Scientific Research Publishing

Open Journal of Statistics, 2024, 14, 413-438
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Nonparametric Feature Screening via the Variance of the Regression Function

Won Chul Song[1], Michael G. Akritas[2]

[1]Department of Mathematics, Milwaukee School of Engineering, Milwaukee, USA
[2]Department of Statistics, Pennsylvania State University, University Park, USA
Email: song@msoe.edu

How to cite this paper: Song, W. and Akritas, M.G. (2024) Nonparametric Feature Screening via the Variance of the Regression Function. *Open Journal of Statistics*, **14**, 413-438.
https://doi.org/10.4236/ojs.2024.144017

Received: July 9, 2024
Accepted: August 23, 2024
Published: August 26, 2024

Abstract

This article develops a procedure for screening variables, in ultra high-dimensional settings, based on their predictive significance. This is achieved by ranking the variables according to the variance of their respective marginal regression functions (RV-SIS). We show that, under some mild technical conditions, the RV-SIS possesses a sure screening property, which is defined by Fan and Lv (2008). Numerical comparisons suggest that RV-SIS has competitive performance compared to other screening procedures, and outperforms them in many different model settings.

Keywords

Sure Independence Screening, Nonparametric Regression, Ultrahigh-Dimensional Data, Variable Selection

1. Introduction

With advances in the data collection technology, ultrahigh-dimensional data can be easily collected in many research areas such as genetic data, microarray data, and high volumn financial data. In these examples, the number of predictors (p) is an exponential function of the the number of the observations (n). In other words, $\log p = O\left(n^a\right)$ for some $a > 0$. The sparsity assumption, that only a small set of covariates has an effect on the response, makes the inference possible in ultrahigh-dimensional data.

The popular variable selection methods may suffer technical difficulties and performance issues when analyzing ultrahigh-dimensional data due to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability [1]. Motivated by this, Fan and Lv [2] recommended that a variable screening procedure be performed prior to variable selection. Working

with a linear model, they introduced sure independence screening (SIS), a variable screening procedure based on Pearson's correlation coefficient. Assuming Gaussian predictors and response variable, they showed that SIS possesses the sure screening property, which means that the true predictors will be chosen with probability one as the sample size approaches to infinity. Since then, several feature screening methods based on SIS have been developed. Fan, Feng and Song [3] introduced a nonparametric screening procedure (NIS), which uses a spline-based nonparametric estimation of the marginal regression functions, and ranks predictors by the Euclidean norm of the estimated marginal regression function (evaluated at the data points). Li, Zhong and Zhu [4] proposed a ranking procedure using the distance correlation (DC-SIS). DC-SIS can be used for grouped predictors and multivariate responses. Li *et al.* [5] propose a robust rank correlation screening (RRCS), which uses a ranking based on Kendall's τ rank correlation coefficient. They show that this procedure can handle semiparametric models under monotonic constraint to the link function. This procedure can be also used when there exists outliers, influence points, or heavy tailed errors. Wang and Deng [6] introduced a model-free feature screening to handle multi-classification problems with both categorical and continuous covariates using Gini impurity to evaluate the predictive power of covariates. Chen and Deng [7] proposed another model-free feature screening for multi-classification using the Maximal Information Coefficient to evaluate the predictive power of the variables.

Variables that are relevant for prediction are of particular interest in most scientific research and its applications. The aforementioned feature screening methods fail to distinguish variables that have predictive significance from those that influence the variance function or other aspects of the conditional distribution of the response. We propose a method that screens out variables without (marginal) predictive significance. The basic idea is that if a variable X_i has no predictive significance, the regression function $E(Y \mid X_i)$ has zero variance. This leads to a method which ranks the predictors according to the sample variance (evaluated at the data points) of the p estimated regression functions, called RV-SIS for *regression variance sure independence screening*. We show that RV-SIS possesses the sure independence screening property under a general nonparametric regression setting. While the proofs use Nadaraya-Watson estimators for the marginal regression functions, the proofs also hold (with mild modifications) for local linear estimators.

We conduct numerical simulation studies to compare the RV-SIS to SIS, DC-SIS, RRCS and NIS. The RV-SIS outperforms SIS, DC-SIS, RRCS and NIS in many different model settings. The RV-SIS procedure shows that it takes less computing time than both DC-SIS and NIS.

We conduct numerical simulation studies to compare the RV-SIS to SIS, DC-SIS, RRCS and NIS. The RV-SIS outperforms SIS, DC-SIS, RRCS and NIS in many different model settings. The RV-SIS procedure shows that it takes less

computing time than both DC-SIS and NIS.

2. Nonparametric Independence Screening via the Variance of the Regression Function

2.1. Preliminaries

Consider a random sample $(\boldsymbol{X}_1, Y_1), \cdots, (\boldsymbol{X}_n, Y_n)$ of iid $(p+1)$-dimensional random vectors, where Y_i is univariate and $\boldsymbol{X}_i = (X_{1i}, \cdots, X_{pi})^{\mathrm{T}}$ is a p-dimensional, $i = 1, \cdots, n$. Let $m(\boldsymbol{X}) = E(Y \mid \boldsymbol{X})$ and write

$$Y = m(\boldsymbol{X}) + \varepsilon, \tag{1}$$

where $\varepsilon = Y - m(\boldsymbol{X})$. For $k = 1, \cdots, p$, we consider p marginal nonparametric regression functions

$$m_k(x) = \mathrm{E}(Y_i \mid X_{ki} = x) \tag{2}$$

of Y on each variable X_k, and define the set of active and inactive predictors by

$$\mathcal{D} = \{k : m_k(x) \text{ is not a constant function}\}, \quad \mathcal{D}^c = \{1, \cdots, p\} - \mathcal{D}, \tag{3}$$

respectively. The proposed screening procedure relies on ranking the significance of the p covariates according to the magnitude of the variance of their respective marginal regression functions,

$$\sigma^2_{m_k} = \mathrm{var}(m_k(X)) \quad \text{for } k = 1, \cdots, p. \tag{4}$$

Note that $\sigma^2_{m_k} > 0$ for $k \in \mathcal{D}$, while $\sigma^2_{m_k} = 0$ for $k \in \mathcal{D}^c$, making $\sigma^2_{m_k}$ a natural quantity to discriminate between the two classes of predictors. In addition, the variance of the regression function appears as the mean shift, under local alternatives, of the procedure for testing the significance of a covariate proposed in Wang, Akritas and Van Keilegom [8]. This suggests that $\sigma^2_{m_k}$ is the best quantity to discriminate between the two classes of predictors.

If $\hat{m}_k$ denotes an estimator of m_k, $\sigma^2_{m_k}$ can be estimated by the sample variance of $\hat{m}_k(X_{k1}), \cdots, \hat{m}_k(X_{kn})$. The methodology described here works with any type of nonparametric estimator of m_k, but the theory has been developed for Nadaraya-Watson type estimators.

For a kernel function $K(\cdot)$ and bandwidth h, set $\hat{m}_k(X_{ki}) = \sum_{j=1}^n Y_j W_{k;i,j}$, where $W_{k;i,j} = K\left(\frac{X_{kj} - X_{ki}}{h}\right) \Big/ \sum_{j=1}^n K\left(\frac{X_{kj} - X_{ki}}{h}\right)$, and

$$\tilde{S}^2_{m_k} = \frac{1}{n}\sum_{i=1}^n \left(\hat{m}_k(X_{ki}) - \frac{1}{n}\sum_{l=1}^n \hat{m}_k(X_{kl})\right)^2 \tag{5}$$

for the estimator of $\sigma^2_{m_k}$. The bandwidth will be of the order $h = cn^{-1/5}$, throughout this paper. The RV-SIS estimates $\mathcal{D}$ by

$$\hat{\mathcal{D}} = \{k : \tilde{S}^2_{m_k} \geq \hat{C}_d, \text{ for } 1 \leq k \leq p\} \tag{6}$$

for some threshold parameter $\hat{C}_d$. Thus, the RV-SIS procedure reduces the dimension of covariate vector from p to $|\hat{\mathcal{D}}|$, where $|\cdot|$ refers the cardinality of a

set. The choice of C_d, which defines the RV-SIS procedure, is discussed below.

2.2. Thresholding Rule

We adopt the idea of the soft thresholding rule by Zhu *et al.* [9] as a method for choosing the threshold parameter C_d. This method consists of randomly generating a vector $\boldsymbol{Z} = \left(X_{p+1}, \cdots, X_{p+d}\right)$ of d auxiliary random variables from the uniform distribution between (0, 1), $X_{p+i} \sim \text{Unif}\left(0,1\right)$ for $i = 1, \cdots, d$, that are independent of both $\boldsymbol{X}$ and $\boldsymbol{Y}$. By design, the auxiliary variables are inactive predictors. The soft thresholding rule chooses the threshold parameter as

$$\hat{C}_d = \max_{j \in \mathcal{B}} \tilde{S}^2_{m_j}, \tag{7}$$

where $\mathcal{B} = \{p+1, \cdots, p+d\}$ denotes the set of indices of the d auxiliary variables.

Theorem 1 provides an upper bound on the probability of selecting inactive predictors from using the proposed soft thresholding rule provided the following *exchangeability condition* holds.

Exchangeability Condition: Let $k \in \mathcal{D}^c$ and $j \in \mathcal{B}$. Then, the probability that $\tilde{S}^2_{m_k}$ is greater than $\tilde{S}^2_{m_j}$ is equal to the probability that $\tilde{S}^2_{m_k}$ is less than $\tilde{S}^2_{m_j}$.

Theorem 1. Under the exchangeability condition, for any integer $r \in (0, p)$ we have

$$P\left(\left|\hat{\mathcal{D}} \cap \mathcal{D}^c\right| \geq r\right) \leq \left(1 - \frac{r}{p+d}\right)^d. \tag{8}$$

For some constants $c > 0$ and $0 < \kappa < 2/5$.

A practical issue using the soft thresholding is how to choose the size of auxiliary variable d. Numerical simulation results suggested that $d = p/2$ works well on simulated data.

The RV-SIS procedure consists of the following steps:

1. Calculate a sample variance of the nonparametric estimator $\tilde{S}^2_{m_k}$ of each covariate for $k = 1, \cdots, p$.

2. Construct d auxiliary random variables and compute a sample variance of the nonparametric estimator $\tilde{S}^2_{m_j}$ of each auxiliary variable for $j = 1, \cdots, d$.

3. Select predictors whose sample variance of the estimator is greater than the maximum sample variance of the auxiliary variables.

2.3. Sure Screening Properties

In this section, we show that the RV-SIS possesses the sure screening property. The sure screening property is fundamental to a feature screening procedure. This property ensures that all active predictors are selected in the screened submodel with probability 1 as the sample size increases. The following conditions are required for technical proofs:

(C1) There exists positive constants t, C_1 and C_2 such that,

(a) $\max_{1\le k\le p} \mathrm{E}\left\{\exp\left(t\left|Y_j - m_k\left(X_{kj}\right)\right|\right)\right\} < C_1 < \infty$,

(b) $\max_{1\le k\le p} \mathrm{E}\left(\exp\left(t\left(X_{ki} - X_{kj}\right)^2\right)\right) < C_2 < \infty$.

(C2) The kernel $K(\cdot)$ has bounded support, is symmetric, and is Lipschitz continuous, *i.e.*, it satisfies, for some $\Lambda_1 < \infty$ and for all $u, u' \in \mathbb{R}$,

$$\left|K(u) - K(u')\right| \le \Lambda_1 |u - u'|.$$

(C3) If $f_k(x)$ denotes the marginal density of the kth predictor, we have

$$\sup_x |x|^s E\left(|Y| \mid X_k = x\right) f_k(x) \le B < \infty \text{ for some } s \ge 1$$

$\sup_x f_k(x) < \infty$, $\inf_x f_k(x) > 0$, and $f_k(x)$ is uniformly continuous, for all $k = 1, \cdots, p$.

(C4) The conditional expected value $m_k(\cdot)$ is a Lipschitz continuous for all $k = 1, \cdots, p$, that is for some $\Lambda_2 < \infty$ and for all $u, u' \in \mathbb{R}$,

$$\left|m_k(u) - m_k(u')\right| \le \Lambda_2 |u - u'|.$$

(C5) For some constants $c > 0$ and $0 < \kappa < 2/5$, we have

$$\min_{k\in\mathcal{D}} \sigma_{m_k}^2 \ge cn^{-\kappa} + C_d,$$

where C_d is $\max_{j\in\mathcal{B}} \sigma_{m_k}^2$.

In words, Condition (C1) requires that the moment generating functions of the absolute value of the error terms of the marginal regressions and the square difference between two covariates, is finite at least for some $t > 0$. Conditions (C2) and (C3) are standard conditions for establishing uniform convergence rates of needed for the kernel density estimator. Condition (C5) sets a lower bound on the variance of the marginal regression functions of the active predictors.

Theorem 2. Let $\sigma_{m_k}^2, \tilde{S}_{m_k}^2, \mathcal{D}, \hat{\mathcal{D}}$, be defined in (4), (5), (3) and (6), respectively.

1. Under condition (C1)~(C4) for any $0 < \kappa < 2/5$ and $0 < \gamma < 2/5 - \kappa$, there exists positive constants c, c_1, and c_2 such that,

$$P\left(\max_{1\le k\le p}\left|\tilde{S}_{m_k}^2 - \sigma_{m_k}^2\right| \ge cn^{-\kappa}\right)$$
$$\le O\left(p\left[n\exp\left(-c_1 n^{4/5-2(\gamma+\kappa)}\right) + n^2\exp\left(-c_2 n^{\gamma}\right)\right]\right)$$

2. Under condition (C1) ~ (C5), c, c_1, c_2, γ and κ as in part 1,

$$P\left(\mathcal{D} \subseteq \hat{\mathcal{D}}\right) \ge 1 - O\left(|\mathcal{D}|\left[n\exp\left(-c_1 n^{4/5-2(\gamma+\kappa)}\right) + n^2\exp\left(-c_2 n^{\gamma}\right)\right]\right),$$

where $|\mathcal{D}|$ is the cardinality of $\mathcal{D}$.

The second part of Theorem 2 shows that the screened submodel includes all active predictors with the probability approaches to 1 with an exponential rate.

3. Numerical Results

3.1. Simulation Studies

Here we present the result of several simulation studies comparing performance

of the SIS, DC-SIS, NIS, RRCS and RV-SIS methods. In all cases, $\boldsymbol{X} = \left(X_1, X_2, \cdots, X_p\right)^{\mathrm{T}}$ comes from a multivariate normal with mean zero and covariance $\Sigma = \left(\sigma_{ij}\right)_{p\times p}$, and $\varepsilon \sim N(0,1)$. We use three different covariance matrices: (i) $\sigma_{ij} = 0.5^{|i-j|}$, (ii) $\sigma_{ij} = 0.8^{|i-j|}$, and (iii) $\sigma_{ij} = 0.5$. We set the dimension of covariates p to be 2000 and the sample size n to be 200. We replicate the experiment 500 times and base the comparisons on the following three criteria.

R1: The 5%, 25%, 50%, 75%, 95% quantiles of the minimum model size that includes all active covariates.

R2: The proportion of times each individual active covariate is selected in models of size $d_1 = \lfloor n/\log n \rfloor$, $d_2 = \lfloor 2n/\log n \rfloor$ and $d_3 = \lfloor 3n/\log n \rfloor$.

R3: The proportion of times all active covariates are selected in models of size $d_1 = \lfloor n/\log n \rfloor$, $d_2 = \lfloor 2n/\log n \rfloor$ and $d_3 = \lfloor 3n/\log n \rfloor$.

Criterion R1 shows the performance of the ranking of the predictors of the different screening procedures. Criterion R2 and R3 shows the accuracy of the different screening procedure if we used the thresholding value suggested by Fan and Lv [2].

To compare the performance of the screening procedure for both linear and nonlinear cases, we used the following four models:

(a) $Y = 2X_1 + 0.5X_2 + 3\cdot 1\{X_{12} < 0\} + 2X_{22} + \varepsilon$

(b) $Y = 1.5X_1 \cdot X_2 + 3\cdot 1\{X_{12} < 0\} + 2X_{22} + \varepsilon$

(c) $Y = 2\cos(2\pi X_1) + 0.5X_2^2 + 3\cdot 1\{X_{12} < 0\} + 2X_{22} + \varepsilon$

(d) $Y = 2\cos(2\pi X_1) X_2^2 + 3X_{12} + 2\exp(1\{X_{22} < 0\}) + \varepsilon$

All models include an indicator variable. Model (a) is linear, model (b) includes an interaction of two active predictors, model(c) is additive but nonlinear, and model(d) is nonlinear with an interaction term.

Tables 1-3 present the simulation results for R1 using each of the above models with $\sigma_{ij} = 0.5^{|i-j|}$, $\sigma_{ij} = 0.8^{|i-j|}$, and $\sigma_{ij} = 0.5$ respectively. **Tables 4-6** presents the simulation results for R2 and R3 with $\sigma_{ij} = 0.5^{|i-j|}$, $\sigma_{ij} = 0.8^{|i-j|}$, and $\sigma_{ij} = 0.5$ respectively.

These results show that the comparisons in term of the three criteria are similar. All procedures perform worse when we use the equal covariance matrix, $\sigma_{ij} = 0.5$. SIS and RRCS perform rather poorly except in Model (a) where all methods have similar performance. For Model (b), NIS performs slightly better than RV-SIS, while RV-SIS performs somewhat better than DC-SIS when $\sigma_{ij} = 0.8^{|i-j|}$, considerably better when $\sigma_{ij} = 0.5^{|i-j|}$, and significantly better when $\sigma_{ij} = 0.5$. In Models(c) and (d) DC-SIS and NIS have similar performance but RV-SIS performs significantly better than either of them.

Finally, **Table 7** presents the execution time, in seconds, of the DC-SIS, NIS and RV-SIS for Model (d). The RV-SIS procedure takes significantly less time than the DC-SIS and slightly less time than the NIS.

Table 1. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size that includes all active covariates when the covariance matrix is $\sigma_{ij} = 0.5^{|i-j|}$.

Model (a)									
SIS					DC-SIS				
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
4.00	4.00	4.00	5.00	7.00	4.00	4.00	4.00	5.00	6.00
NIS					RRCS				
4.00	4.00	4.00	5.00	7.05	4.00	4.00	4.00	5.00	6.00
RV-SIS									
4.00	4.00	4.00	5.00	9.05					
Model (b)									
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS					DC-SIS				
84.60	526.75	1179.00	1655.00	1923.35	9.00	26.00	68.50	169.25	516.50
NIS					RRCS				
4.00	4.00	6.00	14.00	100.20	214.85	786.50	1355.50	1708.75	1931.10
RV-SIS									
4.00	4.00	7.00	22.00	273.20					
Model (c)									
SIS					DC-SIS				
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
232.00	853.50	1363.50	1689.75	1933.00	103.95	316.25	565.00	860.00	1420.50
NIS					RRCS				
55.00	312.25	749.00	1264.25	1786.15	255.65	929.00	1384.50	1732.25	1943.10
RV-SIS									
5.00	15.00	62.50	277.00	1208.10					
Model (d)									
SIS					DC-SIS				
106.90	583.75	1149.50	1628.75	1930.00	102.90	326.25	654.50	1069.00	1583.70
NIS					RRCS				
33.50	389.00	882.00	1463.25	1915.00	231.55	832.25	1337.00	1678.25	1944.05
RV-SIS									
6.00	20.00	89.00	327.25	1144.55					

Table 2. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size that includes all active covariates when the covariance matrix is $\sigma_{ij} = 0.8^{|i-j|}$.

Model (a)									
SIS					DC-SIS				
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
8.00	11.00	17.00	37.25	249.55	6.00	9.00	12.00	17.00	76.05
NIS					RRCS				
6.00	9.00	13.00	26.00	153.40	6.00	9.00	13.00	22.00	141.35
RV-SIS									
5.00	8.00	11.00	26.25	146.60					
Model (b)									
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
SIS					DC-SIS				
29.90	256.50	924.00	1544.75	1935.05	8.00	10.00	13.00	18.00	40.00
NIS					RRCS				
4.00	6.00	8.00	10.00	22.00	111.60	502.50	1133.00	1636.00	1938.10
RV-SIS									
4.00	6.00	7.00	10.00	32.05					
Model (c)									
SIS					DC-SIS				
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
93.90	520.25	1122.50	1647.25	1925.15	40.95	148.00	334.00	629.00	1149.85
NIS					RRCS				
16.00	74.00	239.50	625.00	1454.60	145.85	595.50	1207.00	1585.25	1930.10
RV-SIS									
9.00	17.00	55.50	244.00	978.30					
Model (d)									
SIS					DC-SIS				
34.80	183.75	701.50	1449.25	1899.40	31.90	142.00	344.00	675.25	1322.10
NIS					RRCS				
18.00	106.00	418.00	1111.00	1815.20	83.90	373.50	979.50	1534.25	1930.05
RV-SIS									
9.00	20.00	45.00	171.50	893.40					

Table 3. The 5%, 25%, 50%, 75%, and 95% quantiles of the minimum model size that includes all active covariates when the covariance matrix is $\sigma_{ij} = 0.5$.

Model (a)									
		SIS					DC-SIS		
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
36.00	47.00	55.00	67.00	91.00	36.95	47.00	55.00	67.00	95.00
		NIS					RRCS		
36.00	47.00	55.00	67.00	90.10	36.00	47.00	56.00	68.00	94.00
		RV-SIS							
37.95	47.00	55.00	67.00	94.00					
Model (b)									
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
		SIS					DC-SIS		
145.95	232.00	411.00	910.75	1979.70	122.00	196.75	322.00	651.25	1716.40
		NIS					RRCS		
78.00	119.75	180.00	267.75	919.40	150.95	273.00	449.00	1089.00	2000.00
		RV-SIS							
77.00	119.75	180.50	314.00	1164.50					
Model (c)									
		SIS					DC-SIS		
5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
151.90	245.75	417.00	806.50	1999.05	132.90	227.00	367.50	732.25	1902.45
		NIS					RRCS		
115.95	192.75	310.50	593.25	1713.60	148.95	251.75	451.50	1013.00	2000.00
		RV-SIS							
69.00	99.00	142.00	208.00	456.05					
Model (d)									
		SIS					DC-SIS		
29.00	41.00	54.00	84.00	160.05	27.95	39.00	53.00	77.00	169.10
		NIS					RRCS		
28.00	39.00	52.00	78.00	164.40	27.95	39.00	53.00	82.00	181.35
		RV-SIS							
24.00	31.00	40.00	54.25	101.15					

Table 4. The proportion of times each individual active covariate and all active covariates are selected in models of size $d_1 = [n/\log n]$, $d_2 = [2n/\log n]$ and $d_3 = [3n/\log n]$ when the covariance matrix is $\sigma_{ij} = 0.5^{|i-j|}$.

Model (a)															
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
	SIS					DC-SIS					RRCS				
d1	1.00	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
d3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	NIS					RV-SIS									
d1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99					
d2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00					
d3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00					
Model (b)															
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
	SIS					DC-SIS					RRCS				
d1	0.08	0.08	1.00	1.00	0.03	0.51	0.51	1.00	1.00	0.33	0.03	0.04	1.00	1.00	0.01
d2	0.12	0.14	1.00	1.00	0.05	0.68	0.68	1.00	1.00	0.52	0.07	0.07	1.00	1.00	0.02
d3	0.16	0.17	1.00	1.00	0.06	0.76	0.78	1.00	1.00	0.65	0.09	0.10	1.00	1.00	0.02
	NIS					RV-SIS									
d1	0.95	0.93	1.00	1.00	0.89	0.89	0.88	1.00	1.00	0.80					
d2	0.97	0.96	1.00	1.00	0.94	0.94	0.93	1.00	1.00	0.88					
d3	0.98	0.97	1.00	1.00	0.96	0.95	0.94	1.00	1.00	0.90					
Model (c)															
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
	SIS					DC-SIS					RRCS				
d1	0.01	0.03	1.00	1.00	0.00	0.04	0.13	1.00	1.00	0.01	0.01	0.02	1.00	1.00	0.00
d2	0.03	0.05	1.00	1.00	0.00	0.08	0.26	1.00	1.00	0.04	0.04	0.03	1.00	1.00	0.00
d3	0.06	0.07	1.00	1.00	0.01	0.12	0.37	1.00	1.00	0.06	0.06	0.05	1.00	1.00	0.00
	NIS					RV-SIS									
d1	0.04	0.59	1.00	1.00	0.03	0.94	0.42	1.00	1.00	0.40					
d2	0.09	0.68	1.00	1.00	0.07	0.97	0.54	1.00	1.00	0.53					
d3	0.12	0.74	1.00	1.00	0.10	0.99	0.60	1.00	1.00	0.59					
Model (d)															
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
	SIS					DC-SIS					RRCS				
d1	0.05	0.10	1.00	0.99	0.02	0.04	0.11	1.00	1.00	0.01	0.04	0.04	1.00	1.00	0.01

Continued

d2	0.08	0.17	1.00	1.00	0.04	0.08	0.24	1.00	1.00	0.03	0.06	0.07	1.00	1.00	0.01
d3	0.11	0.21	1.00	1.00	0.06	0.11	0.35	1.00	1.00	0.06	0.08	0.10	1.00	1.00	0.02
			NIS					RV-SIS							
d1	0.09	0.35	1.00	0.99	0.05	0.58	0.61	1.00	0.97	0.35					
d2	0.14	0.41	1.00	1.00	0.09	0.71	0.69	1.00	0.98	0.49					
d3	0.19	0.46	1.00	1.00	0.12	0.79	0.72	1.00	0.99	0.57					

Table 5. The proportion of times each individual active covariate and all active covariates are selected in models of size $d_1 = \lfloor n/\log n \rfloor$, $d_2 = \lfloor 2n/\log n \rfloor$ and $d_3 = \lfloor 3n/\log n \rfloor$ when the covariance matrix is $\sigma_{ij} = 0.8^{|i-j|}$.

							Model (a)								
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
			SIS					DC-SIS					RRCS		
d1	1.00	1.00	0.75	1.00	0.75	1.00	1.00	0.90	1.00	0.90	1.00	1.00	0.84	1.00	0.84
d2	1.00	1.00	0.85	1.00	0.85	1.00	1.00	0.95	1.00	0.95	1.00	1.00	0.91	1.00	0.92
d3	1.00	1.00	0.89	1.00	0.89	1.00	1.00	0.97	1.00	0.97	1.00	1.00	0.94	1.00	0.94
			NIS					RV-SIS							
d1	1.00	1.00	0.82	1.00	0.82	1.00	1.00	0.81	1.00	0.81					
d2	1.00	1.00	0.91	1.00	0.92	1.00	1.00	0.90	1.00	0.90					
d3	1.00	1.00	0.94	1.00	0.94	1.00	1.00	0.92	1.00	0.93					
							Model (b)								
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
			SIS					DC-SIS					RRCS		
d1	0.10	0.11	0.98	1.00	0.07	0.97	0.96	1.00	1.00	0.94	0.02	0.04	1.00	1.00	0.01
d2	0.16	0.18	0.99	1.00	0.11	0.99	0.99	1.00	1.00	0.99	0.07	0.08	1.00	1.00	0.03
d3	0.20	0.23	1.00	1.00	0.15	0.99	1.00	1.00	1.00	0.99	0.11	0.12	1.00	1.00	0.05
			NIS					RV-SIS							
d1	1.00	1.00	0.99	1.00	0.98	1.00	1.00	0.96	1.00	0.96					
d2	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.99	1.00	0.98					
d3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	0.99					
							Model (c)								
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
			SIS					DC-SIS					RRCS		
d1	0.03	0.05	0.99	1.00	0.02	0.09	0.12	1.00	1.00	0.05	0.02	0.03	0.99	1.00	0.01
d2	0.07	0.09	1.00	1.00	0.05	0.17	0.24	1.00	1.00	0.11	0.05	0.07	1.00	1.00	0.03
d3	0.09	0.11	1.00	1.00	0.06	0.27	0.33	1.00	1.00	0.18	0.07	0.09	1.00	1.00	0.04

Continued

	NIS					RV-SIS				
d1	0.16	0.55	1.00	1.00	0.14	0.99	0.43	1.00	1.00	0.43
d2	0.29	0.68	1.00	1.00	0.26	1.00	0.54	1.00	1.00	0.55
d3	0.38	0.74	1.00	1.00	0.35	1.00	0.62	1.00	1.00	0.62

	Model (d)														
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
	SIS					DC-SIS					RRCS				
d1	0.09	0.20	1.00	0.86	0.06	0.07	0.20	1.00	0.99	0.06	0.03	0.09	1.00	0.96	0.02
d2	0.17	0.27	1.00	0.94	0.15	0.17	0.34	1.00	0.99	0.15	0.07	0.14	1.00	0.98	0.05
d3	0.23	0.31	1.00	0.96	0.19	0.24	0.44	1.00	1.00	0.21	0.10	0.19	1.00	0.99	0.08
	NIS					RV-SIS									
d1	0.20	0.38	1.00	0.88	0.13	0.88	0.60	1.00	0.83	0.45					
d2	0.28	0.46	1.00	0.95	0.21	0.94	0.70	1.00	0.92	0.61					
d3	0.34	0.50	1.00	0.96	0.26	0.96	0.76	1.00	0.94	0.69					

Table 6. The proportion of times each individual active covariate and all active covariates are selected in models of size $d_1 = \lfloor n/\log n \rfloor$, $d_2 = \lfloor 2n/\log n \rfloor$ and $d_3 = \lfloor 3n/\log n \rfloor$ when the covariance matrix is $\sigma_{ij} = 0.5$.

	Model (a)														
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
	SIS					DC-SIS					RRCS				
d1	0.09	0.09	0.64	1.00	0.06	0.13	0.12	0.43	1.00	0.06	0.13	0.12	0.40	1.00	0.05
d2	0.85	0.85	1.00	1.00	0.86	0.86	0.86	0.99	1.00	0.86	0.87	0.86	0.98	1.00	0.86
d3	0.99	0.99	1.00	1.00	0.99	0.98	0.98	1.00	1.00	0.99	0.99	0.99	1.00	1.00	0.99
	NIS					RV-SIS									
d1	0.09	0.09	0.71	1.00	0.07	0.10	0.09	0.73	1.00	0.07					
d2	0.87	0.86	1.00	1.00	0.87	0.85	0.85	1.00	1.00	0.85					
d3	0.98	0.98	1.00	1.00	0.98	0.98	0.98	1.00	1.00	0.98					
	Model (b)														
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
	SIS					DC-SIS					RRCS				
d1	0.00	0.00	0.13	1.00	0.00	0.00	0.00	0.08	1.00	0.00	0.00	0.00	0.07	1.00	0.00
d2	0.00	0.00	0.92	1.00	0.00	0.00	0.00	0.86	1.00	0.00	0.00	0.00	0.84	1.00	0.00
d3	0.01	0.01	1.00	1.00	0.01	0.03	0.03	0.99	1.00	0.03	0.01	0.01	0.99	1.00	0.01
	NIS					RV-SIS									
d1	0.00	0.00	0.27	1.00	0.00	0.00	0.00	0.31	1.00	0.00					

Continued

d2	0.04	0.04	0.96	1.00	0.04	0.04	0.04	0.97	1.00	0.04					
d3	0.22	0.21	1.00	1.00	0.23	0.21	0.21	1.00	1.00	0.22					
	Model (c)														
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
	SIS					DC-SIS					RRCS				
d1	0.00	0.00	0.15	1.00	0.00	0.00	0.00	0.07	1.00	0.00	0.00	0.00	0.07	1.00	0.00
d2	0.00	0.00	0.91	1.00	0.00	0.00	0.00	0.83	1.00	0.00	0.00	0.00	0.81	1.00	0.00
d3	0.01	0.01	1.00	1.00	0.01	0.02	0.03	0.97	1.00	0.03	0.02	0.02	0.96	1.00	0.02
	NIS					RV-SIS									
d1	0.00	0.00	0.27	1.00	0.00	0.00	0.00	0.32	1.00	0.00					
d2	0.00	0.00	0.95	1.00	0.00	0.09	0.09	0.96	1.00	0.09					
d3	0.05	0.05	1.00	1.00	0.05	0.32	0.32	1.00	1.00	0.34					
	Model (d)														
	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL	X_1	X_2	X_{12}	X_{22}	ALL
	SIS					DC-SIS					RRCS				
d1	0.19	0.19	1.00	1.00	0.18	0.22	0.24	1.00	1.00	0.22	0.24	0.23	1.00	1.00	0.23
d2	0.71	0.71	1.00	1.00	0.72	0.74	0.74	1.00	1.00	0.74	0.72	0.72	1.00	1.00	0.72
d3	0.88	0.88	1.00	1.00	0.88	0.87	0.87	1.00	1.00	0.88	0.86	0.86	1.00	1.00	0.86
	NIS					RV-SIS									
d1	0.22	0.21	1.00	1.00	0.21	0.45	0.45	1.00	1.00	0.44					
d2	0.73	0.74	1.00	1.00	0.74	0.87	0.87	1.00	1.00	0.87					
d3	0.88	0.88	1.00	1.00	0.88	0.97	0.97	1.00	1.00	0.97					

Table 7. The comparison of execution time of DC-SIS and RV-SIS in seconds for Model (d) when the covariance matrix is $\sigma_{ij} = 0.5^{|i-j|}$.

Model (d)	DC-SIS					NIS					RV-SIS				
	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%	5%	25%	50%	75%	95%
	18.92	19.17	19.30	19.45	19.86	2.32	2.35	2.36	2.38	2.46	1.81	1.82	1.82	1.83	1.90

3.2. Thresholding Simulation

In this section, we use simulations to compare the soft thresholding rule to the hard thresholding approaches for selecting the submodel. We consider following three models relating the response Y to covariates $X_1, X_2, \cdots, X_p$, where $p = 2000$:

(e) $Y = c_1 X_1 + \cdots + c_{25} X_{25} + \varepsilon$

(f) $Y = c_1 X_1 + \cdots + c_{10} X_{10} + \varepsilon$

(g) $Y = c_1 X_1 + c_2 X_2 + c_3 X_3 + c_4 X_4 + c_5 X_5 + \varepsilon$,

where the covariate vector has the p-variate normal distribution with mean zero

and covariance $\Sigma = \left(0.5^{|i-j|}\right)_{p\times p}$, $\varepsilon \sim N(0,1)$, and the coefficients $c_1,\cdots,c_{25}$ were randomly generated from the uniform distribution between (1, 2.5), and kept fixed throughout the simulation. From each of these models, we generated 500 data sets of size $n = 200$.

For the soft thresholding approach, we randomly generate the auxiliary variable $\boldsymbol{Z} = (X_{2001},\cdots,X_{3000})$, where the X_{p+i} are independent Unif (0, 1). For the hard thresholding we consider three model sizes: $d_1 = [n/\log n] = 37$, $d_2 = [2n/\log n] = 75$, $d_3 = [3n/\log n] = 113$. The two approaches are compared in terms of the proportion of each active covariate is selected. We also record the 5%, 25%, 50%, 75% and 95% quantiles of the submodel size using the soft thresholding rule.

The 5%, 25%, 50%, 75% and 95% quantiles of the submodel size using the soft thresholding rule for Models (e), (f) and (g), are presented in Table 9. The proportion that each of the active covariates is selected with the different approaches for Models (e), (f) and (g) are shown in Tables 6-8, respectively.

From Table 9, it is seen that all percentiles decrease as the number of active covariates decreases; this is a nice feature of the soft thresholding approach. Also, for all models, the median submodel size falls between d_1 and d_2, but is always closer to d_1. Regarding the proportion that each active predictor is included in the submodel, Table 6 and Table 7 show that soft thresholding outperforms hard thresholding with d_1 in Model (e), but does slightly worse in Model(f); hard thresholding with d_2 and d_3 outperform soft thresholding. Finally, Table 8 shows that all active predictors were selected 100% of the time by all approaches.

Table 8. The proportion of times each individual active covariate are selected in models of size d_1, d_2, d_3 and using soft thresholding rule for Model (e).

Model (e)									
Hard Threshold with model size d_1									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.718	0.786	0.806	0.880	0.878	0.910	0.582	0.512	0.704	0.918
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
0.862	0.744	0.934	0.892	0.882	0.944	0.770	0.730	0.688	0.930
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}					
0.960	0.948	0.974	0.888	0.388					
Hard Threshold with model size d_2									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.924	0.868	0.974	0.950	0.950	0.986	0.840	0.834	0.816	0.970
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
0.884	0.858	0.922	0.894	0.890	0.918	0.796	0.792	0.754	0.920
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}					
0.986	0.978	0.988	0.952	0.554					

Continued

Hard Threshold with model size d_3									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.884	0.920	0.928	0.958	0.952	0.968	0.776	0.750	0.862	0.970
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
0.946	0.902	0.982	0.966	0.964	0.990	0.904	0.876	0.878	0.984
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}					
0.992	0.984	0.988	0.960	0.634					
Soft Threshold									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
0.746	0.816	0.844	0.898	0.894	0.920	0.650	0.602	0.742	0.928
X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	X_{19}	X_{20}
0.892	0.800	0.942	0.918	0.900	0.954	0.788	0.770	0.750	0.942
X_{21}	X_{22}	X_{23}	X_{24}	X_{25}					
0.966	0.952	0.970	0.918	0.462					

Table 9. The proportion of times each individual active covariate are selected in models of size d_1, d_2, d_3 and using soft thresholding rule for Model (f).

Model (f)									
Hard Threshold with model size d_1									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.976	0.986	0.988
Hard Threshold with model size d_2									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.990	0.992	0.994
Hard Threshold with model size d_3									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.996	1.000
Soft Threshold									
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1.000	1.000	1.000	1.000	1.000	1.000	0.994	0.968	0.978	0.986

Table 10. The proportion of times each individual active covariate are selected in models of size d_1, d_2, d_3 and using soft thresholding rule for Model (g).

Model (g)				
Hard Threshold with model size d_1				
X_1	X_2	X_3	X_4	X_5
1.000	1.000	1.000	1.000	1.000

Continued

Hard Threshold with model size d_2				
X_1	X_2	X_3	X_4	X_5
1.000	1.000	1.000	1.000	1.000
Hard Threshold with model size d_3				
X_1	X_2	X_3	X_4	X_5
1.000	1.000	1.000	1.000	1.000
Soft Threshold				
X_1	X_2	X_3	X_4	X_5
1.000	1.000	1.000	1.000	1.000

Table 11. The 5%, 25%, 50%, 75%, and 95% quantiles of submodel size using soft thresholding rule for Models (e), (f), and (g).

Model (e)				
5%	25%	50%	75%	95%
20.00	37.00	53.00	75.00	116.00
Model (f)				
5%	25%	50%	75%	95%
13.00	26.00	43.00	66.25	109.00
Model (g)				
5%	25%	50%	75%	95%
8.00	19.75	38.00	62.00	100.15

3.3. A Real Data Example

Here we apply the DC-SIS, NIS and RV-SIS methods to identify the most influential genes for over-expression of a G protein-coupled receptor (Ro1) in mice in the Cardiomyopathy microarray dataset [10]. In this data set, which has also been used in [4], $n = 40$ and $p = 6319$, with the covariates corresponding to expression levels of different genes. **Figure 1** shows the scatterplots of the expression levels of two genes versus Ro1, with fitted cubic spline curves. Because these curves, which are typical for most genes, suggest nonlinear effects, we did not apply SIS to this data.

The top two most influential genes identified by RV-SIS, DC-SIS and NIS are (Msa.2877.0, Msa.741.0), (Msa.2134.0, Msa. 2877.0) and (Msa.2877.0, Msa.1166.0), respectively. To compare the models chosen by the three methods, we fit a semiparametric single index model (SIM).

$$Y = g_k\left(\beta_1 X_{k1} + \beta_2 X_{k2}\right) + \varepsilon \quad \text{for} \quad k = 1,2,3,$$

where $\left(X_{k1}, X_{k2}\right), k = 1,2,3$ are the top two variables chosen by RV-SIS, DC-SIS and NIS, respectively, and use the nonparametric coefficient of determination,

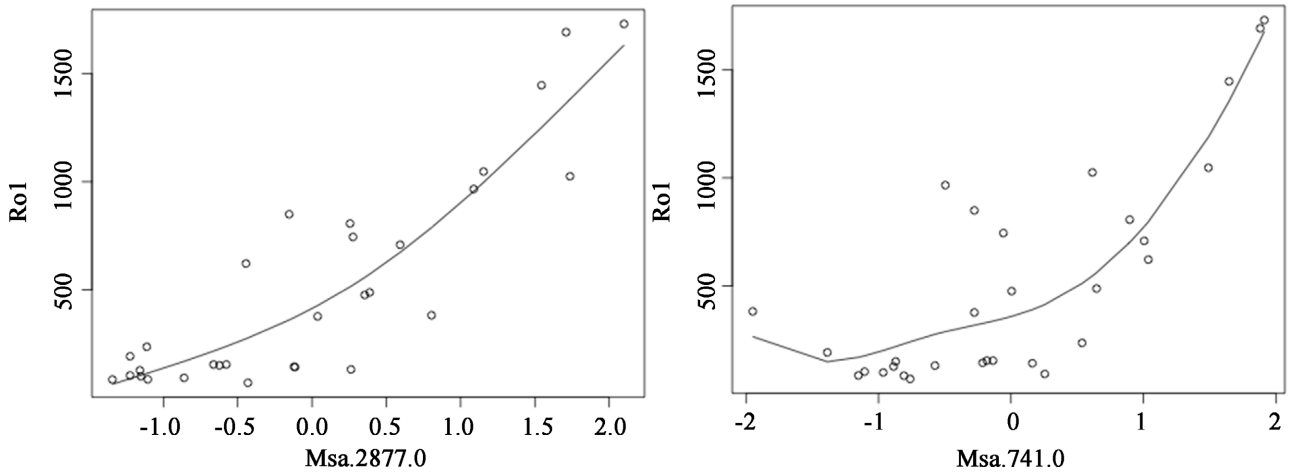

Figure 1. The spline curve of Msa.2877.0 and Msa.741.0.

R^2; see [11]. The R^2-value achieved by RV-SIS, DC-SIS and NIS are 0.927, 0.976 and 0.844, respectively.

The top four most influential genes identified by RV-SIS, DC-SIS and NIS are (Msa.2877.0, Msa.741.0, Msa.1166.0, Msa.26025.0), (Msa.2134.0, Msa. 2877.0, Msa.26025.0, Msa.5583.0) and (Msa.2877.0, Msa.1166.0, Msa.741.0, Msa.18571.0), respectively. Fitting again semiparametric SIMs we obtain R^2-values of 0.9995776, 0.9990484 and 0.9290883 for RV-SIS, DC-SIS and NIS, respectively.

It is seen that, though the selected sets of variables are not identical, RV-SIS, DC-SIS have similar behavior in terms of the nonparametric R^2 criterion, while NIS does somewhat worse.

Kim *et al.* [12] analyzed the ovarian cancer data from The Cancer Genome Atlas (TCGA) to identify the important genes for predicting the ovarian cancer. This data consists of 258 subject and 12,042 gene expressions. We apply RV-SIS and NIS procedures to identify the most influential gene expression for predicting ovarian cancer.

The submodel is selected by the soft thresholding for RV-SIS and by the data driven thresholding using permuted Y then use 99.9th quantile value for NIS. The submodel contains 12 covariates from RV-SIS procedure and 9 covariates from the NIS procedure. We used top 12 covariates from the RV-SIS and top 9 and 12 covariates from the NIS to compare the performance. We fit logistic regression, Random Forest, and Klein and Spady's binary choice estimator (KS) using the submodel selected by RV-SIS and NIS for classification. We record the overall correct classification ratio, specificity, and sensitivity to compare the performance.

Table 12 shows that RV-SIS with Klein and Spady's binary choice estimator perform the best in overall classification, specificity, and sensitivity.

4. Discussion

In this article, we propose the screening procedure, RV-SIS, in a general nonparametric setting. Using a soft thresholding rule for the size of the submodel, it

Table 12. The overall classification rate, sensitivity, and specificity by NIS, RV-SIS, and random forest.

Model	Overall Classification	Specificity	Sensitivity
RV-SIS-KS (12)	0.794	0.664	0.891
NIS-KS (12)	0.755	0.582	0.885
NIS-KS (9)	0.720	0.582	0.824
RV-SIS-Logistic (12)	0.713	0.600	0.797
NIS-Logistic (12)	0.689	0.554	0.790
NIS-Logistic (9)	0.682	0.563	0.770
RV-SIS-RF (12)	0.733	0.655	0.791
NIS-RF (12)	0.744	0.636	0.824
NIS-RF (9)	0.713	0.655	0.757

is shown that RV-SIS possesses the sure screening property.

RV-SIS uses the variance of the marginal regression function in order to rank the predictors. Compared to rankings based on a measure of marginal correlation, the advantage of this ranking is that predictors are ranked according to their predictive significance. Simulations suggest that RV-SIS is more efficient in selecting predictors which influence the response in a nonlinear or nonmonotone fashion; on the other hand, RV-SIS will not select covariates that influence other aspects of the conditional distribution of the response, such as the variance function. The execution time for RV-SIS is competitive compared to other nonparametric methods, making RV-SIS a good candidate for applications to ultra-high-dimensional data.

One issue of practical importance is the choice of the submodel size. Our simulations suggest that soft thresholding has a competitive performance compare to hard thresholding. Moreover, soft thresholding provides an upper bound on the probability of more than r false discoveries. However, thresholding rules do not make a direct link to the false discovery rate. Doing so requires selecting the submodel by suitably determining the cutoff value for the ranking criterion based on its asymptotic distribution. This problem will be addressed in future research.

Similar to other existing screening procedures, RV-SIS relies on a marginal measure between each covariate and the response for ranking of predictors. Due to this, the predictors which are influential jointly but not marginally will not be identified. [13] proposed a process of resuscitation in their partition method for identifying influential predictors that are not identified by marginal observable effects. Resuscitation can also be accomplished by extending the RV-SIS procedure to suitably obtained residuals. This will also be addressed in future research.

Conflicts of Interest

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

References

[1] Fan, J.Q., Samworth, R. and Wu, Y.C. (2009) Ultrahigh Dimensional Feature Selection: BEYOND the linear Model. *The Journal of Machine Learning Research*, **10**, 2013-2038.

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

[3] Fan, J., 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

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

[5] Li, G.R., Peng, H., Zhang, J., Zhu, L.X., *et al*. (2014) Robust Rank Correlation Based Screening. *The Annals of Statistics*, **40**, 1846-1877.

[6] Wang, Z. and Deng, G. (2022) Model-Free Feature Screening Based on Gini Impurity for Ultrahigh-Dimensional Multiclass Classification. *Open Journal of Statistics*, **12**, 711-732. https://doi.org/10.4236/ojs.2022.125042

[7] Chen, T. and Deng, G. (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

[8] Wang, L., Akritas, M.G. and Van Keilegom, I. (2008) An Anova-Type Nonparametric Diagnostic Test for Heteroscedastic Regression Models. *Journal of Nonparametric Statistics*, **20**, 365-382. https://doi.org/10.1080/10485250802066112

[9] Zhu, L., Li, L., Li, R. and Zhu, L. (2011) Model-Free Feature Screening for Ultrahigh-Dimensional Data. *Journal of the American Statistical Association*, **106**, 1464-1475. https://doi.org/10.1198/jasa.2011.tm10563

[10] Segal, M.R., Dahlquist, K.D. and Conklin, B.R. (2003) Regression Approaches for Microarray Data Analysis. *Journal of Computational Biology*, **10**, 961-980. https://doi.org/10.1089/106652703322756177

[11] Doksum, K. and Samarov, A. (1995) Nonparametric Estimation of Global Functionals and a Measure of the Explanatory Power of Covariates in Regression. *The Annals of Statistics*, **23**, 1443-1473. https://doi.org/10.1214/aos/1176324307

[12] Kim, D., Li, R., Dudek, S.M., Frase, A.T., Pendergrass, S.A. and Ritchie, M.D. (2014) Knowledge-Driven Genomic Interactions: An Application in Ovarian Cancer. *BioData Mining*, **7**, Article No. 20. https://doi.org/10.1186/1756-0381-7-20

[13] Chernoff, H., Lo, S. and Zheng, T. (2009) Discovering Influential Variables: A Method of Partitions. *The Annals of Applied Statistics*, **3**, 1335-1369. https://doi.org/10.1214/09-aoas265

[14] Serfling, R.J. (2009) Approximation Theorems of Mathematical Statistics. Wiley.

[15] Hansen, B.E. (2008) Uniform Convergence Rates for Kernel Estimation with Dependent Data, Econometric Theory. Cambridge University Press.

Appendix

A1. Some Lemmas

In all that follows, $f(x)$ is a generic notation for any of the marginal densities $f_k(x)$. Lemmas 1, 2, 3, and 4 are used to prove the Theorem 2.

Lemma 1. For any random variable X which has a moment generating function $\mathrm{E}\{\exp(tX)\}$ for $0 < t < t_0$,

$$P(X - E(X) \geq \varepsilon) \leq \exp(-t\varepsilon)\mathrm{E}\{\exp(t(X - \mathrm{E}(X)))\}, \quad t > 0$$

If $P(|X| \leq M) = 1$, then,

$$\mathrm{E}\{\exp(t(X - \mathrm{E}(X)))\} \leq \exp\left(\frac{1}{2}t^2M^2\right), \quad t > 0$$

Proof. It follows directly from Theorem 5.6.1.A of [14] (2009, pp 201).□

Lemma 2. Suppose $\hat{f}(x)$ be the kernel density estimator of $f(x)$. Under conditions (C2) and (C3), and $h = O(1)$, we have

$$\sup_{x\in\mathbb{R}}|\hat{f}(x) - f(x)| = O\left(\left(\frac{\log(n)}{nh}\right)^{1/2} + h^2\right), \text{ almost surely.}$$

Proof. It follows by writing $|\hat{f}(x) - f(x)| \leq |\hat{f}(x) - E\hat{f}(x)| + |E\hat{f}(x) - f(x)|$, using Theorem 5 of [15] with $Y \equiv 1$ to get $\sup_x|\hat{f}(x) - E\hat{f}(x)| = O\left((\log(n)/nh)^{1/2}\right)$, and $|E\hat{f}(x) - f(x)| = O(h^2)$, which follows by a direct calculation.

Lemma 3. Let $W_j(x) = K\left(\frac{x - X_j}{h}\right)\Big/\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$ be the weight function of the Nadaraya-Watson estimator. Then, under the same assumptions as in Lemma 2, we have

$$\sum_{j=1}^n W_j^2(x) = O\left(\frac{1}{nh}\right), \text{ almost surely.}$$

Proof. Noting that $K^2(\cdot)\Big/\int K^2(u)\mathrm{d}u$ is a symmetric kernel function, by Lemma 2 it is easily seen that

$$\sum_{j=1}^n W_j^2(x) = \frac{1}{nh}\frac{\frac{1}{nh}\sum_{j=1}^n K^2\left(\frac{x - X_j}{h}\right)}{\left(\frac{1}{nh}\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)\right)^2} \quad \square$$

$$= O\left(\frac{1}{nh}\right), \text{ almost surely.}$$

Lemma 4. Under condition (C1)-(a), (C2), (C3), and (C4) and any $0 < \gamma < 2/5$ there exists positive constants c_1, and c_2 such that,

$$P\left(\max_i|\hat{m}(X_i) - m(X_i)| > \varepsilon\right) \leq O\left(n\exp\left(-c_1\varepsilon^2n^{4/5-2\gamma}\right) + n^2\exp\left(-c_2n^\gamma\right)\right)$$

Proof. By adding and subtracting $\sum_j m(X_j)W_j(X_i)$ we have the inequality

$$
\begin{aligned}
&P\left(\left|\hat{m}\left(X_i\right)-m\left(X_i\right)\right|>\varepsilon\right)\\
&\le P\left(\left|\sum_{j=1}^{n}\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)\right|>\frac{\varepsilon}{2}\right)\\
&\quad+P\left(\left|\sum_{j=1}^{n}\left(m\left(X_j\right)-m\left(X_i\right)\right)W_j\left(X_i\right)\right|>\frac{\varepsilon}{2}\right)\\
&\equiv A+B.
\end{aligned}
$$

Note that the dependence of A and B on i is suppressed for convenience. Consider first A. Letting $I_{1j}=I\left\{\left|y_j-m\left(X_j\right)\right|\le M\right\}$, where M will be allowed to tend to ∞ with n, and $I_{2j}=1-I_{1j}$, and noting that $\mathrm{E}\left[\sum_{j=1}^{n}\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)\right]=0$, we have the following inequality

$$
\begin{aligned}
A&\le P\left(\left|\sum_{j=1}^{n}\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)I_{1j}-\sum_{j=1}^{n}\mathrm{E}\left(\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)I_{1j}\right)\right|>\frac{\varepsilon}{4}\right)\\
&\quad+P\left(\left|\sum_{j=1}^{n}\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)I_{2j}-\sum_{j=1}^{n}\mathrm{E}\left(\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)I_{2j}\right)\right|>\frac{\varepsilon}{4}\right)\\
&\equiv A_1+A_2.
\end{aligned}
$$

Arguing conditionally on $\left(X_1,\cdots,X_n\right)$, and using Markov's inequality and Lemma 1,

$$
\begin{aligned}
&P\left(\sum_{j=1}^{n}\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)I_{1j}-\sum_{j=1}^{n}\mathrm{E}\left(\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)I_{1j}\right)>\frac{\varepsilon}{4}\right)\\
&\le\exp\left(-t_1\frac{\varepsilon}{4}\right)\prod_{j=1}^{n}\mathrm{E}\left(\exp\left(t_1\left[\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)I_{1j}-\mathrm{E}\left(\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)I_{1j}\right)\right]\right)\right)\\
&\le\exp\left(-t_1\frac{\varepsilon}{4}\right)\prod_{j=1}^{n}\exp\left(\frac{1}{2}t_1^2W_j^2\left(X_i\right)M^2\right)\\
&=\exp\left(-t_1\frac{\varepsilon}{4}\right)\exp\left(\frac{1}{2}t_1^2M^2\sum_{j=1}^{n}W_j^2\left(X_i\right)\right)\\
&=\exp\left(-\frac{1}{32}\frac{\varepsilon^2}{\sum_j W_j^2\left(X_i\right)M^2}\right),\ \text{by choosing } t_1=\frac{\varepsilon}{4\sum_{j=1}^{n}W_j^2\left(X_i\right)M^2}\\
&\simeq\exp\left(-\frac{1}{32}\frac{\varepsilon^2nh}{M^2}\right),\ \text{by Lemma 3.}
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
&P\left(\sum_{j=1}^{n}\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)I_{1j}-\sum_{j=1}^{n}\mathrm{E}\left(\left(y_j-m\left(X_j\right)\right)W_j\left(X_i\right)I_{1j}\right)<-\frac{\varepsilon}{4}\right)\\
&\le\exp\left(-\frac{1}{32}\frac{\varepsilon^2nh}{M^2}\right)
\end{aligned}
$$

Thus, also unconditionally, we have that for each i,

$$
A_1\le 2\exp\left(-\frac{1}{32}\frac{\varepsilon^2nh}{M^2}\right)
$$

For the A_2 part,

$$A_2 \le P\left(\left|\sum_{j=1}^{n}\left(y_j - m\left(X_j\right)\right)W_j\left(X_i\right)I_{2j}\right| + \sum_{j=1}^{n}\left|\mathrm{E}\left(\left(y_j - m\left(X_j\right)\right)W_j\left(X_i\right)I_{2j}\right)\right| > \frac{\varepsilon}{4}\right)$$

We first show that $\sum_{j=1}^{n}\left|\mathrm{E}\left(\left(y_j - m\left(X_j\right)\right)W_j\left(X_i\right)I_{2j}\right)\right|$ is bounded by $\varepsilon/8$ for n large enough. By the Cauchy-Schwartz and Markov inequalities, we have

$$\begin{aligned}
&\left|\mathrm{E}\left[\left(y_j - m\left(X_j\right)\right)W_j\left(X_i\right)I_{2j}\right]\right| \\
&\le \sqrt{\mathrm{E}\left[\left\{\left(y_j - m\left(X_j\right)\right)W_j\left(X_i\right)\right\}^2\right]P\left(\left|y_j - m\left(X_j\right)\right| > M\right)} \\
&\le \sqrt{\mathrm{E}\left[\left\{y_j - m\left(X_j\right)\right\}^2 W_j^2\left(X_i\right)\right]\exp\left(-tM\right)\mathrm{E}\left\{\exp\left(t\left|y_j - m\left(X_j\right)\right|\right)\right\}}
\end{aligned}$$

By condition (C1)-(a), there exists a constant t such that $\mathrm{E}\left\{\exp\left(t\left|y_j - m\left(X_j\right)\right|\right)\right\} < C_1$. Also, by Lemma 2, $\mathrm{E}\left[\left\{y_j - m\left(X_j\right)\right\}^2 W_j^2\left(X_i\right)\right] = O\left(1/(nh)^2\right)$, uniformly in i. Then, by choosing $M = n^{\gamma}$, some $\gamma > 0$, we have $\sum_{j=1}^{n}\left|\mathrm{E}\left(\left(y_j - m\left(X_j\right)\right)W_j\left(X_i\right)I_{2j}\right)\right| < \varepsilon/8$, for n large enough. Hence, for n large enough,

$$A_2 \le P\left(\left|\sum_{j=1}^{n}\left(y_j - m\left(X_j\right)\right)W_j\left(X_i\right)I_{2j}\right| > \frac{\varepsilon}{8}\right).$$

To bound this, note first that

$$\left\{\left|\sum_{j=1}^{n}\left(y_j - m\left(X_j\right)\right)W_j\left(X_i\right)I_{2j}\right| > \varepsilon/8\right\} \subset \bigcup_{j=1}^{n}\left\{\left|y_j - m\left(X_j\right)\right| > M\right\}.$$

Indeed, if the event on the left hand side holds it must be that $\left|y_j - m\left(X_j\right)\right| > M$ for at least one j since, otherwise $\left|\left(y_j - m\left(X_j\right)\right)W_j\left(X_i\right)I_2\right| = 0$ for all j which contradicts $\left|\sum_{j=1}^{n}\left(y_j - m\left(X_j\right)\right)W_j\left(X_i\right)I_2\right| > \varepsilon/8$. Thus, by condition (C1)-(a), it follows that

$$\begin{aligned}
A_2 &\le P\left(\bigcup_{j}\left\{\left|y_j - m\left(X_j\right)\right| > M\right\}\right) \\
&\le nP\left(\left|y_j - m\left(X_j\right)\right| > M\right) \\
&\le n\exp\left(-tM\right)\mathrm{E}\left[\exp\left(t\left|y_j - m\left(X_j\right)\right|\right)\right] \\
&= nC_1\exp\left(-tM\right)
\end{aligned}$$

Then by choosing $M = n^{\gamma}$, $0 < \gamma < 2/5$, we have

$$\begin{aligned}
A &\le 2\exp\left(-\frac{1}{32}\frac{\varepsilon^2 nh}{M^2}\right) + nC_1\exp\left(-tM\right) \\
&= 2\exp\left(-\frac{1}{32}\varepsilon^2 n^{1-2\gamma}h\right) + nC_1\exp\left(-tn^{\gamma}\right)
\end{aligned} \tag{9}$$

Consider now part B. By condition (C4) and for n large enough, we have

$$
\begin{aligned}
B &\le P\left(\sum_{j=1}^{n}\left|\left(m\left(X_j\right)-m\left(X_i\right)\right)W_j\left(X_i\right)\right|>\frac{\varepsilon}{2}\right)\\
&\le P\left(\sum_{j=1}^{n}\left|\Lambda_2\left(X_j-X_i\right)W_j\left(X_i\right)\right|>\frac{\varepsilon}{2}\right) \qquad (10)\\
&\le P\left(\Lambda_2 h\sum_{j=1}^{n}W_j\left(X_i\right)>\frac{\varepsilon}{2}\right)=P\left(\Lambda_2 h>\frac{\varepsilon}{2}\right)=0.
\end{aligned}
$$

Therefore, by (9) and (10), we have that for all n large enough

$$
P\left(\left|\hat{m}\left(X_i\right)-m\left(X_i\right)\right|>\varepsilon\right)\le 2\exp\left(-\frac{1}{32}\varepsilon^2 n^{1-2\gamma}h\right)+nC_1\exp\left(-tn^{\gamma}\right).
$$

It follows that under condition (C1)-(a), (C2), (C3), and (C4), and any $0<\gamma<2/5$, there exists positive constants c_1 and c_2 such that,

$$
\begin{aligned}
P\left(\max_i\left|\hat{m}\left(X_i\right)-m\left(X_i\right)\right|>\varepsilon\right) &\le O\left(n\exp\left(-c_1\varepsilon^2 n^{1-2\gamma}h\right)+n^2\exp\left(-c_2 n^{\gamma}\right)\right)\\
&\le O\left(n\exp\left(-c_1\varepsilon^2 n^{4/5-2\gamma}\right)+n^2\exp\left(-c_2 n^{\gamma}\right)\right),
\end{aligned}
$$

by substituting $n^{-1/5}$ for h.

A2. Proof of Theorem 2 for Part 1 Write

$$
\begin{aligned}
P\left(\left|\tilde{S}_{m_k}^2-\sigma_{m_k}^2\right|\ge\varepsilon\right) &\le P\left(\left|\tilde{S}_{m_k}^2-S_{m_k}^2\right|\ge\varepsilon/2\right)+P\left(\left|S_{m_k}^2-\sigma_{m_k}^2\right|\ge\varepsilon/2\right)\\
&\equiv T_1+T_2,
\end{aligned}
$$

where $S_{m_k}^2=\frac{1}{n}\sum_{i=1}^{n}\left[m_k\left(X_i\right)-\left(\frac{1}{n}\sum_{l=1}^{n}m_k\left(X_i\right)\right)\right]^2$. For convenience in notation, we will omit the subscript k from m_k and X_{kj}, $j=1,\cdots,n$, for the rest of this proof. For T_1 we have

$$
\begin{aligned}
T_1 &= P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}^2\left(x_i\right)-m^2\left(X_i\right)\right)-\left(\frac{1}{n}\sum_{l=1}^{n}\hat{m}\left(X_i\right)\right)^2-\left(\frac{1}{n}\sum_{l=1}^{n}m\left(X_i\right)\right)^2\right|>\frac{\varepsilon}{2}\right)\\
&\le P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}\left(X_i\right)-m\left(X_i\right)\right)\left(\hat{m}\left(X_i\right)+m\left(X_i\right)\right)\right|>\frac{\varepsilon}{4}\right)\\
&\quad+P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}\left(X_i\right)-m\left(X_i\right)\right)\left(\frac{1}{n}\sum_{l=1}^{n}\left(\hat{m}\left(X_i\right)+m\left(X_i\right)\right)\right)\right|>\frac{\varepsilon}{4}\right)\\
&\le P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}\left(X_i\right)-m\left(X_i\right)\right)^2\right|>\frac{\varepsilon}{8}\right)\\
&\quad+P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}\left(X_i\right)-m\left(X_i\right)\right)\left(2m\left(X_i\right)\right)\right|>\frac{\varepsilon}{8}\right)\\
&\quad+P\left(\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}\left(X_i\right)-m\left(X_i\right)\right)\right]^2>\frac{\varepsilon}{8}\right)\\
&\quad+P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}\left(X_i\right)-m\left(X_i\right)\right)\frac{1}{n}\sum_{i=1}^{n}2m\left(X_i\right)\right|>\frac{\varepsilon}{8}\right)\\
&\equiv A_1+A_2+A_3+A_4.
\end{aligned}
$$

The following inequalities all follow by Lemma 4 (so that $0<\gamma<2/5$):

$$\begin{aligned}
A_1 &\le P\left(\max_i\left(\hat{m}(X_i)-m(X_i)\right)^2>\frac{\varepsilon}{8}\right)\\
&\le O\left(n\exp\left(-c_1\varepsilon^2 n^{4/5-2\gamma}\right)+n^2\exp\left(-c_2 n^{\gamma}\right)\right),\\
A_2 &\le P\left(\max_i\left|\left(\hat{m}(X_i)-m(X_i)\right)\right|>\frac{\varepsilon}{16\sup_x|m(x)|}\right)\\
&\le O\left(n\exp\left(-c_1\varepsilon^2 n^{4/5-2\gamma}\right)+n^2\exp\left(-c_2 n^{\gamma}\right)\right),\\
A_3 &\le P\left(\max_i\left|\left(\hat{m}(X_i)-m(X_i)\right)\right|>\varepsilon\right)\\
&\le O\left(n\exp\left(-c_1\varepsilon^2 n^{4/5-2\gamma}\right)+n^2\exp\left(-c_2 n^{\gamma}\right)\right),\\
A_4 &\le P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}(X_i)-m(X_i)\right)\right|>\frac{\varepsilon}{16\sup_x|m(x)|}\right)\\
&\le O\left(n\exp\left(-c_1\varepsilon^2 n^{4/5-2\gamma}\right)+n^2\exp\left(-c_2 n^{\gamma}\right)\right).
\end{aligned}$$

Combining the above we have

$$T_1 \le O\left(n\exp\left(-c_1\varepsilon^2 n^{4/5-2\gamma}\right)+n^2\exp\left(-c_2 n^{\gamma}\right)\right). \tag{11}$$

Consider now T_2, and let $h(X_i,X_j)$ be the kernel of the *U*-statistic $U_m=\left[n/(n-1)\right]S_m^2$. For a constant *M*, we decompose U_m as $U_m=U_{1m}+U_{2m}$, where

$$U_{1m}=\frac{1}{n(n-1)}\sum_{i\ne j}h(X_i,X_j)I\left\{h(X_i,X_j)\le M\right\} \text{ and } U_{2m}=U_m-U_{1m}.$$

Similarly, we decompose $\sigma_m^2=\mathrm{E}(U_m)$ as $\sigma_m^2=\sigma_{1m}^2+\sigma_{1m}^2$, where

$$\sigma_{1m}^2=\mathrm{E}\left(h(X_i,X_j)I\left\{h(X_i,X_j)\le M\right\}\right) \text{ and } \sigma_{2m}^2=\sigma_m^2-\sigma_{1m}^2.$$

Then we have following inequality

$$\begin{aligned}
T_2 &= P\left(\left|\frac{n-1}{n}U_m-\sigma_m^2\right|\ge\varepsilon/2\right)\\
&\le P\left(\left|\frac{n-1}{n}\left(U_{1m}-\sigma_{1m}^2\right)\right|\ge\varepsilon/4\right)+P\left(\left|\frac{n-1}{n}\left(U_{2m}-\sigma_{2m}^2\right)-\frac{1}{n}\sigma_m^2\right|\ge\varepsilon/4\right) \qquad (12)\\
&\equiv C_1+C_2
\end{aligned}$$

By Lemma 1 we have that for any $t>0$,

$$P\left(\frac{n-1}{n}\left(U_{1m}-\sigma_{1m}^2\right)\ge\varepsilon/4\right)\le\exp\left(-\frac{t\varepsilon n}{4(n-1)}\right)\exp\left(-t\sigma_{1m}^2\right)\mathrm{E}\left(\exp\left(tU_{1m}\right)\right). \tag{13}$$

Next, using the representation $U_{m1}=\frac{1}{n!}\sum_{n!}W\left(X_{i_1},\cdots,X_{i_n}\right)$, where

$W\left(X_1,\cdots,X_n\right)=\frac{1}{m}\sum_{i=1}^{m}h\left(X_{2i-1},X_{2i}\right)I\left\{h\left(X_{2i-1},X_{2i}\right)\le M\right\}$ is an average of $m=\left[n/2\right]$ i.i.d random variables, and $\uparrow\sum_{n!}$ denotes the summation over all

possible permutations of $(1,\cdots,n)$ (cf. Serfling [14], 1981, pp. 180-181), we have

$$\begin{aligned}\mathrm{E}\left(\exp\left(tU_{1m}\right)\right) &= \mathrm{E}\left(\exp\left(\frac{t}{n!}\sum_{n!}W\left(X_{i_1},\cdots,X_{i_n}\right)\right)\right)\\ &\le \frac{1}{n!}\sum_{n!}\mathrm{E}\left[\exp\left\{tW\left(X_{i_1},\cdots,X_{i_n}\right)\right\}\right]\\ &= \mathrm{E}\left(\exp\left(\sum_{i=1}^{m}\frac{t}{m}h\left(X_{2i-1},X_{2i}\right)I\left\{h\left(X_{2i-1},X_{2i}\right)\le M\right\}\right)\right)\\ &= \mathrm{E}^m\left(\exp\left(\frac{t}{m}h\left(X_{2i-1},X_{2i}\right)I\left\{h\left(X_{2i-1},X_{2i}\right)\le M\right\}\right)\right),\end{aligned}$$

where Jensen's inequality was also used. Substituting this in (13) we have,

$$\begin{aligned}&P\left(\frac{n-1}{n}\left(U_{1m}-\sigma_{1m}^2\right)\ge\varepsilon/4\right)\\ &\le \exp\left(-\frac{tn\varepsilon}{4(n-1)}\right)\mathrm{E}^m\left(\exp\left(\frac{t}{m}\left(h\left(X_{2i-1},X_{2i}\right)I\left\{h\left(X_{2i-1},X_{2i}\right)\le M\right\}-\sigma_{m1}^2\right)\right)\right)\\ &\le \exp\left(-\frac{t\varepsilon n}{4(n-1)}+\frac{t^2M^2}{2m}\right)\ \text{by Lemma 1}\\ &\le \exp\left(-\frac{n^2\varepsilon^2 m}{32M^2(n-1)^2}\right)\ \text{by choosing } t\\ &= \frac{n\varepsilon m}{4M^2(n-1)}\end{aligned}$$

Therefore, for C_1 given in (12) we have

$$\begin{aligned}C_1 &\le 2\exp\left(-\frac{n^2\varepsilon^2 m}{32M^2(n-1)^2}\right)\\ &\le 2\exp\left(-\frac{\varepsilon^2 n}{64M^2}\right)\end{aligned}\tag{14}$$

Consider now C_2 given in (12). Note first that $\sigma_m^2/n<\varepsilon/16$ for all n sufficient large. Also, by the Cauch-Schwartz and Markov inequalities, we have

$$\begin{aligned}\sigma_{2m}^2 &\le \sqrt{E\left(h^2\left(X_i,X_j\right)\right)P\left(h\left(X_i,X_j\right)>M\right)}\\ &\le \sqrt{E\left(h^2\left(X_i,X_j\right)\right)\exp(-tM)E\left(\exp\left(th\left(X_i,X_j\right)\right)\right)}\end{aligned}$$

so that, by choosing $M=n^{\gamma}$, $\gamma>0$, condition (C1)-(b) yields $(n-1)\sigma_{2m}^2/n<\varepsilon/16$ for n sufficient large. Thus, for n large enough,

$$C_2 \le P\left(\left|\frac{n-1}{n}U_{2m}\right|>\varepsilon/8\right).$$

To bound this, observe that $\left\{\left|\frac{n-1}{n}U_{2m}\right|\ge\varepsilon/8\right\}\subseteq\bigcup_{i\ne j}\left\{\left|h\left(X_i,X_j\right)\right|\ge M\right\}$. Thus, by Markov's inequality and condition (C1)-(b), it follows that

$$\begin{aligned} C_2 &\le P\left(\bigcup_{i\neq j}\left|h\left(X_i, X_j\right)\right| \ge M\right) \\ &\le n^2 \exp(-tM)\,\mathrm{E}\left(\exp\left(t\left|h\left(X_i, X_j\right)\right|\right)\right) \\ &\le n^2 C_3 \exp\left(-tn^{\gamma}\right) \end{aligned} \tag{15}$$

Combining (12), (14) with $M = n^{\gamma}$, for $\gamma < 1/2$, and (15), we have

$$\begin{aligned} T_2 &\le 2\exp\left(-\frac{\varepsilon^2 n^{1-2\gamma}}{64}\right) + n^2 \exp\left(-tn^{\gamma}\right) \\ &= O\left(\exp\left(-c_3\varepsilon^2 n^{1-2\gamma}\right) + n^2 \exp\left(-c_4 n^{\gamma}\right)\right), \end{aligned} \tag{16}$$

for some positive constants c_3 and c_4.

By (11), (11) and (16), for $0 < \gamma < 2/5$ we have

$$P\left(\left|\tilde{S}^2 - \sigma_m^2\right| \ge \varepsilon\right) = O\left(n\exp\left(-c_1\varepsilon^2 n^{4/5-2\gamma}\right) + n^2\exp\left(-c_2 n^{\gamma}\right)\right)$$

It follows that for $0 < \gamma < 2/5$

$$\begin{aligned} P\left(\max_k\left|\tilde{S}_k^2 - \sigma_{m_k}^2\right| \ge \varepsilon\right) &\le O\left(p\left[n\exp\left(-c_1\varepsilon^2 n^{4/5-2\gamma}\right) + n^2\exp\left(-c_2 n^{\gamma}\right)\right]\right) \\ &= O\left(p\left[n\exp\left(-c_1 n^{4/5-2(\gamma+\kappa)}\right) + n^2\exp\left(-c_2 n^{\gamma}\right)\right]\right) \end{aligned}$$

The last equality holds by choosing $\varepsilon = cn^{-\kappa}$ for a constant $c > 0$, $0 < \kappa < 2/5$ and $0 < \gamma < 2/5 - \kappa$.

For part 2 of Theorem 2, if $\mathcal{D} \not\subseteq \hat{\mathcal{D}}$, then there must exists some $k \in \mathcal{D}$ such that $\tilde{S}_k^2 < C_d$. $k \notin \hat{\mathcal{D}}$. It follows from condition (C5) that $\sigma_{m_k}^2 - \tilde{S}_k^2 > cn^{-\kappa}$ for some $k \in \mathcal{D}$. Thus, $\left\{\mathcal{D} \not\subseteq \hat{\mathcal{D}}\right\} \subseteq \max_{k\in\mathcal{D}}\left\{\left|\tilde{S}_k^2 - \sigma_{mk}^2\right| > cn^{-\kappa}\right\}$. Using part 1 of this theorem we have

$$\begin{aligned} P\left(\mathcal{D} \subseteq \hat{\mathcal{D}}\right) &\ge 1 - P\left(\min_{k\in\mathcal{D}}\left|\tilde{S}_k^2 - \sigma_{mk}^2\right| > cn^{-\kappa}\right) \\ &= 1 - |\mathcal{D}|\,P\left(\left|\tilde{S}_k^2 - \sigma_{mk}^2\right| > cn^{-\kappa}\right) \\ &\ge 1 - O\left(|\mathcal{D}|\left[\exp\left(-n\exp\left(-c_1 n^{4/5-2(\gamma+\kappa)}\right) + n^2\exp\left(-c_2 n^{\gamma}\right)\right)\right]\right). \end{aligned}$$

□