Hindawi
Journal of Mathematics
Volume 2022, Article ID 7584374, 11 pages
https://doi.org/10.1155/2022/7584374

Research Article

Feature Screening via Mutual Information Learning Based on Nonparametric Density Estimation

Shengbin Zhou,[1] Tao Wang ,[1] and Yejin Huang[2]

[1] *School of Mathematical Sciences, Harbin Normal University, Harbin 150025, China*
[2] *China Securities Data Co., Ltd, Beijing 100032, China*

Correspondence should be addressed to Tao Wang; wangtaohrb@gmail.com

Received 31 October 2021; Revised 12 January 2022; Accepted 15 January 2022; Published 8 February 2022

Academic Editor: Xiaolong Qin

With the advent of the era of big data, feature selection in high- or ultra-high-dimensional data is increasingly important in statistics and machine learning fields. In this paper, we propose a marginal utility measure screening method MI-SIS based on mutual information. The proposed marginal utility measure has several appealing features compared with the existing independence screening methods. Firstly, the proposed procedure is model-free without specifying any relationship between the predictors and the response and is valid under a wide range of model settings including parametric and nonparametric models. Secondly, it is suitable for various combinations of the continuous and categorical of predictors and response in our new method. Finally, the new procedure has a good performance in discovering a weak signal in the finite sample and its computation is simple and easy to implement. We establish the sure screening property for the proposed procedure with mild conditions. Simulation experiments and real data applications are presented to illustrate the finite sample performance of the proposed procedures.

1. Introduction

With the development of the information and technology, more and more data were collected in many scientific areas. The number of predictors p can be bigger than the number of samples n. Theoretically, classical results allow that p may diverge at exponential rate of the sample size. The difficulty in ultra-high-dimensional data analysis is obvious; the most typical difficulty is a large amount of calculation. Fan and Lv [1] proposed a relatively new approach to deal with this problem; they proposed a sure independent screening (SIS, hereafter) method based on the Pearson correlation. Since the work of Fan and Lv [1] on sure independent screening, there has been a lot of work on feature screening. Wang [2] proposed Forward Regression for linear models; Fan et al. [3] designed a screening method based on the marginal likelihood estimators for generalized linear models and robust regression. Fan and Song [4] established the sure independence screening property under the background of generalized linear models. Within the scope of the ultra-high-dimensional nonparametric modeling, various feature screening methods have also been proposed. Those methods include but are not limited to the following: nonparametric independence screening (Fan et al. [5]), conditional correlation sure independence screening (Liu et al. [6]), iterative nonparametric independence screening (Fan et al. [7]), and others. All these methods are based on specific model assumption.

The model-based screening methods have a good performance when the underling model is correctly specified. However, it is challenging to specify a correct model for ultra-high-dimensional data analysis. Model-based screening methods may perform poorly in the case of model misspecification. In order to overcome this problem, great efforts have been made to relax the model assumption and make the screening methods less model-dependent. Thus, new model-free screening methods have been proposed in the latest literature. See the works of Zhu et al. [8], Li et al. [9], He et al. [10], Mai and Zou [11], Shao and Zhang [12], Cui et al. [13], and others. In particular, Zhu et al. [8]

proposed a sure independent ranking and screening (SIRS) method to screen predictors in multi-index models. They further showed that SIRS enjoys the ranking consistency property. Li et al. [9] developed a distance correlation based screening methods (DC-SIS). DC-SIS does not require model specification, and it can deal with grouped predictor variables. They further demonstrated that DC-SIS procedure possessed the sure screening property. Shao and Zhang [12] proposed a martingale difference correlation based screening method (MDC-SIS), which is also a model-free screening method; MDC-based screening methods can extend to quantile regression, and the sure screening property was also established.

In ultra-high-dimensional classification problem, Cui et al. [13] proposed a screening method for classification problem (MV-SIS); this method can be applied to the situation where one of the predictors and the response are categorical. Huang et al. [14] proposed a screening method based on Pearson Chi-square (PC-SIS); PC-SIS is applicable in case the response and the predictors are categorical; by categorizing the continuous variables, both MV-SIS and PC-SIS have a wider applicability. We were inspired by the simplicity of MV-SIS and PC-SIS; in this paper, we proposed a new screening method based on mutual information. It has advantages over PC-SIS and MV-SIS. It does not require categorizing the continuous variables; categorizing the continuous variables will lose some information especially in small samples case and we do not know how many categories we should use to categorize the continuous variables in advance. Cui et al. [13] thought that the number of categories could be treated as tuning parameter, and they could be determined by cross validation; obviously, the calculation will increase especially in the higher-dimensional case.

Recently, the studies on the independent screening methods are still booming. We just list a few relevant researches; for example, Zhang et al. [15] proposed a Gini correlation screening (GCS) method to select the important variables in ultra-high-dimensional data. Zhou and Zhu [16] proposed a modified martingale difference correlation to improve some drawbacks of martingale difference correlation. Dai et al. [17] proposed a feature selection method based on kernel density estimation for interval-valued data. An et al. [18] proposed a new model for supervised multiclass feature selection which has the $l_{2,1}$−norm in both the fidelity loss and the regularization terms with an additional $l_{2,0}$−constraint. Cuong et al. [19] established fundamental qualitative properties of the minimum sum-of-squares clustering problem and proved that the problem always has a global solution and the global solution set is finite. For more details, we refer to the selective survey by Kamolov [20, 21].

In this paper, we propose a new model-free screening method which has a wider application, the mutual information based screening method; we refer to our method as MI-SIS. Table 1 shows the application scope and algorithm complexity of different feature screening methods. We systematically study the theoretical properties of MI-SIS and establish the sure screening property for the proposed procedure with mild conditions. The new procedure has a good performance in discovering a weak signal in the finite sample. MI-SIS is comparable with MV-SIS and PC-SIS corresponding to the application scope of these methods. In the case that both the predictors and response are categorical variables, the MI index is similar to the PC index proposed by Huang et al. [14]; thus, the efficiency of MI-SIS will be the same as that of PC-SIS in this situation. Moreover, to enhance the performance of the proposed method in finite sample, we conduct three Monte Carlo simulations and conduct a real data analysis. MI-SIS has a good performance in all the simulations.

The rest of this paper is organized as follows. In Section 2, we propose MI-SIS procedure and further study the theoretical properties of the novel approach. In Section 3, we conduct three Monte Carlo simulation studies to examine the performance of MI-SIS in finite samples, especially in very small sample case. We also analyze real data, and the result is very impressive. All technical proofs are given in Appendix A.

2. Independence Screening Using Mutual Information

2.1. Mutual Information. The mutual information between two random variables X and Y is defined in terms of their joint probability distribution $p(X,Y)$ as

$$\mathrm{MI}(X;Y)=\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y. \tag{1}$$

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

The MI marginal measure can be estimated by letting $\widehat{w}=\widehat{\mathrm{M}}\mathrm{I}(X,Y)$. The estimator of the mutual information based on a nonparametric density estimator is illustrated in the following. Let $\{(X_i,Y_i)\colon 1\le i\le n\}$ be a random sample of size n from the population (X,Y). We assume that (X,Y) has continuous joint pdf. Define

$$\begin{aligned}
\widehat{p}(x)&=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right),\\
\widehat{p}(y)&=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{y-Y_i}{h}\right),\\
\widehat{p}(x,y)&=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_1h_2}K\left(\frac{x-X_i}{h_1},\frac{y-X_i}{h_2}\right),\\
\widehat{w}&=\widehat{\mathrm{M}}\mathrm{I}(X,Y)=\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{p}\left(X_i,Y_j\right)\log\frac{\widehat{p}\left(X_i,Y_j\right)}{\widehat{p}(X_i)\widehat{p}\left(Y_j\right)},
\end{aligned} \tag{2}$$

where the function $K(\cdot)$ is kernel function, h and h_i $(i=1,2)$ are bandwidth in nonparametric density estimate, and in practice the kernel functions we often use are Gaussian kernel function and Epanechnikov kernel function. MI marginal measure under other circumstances is given in the following. When X is continuous and Y is categorical,

Table 1: Application scope and algorithm complexity of different feature screening methods.

	Y continuous X continuous	Y continuous X categorical	Y categorical X continuous	Y categorical X categorical	Complexity
MV-SIS		✓	✓		$O(Rn^2p)$
PC-SIS				✓	$O(K^2p)$
MI-SIS	✓	✓	✓	✓	$O(n^3p)$

$$\widehat{w} = \widehat{\mathrm{MI}}(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{K}\widehat{p}(X_i, j)\log\left(\frac{\widehat{p}(X_i, j)}{\widehat{p}(X_i)\,\widehat{p}_j}\right), \tag{3}$$

where $\widehat{p}(X_i, j) = \widehat{p}_j\widehat{p}(X_i|Y = j)$ and $\widehat{p}_j = (\sum_{i=1}^{n} I(Y_i = j))/n$; when X is categorical and Y is also categorical,

$$\widehat{w} = \widehat{\mathrm{MI}}(X,Y) = \sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\widehat{p}(X_k = k_1, Y = k_2)\log \left(\frac{\widehat{p}(X = k_1, Y = k_2)}{\widehat{p}(X = k_1)\widehat{p}(Y = k_2)}\right), \tag{4}$$

where $\widehat{p}(X = k_1, Y = k_2) = (1/n)\sum_{i=1}^{n} I(X_i = k_1)I(Y_i = k_2)$. In this case, MI index is very similar to PC index [14]; when X is categorical and Y is continuous,

$$\widehat{w} = \widehat{\mathrm{MI}}(X,Y) = \sum_{i=1}^{K}\sum_{j=1}^{n}\widehat{p}\left(i, Y_j\right)\log\left(\frac{\widehat{p}\left(i, Y_j\right)}{\widehat{p}_i\widehat{p}\left(Y_j\right)}\right), \tag{5}$$

where $\widehat{p}(i, Y_j) = \widehat{p}_i\widehat{p}(Y_j|X = i)$ and $\widehat{p}_i = (\sum_{j=1}^{n} I(X_j = i))/n$.

2.2. An Independence Ranking and Screening Procedure. We now propose a new model-free sure independence screening using $\mathrm{MI}(X,Y)$ for ultra-high-dimensional data analysis. Let Y be the response with support Ψ_y and Y can be either discrete random variable or continuous random variable, and $X = (X_1, \ldots, X_p)$ denotes the predictor vector, where $p \gg n$ and n is the sample size. Without specifying a regression model, define the active predictors index subset by

$$D = \left\{k\colon\ p(Y|X) \text{ functionally depends on } X_k \text{ for some } y \in \Psi_y\right\}, \tag{6}$$

and define the inactive predictors index subset by

$$I = \left\{k\colon\ p(Y|X) \text{ does not functionally depend on } X_k \text{ for any } y \in \Psi_y\right\}. \tag{7}$$

With the above notation, we can specify the active predictors as $X_D = \{X_k\colon k \in D\}$ and the inactive predictors as $X_I = \{X_k\colon k \in I\}$. Our main purpose is to accurately find the active predictors index subset D.

We now set out to calculate MI of each predictor and the response, $w_k = \mathrm{MI}(X_k, Y)$, $k = 1, \ldots, p$. Note that $w_k = 0$ only if $X_k \in X_I$; this also implies that the predictor X_k is statistically independent of Y; thus we can use MI index as a dependence measure to screen the predictors. The MI-based method is model-free because it involves only marginal density and joint density of the random variables. This index can characterize both linear and nonlinear relationships between the response and predictors.

The primary objective of feature screening in ultra-high-dimensional data analysis is to find a reduced model with a small scale which can contain the true model D with high probability. In this paper, we propose using the index $\widehat{w}_k$ to select a moderate model

$$\widehat{D} = \{k\colon\ \widehat{w}_k \geq cn^{-\tau},\ \text{for}\quad 1 \leq k \leq p\}, \tag{8}$$

where c and τ are predetermined positive values. In practice, we often select the reduced model using another formula: $\widehat{D}^* = \{k\colon\ \widehat{w}_k \text{ is among the top } d \text{ largest of all}\}$. Obviously, $\left\{X_k\colon\ k \in \widehat{D}^*\right\}$ are the most likely relevant predictors with the response. Thus, we can use the predictors in $\left\{X_k\colon\ k \in \widehat{D}^*\right\}$ to estimate the true model. For ease of presentation, we call the above procedure MI-SIS procedure for short.

In the following, we will establish the theoretical properties of the proposed independence screening procedure; Fan and Lv [1] and Ji and Jin [22] demonstrated that the sure screening property guaranteed the effectiveness of the class of independence screening procedure. Therefore, to establish the sure screening property for MI-SIS is essential. The three following conditions are assumed to guarantee that the MI-SIS procedure has sure screening property. They are imposed mainly to facilitate the technical proofs, although they may not be the weakest ones.

($C1$) Suppose that $X = (x_1, \ldots, x_p)$, and x_i come from the distribution F_i which is unknown but has a Lebesgue pdf $f_i, i = 1, \ldots, p$, and some conditions present in Lemma A.3 in the Appendix.

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

$$\sup_{1\leq k\leq p}\sum_{i=1}^{n}\log\frac{p(X_{ik}, Y_i)}{p(X_{ik})p(Y_i)} = O(n^{\kappa}), a.e. \tag{9}$$

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

($C4$) Both X and Y satisfy the subexponential tail probability uniformly in p. That is, there exists a positive constant μ_0 such that, for all $0 < \mu \leq \mu_0$,

$$\sup\nolimits_p\max\nolimits_{1\leq k\leq p} E\left\{\exp\left(\mu\|X_k\|_1^2\right)\right\} < \infty, \text{ and } E\left\{\exp\left(\mu\|Y\|_q^2\right)\right\} < \infty. \tag{10}$$

Theorem 1 (Sure Screening Property). *Under conditions (C1)-(C2), there exists the positive constant C_1 such that*

$$P\left(\max_{1\le k\le p}\left|\widehat{w}_k - w_k\right| \ge cn^{-\tau}\right) \le O(p)\exp\left\{-C_1 n^{1-2\tau}\right\}. \quad (11)$$

Further, we have that

$$P\left(D \subseteq \widehat{D}^*\right) \ge 1 - O\left(s_n \exp\left(-C_1 n^{1-2\tau}\right)\right). \quad (12)$$

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

Theorem 1 indicates that we can deal with the ultra-high-dimensional case with $\log(p) = O(n^{1-2\tau})$, $\tau > 0$.

3. Numerical Studies

In this section, we first assess the finite sample performance of the proposed MI-SIS by Monte Carlo simulation studies. Then, we use real data to analyze the sure screening property of our proposed method. All of our simulation studies were performed in the R language.

Example 1 (X is continuous and Y is categorical). In this example, we simulate a quadratic discriminant analysis problem with ultra-high-dimensional predictors by following the similar idea in Cui et al. [13] or Pan et al. [23]. Our simulation example is slightly different from theirs; we conduct a quadratic discriminant analysis in which the categorical response Y comes from two distributions which have very small difference. We generate Y from a discrete uniform distribution with R categories, where $P(Y_i = r) = 1/R$, with $r = 1, \ldots, R$; given $Y_i = r$, the ith predictor X_i is then generated by letting $X_i = \mu_r + \varepsilon_i + \eta_i$, where $\mu_r = (\mu_{r1}, \ldots, \mu_{rp}) = E(X_i|Y_i = r)$ is the mean of the r categories, and its rth component $\mu_{rr} = \mu$, but its other components are all zeros. μ is relatively a small number; in this example, we conduct $|\mu| \le 0.1$, and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ip})$ are the p–dimensional error terms, and we assume that $\varepsilon_{ij} \sim N(0,1)$, $1 \le j \le p$, and $\eta_i = (0, \ldots, \eta_{rr}, \ldots, 0)$, with only the rth component $\eta_{rr} \sim N(0,3)$; η_{rr} can take other symmetric distributions, such as t-distributions; in this case, we only illustrate the simulation result with $\eta_{rr} \sim N(0,3)$. The shape of the conditional density given Y is shown in Figure 1(a), and that of the marginal density of X_j, $j = 1, 2, 3$, is shown in Figure 1(b).

In order to illustrate the performance of the novel approach, we compare our result with the existing methods PC-SIS [14] and MV-SIS [13]. The MV-SIS method can be directly applied in this situation. But the PC-SIS method only applies to the situation that both the response and the predictors are categorical variables; in this case, we need to categorize the predictors. Cui et al. [13] proposed a specific procedure to categorize continuous variables. The procedure can be described as follows: assuming that X_j is a continuous variable, we define a new vector X_j^* using the percentiles $\{\tau_1, \ldots, \tau_{k_n}\}$ of X_j; let $X_j^* = kI(\tau_k \le X_{ij} \le \tau_{k+1})$, $i = 1, \ldots, n$, $k = 1, \ldots, k_n$, where $I(\cdot)$ denotes an indicator function. Then we immediately face a problem of how many categories we should use in categorizing X_j. In their paper, they suggest $k_n = O(n^{1/5})$. In this simulation, we take $R = 2, n = 50, p = 500$ and $R = 2, n = 100, p = 1000$ separately, and, in each case, we let $\mu = 0.05$ and $k_n = c[n^{1/5}], c = 1, 2, 3$. We repeat each experiment 100 times and define some evaluation index to illustrate their performance. MMMS is short for the median of the minimum model size to include all the active predictors; we also report the robust estimate of its standard deviation RSD (=IQR/1.34, IQR stands for interquartile range) in the parentheses. P_j denotes that an individual active predictor is selected for a given model size $d = [n/\log(n)]$ in the 100 simulations, where $[x]$ denotes the integer part of x. P_a denotes that all active predictors are selected for a given model size d in the 100 simulations. The MMMS is an index to assess the model complexity of an underling procedure. The closer to the true model size it is, the better the screening procedure is. The sure screening property ensures that P_j and P_a are asymptotic to 1 when the estimated model size d is sufficiently large. We report the detailed result in Table 2.

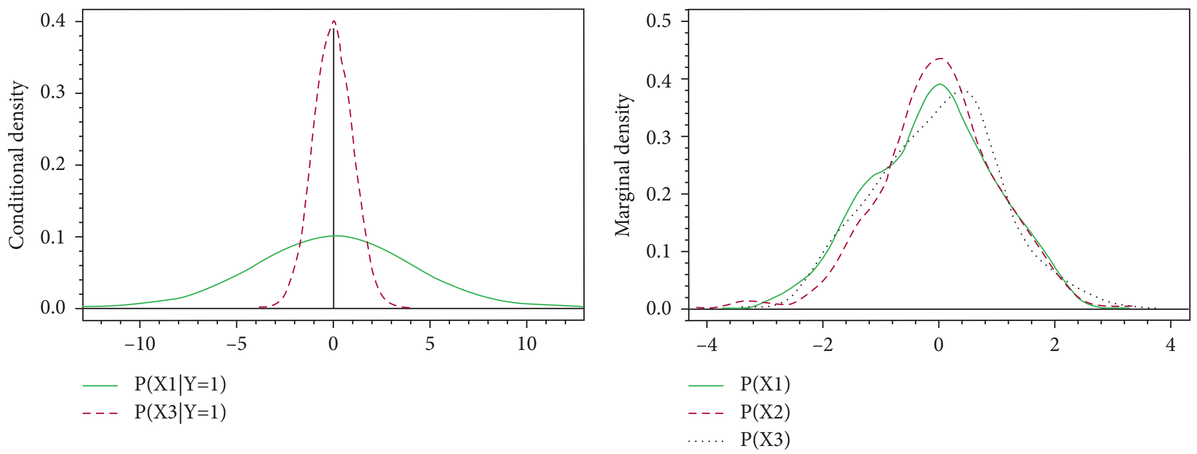
Table 2 implies that the procedure MI-SIS performs reasonably well when the sample size is relatively small; thus, we can conclude that MI-SIS can capture more subtle signals in this situation. Table 2 also shows that the PC-SIS and MV-SIS have a poor performance in this situation, and the results are basically unchanged, despite the fact that we select three k_n values for PC-SIS. With the sample size becoming large, we can find that all the three procedures have a good performance. But the result of PC-SIS is related with the selected k_n; note that k_n cannot be too small.

Example 2 (both X and Y are continuous). This example is designed to compare the performance of MI-SIS with those of PC-SIS and MV-SIS in the case that both the response and the predictors are continuous variables. We consider the two following models:

(2.a) $Y = 5X_1 + 4X_2 + 3X_3 + \varepsilon$

(2.b) $Y = 5X_1X_2 + 3I(X_2 < 0) + 4\sin(X_3) + \varepsilon$

where (2.a) is a linear regression; the relationship between $X_i, i = 1, 2, 3$, and Y is also linear, while in model (2.b) the relationship between $X_i, i = 1, 2, 3$, and Y is nonlinear, and $I(X_2 < 0)$ is an indicator function. In addition, model (2.b) contains a sine function of X_3 and an interaction term X_1X_2. All of the three methods are model-free; thus they can be directly applied in these two models. But we need to categorize the continuous response for MV-SIS and categorize both the continuous response and the predictors for PC-SIS. In this example, we set $k_n = 4$ when we categorize the continuous variables. Let $(n, p) = (100, 500)$ and $\varepsilon_i \sim N(0,1)$, and, for each model, we consider two scenarios to assess the performances of the three methods: $\text{cov}(X) = I$ and $\text{cov}(X) = \sum$, where $\sum = (0.5^{|i-j|})$. Moreover, in order to show defects of categorizing, we use formula (3) to calculate MI; we also need to categorize the continuous response in this case, and we name this method MI-SIS2 in Table 3. Table 3 presents the simulation result for P_j, $j = 1, 2, 3$, and P_a. The performances of the MV-SIS, PC-SIS, MI-SIS, and MI-SIS2 are very similar in model (2.a). But MI-SIS outperforms the other procedures in model (2.b). We can



(a)

(b)

Figure 1: (a) The conditional density of X_1, $P(X_1|Y=1)$, and the conditional density of X_2, $P(X_2|Y=2)$; note that $P(X_1|Y=1)$ is identical to $P(X_2|Y=2)$, so we only draw a line. (b) The marginal density of all the predictors. Under our assumption, the marginal densities of X_j are nearly all the same except a small mean difference.

Table 2: Simulation result for Example 1.

		Method	Categories	P_1	P_2	P_a	MMMS
$P = 500$	$n = 50$ $d = 12$	PC-SIS	$c = 1$	0.10	0.07	0.01	265.0 (137.3)
			$c = 2$	0.64	0.69	0.45	45.5 (17.9)
			$c = 3$	0.75	0.79	0.59	14.0 (8.2)
		MV-SIS	—	0.35	0.39	0.18	73.0 (85.8)
		MI-SIS	—	0.98	0.95	0.94	2.0 (0.7)
	$n = 100$ $d = 21$	PC-SIS	$c = 1$	0.10	0.06	0.01	304.5 (126.3)
			$c = 2$	0.97	0.96	0.93	2.0 (0.7)
			$c = 3$	1.00	1.00	1.00	2.0 (0.0)
		MV-SIS	—	0.90	0.92	0.83	4.0 (1.5)
		MI-SIS	—	1.00	1.00	1.00	2.0 (0.0)
$P = 1000$	$n = 50$ $d = 12$	PC-SIS	$c = 1$	0.06	0.05	0.01	393.0 (194.0)
			$c = 2$	0.55	0.59	0.30	127.0 (126.9)
			$c = 3$	0.64	0.64	0.40	99.0 (22.4)
		MV-SIS	—	0.16	0.15	0.02	285.5 (96.3)
		MI-SIS	—	0.92	0.91	0.84	7.0 (4.4)
	$n = 100$ $d = 21$	PC-SIS	$c = 1$	0.06	0.08	0.00	453.0 (97.0)
			$c = 2$	0.96	0.95	0.92	2.0 (0.7)
			$c = 3$	1.00	0.98	0.98	2.0 (0.0)
		MV-SIS	—	0.84	0.82	0.72	9.5 (4.5)
		MI-SIS	—	1.00	1.00	1.00	2.0 (0.0)

The user-specified number $k_n = c[n^{1/5}]$.

Table 3: The proportions of P_j and P_a in Example 2.

Model	Parameter	cov (X)	Method	P_1	P_2	P_3	P_a
(a)	$p = 500$ $n = 100$ $d = 21$ $k = 4$	$\sum = I$	PC-SIS	1.00	0.96	0.69	0.65
			MV-SIS	1.00	1.00	0.87	0.87
			MI-SIS2	1.00	0.88	0.73	0.55
			MI-SIS	1.00	0.98	0.58	0.57
		$\sum = 0.5^{\lvert i-j\rvert}$	PC-SIS	1.00	1.00	0.98	0.98
			MV-SIS	1.00	1.00	1.00	1.00
			MI-SIS2	1.00	1.00	0.96	0.96
			MI-SIS	1.00	1.00	1.00	1.00

Table 3: Continued.

Model	Parameter	cov (X)	Method	P_1	P_2	P_3	P_a
(b)	$p = 500$ $n = 100$ $d = 21$ $k = 4$	$\sum = I$	PC-SIS	0.35	0.59	0.99	0.20
			MV-SIS	0.39	0.56	1.00	0.28
			MI-SIS2	0.46	0.64	0.99	0.29
			MI-SIS	0.99	1.00	0.94	0.93
		$\sum = 0.5^{\lvert i-j\rvert}$	PC-SIS	0.83	0.80	0.91	0.57
			MV-SIS	0.42	0.48	0.98	0.25
			MI-SIS2	0.80	0.82	0.92	0.58
			MI-SIS	1.00	1.00	0.91	0.91

conclude that MI-SIS has a better performance than other procedures in the case where both the response and the predictors are continuous variables.

Example 3 (both X and Y are categorical). In this case, our proposed index reduces to the form in (4); in order to assess the effect of our index with PC-SIS, we borrow an example from Huang et al. [14]; in their paper, the following example was used to assess the effect of PC-SIS in the case of the predictors without interaction. The category response $Y_i \in \{1, 2, 3, 4\}$, and $P(Y_i = k) = 1/4$, for every $1 \leq k \leq 4$. Define the true model as $S_T = \{1, 2, \ldots, 10\}$ with $|S_T| = 10$. Next, conditional on Y_i, we generate the predictor as $P(X_{ij} = 1|Y_i = k) = \theta_{kj}$ for every $1 \leq k \leq 4$ and $j \in S_T$. The specific values of θ_{kj} are shown in Table 4. For $j \notin S_T$, let $\theta_{kj} = 0.5$, for every $1 \leq k \leq 4$.

Under this mechanism, we infer that the predictors $X_j, j > 10$, are independent of the response, because $P(X_{ij} = 1|Y_i = k, k \notin S_T) = \theta_{kj} = 0.5$, and $P(Y_i = k) = 1/4$; we can calculate $P(X_{ij} = 1) = \sum P(X_{ij} = 1|Y_i = k)P(Y_i = k) = 0.5$; the conditional mass functions are identical to the unconditional mass functions, so $X_j, j > 10$, and Y are independent. In this example, in order to systematically study the gradual equivalence about MI-SIS and PC-SIS proposed by Huang et al. (2014), we designed four cases and performed simulation 100 times for each case. The gradual equivalence is a very strong property which needs us to subsequently define several evaluation indexes to assess. $P_a - \mathrm{MI}$ denotes the proportion that all active predictors are included in $\widehat{D}^*_{\mathrm{MI}}$ for a given model size $d = [n/\log(n)]$, while $P_a - \mathrm{PC}$ denotes the proportion that all active predictors are included in $\widehat{D}^*_{\mathrm{PC}}$ at the same model size. More specifically, we let $\mathrm{PCS} = \sum I(\widehat{D}^*_{\mathrm{MI}} \cap D = \widehat{D}^*_{\mathrm{PC}} \cap D)/n$ denote the proportion of the identical correct predictors selected by the methods of PC-SIS and MI-SIS, while we let $\mathrm{PFS} = \sum I(\widehat{D}^*_{MI} \cap D^c = \widehat{D}^*_{\mathrm{PC}} \cap D^c)/n$ denote the proportion of the identical false selected predictors.

The detailed results are shown in Table 5; the four simulation results are very similar; this result is not out of our expectation. Both the MI-SIS and the PC-SIS methods are relatively very efficient in selecting the active predictors; the values of MMMS (RSD) and $P_a - \mathrm{MI}$ and $P_a - \mathrm{PC}$ are very close to 1. The value of PCS is approximately equal to 1 as n goes to infinity; this result demonstrates that MI-SIS and the PC-SIS have gradual equivalence. Both can almost surely select the active predictors, even if they do not select all the true predictors. However, PFS has a relatively poor performance; the reason may be that when n goes to infinity, the cardinality of $\widehat{D}^*$ will also be very large; due to the randomness, the probability of the false selected predictors will vary a lot.

Table 4: Probability specification for Example 3.

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

Example 4 (X is categorical and Y is continuous). In this example, we use the same setting as Example 2 but add some new categorical variables. The models we considered are as follows:

(2.a) $Y = 5X_1 + 4X_2 + 3X_3 + 2D_1 + 3D_2 + \varepsilon$

(2.b)
$Y = 5X_1X_2 + 3I(X_2 < 0) + 4\sin(X_3) + 2D_1 + 3D_2 + \varepsilon$

where $X_j, j = 1, 2, 3$, are continuous variables and $D_j, j = 1, 2$, are categorical variables. We consider the following scenarios with $n = 100, p = 500$. Firstly, we generate p_1 continuous variable with $\operatorname{cov}(X) = I_{p_1}$ and $\operatorname{cov}(X) = \Sigma_{p_1 \times p_1}$, where $\Sigma_{p_1 \times p_1} = (0.5^{|i-j|})_{p_1 \times p_1}$. Then, we generate p_2 categorical variables with $P(D_{ij} = 1) = \theta_j$ and θ_j follows the uniform distribution on 0 and 1. Note that $p_1 + p_2 = p$ and $p_1/p_2 = 2/3$, 1, 0, 0. The simulation result is presented in Table 6 which shows that the MI-SIS method also outperforms its competitors when X is categorical and Y is continuous.

4. Real Data Analysis

In this section, we illustrate the proposed MI-SIS procedure with an application to detect the important features about the voice of Parkinson's patients using LSVT dataset. The dataset is available at http://archive.ics.uci.edu/ml/datasets/LSVT+Voice+Rehabilitation. The LSVT (Lee Silverman voice treatment) Voice Rehabilitation Dataset comes from UCI website; this dataset was created by Tsanas et al., and it was first analyzed by [24]. The goal of the study is to improve the effectiveness of rehabilitative speech treatment by appropriate statistical algorithms for LSVT Companion

Table 5: Detailed simulation results for Example 3.

p	n	MMMS (RSD)	P_a – MI	P_a – PC	PCS	PFS
1000	100	10.0 (0.0)	0.96	0.96	1.00	0.62
	200	10.0 (0.0)	1.00	1.00	1.00	0.67
5000	100	10.0 (0.74)	0.86	0.86	0.96	0.31
	200	10.0 (0.0)	1.00	1.00	1.00	0.43

Table 6: The proportions of P_j and P_a in Example 4.

Model	Parameter	cov (X)	Method	P_1	P_2	P_3	P_a
(a)	$p = 500$ $n = 100$ $d = 21$ $k = 4$	$\sum = I$	PC-SIS	1.00	0.95	0.68	0.64
			MV-SIS	1.00	0.99	0.86	0.86
			MI-SIS2	1.00	0.89	0.74	0.56
			MI-SIS	1.00	0.97	0.59	0.58
		$\sum = 0.5^{\vert i-j\vert}$	PC-SIS	1.00	1.00	0.98	0.98
			MV-SIS	1.00	1.00	1.00	1.00
			MI-SIS2	1.00	1.00	0.96	0.96
			MI-SIS	1.00	1.00	1.00	1.00
(b)	$p = 500$ $n = 100$ $d = 21$ $k = 4$	$\sum = I$	PC-SIS	0.34	0.57	0.97	0.20
			MV-SIS	0.38	0.54	1.00	0.26
			MI-SIS2	0.46	0.65	0.98	0.28
			MI-SIS	0.98	1.00	0.95	0.92
		$\sum = 0.5^{\vert i-j\vert}$	PC-SIS	0.84	0.81	0.90	0.56
			MV-SIS	0.41	0.46	0.97	0.24
			MI-SIS2	0.81	0.83	0.91	0.57
			MI-SIS	1.00	1.00	0.90	0.90

system. These algorithms can automatically be used for detecting whether the characteristics of the voice are acceptable or not. An efficient algorithm can automatically assess the effectiveness of the LSVT Companion system during use of software away from expert clinical guidance.

This dataset contains 126 samples from 14 participants, 309 predictors, and one response. The predictors are the features of the voice. The response is a binary variable; one means "unacceptable" and zero means "acceptable." "Unacceptable" means that a clinician thought the voice was not persisting during in-person rehabilitation treatment. More details about the predictors can be found in [24]. Therein, the authors demonstrated that the algorithm they proposed could correctly replicate the experts' binary assessment with approximately 90% accuracy.

Here, we use two penalized regression models to analyze this dataset. We first established a penalized logistic model using all the predictors by minimizing the penalized likelihood

$$\begin{aligned}
\beta &= \arg\min_{\beta} l(\beta) = -L(\beta) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|, \\
L(\beta) &= \sum_{i=1}^{n} y_i \log\left(p_i\right) + \left(1 - y_i\right) \log\left(1 - p_i\right), \qquad (13) \\
p_i &= \frac{e^{\beta_0 + x_i^T \beta}}{1 + e^{\beta_0 + x_i^T \beta}}.
\end{aligned}$$

For the second model, we propose first applying MI-SIS to screening $d = 2[n/\log(n)] = 30$ predictors; thereafter, we extend the predictor variable space by adding the interaction terms of the screened predictors. Then we apply the penalized logistic model to the new feature space, and in this way, we can explore the nonlinear relationship of the screened predictors. In the paper, these two models will be referred to as penalized-logistic model and MI-SIS-penalized-logistic model, respectively, for simplicity.

The top-right and bottom-right figures in Figure 2 show the penalized-logistic coefficient paths about the two models. The top-left and bottom-left figures are the CV error for each λ; the hyperparameter λ is selected by 5-fold cross validation, and the best λ will be obtained at the minimum of the CV error curve. We summarize the classification result in Table 7. From the confusion matrix, we conclude that both the penalized-logistic model and MI-SIS-penalized-logistic model have a better performance. For penalized-logistic model, it finally selects only 13 predictors, with the best $\lambda = e^{-3.4}$, and its classification rate is 91.27%; for MI-SIS-penalized-logistic model, it finally selects 25 predictors on the new feature space, with the best $\lambda = e^{-4.4}$, and its classification rate is 96.83%. This example further demonstrates that the MI-SIS with a penalized-logistic model is more enjoyable in real data analysis.

5. Discussion

In this paper, we proposed a new independent feature screening method based on the mutual information, that is, MI-SIS. The proposed procedure is model-free, and the sure screening property was established when the number of the predictors diverges with an exponential rate of the samples. The new procedure has a good performance in discovering a

Table 7: The confusion matrix for classification problem in Example 4.

	MI-SIS-penalized-logistic model		Penalized-logistic model	
Y	0	1	0	1
0	39	1	35	4
1	3	83	7	80

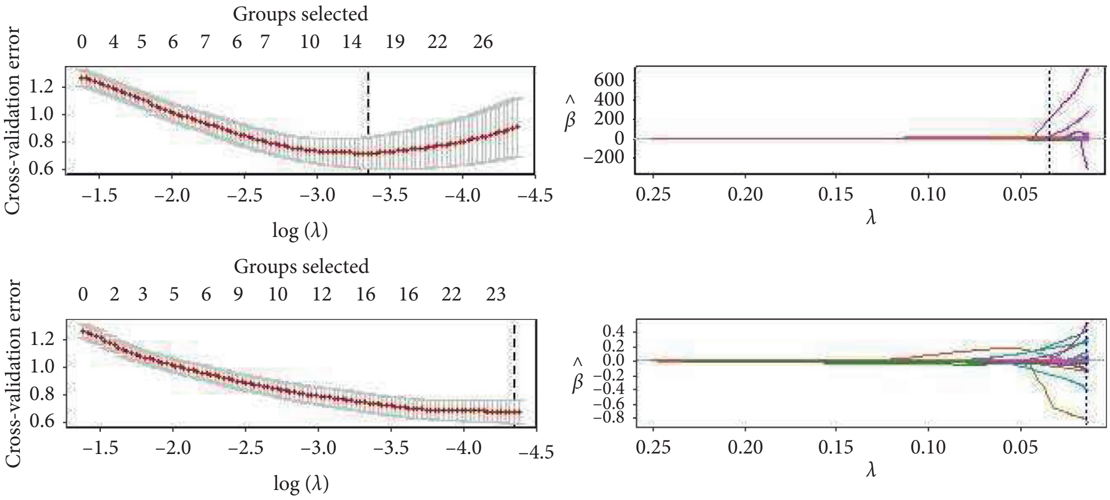


Figure 2: The top-right and bottom-right figures are estimated coefficient in penalized-logistic models. The top-left and bottom-left figures are the CV error during the selection procedure of hyperparameter λ.

weak signal in the finite sample. Similar to Fan and Lv [1], we select a cutoff d for MI-SIS. How to choose d is a very important and tough problem. How to choose d for MI-SIS more reasonably is a good topic desiring more discussion and reasonable d plays an important role in feature screening methods.

Theoretically, we ignore the marginal dependence between the predictors; marginal dependent problem may be a trouble during the feature screening procedure. How to deal with the marginal dependence between the predictors remains a question. Similar to the ISIS, we adopt the idea to MI-SIS and develop an iterative procedure about MI-SIS. Due to the defect of nonparametric estimation in small sample, the iterative MI-SIS performs poorly, so we do not post the result in this paper. To overcome the marginal dependence problem, more novel research needs to be developed.

Appendix

A. Proof of Theoretical Result

In order to prove Theorem 1, we need the three following lemmas. The first two lemmas provide us two exponential inequalities, and their proofs can be found in [25].

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

$$E[\exp\{s(Y-\mu)\}] \le \exp\left\{\frac{s^2(b-a)^2}{8}\right\}. \tag{A.1}$$

Lemma A.2. *Let $h(Y_1,\ldots,Y_m)$ be a kernel of the U statistics U_n, and $\theta = E\{h(Y_1,\ldots,Y_m)\}$. If $a \le h(Y_1,\ldots,Y_m) \le b$, then, for any $t>0$ and $n \ge m$,*

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

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

Lemma A.2 is the unilateral tail inequality of U_n; we can easily get the bilateral tail inequality of U_n due to its symmetry.

$$P\left(\left|U_n-\theta\right| \ge t\right) \le 2\exp\left(\frac{-2[n/m]t^2}{(b-a)^2}\right). \tag{A.3}$$

Lemma A.3. *(the asymptotic property of nonparametric density estimators). Suppose that $f''(x)$ exists and $h = cn^{-(1/5)}$; then*

$$n^{2/5}\{\widehat{p}(x) - p(x)\} \xrightarrow{L} N\left(\frac{c^2}{2}f''(x)\mu_2(K), \frac{1}{c}f(x)\|K\|_2^2\right). \tag{A.4}$$

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

Lemma A.3 directly implies that $\widehat{p}(x) \xrightarrow{p} p(x)$. Under some more strict conditions, we have the strong uniform convergence of $\widehat{p}(x)$.

$$\lim_{n\longrightarrow\infty} \sup_x |\widehat{p}(x) - p(x)| = 0, a.e. \tag{A.5}$$

More details about the strong uniform convergence can be found, for example, in [26] or [27].

Proof of Theorem 1. First, we show for each k that the following inequality holds:

$$P\left\{\left|\widehat{w}_k - w_k\right| \geq cn^{-\tau}\right\} \leq O\left(\exp\left(-C_1 n^{1-2\tau}\right)\right). \qquad (A.6)$$

This is because

$$\begin{aligned}\left|\widehat{w}_k - w_k\right| &= \left|\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{p}\left(X_{ik}, Y_j\right)\log\frac{\widehat{p}\left(X_{ik}, Y_j\right)}{\widehat{p}\left(X_{ik}\right)\widehat{p}\left(Y_j\right)} - \int p\left(x_k, y\right)\log\frac{p\left(x_k, y\right)}{p\left(x_k\right)p\left(y\right)}\mathrm{d}x_k\mathrm{d}y\right| \\ &= \Big|\sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{p}\left(X_{ik}, Y_j\right)\log\frac{\widehat{p}\left(X_{ik}, Y_j\right)}{\widehat{p}\left(X_{ik}\right)\widehat{p}\left(Y_j\right)} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\log\frac{\widehat{p}\left(X_{ik}, Y_j\right)}{\widehat{p}\left(X_{ik}\right)\widehat{p}\left(Y_j\right)} \\ &\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\log\frac{\widehat{p}\left(X_{ik}, Y_j\right)}{\widehat{p}\left(X_{ik}\right)\widehat{p}\left(Y_j\right)} - \int p\left(x_k, y\right)\log\frac{p\left(x_k, y\right)}{p\left(x_k\right)p\left(y\right)}\mathrm{d}x_k\mathrm{d}y\Big| \\ &= \left|M_{k,1} + M_{k,2}\right|.\end{aligned} \qquad (A.7)$$

By Lemma A.3 and strong law of large numbers, it shows the convergence of

$$\begin{aligned}M_{k,1} &= \sum_{i=1}^{n}\sum_{j=1}^{n}\widehat{p}\left(X_{ik}, Y_j\right)\log\frac{\widehat{p}\left(X_{ik}, Y_j\right)}{\widehat{p}\left(X_{ik}\right)\widehat{p}\left(Y_j\right)} \\ &\quad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\log\frac{\widehat{p}\left(X_{ik}, Y_j\right)}{\widehat{p}\left(X_{ik}\right)\widehat{p}\left(Y_j\right)} \longrightarrow 0, a.e.\end{aligned} \qquad (A.8)$$

Next, we will establish the bound of the second term.

Define $h(X_{ik}, Y_j; X_k, Y) = \log((\widehat{p}(X_{ik}, Y_j))/(\widehat{p}(X_{ik})\widehat{p}(Y_j)))$ as the kernel of the U statistics of $I_{k,2}^{*}$, where we define $M_{k,2} = I_{k,2} - w_k$, and $I_{k,2}^{*} = I_{k,2} - (1/n^2)\sum_{i=j}\log((\widehat{p}(X_{ik}, Y_i))/(\widehat{p}(X_{ik})\widehat{p}(Y_i)))$. With Markov's inequality, we can ensure that

$$\begin{aligned}P\left(I_{k,2}^{*} - \mathrm{Eh} > \varepsilon\right) \leq\ &\exp(-t\varepsilon)\exp \\ &(-t\mathrm{Eh})E\left\{\exp\left(tI_{k,2}^{*}\right)\right\},\ \text{for any}\ t > 0,\end{aligned} \qquad (A.9)$$

where $\mathrm{Eh} = \int\log((\widehat{p}(X_{ik}, Y_j))/(\widehat{p}(X_{ik})\widehat{p}(Y_j)))p(x, y)\,\mathrm{d}x\mathrm{d}y$. As Li et al. (2012) used the technique to deal with the U statistics in their paper, together with condition $(C2)$, it is entailed immediately that

$$P\left(I_{k,2}^{*} - Eh > \varepsilon\right) \leq\ \exp\left(\frac{-t\varepsilon + t^2}{8n}\right). \qquad (A.10)$$

By choosing $t = 4n\varepsilon$, we have $P(I_{k,2}^{*} - \mathrm{Eh} > \varepsilon) \leq \exp(-2n\varepsilon^2)$; therefore, due to the symmetry of U statistics, we can obtain the bilateral tail inequality

$$P\left(\left|I_{k,2}^{*} - \mathrm{Eh}\right| > \varepsilon\right) \leq 2\exp\left(-2n\varepsilon^2\right). \qquad (A.11)$$

Using the relationship between $I_{k,2}^{*}$ and $I_{k,2}$, we can show that

$$\begin{aligned}P\left(\left|M_{k,2}\right| > 2\varepsilon\right) &= P\left(\left|I_{k,2} - w_k\right| > 2\varepsilon\right) \\ &= P\left(\left|I_{k,2}^{*} + \frac{1}{n^2}\sum_{i=j}\log\frac{\widehat{p}\left(X_{ik}, Y_i\right)}{\widehat{p}\left(X_{ik}\right)\widehat{p}\left(Y_i\right)} - w_k\right| > 2\varepsilon\right).\end{aligned} \qquad (A.12)$$

Under condition $(C2)$, for $\varepsilon > 0$, we can take a large N_1; when $n > N_1$, $(1/n^2)\sum_{i=j}\log((\widehat{p}(X_{ik}, Y_i))/(\widehat{p}(X_{ik})\widehat{p}(Y_i))) < (\varepsilon/3)$; we can easily prove that

$$P\left(\left|I_{k,2} - w_k\right| > 2\varepsilon\right) \leq P\left(\left|I_{k,2}^{*} - w_k\right| > \frac{5}{3}\varepsilon\right). \qquad (A.13)$$

Note that

$$\left|I_{k,2} - w_k\right| = \left|I_{k,2} - \mathrm{Eh} + \mathrm{Eh} - w_k\right|. \qquad (A.14)$$

Similarly, we can use the above skill and take larger N_2; when $n > N_2|w_k - \mathrm{Eh}| < (\varepsilon/3)$, this can directly show that

$$P\left(\left|I_{k,2}^{*} - w_k\right| > \frac{5}{3}\varepsilon\right) \leq P\left(\left|I_{k,2}^{*} - \mathrm{Eh}\right| > \frac{4}{3}\varepsilon\right). \qquad (A.15)$$

Let $\varepsilon = cn^{-\tau}$, where $0 < \tau < 1/2$; from inequality (A.11), together with property (A.1), and Bonferroni's inequality, it is implied that

$$P\left\{\left|\widehat{w}_k - w_k\right| \geq cn^{-\tau}\right\} \leq 2\ \exp\left(-2c^2 n^{1-2\tau}\right) \qquad (A.16)$$

We thus have

$$\begin{aligned}P\left\{\max_{1\leq k\leq p}\left|\widehat{w}_k - w_k\right| \geq cn^{-\tau}\right\} &\leq 2p\exp\left(-2c^2 n^{1-2\tau}\right) \\ &= O\left(p\left[\exp\left(-C_1 n^{1-2\tau}\right)\right]\right).\end{aligned} \qquad (A.17)$$

Then, we proof the second part of Theorem 1.

If $D \subset \widehat{D}^{*}$, this implies that there must exist some $k \in D$ satisfying the fact that $\widehat{w}_k < cn^{-\tau}$. Using condition $(C3)$,

$\widehat{w}_k < cn^{-\tau}$ implies that $|\widehat{w}_k - w_k| > cn^{-\tau}$ for some $k \in D$. Thus, the event $\left\{D \subset \widehat{D}^*\right\} \subseteq \{|\widehat{w}_k - w_k| > cn^{-\tau}$, for some $k \in D\}$; we take complement on both sides, and we get $\left\{\max_{k\in D} |\widehat{w}_k - w_k| \le cn^{-\tau}\right\} \subseteq \left\{D \subseteq \widehat{D}^*\right\}$. Therefore,

$$\begin{aligned}
P\left(D \subseteq \widehat{D}^*\right) &\ge P\left\{\max_{k\in D}\left|\widehat{w}_k - w_k\right| \le cn^{-\tau}\right\} \\
&= 1 - P\left\{\min_{k\in D}\left|\widehat{w}_k - w_k\right| \ge cn^{-\tau}\right\} \\
&= 1 - s_n P\left\{\left|\widehat{w}_k - w_k\right| \ge cn^{-\tau}\right\} \\
&\ge 1 - O\left(s_n\left[\exp\left(-C_1 n^{1-2\tau}\right)\right]\right).
\end{aligned} \tag{A.18}$$

In the above equation, s_n is the cardinality of D. This is the end of the proof. □

Data Availability

The LSVT (Lee Silverman voice treatment) Voice Rehabilitation Dataset was adopted to illustrate the proposed MI-SIS procedure in Section 4. The dataset is available at UCI website: http://archive.ics.uci.edu/ml/datasets/LSVT+Voice+Rehabilitation.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was funded by Doctoral Program of Harbin Normal University, China (grant no. XKB201805).

References

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

[2] H. Wang, “Forward regression for ultra-high dimensional variable screening,” *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1512–1524, 2009.

[3] J. Fan, R. Samworth, and Y. Wu, “Ultrahigh dimensional feature selection: beyond the linear model,” *Journal of Machine Learning Research: JMLR*, vol. 10, pp. 2013–2038, 2009.

[4] J. Fan and R. Song, “Sure independent screening in generalized linear models with NP dimensionality,” *Annals of Statistics*, vol. 38, no. 6, pp. 3567–3604, 2010.

[5] J. Fan, Y. Feng, and R. Song, “Nonparametric independence screening in sparse ultra-high-dimensional additive models,” *Journal of the American Statistical Association*, vol. 106, no. 494, pp. 544–557, 2011.

[6] J. Liu, R. Li, and R. Wu, “Feature selection for varying coefficient models with ultrahigh-dimensional covariates,” *Journal of the American Statistical Association*, vol. 109, no. 505, pp. 266–274, 2014.

[7] J. Fan, Y. Ma, and W. Dai, “Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models,” *Journal of the American Statistical Association*, vol. 109, no. 507, pp. 1270–1284, 2014.

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

[9] R. Li, W. Zhong, and L. Zhu, “Feature screening via distance correlation learning,” *Journal of the American Statistical Association*, vol. 107, no. 499, pp. 1129–1139, 2012.

[10] X. He, L. Wang, and H. Hong, “Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data,” *Annals of Statistics*, vol. 41, no. 1, pp. 342–369, 2013.

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

[12] X. Shao and J. Zhang, “Martingale difference correlation and its use in high-dimensional variable screening,” *Journal of the American Statistical Association*, vol. 109, no. 507, pp. 1302–1318, 2014.

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

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

[15] J.-y. Zhang, X.-f. Liu, R.-q. Zhang, and H. Wang, “Gini correlation for feature screening,” *Acta Mathematicae Applicatae Sinica, English Series*, vol. 37, no. 3, pp. 590–601, 2021.

[16] J. Zhou and L. Zhu, “Modified martingale difference correlations,” *Journal of Nonparametric Statistics*, vol. 33, no. 2, pp. 359–386, 2021.

[17] J. Dai, Y. Liu, J. Chen, and X. Liu, “Fast feature selection for interval-valued data through kernel density estimation entropy,” *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 12, pp. 2607–2624, 2020.

[18] N. T. An, P. D. Dong, and X. Qin, “Robust feature selection via nonconvex sparsity-based methods,” *Journal of Nonlinear Variable Analysis*, vol. 5, no. 1, pp. 59–77, 2021.

[19] T. H. Cuong, J.-C. Yao, and N. D. Yen, “Qualitative properties of the minimum sum-of-squares clustering problem,” *Optimization*, vol. 69, no. 9, pp. 2131–2154, 2020.

[20] S. Kamolov, “Feature selection: state-of-the-art survey,” *Annals of Mathematics and Computer Science*, vol. 4, pp. 48–54, 2021.

[21] J. Y. Liu, W. Zhong, and R. Z. Li, “A selective overview of feature screening for ultrahigh-dimensional data,” *Science China Mathematics*, vol. 58, no. 10, pp. 2033–2054, 2015.

[22] P. Ji and J. Jin, “UPS delivers optimal phase diagram in high-dimensional variable selection,” *Annals of Statistics*, vol. 40, no. 1, pp. 73–103, 2012.

[23] R. Pan, H. Wang, and R. Li, “Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening,” *Journal of the American Statistical Association*, vol. 111, no. 513, pp. 169–179, 2016.

[24] A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig, “Objective automatic assessment of rehabilitative speech treatment in Parkinson’s disease,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. 181–190, 2014.

[25] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*, John Wily and Sons Inc, New York, NY, USA, 1980.

[26] R. Dias, N. L. Garcia, and A. Z. Zambom, “Monte Carlo algorithm for trajectory optimization based on Markovian readings,” *Computational Optimization and Applications*, vol. 51, no. 1, pp. 305–321, 2012.

[27] P. J. Bickel and M. Rosenblatt, “On some global measures of the deviations of density function estimates,” *Annals of Statistics*, vol. 1, no. 6, pp. 1071–1095, 1973.