables differ. To examine the validity and viability of the proposed methods, we created four simulation experiments that tested them using the four kinds of data types

of the covariates mentioned above. Of these, the similarities are that the response variable is binary and the type of distribution of the response variable is the same, while the differences are in the type of covariates and the generation method of the covariates. Besides, to determine whether varying the number of slices will affect the effectiveness of the suggested methods, we attempt to slice the continuous variables in the simulation experiments using varying numbers of slices. This can also provide some references for choosing the optimal categories of the continuous variables in practical applications.

4.1. Evaluation indexes

Three evaluation indexes were used to compare the effectiveness of the variable screening method. The first evaluation index is CP, coverage 1, which indicates the proportion of true significant covariates screened for inclusion in the set of significant covariates. And, with CP $\in [0, 1]$, when the value is closer to 1, the more true significant covariates are selected to be included in the set of significant covariates. The second evaluation index is CPa, coverage 2, which indicates whether the selected set of significant covariates contains all the true significant covariates, where CPa= $0, 1$, so the average value of CPa takes the range of $[0, 1]$. The average value being closer to 1 indicates that the selection of significant covariates has a higher probability of including all of the actual significant variables. CP1 and CPa1 are used to denote the index values when the first $[n/\log n]$ variables are screened as the set of significant covariates, and CP2 and CPa2 are used to denote the index values when the first $2\,[n/\log n]$ variables are selected as the set of significant covariates. The third evaluation index is the MMS, which represents the minimum model size at which all important variables will be screened and expresses the performance of the method by calculating the 5%, 25%, 50%, 75%, and 95% quantile points of the MMS. The value of the quartile of the MMS is in the range of $[0, 1]$, and the lower the value means that the screening method can select the truly essential variables while reducing dimensionality. In this paper, the final evaluation index is expressed as the average of the indexes of 100 simulation experiments. Hence, we calculated the standard deviation, where lower values indicate higher method stability as well as the feasibility of using the average value for evaluation.

4.2. Simulation experiments and results

4.2.1. Simulation 1

The response variable is binary categorical, all covariates are categorical, and each covariate has the same number of categories. We refer to the simulation experiment in Ni and Fang [16] and consider both balanced and unbalanced distributions for the response variable: (1) balanced, $p_r = \Pr(Y = r) = 1/R$, with $r = 1, \ldots, R$, and $R = 2$; (2) unbalanced, $p_r = 2\,[1 + (R - r)/(R - 1)]\,/3R$ with $\max\limits_{1\le r\le R} p_r = 2 \min\limits_{1\le r\le R} p_r$. Define the set of true imporant variables as D, where $D = \{1, 2, \ldots, 10\}$ with $d_0 = |D| = 10$. Conditional on Y, the relevant categorical covariates are generated as $\Pr\left(x_{ij} = (1, 2, 3, 4) \mid y_i = r\right) = \left(\theta_{rj}/2, \left(1 - \theta_{rj}\right)/2, \theta_{rj}/2, \left(1 - \theta_{rj}\right)/2\right)$ for $1 \le r \le R$ and $1 \le j \le d_0$, where θ_{rj} is given in Table 1. And, $\theta_{rj} = 0.5$ when $1 \le r \le R$, $d_0 < j \le p$. We take the dimensions of the covariates $p = 1000$ and $p = 2000$ with sample sizes of $n = 200$ and $n = 400$.

Table 1. Parameter specification for the simulations.

	θ_{rj}									
j	1	2	3	4	5	6	7	8	9	10
$r = 1$	0.2	0.8	0.7	0.2	0.2	0.9	0.1	0.1	0.7	0.7
$r = 2$	0.9	0.3	0.3	0.7	0.8	0.4	0.7	0.6	0.4	0.1

Tables 2 and 3 show that CP and CPa for all methods are higher when Y is a balanced distribution than when Y is an unbalanced distribution, and the MMS for all methods is closer to the number of significant variables $d_0 = 10$. When the dimensionality of the covariates is $p = 1000$, variable screening performs better than when $p = 2000$, indicating that, for a given sample size, variable screening may become more challenging as the dimensionality of the covariates rises. Because all covariates are 4-categorical data in this simulation, the effects of JS-SIS and AJ-SIS are the same. The coverage CP and CPa of variable screening in JS-SIS and AJ-SIS are almost the same as in IG-SIS, except that the MMS in IG-SIS is a little closer to $d_0 = 10$ than in JS-SIS and AJ-SIS. Besides, through Simulation 1, it can be seen that JS-SIS and AJS-SIS can effectively screen out important variables, indicating that JS-SIS and AJS-SIS apply to data where the covariates are all categorical variables with the same category.

Table 2. Results from Simulation 1 when Y is a balanced distribution.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
balanced Y, p=1000, n=200									
JS-SIS	0.998(0.001)	0.999(0.001)	0.98(0.014)	0.99(0.01)	1.45	3.25	5.5	7.75	11.368
AJS-SIS	0.998(0.001)	0.999(0.001)	0.98(0.014)	0.99(0.01)	1.45	3.25	5.5	7.75	11.368
IG-SIS	0.998(0.001)	0.999(0.001)	0.98(0.014)	0.99(0.01)	1.45	3.25	5.5	7.75	11.368
balanced Y, p=1000, n=400									
JS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.583
AJS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.583
IG-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.583
balanced Y, p=2000, n=200									
JS-SIS	0.992(0.003)	0.996(0.002)	0.92(0.027)	0.96(0.02)	1.45	3.25	5.5	7.75	16.04
AJS-SIS	0.992(0.003)	0.996(0.002)	0.92(0.027)	0.96(0.02)	1.45	3.25	5.5	7.75	16.04
IG-SIS	0.992(0.003)	0.996(0.002)	0.92(0.027)	0.96(0.02)	1.45	3.25	5.5	7.75	16.029
balanced Y, p=2000, n=400									
JS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.599
AJS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.599
IG-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.599

The numbers in parentheses are the corresponding standard deviations.

Table 3. Results from Simulation 1 when Y is an unbalanced distribution.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
unbalanced Y, p=1000, n=200									
JS-SIS	0.991(0.003)	0.995(0.002)	0.91(0.029)	0.95(0.022)	1.45	3.25	5.5	7.75	17.113
AJS-SIS	0.991(0.003)	0.995(0.002)	0.91(0.029)	0.95(0.022)	1.45	3.25	5.5	7.75	17.113
IG-SIS	0.992(0.003)	0.995(0.002)	0.92(0.029)	0.95(0.022)	1.45	3.25	5.5	7.75	16.865
unbalanced Y, p=1000, n=400									
JS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.748
AJS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.748
IG-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.742
unbalanced Y, p=2000, n=200									
JS-SIS	0.981(0.004)	0.989	0.81(0.039)	0.89(0.031)	1.45	3.25	5.5	7.75	24.204
AJS-SIS	0.981(0.004)	0.989	0.81(0.039)	0.89(0.031)	1.45	3.25	5.5	7.75	24.204
IG-SIS	0.982(0.004)	0.989	0.82(0.039)	0.89(0.031)	1.45	3.25	5.5	7.75	23.748
unbalanced Y, p=2000, n=400									
JS-SIS	0.999(0.001)	0.999(0.001)	0.99(0.01)	0.99(0.01)	1.45	3.25	5.5	7.75	11.436
AJS-SIS	0.999(0.001)	0.999(0.001)	0.99(0.01)	0.99(0.01)	1.45	3.25	5.5	7.75	11.436
IG-SIS	0.999(0.001)	0.999(0.001)	0.99(0.01)	0.99(0.01)	1.45	3.25	5.5	7.75	11.376

The numbers in parentheses are the corresponding standard deviations.

4.2.2. Simulation 2

The response variables are set the same as in Simulation 1. The covariates were categorical, and each covariate had a different number of categories, set at 2, 4, 6, 8, and 10. Define the set of important variables as $D = \{j = [j'p/10], j' = 1, 2, \ldots, 10\}$. Referring to the simulation experiment setup in [23], the latent variables $z_i = \left(z_{i,1}, \ldots, z_{i,p}\right)$ are generated under the condition y_i. Generate covariate $x_{i,j}$ by $f_j\left(\varepsilon_{i,j} + \mu_{i,j}\right)$, where $1 \leq j \leq p$ and $f_j(\cdot)$ is the quantile function of the standard normal distribution. And $\varepsilon_{i,j} \sim N(0, 1)$, $\mu_{i,j} = 1.5 \times (-0.9)^r$ when $j \in D$, and $\mu_{i,j} = 0$ when $j \notin D$.

The specific steps for generating covariate data are as follows:

$$f_j\left(\varepsilon_{i,j} + \mu_{i,j}\right) = I\left(z_{i,j} > z_{\left(\frac{j''}{L}\right)}\right) + 1, (j'' = 1, 2, \ldots, L-1).$$

If $1 \leq j \leq 400$, then $L = 2$; if $401 \leq j \leq 800$, then $L = 4$; if $801 \leq j \leq 1200$, then $L = 6$; if $1201 \leq j \leq 1600$, then $L = 8$; if $1601 \leq j \leq 2000$, then $L = 10$.

This makes the number of covariates the same for two categorical, four categorical, six categorical, eight categorical, and ten categorical. We take $p = 2000$, $n = 160, 240, 320$.

Table 4 displays the outcomes of the simulation. The performance metrics of all methods for all conditions are exactly the same; the coverage CP and CPa are 1, and the MMS values are close to $d_0 = 10$. Furthermore, Simulation 2 indicates that JS-SIS and AJS-SIS apply to data where

the covariates are all categorical variables with different categories since it clearly shows that these two methods are effective at selecting significant variables.

Table 4. Results for Simulation 2.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
balanced Y, p=2000, n=160									
JS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
balanced Y, p=2000, n=240									
JS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
balanced Y, p=2000, n=320									
JS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
unbalanced Y, p=2000, n=160									
JS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
unbalanced Y, p=2000, n=240									
JS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
unbalanced Y, p=2000, n=320									
JS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55

The numbers in parentheses are the corresponding standard deviations.

4.2.3. Simulation 3

The response variables are set the same as in Simulation 1. The covariates are continuous variables, which we use the quantile function of the standard normal distribution to cut into categorical data with different numbers of slices $J_k = 4, 8, 10$, respectively, and define the methods corresponding to the number of slices as JS-SIS-4, AJS-SIS-4, IG-SIS-4; JS-SIS-8, AJS-SIS-8, IG-SIS-8; JS-SIS-10, AJS-SIS-10, IG-SIS-10. The set of important variables is set up as in Simulation 1. We use the normal distribution to generate covariates, where $\mathbf{x}_i = \left\{x_{i1}, x_{i2}, \ldots, x_{ip}\right\} \in \mathbb{R}^p$ and $x_{ij}(j = 1, 2, \ldots, p)$ are distributed as $N(\mu_{ij}, 1)$ with $\mu_i = \left\{\mu_{i1}, \mu_{i2}, \ldots, \mu_{ip}\right\}$. When $Y = r$ and $j \in D$, $\mu_{ij} = (-1)^r\theta_{rj}$, otherwise $j \notin D$, $\mu_{ij} = 0$. We take $p = 5000$ and $n = 400, 600, 800$.

As can be seen from Tables 5 and 6, there are significant differences in the performance of the methods only when the sample size is relatively small. Therefore, the simulation results for the specific analyzed sample size of $n = 400$ are as follows: By comparing the performance of different methods applying different numbers of slices, it is found that the CP and CPa values are the same for all three methods, and only the MMS values are slightly different. Specifically, JS-SIS and AJ-SIS have smaller MMS values than IG-SIS when Y is a balanced distribution, while IG-SIS has smaller MMS values than JS-SIS and AJS-SIS when Y is an unbalanced distribution. The performance indexes of JS-SIS and AJS-SIS are the same because the same number of slices is used for dividing all covariates. By comparing the two different distributions of Y, it is discovered that when Y is unbalanced, all methods' CP and CPa values are higher than when Y is balanced, and all methods' MMS values are lower. All approaches perform better when smaller slices are applied to continuous variables, according to comparisons between applications of various slice counts. In addition, it is evident from Simulation 3 that JS-SIS and AJS-SIS are capable of efficiently selecting significant variables, suggesting that they apply to data in which the covariates are all continuous.

4.2.4. Simulation 4

The response variables are set the same as in Simulation 1. There are two kinds of covariates: Continuous and categorical, and the treatment of continuous covariates is the same as in Simulation 3. The set of important variables is $D = \{j = [j'p/20], j' = 1, 2, \ldots, 20\}$. To generate the covariates, the latent variables $z_i = \left(z_{i,1}, \ldots, z_{i,p}\right)$ are first generated through the normal distribution by the same process as that used to generate the covariates in Simulation 3. We then refer to [23] and generate categorical and continuous covariates, where the first $1/4$ of the covariates are four categorical, the middle $1/4$ to $1/2$ of the covariates are ten categorical, and the remaining $1/2$ of the covariates are continuous. We take $p = 5000$ and $n = 400, 600, 800$.

As can be seen from Tables 7 and 8, the results of Simulations 4 and 3 are similar. Therefore, the simulation results for the specific analyzed sample size of $n = 400$ are as follows: By comparing the performance of various methods used with various slice numbers, it is discovered that when Y is a balanced distribution, the CP and CPa values of the JS-SIS are smaller than those of the AJ-SIS and IG-SIS and fluctuate more, whereas those of the AJ-SIS and IG-SIS are the same and fluctuate much less. Regarding the MMS values, the MMS values of JS-SIS, although smaller than those of AJ-SIS and IG-SIS when the number of slices is large, fluctuate more, while the MMS values of AJ-SIS and IG-SIS are roughly the same size. By comparing two different distributions of Y, all methods perform better when Y is unbalanced than when Y is balanced, and all methods show better performance when the number of slices is small. And, as shown in Simulation 4, JS-SIS and AJS-SIS can effectively screen for significant variables, which means they are appropriate for data with both continuous and categorical covariates where the categories of the categorical variables differ.

Table 5. Results from Simulation 3 when Y is a balanced distribution.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
balanced Y, p=5000, n=400									
JS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.819
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.819
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.808
JS-SIS-8	0.999(0.001)	0.999(0.001)	0.99(0.01)	0.99(0.01)	1.45	3.25	5.5	7.75	10.787
AJS-SIS-8	0.999(0.001)	0.999(0.001)	0.99(0.01)	0.99(0.01)	1.45	3.25	5.5	7.75	10.787
IG-SIS-8	0.999(0.001)	0.999(0.001)	0.99(0.01)	0.99(0.01)	1.45	3.25	5.5	7.75	10.793
JS-SIS-10	0.997(0.002)	0.999(0.001)	0.97(0.017)	0.99(0.01)	1.45	3.25	5.5	7.75	11.326
AJS-SIS-10	0.997(0.002)	0.999(0.001)	0.97(0.017)	0.99(0.01)	1.45	3.25	5.5	7.75	11.326
IG-SIS-10	0.997(0.002)	0.999(0.001)	0.97(0.017)	0.99(0.01)	1.45	3.25	5.5	7.75	11.337
balanced Y, p=5000, n=600									
JS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
JS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
balanced Y, p=5000, n=800									
JS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
JS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55

The numbers in parentheses are the corresponding standard deviations.

Table 6. Results from Simulation 3 when Y is an unbalanced distribution.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
unbalanced Y, p=5000, n=400									
JS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.61
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.61
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.61
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	10.045
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	10.045
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.984
JS-SIS-10	0.999(0.001)	1(0)	0.99(0.01)	1(0)	1.45	3.25	5.5	7.75	10.428
AJS-SIS-10	0.999(0.001)	1(0)	0.99(0.01)	1(0)	1.45	3.25	5.5	7.75	10.428
IG-SIS-10	0.999(0.001)	1(0)	0.99(0.01)	1(0)	1.45	3.25	5.5	7.75	10.33
unbalanced Y, p=5000, n=600									
JS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
JS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
unbalanced Y, p=5000, n=800									
JS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
JS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
AJS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55
IG-SIS-10	1(0)	1(0)	1(0)	1(0)	1.45	3.25	5.5	7.75	9.55

The numbers in parentheses are the corresponding standard deviations.

Table 7. Results from Simulation 4 when Y is a balanced distribution.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
balanced Y, p=5000, n=400									
JS-SIS-4	0.998(0.001)	0.999(0.001)	0.97(0.017)	0.98(0.014)	1.95	5.75	10.5	15.25	20.112
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.127
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.126
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.18
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.2
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.2
JS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.273
AJS-SIS-10	1(0.001)	1(0)	0.99(0.01)	1(0)	1.95	5.75	10.5	15.25	19.34
IG-SIS-10	1(0.001)	1(0)	0.99(0.01)	1(0)	1.95	5.75	10.5	15.25	19.341
balanced Y, p=5000, n=600									
JS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
JS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
balanced Y, p=5000, n=800									
JS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
JS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05

The numbers in parentheses are the corresponding standard deviations.

Table 8. Results from Simulation 4 when Y is an unbalanced distribution.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
unbalanced Y, p=5000, n=400									
JS-SIS-4	1(0.001)	1(0)	0.99(0.01)	1(0)	1.95	5.75	10.5	15.25	19.499
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.065
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.065
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.123
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.107
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.108
JS-SIS-10	1(0.001)	1(0)	0.99(0.01)	1(0)	1.95	5.75	10.5	15.25	19.122
AJS-SIS-10	1(0.001)	1(0)	0.99(0.01)	1(0)	1.95	5.75	10.5	15.25	19.157
IG-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.144
unbalanced Y, p=5000, n=600									
JS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.063
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
JS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
unbalanced Y, p=5000, n=800									
JS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-4	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
JS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-8	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
JS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
AJS-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05
IG-SIS-10	1(0)	1(0)	1(0)	1(0)	1.95	5.75	10.5	15.25	19.05

The numbers in parentheses are the corresponding standard deviations.

4.3. Computational time cost

As the dimensionality of the data increases, the computational time also increases, so more efficient algorithms help to improve computational efficiency. We evaluated the computational time costs of the three methods. The median running time of each algorithm is obtained as a comparison index through simulation experiments, where the covariate X is set similarly to simulation experiment 2, and Y is set to a balanced distribution. The set of significant variables is $D = \{j = [j'p/10], j' = 1, 2, \ldots, 10\}$

such that 1/5 of the significant covariates are two categorical, 1/5 are four categorical, 1/5 are six categorical, 1/5 are eight categorical, and 1/5 are ten categorical. In this simulation experiment, the control sample size was constant at 400, the dimensionality of the covariates was increased from 1,000 to 10,000 at a rate of 1,000 per increase, and the experiment was repeated 100 times. Then, the median running time of the three methods in 100 experiments was calculated. All calculations were done on a Windows 10 computer with an Intel Core i7-8700 3.20 GHz CPU.

Table 9 shows the median values of the running time for the three methods, and it can be seen that, due to the linear variation of p, the running time also shows a linear trend of variation, increasing as p increases. The running times of the three methods do not differ much at low dimensions, and as p increases, the running times of JS-SIS and AJS-SIS are significantly shorter than those of IG-SIS. And, in every case, JS-SIS has the shortest run time, AJS-SIS has the second shortest run time, and IG-SIS has the longest run time.

Table 9. Simulation results for calculating the cost of time.

p	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
JS-SIS	1.557	3.127	4.705	6.233	7.847	9.426	11.034	12.644	14.277	15.833
	(0.004)	(0.005)	(0.006)	(0.008)	(0.008)	(0.009)	(0.002)	(0.008)	(0.011)	(0.009)
AJS-SIS	1.616	3.234	4.865	6.435	8.118	9.750	11.416	13.089	14.778	16.431
	(0.003)	(0.005)	(0.006)	(0.007)	(0.006)	(0.008)	(0.006)	(0.008)	(0.009)	(0.009)
IG-SIS	1.684	3.372	5.703	6.711	8.465	10.178	12.061	13.642	15.398	17.131
	(0.003)	(0.004)	(0.018)	(0.007)	(0.005)	(0.017)	(0.009)	(0.007)	(0.008)	(0.008)

The numbers in parentheses are the corresponding standard deviations.

4.4. Comprehensive analysis of simulation results

Overall, the suggested approaches' CP and CPa values are nearly equal to 1, the quartiles of MMS are nearly at the model size of the real essential variables, and the standard deviations of CP, CPa, and computation time are close to 0. This indicates that: (1) our suggested methods are efficient at screening significant variables; (2) they are very stable; (3) it is feasible to represent the methods' performance using the average of the indexes; (4) they perform well with a broad range of data types and can be applied to situations where the data contain categorical covariates with the same categories, categorical covariates with different categories, continuous covariates, and both continuous and categorical covariates appearing in the data, where the categories of the categorical covariates differ.

The specific analysis is as follows: The JS-SIS and AJS-SIS methods proposed in this paper are very similar to IG-SIS in terms of performance. When the sample size is small, there is a difference in performance between the methods, and the performance of JS-SIS is affected more by the number of slices compared to AJ-SIS and IG-SIS, which are more adaptive to the number of slices and are more robust. However, the performance of all techniques converges to the same as the number of screening variables or sample size grows. Both CP and CPa increase and converge to 1, the MMS value gets closer to $d_0 = 10$, and all method performances are independent of the distribution of the response variable Y and the number of slices.

5. Real data analysis

5.1. Experimental study with real data

5.1.1. Data description

To assess our proposed method, two popular public datasets [24, 25] for cancer classification were utilized. The two datasets are high-dimensional, the variables are of continuous type, and the response variables are binary with values of 0 (normal) and 1 (cancer). The first is the prostate cancer dataset, where the distribution of response variables is roughly balanced, and the second is the B-cell lymphoma dataset, where the distribution of response variables is roughly unbalanced. Details about the two datasets are listed in Table 10.

Table 10. Details of the used datasets.

Dataset type	Number of samples	Number of variables	Classification of samples
Prostate	102	5966	52 Tumor/ 50 Non-tumor
DLBCL	77	6286	58 DLBCL/ 19 FL

5.1.2. Performance evaluation criteria

The dataset is divided into two parts using a 7:3 random ratio, with 70% of the data used as the training dataset and the other 30% as the test dataset. Then, the number of slices $J_k = 4, 8, 10$ is taken to slice the continuous covariates in the training set, respectively. This process can also be thought of as cross-validation, whereby we choose the number of slices at which the methods perform optimally as the optimal categories for the continuous variable. On the training set, variables were screened using the JS-SIS, AJS-SIS, and IG-SIS screening approaches; on the test set, support vector machines were used to assess how well the variables were classified using these techniques. We utilized ten-fold cross-validation to reduce the impact of randomly divided data in the dataset on the model accuracy and repeated it 100 times while testing the classification effect to reduce the random error.

We used two evaluation indexes to assess classification effectiveness: Classification accuracy (CA) and the geometric mean (G-mean) of specificity (SPE) and sensitivity (SEN).

5.1.3. Results and discussion

Tables 11 and 12 show the categorization effects of the variables screened by applying the JS-SIS, AJS-SIS, and IG-SIS methods to the two datasets, respectively. CA1 (G-mean1) and CA2 (G-mean2) denote the index values when the number of screening variables is the first $[n/\log n]$ and the first $2\,[n/\log n]$, respectively.

Table 11. The results of the Prostate dataset.

Method	CA1	CA2	G-mean1	G-mean2
JS-SIS-4	0.884(0.008)	0.964(0.005)	0.873(0.01)	0.968(0.005)
AJS-SIS-4	0.894(0.007)	0.921(0.006)	0.888(0.009)	0.923(0.006)
IG-SIS-4	0.894(0.007)	0.921(0.006)	0.888(0.009)	0.923(0.006)
JS-SIS-8	0.926(0.008)	0.937(0.006)	0.926(0.008)	0.942(0.006)
AJS-SIS-8	0.925(0.007)	0.932(0.006)	0.927(0.008)	0.935(0.006)
IG-SIS-8	0.925(0.007)	0.932(0.006)	0.927(0.008)	0.935(0.006)
JS-SIS-10	0.895(0.01)	0.947(0.006)	0.887(0.011)	0.951(0.006)
AJS-SIS-10	0.896(0.01)	0.94(0.006)	0.897(0.011)	0.94(0.007)
IG-SIS-10	0.898(0.009)	0.94(0.006)	0.897(0.01)	0.94(0.007)

The numbers in parentheses are the corresponding standard deviations.

Table 11 shows the experimental results for the Prostate dataset. Comparison of the categorical prediction performance of the three methods when screening different numbers of variables for each method is as follows: When the number of screening variables is $[n/\log n]$, all three methods have similar CA and G-mean values for the same number of slices, while when the number of screening variables is $2\,[n/\log n]$, JS-SIS has significantly higher CA and G-mean values than AJS-SIS and IG-SIS for the same number of slices, and the CA and G-mean values of AJS-SIS and IG-SIS are the same. Then, comparing the categorical prediction performance of the three methods for variables screened in different numbers of slices, it is found that the fluctuation of CA and G-mean values with the number of slices is higher in JS-SIS than in AJS-SIS and IG-SIS, while the fluctuation of CA and G-mean values is about the same in AJS-SIS and IG-SIS. Overall, JS-SIS is more impacted by the number of screens and slices than IG-SIS is, and AJS-SIS and IG-SIS are both impacted in a similar way. Moreover, regarding the performance of the categorical prediction accuracy of the screened variables, JS-SIS outperforms AJS-SIS and IG-SIS as the number of variables screened increases, while AJS-SIS and IG-SIS perform almost as well.

Table 12. The results of the DLBCL dataset.

Method	CA1	CA2	Gmean1	Gmean2
JS-SIS-4	0.755(0.015)	0.87(0.013)	0.602(0.034)	0.868(0.014)
AJS-SIS-4	0.779(0.014)	0.87(0.013)	0.629(0.033)	0.868(0.014)
IG-SIS-4	0.725(0.013)	0.85(0.012)	0.546(0.033)	0.819(0.022)
JS-SIS-8	0.756(0.011)	0.887(0.008)	0.659(0.025)	0.85(0.013)
AJS-SIS-8	0.754(0.013)	0.91(0.011)	0.69(0.024)	0.884(0.014)
IG-SIS-8	0.718(0.015)	0.89(0.009)	0.653(0.024)	0.873(0.013)
JS-SIS-10	0.779(0.014)	0.775(0.012)	0.624(0.032)	0.624(0.029)
AJS-SIS-10	0.755(0.015)	0.87(0.011)	0.63(0.028)	0.826(0.019)
IG-SIS-10	0.789(0.012)	0.833(0.013)	0.651(0.029)	0.756(0.022)

The numbers in parentheses are the corresponding standard deviations.

The experimental findings for the DLBCL dataset are displayed in Table 12. Comparing the

categorical prediction performance of the three methods with different numbers of variables screened separately, the CA and G-mean values of the three methods are significantly different when the number of screened variables is $[n/\log n]$, respectively. However, when the number of screened variables is $2[n/\log n]$, AJS-SIS's CA and G-mean values are significantly higher than those of JS-SIS and IG-SIS, while JS-SIS and IG-SIS were also significantly different. Comparing the categorical prediction performance of the three methods for variables screened in different numbers of slices, the fluctuations of CA and G-mean values are the largest for JS-SIS, while the fluctuations of CA and G-mean values for AJS-SIS are smaller than for IG-SIS. JS-SIS and AJS-SIS are both more and less affected by the number of screenings and slices, respectively, as compared to IG-SIS. AJS-SIS outperforms JS-SIS and IG-SIS in terms of the categorical prediction accuracy of the screened variables.

5.1.4. Comprehensive analysis of real experimental results

Combining the experimental results of the two real datasets, it can be seen that the methods JS-SIS and AJS-SIS proposed in this paper are very similar to IG-SIS in terms of performance, but AJS-SIS is more robust in terms of performance than JS-SIS and IG-SIS, where JS-SIS is a little bit weaker in terms of robustness compared to IG-SIS. Further, all methods perform better in the Prostate dataset than in the DLBCL dataset. This may be because the DLBCL dataset variables have larger dimensions than the Prostate dataset variables, but smaller sample sizes than the Prostate dataset variables, which may make it more challenging to screen the DLBCL dataset variables. Finally, the predictive effectiveness of the variables screened by all methods improved with the number of variables screened. Overall, experiments with real data illustrate that our methods can be applied to real datasets.

5.2. Applying methods to real data

Based on both numerical simulations and experiments with real data, it is shown that the methods proposed in this paper are well able to screen out important variables. Thus, in this section, we use a biological dataset with detailed information to further explore the specifics of the suggested approaches for screening variables in practical applications. This dataset is available from the R package "colonCA" (https://bioconductor.org/packages/release/data/experiment/html/colonCA.html). Table 13 displays detailed information about the dataset. The sample categories in this dataset are binary variables, and the gene variables are continuous.

Table 13. Details of the used datasets.

Dataset type	Number of samples	Number of variables	Classification of samples
Colon	62	2000	40 Tumor/ 22 Normal

The results of earlier experiments indicate that good results can be obtained when the number of variables screened is $[n/\log n]$. For this reason, we select the essential variables as $[n/\log n]$ when applying the real dataset. Then, the training set and test set were divided in the same manner as the real-data experiments in Section 5. The significant gene variables were screened on the training set using the IG-SIS method and our suggested approaches, with the results displayed in Tables 14 and 15, respectively. Lastly, the test set is utilized to examine the classification effect of the selected gene variables, as shown in Table 16.

Table 14. The top $[n/\log n]$ genes were screened by the JS-SIS method and the AJS-SIS method based on the Colon dataset.

JS-SIS-4	JS-SIS-8	JS-SIS-10	AJS-SIS-4	AJS-SIS-8	AJS-SIS-10
Hsa.549	Hsa.8147	Hsa.36354	Hsa.549	Hsa.8147	Hsa.36354
Hsa.8147	Hsa.1410	Hsa.1588	Hsa.8147	Hsa.692	Hsa.627
Hsa.1410	Hsa.831	Hsa.2588	Hsa.1410	Hsa.1410	Hsa.8147
Hsa.6814	Hsa.549	Hsa.2715	Hsa.6814	Hsa.831	Hsa.2588
Hsa.1682	Hsa.692	Hsa.549	Hsa.1682	Hsa.549	Hsa.2715
Hsa.1832	Hsa.1588	Hsa.627	Hsa.1832	Hsa.2097	Hsa.1588
Hsa.3016	Hsa.823	Hsa.831	Hsa.3016	Hsa.1588	Hsa.549
Hsa.36689	Hsa.6814	Hsa.6814	Hsa.36689	Hsa.823	Hsa.831
Hsa.544	Hsa.1660	Hsa.1776	Hsa.544	Hsa.6814	Hsa.6814
Hsa.5971	Hsa.733	Hsa.37937	Hsa.5971	Hsa.37937	Hsa.37937
Hsa.2097	Hsa.951	Hsa.31943	Hsa.2097	Hsa.1660	Hsa.37541

Table 15. The top $[n/\log n]$ genes were screened by the IG-SIS method based on the Colon dataset.

IG-SIS-4	IG-SIS-8	IG-SIS-10
Hsa.8147	Hsa.8147	Hsa.36354
Hsa.549	Hsa.692	Hsa.8147
Hsa.1410	Hsa.1410	Hsa.627
Hsa.1832	Hsa.831	Hsa.2715
Hsa.6814	Hsa.549	Hsa.2588
Hsa.3016	Hsa.1832	Hsa.1588
Hsa.692.2	Hsa.2097	Hsa.549
Hsa.1682	Hsa.37937	Hsa.831
Hsa.36689	Hsa.823	Hsa.37937
Hsa.544	Hsa.1588	Hsa.6814
Hsa.2291	Hsa.6814	Hsa.31943

Table 16. The results of the Colon dataset.

Method	CA1	Gmean1
JS-SIS-4	0.811(0.012)	0.684(0.025)
AJS-SIS-4	0.801(0.012)	0.657(0.027)
IG-SIS-4	0.769(0.011)	0.557(0.032)
JS-SIS-8	0.759(0.012)	0.606(0.03)
AJS-SIS-8	0.799(0.013)	0.715(0.018)
IG-SIS-8	0.803(0.011)	0.709(0.019)
JS-SIS-10	0.807(0.013)	0.686(0.03)
AJS-SIS-10	0.789(0.01)	0.644(0.025)
IG-SIS-10	0.789(0.01)	0.631(0.027)

The numbers in parentheses are the corresponding standard deviations.

As can be observed from Tables 14 and 15, "Hsa.549" and "Hsa.6814" are the same gene variables among the $[n/\log n]$ gene variables selected by each technique at $J_k = 4, 8, 10$, indicating that they might be the most significant gene variables. Then, according to Table 16, from the perspective of data analysis, it is evident that the Colon dataset exhibits optimal performance for the JS-SIS and AJS-SIS methods at $J_k = 4$, and optimal performance for IG-SIS at $J_k = 8$. Generally, the important gene variables selected by the proposed methods have good classification prediction performance, indicating the utility of the proposed methods.

6. Conclusions

In this research, we established a model-free feature screening procedure based on JS divergence for binary categorical response variables, implying less restrictive data assumptions. We also suggested two different feature screening techniques for binary response variables in various scenarios. When the number of categories for each covariate is the same, the method based on JS divergence is used. Additionally, when the number of categories in each covariate is different, we investigated the AJ-SIS method for screening variables, which uses the logarithmic factor of the number of categories to adjust the JS divergence, which is used for feature screening. Afterward, theoretical proof showed that JS-SIS and AJS-SIS have sure screening and ranking consistency properties. Then, the screening performance of the proposed methods was evaluated through simulation experiments and real data analysis, which showed their effectiveness, availability, and practicality. It is also evident that a variety of data may be widely applicable to our suggested approaches and that they have good screening performance when the data contains categorical covariates, continuous covariates, and both continuous and categorical covariates appearing in the data. We suggested experimenting with several different numbers of slices and applying cross-validation to determine the optimal categories for continuous variables when dealing with continuous covariates. We can see that, in feature screening, the performance of the proposed methods JS-SIS and AJS-SIS in this paper is similar to IG-SIS. But, AJS-SIS performs better than IG-SIS when the covariates have a varying number of categories, particularly when the sample size is small. Moreover, in terms of computational time, JS-SIS and AJS-SIS are both shorter

than IG-SIS. In addition to this, our method's perspective is appealing, and since JS divergence does not require the probability distributions to be absolutely continuous and has the advantages of symmetry, non-negativity, and boundedness, it can measure the differences between probability distributions very effectively. Finally, the methods proposed in this work only take into account situations when the response variable is binary, although, in reality, multicategorical responses are very frequent. Therefore, we will extend JS-SIS and AJS-SIS to situations in which the response variable is multicategorical in our future research.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

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

The authors would like to thank the editor and anonymous referees for their constructive comments and suggestions.

Conflict of interest

The author certifies that the publication of this paper does not involve any conflicts of interest.

References

1. J. Q. Fan, J. C. Lv, Sure independence screening for ultrahigh dimensional feature space, *J. R. Statist. Soc. B.*, **70** (2008), 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x
2. 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
3. G. X. Li, H. Peng, J. Zhang, L. X. Zhu, Robust rank correlation based screening, *Ann. Statist.*, **40** (2012), 1846–1877. https://doi.org/10.1214/12-AOS1024
4. J. Q. Fan, R. Song, Sure independence screening in generalized linear models with NP-dimensionality, *Ann. Statist.*, **38** (2010), 3567–3604. https://doi.org/10.1214/10-AOS798
5. J. Q. Fan, Y. Feng, R. Song, Nonparametric independence screening in sparse ultra-high-dimensional additive models, *J. Am. Stat. Assoc.*, **106** (2011), 544–557. https://doi.org/10.1198/jasa.2011.tm09779
6. J. Y. Liu, R. Z. Li, R. L. Wu, Feature selection for varying coefficient models with ultrahigh-dimensional covariates, *J. Am. Stat. Assoc.*, **109** (2014), 266–274. https://doi.org/10.1080/01621459.2013.850086
7. H. Liang, H. S. Wang, C. L. Tsai, Profiled forward regression for ultrahigh dimensional variable screening in semiparametric partially linear models, *Stat. Sinica*, **22** (2012), 531–554. https://doi.org/10.5705/ss.2010.134

8. 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
9. R. Z. Li, W. Zhong, L. P. Zhu, Feature screening via distance correlation learning, *J. Am. Stat. Assoc.*, **107** (2012), 1129–1139. https://doi.org/10.1080/01621459.2012.695654
10. X. He, L. Wang, H. G. Hong, Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data, *Ann. Statist.*, **41** (2013), 342–369. https://doi.org/10.1214/13-AOS1087
11. W. L. Pan, X. Q. Wang, W. N. Xiao, H. T. Zhu, A generic sure independence screening procedure, *J. Am. Stat. Assoc.*, **114** (2018), 928–937. https://doi.org/10.1080/01621459.2018.1462709
12. J. Q. Fan, Y. Y. Fan, High-dimensional classification using features annealed independence rules, *Ann. Statist*, **36** (2008), 2605–2637. https://doi.org/10.1214/07-AOS504
13. 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
14. H. J. Cui, R. Z. Li, W. Zhong, Model-free feature screening for ultrahigh dimensional discriminant analysis, *J. Am. Stat. Assoc.*, **110** (2015), 630–641. https://doi.org/10.1080/01621459.2014.920256
15. D. Y. Huang, R. Z. Li, H. S. 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
16. L. Ni, F. Fang, Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification, *J. Nonparametr Stat.*, **28** (2016), 515–530. https://doi.org/10.1080/10485252.2016.1167206
17. F. Y. Xiao, Multi-sensor data fusion based on the belief divergence measure of evidences and the belief entropy, *Inform. Fusion.*, **46** (2019), 23–32. https://doi.org/10.1016/j.inffus.2018.04.003
18. F. Y. Xiao, A new divergence measure for belief functions in D-S evidence theory for multisensor data fusion, *Inform. Sciences*, **514** (2020), 462–483. https://doi.org/10.1016/j.ins.2019.11.022
19. F. Y. Xiao, GEJS: A generalized evidential divergence measure for multisource information fusion, *IEEE T. Syst. Man Cy-S.*, **53** (2022), 2246–2258. https://doi.org/10.1109/TSMC.2022.3211498
20. F. Y. Xiao, J. H. Wen, W. Pedrycz, Generalized divergence-based decision making method with an application to pattern classification, *IEEE T. Knowl. Data En.*, **35** (2022), 6941–6956. https://doi.org/10.1109/TKDE.2022.3177896
21. J. Lin, Divergence measures based on the shannon entropy, *IEEE Trans. Inform. Theory*, **37** (1991), 145–151. https://doi.org/10.1109/18.61115
22. C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.*, **27** (1948), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
23. H. J. He, G. M. Deng, Grouped feature screening for ultra-high dimensional data for the classification model, *J. Stat. Comput. Sim.*, **92** (2022), 974–997. https://doi.org/10.1080/00949655.2021.1981901
24. D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, et al., Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, **1** (2002), 203–209. https://doi.org/10.1016/S1535-6108(02)00030-2

25. M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, et al., Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, *Nat. Med.*, **8** (2002), 68–74. https://doi.org/10.1038/news011227-7

26. W. Hoeffding, Probability inequalities for sums of bounded random variables, *J. Am. Stat. Assoc.*, **58** (1963), 13–30. https://doi.org/10.1080/01621459.1963.10500830

Appendix

To prove Theorem 3.1, the following four lemmas are first introduced.

Lemma 1. *Suppose that $x_1, x_2, \ldots, x_n$ are mutually independent random variables with sample size n and $\Pr(x_i \in [a_i, b_i]) = 1, 1 \le i \le n$, where a_i, b_i are constants. Let $\bar{x} = 1/n \sum_{i=1}^{n} x_i$. Then there exists a constant t for which the following inequality holds:*

$$\Pr(|\bar{x} - E(\bar{x}) \ge t|) \le 2\exp\left(-2nt^2 \Big/ \sum_{i=1}^{n}(b_i - a_i)^2\right).$$

The proof of Lemma 1 is given in [26].

Lemma 2. *Suppose a and b are two bounded random variables, and there exist constants $M_1 > 0, M_2 > 0$ such that $|a| \le M_1, |b| \le M_2$. Given a sample size n, the estimates corresponding to a, b can be obtained as $\hat{A}, \hat{B}$. Suppose that, $\forall \varepsilon \in (0, 1)$, there exist positive constants c_1, c_2 and s such that*

$$\Pr\left(\left|\hat{A} - a\right| \ge \varepsilon\right) \le c_1\left(1 - \frac{\varepsilon s}{c_1}\right)^n,$$

$$\Pr\left(\left|\hat{B} - b\right| \ge \varepsilon\right) \le c_2\left(1 - \frac{\varepsilon s}{c_2}\right)^n.$$

Then, we have

$$\Pr\left(\left|\hat{A}\hat{B} - ab\right| \ge \varepsilon\right) \le C_1\left(1 - \frac{\varepsilon s}{C_1}\right)^n,$$

$$\Pr\left(\left|\hat{A}^2 - a^2\right| \ge \varepsilon\right) \le C_2\left(1 - \frac{\varepsilon s}{C_2}\right)^n,$$

$$\Pr\left(\left|\left(\hat{A} - a\right) - \left(\hat{B} - b\right)\right| \ge \varepsilon\right) \le C_3\left(1 - \frac{\varepsilon s}{C_3}\right)^n,$$

where, $C_1 = \max\{2c_1 + c_2, c_1 + 2c_2 + 2c_2M_1, 2c_2M_2\}$, $C_2 = \max\{3c_1 + 2c_1M_1, 2c_2M_2\}$, $C_3 = \max\{2c_1, 2c_2, c_1 + c_2\}$.

Furthermore, assuming that b is bounded and non-zero, and that there exists $M_3 > 0$ such that $|b| \ge M_3$, then we have

$$\Pr\left(\left|\frac{\hat{A}}{\hat{B}} - \frac{\hat{a}}{\hat{b}}\right| \ge \varepsilon\right) \le C_4\left(1 - \frac{\varepsilon s}{C_4}\right)^n$$

where $C_4 = \max\{c_1 + c_2 + c_5, c_2/M_4, 2c_2M_1/(M_2M_4)\}$, $c_5 > 0$ and $M_4 > 0$.

The proof of Lemma 2 is given in [15].

Lemma 3. *When the covariates are categorical, we know that* $e_j \geq 0$. *Only if* $\Pr\left(\mathbf{x}_j = l \mid Y = 1\right) = \Pr\left(\mathbf{x}_j = l \mid Y = 2\right)$, *we have* $e_j = 0$, *that is, Y and* $\mathbf{x}_j$ *are independent.*

Proof of Lemma 3: Let $G = \Pr\left(\mathbf{x}_j = l \mid Y = 1\right)$, $Q = \Pr\left(\mathbf{x}_j = l \mid Y = 2\right)$, $M = \frac{1}{2}(G + Q)$. Define $f(x) = x\log(x)$. Since $f(x)$ is a hypoconvex function, and through Jensen's inequality, we can get

$$\begin{aligned} H(M) =& -\sum_{j=1}^{p} M\log(M) \\ =& -\sum_{j=1}^{p} f(M) \\ \geq& -\left(\sum_{j=1}^{p}\frac{1}{2}f(G) + \sum_{j=1}^{p}\frac{1}{2}f(Q)\right) \\ =& \frac{1}{2}H(G) + \frac{1}{2}H(Q). \end{aligned}$$

Thus,

$$\begin{aligned} e_j =& \frac{1}{2}\sum_{j=1}^{p} G\log(G) - \frac{1}{2}\sum_{j=1}^{p} G\log(M) + \frac{1}{2}\sum_{j=1}^{p} Q\log(Q) - \frac{1}{2}\sum_{j=1}^{p} Q\log(M) \\ =& -\frac{1}{2}\sum_{j=1}^{p}(G+Q)\left(\frac{G}{G+Q}\log(M) + \frac{Q}{G+Q}\log(M)\right) + \frac{1}{2}\sum_{j=1}^{p} G\log(G) + \frac{1}{2}\sum_{j=1}^{p} Q\log(Q) \\ =& -\frac{1}{2}\sum_{j=1}^{p}(G+Q)\log(M) + \frac{1}{2}\sum_{j=1}^{p} G\log(G) + \frac{1}{2}\sum_{j=1}^{p} Q\log(Q) \\ =& H(M) - \frac{1}{2}H(G) - \frac{1}{2}H(Q) \\ \geq& 0. \end{aligned}$$

So, e_j is larger than or equal to 0. The equation holds if and only if $G = Q$. And, when $G = Q$,

$$\begin{aligned} \Pr(\mathbf{x}_j = l) =& \Pr(Y = 1)G + \Pr(Y = 2)Q \\ =& G \\ =& Q. \end{aligned}$$

By the condition of independence, it can be inferred that Y and $\mathbf{x}_j$ are independent. Thus, when there is a nonlinear relationship between Y and $\mathbf{x}_j$, the conditional probability distribution between them will not satisfy independence, and thus the JS divergence will be larger than zero.

Lemma 4. *When the covariates are continuous, it follows from the proof of Proposition 2.2 in [16] that for a continuous variable X, there exists a sequence* $\{x_m, m = 1, 2, \dots\}$ *such that* x_m *is a quantile of X and* $\lim_{n\to\infty} x_m = X$, *and there exist* J_k *and J such that* $x = q_{(J)}$. *Then,* $\Pr(X \leq x \mid Y = r) = J/J_k$ *does not depend on r. There is* $e_j \geq 0$, *and when Y and* $\mathbf{x}_j$ *are independent,* $e_j = 0$.

The proof of Lemma 4 is similar to the proof of Proposition 2.2 in [16], so it is omitted here.

Proof of Theorem 3.1:

$$\begin{aligned}
e_j =& JS\left(G \,\|\, Q\right)\\
=& \frac{1}{2}\sum_{j=1}^{p} G\log\left(\frac{G}{M}\right) + \frac{1}{2}\sum_{j=1}^{p} Q\log\left(\frac{Q}{M}\right)\\
=& \frac{1}{2}\sum_{j=1}^{p} G\log(G) - \frac{1}{2}\sum_{j=1}^{p} G\log(M) + \frac{1}{2}\sum_{j=1}^{p} Q\log(Q) - \frac{1}{2}\sum_{j=1}^{p} Q\log(M)\\
=& \frac{1}{2}\left(H\left(G,M\right) - H(G)\right) + \frac{1}{2}\left(H\left(Q,M\right) - H(Q)\right).
\end{aligned}$$

According to the definitions of e_j and $\hat{e}_j$ we have

$$\begin{aligned}
\left|e_j - \hat{e}_j\right| =& \frac{1}{2}\left|\left[\left(H\left(G,M\right) - H(G)\right)\right) + \left(H\left(Q,M\right) - H(Q)\right)\right]\right.\\
&\left. - \left[\left(\hat{H}\left(G,M\right) - \hat{H}(G)\right) + \left(\hat{H}\left(Q,M\right) - \hat{H}(Q)\right)\right]\right|\\
=& \frac{1}{2}\left|\left(H\left(G,M\right) - \hat{H}\left(G,M\right)\right) + \left(H\left(Q,M\right) - \hat{H}\left(Q,M\right)\right)\right.\\
&\left. - \left(H(G) - \hat{H}(G)\right) - \left(H(Q) - \hat{H}(Q)\right)\right|\\
\leq& \frac{1}{2}\left|H\left(G,M\right) - \hat{H}\left(G,M\right)\right| + \frac{1}{2}\left|H\left(Q,M\right) + \hat{H}\left(Q,M\right)\right|\\
&+ \frac{1}{2}\left|H(G) - \hat{H}(G)\right| + \frac{1}{2}\left|H(Q) - \hat{H}(Q)\right|
\end{aligned}$$

and

$$\begin{aligned}
\Pr\left(\left|e_j - \hat{e}_j\right| > \varepsilon\right) \leq& \Pr\left(\frac{1}{2}\left\|H\left(G,M\right) - \hat{H}\left(G,M\right)\right| + \left|H\left(Q,M\right) + \hat{H}\left(Q,M\right)\right|\right.\\
&\left. + \left|H(G) - \hat{H}(G)\right| + \left|H(Q) - \hat{H}(Q)\right\| > \varepsilon\right)\\
\leq& \Pr\left(\left|H\left(G,M\right) - \hat{H}\left(G,M\right)\right| > \frac{\varepsilon}{2}\right) + \Pr\left(\left|H\left(G\right) - \hat{H}\left(G\right)\right| > \frac{\varepsilon}{2}\right)\\
&+ \Pr\left(\left|H\left(Q,M\right) - \hat{H}\left(Q,M\right)\right| > \frac{\varepsilon}{2}\right) + \Pr\left(\left|H\left(Q\right) - \hat{H}\left(Q\right)\right| > \frac{\varepsilon}{2}\right)\\
=:& E_{j1} + E_{j2} + E_{j3} + E_{j4}.
\end{aligned}$$

Next, we prove that $E_{j1} \le 2L\exp\left\{-n\varepsilon^2\Big/2L^2\right\}$.

$$\begin{aligned}
&\Pr\left(\left|H(G,M)-\hat{H}(G,M)\right|>\frac{\varepsilon}{2}\right)\\
=&\Pr\left(\left|\sum_{l=1}^{L}\hat{p}\left(\mathbf{x}_j=l|Y=1\right)\log\left(\frac{\hat{p}\left(\mathbf{x}_j=l|Y=1\right)+\hat{p}\left(\mathbf{x}_j=l|Y=2\right)}{2}\right)\right.\right.\\
&\left.\left.-\sum_{l=1}^{L}p\left(\mathbf{x}_j=l|Y=1\right)\log\left(\frac{p\left(\mathbf{x}_j=l|Y=1\right)+p\left(\mathbf{x}_j=l|Y=2\right)}{2}\right)\right|>\frac{\varepsilon}{2}\right)\\
\le&L\max_{l}\Pr\left(\left|\hat{p}\left(\mathbf{x}_j=l|Y=1\right)\log\left(\frac{\hat{p}\left(\mathbf{x}_j=l|Y=1\right)+\hat{p}\left(\mathbf{x}_j=l|Y=2\right)}{2}\right)\right.\right.\\
&\left.\left.-p\left(\mathbf{x}_j=l|Y=1\right)\log\left(\frac{p\left(\mathbf{x}_j=l|Y=1\right)+p\left(\mathbf{x}_j=l|Y=2\right)}{2}\right)\right|>\frac{\varepsilon}{2L}\right).
\end{aligned}$$

Using the sample frequency to estimate the probability, we have

$$\hat{p}\left(\mathbf{x}_j=l\mid Y=1\right)=\sum_{i=1}^{n}I\left(x_{ij}=l\right)I\left(y_i=1\right)\Big/\sum_{i=1}^{n}I\left(y_i=1\right),$$

$$\hat{p}\left(\mathbf{x}_j=l\mid Y=2\right)=\sum_{i=1}^{n}I\left(x_{ij}=l\right)I\left(y_i=2\right)\Big/\sum_{i=1}^{n}I\left(y_i=2\right),$$

$$p\left(\mathbf{x}_j=l\mid Y=1\right)=E\left(I\left(x_{ij}=l\right)I\left(y_i=1\right)\right)\Big/p\left(I\left(y_i=1\right)\right),$$

$$p\left(\mathbf{x}_j=l\mid Y=2\right)=E\left(I\left(x_{ij}=l\right)I\left(y_i=2\right)\right)\Big/p\left(I\left(y_i=2\right)\right).$$

So, we have

$$\begin{aligned}
&\Pr\left(\left|\hat{p}\left(\mathbf{x}_j=l\mid Y=1\right)-p\left(\mathbf{x}_j=l\mid Y=1\right)\right|>\varepsilon_1\right)\\
&=\Pr\left(\left|\sum_{i=1}^{n}I\left(x_{ij}=l\right)I\left(y_i=1\right)\Big/\sum_{i=1}^{n}I\left(y_i=1\right)-E\left(I\left(x_{ij}=l\right)I\left(y_i=1\right)\right)\Big/p\left(I\left(y_i=1\right)\right)\right|>\varepsilon_1\right)\\
&=:\Pr\left(\left|\frac{S_n}{T_n}-\frac{s_n}{t_n}\right|\ge\varepsilon_1\right)
\end{aligned}$$

and because S_n, T_n is an estimate of s_n, t_n, it follows from Lemmas 1 and 2 that

$$\Pr\left(|S_n-s_n|>\varepsilon_2\right)\ge 2\exp\left\{-2n\varepsilon_2^2\right\},$$

$$\Pr\left(|T_n-t_n|>\varepsilon_2\right)\ge 2\exp\left\{-2n\varepsilon_2^2\right\}.$$

Thus, there is a convergence of $\hat{p}\left(\mathbf{x}_j = l \mid Y = 1\right)$ by probability to its probability function $p\left(\mathbf{x}_j = l \mid Y = 1\right)$, that is

$$\Pr\left(\left|\hat{p}\left(\mathbf{x}_j = l \mid Y = 1\right) - p\left(\mathbf{x}_j = l \mid Y = 1\right)\right| > \varepsilon_1\right) \leq 2\exp\left\{-2n\varepsilon_1^2\right\}.$$

Similarly, $\hat{p}\left(\mathbf{x}_j = l \mid Y = 2\right)$ also converges to $p\left(\mathbf{x}_j = l \mid Y = 2\right)$.

It can also be shown that $\log\left(\hat{p}\left(\mathbf{x}_j = l \mid Y = 1\right)\right)$ converges probabilistically to $\log\left(p\left(\mathbf{x}_j = l \mid Y = 1\right)\right)$. Let $\hat{p}^* = \hat{p}\left(\mathbf{x}_j = l \mid Y = 1\right)$, $p^* = p\left(\mathbf{x}_j = l \mid Y = 1\right)$:

$$\begin{aligned}
&\Pr\left(\left|\log\left(\hat{p}^*\right) - \log\left(p^*\right)\right| > \varepsilon_3\right) \\
&= \Pr\left(\left|\log((\hat{p}^* - p^*) + p^*) - \log(p^*)\right| > \varepsilon_3\right) \\
&\leq \Pr\left(\left|\log(p^*) + \frac{1}{p^*}(\hat{p}^* - p^*) + o(\hat{p}^* - p^*) - \log(p^*)\right| > \varepsilon_3\right) \\
&\leq \Pr\left(|\hat{p}^* - p^*| > \varepsilon_3 p^* - o(\hat{p}^* - p^*)\right)
\end{aligned}$$

Thus, we can get that $\log\left(\hat{p}\left(\mathbf{x}_j = l \mid Y = 1\right)\right)$ converges to $\log\left(p\left(\mathbf{x}_j = l \mid Y = 1\right)\right)$ with probability.

In a similar proof, we can get that $\log(\hat{p}\left(\mathbf{x}_j = l \mid Y = 2\right))$ converges to $\log(p\left(\mathbf{x}_j = l \mid Y = 2\right))$ with probability and $\log\left(\frac{\hat{p}\left(\mathbf{x}_j = l|Y=1\right) + \hat{p}\left(\mathbf{x}_j = l|Y=2\right)}{2}\right)$ converges to $\log\left(\frac{p\left(\mathbf{x}_j = l|Y=1\right) + p\left(\mathbf{x}_j = l|Y=2\right)}{2}\right)$ with probability.

Thus, we can obtain $E_{j1} \leq 2L\exp\left\{-n\varepsilon^2\middle/2L^2\right\}$. Similarly, the other three parts can be proved: $E_{j2} \leq 2L\exp\left\{-n\varepsilon^2\middle/2L^2\right\}$, $E_{j3} \leq 2L\exp\left\{-n\varepsilon^2\middle/2L^2\right\}$, $E_{j4} \leq 2L\exp\left\{-n\varepsilon^2\middle/2L^2\right\}$.

Thus, for $0 < \epsilon_4 < 1$, there is

$$\Pr\left(\left|e_j - \hat{e}_j\right| > \epsilon_4\right) \leq 8L\exp\left\{-n\epsilon_4^2\middle/2L^2\right\}. \tag{1}$$

For $0 \leq \tau < 1/2$, there exists a positive number c with

$$\Pr\left(\left|e_j - \hat{e}_j\right| > cn^{-\tau}\right) \leq 8L\exp\left\{-c^2n^{1-2\tau}\middle/2L^2\right\} \tag{2}$$

and then

$$\begin{aligned}
\Pr\left(\max_{1\leq j\leq p}\left|e_j - \hat{e}_j\right| > cn^{-\tau}\right) &\leq \Pr\left(\bigcup_{j=1}^{p}\left|e_j - \hat{e}_j\right| > cn^{-\tau}\right) \\
&\leq p\Pr\left(\left|e_j - \hat{e}_j\right| > cn^{-\tau}\right) \\
&\leq 8pL\exp\left\{-c^2n^{1-2\tau}\middle/2L^2\right\}.
\end{aligned} \tag{3}$$

By (3), for $0 < \delta < 1 - 2\tau$, we have

$$\Pr\left(\max_{1\leq j\leq p}\left|e_j - \hat{e}_j\right| > cn^{-\tau}\right) \to 0 \tag{4}$$

with $n \to \infty$. Thus, by (4), we have

$$\begin{aligned}\Pr(D \subseteq \hat{D}) &\geq \Pr(\left|e_j - \hat{e}_j\right| > cn^{-\tau}, \forall j \in D)\\ &\geq \Pr\left(\max_{j\in D}\left|e_j - \hat{e}_j\right| > cn^{-\tau}\right)\\ &\geq 1 - d_0 \Pr(\left|e_j - \hat{e}_j\right| > cn^{-\tau})\\ &\geq 1 - 8d_0 L\exp\left\{-c^2 n^{1-2\tau}\Big/2L^2\right\}.\end{aligned} \tag{5}$$

Therefore, $\Pr(D \subseteq \hat{D}) \to 1$, with $n \to \infty$.

Thus, by Theorem 3.1 the sure screening property holds under conditions (C1)–(C3).

Proof of Theorem 3.4:

Since $\min_{j\in D} e_j - \max_{j\in D^c} e_j > 0$, there exists $\delta > 0$ such that $\min_{j\in D} e_j - \max_{j\in D^c} e_j = \delta$, and then we have

$$\begin{aligned}\Pr\left(\min_{j\in D}\hat{e}_j \leq \max_{j\in D^c}\hat{e}_j\right) &= \Pr\left(\min_{j\in D}\hat{e}_j - \max_{j\in D^c} e_j \leq \max_{j\in D^c}\hat{e}_j - \max_{j\in D^c} e_j\right)\\ &= \Pr\left(\min_{j\in D}\hat{e}_j - \min_{j\in D} e_j + \delta \leq \max_{j\in D^c}\hat{e}_j - \max_{j\in D^c} e_j\right)\\ &= \Pr\left(\min_{j\in D}\hat{e}_j - \min_{j\in D} e_j - \max_{j\in D^c}\hat{e}_j + \max_{j\in D^c} e_j \leq -\delta\right)\\ &= \Pr\left(\left|\left(\min_{j\in D}\hat{e}_j - \max_{j\in D^c}\hat{e}_j\right) - \left(\min_{j\in D} e_j - \max_{j\in D^c} e_j\right)\right| \geq \delta\right)\\ &\leq \Pr\left(\max_{1\leq j\leq p}\left|e_j - \hat{e}_j\right| \geq \delta/2\right)\\ &\leq 8pL\exp\left\{-n\delta^2\Big/2\right\}.\end{aligned}$$

From Fatou's Lemma, we have

$$\begin{aligned}&\Pr\left\{\liminf_{n\to\infty}\left(\min_{j\in D}\hat{e}_j - \max_{j\in D^c}\hat{e}_j\right) \leq 0\right\}\\ &\leq \liminf_{n\to\infty}\Pr\left\{\left(\min_{j\in D}\hat{e}_j - \max_{j\in D^c}\hat{e}_j\right) \leq 0\right\}\\ &= 0.\end{aligned}$$

So,

$$\Pr\left\{\liminf_{n\to\infty}\left(\min_{j\in D}\hat{e}_j - \max_{j\in D^c}\hat{e}_j\right) \leq 0\right\} = 1. \tag{6}$$

Thus, Theorem 3.4 holds.

Proof of Theorem 3.2: Let $F_k(x \mid y)$ be the cumulative distribution function of (X_k, Y) and $\hat{F}_k(x \mid y)$ be the empirical cumulative distribution function of (X_k, Y). Then, by the proof of Lemma A.2 in [16], we can similarly show that, under conditions (C4) and (C5), $\forall \epsilon_5 > 0, 1 \leq r \leq R, 1 \leq J \leq J_k$, we have

$$\Pr\left(\left|\hat{F}_k(\hat{q}_{k,(J)} \mid r) - F_k(q_{k,(J)} \mid r)\right| > \epsilon_5\right) \leq c_6\exp\left\{-c_7 n^{1-2\rho}\epsilon_5^2\right\}$$

where $c_6 = 3c_8$ and $c_7 = \min\left\{1/2, c_4^2/2c_3^2\right\}$ are positive constants.

Thus, $\hat{F}_k(\hat{q}_{k,(J)} \mid r)$ converges to $F_k(q_{k,(J)} \mid r)$ with probability. The proof of the other parts is the same as in Theorem 3.1. Then, for $0 < \epsilon_6 < 1$, we have

$$\Pr\left(|e_k - \hat{e}_k| > \epsilon_6\right) \leq 4c_6 J_k \exp\left\{\frac{-c_7 n^{1-2\rho}\epsilon_6^2}{4J_k^2}\right\}. \tag{7}$$

Equation (7) is similar to the proof process of Eq (1) and will not be proved here.

Under condition (C6), there exists a constant c_5 such that

$$\begin{aligned}\Pr\left(\max_{1\leq j\leq p}|e_k - \hat{e}_k| > c_5 n^{-\tau}\right) &\leq \Pr\left(\bigcup_{j=1}^{p}|e_k - \hat{e}_k| > c_5 n^{-\tau}\right)\\ &\leq p\Pr\left(|e_k - \hat{e}_k| > c_5 n^{-\tau}\right)\\ &\leq 4c_6 p J_k \exp\left\{\frac{-c_7 c_5^2 n^{1-2\rho-2\tau}}{4J_k^2}\right\}.\end{aligned} \tag{8}$$

With $n \to \infty$, we have

$$\Pr\left(\max_{1\leq k\leq p}|e_k - \hat{e}_k| > c_5 n^{-\tau}\right) \to 0$$

$$\begin{aligned}\Pr(D \subseteq \hat{D}) &\geq \Pr(|e_k - \hat{e}_k| > c_5 n^{-\tau}, \forall k \in D)\\ &\geq \Pr\left(\max_{k\in D}|e_k - \hat{e}_k| > c_5 n^{-\tau}\right)\\ &\geq 1 - d_0\Pr(|e_k - \hat{e}_k| > c_5 n^{-\tau})\\ &\geq 1 - 4c_6 d_0 J_k \exp\left\{\frac{-c_7 c_5^2 n^{1-2\rho-2\tau}}{4J_k^2}\right\}.\end{aligned} \tag{9}$$

Therefore, $\Pr(D \subseteq \hat{D}) \to 1$, $n \to \infty$.

Thus, the proof of Theorem 3.5 is similar to that of Eq (6), and we have

$$\Pr\left\{\liminf_{n\to\infty}\left(\min_{k\in D}\hat{e}_k - \max_{k\in D^c}\hat{e}_k\right) \leq 0\right\} = 1. \tag{10}$$

Theorems 3.3 and 3.6 combine Theorems 3.1, 3.2, 3.4 and 3.5, and the proof process is similar to them, so they will not be proven in detail.