Scientific Research Publishing

Open Journal of Statistics, 2023, 13, 917-940
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Model-Free Feature Screening via Maximal Information Coefficient (MIC) for Ultrahigh-Dimensional Multiclass Classification

Tingting Chen[1], Guangming Deng[1,2]

[1]College of science, 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: Chen, T.T. and Deng, G.M. (2023) Model-Free Feature Screening via Maximal Information Coefficient (MIC) for Ultrahigh-Dimensional Multiclass Classification. *Open Journal of Statistics*, **13**, 917-940.
https://doi.org/10.4236/ojs.2023.136046

Received: December 6, 2023
Accepted: December 25, 2023
Published: December 28, 2023

Abstract

It is common for datasets to contain both categorical and continuous variables. However, many feature screening methods designed for high-dimensional classification assume that the variables are continuous. This limits the applicability of existing methods in handling this complex scenario. To address this issue, we propose a model-free feature screening approach for ultra-high-dimensional multi-classification that can handle both categorical and continuous variables. Our proposed feature screening method utilizes the Maximal Information Coefficient to assess the predictive power of the variables. By satisfying certain regularity conditions, we have proven that our screening procedure possesses the sure screening property and ranking consistency properties. To validate the effectiveness of our approach, we conduct simulation studies and provide real data analysis examples to demonstrate its performance in finite samples. In summary, our proposed method offers a solution for effectively screening features in ultra-high-dimensional datasets with a mixture of categorical and continuous covariates.

Keywords

Ultrahigh-Dimensional, Feature Screening, Model-Free, Maximal Information Coefficient (MIC), Multiclass Classification

1. Introduction

With the advancement of data acquisition tools and the improvement of computer storage capacity, ultra-high dimensional data has been widely applied in various scientific research fields, especially in genomics, tumor classification,

machine learning and other fields. The traditional data screening methods can no longer be applied, and the existing feature screening methods for ultra-high dimensional data have their own limitations. Among them, when the covariates are both categorical and continuous variables, there are fewer research methods and the screening effect needs to be improved, so there is an urgent need to develop new theoretical and statistical methods to deal with ultra-high dimensional data. The pioneering work by Fan and Lv (2008) [1] introduced the concept of sure independence screening (SIS) in their seminal paper. Specifically, for linear regressions, they demonstrated that the approach based on Pearson correlation learning exhibits a sure screening property. This means that even when the number of predictors (p) grows at a much faster rate than the number of observations (n) with logarithm of p equal to $O\left(n^{\alpha}\right)$ for some $\alpha \in \left(0, \frac{1}{2}\right)$, all relevant predictors can be selected with a probability approaching one (2009) [2].

Numerous approaches have been developed in recent years for feature screening in ultrahigh-dimensional data. Wang (2009) [3] introduced forward regression as a method for handling such data. Fan and Song (2010) [4] applied maximum marginal likelihood estimates or maximum marginal likelihood to ultrahigh-dimensional screening in generalized linear models. Fan *et al.* (2011) [5] extended correlation learning to marginal nonparametric learning. Li *et al.* (2012) [6] presented a robust rank correlation screening method based on the Kendall τ correlation coefficient. He *et al.* (2013) [7] developed a quantile-adaptive framework for nonlinear variable screening in high-dimensional heterogeneous data. Fan *et al.* (2014) [8] introduced nonparametric independence screening, which selects variables based on the nonparametric marginal contributions of each covariate given the exposure variable. Nandy *et al.* (2022) [9] introduced covariate information number sure independence screening, which incorporates a marginal utility connected to the traditional Fisher information. Tong *et al.* (2022) [10] propose a model-free conditional feature screening method for ultra-high-dimensional data based on false discovery rate (FDR) control, which does not require a specific functional form of the regression function and is robust to heavy-tail responses and predictors.

To tackle the challenge of ultrahigh-dimensional feature screening in classification problems, Fan and Fan (2008) [11] introduced the t-test statistic for the two-sample mean problem as a marginal utility for feature screening and established its theoretical properties. Mai and Zou (2013) [12] applied the Kolmogorov filter to ultrahigh-dimensional binary classification. Cui *et al.* (2015) [13] proposed a screening procedure that utilizes empirical conditional distribution functions. Lai *et al.* (2017) [14] developed a feature screening procedure based on the expected conditional Kolmogorov filter for binary classification problems.

However, the aforementioned screening methods assume that the data types are continuous. For categorical covariates, Huang *et al.* (2014) [15] devised a model-free discrete feature screening method based on Pearson Chi-square statistics and demonstrated its sure screening property, as mentioned in Fan *et al.*

(2009) [2]. When all the covariates are binary, Ni and Fang (2016) [16] proposed a model-free feature screening procedure based on information entropy theory for multi-class classification. Ni *et al.* (2017) [17] further extended this by introducing a feature screening procedure based on weighted Adjusted Pearson Chi-square for multi-class classification. Sheng and Wang (2020) [18] introduced a novel model-free feature screening method based on the classification accuracy of marginal classifiers for ultrahigh-dimensional classification. Anzarmou *et al.* (2022) [19] presented a new model-free interaction screening method called Kendall Interaction Filter (KIF) for classification in high-dimensional settings.

Based on the aforementioned research on classification models, this paper introduces a model-free feature screening approach for ultrahigh-dimensional multi-classification that accommodates both categorical and continuous covariates. The proposed method utilizes the maximal information coefficient (MIC) to evaluate the predictive power of the covariates. For screening categorical covariates, we employ the maximal information coefficient (MIC) index, which is equivalent to information gain [16]. The feature screening procedure proposed in this paper is based on maximal information coefficient, specifically referred to as Maximal Information Coefficient Sure Independence Screening (MIC-SIS). The maximum mutual information coefficient can be directly used to categorize the data without any cut-off processing, and it overcomes the disadvantages of information gain, such as the difficulty of calculating the joint probability, not belonging to the measurement method, no way to normalize it, and not being able to compare the results of different data and it, and the screening results are more robust and have lower computational complexity. It first finds an optimal discretization method, and then converts the mutual information value into a measurement method, and the value range is between [0, 1]. MIC has the advantages of wide application range, low computational complexity and strong robustness. The MIC-SIS method is rigorously proven to possess the sure screening property, as originally proposed by Fan and Lv [1], ensuring that all significant features can be identified. Through simulation results, the MIC-SIS approach demonstrates the satisfaction of the sure screening property when compared to existing feature screening methods.

The paper is organized as follows: Section 2 provides a detailed description of the proposed MIC-SIS method. Section 3 establishes the sure screening property of the method. In Section 4, numerical simulations and a real data analysis example are presented to assess the sure screening property of our approach. Concluding remarks are provided in Section 5, and all proofs are included in the Appendix.

2. Feature Screening Procedure

Firstly, we introduce the concept of the maximal information coefficient (MIC), and subsequently, we propose a screening procedure that is based on the maximal information coefficient.

2.1. Maximal Information Coefficient (MIC)

The fundamental principle underlying the maximal information coefficient (MIC) is based on the concept of mutual information. Mutual Information (MI) [20] is a valuable measure in information theory, quantifying the amount of information contained in one random variable regarding another random variable. It represents the reduction in uncertainty of a random variable due to the knowledge of another random variable. In decision trees, mutual information and information gain (IG) are essentially equivalent.

The mutual information between two random variables X and Y is defined based on their joint probability distribution $p(X,Y)$ as follows

$$MI(X,Y)=\int p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y.$$

$MI(X,Y)$ is always nonnegative and $MI(X,Y)=0$ if and only if X and Y are independent.

When covariate X is continuous and Y is a categorical response with R classes $\{1,2,\cdots,R\}$,

$$MI(X,Y)=\sum_{i=1}^{n}\sum_{r=1}^{R}p(X_i,r)\log\left(\frac{p(X_i,r)}{p(X_i)p_r}\right).$$

where $p(X_i,r)=p_rP(X_i\mid Y=r)$ and $p_r=\frac{1}{n}\sum_{i=1}^{n}I(Y_i=r)$. In practice probability functions are usually calculated using Gaussian kernel functions.

When covariate $X=\{X_1,X_2,\cdots,X_p\}$ is a vector of p dimension with J categories, where $j=\{1,2,\cdots,J\}$, and Y is also categorical with R classes $\{1,2,\cdots,R\}$,

$$MI(X,Y)=\sum_{j=1}^{J}\sum_{r=1}^{R}p(X=j,Y=r)\log\left(\frac{p(X=j,Y=r)}{p(X=j)p(Y=r)}\right).$$

where $p(X=j,Y=r)=\frac{1}{n}\sum_{i=1}^{n}I(X_i=j,Y_i=r);r=1,2,\cdots,R;j=1,2,\cdots,J$.

The concept behind MIC is to discretize the relationship between two variables and represent it in two-dimensional space using a scatterplot. A data set consisting of data points with two attributes is distributed in a two-dimensional space. A grid of a multiplied by b is used to divide the data space, and the frequency of data points falling in each (x,y) cell is estimated as $p(x,y)$, which greatly reduces the computational complexity of the joint probability and successfully solves the difficult problem of estimating the joint probability in mutual information.

$$p(x,y)=\frac{\text{number of data points in the }(x,y)\text{ grid}}{\text{total number of data points}}.$$

Since there are several ways to partition the data points using an $a\times b$ grid, our goal is to find the partitioning method that maximizes the mutual information. The mutual information values are normalized using a normalization factor

that maps them to the $[0,1]$ interval. Finally, the mesh resolution that maximises the normalised mutual information is determined as the MIC measure. The formula for *MIC* is expressed in the following equation:

$$MIC(X,Y)=\max_{a*b<B}\frac{MI(X,Y)}{\log_2\min(a,b)}.$$

In the above equation, *a*, *b* is the number of grids divided in the *x*, *y* direction, which is essentially the grid distribution, and *B* is the variable. According to Reshef D N *et al* [21], the grid resolution is typically limited to $a\times b<B$, where the size setting of *B* is often chosen to be approximately 0.6 times the power of the data volume.

$MIC(X,Y)$ is always nonnegative and $MIC(X,Y)=0$ if and only if *X* and *Y* are independent.

2.2. An Independence Ranking and Screening Procedure

We propose a novel model-free sure independence screening method utilizing the maximal information coefficient $MIC(X,Y)$ for analyzing ultrahigh-dimensional data. In this context, *Y* represents the response variable with support Ψ_y, and $X=(X_1,\cdots,X_p)$ denotes the predictor vector, where *p* is significantly larger than the sample size *n*. Without specifying a particular regression model, we define the subset of active predictor indices as follows:

$$D=\{k: p(Y\mid X) \text{ functionally depends on } X_k \text{ for some } y\in\Psi_y\},$$

and define the subset of inactive predictor indices by

$$I=\{k: p(Y\mid X) \text{ does not functionally depend on } X_k \text{ for any } y\in\Psi_y\}.$$

Using the notation mentioned above, we can define the active predictors as $X_D=\{X_k: k\in D\}$ and the inactive predictors as $X_I=\{X_k: k\in I\}$. Our primary objective is to accurately identify the subset of active predictor indices, denoted as *D*.

The *MI* marginal measure can be estimated by letting $\widehat{MI}(X,Y)$. When covariate *X* is continuous and *Y* is a categorical response with *R* classes $\{1,2,\cdots,R\}$,

$$\hat{\omega}_k=\widehat{MI}(X_{ik},Y)=\sum_{i=1}^{n}\sum_{r=1}^{R}\hat{p}(X_{ik},r)\log\left(\frac{\hat{p}(X_{ik},r)}{\hat{p}(X_{ik})\hat{p}_r}\right).$$

where $\hat{p}(X_{ik},r)=\hat{p}_r\hat{P}(X_{ik}\mid Y=r)$ and $\hat{p}_r=\frac{1}{n}\sum_{i=1}^{n}I(Y_i=r)$. Consider a covariate vector $X=\{X_1,X_2,\cdots,X_p\}$ of dimension *p*, where each component X_k takes on J_k categories, represented by $J_k=\{1,2,\cdots,J\}$. Furthermore, the response variable *Y* is also categorical, with *R* classes denoted by $\{1,2,\cdots,R\}$,

$$\hat{\omega}_k=\widehat{MI}(X_k,Y)=\sum_{j=1}^{J_k}\sum_{r=1}^{R}\hat{p}(X_k=j,Y=r)\log\left(\frac{\hat{p}(X_k=j,Y=r)}{\hat{p}(X_k=j)\hat{p}(Y=r)}\right).$$

where $\hat{p}(X_k=j,Y=r)=\frac{1}{n}\sum_{i=1}^{n}I(X_{ik}=j,Y_i=r)$;

$\hat{p}(X_k = j) = \frac{1}{n}\sum_{i=1}^{n} I(X_{ik} = j)$; $\hat{p}(Y = r) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = r)$; $i = 1, 2, \cdots, n$; $r = 1, 2, \cdots, R$; $j = 1, 2, \cdots, J_k$.

The *MIC* marginal measure can be estimated by letting $\widehat{MIC}(X, Y)$. Then

$$\hat{\omega}_k^* = \widehat{MIC}(X_k, Y) = \max_{a*b<B} \frac{\widehat{MI}(X_k, Y)}{\log_2 \min(a,b)} = \max_{a*b<B} \frac{\hat{\omega}_k}{\log_2 \min(a,b)}.$$

Our goal is to calculate the maximal information coefficient *MIC* between each predictor and the response variable, denoted as $\hat{\omega}_k^* = \widehat{MIC}(X_k, Y)$ for $k = 1, 2, \cdots, p$. Note that $\hat{\omega}_k^* = 0$ if and only if $X_k \in X_I$, this also indicates that predictor X_k is statistically independent of *Y*. Therefore, the *MIC* index can be utilized as a measure of dependence to screen the predictors. The MIC-based approach is considered model-free because it solely relies on the marginal and joint densities of the random variables. This index can effectively capture both linear and nonlinear relationships between the response and predictors.

In ultra-high-dimensional data analysis, the primary objective of feature screening is to identify a reduced model with a small number of predictors that can still encompass the true model *D* with a high probability. According to the deterministic screening property proposed by Fan and Lv [2], as the amount of data n tends to infinity, the probability that the model converges to the true model must converge to 1, so that the screened covariates are guaranteed to be valid. In this paper, we propose utilizing the $\hat{\omega}_k^*$ index to select a moderate-sized model

$$\hat{D} = \left\{k : \hat{\omega}_k^* \geq cn^{-\tau}, \text{ for } 1 \leq k \leq p\right\}$$

where c and τ are predetermined positive values. In practice, we often select the reduced model using another formula:

$$\hat{D}^* = \left\{k : \hat{\omega}_k^* \text{ is among the top } d \text{ largest of all}\right\}$$

It is evident that the set of predictors $\left\{X_k : k \in \hat{D}^*\right\}$ represents the most likely relevant predictors associated with the response variable. Consequently, we can employ the predictors in $\left\{X_k : k \in \hat{D}^*\right\}$ to estimate the true model. To simplify the description, we refer to the aforementioned procedure as the MIC-SIS (Maximal Information Coefficient Sure Independence Screening) procedure.

3. Feature Screening Property

In the subsequent sections, we will establish the theoretical properties of the proposed independence screening procedure. Previous studies by Fan and Lv [1], Ji and Jin [22], Zhou and Wang [20] and Ni and Fang [16] have demonstrated that the sure screening property ensures the effectiveness of the independence screening procedure. Hence, it is crucial to establish the sure screening property for MIC-SIS. The following conditions are assumed to guarantee the sure screening property of the MIC-SIS procedure. Although they may not be

the weakest conditions, they are primarily imposed to facilitate the technical proofs.

(C1) Let $X=\left(x_1,x_2,\cdots,x_p\right)$, where x_i is drawn from an unknown distribution F_i. Each distribution F_i has an unknown Lebesgue probability density function pdf f_i, with $i=1,2,\cdots,p$. The conditions specified in Lemma 3 in the Appendix apply to these distributions.

(C2) There exists a positive constant $0<\kappa<2$, such that

$$\sup_{1\le k\le p}\sum_{i=1}^{n}\sum_{r=1}^{R}\log\frac{p\left(X_{ik},Y_r\right)}{p\left(X_{ik}\right)p\left(Y_r\right)}=O\left(n^{\kappa}\right),\ a.e.$$

(C3) There exists a positive constant $c>0$ and τ; the minimum MIC of the active predictors satisfies $\min_{k\in D}\omega_k^*\ge 2cn^{-\tau}$;

(C4) Both X and Y exhibit sub exponential tail probabilities that hold uniformly in p. Specifically, there exists a positive constant μ_0 such that, for all $0<\mu\le\mu_0$, the following condition holds:

$$\sup_p \max_{1\le k\le p} E\left\{\exp\left(\mu\left\|X_k\right\|_1^2\right)\right\}<\infty,\ E\left\{\exp\left(\mu\left\|Y\right\|_q^2\right)\right\}<\infty$$

(C5) There exist two positive constants c_1 and c_2 such that, $c_1/R\le p\left(Y=r\right)\le c_2/R$, $c_1+c_2\le R$, $c_1/R\le p\left(X_k=j,Y=r\right)\le c_2/R$ and $c_1/J_k\le p\left(X_k=j\right)\le c_2/J_k$ for every $1\le J_k\le J$, $1\le r\le R$, and $1\le k\le p$.

(C6) There exist a positive constant c_3, such that $0<f_k\left(x\mid Y=r\right)<c_3$ for any $1\le r\le R$, and x in the domain of X_k, where $f_k\left(x\mid Y=r\right)$ is the Lebesgue density function of X_k conditional on $Y=r$.

(C7) There exist a positive constant c_4 and $0\le\rho<1/2$ such that $f_k\left(x\right)\ge c_4n^{-\rho}$ for any $1\le k\le p$ and x in the domain of X_k, where $f_k\left(x\right)$ is the Lebesgue density function of X_k. Furthermore, $f_k\left(x\right)$ is continuous in the domain of X_k.

(C8) $\liminf_{n\to\infty}\left\{\min_{k\in D}\hat{\omega}_k^*-\max_{k\in I}\hat{\omega}_k^*\right\}>\delta$, where $\delta>0$ is a constant.

According to Ji and Jin [22] and conditions in the Appendix, conditions (C1) and (C2) ensure that the estimated probabilities converge strongly and uniformly to the true probabilities. According to Fan and Lv [1] and Cui [13], condition (C3) allows the minimum true signal to disappear to zero in the order of $n^{-\tau}$ as the sample size goes to infinity. According to the sure screening property proposed by Zhou and Wang [20], then condition (C4) is established. Condition (C5) guarantees that the proportion of each class of variables cannot be either extremely small or extremely large. A similar assumption is also made in condition (C5) in Huang [15] and Cui [13]. To ensure that the sample percentiles are close to the true percentiles, condition (C6) excludes the extreme case that some X_k put heavy mass in a small range. Condition (C7) requires the $n^{-\rho}$ as a lower bound on the density. According to Cui [13], it is easy to show that $\omega_k^*>0$ for $k\in D$ and $\omega_k^*=0$ for $k\in I$ naturally holds. Thus, condition (C8) is established, and MIC index is able to separate active and inactive predictors well at the population level.

Theorem 1 (Sure Screening Property).

Under conditions(C1) - (C4), there exists the positive constant C_1 such that

$$P\left(\hat{\omega}_k^* - \omega_k^* \geq cn^{-\tau}\right) \leq O(p)\exp\left\{-C_1 n^{1-2\tau}\right\}$$

Further, we have that

$$P\left(D \subseteq \hat{D}^*\right) \geq 1 - O\left(s_n \exp\left(-C_1 n^{1-2\tau}\right)\right).$$

In the equation above, s_n represents the cardinality of D. According to Theorem 1, it implies that we can handle the ultra-high-dimensional scenario where the logarithm of p is on the order of $O\left(n^{1-2\tau}\right)$, with $\tau > 0$.

Theorem 2 (Ranking consistency property).

Under conditions(C5) - (C8), if $\log \frac{RJ}{\log n} = O(1)$ and

$\frac{\max\{\log P, \log n\} R^4 J^4}{n^{1-2\rho}} = o(1)$, then

$$\liminf_{n\to\infty} \left\{\min_{k\in D} \omega_k^* - \max_{k\in I} \omega_k^*\right\} > 0, a.s.$$

Theorem 2 demonstrates that the proposed screening index effectively distinguishes between active and inactive covariates at the sample level.

4. Numerical Studies

4.1. Simulation Results

In this subsection, we conduct three simulation studies to demonstrate the finite sample performance of our group screening methods as described in Section 2. We compare the performance of MIC-SIS with that of IG-SIS [16] and APC-SIS [17] using the following evaluation criteria:

1) MMS (Minimal Model Size): This criterion represents the smallest model size that includes all active covariates. The results are presented for various proportions of MMS, such as 5%, 25%, 50%, 75%, and 95%.

2) CP1, CP2, and CP3: These criteria indicate the probabilities that a given model size, specifically $\lceil n/\log n \rceil$, $2\lceil n/\log n \rceil$, and $3\lceil n/\log n \rceil$, respectively, cover all active covariates.

3) CPa: This criterion evaluates whether the indicators of the selected model cover all active covariates.

By comparing these evaluation criteria, we can assess and compare the performance of MIC-SIS, IG-SIS, and APC-SIS in terms of their ability to identify and include active covariates in the model.

Model 1: categorical covariates and binary response

We begin by examining the response variables with different categories. Following Ni and Fang [16], we consider a binary response model where $R = 2$, and all covariates are categorical. We consider two distributions for the response variable y_i:

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

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

The true model is defined at $D=\{1,2,\cdots,20\}$ with $d_0 = |D| = 20$. Condition on y_i, a latent variable z_i is generated as $z_i = (z_{i,1}, z_{i,2}, \cdots, z_{i,p})$, where $z_{i.k} \sim N(\mu_{rk}, 1)$ for $1 \le k \le p$. We then construct the active covariates as follows:

1) If $k > d_0$, then $\mu_{rk} = 0$;

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

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

Next, we generate the covariates by applying the quantile of the standard normal distribution. The specific approach is as follows:

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

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

Where αth percentile of the standard normal distribution is z_α.

Therefore, out of all p covariates, half of them belong to two categories, while the other half belong to five categories. Following the approach in Ni and Fang [17], we consider $p = 1000$ and $p = 5000$, with sample sizes $n = 200$ and $n = 400$ in this model.

Table 1 shows the evaluation criteria for Model 1 based on 100 simulations. The results show the effectiveness of the proposed MIC-SIS method. As the sample size n increases, MIC-SIS approaches the true model size $d_0 = 20$ in terms of MMS, and the coverage probability increases toward 1. MMS performs better in the unbalanced response compared to the balanced response when considering different response structures. MIC-SIS and IG-SIS show similar performance, with MIC-SIS slightly outperforming APC-SIS at higher coverage probabilities.

Model 2: categorical covariates and multi-class response

We further investigate the classification of more covariates, where the response variable y_i has multiple classes with $R = 10$. We consider two distributions for y_i:

1) Balanced, $P(y_i = r) = \frac{1}{R}$;

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

Out of the $p = 2000$ covariates, the minimum set of active covariates is represented by $X^D = \{X_{200}, X_{400}, X_{600}, X_{800}, X_{1000}, \cdots, X_{2000}\}$, with a total of $d_0 = 20$ active covariates. Conditional on y_i, the latent variable $z_i = (z_{i,1}, z_{i,2}, \cdots, z_{i,p})$ is generated, where $z_{i.k}$ follows a standard normal distribution $N(\mu_{i,k}, 1)$ for covariate X_k. Each covariate $x_{i,k}$ is defined as $z_i = (z_{i,1}, z_{i,2}, \cdots, z_{i,p})$, where $\varepsilon_{i,k}$ follows a standard normal distribution $N(0,1)$, and $f_k(\cdot)$ represents the quantile function of the standard normal

Table 1. Simulation results for model 1.

Condition	MMS					CP			
	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
Balanced Y, $n = 200$, $p = 1000$									
MIC-SIS	20.0	20.0	20.0	21.0	22.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.0	20.0	21.0	22.0	1.000	1.000	0.000	0.000
Balanced Y, $n = 400$, $p = 1000$									
MIC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	0.000	0.000
Balanced Y, $n = 200$, $p = 5000$									
MIC-SIS	20.0	21.0	21.0	24.0	28.1	0.994	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	21.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.8	21.0	23.0	28.0	0.996	1.000	0.000	0.000
Balanced Y, $n = 400$, $p = 5000$									
MIC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	0.000	0.000
UnBalanced Y, $n = 200$, $p = 1000$									
MIC-SIS	21.0	23.0	26.0	28.0	32.1	0.974	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	21.0	1.000	1.000	1.000	1.000
APC-SIS	21.0	23.0	25.0	27.0	30.1	0.984	1.000	0.000	0.000
UnBalanced Y, $n = 400$, $p = 1000$									
MIC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	0.000	0.000
UnBalanced Y, $n = 200$, $p = 5000$									
MIC-SIS	32.0	40.8	47.0	59.0	96.4	0.906	0.978	0.995	0.920
IG-SIS	20.0	20.0	20.5	21.0	23.1	1.000	1.000	1.000	1.000
APC-SIS	29.0	35.0	43.0	53.0	89.1	0.914	0.983	0.000	0.000
UnBalanced Y, $n = 400$, $p = 5000$									
MIC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS	20.0	20.0	20.0	20.0	20.0	1.000	1.000	0.000	0.000

distribution. Based on this, we construct the active covariates by defining $\mu_{i,k}$:

1) If $X \in X^D$ and $y_i = r$, then $\mu_{i,k} = 1.5 \times (-0.9)^r$;

2) If $X \notin X^D$ and $y_i = r$, then $\mu_{i,k} = 0$.

Next, we generate the covariates by applying the quantile function $f_k(\cdot)$ to the defined parameters. We consider $p = 2000$ and sample sizes $n = 300, 400$ and 500 for this model. The specific approach is as follows:

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

2) For $400 < k \le 800$, then ($f_k(\varepsilon_{i,k} + \mu_{i,k}) = I\left(z_{i,k} > z_{\frac{i}{4}}\right) + 1$;

3) For $800 < k \le 1200$, then $f_k(\varepsilon_{i,k} + \mu_{i,k}) = I\left(z_{i,k} > z_{\frac{1}{6}}\right) + 1$;

4) For $1200 < k \le 1600$, then $f_k(\varepsilon_{i,k} + \mu_{i,k}) = I\left(z_{i,k} > z_{\frac{j}{8}}\right) + 1$;

5) For $1600 < k \le 2000$, then $f_k(\varepsilon_{i,k} + \mu_{i,k}) = I\left(z_{i,k} > z_{\frac{j}{i0}}\right) + 1$.

Among the $p = 2000$ covariates, each category 2, 4, 6, 8, and 10 constitutes one-fifth of the total.

Table 2 shows the results of the evaluation criteria for 100 simulations of Model 2. The following conclusions can be drawn:

Both methods perform poorly in the more complex Model 2 compared to Model 1. MIC-SIS and IG-SIS perform similarly. As the sample size n increases, MIC-SIS approaches the true model size $d_0 = 10$ in MMS and the coverage probability reaches 1. When the sample size is 300, the coverage probability of APC-SIS is lower compared to MIC-SIS. By comparing the responses of different structures, the unbalanced response has a better MMS performance than the balanced response, the performance of MIC-SIS and IG-SIS is more stable with less fluctuation in MMS. In conclusion, these results highlight the effectiveness and robustness of MIC-SIS and IG-SIS in dealing with Model 2.

Model 3: continuous and categorical covariates

Finally, we consider a more complex example where the response variable y_i is multi-class with $R = 4$. We examine two distributions for y_i:

1) Balanced, $P(y_i = r) = \frac{1}{R}$;

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

In this model, we consider $p = 5000$ and sample sizes $n = 400, 600, 800$. The true model is defined as $X^D = \left\{X_k : k = \left[\frac{k'p}{20}\right], k' = 1, 2, \cdots, 20\right\}$ with $d_0 = 20$. Conditioned on y_i, the latent variable is generated as $z_i = (z_{i,1}, z_{i,2}, \cdots, z_{i,p})$, where $z_{i.k} \sim N(\mu_{i,k}, 1)$ for $1 \le k \le p$. For the covariates X_k, we have $x_{i,k} \sim N(\mu_{i,k}, 1)$ for $1 \le k \le p$, where

Table 2. Simulation results for model 2.

Condition	MMS					CP			
	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
Balanced Y, $n = 300$, $P = 2000$									
MIC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	12.0	17.0	21.0	30.0	40.0	0.989	1.000	0.000	0.000
Balanced Y, $n = 400$, $P = 2000$									
MIC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	0.000	0.000
Balanced Y, $n = 500$, $P = 2000$									
MIC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	0.000	0.000
UnBalanced Y, $n = 300$, $P = 2000$									
MIC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	10.0	11.0	12.0	13.0	1.000	1.000	0.000	0.000
UnBalanced Y, $n = 400$, $P = 2000$									
MIC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	0.000	0.000
UnBalanced Y, $n = 500$, $P = 2000$									
MIC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
IG-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	1.000	1.000
APC-SIS	10.0	10.0	10.0	10.0	10.0	1.000	1.000	0.000	0.000

$\mu_{i,k} \sim (-1)^r \theta_{rk} \mu_{i,k} \sim (-1)^r \theta_{rk}$ if $y_i = r$ and $k \in D$. The values of θ_{rk}, as given in **Table 3** by Ni and Fang [16], determine the active covariates. Specifically, $\mu_{i,k} = 0$ when $k \notin D$.

An active covariate is established by defining $\mu_{i,k}$:

1) For $k \le \left[\frac{5p}{20}\right]$, then $x_{i,k} = j$, if $z_{i,k} \in \left(z_{\frac{j-1}{4}}, z_{\frac{j}{4}}\right], j = 1,2,3,4$;

2) For $\left[\frac{5p}{20}\right] < k \le \left[\frac{10p}{20}\right]$, then $x_{i,k} = j$, if $z_{ik} \in \left(z_{\frac{j-1}{0}}, z_{\frac{j}{10}}\right], j = 1,2,\cdots,10$;

3) For $\left[\frac{10p}{20}\right] < k \le p$, then $x_{i,k} = z_{i,k}$.

Table 3. Parameter specification of Model 3.

θ_{rk}	K									
	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	0.4
$r=3$	0.1	0.9	0.9	0.1	0.3	0.1	0.4	0.3	0.6	0.6

Among all the p covariates, four categories and ten categories account for one-fifth each, respectively, while the remaining covariates are continuous. Similarly, there are 5 active covariates in each of the four categories and ten categories, with the remaining active covariates being continuous, accounting for half of the total. For the continuous covariates, we applied different subdivisions with $J_k = 4, 8$ and 10. Accordingly, we define the corresponding approaches as MIC-SIS-4, IG-SIS-4, APC-SIS-4, MIC-SIS-8, IG-SIS-8, APC-SIS-8, MIC-SIS-10, IG-SIS-10, and APC-SIS-10.

Table 4 and Table 5 present the simulation results based on over 100 simulations for the balanced and unbalanced cases, respectively. The following observations can be made: As the sample size n increases, MIC-SIS approaches the true model size $d_0 = 20$ in terms of MMS, and both approach 1 in terms of coverage probability. The coverage probability of MIC-SIS is similar to that of IG-SIS for all five indices, demonstrating the characteristic screening properties of MIC-SIS. For MMS, the unbalanced response outperforms the balanced response when comparing response structures. In addition, both MIC-SIS and IG-SIS exhibit robust performance, as evidenced by the small range of variation in MMS for both response types. When applying different breakdowns to the continuous covariates, MIC-SIS and IG-SIS outperform the other methods in terms of coverage probability and MMS when comparing response structures.

4.2. Real Data

In this subsection, we analyze a real dataset obtained from the feature selection database of Arizona State University (http://featureselection.asu.edu/). The dataset, called GLIOMA biological data, consists of 50 samples and 4434 features. The data is unbalanced due to the response variable, with class sizes of 14, 7, 14, and 15. The covariates in this dataset are both continuous and multiclass. We randomly divided the data into two parts, with 90% used as training data and 10% used as test data. The training data consists of 45 samples, while the test data consists of 5 samples. The dimensionality of both the training and test data is $p = 4434$.

To assess the performance of MIC-SIS, PG-SIS, IG-SIS, and APC-SIS, we employ three classification approaches: Support Vector Machine (SVM) [23], Random Forest (RF), and Decision Tree (DT). We utilize a ten-fold cross-validation to address potential issues related to varying training data that could affect the

accuracy of the models. These classification approaches are applied to the selected active covariates obtained from the aforementioned screening methods. The evaluation metrics commonly used in such analyses include accuracy, recall,

Table 4. Simulation results for model 3 (Balanced Y).

Condition	MMS					CP			
	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
Balanced Y, $n = 400$, $P = 5000$									
MIC-SIS-4	20.0	20.0	20.0	22.0	23.0	1.000	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	20.0	20.0	21.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	20.0	20.0	21.0	21.1	1.000	1.000	1.000	1.000
MIC-SIS-8	20.0	20.0	21.0	22.0	24.0	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	20.0	21.0	21.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	20.0	20.0	21.0	22.0	1.000	1.000	1.000	1.000
MIC-SIS-10	20.0	21.0	22.0	23.3	26.0	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	20.0	20.0	21.0	22.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	20.0	20.0	21.0	21.1	1.000	1.000	1.000	1.000
Balanced Y, $n = 600$, $P = 5000$									
MIC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
MIC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
MIC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
Balanced Y, $n = 800$, $P = 5000$									
MIC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
MIC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
MIC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000

Table 5. Simulation results for model 3 (Unbalanced Y).

Condition	MMS					CP			
	5%	25%	50%	75%	95%	CP1	CP2	CP3	CPa
	Unbalanced Y, $n = 400$, $P = 5000$								
MIC-SIS-4	20.0	21.0	22.0	22.0	24.0	1.000	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	21.0	22.0	23.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	21.0	22.0	23.0	24.0	1.000	1.000	1.000	1.000
MIC-SIS-8	20.0	21.0	22.0	23.0	24.6	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	21.0	22.0	23.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	21.0	22.0	23.0	24.0	1.000	1.000	1.000	1.000
MIC-SIS-10	21.5	23.0	25.0	26.0	28.0	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	21.0	22.0	23.0	24.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	21.0	22.0	23.0	24.0	1.000	1.000	1.000	1.000
	Unbalanced Y, $n = 600$, $P = 5000$								
MIC-SIS-4	20.0	20.0	20.0	20.0	21.0	1.000	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
MIC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
MIC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
	Unbalanced Y, $n = 800$, $P = 5000$								
MIC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-4	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
MIC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-8	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
MIC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
IG-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000
APC-SIS-10	20.0	20.0	20.0	20.0	20.0	1.000	1.000	1.000	1.000

F-measure, and G-mean. In this paper, we specifically utilize G-mean and F-measure [24] to assess the performance of the models on both the training and test data. The performance of MIC-SIS for unbalanced data is presented in Table 6.

Table 6. Analysis results for real data example.

	screening method	response			
		1	2	3	4
classification method		SVM			
G-mean (train data)	APC-SIS	1.0000	0.9304	0.9709	0.9713
	IG-SIS	1.0000	0.9025	0.9093	1.0000
	PG-SIS	0.9853	0.9378	0.9514	0.9946
	MIC-SIS	1.0000	0.8476	0.9504	0.9827
G-mean (test data)	APC-SIS	0.9775	0.9913	0.9564	0.9379
	IG-SIS	1.0000	0.9439	0.8736	0.9922
	PG-SIS	0.9678	0.9779	0.9173	0.9739
	MIC-SIS	1.0000	0.8988	0.9118	0.9302
F-measure (train data)	APC-SIS	0.7673	0.5958	0.7124	0.7101
	IG-SIS	0.6924	0.3969	0.4121	0.7004
	PG-SIS	0.7307	0.6095	0.6511	0.7469
	MIC-SIS	0.6544	0.6053	0.5324	0.6176
F-measure (test data)	APC-SIS	0.5424	0.3233	0.5333	0.4033
	IG-SIS	0.5850	0.1667	0.2167	0.5505
	PG-SIS	0.4533	0.2667	0.3967	0.5057
	MIC-SIS	0.4988	0.3000	0.2967	0.3600
classification method		DT			
G-mean (train data)	APC-SIS	0.9909	0.8898	0.9489	0.9872
	IG-SIS	0.9945	0.8743	0.9437	0.9917
	PG-SIS	0.9815	0.8902	0.9578	0.9835
	MIC-SIS	0.9944	0.8707	0.9454	0.9758
G-mean (test data)	APC-SIS	0.9913	0.9626	0.9371	0.9774
	IG-SIS	0.9913	0.9200	0.9371	0.9862
	PG-SIS	0.9862	0.8963	0.8838	0.9609
	MIC-SIS	1.0000	0.8942	0.8961	0.9059
F-measure (train data)	APC-SIS	0.6743	0.2915	0.5757	0.6648
	IG-SIS	0.6668	0.1792	0.5466	0.6613
	PG-SIS	0.6424	0.2807	0.5825	0.6468
	MIC-SIS	0.6450	0.1689	0.5276	0.6051
F-measure (test data)	APC-SIS	0.5457	0.1967	0.4000	0.5367
	IG-SIS	0.5790	0.0500	0.4333	0.5471
	PG-SIS	0.4667	0.0500	0.2933	0.4333
	MIC-SIS	0.5000	0.0500	0.2667	0.3467

Continued

classification method		RF			
G-mean (train data)	APC-SIS	1.0000	0.9458	1.0000	1.0000
	IG-SIS	1.0000	0.9458	1.0000	1.0000
	PG-SIS	0.9923	0.9421	0.9782	1.0000
	MIC-SIS	1.0000	0.9458	1.0001	1.0072
G-mean (test data)	APC-SIS	1.0000	0.9807	0.9540	0.9835
	IG-SIS	1.0000	0.9894	0.9523	0.9453
	PG-SIS	0.9871	0.9524	0.9384	0.9774
	MIC-SIS	0.9826	0.9645	0.9234	0.9173
F-measure (train data)	APC-SIS	1.0000	1.0000	1.0000	1.0000
	IG-SIS	1.0000	1.0000	1.0000	1.0000
	PG-SIS	0.8603	0.7671	0.8417	0.8725
	MIC-SIS	1.0000	1.0000	1.0000	1.0000
F-measure (test data)	APC-SIS	0.6624	0.3500	0.6300	0.6300
	IG-SIS	0.6124	0.3567	0.5733	0.4667
	PG-SIS	0.5967	0.2833	0.4933	0.5733
	MIC-SIS	0.4857	0.1800	0.3733	0.3200

According to the results of Table 6, we can get that among all the classification methods, MIC-SIS consistently outperforms the others, exhibiting higher G-mean and F-measure values that are closer to 1. In summary, the proposed MIC-SIS method demonstrates superior performance.

5. Conclusions

In practical scenarios, it is common to encounter datasets with a combination of continuous and categorical covariates, along with categorical responses. However, the available screening methods for such cases are limited. To address this problem, we introduce a new method called MIC-SIS (Maximum Information Coefficient-based Screening), which does not require continuous variables to be sliced and can be applied directly to a wide range of variables, overcoming the shortcomings of existing methods. In this paper, we demonstrate that MIC-SIS has theoretical properties such as deterministic screening and ranking consistency, and that no modeling is required. Through numerical simulations, we find that MIC-SIS can effectively screen covariates with better screening and lower computational complexity than existing methods. It also performs well in the empirical analysis of GLIOMA data.

One of the current challenges in covariate screening is missing data. It is common to have missing or incorrect data during the data collection process, which can affect the variable screening results. In future work, we aim to develop

a new approach to feature screening that can either deal with missing variables prior to screening, or use classification models to screen features based on response variables.

Acknowledgements

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

Availability of Data

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

Conflicts of Interest

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

References

[1] 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

[2] 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.

[3] Wang, H.S. (2009) Forward Regression for Ultra-High Dimensional Variable Screening. *Journal of the American Statistical Association*, **104**, 1512-1524. https://doi.org/10.1198/jasa.2008.tm08516

[4] 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

[5] 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

[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] He, X.M., Wang, L. and Hong, H.G. (2013) Quantile-Adaptive Model-Free Variable Screening for High-Dimensional Heterogeneous Data. *The Annals of Statistics*, **41**,342-369. https://doi.org/10.1214/13-AOS1087

[8] Fan, J.Q., Ma, Y.B. and Dai, W. (2014) Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Varying Coefficient Models. *Journal of the American Statistical Association*, **109**, 1270-1284. https://doi.org/10.1080/01621459.2013.879828

[9] Nandy, D., Chiaromonte, F. and Li, R.Z. (2022) Covariate Information Number for Feature Screening in Ultrahigh-Dimensional Supervised Problems. *Journal of the American Statistical Association*, **117**, 1516-1529. https://doi.org/10.1080/01621459.2020.1864380

[10] Tong, Z.X., Cai, Z.R., Yang, S.S. and Li, R.Z. (2022) Model-Free Conditional Feature Screening with FDR Control. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2022.2063130

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

[12] 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

[13] 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

[14] Lai, P., Song, F.L. and Chen, K.W. (2017) Model Free Feature Screening with Dependent Variable in Ultrahigh Dimensional Binary Classification. *Statistics & Probability Letters*, **125**, 141-148. https://doi.org/10.1016/j.spl.2017.02.011

[15] 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

[16] 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

[17] Ni, L., Fang, F. and Wan, F.J. (2017) Adjusted Pearson Chi-Square Feature Screening for Multi-Classification with Ultrahigh Dimensional Data. *Metrika*, **80**, 805-828. https://doi.org/10.1007/s00184-017-0629-9

[18] Sheng, Y. and Wang, Q.H. (2020) Model-Free Feature Screening for Ultrahigh Dimensional Classification. *Journal of Multivariate Analysis*, **178**, Article ID: 104618. https://doi.org/10.1016/j.jmva.2020.104618

[19] Anzarmou, Y., Mkhadri, A. and Oualkacha, K. (2022) The Kendall Interaction Filter for Variable Interaction Screening in High Dimensional Classification Problems. *Journal of Applied Statistics*, **50**, 1496-1514.

[20] Zhou, S.B., Wang, T. and Huang, Y.J. (2022) Feature Screening via Mutual Information Learning Based on Nonparametric Density Estimation. *Journal of Mathematics*, **2022**, Article ID: 7584374. https://doi.org/10.1155/2022/7584374

[21] Reshef, D.N., Reshef, Y.A. and Finucane, H.K. (2011) Detecting Novel Associations in Large Data Sets. *Science*, **334**, 1518-1524. https://doi.org/10.1126/science.1205438

[22] Ji, P.S. and Jin, J.S. (2012) UPS Delivers Optimal Phase Diagram in High-Dimensional Variable Selection. *The Annals of Statistics*, **40**, 73-103. https://doi.org/10.1214/11-AOS947

[23] Suykens, J.A.K. and Vandewalle, J. (1999) Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, **9**, 293-300. https://doi.org/10.1023/A:1018628609742

[24] He, H.J. and Deng, G.M. (2022) Grouped Feature Screening for Ultra-High Dimensional Data for the Classification Model. *Journal of Statistical Computation and Simulation*, **92**, 974-997. https://doi.org/10.1080/00949655.2021.1981901

[25] Serfling, R.J. (1980) Approximation Theorems of Mathematical Statistics. John Wiley & Sons Inc., New York. https://doi.org/10.1002/9780470316481

[26] Dias, R., Garcia, N.L. and Zambom, A.Z. (2012) Monte Carlo Algorithm for Trajectory Optimization Based on Markovian Readings. *Computational Optimization*

and Applications, **51**, 305-321. https://doi.org/10.1007/s10589-010-9337-3

[27] Davis, R.A., Lii, K.S. and Politis, D.N. (2011) Selected Works of Murray Rosenblatt. Springer, New York. https://doi.org/10.1007/978-1-4419-8339-8

[28] 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

Appendix

Proof of Theoretical Result.

To establish the validity of Theorem 1, we rely on three accompanying lemmas. The first two lemmas furnish us with exponential inequalities, and their detailed proofs can be found in [25].

Lemma 1. Let $\mu = E(Y)$. If $P(a \le Y \le b) = 1$, then

$$E\left[\exp\left\{s(Y-\mu)\right\}\right] \le \exp\left\{\frac{s^2(b-a)^2}{8}\right\}.$$

Lemma 2. Let $h(Y_1,\cdots,Y_m)$ be a kernel function for the U statistics U_n, and $\theta = E\left\{h(Y_1,\cdots,Y_m)\right\}$. If $a \le h(Y_1,\cdots,Y_m) \le b$ holds, then for any $t>0$ and $n \ge m$, we have the following inequality

$$P(U_n - \theta \ge t) \le \exp\left(\frac{-2\left[\frac{n}{m}\right]t^2}{(b-a)^2}\right).$$

where $[n/m]$ denotes the integer part of n/m.

Lemma 2 represents the one-sided tail inequality of U_n. As a result of its symmetry, we can readily derive the two-sided tail inequality of U_n

$$P(|U_n - \theta| \ge t) \le 2\exp\left(\frac{-2\left[\frac{n}{m}\right]t^2}{(b-a)^2}\right).$$

Lemma 3. (The asymptotic property of nonparametric density estimators). Suppose that $f''(x)$ exists and $h = cn^{-\frac{1}{5}}$, then

$$n^{\frac{2}{5}}\left\{\hat{p}(x) - p(x)\right\} \xrightarrow{L} N\left(\frac{c^2}{2}f''(x)\mu_2(K), \frac{1}{c}f(x)\|K\|_2^2\right).$$

From the above equation, $\mu_2(K) = \int s^2 K(s)\mathrm{d}s$ and $\|K\|_2^2 = \int K^2(s)\mathrm{d}s$.

Lemma 3 directly implies that $\hat{p}(x) \xrightarrow{p} p(x)$. Under certain additional stringent conditions, we can achieve strong uniform convergence of $\hat{p}(x)$.

$$\lim_{n\to\infty}\sup_x\left|\hat{p}(x) - p(x)\right| = 0, \ a.e.$$

For more detailed information regarding the strong uniform convergence, one can refer to references such as [26] or [27]. These sources provide further insights and explanations on the topic.

Proof of Theorem 1. We begin by demonstrating that the following inequality holds for each k:

$$P\left\{\left|\hat{\omega}_k^* - \omega_k^*\right| \ge cn^{-\tau}\right\} \le O\left(\exp\left(-C_1 n^{1-2\tau}\right)\right).$$

Since $\hat{\omega}_k^*$ is obtained by normalizing $\hat{\omega}_k$ without altering its properties, we only need to establish that the aforementioned inequality holds for each k:

$$P\left\{\max_{a*b<B}\frac{\left|\hat{\omega}_k-\omega_k\right|}{\log_2\min(a,b)}\geq cn^{-\tau}\right\}\leq O\left(\exp\left(-C_1n^{1-2\tau}\right)\right).$$

Because

$$\begin{aligned}&\max_{a*b<B}\log_2\min(a,b)\left|\hat{\omega}_k-\omega_k\right|=\left|\hat{\omega}_k-\omega_k\right|\\&=\left|\sum_{i=1}^{n}\sum_{r=1}^{R}\hat{p}(X_{ik},Y_r)\log\frac{\hat{p}(X_{ik},Y_r)}{\hat{p}(X_{ik})\hat{p}(Y_r)}-\int p(x_k,y)\log\frac{p(x_k,y)}{p(x_k)p(y)}\mathrm{d}x_k\mathrm{d}y\right|\\&=\left|\sum_{i=1}^{n}\sum_{r=1}^{R}\hat{p}(X_{ik},Y_r)\log\frac{\hat{p}(X_{ik},Y_r)}{\hat{p}(X_{ik})\hat{p}(Y_r)}-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{r=1}^{R}\log\frac{\hat{p}(X_{ik},Y_r)}{\hat{p}(X_{ik})\hat{p}(Y_r)}\right.\\&\left.+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{r=1}^{R}\log\frac{\hat{p}(X_{ik},Y_r)}{\hat{p}(X_{ik})\hat{p}(Y_r)}-p(x_k,y)\log\int\frac{p(x_k,y)}{p(x_k)p(y)}\mathrm{d}x_k\mathrm{d}y\right|\\&=\left|M_{k,1}+M_{k,2}\right|.\end{aligned}$$

Using Lemma 3 and the strong law of large numbers, we observe the convergence of

$$M_{k,1}=\sum_{i=1}^{n}\sum_{r=1}^{R}\hat{p}(X_{ik},Y_r)\log\frac{\hat{p}(X_{ik},Y_r)}{\hat{p}(X_{ik})\hat{p}(Y_r)}-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{r=1}^{R}\log\frac{\hat{p}(X_{ik},Y_r)}{\hat{p}(X_{ik})\hat{p}(Y_r)}\to 0,\ a.e.$$

Next, our goal is to establish an upper bound for the second term.

Define $h(X_{ik},Y_r;X_k,Y)=\log\frac{\hat{p}(X_{ik},Y_r)}{\hat{p}(X_{ik})\hat{p}(Y_r)}$ as the kernel of the U statistics of $I_{k,2}^*$, where we define $M_{k,2}=I_{k,2}-\omega_k$, and

$$I_{k,2}^*=I_{k,2}-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{r=1}^{R}\log\frac{\hat{p}(X_{ik},Y_r)}{\hat{p}(X_{ik})\hat{p}(Y_r)}.$$

By applying Markov's inequality, we can ensure that:

$$P\left(I_{k,2}^*-Eh>\varepsilon\right)\leq\exp(-t\varepsilon)\exp(-tEh)E\left[\exp\left(tI_{k,2}^*\right)\right],\text{ for any } t>0.$$

where $Eh=\int p(x_k,y)\log\frac{p(x_k,y)}{p(x_k)p(y)}\mathrm{d}x_k\mathrm{d}y$.

Following the approach utilized by Li *et al.* (2012) [28] to handle the U statistics, and considering condition (C2), we can immediately deduce that:

$$P\left(I_{k,2}^*-Eh>\varepsilon\right)\leq\exp\left(\frac{-t\varepsilon+t^2}{8n}\right).$$

By selecting $t=4n\varepsilon$, we obtain $P\left(I_{k,2}^*-Eh>\varepsilon\right)\leq\exp\left(-2n\varepsilon^2\right)$. Consequently, taking into account the symmetry of U statistics, we can derive the bilateral tail inequality:

$$P\left(\left|I_{k,2}^*-Eh\right|>\varepsilon\right)\leq 2\exp\left(-2n\varepsilon^2\right).$$

By utilizing the relationship between $I_{k,2}^*$ and $I_{k,2}$, we can demonstrate that

$$P\left(\left|M_{k,2}\right|>2\varepsilon\right)=P\left(\left|I_{k,2}-\omega_k^*\right|>2\varepsilon\right)$$

$$=P\left(\left|I_{k,2}^*+\frac{1}{n^2}\sum_{i=r}\log\frac{\hat{p}\left(X_{ik},Y_r\right)}{\hat{p}\left(X_{ik}\right)\hat{p}\left(Y_r\right)}-\omega_k^*\right|>2\varepsilon\right).$$

Under condition (C2), for $\varepsilon>0$, we can choose a sufficiently large N_1 such that when $n>N_1$, $\frac{1}{n^2}\sum_{i=r}\log\frac{\hat{p}\left(X_{ik},Y_r\right)}{\hat{p}\left(X_{ik}\right)\hat{p}\left(Y_r\right)}<\frac{\varepsilon}{3}$. Furthermore, we can easily establish that

$$P\left(\left|I_{k,2}-\omega_k^*\right|>2\varepsilon\right)\le P\left(\left|I_{k,2}^\star-\omega_k^*\right|>\frac{5}{3}\varepsilon\right).$$

Note that

$$\left|I_{k,2}-\omega_k^*\right|=\left|I_{k,2}-Eh+Eh-\omega_k^*\right|.$$

Similarly, employing the same technique and selecting a larger N_2, we can ensure that when $n>N_2$, $\left|\omega_k^*-Eh\right|<\frac{\varepsilon}{3}$. This directly implies that

$$P\left(\left|I_{k,2}^*-\omega_k^*\right|>\frac{5}{3}\varepsilon\right)\le P\left(\left|I_{k,2}^*-Eh\right|>\frac{4}{3}\varepsilon\right).$$

Let $\varepsilon=cn^{-\tau}$, $0<\tau<\frac{1}{2}$. By employing $P\left(\left|I_{k,2}^*-Eh\right|>\varepsilon\right)\le 2\exp\left(-2n\varepsilon^2\right)$, together with Lemma 1 and Bonferroni's inequality, we can deduce that

$$P\left\{\left|\hat{\omega}_k^*-\omega_k^*\right|\ge cn^{-\tau}\right\}\le 2\exp\left(-2c^2n^{1-2\tau}\right).$$

We thus have

$$P\left\{\max_{1\le k\le p}\left|\hat{\omega}_k^*-\omega_k^*\right|\ge cn^{-\tau}\right\}\le 2p\exp\left(-2c^2n^{1-2\tau}\right)=O\left(p\left[\exp\left(-C_1n^{1-2\tau}\right)\right]\right).$$

Next, we prove the second part of Theorem 1.

If $D\subseteq\hat{D}^*$, it implies that there must exist some $k\in D$ such that $\hat{\omega}_k^*=cn^{-\tau}$. By utilizing condition (C3), we can deduce that if $\hat{\omega}_k^*=cn^{-\tau}$ holds for some $k\in D$, then $\hat{\omega}_k^*=cn^{-\tau}$ also holds for some $k\in D$. Thus, the event $\left\{D\subseteq\hat{D}^*\right\}$ is a subset of the event $\left\{\left|\hat{\omega}_k^*-\omega_k^*\right|\ge cn^{-\tau}, \text{for some} k\in D\right\}$. Taking the complement on both sides, we obtain $\left\{\max_{k\in D}\left|\hat{\omega}_k^*-\omega_k^*\right|\le cn^{-\tau}\right\}$ is a subset of $\left\{D\subseteq\hat{D}^*\right\}$. Therefore, we have:

$$P\left(D\subseteq\hat{D}^*\right)\ge P\left\{\max_{k\in D}\left|\hat{\omega}_k^*-\omega_k^*\right|\le cn^{-\tau}\right\}$$

$$=1-P\left\{\min_{k\in D}\left|\hat{\omega}_k^*-\omega_k^*\right|\ge cn^{-\tau}\right\}$$

$$=1-s_nP\left\{\left|\hat{\omega}_k^*-\omega_k^*\right|\ge cn^{-\tau}\right\}$$

$$\ge 1-O\left(s_n\left[\exp\left(-C_1n^{1-2\tau}\right)\right]\right).$$

In the above equation, s_n is the cardinality of D.

Now we prove Theorem 2. The proofs for Lemma 4 and Lemma 5 can be

found in Ni and Fang (2016) [16].

Lemma 4. For categorical covariates X_k and response Y,under condition (C5), for any $0<\varepsilon<1$, we have $P\left(\left|\hat{\omega}_k^*-\omega_k^*\right|>2\varepsilon\right)\le O(RJ\sim)\exp\left\{-c_5\frac{n\varepsilon^2}{R^4J^4}\right\}$, where c_5 represents a positive constant.

Lemma 5. For continuous covariates X_k and response Y, under condition (C5), (C6) and (C7), for any $0<\varepsilon<1$, we have $P\left(\left|\hat{\omega}_k^*-\omega_k^*\right|>2\varepsilon\right)\le O(RJ\sim)\exp\left\{-c_6\frac{n^{1-2\rho}\varepsilon^2}{R^4J^4}\right\}$, where c_6 represents a positive constant.

Under Conditions (C5)-(C8) and by Lemmas 4 and 5, if $\log\frac{RJ}{\log n}=O(1)$ and $\frac{\max\{\log P,\log n\}R^4J^4}{n^{1-2\rho}}=o(1)$, we get

$$
\begin{aligned}
&P\left(\min_{k\in D}\hat{\omega}_k^*-\max_{k\in I}\hat{\omega}_k^*<\frac{\delta}{2}\right)\\
&\le P\left(\left(\min_{k\in D}\hat{\omega}_k^*-\max_{k\in I}\hat{\omega}_k^*\right)-\left(\min_{k\in D}\omega_k^*-\max_{k\in I}\omega_k^*\right)<-\frac{\delta}{2}\right)\\
&\le P\left(\left|\left(\min_{k\in D}\hat{\omega}_k^*-\max_{k\in I}\hat{\omega}_k^*\right)-\left(\min_{k\in D}\omega_k^*-\max_{k\in I}\omega_k^*\right)\right|>\frac{\delta}{2}\right)\\
&\le P\left(\max_{1\le k\le p}\left|\hat{\omega}_k^*-\omega_k^*\right|>\frac{\delta}{4}\right)\le O(RJ_k)\,p\exp\left\{-c_5\frac{n^{1-2\rho}}{R^4J_k^4}\right\}\\
&=O\left(\exp\left\{\log RJ+\log p-c_7\frac{n^{1-2\rho}}{R^4J^4}\right\}\right)
\end{aligned}
$$

where $c_7=\min\{c_5,c_6\}(\delta/4)^2$. Since $\log\frac{RJ}{\log n}=O(1)$, there exists a positive constant c_8 such that $\log(RJ)\le c_8\log n$. Also, $\frac{\max\{\log P,\log n\}R^4J^4}{n^{1-2\rho}}=o(1)$ implies that $\log p\le\frac{\frac{1}{2}c_7n^{1-2\rho}}{R^4J^4}$ and $\frac{\frac{1}{2}c_7n^{1-2\rho}}{R^4J^4}\ge(c_8+2)\log n$ for large n. Then there exists a n_0 such that

$$
\begin{aligned}
&\sum_{n=n_0}^{\infty}\exp\left\{\log RJ+\log p-c_7\frac{n^{1-2\rho}}{R^4J^4}\right\}\le\sum_{n=n_0}^{\infty}\exp\left\{c_8\log n-\frac{\frac{1}{2}c_7n^{1-2\rho}}{R^4J^4}\right\}.\ \text{Accord-}\\
&\le\sum_{n=n_0}^{\infty}\exp\left\{c_8\log n-(c_8+2)\log n\right\}=\sum_{n=n_0}^{\infty}n^{-2}<\infty
\end{aligned}
$$

ing to Ni and Fang (2016) [16] and by the Borel Cantelli Lemma, we can get $\liminf_{n\to\infty}\left\{\min_{k\in D}\omega_k^*-\max_{k\in I}\omega_k^*\right\}\ge\frac{\delta}{2}>0$, *a.s.*

This is the end of the proof.