Scientific Research Publishing

Open Journal of Statistics, 2023, 13, 822-849
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Model-Free Ultra-High-Dimensional Feature Screening for Multi-Classified Response Data Based on Weighted Jensen-Shannon Divergence

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

[1]School of Mathematics and Statistics, Guilin University of Technology, Guilin, China
[2]Applied Statistics Institute, Guilin University of Technology, Guilin, China
Email: *dgm@glut.edu.cn

How to cite this paper: Jiang, Q.Q. and Deng, G.M. (2023) Model-Free Ultra-High-Dimensional Feature Screening for Multi-Classified Response Data Based on Weighted Jensen-Shannon Divergence. *Open Journal of Statistics*, **13**, 822-849.
https://doi.org/10.4236/ojs.2023.136042

Received: October 23, 2023
Accepted: December 11, 2023
Published: December 14, 2023

Abstract

In ultra-high-dimensional data, it is common for the response variable to be multi-classified. Therefore, this paper proposes a model-free screening method for variables whose response variable is multi-classified from the point of view of introducing Jensen-Shannon divergence to measure the importance of covariates. The idea of the method is to calculate the Jensen-Shannon divergence between the conditional probability distribution of the covariates on a given response variable and the unconditional probability distribution of the covariates, and then use the probabilities of the response variables as weights to calculate the weighted Jensen-Shannon divergence, where a larger weighted Jensen-Shannon divergence means that the covariates are more important. Additionally, we also investigated an adapted version of the method, which is to measure the relationship between the covariates and the response variable using the weighted Jensen-Shannon divergence adjusted by the logarithmic factor of the number of categories when the number of categories in each covariate varies. Then, through both theoretical and simulation experiments, it was demonstrated that the proposed methods have sure screening and ranking consistency properties. Finally, the results from simulation and real-dataset experiments show that in feature screening, the proposed methods investigated are robust in performance and faster in computational speed compared with an existing method.

Keywords

Ultra-High-Dimensional, Multi-Classified, Weighted Jensen-Shannon Divergence, Model-Free, Feature Screening

1. Introduction

In fields like tumor classification, genomics, and machine learning, the issue of processing ultra-high-dimensional data is frequently faced. According to [1] definition of ultra-high-dimensional data, it is assumed that the sample size and the dimensionality of the covariates are n and p, respectively. There exists a constant $\alpha \in (0, 1/2)$ such that $\ln p = O(n^{\alpha})$, and at this point, p exhibits an exponential order of increase with the sample size n. Furthermore, this data tends to be sparse; the number of variables tends to be very high, while the number of variables that have a significant impact is very small. Therefore, in the problem of ultra-high-dimensional data analysis, the development of fast and effective variable screening methods to rapidly reduce ultra-high-dimensional data to reasonable dimensions is very important research.

To address this problem, Fan and Lv [2] first proposed the SIS method for variable screening of ultra-high-dimensional data. Afterwards, many scholars of statistics studied the problem and established a series of feature screening methods. The ultra-high-dimensional feature screening methods that have been developed are: Fan and Song [3] proposed a method of variable screening (MMLE), whereby variable screening is performed by ranking the very large marginal likelihood estimates in a generalized linear model. Fan *et al.* [4] proposed a nonparametric independent screening (NIS) method to investigate variable screening methods in additivity models, using B-spline basis functions to fit the edge nonparametric components. Subsequently, besides the additive model, the variable coefficient model is another widely used nonparametric model. Liu *et al.* [5] further proposed a new variable screening method based on conditional correlation coefficients for variable coefficient models. For the semiparametric model, Li *et al.* [6] utilized a robust rank correlation screening (RRCS) method based on the Kendall τ correlation coefficient. Most of these methods imply the assumption that the response variable is continuous, but ultra-high-dimensional data with categorical response variables are increasingly appearing in various fields of scientific research, and if traditional categorization methods such as logistic regression, decision trees, and support vector machines are used to solve this kind of problem, they will encounter problems such as long time-consuming, high computational costs, and reduced prediction accuracy. Based on this, many researchers have proposed various screening methods for ultra-high-dimensional data where the response variable is categorical. Fan and Fan [7] suggested a feature screening technique based on marginal t-tests for normal distributions for response variables that are binary. However, the robustness of this method is low, so Mai and Zou [8] proposed a screening method for ultra-high-dimensional binary categorical variables based on the Kolmogorov-Smirnov statistic. In practice, the response variable is multi-categorical, which is also very common. For response variables that are multi-classified, Cui *et al.* [9] establish a robust screening method by constructing the distance between the global distribution function and the conditional distribution function. Huang *et al.* [10] proposed an ultra-high-dimensional multi-classified variable screening method based on Pearson's chi-square statistic (PC-SIS).

The above methods of variable screening are based on the correlation between explanatory and response variables. With the development of information theory and the disciplinary integration with statistics, the characteristics of information entropy and the entropy family are recognized and applied by researchers. Ni and Fang [11] proposed a method for ultra-high-dimensional variable screening based on information gain (IG-SIS) from the perspective of information quantity. Jensen-Shannon divergence is an information theory-based concept that plays an important role in calculating similarity and comparing differences in probability distributions and is characterized by non-negativity and symmetry. For ease of reading, Jensen-Shannon divergence is abbreviated to JS divergence in this paper. When the response variable is a binary categorical variable, there are two conditional probability distributions for $\mathbf{x}_j$ given Y. The degree of difference between these two conditional probability distributions can be measured using JS divergence, and the magnitude of JS divergence represents the degree of strength of the correlation between $\mathbf{x}_j$ and Y.

Therefore, on the basis of the above research on ultra-high-dimensional feature screening for response variables that are categorical, in this paper, from a new perspective, we propose a model-free ultra-high-dimensional feature screening method for multi-classified response data based on weighted JS divergence, defined as WJS-SIS. The idea of the method is to first calculate the JS divergence between the conditional probability distribution of $\mathbf{x}_j$ and the unconditional probability distribution of $\mathbf{x}_j$ given $Y = r\left(r = 1, 2, \cdots, R\right)$ between the conditional probability distribution of $\mathbf{x}_j$ and the unconditional probability distribution of $\mathbf{x}_j$ conditionally, and then use $\Pr\left(Y = r\right)$ as the weight to calculate the weighted JS divergence. And, when the number of categories in each covariate is different, using the logarithmic factor of the number of categories in each covariate to adjust the weighted JS divergence is also proposed to measure the relationship between the covariates and the response variable, defined as AWJS-SIS. Theoretically, both WJS-SIS and AWJS-SIS have sure screening properties and ranking consistency, and from the results of Monte Carlo simulations and real data experiments, they have significant effects on screening ultra-high-dimensional multi-classified response variable data. At the same time, they are model-free screening methods that do not depend on any model assumptions.

The rest of the paper is organized as follows: Section 2 describes the proposed WJS-SIS and AWJS-SIS methods in detail. Section 3 describes the screening and ranking consistency of the methods. Section 4 and Section 5 give the simulation study and an experiment with real data, respectively. Section 6 draws conclusions. All theorem proofs are given in the appendix.

2. Method

2.1. Basic Assumption

Suppose $X = \left(x_{i1}, x_{i2}, \cdots, x_{ij}\right)$ is an $N \times P$-dimensional covariate matrix, where X obeys the assumption of independent identical distribution, let $\mathbf{x}_j = \left\{x_{1j}, x_{2j}, \cdots, x_{ij}\right\}, i = 1,2,\cdots,N; j = 1,2,\cdots,P$. And $Y = \left(y_1, y_2, \cdots, y_N\right)$ is an $N \times 1$-dimensional categorical response variable.

Define D as the set of important covariates, D^c as the set of unimportant covariates, and $|D| = d_0$ as the number of variables in the set of important covariates, which is expressed as:

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

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

2.2. Information Entropy

Information entropy is a measure of the index of information proposed by [12]. When the covariate $\mathbf{x}_j \in \{1,2,\cdots,L\}$ and the response variable $Y \in \{1,2,\cdots,R\}$, the information entropy of the $\mathbf{x}_j$ and Y are:

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

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

where the logarithmic base is 2, and $0 \times \log 0 = 0$. And, where the expressions for p_r and $p_{j,l}$ are as follows:

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

$$\hat{p}_r = \frac{\sum_{i=1}^{N} I\{y_i = r\}}{N},$$

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

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

The conditional information entropy of $\mathbf{x}_j$ given Y is defined as:

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

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

where

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

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

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

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

But when the covariate $\mathbf{x}_j$ is a continuous variable, using standard normal distribution quantiles to cut $\mathbf{x}_j$ into categorical data:

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

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

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

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

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

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

where $q_{(j)}$ be the j/J quantile, and $j = 1, 2, \cdots, J$, $q_{(0)} = -\infty$, $q_{(J)} = +\infty$.

2.3. IG-SIS

Ni and Fang [11] proposed the feature screening method of IG-SIS, which is based on the principle of using the difference between the information entropy of y and the conditional information entropy of Y given $\mathbf{x}_j$ to measure the importance of $\mathbf{x}_j$.

The strength of the correlation between Y and $\mathbf{x}_j$ can be represented by the information gain, and the expression is as follows:

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

and

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

The higher the difference, the more important $\mathbf{x}_j$ is.

2.4. WJS-SIS

Finding an index to measure the relationship between response variables and covariates is the core of ultra-high-dimensional feature screening and the key to big data processing. From the perspective of the distribution of the data, for univariate feature screening, the relationship between variables can be measured by comparing the distribution of the data. Moreover, the current screening methods for categorical variables, in addition to the use of traditional statistical indexes to measure the relationship between the variables, also combine methods from other disciplines. For example, some studies have quantified some measure of the amount of information as an index for feature screening. The Jensen-Shannon divergence mentioned in this paper is based on an information-theoretic concept that is important in calculating similarity and comparing differences in probability distributions and has the properties of non-negativity and symmetry: assuming that there are two distributions A and B, then $JS(A \| B) = JS(B \| A)$. Thus, for ultra-high-dimensional feature screening for multi-classified response variable data, we can utilize the Jensen-Shannon divergence to measure the relationship between response variables and covariates.

When the response variable is a binary categorical variable, there are two conditional probability distributions for $\mathbf{x}_j$ given Y. The degree of difference between these two conditional probability distributions can be measured using JS divergence, and the magnitude of JS divergence represents the degree of strength of the correlation between $\mathbf{x}_j$ and Y. In practice, the response variable is more than just a binary categorization case and more often involves multicategorization.

Therefore, in this paper, a model-free ultra-high-dimensional feature screening method for multicategorical response data with weighted JS divergence is investigated from the perspective of JS divergence for the case where the response variable is multicategorical.

First, separately calculate the JS divergence between conditional probability distributions for $\mathbf{x}_j$ given $Y = r\left(r = 1, 2, \cdots, R\right)$ and the probability distribution of $\mathbf{x}_j$, and then $\Pr\left(Y = r\right)$ is used as the weight to obtain the weighted JS divergence.

Assume that $U = \Pr\left(\mathbf{x}_j = l \mid Y = r\right)$ and $V = \Pr\left(\mathbf{x}_j = l\right)$, and $M = \frac{1}{2}\left(U + V\right)$ is the average probability distribution of U and V. If $\mathbf{x}_j$ is a continuous variable, U and V are defined as follows: $U = \Pr\left(\mathbf{x}_j \in \left(q_{(j-1)}, q_{(j)}\right] \mid Y = r\right)$, $V = \Pr\left(\mathbf{x}_j \in \left(q_{(j-1)}, q_{(j)}\right]\right)$.

Then, the weighted JS divergence of U and V is defined as:

$$
\begin{aligned}
e_j &= \sum_{r=1}^{R} \Pr\left(Y = r\right) JS\left(U \| V\right) \\
&= \sum_{r=1}^{R} \Pr\left(Y = r\right) \left(\frac{1}{2} \sum_{j=1}^{P} U \log\left(\frac{U}{M}\right) + \frac{1}{2} \sum_{j=1}^{P} V \log\left(\frac{V}{M}\right) \right) \\
&= \frac{1}{2} \sum_{r=1}^{R} \Pr\left(Y = r\right) \left(\sum_{j=1}^{P} U \log\left(U\right) - \sum_{j=1}^{P} U \log\left(M\right) \right. \\
&\quad \left. + \sum_{j=1}^{P} V \log\left(V\right) - \sum_{j=1}^{P} V \log\left(M\right) \right) \\
&= \frac{1}{2} \sum_{r=1}^{R} \Pr\left(Y = r\right) \left(\left(H\left(U, M\right) - H\left(U\right) \right) + \left(H\left(V, M\right) - H\left(V\right) \right) \right),
\end{aligned}
$$

and

$$\begin{aligned}\hat{e}_j &= JS\left(\hat{U} \parallel \hat{V}\right) \\ &= \frac{1}{2}\sum_{r=1}^{R}\Pr\left(Y=r\right)\left(\left(H\left(\hat{U},\hat{M}\right)-H\left(\hat{U}\right)\right)+\left(H\left(\hat{V},\hat{M}\right)-H\left(\hat{V}\right)\right)\right).\end{aligned} \tag{2}$$

2.5. AWJS-SIS

When the number of categories of covariates is large, the directly computed weighted JS divergence values may be large, which makes it possible that unimportant variables due to a large number of categories may be incorrectly selected. To address this problem, this paper refers to Ni and Fang [11] using $\left(\log J_k\right)^{-1}$ to construct the adjusted weighted JS divergence for variable selection. Where J_k represents the number of categorical categories L of $\mathbf{x}_j$ or the number of categories in which $\mathbf{x}_j$ is sliced by a standard normally distributed quantile.

The adjusted weighted JS divergence of U and V is defined as:

$$\begin{aligned}w_j &= e_j/\log J_k \\ &= \frac{\frac{1}{2}\sum_{r=1}^{R}\Pr\left(Y=r\right)\left(\left(H\left(U,M\right)-H\left(U\right)\right)+\left(H\left(V,M\right)-H\left(V\right)\right)\right)}{\log J_k},\end{aligned} \tag{3}$$

and

$$\begin{aligned}\hat{w}_j &= \hat{e}_j/\log J_k \\ &= \frac{\frac{1}{2}\sum_{r=1}^{R}\Pr\left(Y=r\right)\left(\left(H\left(\hat{U},\hat{M}\right)-H\left(\hat{U}\right)\right)+\left(H\left(\hat{V},\hat{M}\right)-H\left(\hat{V}\right)\right)\right)}{\log J_k}.\end{aligned} \tag{4}$$

3. Theoretical Properties

In [2], it is shown that a good feature screening method should satisfy the properties of sure screening and ranking consistency. Sure screening is the basis of feature screening, which means being able to screen all important variables with a probability of 1 when the sample size is sufficient, which ensures that the truly important variables will theoretically be screened in their entirety. Ranking consistency means that the indexes of all significant variables are ranked before the indexes of all insignificant variables, which ensures that when selecting the top d_n variables, important variables can be screened out reasonably and robustly. This subsection will illustrate the theoretical properties of the methods proposed in this paper under certain conditions, which are as follows:

Condition 1 (C1). $P = o\left(\exp\left(N^{\delta}\right)\right), \delta \in (0,1)$, this indicates that the variable dimension P is an exponential multiple of the sample capacity *n*.

Condition 2 (C2). There have $c_1 > 0$, $c_2 > 0$, such that $0 < c_1 \le p_{j,lr} \le c_2 < 1$, $\forall l, r \in \{0,1\}$, $\forall k = 1,2,\cdots,P$.

Condition 3 (C3). There has a constant $c > 0$ and $0 \le \tau < 1/2$, such that $\min_{j\in D} e_j \ge 2cN^{-\tau}$.

Condition 4 (C4). There has a constant c_3 for $\forall 1 \le r \le R$ such that

$0 < f_k(x \mid Y = r) < c_3$, and x is in the domain of definition of X_k, where under the condition $Y = r$, $f_k(x \mid Y = r)$ is the Lebesgue density function of X_k.

Condition 5 (C5). There have a constant c_4 and $\forall 1 \le \rho \le 1/2$ such that $c_4 n^{-\rho} \le f_k(x) < c_5$, and x is in the domain of definition of X_k for $\forall 1 \le k \le \rho$, where $f_k(x)$ is continuous in the domain of definition of X_k, and $f_k(x)$ is the Lebesgue density function of X_k.

Condition 6 (C6). $J = \max_{1 \le j \le P} J_k = O(N^{\kappa})$, $\kappa > 0$, $\forall 1 \le \tau \le 1/2$ and $\forall 1 \le \rho \le 1/2$ with $2\tau + 2\rho < 1$.

The literature on ultra-high-dimensional feature screening approaches, such as [2] [13], and [14], typically includes the aforementioned six requirements. Condition (C1) demonstrates that it is a feature screening method applied to ultra-high-dimensional problems. Condition (C2) demonstrates that the marginal probabilities of the response variable and the covariate are bounded by an upper and a lower limit, preventing the worst-case scenario of the screening method failing. This worst-case situation is due to a flaw in the Jensen-Shannon divergence. When the two distributions do not overlap at all, even if the centers of the two distributions are as close as possible, their Jensen-Shannon divergence is constant, and at this point, the Jensen-Shannon divergence fails to measure the extent of the difference between the two distributions and thus the importance of the covariates. And Condition (C3) demonstrates that the values of the indexes corresponding to the really important variables are bounded by a lower value. Condition (C4) eliminates an extreme scenario in which some X_k places a large mass in a small range, ensuring that the sample percentile is close to the genuine percentile. Condition (C5) shows the lower bound of the density must be of order $n^{-\rho}$ in order. The presence of Condition (C6) guarantees a certain rate of divergence in the number of covariate categories. When the response is multi-classified and all covariates are discrete, we provide the theoretical properties of the feature screening technique WJS-SIS under these six conditions.

Because $w_j = e_j / \log J_k$, $\hat{w}_j = \hat{e}_j / \log J_k$, and $\log J_k \ge \log 2 \ge 1/2$, it follows that $\Pr\left(\left|w_j - \hat{w}_j\right| > \varepsilon\right) = \Pr\left(\left|e_j - \hat{e}_j\right| > \varepsilon/2\right)$. So, this study provides a thorough theoretical proof for the index e_j of weighted JS divergence-based sure screening and ranking consistency for feature screening.

The Properties of Sure Screening and Ranking Consistency

Categorical covariates are subscripted with the letter *j*, while continuous covariates are subscripted with the letter *k*. If the covariate is categorical, *L* represents the number of categories, while J_k represents the number of categories if the covariate is continuous.

When the covariates are categorical variables, there are theorems 3.1 and 3.2.

Theorem 3.13 In the (C1) - (C2) conditions, $0 \le \tau < 1/2$, there have $c > 0$ and $C > 0$ with

$$\Pr\left(\max_{1 \le j \le P}\left|e_j - \hat{e}_j\right| > cN^{-\tau}\right) \le 12PRL\exp\left\{-CN^{1-2\tau}\right\},$$

when $0 < a < 1-2\tau$, $\Pr\left(\max_{1\le j\le P}\left|e_j - \hat{e}_j\right| \ge cN^{-\tau}\right) \to 0, N \to \infty$. And under the (C1) - (C3) conditions, when $N \to \infty$, such that

$$\Pr\left(D \subseteq \hat{D}\right) \ge 1 - 12 d_0 RL \exp\left\{-CN^{1-2\tau}\right\} \to 1.$$

Theorem 3.2. In the (C1) - (C3) conditions, assume that $\min_{j\subseteq D} \hat{e}_j - \max_{j\subseteq D^c} \hat{e}_j > 0$, then there have

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

When the covariates are continuous, there are theorems 3.3 and 3.4.

Theorem 3.3. Under the conditions (C1), (C2), (C4), (C5), and (C6), there have constants $c_{11} > 0$, $C_1 > 0$ and there are

$$\Pr\left(\max_{1\le j\le P}\left|e_k - \hat{e}_k\right| > c_{10} N^{-\tau}\right) \le 6c_6 PRJ_k \exp\left\{-C_1 N^{1-2\rho-2\tau}\right\}. \tag{5}$$

When $N \to \infty$, there is

$$\Pr\left(D \subseteq \hat{D}\right) \ge 1 - 6c_6 d_0 RJ_k \exp\left\{-C_1 N^{1-2\rho-2\tau}\right\} \to 1.$$

Theorem 3.4. Assume that $\min_{k\subseteq D} \hat{e}_k - \max_{k\subseteq D^c} \hat{e}_k > 0$, under the conditions (C1), (C3), (C4), (C5), and (C6), there have

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

When the covariates are continuous and categorical covariates coexist, there are theorems 3.5 and 3.6.

Theorem 3.5. Under the conditions (C1), (C2), (C4), (C5), and (C6), there have constants $c_{11} > 0$, $C_2 > 0$ and $C_3 > 0$ and there are

$$\begin{aligned} &\Pr\left(\max_{1\le j\le P}\left(\left|e_j - \hat{e}_j\right| + \left|e_k - \hat{e}_k\right|\right) > c_{11} N^{-\tau}\right) \\ &\le 12 P_1 RL \exp\left\{-C_2 N^{1-2\tau}\right\} + 6c_6 P_2 RJ_k \exp\left\{-C_3 N^{1-2\rho-2\tau}\right\}, \end{aligned} \tag{6}$$

where $P_1 + P_2 = P$. When $N \to \infty$, there is

$$\Pr\left(D \subseteq \hat{D}\right) \ge 1 - 12 d_1 RL \exp\left\{-C_2 N^{1-2\tau}\right\} - 6c_6 d_2 RJ_k \exp\left\{-C_3 N^{1-2\rho-2\tau}\right\} \to 1,$$

where $d_1 + d_2 = d_0$.

Theorem 3.6. Assume that $\min_{j\subseteq D} \hat{e}_j - \max_{j\subseteq D^c} \hat{e}_j > 0$ and $\min_{k\subseteq D} \hat{e}_k - \max_{k\subseteq D^c} \hat{e}_k > 0$, under the conditions (C1), (C2), (C4), (C5), and (C6), there have

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

The appendix contains a thorough proof of the theoretical portion.

4. Numerical Simulation

4.1. Evaluation Indexes

The first evaluation indexes are CP1 and CP2, which represent the proportion of true important covariates that are selected into the set of significant covariates when the top $[N/\log N]$ and top $2[N/\log N]$ variables are selected as the set of significant covariates, respectively. The second evaluation indexes are CPa1 and CPa2, which show whether the selected set of important covariates contains all the true important covariates when the number of the important covariates set is $[N/\log N]$ and $2[N/\log N]$, respectively. The third evaluation index is the MMS, which represents the minimum model size that will be selected for all important variables. Each simulation was conducted 100 times. All calculations in this paper were performed in R software.

4.2. Simulation Experiments and Results

4.2.1. Simulation 1

The response variable and all covariates are four-categorical variables. Where, for the response variable Y, both balanced and unbalanced distributions are considered: balanced, $p_r = \Pr(Y = r) = 1/R$, with $r = 1,\cdots,R$, and $R = 2$; unbalanced, $p_r = 2\left[1 + (R - r)/(R - 1)\right]/3R$ with $\max\limits_{1 \le r \le R} p_r = 2\min\limits_{1 \le r \le R} p_r$. Define $D = \{1,2,\cdots,d_0\}$ as the true imporant variables set, where $d_0 = |D| = 10$. X is generated by the conditional probability of X given by Y. $\Pr\left(x_{ij} = (1,2,3,4) \mid y_i = r\right) = \left(\theta_{rj}/2, \left(1-\theta_{rj}\right)/2, \theta_{rj}/2, \left(1-\theta_{rj}\right)/2\right)$ 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 $1 \le r \le R$, $d_0 < j \le P$. The dimensionality of the covariates was set to $P = 2000$, and the sample sizes were set to $N = 300$, $N = 400$, and $N = 500$.

The simulation results are shown in Table 2, where the performance indexes of all the methods for all conditions are the same, with coverage CP and CPa both being 1 and the values of the MMS values being close to $d_0 = 10$.

4.2.2. Simulation 2

The response variable and all covariates are categorical variables, where the response variable is set up as in Simulation 1, and the categories of the covariates are set up as categories 2, 4, 6, 8, and 10, respectively. Similarly, define $D = \left\{ j = [j'P/10], j' = 1,2,\cdots,10 \right\}$ as the set of important variables. The covariate data were generated through the quantile of the standard normal distribution $f_j(\cdot)$. Define $x_{i,j}$ by $f_j\left(\varepsilon_{i,j} + \mu_{i,j}\right)$, where $\varepsilon_{i,j} \sim N(0,1)$, $1 \le j \le P$. And when $j \in D$, $\mu_{i,j} = 1.5 \times (-0.9)^r$, otherwise $\mu_{i,j} = 0$. The following are the precise steps for creating covariates:

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
$r = 3$	0.1	0.9	0.6	0.1	0.3	0.1	0.4	0.3	0.6	0.4
$r = 4$	0.7	0.2	0.1	0.6	0.7	0.6	0.8	0.9	0.1	0.8

Table 2. Results for simulation 1.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
balanced Y, $P = 2000$, $N = 300$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
balanced Y, $P = 2000$, $N = 400$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
balanced Y, $P = 2000$, $N = 500$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
unbalanced Y, $P = 2000$, $N = 300$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
unbalanced Y, $P = 2000$, $N = 400$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
unbalanced Y, $P = 2000$, $N = 500$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)

The numbers in parentheses are the corresponding standard deviations.

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

The values of L are 2, 4, 6, 8, and 10, which correspond to $1 \le j \le 400$, $400 < j \le 800$, $800 < j \le 1200$, $1200 < j \le 1600$, $1600 < j \le 2000$. We set $P = 2000$ and $N = 160, 240, 320$.

The simulation results are displayed in Table 3, and all techniques' performance indexes for every situation are same, with coverage CP and CPa both being 1, and MMS values that are nearly equal to $d_0 = 10$.

Table 3. Results for simulation 2.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
balanced Y, $P = 2000$, $N = 300$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
balanced Y, $P = 2000$, $N = 400$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
balanced Y, $P = 2000$, $N = 500$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
unbalanced Y, $P = 2000$, $N = 300$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
unbalanced Y, $P = 2000$, $N = 400$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
unbalanced Y, $P = 2000$, $N = 500$									
WJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
AWJS-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)
IG-SIS	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25 (0)	5.5 (0)	7.75 (0)	9.55 (0)

The numbers in parentheses are the corresponding standard deviations.

4.2.3. Simulation 3

The covariates are continuous variables, and the response variables are set up as in Simulation 1. We use the standard normal distribution of quantile function to slice the covariates into categorical data, where $J_K = 4,8,10$, and define the methods as WJS-SIS-4, AWJS-SIS-4, IG-SIS-4; WJS-SIS-8, AWJS-SIS-8, IG-SIS-8; and WJS-SIS-10, AWJS-SIS-10, IG-SIS-10, respectively. The essential variables are set up in the same way as in Simulation 1. Generate X using the standard normal distribution $N\left(\mu_{ij},1\right)$ with $\mu_i = \left\{\mu_{i1},\mu_{i2},\cdots,\mu_{iP}\right\}$ and assume $\mathbf{x}_i = \left\{x_{i1},x_{i2},\cdots,x_{iP}\right\} \in \mathbb{R}^P$, and $x_{ij},\left(j=1,2,\cdots,P\right)$. Where $j \in D$,

$\mu_{ij} = (-1)^r \theta_{rj}$, otherwise, $\mu_{ij} = 0$. We set $P = 5000$ and $N = 400, 600, 800$.

Because this slice is used to divide all covariates after a particular number of slices are chosen each time, the performance indexes of WJS-SIS and AWJS-SIS are the same.

Table 4 displays the simulation results for a balanced distribution of *Y*. Since

Table 4. Results for simulation 3: balanced *Y*.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
				balanced *Y*, *P* = 5000, *N* = 400					
WJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25	5.5	7.75	9.605
AWJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25	5.5	7.75	9.605
IG-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25	5.5	7.75	9.577
WJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25	5.5	7.75	9.759
AWJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25	5.5	7.75	9.759
IG-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25	5.5	7.75	9.764
WJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25	5.5	7.75	9.914
AWJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25	5.5	7.75	9.914
IG-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.45 (0)	3.25	5.5	7.75	9.813
				balanced *Y*, *P* = 5000, *N* = 600					
WJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
AWJS-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
WJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
AWJS-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
WJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
AWJS-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					
WJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
AWJS-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
WJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
AWJS-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
WJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
AWJS-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.

all covariates are divided using the slices each time a certain number of slices is chosen, the performance indexes of WJS-SIS and AWJS-SIS are identical. The coverage index values CP and CPa for the three methods are 1 in all cases. Regarding the MMS values, the 95% quantile values of MMS are slightly different only at $N = 400$, where IG-SIS is smaller than those of WJS-SIS and AWJS-SIS at $J_k = 4$ and 10, and WJS-SIS and AWJS-SIS are smaller than those of IG-SIS at $J_k = 8$; and the 95% quantile values of MMS of all the three methods are increasing with J_k increases. Table 5 shows the simulation results when Y is an unbalanced distribution: with respect to the CP and CPa values, IG-SIS is slightly higher than WJS-SIS and AWJS-SIS at $N = 400$, and 1 for all methods in all other cases. With regard to the MMS values, IG-SIS has a smaller MMS than WJS-SIS and AWJS-SIS at $N = 400, 600$ for the 95% percentile of the MMS values are smaller, and at $N = 800$, the MMS values are the same for all methods. All methods perform better when the number of slices is small.

4.2.4. Simulation 4

The covariates are categorical and continuous, with continuous covariates treated the same as in Simulation 3 regarding handling. The response variables are set up as in Simulation 1. Set the essential variables set is $D = \left\{ j = \left[j'P/20 \right], j' = 1, 2, \cdots, 20 \right\}$. Generating the latent variables $z_i = \left(z_{i,1}, \cdots, z_{i,P} \right)$ in the same way of Simulation 3 generating covariates and then generating categorical and continuous covariates: 1) For $P \le 1250$, then $x_{ij} = j''$, if $z_{ij} \in \left(z_{(j''-1)/4}, z_{j''/4} \right]$, $j'' = 1, 2, 3, 4$; 2) For $1250 < P \le 2500$, then $x_{ij} = j'''$, if $z_{ij} \in \left(z_{(j'''-1)/10}, z_{j'''/10} \right]$, $j''' = 1, \cdots, 10$; 3) For $2500 < P \le 5000$, then $x_{ij} = z_{ij}$. We set $P = 5000$ and $N = 400, 600, 800$.

Table 6 displays the simulation results for a balanced distribution of Y. The CP and CPa values for AWJS-SIS and IG-SIS are the same and larger than those for WJS-SIS. Regarding the MMS values, the 75% and 95% quantile values of AWJS-SIS and IG-SIS are smaller than those of IG-SIS at N = 400 and 600; whereas AWJS-SIS is smaller than IG-SIS at $N = 400$ and $J_k = 8$, and larger than IG-SIS at $N = 400$ and $J_k = 4$ and 10, and at $N = 600$, the MMS values of AWJS-SIS and IG-SIS have the same MMS value; at $N = 800$, all methods have the same MMS value. Table 7 shows the simulation results when Y is an unbalanced distribution: At $N = 800$, all methods have the same performance index value. At $N = 400$, AWJS-SIS and IG-SIS are both larger than WJS-SIS, with AWJS-SIS being somewhat smaller than IG-SIS. At $N = 600$, all methods are equal in terms of the CP and CPa values. All methods perform better when the number of slices is large.

4.3. Computational Time Cost

We obtained the median running time of each algorithm through a simulation experiment, where the covariates X and the set of significant variables were set

Table 5. Results for simulation 3: unbalanced Y.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
unbalanced Y, $P = 5000$, $N = 400$									
WJS-SIS-4	0.999 (0.001)	1 (0)	0.99 (0.01)	1 (0)	1.45	3.25	5.5	7.75	9.988
AWJS-SIS-4	0.999 (0.001)	1 (0)	0.99 (0.01)	1 (0)	1.45	3.25	5.5	7.75	9.988
IG-SIS-4	1 (0)	1 (0)	1 (0.01)	1 (0)	1.45	3.25	5.5	7.75	9.621
WJS-SIS-8	0.998 (0.001)	0.999 (0.001)	0.98 (0.014)	0.99 (0.01)	1.45	3.25	5.5	7.768	12.719
AWJS-SIS-8	0.998 (0.001)	0.999 (0.001)	0.98 (0.014)	0.99 (0.01)	1.45	3.25	5.5	7.768	12.719
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	7.758	10.83
WJS-SIS-10	0.997 (0.002)	0.999 (0.001)	0.97 (0.017)	0.99 (0.01)	1.45	3.25	5.5	7.765	20.692
AWJS-SIS-10	0.997 (0.002)	0.999 (0.001)	0.97 (0.017)	0.99 (0.01)	1.45	3.25	5.5	7.765	20.692
IG-SIS-10	0.999 (0.001)	0.999 (0.001)	0.99 (0.01)	0.99 (0.01)	1.45	3.25	5.5	7.75	14.867
unbalanced Y, $P = 5000$, $N = 600$									
WJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
AWJS-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
WJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.561
AWJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.561
IG-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
WJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.572
AWJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.572
IG-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
unbalanced Y, $P = 5000, N = 800$									
WJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
AWJS-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
WJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
AWJS-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
WJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.45	3.25	5.5	7.75	9.55
AWJS-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 for simulation 4: balanced Y.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
balanced Y, $P = 5000$, $N = 400$									
WJS-SIS-4	0.998 (0.001)	1 (0.001)	0.95 (0.022)	0.99 (0.01)	1.95	5.75	10.5	15.265	20.807
AWJS-SIS-4	1 (0.001)	1 (0)	0.99 (0.01)	1	1.95 (0)	5.75	10.5	15.25	19.134
IG-SIS-4	1 (0.001)	1 (0)	0.99 (0.01)	1	1.95 (0)	5.75	10.5	15.25	19.141
WJS-SIS-8	0.998 (0.001)	0.999 (0.001)	0.95 (0.002)	0.98 (0.014)	1.95	5.75	10.5	15.252	19.952
AWJS-SIS-8	1 (0.001)	1 (0)	0.99 (0.01)	1 (0)	1.95	5.75	10.5	15.25	19.214
IG-SIS-8	1 (0.001)	1 (0)	0.99 (0.01)	1 (0)	1.95	5.75	10.5	15.25	19.218
WJS-SIS-10	0.994 (0.002)	0.998 (0.001)	0.88 (0.033)	0.95 (0.022)	1.95	5.75	10.5	15.252	21.329
AWJS-SIS-10	0.999 (0.001)	1 (0.001)	0.98 (0.014)	0.99 (0.01)	1.95	5.75	10.5	15.25	19.299
IG-SIS-10	0.999 (0.001)	1 (0.001)	0.98 (0.014)	0.99 (0.01)	1.95	5.75	10.5	15.25	19.298
balanced Y, $P = 5000$, $N = 600$									
WJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.087
AWJS-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
WJS-SIS-8	1 (0.001)	1 (0)	0.99 (0.01)	1 (0)	1.95	5.75	10.5	15.25	19.097
AWJS-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
WJS-SIS-10	1 (0.001)	1 (0.001)	0.99 (0.01)	0.99 (0.01)	1.95	5.75	10.5	15.25	19.146
AWJS-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, $P = 5000$, $N = 800$									
WJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.05
AWJS-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
WJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.05
AWJS-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
WJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.05
AWJS-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 7. Results for simulation 4: unbalanced Y.

Method	CP		CPa		MMS				
	CP1	CP2	CPa1	CPa2	5%	25%	50%	75%	95%
				unbalanced Y, $P = 5000$, $N = 400$					
WJS-SIS-4	0.991 (0.002)	0.997 (0.001)	0.83 (0.038)	0.94 (0.024)	1.95	5.75	10.5	15.345	25.135
AWJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.302
IG-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.096
WJS-SIS-8	0.996 (0.001)	0.998 (0.001)	0.91 (0.029)	0.97 (0.017)	1.95	5.75	10.5	15.255	21.4
AWJS-SIS-8	1 (0.001)	1 (0)	0.99 (0.01)	1 (0)	1.95	5.75	10.5	15.25	19.529
IG-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.129
WJS-SIS-10	0.989 (0.002)	0.995 (0.002)	0.79 (0.041)	0.9 (0.03)	1.95	5.75	10.5	15.275	24.033
AWJS-SIS-10	0.998 (0.001)	1 (0)	0.97 (0.017)	1 (0)	1.95	5.75	10.5	15.25	19.767
IG-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.15
				unbalanced Y, $P = 5000$, $N = 600$					
WJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.084
AWJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.055
IG-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.05
WJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.074
AWJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.056
IG-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.051
WJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.121
AWJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.056
IG-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.051
				unbalanced Y, $P = 5000$, $N = 800$					
WJS-SIS-4	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.05
AWJS-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
WJS-SIS-8	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.05
AWJS-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
WJS-SIS-10	1 (0)	1 (0)	1 (0)	1 (0)	1.95	5.75	10.5	15.25	19.05
AWJS-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. The results of computational time cost.

P	1000	2000	3000	4000	5000	6000	7000	8000	9000	10,000
WJS-SIS	1.408	2.809	4.212	5.632	6.915	8.287	9.686	11.095	12.544	13.929
	(0.003)	(0.007)	(0.005)	(0.015)	(0.010)	(0.006)	(0.007)	(0.008)	(0.009)	(0.008)
AWJS-SIS	1.464	2.925	4.380	5.861	7.196	8.618	10.083	11.559	13.021	14.490
	(0.004)	(0.004)	(0.005)	(0.014)	(0.010)	(0.006)	(0.006)	(0.008)	(0.006)	(0.007)
IG-SIS	1.692	3.382	5.069	6.774	8.298	9.945	11.633	13.323	15.047	16.705
	(0.010)	(0.005	(0.005)	(0.014)	(0.011)	(0.006)	(0.007)	(0.009)	(0.007)	(0.007)

The numbers in parentheses are the corresponding standard deviations.

up as in simulation experiment 2, and Y was set up as a balanced distribution. The experiment was set up with $N = 400$, $P = 1000, 2000, \cdots, 10000$, and repeat the experiment 100 times. An Intel Core i7-8700 machine running Windows 10 at 3.20 GHz was used for all calculations. Table 8 shows the median runtime for the three methods, which increases as P increases and is consistently faster to compute for WJS-SIS and AWJS-SIS than for IG-SIS.

4.4. Comprehensive Analysis of Simulation Results

The main argument is that, in terms of performance, the approaches of WJS-SIS and AWJS-SIS in this study are extremely comparable to IG-SIS. There is a difference in performance between the approaches when the sample size is small, and the performance of WJS-SIS is more affected by the slices than that of AWJ-SIS and IG-SIS, which are more robust and whose performance is more adaptive to the number of slices. But all methods perform as well as they do as the number of variables screened increases or as the sample size increases, and all are able to screen out all the important variables, and the true model size is close to the number of important variables. In terms of computing time, IG-SIS is longer than WJS-SIS and AWJS-SIS.

5. Experimental Study with Real Data

In real life, ultra-high-dimensional data with multi-class response variables is common, and feature screening of such data can achieve the effects of data dimensionality reduction, feature mining, and variable selection. The methods proposed in this paper can be applied in different fields and can improve the efficiency and accuracy of data analysis, reveal the information behind the data, and help in the construction of decision-making and prediction models. For example, in the medical field, it can be applied to analyze gene expression data and help identify genes associated with diseases. In the financial field, it can help identify key factors that affect stock or commodity prices. In image processing, it can be used for tasks such as feature extraction and target recognition. In practical implementation, based on Fan and Lv [2], the number of important variables is generally selected as the first $d_n = \gamma\left[N/\log N\right]$. In this paper, we analyze the case when $\gamma = 1, 2$.

We analyzed the TOX-171 micro-integrated columns biological dataset from the Arizona State University feature selection database (http://featureselection.asu.edu/) with 171 samples and 5748 features, with four classes of response variables and roughly unbalanced distributions, with covariates of continuous type. We randomly divide the dataset in a 7:3 ratio, where 70% of the data is used as the training dataset and the remaining 30% as the test dataset. As randomly dividing datasets may bring the potential problem of model prediction accuracy degradation, to address this problem, we used ten-fold cross-validation to train the model and repeated the experiment 100 times to take the average of the evaluation indexes and calculate the standard deviation of the evaluation indexes. The smaller the standard deviation, the more stable it is, which means that the average of the evaluation indexes is desirable. On the training and test sets, respectively, variables screened using the three methods were tested for categorization using a support vector machine, and the values of the geometric mean (G-mean) for categorization accuracy (CA), specificity (SPE), and sensitivity (SEN) were calculated.

Table 9 and Table 10 show the corresponding index values when the number of selected variables is $\left[N/\log N\right]$ and $2\left[N/\log N\right]$, respectively. Combining Table 9 and Table 10, it can be seen that in both the training and test sets, all methods perform better when J_k is relatively large, and the CA and G-mean values of AWJS-SIS are always higher than those of WJS-SIS, and the CA and G-mean values of AWJS-SIS are higher than those of IG-SIS at $J_k = 8$. As well, the classification of the method is better as the screening variables increase.

6. Conclusion

In this paper, from the perspective of introducing JS (Jensen-Shannon) divergence to measure the importance of covariates, for the case where Y is multi-classified, this paper constructs model-free ultra-high-dimensional feature

Table 9. The result when screening the first $\left[N/\log N\right]$ variables.

Method	test data_CA	test data_Gmean	train data_CA	train data_Gmean
WJS-SIS-4	0.761 (0.009)	0.682 (0.014)	0.994 (0.001)	0.995 (0.001)
AWJS-SIS-4	0.769 (0.008)	0.72 (0.011)	0.994 (0.001)	0.995 (0.001)
IG-SIS-4	0.772 (0.008)	0.716 (0.014)	0.999 (0.001)	0.999 (0)
WJS-SIS-8	0.809 (0.008)	0.721 (0.016)	1 (0)	1 (0)
AWJS-SIS-8	0.821 (0.008)	0.738 (0.013)	1 (0)	1 (0)
IG-SIS-8	0.805 (0.007)	0.684 (0.015)	1 (0)	1 (0)
WJS-SIS-10	0.854 (0.007)	0.726 (0.019)	1 (0)	1 (0)
AWJS-SIS-10	0.872 (0.006)	0.782 (0.015)	1 (0)	1 (0)
IG-SIS-10	0.877 (0.006)	0.809 (0.015)	1 (0)	1 (0)

The numbers in parentheses are the corresponding standard deviations.

Table 10. The result when screening the first $2[N/\log N]$ variables.

Method	test data_CA	test data_Gmean	train data_CA	train data_Gmean
WJS-SIS-4	0.857 (0.008)	0.832 (0.009)	1 (0)	1 (0)
AWJS-SIS-4	0.857 (0.008)	0.832 (0.009)	1 (0)	1 (0)
IG-SIS-4	0.849 (0.007)	0.823 (0.009)	1 (0)	1 (0)
WJS-SIS-8	0.84 (0.007)	0.782 (0.011)	1 (0)	1 (0)
AWJS-SIS-8	0.84 (0.007)	0.782 (0.011)	1 (0)	1 (0)
IG-SIS-8	0.823 (0.006)	0.739 (0.012)	1 (0)	1 (0)
WJS-SIS-10	0.909 (0.006)	0.839 (0.012)	1 (0)	1 (0)
AWJS-SIS-10	0.914 (0.006)	0.844 (0.012)	1 (0)	1 (0)
IG-SIS-10	0.929 (0.005)	0.872 (0.01)	1 (0)	1 (0)

The numbers in parentheses are the corresponding standard deviations.

screening methods for multi-classified response data based on weighted JS divergence under different scenarios, using the WJS-SIS method when the number of categories in each covariate is the same and the AWJS-SIS method with adjusted weighted JS divergence when the number of categories in each covariate is different. Theoretically, both WJS-SIS and AWJS-SIS have sure screening properties and ranking consistency. Then, from the Monte Carlo simulation results and experiments with real data, WJS-SIS and AWJS-SIS have a significant effect on feature screening, and the performance is very similar to that of IG-SIS, but WJS-SIS is a little bit weaker in terms of robustness, whereas AWJ-SIS and IG-SIS are robust a little stronger, and both WJS-SIS and AWJS-SIS are faster than IG-SIS in terms of computation time. Finally, the approaches proposed in this paper utilize the Jensen-Shannon divergence to measure the importance of covariates from the perspective of information quantity, which is different from the traditional statistical indicators, which may provide a reference for methodological research in the field of multi-class variable screening for ultra-high-dimensional data. And the methods proposed in this work only take into consideration the correlation between the response variable and the covariates; they do not account for the presence of a high covariate correlation. Therefore, in future studies for ultra-high-dimensional variable screening, the element of covariate correlation will be incorporated.

Acknowledgements

The study was supported by the National Natural Science Foundation of China [grant number 71963008].

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Fan, J.Q., Samworth, R. and Wu, Y.C. (2009) Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. *Journal of Machine Learning Research*, **10**, 2013-2038. http://arxiv.org/abs/0812.3201

[2] Fan, J.Q. and Lv, J.C. (2008) Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **70**, 849-911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

[3] Fan, J.Q. and Song, R. (2010) Sure Independence Screening in Generalized Linear Models with NP-Dimensionality. *The Annals of Statistics*, **38**, 3567-3604. https://doi.org/10.1214/10-AOS798

[4] Fan, J.Q., Feng, Y. and Song, R. (2011) Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models. *Journal of the American Statistical Association*, **106**, 544-557. https://doi.org/10.1198/jasa.2011.tm09779

[5] Liu, J.Y., Li, R.Z. and Wu, R.L. (2014) Feature Selection for Varying Coefficient Models with Ultrahigh-Dimensional Covariates. *Journal of the American Statistical Association*, **109**, 266-274. https://doi.org/10.1080/01621459.2013.850086

[6] Li, G.R., Peng, H., Zhang, J. and Zhu, L.X. (2012) Robust Rank Correlation Based Screening. *The Annals of Statistics*, **40**, 1846-1877. https://doi.org/10.1214/12-AOS1024

[7] Fan, J.Q. and Fan, Y.Y. (2008) High Dimensional Classification Using Features Annealed Independence Rules. *Annals of Statistics*, **36**, 2605-2637. https://doi.org/10.1214/07-AOS504

[8] Mai, Q. and Zou, H. (2013) The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification. *Biometrika*, **100**, 229-234. https://doi.org/10.1093/biomet/ass062

[9] Cui, H.J., Li, R.Z. and Zhong, W. (2015) Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis. *Journal of the American Statistical Association*, **110**, 630-641. https://doi.org/10.1080/01621459.2014.920256

[10] Huang, D.Y., Li, R.Z. and Wang, H.S. (2014) Feature Screening for Ultrahigh Dimensional Categorical Data with Applications. *Journal of Business & Economic Statistics*, **32**, 237-244. https://doi.org/10.1080/07350015.2013.863158

[11] Ni, L. and Fang, F. (2016) Entropy-Based Model-Free Feature Screening for Ultrahigh-Dimensional Multiclass Classification. *Journal of Nonparametric Statistics*, **28**, 515-530. https://doi.org/10.1080/10485252.2016.1167206

[12] Shannon, C.E. (1948) A Mathematical Theory of Communication. *The Bell System Technical Journal*, **27**, 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[13] Li, R.Z., Zhong, W. and Zhu, L.P. (2012) Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association*, **107**, 1129-1139. https://doi.org/10.1080/01621459.2012.695654

[14] Cui, H.J. and Zhong, W. (2019) A Distribution-Free Test of Independence Based on Mean Variance Index. *Computational Statistics & Data Analysis*, **139**, 117-133. https://doi.org/10.1016/j.csda.2019.05.004

[15] Hoeffding, W. (1963) Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, **58**, 13-30. https://doi.org/10.1080/01621459.1963.10500830

[16] Lin, J.H. (1991) Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, **37**, 145-151. https://doi.org/10.1109/18.61115

Appendix

The following four lemmas are initially introduced in order to demonstrate Theorem 3.1.

Lemma 1. Suppose there are mutually independent random variables $x_1, x_2, \cdots, x_N$ with sample size N and $\Pr\left(x_i \in [a_i, b_i]\right) = 1, 1 \le i \le N$, where a_i, b_i are constants. If we assume that $\bar{x} = 1/N \sum_{i=1}^{N} x_i$, then there has a constant t for which the inequality holds:

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

In [15], Lemma 1's proof is presented.

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

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

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

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

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

Besides, assuming that b is bounded and non-zero, and that there has $M_3 > 0$ such that $|b| \ge M_3$, then there exist

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

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

In [10], Lemma 2's proof is presented.

Lemma 3. If the covariates are categorical, we can get that $e_j \ge 0$. And $e_j = 0$ only when $\Pr\left(\mathbf{x}_j = l \mid Y = r\right) = \Pr\left(\mathbf{x}_j = l\right)$, Y and $\mathbf{x}_j$ are independent.

In [16], Lemma 3's proof is presented.

Lemma 4. If the covariates are continuous, there is $e_j \ge 0$, when Y and $\mathbf{x}_j$

are independent, $e_j = 0$.

Lemma 4's proof is omitted here because it is similar to Proposition 2.2's proof in [11].

Theorem 3.1 proof:

Let $U = \Pr\left(\mathbf{x}_j = l \mid Y = r\right)$, $V = \Pr\left(\mathbf{x}_j = l\right)$, $M = \frac{1}{2}(U+V)$ then

$$\begin{aligned}
e_j &= \sum_{r=1}^{R} \Pr(Y=r) JS(U \parallel V) \\
&= \sum_{r=1}^{R} \Pr(Y=r) \left(\frac{1}{2} \sum_{j=1}^{P} U \log\left(\frac{U}{M}\right) + \frac{1}{2} \sum_{j=1}^{P} V \log\left(\frac{V}{M}\right) \right) \\
&= \frac{1}{2} \sum_{r=1}^{R} \Pr(Y=r) \left(\sum_{j=1}^{P} U \log(U) - \sum_{j=1}^{P} U \log(M) \right. \\
&\quad \left. + \sum_{j=1}^{P} V \log(V) - \sum_{j=1}^{P} V \log(M) \right) \\
&= \frac{1}{2} \sum_{r=1}^{R} \Pr(Y=r) \Big(\big(H(U,M) - H(U)\big) + \big(H(V,M) - H(V)\big) \Big)
\end{aligned}$$

The definitions of e_j and $\hat{e}_j$ state that there are

$$\begin{aligned}
\left|e_j - \hat{e}_j\right| &= \frac{1}{2} \left| \sum_{r=1}^{R} \Pr(Y=r) \Big[\big(H(U,M) - H(U)\big) + \big(H(V,M) - H(V)\big) \Big] \right. \\
&\quad \left. - \sum_{r=1}^{R} \hat{\Pr}(Y=r) \Big[\big(\hat{H}(U,M) - \hat{H}(U)\big) + \big(\hat{H}(V,M) - \hat{H}(V)\big) \Big] \right| \\
&= \frac{1}{2} \left| \sum_{r=1}^{R} \Big[\hat{\Pr}(Y=r) \left|\hat{H}(U,M) - \hat{H}(U)\right| - \hat{\Pr}(Y=r) \left|H(U,M) - H(U)\right| \right. \\
&\quad + \hat{\Pr}(Y=r) \left|H(U,M) - H(U)\right| - \Pr(Y=r) \left|H(U,M) - H(U)\right| \\
&\quad + \hat{\Pr}(Y=r) \left|\hat{H}(V,M) - \hat{H}(V)\right| - \hat{\Pr}(Y=r) \left|H(V,M) - H(V)\right| \\
&\quad \left. + \hat{\Pr}(Y=r) \left|H(V,M) - H(V)\right| - \Pr(Y=r) \left|H(V,M) - H(V)\right| \Big] \right| \\
&= \frac{1}{2} \left| \sum_{r=1}^{R} \Pr(Y=r) \Big[\big(H(U,M) - H(U)\big) + \big(H(V,M) - H(V)\big) \Big] \right. \\
&\quad \left. - \sum_{r=1}^{R} \hat{\Pr}(Y=r) \Big[\big(\hat{H}(U,M) - \hat{H}(U)\big) + \big(\hat{H}(V,M) - \hat{H}(V)\big) \Big] \right| \\
&= \frac{1}{2} \left| \sum_{r=1}^{R} \Big[\Big(\hat{\Pr}(Y=r) \left|\hat{H}(U,M) - \hat{H}(U)\right| - \left|H(U,M) - H(U)\right| \Big) \right. \\
&\quad + \Big(\hat{\Pr}(Y=r) - \Pr(Y=r) \Big) \left|H(U,M) - H(U)\right| \\
&\quad + \hat{\Pr}(Y=r) \Big(\left|\hat{H}(V,M) - \hat{H}(V)\right| - \left|H(V,M) - H(V)\right| \Big) \\
&\quad \left. + \Big(\hat{\Pr}(Y=r) - \Pr(Y=r) \Big) \left|H(V,M) - H(V)\right| \Big] \right| \\
&\le \frac{1}{2} \left| \sum_{r=1}^{R} \hat{\Pr}(Y=r) \Big(\left|\hat{H}(U,M) - H(U,M)\right| + \left|\hat{H}(U) - H(U)\right| \Big) \right. \\
&\quad + \sum_{r=1}^{R} \Big(\hat{\Pr}(Y=r) - \Pr(Y=r) \Big) \left|H(U,M) - H(U)\right| \\
&\quad + \sum_{r=1}^{R} \hat{\Pr}(Y=r) \Big(\left|\hat{H}(V,M) - H(V,M)\right| + \left|\hat{H}(V) - H(V)\right| \Big)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{r=1}^{R}\left(\hat{\Pr}\left(Y=r\right)-\Pr\left(Y=r\right)\right)\left|H\left(V,M\right)-H\left(V\right)\right|\Bigg| \\
&\le \frac{1}{2}\Bigg|\sum_{r=1}^{R}\left|\hat{H}\left(U,M\right)-H\left(U,M\right)\right|+\sum_{r=1}^{R}\left|\hat{H}\left(U\right)-H\left(U\right)\right| \\
&+\sum_{r=1}^{R}\left|\hat{H}\left(V,M\right)-H\left(V,M\right)\right|+\sum_{r=1}^{R}\left|\hat{H}\left(V\right)-H\left(V\right)\right| \qquad (7)\\
&+2\sum_{r=1}^{R}\left(\hat{\Pr}\left(Y=r\right)-\Pr\left(Y=r\right)\right)\Bigg|,
\end{aligned}
$$

and

$$
\begin{aligned}
&\Pr\left(\left|e_j-\hat{e}_j\right|>\varepsilon\right) \\
&\le \Pr\Bigg(\frac{1}{2}\Bigg|\sum_{r=1}^{R}\left|\hat{H}\left(U,M\right)-H\left(U,M\right)\right|+\sum_{r=1}^{R}\left|\hat{H}\left(U\right)-H\left(U\right)\right| \\
&\quad+\sum_{r=1}^{R}\left|\hat{H}\left(V,M\right)-H\left(V,M\right)\right|+\sum_{r=1}^{R}\left|\hat{H}\left(V\right)-H\left(V\right)\right| \\
&\quad+2\sum_{r=1}^{R}\left(\Pr\left(Y=r\right)-\Pr\left(Y=r\right)\right)\Bigg|>\varepsilon\Bigg) \\
&\le \Pr\left(\sum_{r=1}^{R}\left|\hat{H}\left(U,M\right)-H\left(U,M\right)\right|>\frac{\varepsilon}{3}\right)+\Pr\left(\sum_{r=1}^{R}\left|\hat{H}\left(U\right)-H\left(U\right)\right|>\frac{\varepsilon}{3}\right) \\
&\quad+\Pr\left(\sum_{r=1}^{R}\left|\hat{H}\left(V,M\right)-H\left(V,M\right)\right|>\frac{\varepsilon}{3}\right)+\Pr\left(\sum_{r=1}^{R}\left|\hat{H}\left(V\right)-H\left(V\right)\right|>\frac{\varepsilon}{3}\right) \\
&\quad+2\Pr\left(\sum_{r=1}^{R}\left(\Pr\left(Y=r\right)-\Pr\left(Y=r\right)\right)>\frac{\varepsilon}{3}\right) \\
&=: E_{j1}+E_{j2}+E_{j3}+E_{j4}+E_{j5}.
\end{aligned}
$$

To prove that E_{j1} Part at first:

$$
\begin{aligned}
&\Pr\left(\left|H\left(U,M\right)-\hat{H}\left(U,M\right)\right|>\frac{\varepsilon}{3}\right) \\
&=\Pr\Bigg(\Bigg|\sum_{l=1}^{L}\hat{p}\left(\mathbf{x}_j=l\,|\,Y=r\right)\log\left(\frac{\hat{p}\left(\mathbf{x}_j=l\,|\,Y=r\right)+\hat{p}\left(\mathbf{x}_j=l\right)}{2}\right) \\
&\quad-\sum_{l=1}^{L}p\left(\mathbf{x}_j=l\,|\,Y=r\right)\log\left(\frac{p\left(\mathbf{x}_j=l\,|\,Y=1\right)+p\left(\mathbf{x}_j=l\right)}{2}\right)\Bigg|>\frac{\varepsilon}{3}\Bigg) \\
&\le R\max_{r}\Pr\Bigg(\Bigg|\sum_{l=1}^{L}\hat{p}\left(\mathbf{x}_j=l\,|\,Y=r\right)\log\left(\frac{\hat{p}\left(\mathbf{x}_j=l\,|\,Y=r\right)+\hat{p}\left(\mathbf{x}_j=l\right)}{2}\right) \\
&\quad-\sum_{l=1}^{L}p\left(\mathbf{x}_j=l\,|\,Y=r\right)\log\left(\frac{p\left(\mathbf{x}_j=l\,|\,Y=r\right)+p\left(\mathbf{x}_j=l\right)}{2}\right)\Bigg|>\frac{\varepsilon}{3R}\Bigg) \\
&\le RL\max_{r,l}\Pr\Bigg(\Bigg|\hat{p}\left(\mathbf{x}_j=l\,|\,Y=r\right)\log\left(\frac{\hat{p}\left(\mathbf{x}_j=l\,|\,Y=r\right)+\hat{p}\left(\mathbf{x}_j=l\right)}{2}\right) \\
&\quad-p\left(\mathbf{x}_j=l\,|\,Y=r\right)\log\left(\frac{p\left(\mathbf{x}_j=l\,|\,Y=r\right)+p\left(\mathbf{x}_j=l\right)}{2}\right)\Bigg|>\frac{\varepsilon}{3RL}\Bigg).
\end{aligned}
$$

By estimating the probability with the sample frequency, we have

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

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

$$\hat{p}\left(\mathbf{x}_j = l\right) = \frac{\sum_{i=1}^{N} I\left\{x_{ij} = l\right\}}{N}$$

$$p\left(\mathbf{x}_j = l\right) = p\left(I\left\{x_{ij} = l\right\}\right)$$

$$\hat{p}\left(Y = r\right) = \frac{\sum_{i=1}^{N} I\left\{y_i = r\right\}}{N}$$

$$p\left(Y = r\right) = p\left(I\left\{Y = r\right\}\right)$$

Thus, there is

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

furthermore, it follows from Lemma 1 and Lemma 2 that since S_n, T_n are estimates of s_n, t_n:

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

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

So, $\hat{p}\left(\mathbf{x}_j = l \mid Y = r\right) \xrightarrow{P} p\left(\mathbf{x}_j = l \mid Y = r\right)$:

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

Additionally, may be proved that $\log\left(\hat{p}\left(\mathbf{x}_j = l \mid Y = 1\right)\right) \xrightarrow{P} \log\left(p\left(\mathbf{x}_j = l \mid Y = 1\right)\right)$. Assume $\hat{p}^* = \hat{p}\left(\mathbf{x}_j = l \mid Y = 1\right)$, $p^* = p\left(\mathbf{x}_j = l \mid Y = 1\right)$:

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

Then, $\log\left(\hat{p}\left(\mathbf{x}_j = l \mid Y = 1\right)\right) \xrightarrow{P} \log\left(p\left(\mathbf{x}_j = l \mid Y = 1\right)\right)$.

We can obtain that

$\log\left(\frac{\hat{p}\left(\mathbf{x}_j=l\mid Y=r\right)+\hat{p}\left(\mathbf{x}_j=l\right)}{2}\right)\xrightarrow{P}\log\left(\frac{p\left(\mathbf{x}_j=l\mid Y=r\right)+p\left(\mathbf{x}_j=l\right)}{2}\right)$ in a proof similar to this one.

So, we can get $E_{j1}\le 2RL\exp\left\{-2N\varepsilon^2/9R^2L^2\right\}$. Similarly, it can be shown that E_{j2}, E_{j3}, and E_{j4} are all $\le 2RL\exp\left\{-2N\varepsilon^2/9R^2L^2\right\}$.

For the E_{j5} part:

$$\Pr\left(\sum_{r=1}^{R}\left(\hat{\Pr}\left(Y=r\right)-\Pr\left(Y=r\right)\right)>\frac{\varepsilon}{3}\right)\le 2R\exp\left\{-2N\varepsilon^2/9R^2\right\}$$

Prove the E_{j5} Part:

According to Lemma 1 and Lemma 2, it can also be shown that $\hat{p}\left(Y=r\right)$ converges to $p\left(Y=r\right)$ with probability, then

$$\begin{aligned}&\Pr\left(\sum_{r=1}^{R}\left(\hat{\Pr}\left(Y=r\right)-\Pr\left(Y=r\right)\right)>\frac{\varepsilon}{3}\right)\\&\le R\max_{r}\Pr\left(\left|\frac{1}{N}\sum_{i=1}^{N}f\left(y_i=r\right)-E\left(y_i=r\right)\right|>\frac{\varepsilon}{3}\right)\\&\le 2R\exp\left\{-2N\varepsilon^2/9R^2\right\}.\end{aligned}$$

For $0<\varepsilon_4<1$, thus, there is

$$\Pr\left(\left|e_j-\hat{e}_j\right|>\varepsilon_4\right)\le 8RL\exp\left\{-2N\varepsilon_4^2/9R^2L^2\right\}+4R\exp\left\{-2N\varepsilon_4^2/9R^2\right\}. \tag{8}$$

In the (C1) - (C3) condition, there exists $c>0$ and $C>0$ with

$$\begin{aligned}\Pr\left(\left|e_j-\hat{e}_j\right|>cN^{-\tau}\right)&\le 8RL\exp\left\{-2c^2N^{1-2\tau}/9R^2L^2\right\}\\&\quad+4R\exp\left\{-2c^2N^{1-2\tau}/9R^2\right\}\\&\le 12RL\exp\left\{-CN^{1-2\tau}\right\},\end{aligned} \tag{9}$$

then,

$$\begin{aligned}\Pr\left(\max_{1\le j\le P}\left|e_j-\hat{e}_j\right|>cN^{-\tau}\right)&\le\Pr\left(\bigcup_{j=1}^{P}\left|e_j-\hat{e}_j\right|>cN^{-\tau}\right)\\&\le P\Pr\left(\left|e_j-\hat{e}_j\right|>cN^{-\tau}\right)\\&\le 12PRL\exp\left\{-CN^{1-2\tau}\right\}.\end{aligned} \tag{10}$$

when $N\to\infty$ and $0<a<1-2\tau$, there has

$$\Pr\left(\max_{1\le j\le P}\left|e_j-\hat{e}_j\right|>cN^{-\tau}\right)\to 0.$$

Then,

$$\begin{aligned}\Pr\left(D\subseteq\hat{D}\right)&\ge\Pr(\left|e_j-\hat{e}_j\right|>cN^{-\tau},\forall j\in D)\\&\ge\Pr\left(\max_{j\in D}\left|e_j-\hat{e}_j\right|>cN^{-\tau}\right)\\&\ge 1-d_0\Pr\left(\left|e_j-\hat{e}_j\right|>cN^{-\tau}\right)\\&\ge 1-12d_0RL\exp\left\{-CN^{1-2\tau}\right\},\end{aligned} \tag{11}$$

so, $\Pr\left(D \subseteq \hat{D}\right) \to 1$, with $N \to \infty$.

Therefore, in the conditions (C1) - (C3), Theorem 0.1 sure screening property holds.

Theorem 0.2 proof:

Because of $\min_{j\in D} e_j - \max_{j\in D^c} e_j > 0$, there has $\delta > 0$ such that $\min_{j\in D} e_j - \max_{j\in D^c} e_j = \delta$, and after that, there have

$$\begin{aligned}
\Pr\left(\min_{j\in D}\hat{e}_j \le \max_{j\in D^c}\hat{e}_j\right) &= \Pr\left(\min_{j\in D}\hat{e}_j - \max_{j\in D^c} e_j \le \max_{j\in D^c}\hat{e}_j - \max_{j\in D^c} e_j\right) \\
&= \Pr\left(\min_{j\in D}\hat{e}_j - \min_{j\in D} e_j + \delta \le \max_{j\in D^c}\hat{e}_j - \max_{j\in D^c} e_j\right) \\
&= \Pr\left(\min_{j\in D}\hat{e}_j - \min_{j\in D} e_j - \max_{j\in D^c}\hat{e}_j + \max_{j\in D^c} e_j \le -\delta\right) \\
&= \Pr\left(\left|\left(\min_{j\in D}\hat{e}_j - \max_{j\in D^c}\hat{e}_j\right) - \left(\min_{j\in D} e_j - \max_{j\in D^c} e_j\right)\right| \ge \delta\right) \\
&\le \Pr\left(\max_{1\le j\le P}\left|e_j - \hat{e}_j\right| \ge \delta/2\right) \\
&\le 12PRL\exp\left\{-CN^{1-2\tau}\right\}.
\end{aligned}$$

From Fatou's Lemma we can get

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

Thus,

$$\Pr\left\{\liminf_{N\to\infty}\left(\min_{j\in D}\hat{e}_j - \max_{j\in D^c}\hat{e}_j\right) \le 0\right\} = 1. \tag{12}$$

Therefore, Theorem 3.2 holds.

Theorem 3.3 proof:

Assume that $\hat{F}_k(x \mid y)$ is (X_k, Y)'s empirical cumulative distribution function and that $F_k(x \mid y)$ is the cumulative distribution function of (X_k, Y). And let $F_k(x)$ be the cumulative distribution function of x_k, and $\partial F_k(x)/\partial x = f_k(x)$. Then, using LEMMA.A.2 in [11] as evidence, we can similarly demonstrate that, for $\forall \varepsilon_5, \varepsilon_6 > 0, 1 \le r \le R, 1 \le j \le J_k$, given the conditions (C4) and (C5), there are

$$\Pr\left(\left|\hat{F}_k\left(\hat{q}_{k,(j)} \mid r\right) - F_k\left(q_{k,(j)} \mid r\right)\right| > \varepsilon_5\right) \le c_6\exp\left\{-c_7 N^{1-2\rho}\varepsilon_5^2\right\},$$

$$\Pr\left(\left|\hat{F}_k\left(\hat{q}_{k,(j)}\right) - F_k\left(q_{k,(j)}\right)\right| > \varepsilon_6\right) \le c_6\exp\left\{-c_9 N^{1-2\rho}\varepsilon_6^2\right\}$$

where $c_6 = 3c_8$ and $c_7 = \min\left\{1/2, c_4^2/2c_3^2\right\}$, $c_9 = \min\left\{1/2, c_4^2/2c_5^2\right\}$ are positive constants.

So, $\hat{F}_k\left(\hat{q}_{k,(j)} \mid r\right) \xrightarrow{P} F_k\left(q_{k,(j)} \mid r\right)$ and $\hat{F}_k\left(\hat{q}_{k,(j)}\right) \xrightarrow{P} F_k\left(q_{k,(j)}\right)$, respectively.

Then, for $0 < \varepsilon_7 < 1$, there has

$$\Pr\left(\left|e_k - \hat{e}_k\right| > \varepsilon_7\right) \le 4c_6 RJ_k \exp\left\{\frac{-c_7 N^{1-2\rho}\varepsilon_7^2}{9R^2 J_k^2}\right\} + 2c_6 RJ_k \exp\left\{\frac{-c_9 N^{1-2\rho}\varepsilon_7^2}{9R^2 J_k^2}\right\}. \quad (13)$$

Equation (13) will not be proven here because it is similar to the proof of Equation (8).

There are constants c_{10} and C_1 under condition (C6) such that

$$\begin{aligned}\Pr\left(\left|e_j - \hat{e}_j\right| > c_{10}N^{-\tau}\right) &\le 4c_6 PRJ_k \exp\left\{\frac{-c_7 c_{10}^2 N^{1-2\rho-2\tau}}{9R^2 J_k^2}\right\} \\ &\quad + 2c_6 PRJ_k \exp\left\{\frac{-c_9 c_{10}^2 N^{1-2\rho-2\tau}}{9R^2 J_k^2}\right\} \\ &\le 6c_6 PRJ_k \exp\left\{-C_1 N^{1-2\rho-2\tau}\right\},\end{aligned} \quad (14)$$

then,

$$\begin{aligned}\Pr\left(\max_{1\le j\le P}\left|e_k - \hat{e}_k\right| > c_{10}N^{-\tau}\right) &\le \Pr\left(\bigcup_{j=1}^{P}\left|e_k - \hat{e}_k\right| > c_{10}N^{-\tau}\right) \\ &\le P\Pr\left(\left|e_k - \hat{e}_k\right| > c_{10}N^{-\tau}\right) \\ &\le 6c_6 PRJ_k \exp\left\{-C_1 N^{1-2\rho-2\tau}\right\},\end{aligned} \quad (15)$$

with $N \to \infty$, there are

$$\Pr\left(\max_{1\le k\le P}\left|e_k - \hat{e}_k\right| > c_{10}N^{-\tau}\right) \to 0$$

$$\begin{aligned}\Pr\left(D \subseteq \hat{D}\right) &\ge \Pr\left(\left|e_k - \hat{e}_k\right| > c_{10}N^{-\tau}, \forall k \in D\right) \\ &\ge \Pr\left(\max_{k\in D}\left|e_k - \hat{e}_k\right| > c_{10}N^{-\tau}\right) \\ &\ge 1 - d_0 \Pr\left(\left|e_k - \hat{e}_k\right| > c_{10}N^{-\tau}\right) \\ &\ge 1 - 6c_6 d_0 RJ_k \exp\left\{-C_1 N^{1-2\rho-2\tau}\right\}.\end{aligned} \quad (16)$$

So, $\Pr\left(D \subseteq \hat{D}\right) \to 1$, $N \to \infty$, Theorem 0.3 holds.

Therefore, a proof similar to that of Equation (12), we have:

$$\Pr\left\{\liminf_{N\to\infty}\left(\min_{k\in D}\hat{e}_k - \max_{k\in D^c}\hat{e}_k\right) \le 0\right\} = 1. \quad (17)$$

So, Theorem 3.4 holds.

Thus, based on Equations ((10), (11), (15), (16)), and the proof of Theorem 3.5 and Theorem 3.6 are similar to that of Theorem 3.1 and Theorem 3.2, so they will not be proved in detail.