Statistics and Its Interface Volume 12 (2019) 239–251

Model-free conditional feature screening with exposure variables

Yeqing Zhou, Jingyuan Liu*,†, Zhihui Hao, and Liping Zhu‡

In high dimensional analysis, effects of explanatory variables on responses sometimes rely on certain exposure variables, such as time or environmental factors. In this paper, to characterize the importance of each predictor, we utilize its conditional correlation given exposure variables with the empirical distribution function of response. A model-free conditional screening method is subsequently advocated based on this idea, aiming to identify significant predictors whose effects may vary with the exposure variables. The proposed screening procedure is applicable to any model form, including that with heteroscedasticity where the variance component may also vary with exposure variables. It is also robust to extreme values or outlier. Under some mild conditions, we establish the desirable sure screening and the ranking consistency properties of the screening method. The finite sample performances are illustrated by simulation studies and an application to the breast cancer dataset.

Keywords and phrases: Conditional screening, Feature screening, Exposure variable, Model-free, Sure screening property, Variable selection.

1. INTRODUCTION

Ultrahigh dimensional data arise in many frontier areas, such as genetics, imaging, economics and finance. In these areas, quite often tremendous amount of explanatory variables are collected, while only a few predictors are truly important to the response. To identify these truly active predictors, a variety of variable selection methods were studied based on different models. One appealing method to select important variables and reduce the predictor dimensionality is the two-stage approach: feature screening methods are first conducted to roughly rule out the marginally unimportant predictors, and subsequent regularized regression approaches are applied to recover the final sparse models. In the screening stage, Fan and Lv (2008) first proposed a sure independent screening procedure (SIS) based on the marginal Pearson correlation in the context of linear models. Its appealing sure screening property urges statisticians to extend the idea of SIS under different settings, including the generalized linear models (Fan and Song, 2010), semiparametric models (Li et al., 2012a), nonparametric additive models (Fan et al., 2011; He et al., 2013). Fan and Song (2010) stated that the validity of screening methods usually relies on the correct underlying-model specification, which motivates researchers to propose screening methods at a model-free basis, such as SIRS (Zhu et al., 2011), DC-SIS (Li et al., 2012b) and BCor-SIS (Pan et al., 2018). See Liu et al. (2015) for a selective overview of the screening methods.

However, the effects of predictors on the response are sometimes dependent on certain exposure variables in an unknown pattern, such as time or some environmental indices. For instance, in human genetics research, gene effects on certain phenotype, say body mass index, may reply on the current age of people. When the predictors affect the response via one or more exposure variables, the corresponding effects are often depicted by the interactions between predictors and the exposure variables in linear models, or by the nonparametric coefficient functions in varying coefficient models. In the varying coefficient model, the dependence between predictors and response can be marginally assessed by the conditional Pearson correlation given exposure variables, due to the linearity of varying coefficient models given exposure variables. Therefore, to reduce the dimensionality of such ultrahigh dimensional varying coefficient models, Liu et al. (2014) and Fan et al. (2014) studied several conditional screening methods, based on such the partial correlation and the conditional correlation learning.

In analysis of ultrahigh dimensional data, unfortunately, we are often lack of prior information on the regression structure (Zhu et al., 2011), and the aforementioned linearity can be easily violated. In addition, extreme values or outliers often take a non-negligible role when tremendous amount of data are collected, destroying the nice data structure for applying the methods to the well-designed varying coefficient models. Under some other circumstances, predictors might be responsible for the variance, rather than the mean of response and the exposure variables may also play a role in the effects on the variance component. To address these issues,

arXiv: 1804.03637
*Corresponding author.
†The research of Jingyuan Liu is supported in part by National Natural Science Foundation of China (NSFC, 11771361), JAS14007, and Fundamental Research Funds for the Scientific Research Foundation for the Returned Overseas Chinese Scholars.
‡The research of Liping Zhu is supported in part by National Natural Science Foundation of China (NSFC, 11371236, 11422107), and Henry Fok Education Foundation Fund of Young College Teachers (141002).

Wen et al. (2018) developed a model-free screening method based on conditional distance correlation learning (Wang et al., 2015). However, the performance of this method is easily influenced by the presence of extreme values or outliers in the observations. Thus, we are motivated to use a robust metric to measure the conditional association between the predictors and response given exposure variables and apply it to the feature screening procedures. We adopt the conditional correlation between the predictor and indicator function of response given the exposure variables. It employs the conditional rank instead of the original observed value of the response and thus stays invariant after strictly monotone transformation of the response. In estimation, the standard Nadaraya-Watson estimator is applied, which is easy to implement. Using the metric as a marginal score function, we further develop a model-free conditional sure independence screening procedure. The sure screening property (Fan and Lv, 2008) and ranking consistency property (Zhu et al., 2011) of the screening procedure are carefully studied. We conduct extensive simulations to illustrate our proposed method is effective to detect both linear and nonlinear conditional relations between the predictors and response given exposure variables, and ranks the important predictors above the unimportant ones with an overwhelming probability.

The rest of this paper is organized as follows. In Section 2, we propose a model-free conditional feature screening procedure based on the correlation learning, with a careful study of its theoretical properties. In Section 3, we conduct Monte Carlo simulations to evaluate the finite sample performance of our proposals, and apply the method to analyze the breast cancer data. A discussion is given in Section 4. All technical proofs are relegated to the Appendix.

2. CONDITIONAL SURE INDEPENDENCE RANKING AND SCREENING

2.1 Some preliminaries

Suppose $Y \in \mathbb{R}^1$ is the response variable, $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ is the associated predictor vector and $u \in \mathbb{U}$ is the exposure variable, where $\mathbb{U}$ is the support of u. We assume $\mathbb{U}$ to be bounded with finite constants. Given the exposure variables u, we define the set of active predictors without model-specification:

$$(1)\quad \mathcal{A} = \{k : \text{Given } u \in \mathbb{U},\ F(y \mid \mathbf{x}, u) \text{ varies with } X_k \text{ for some } y \in \mathbb{R}^1\}$$

where $F(y \mid \mathbf{x}, u)$ stands for the conditional distribution function of Y given $\mathbf{x}$ and u. (1) indicates that the truly active predictors affect the response variable through its distribution function, which may also depend on u. The set of inactive predictors is denoted by $\mathcal{A}^c$, the complementary set of $\mathcal{A}$. A screening method aims at removing as many predictors inactive predictors $\mathbf{x}_{\mathcal{A}^c}$ as possible while retaining all the active predictors $\mathbf{x}_{\mathcal{A}}$. Thus, we need to adopt a reasonable metric to measure the relative importance of each predictor conditioning on the exposure variables u.

We briefly review the sure independent ranking and screening procedure (Zhu et al., 2011, SIRS), which identifies active predictors satisfying $F(y \mid \mathbf{x}) = F(y \mid \mathbf{x}_{\mathcal{A}})$ for all $y \in \mathbb{R}^1$. For easy illustration of its rationale, assume that $\mathbf{x}$ follows standard multivariate normal distribution and each predictor is standardized. The conditional distribution Y given $\mathbf{x}$ varies with $\mathbf{x}_{\mathcal{A}}$, and stay constant with $\mathbf{x}_{\mathcal{A}}^c$. Thus it is natural to expect that $E\{\partial F(y \mid X_k)/\partial X_k\}$ to be non-zero for $k \in \mathcal{A}$ and zero for $k \in \mathcal{A}^c$. The normality assumption implies that $E\{\partial F(y \mid X_k)/\partial X_k\} = E\{X_k I(Y \le y)\}$, where $I(\cdot)$ is an indicator function. Thus, by defining $\rho_k(y) = E\{X_k I(Y \le y)\}$, Zhu et al. (2011) employs $E\{\rho_k^2(Y)\}$ to rank the relative importance of predictors. The indicator function in $\rho_k(y)$ ensures the robustness of the method to extreme values and outliers.

When exposure variable u is involved, however, the distribution of Y, as well as its association with the predictors, may vary with u. Under this circumstance, only considering the marginal expectations in $\rho_k(y)$ may miss important u-varying information. Instances indeed exist (Liu et al., 2014) where marginal screening procedures fail to detect those predictors with varying effects of u. To address this issue, we define the conditional correlation between the predictor X_k for $k = 1, \ldots, p$ and the indicator function of response Y conditioning on u as follows.

$$\begin{aligned}\Omega_k(u, y) &= \mathrm{corr}(X_k, I(Y \le y)|u) \\ &= \frac{\mathrm{cov}(X_k, I(Y \le y)|u)}{\sqrt{\mathrm{var}(X_k|u)\mathrm{var}(I(Y \le y)|u)}}.\end{aligned}$$

Then the marginal utility for screening becomes

$$\omega_k = E_y\left[E_u\left\{\Omega_k^2(u, y)\right\}\right],\ k = 1, \ldots, p.$$

To estimate ω_k based on the random sample $\{(\mathbf{x}_i, Y_i, u_i), i = 1, \ldots, n\}$, we adopt Nadaraya-Watson estimator for each conditional mean used to compute $\Omega_k(u, y)$. Specially,

$$\begin{aligned}\widehat{E}(X_k \mid u_j) &= \frac{n^{-1}\sum_{i=1}^n K_h(u_i - u_j)X_{ik}}{n^{-1}\sum_{i=1}^n K_h(u_i - u_j)}, \\ \widehat{E}(X_k^2 \mid u_j) &= \frac{n^{-1}\sum_{i=1}^n K_h(u_i - u_j)X_{ik}^2}{n^{-1}\sum_{i=1}^n K_h(u_i - u_j)}, \\ \widehat{E}\{I(Y \le Y_l) \mid u_j\} &= \frac{n^{-1}\sum_{i=1}^n K_h(u_i - u_j)I(Y_i \le Y_l)}{n^{-1}\sum_{i=1}^n K_h(u_i - u_j)}, \\ \widehat{E}\{X_k I(Y \le Y_l) \mid u_j\} &= \frac{n^{-1}\sum_{i=1}^n K_h(u_i - u_j)X_{ik}I(Y_i \le Y_l)}{n^{-1}\sum_{i=1}^n K_h(u_i - u_j)},\end{aligned}$$

where $K_h(u) = K(u/h)/h$, $K(\cdot)$ is a kernel function and h is the bandwidth. Then a natural estimator of $\hat{\omega}_k$ is

$$\hat{\omega}_k = \frac{1}{n^2}\sum_{l=1}^n\sum_{j=1}^n \frac{\widehat{\mathrm{cov}}^2\{X_k, I(Y \le Y_l) \mid u_j\}}{\widehat{\mathrm{var}}(X_k \mid u_j)\widehat{\mathrm{var}}\{I(Y \le Y_l) \mid u_j\}},$$

where $\widehat{\text{cov}}\{X_k, I(Y \le Y_l) \mid u_j\}$ can be estimated through $\widehat{E}\{X_k I(Y \le Y_l) \mid u_j\} - \widehat{E}(X_k \mid u_j)\widehat{E}\{I(Y \le Y_l) \mid u_j\}$. The variance term can be estimated by similar patterns. Based on the sample estimation of $\hat{\omega}_k$, we conduct the screening criterion to identify the active set indexed by

$$\widehat{\mathcal{A}} = \{k : \hat{\omega}_k \text{ ranks in the top } d \text{ among all } \hat{\omega}\text{'s, for } 1 \le k \le p\},$$

where d is the user-specified threshold value. We refer the proposed conditional sure independent ranking and screening procedure as C-SIRS in the paper.

3. THEORETICAL PROPERTIES

We establish several appealing properties for our proposed screening procedure. Denote $\lambda_{\max}(\mathbf{C})$ and $\lambda_{\min}(\mathbf{C})$ for the largest and smallest eigenvalues of a matrix $\mathbf{C}$, respectively. Write $\mathbf{v}^2 = \mathbf{v}\mathbf{v}^{\mathrm{T}}$ for a vector $\mathbf{v}$, and $\Omega_{\mathcal{A}}(u, y) = \text{corr}\{\mathbf{x}_{\mathcal{A}}, I(Y \le y)|u\}$. Say $a_n - b_n$ uniformly in n if $\liminf_{n\longrightarrow\infty} a_n - b_n > 0$. We consider that given u, $F(y \mid \mathbf{x}, u)$ depends on $\mathbf{x}$ only through $\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}(u)$ for some $|\mathcal{A}| \times K$ matrix $\boldsymbol{\beta}_{\mathcal{A}}(u)$.

The following three conditions are required for Theorem 1.

(A1) $E\{\mathbf{x}|\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}(u), u\} = \text{cov}\{\mathbf{x}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}|u\}\boldsymbol{\beta}_{\mathcal{A}}(u)[\text{cov}\{\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}(u) \mid u\}]^{-1}\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathbf{x}_{\mathcal{A}}$ holds uniformly for u.

(A2) $\min_{k\in\mathcal{A}} \omega_k > E[\lambda_{\max}\{\text{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}^c}^{\mathrm{T}} \mid u)\text{cov}(\mathbf{x}_{\mathcal{A}^c}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}} \mid u)\} \lambda_{\max}\{\Omega_{\mathcal{A}}^2(u, y)\}/\lambda_{\min}^2\{\text{cov}(\mathbf{x}_{\mathcal{A}}, \mathbf{x}_{\mathcal{A}}^{\mathrm{T}} \mid u)\}]$ holds uniformly for n.

(A3) $\mathbf{x}$ and Y are independent conditioning on $\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}(u), u$.

Condition (A1) is referred to as the conditional linearity condition. Condition (A2) is the crucial assumption to guarantee the satisfactory performance of our proposal, which requires the minimal signal of the active predictors not too small. It also does not allow strong correlation between $\mathbf{x}_{\mathcal{A}}$ and $\mathbf{x}_{\mathcal{A}^c}$, or among $\mathbf{x}_{\mathcal{A}}$ themselves given the exposure variable u. Note that (A2) holds automatically if $\mathbf{x}_{\mathcal{A}}$ and $\mathbf{x}_{\mathcal{A}^c}$ are uncorrelated conditioning on u. Similar conditions are assumed in Zhu et al. (2011) and Liu et al. (2014). Condition (A3) dictates that Y relies on $\mathbf{x}$ via the linear combinations $\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}(u)$.

Theorem 1. *Suppose conditions (A1), (A2) and (A3) hold, then we have*

$$\liminf_{n\longrightarrow\infty}\left\{\min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{A}^c}\omega_k\right\} > 0.$$

Theorem 1 illustrates that signals between the important predictors and the unimportant ones are distinguishable, which is a prerequisite for the ranking consistency property. We assume the following regularity conditions to derive the theoretical properties of C-SIRS. Define $E(\cdot \mid u) = g(\cdot \mid u)/f(u)$.

(C1) *(The Kernel Function)* The kernel $K(\cdot)$ is a density function with compact support. It is symmetric about zero and Lipschitz continuous. In addition, it satisfies

$$\int_{-1}^{1} K(t)dt = 1, \quad \int_{-1}^{1} t^{i-1}K(t)dt = 0,\ 0 \le i \le m-1,$$
$$0 \ne \int t^m K(t)dt = \nu_m < \infty$$

It is bounded uniformly such that

$$\sup_{u\in\mathbb{U}}|K(u)| = M_k < \infty.$$

(C2) *(The Density)* The probability density functions of u, denoted by $f(u)$ has continuous second-order derivative on $\mathbb{U}$.

(C3) *(The Derivatives)* The $(m-1)$-th derivatives of both $g(\cdot \mid u)$, $f(u)$ are locally Lipschitz-continuous with respect to u.

(C4) *(The Bandwidth)* The bandwidth h satisfies $h = O(n^{-\theta})$, for some θ which satisfies $(4m)^{-1} < \theta < 1/4$.

(C5) *(The Moment Condition)* There exists a positive constant s_0 such that

$$\sup_{u\in\mathbb{U}}\max_{1\le k\le p} E\{\exp(sX_k^2) \mid u\} < \infty, \text{ for } 0 < s \le s_0.$$

Further assume that $E(X_k|u)$ and $E(X_k^2|u)$, their first-order and second-order derivatives are finite uniformly in $u \in \mathbb{U}$.

Theorem 2. *(Sure Screening Property) Under the conditions (C1)-(C5), for any $0 < \gamma+\theta \le 1/4$ and $0 < \gamma \le m\theta$, if p satisfies $n^3 p\exp(-cn^{1/2-2\gamma-2\theta}) \to 0$ and $\min_{k\in\mathcal{A}}\omega_k \ge 2cn^{-\gamma}$ for some $c > 0$, then*

$$Pr(\mathcal{A} \subseteq \widehat{\mathcal{A}}) \ge 1 - O\left\{n^3|\mathcal{A}|\exp(-cn^{1/2-2\gamma-2\theta})\right\},$$

where c is a generic constant and $|\mathcal{A}|$ is cardinality of $\mathcal{A}$.

The sure screening property (Fan and Lv, 2008) of the C-SIRS procedure ensures that all truly active predictors can be retained after screening with the probability approaching to one.

Theorem 3. *Under conditions (C1)-(C5), in addition to conditions (A1)-(A3), if p satisfies $n^3 p\exp(-cn^{1/2-2\theta}) \to 0$ for some positive c, then*

$$\liminf_{n\longrightarrow\infty}\left\{\min_{k\in\mathcal{A}}\widehat{\omega}_k - \max_{k\in\mathcal{A}^c}\widehat{\omega}_k\right\} > 0 \text{ in probability.}$$

The ranking consistency guarantees the active predictors ranked in the top, prior to the inactive ones, with an overwhelming probability.

4. NUMERICAL STUDIES

4.1 The performance of conditional screening

In this section, we investigate the finite sample performance of our proposed screening procedure through Monte Carlo simulations, and also compare it with five screening methods including SIRS (Zhu et al., 2011), DC-SIS (Li et al., 2012b), DT-SIR (Lin et al., 2018), BCor-SIS (Pan et al., 2018) and CC-SIS (Liu et al., 2014), CDC-SIS (Wen et al., 2018). Under all model settings, we draw $\mathbf{x} = (X_1, X_2, \cdots, X_p)^{\mathrm{T}}$ and an intermediate variable u^* from the multivariate normal distribution with mean zero and AR covariance matrix $\mathbf{\Sigma} = (0.5^{|i-j|})_{(p+1)\times(p+1)}$. Then the exposure variable u is obtained from $u = \Phi(u^*)$, where $\Phi(\cdot)$ is cumulative distribution function of the standard normal distribution $\mathcal{N}(0,1)$. We set the sample size $n = 200$ and fix $p = 1000$. Each experiment is repeated 1000 times. We adopt the Epanechnikov kernel $K(u) = 0.75(1-u^2)_+$ in both simulations and real data analysis.

We evaluate the finite-sample performance through the following four criteria:

1. R_k: The average of the ranks of each important predictor out of 1000 replications.
2. $\mathcal{S}$: The minimum model size to ensure that all important predictors are included after screening. We expect it to be as close as the the number of truly active predictors. We report the 5%, 25%, 50%, 75% and 95% quantiles of $\mathcal{S}$ out of 1000 replications.
3. $\mathcal{P}_a$: The proportion of all active predictors selected after screening for a given model size out of 1000 replications. We consider three screened model size $d = \nu[n^{4/5}/\log(n^{4/5})]$ varying $\nu = 1, 2$ and 3. The corresponding d is 16, 32 and 48, respectively. We expect it to be as close to one as possible.
4. $\mathcal{P}_k$: The proportion of each active predictor selected after screening for a given model size out of 1000 replications.

Example 1. We first generate the response Y from the generalized varying coefficient models respectively:

- *Case 1*: $\text{logit}\{E(Y \mid \mathbf{x}, u)\} = \mathbf{x}^T\boldsymbol{\beta}(u)$;
- *Case 2*: $\log\{E(Y \mid \mathbf{x}, u)\} = \mathbf{x}^T\boldsymbol{\beta}(u)$,

where $\boldsymbol{\beta} = \{\beta_1(u), \ldots, \beta_{1000}(u)\}^T$. β_j is nonzero when X_j is the active predictor and remains zero otherwise. We set he active predictors index to be $\{2, 100, 400, 600, 1000\}$, with corresponding coefficient functions $\beta_2(u) = 2I(u > 0.4)$, $\beta_{100}(u) = 1 + u$, $\beta_{400}(u) = (2-3u)^2, \beta_{600}(u) = 3\sin(2\pi u)$, $\beta_{1000}(u) = \exp\{u/(u+1)\}$. The first model is logistic varying coefficient model while the second one is the Poisson varying coefficient model.

Table 1. The mean of R_k of each active predictor for Example 1

	Method	R_2	R_{100}	R_{400}	R_{600}	R_{1000}
	SIRS	30.89	8.22	109.32	491.70	18.37
	BCor-SIS	38.26	11.84	96.90	504.99	23.30
	DT-SIR	187.60	85.87	259.61	497.15	115.21
Case1	DC-SIS	18.11	6.09	67.89	512.58	10.73
	CDC-SIS	18.25	4.34	61.82	457.82	10.26
	CC-SIS	9.85	9.63	7.63	2.49	17.68
	C-SIRS	10.41	7.15	5.69	3.02	14.34
	SIRS	7.05	2.44	123.44	462.06	3.44
	BCor-SIS	11.64	3.34	38.20	250.44	6.18
	DT-SIR	89.67	22.00	158.81	478.59	37.37
Case2	DC-SIS	17.53	3.66	163.17	164.99	6.36
	CDC-SIS	104.78	76.95	203.15	142.34	107.58
	CC-SIS	21.81	10.00	65.09	1.56	19.07
	C-SIRS	3.96	4.67	35.12	1.08	6.75

Table 2. The quantiles of the minimum model size $\mathcal{S}$ for Example 1

	Method	5%	25%	50%	75%	95%
	SIRS	77.00	277.75	513.00	753.00	953.00
	BCor-SIS	80.85	297.00	532.50	756.00	953.05
	DT-SIR	206.85	436.00	642.00	824.25	957.05
Case1	DC-SIS	83.00	285.75	526.00	773.00	959.00
	CDC-SIS	75.95	245.75	463.00	691.25	898.00
	CC-SIS	5.00	7.00	12.00	27.00	137.05
	C-SIRS	5.00	6.00	10.50	22.00	96.05
	SIRS	75.90	255.75	488.00	725.75	942.05
	BCor-SIS	17.95	81.00	192.50	396.00	765.30
	DT-SIR	104.00	308.75	514.00	762.00	962.00
Case2	DC-SIS	33.00	93.75	206.00	389.25	798.05
	CDC-SIS	69.00	191.75	336.50	547.25	810.20
	CC-SIS	7.00	19.00	45.00	116.00	345.10
	C-SIRS	5.00	7.00	12.00	32.25	169.15

We report the simulation results of R_k and $\mathcal{S}$ in Table 1 and Table 2, respectively. The proposed C-SIRS method outperforms other competitors under both model settings. The rank of each predictor is on the top while the median of $\mathcal{S}$ is close to the number of truly active predictors, indicating that our proposal achieves a high accuracy in ranking. CC-SIS also performs satisfactorily as it is designed for the varying coefficient model. The marginal screening methods, including SIRS, BCor-SIS, DT-SIR and DC-SIS, are not able to detect the predictor X_{600}. This is mainly because the expectation of $\beta_{600}(u)$ is zero, making the predictor X_{600} marginally independent but conditional related to the response. The simulation results of selection proportions $\mathcal{P}_a$ and $\mathcal{P}_k$ are summarized in Table 3. The C-SIRS method selects all important predictors with high probability, indicating its sure screening property. The $\mathcal{P}_a$ of SIRS, BCor-SIS, DT-SIR, DC-SIS and CDC-SIS methods are negligible even for the largest submodel size $d = 48$.

Table 3. The proportions of $\mathcal{P}_a$ and $\mathcal{P}_k$ given the model size d for Example 1

	d	Method	$\mathcal{P}_2$	$\mathcal{P}_{100}$	$\mathcal{P}_{400}$	$\mathcal{P}_{600}$	$\mathcal{P}_{1000}$	$\mathcal{P}_a$
		SIRS	0.71	0.91	0.38	0.01	0.81	0.00
		BCor-SIS	0.67	0.89	0.44	0.02	0.78	0.01
		DT-SIR	0.23	0.45	0.13	0.02	0.38	0.00
	16	DC-SIS	0.81	0.93	0.52	0.01	0.87	0.00
		CDC-SIS	0.82	0.96	0.56	0.01	0.90	0.00
		CC-SIS	0.90	0.91	0.91	0.99	0.83	0.60
		C-SIRS	0.90	0.93	0.95	0.99	0.84	0.65
		SIRS	0.79	0.95	0.52	0.03	0.88	0.01
		BCor-SIS	0.78	0.94	0.57	0.03	0.85	0.02
		DT-SIR	0.35	0.57	0.19	0.03	0.49	0.00
Case1	32	DC-SIS	0.90	0.97	0.64	0.03	0.93	0.01
		CDC-SIS	0.88	0.98	0.67	0.02	0.94	0.01
		CC-SIS	0.95	0.95	0.96	1.00	0.89	0.77
		C-SIRS	0.96	0.96	0.98	1.00	0.91	0.81
		SIRS	0.85	0.97	0.59	0.04	0.91	0.02
		BCor-SIS	0.82	0.96	0.63	0.05	0.88	0.03
		DT-SIR	0.40	0.62	0.24	0.05	0.55	0.00
	48	DC-SIS	0.91	0.98	0.72	0.05	0.95	0.03
		CDC-SIS	0.91	0.99	0.75	0.04	0.96	0.02
		CC-SIS	0.97	0.97	0.98	1.00	0.92	0.84
		C-SIRS	0.97	0.97	0.99	1.00	0.94	0.87
		SIRS	0.93	0.99	0.33	0.02	0.98	0.00
		BCor-SIS	0.89	0.97	0.67	0.08	0.95	0.04
		DT-SIR	0.46	0.77	0.25	0.02	0.68	0.00
	16	DC-SIS	0.80	0.98	0.28	0.15	0.94	0.02
		CDC-SIS	0.41	0.52	0.23	0.29	0.42	0.00
		CC-SIS	0.75	0.87	0.45	0.99	0.79	0.19
		C-SIRS	0.98	0.98	0.66	1.00	0.94	0.59
		SIRS	0.97	0.99	0.47	0.04	0.99	0.02
		BCor-SIS	0.94	0.99	0.77	0.15	0.97	0.10
		DT-SIR	0.56	0.85	0.36	0.04	0.78	0.00
Case2	32	DC-SIS	0.88	0.99	0.40	0.24	0.97	0.05
		CDC-SIS	0.52	0.61	0.31	0.41	0.51	0.01
		CC-SIS	0.86	0.93	0.60	1.00	0.88	0.41
		C-SIRS	0.99	0.99	0.78	1.00	0.97	0.74
		SIRS	0.98	1.00	0.56	0.06	0.99	0.03
		BCor-SIS	0.96	0.99	0.83	0.21	0.98	0.16
		DT-SIR	0.63	0.89	0.43	0.05	0.81	0.01
	48	DC-SIS	0.92	0.99	0.47	0.33	0.98	0.11
		CDC-SIS	0.59	0.67	0.36	0.47	0.57	0.02
		CC-SIS	0.91	0.95	0.68	1.00	0.92	0.52
		C-SIRS	1.00	1.00	0.83	1.00	0.99	0.82

Example 2. We then consider following four models, in which the response Y depends on the predictors nonlinearly with a given u:

- *Case 1*: $Y = \exp\{\mathbf{x}^T\boldsymbol{\beta}(u)\} + \varepsilon$;
- *Case 2*: $Y = \exp\{\mathbf{x}^T\boldsymbol{\beta}(u) + \varepsilon\}$;
- *Case 3*: $Y = \beta_2(u)\exp(X_2) + \beta_{100}(u)X_{100}^3 + 2\beta_{400}(u)X_{400}I(X_{400} < 2) + \beta_{600}(u)X_{600} + 1.5\beta_{1000}(u)X_{1000} + \varepsilon$;
- *Case 4*: $Y = \beta_2(u)X_2 + \beta_{100}(u)X_{100} + \beta_{400}(u)X_{400} + \beta_{1000}(u)X_{1000} + 2\exp\{\beta_{600}(u)X_{600}\}\varepsilon$,

where the $\boldsymbol{\beta}(u)$ setting remains identical as Example 1 and the error term is independently generated from $t(1)$ for Case 1 to Case 3 and $\mathcal{N}(0,1)$ for Case 4. Notice that Case 4 demonstrates the heteroscedasticity issue, where variance component is affected by X_{600}, and the effect vary with u.

Table 4. The mean of R_k of each active predictor for Example 2

	Method	R_2	R_{100}	R_{400}	R_{600}	R_{1000}
	SIRS	24.64	2.76	61.90	485.22	9.39
	BCor-SIS	25.98	10.82	82.19	233.65	20.97
	DT-SIR	121.18	40.52	198.86	510.08	60.21
Case1	DC-SIS	104.67	65.33	277.81	171.04	146.13
	CDC-SIS	108.22	79.91	206.46	143.12	111.44
	CC-SIS	188.26	167.32	131.75	56.50	169.04
	C-SIRS	5.69	8.36	10.61	1.81	6.97
	SIRS	22.78	3.29	53.03	478.24	7.70
	BCor-SIS	38.16	15.31	96.98	226.20	29.06
	DT-SIR	153.55	61.14	225.41	495.37	99.94
Case2	DC-SIS	362.10	520.40	560.70	579.80	538.20
	CDC-SIS	497.71	484.76	506.99	500.14	487.19
	CC-SIS	231.46	197.64	143.85	71.32	188.68
	C-SIRS	9.63	6.71	18.70	2.39	12.31
	SIRS	9.39	1.24	31.41	439.37	7.57
	BCor-SIS	13.73	2.56	33.34	312.92	22.19
	DT-SIR	84.96	9.54	157.90	462.88	84.07
Case3	DC-SIS	20.31	8.49	52.85	407.09	31.37
	CDC-SIS	18.45	7.61	50.79	389.78	29.05
	CC-SIS	123.35	34.00	151.27	197.23	177.44
	C-SIRS	6.49	1.34	8.67	19.35	8.22
	SIRS	21.89	4.01	46.42	457.25	6.79
	BCor-SIS	36.70	8.54	82.43	9.26	22.08
	DT-SIR	192.56	102.51	259.23	512.11	130.13
Case4	DC-SIS	226.48	172.32	295.15	5.51	202.70
	CDC-SIS	234.92	175.71	318.88	4.99	204.87
	CC-SIS	452.48	465.33	486.95	74.69	469.96
	C-SIRS	7.99	7.39	17.09	8.78	14.56

The simulation results are tabulated in Table 4 for R_k, Table 5 for $\mathcal{S}$ and Table 6, 7 for $\mathcal{P}_a$ and $\mathcal{P}_k$. It can be clearly seen that C-SIRS method is still the obvious winner, and performs well in both ranking and selection. CC-SIS method is not able to capture the nonlinear conditional dependence, and the rankings are not accurate. The performances of DC-SIS and CDC-SIS methods are still poor, because of the heavy distribution of the response. We observe that SIRS, BCor-SIS and DT-SIR methods also fail to identify the X_{600} in most cases, which is in accordance with Example 1.

Table 8 reports the average computation time, in seconds, along with the standard deviation (SD). All simulations have been carried out on a laptop computer with Intel Xeon 2.60 GHz processor. As a screening procedure involving exposure variables, we have to utilize smoothing techniques to estimate the conditional correlations, making it slower than SIRS, DC-SIS, DT-SIR, among other methods that do not need to do smoothing (and thus cannot deal with the exposure variables). However, it is still reasonably fast compared

Table 5. The quantiles of the minimum model size $\mathcal{S}$ for Example 2

	Method	5%	25%	50%	75%	95%
Case1	SIRS	68.95	259.75	473.50	728.00	933.10
	BCor-SIS	22.00	92.00	226.50	434.50	777.20
	DT-SIR	151.00	389.00	607.00	805.00	955.05
	DC-SIS	82.00	196.75	359.00	583.25	844.05
	CDC-SIS	72.95	194.00	353.00	550.00	822.05
	CC-SIS	65.95	185.00	328.00	568.75	865.00
	C-SIRS	5.00	6.00	9.00	21.00	92.00
Case2	SIRS	95.00	203.15	495.20	702.00	925.45
	BCor-SIS	31.00	98.00	232.50	448.00	796.05
	DT-SIR	183.90	395.75	605.00	798.00	952.05
	DC-SIS	42.75	126.25	251.50	505.20	879.55
	CDC-SIS	563.95	762.75	872.00	946.00	989.00
	CC-SIS	45.00	162.00	346.00	578.95	882.00
	C-SIRS	5.00	7.50	13.25	42.50	199.65
Case3	SIRS	43.90	194.00	392.00	667.00	920.10
	BCor-SIS	24.95	103.75	256.00	509.25	875.20
	DT-SIR	116.00	319.00	527.00	729.00	939.05
	DC-SIS	50.00	178.75	369.50	633.75	911.05
	CDC-SIS	42.00	174.00	376.00	589.50	892.05
	CC-SIS	19.00	130.00	353.50	645.00	918.25
	C-SIRS	5.00	6.00	10.00	23.00	114.00
Case4	SIRS	59.95	237.00	436.50	689.25	933.00
	BCor-SIS	7.00	21.00	53.00	156.25	473.00
	DT-SIR	222.95	471.00	660.00	840.25	965.00
	DC-SIS	68.95	228.75	407.00	617.00	842.10
	CDC-SIS	98.00	267.75	442.50	629.00	844.05
	CC-SIS	432.90	673.75	811.50	918.00	984.00
	C-SIRS	6.00	9.00	16.00	36.00	154.00

Table 6. The proportions of $\mathcal{P}_a$ and $\mathcal{P}_k$ given the model size d for Example 2

	d	Method	$\mathcal{P}_2$	$\mathcal{P}_{100}$	$\mathcal{P}_{400}$	$\mathcal{P}_{600}$	$\mathcal{P}_{1000}$	$\mathcal{P}_a$
Case1	16	SIRS	0.87	0.97	0.63	0.01	0.93	0.00
		BCor-SIS	0.75	0.90	0.48	0.14	0.81	0.03
		DT-SIR	0.35	0.641	0.19	0.03	0.57	0.00
		DC-SIS	0.40	0.54	0.23	0.24	0.34	0.00
		CDC-SIS	0.40	0.51	0.23	0.29	0.41	0.00
		CC-SIS	0.18	0.28	0.32	0.63	0.22	0.00
		C-SIRS	0.92	0.97	0.90	1.00	0.91	0.72
	32	SIRS	0.92	0.99	0.76	0.03	0.96	0.02
		BCor-SIS	0.84	0.94	0.59	0.23	0.87	0.08
		DT-SIR	0.46	0.74	0.28	0.05	0.67	0.00
		DC-SIS	0.50	0.64	0.30	0.33	0.45	0.00
		CDC-SIS	0.52	0.60	0.30	0.41	0.50	0.01
		CC-SIS	0.28	0.38	0.48	0.75	0.33	0.01
		C-SIRS	0.98	0.99	0.97	1.00	0.96	0.90
	48	SIRS	0.93	0.99	0.83	0.05	0.99	0.03
		BCor-SIS	0.89	0.96	0.66	0.29	0.91	0.14
		DT-SIR	0.52	0.79	0.35	0.06	0.72	0.00
		DC-SIS	0.61	0.68	0.33	0.38	0.51	0.01
		CDC-SIS	0.58	0.66	0.35	0.46	0.57	0.02
		CC-SIS	0.37	0.44	0.53	0.78	0.44	0.01
		C-SIRS	1.00	0.99	0.99	1.00	0.98	0.96
Case2	16	SIRS	0.82	0.97	0.61	0.02	0.92	0.01
		BCor-SIS	0.72	0.86	0.46	0.12	0.77	0.02
		DT-SIR	0.27	0.55	0.16	0.02	0.42	0.00
		DC-SIS	0.38	0.53	0.26	0.28	0.31	0.00
		CDC-SIS	0.01	0.02	0.01	0.03	0.02	0.00
		CC-SIS	0.16	0.23	0.34	0.59	0.21	0.01
		C-SIRS	0.92	0.93	0.81	0.99	0.86	0.58
	32	SIRS	0.89	0.99	0.72	0.05	0.96	0.03
		BCor-SIS	0.80	0.91	0.57	0.20	0.83	0.06
		DT-SIR	0.38	0.66	0.24	0.04	0.53	0.00
		DC-SIS	0.51	0.65	0.32	0.36	0.43	0.01
		CDC-SIS	0.03	0.04	0.03	0.05	0.04	0.00
		CC-SIS	0.23	0.36	0.45	0.71	0.34	0.01
		C-SIRS	0.95	0.97	0.89	1.00	0.92	0.74
	48	SIRS	0.92	0.99	0.79	0.07	0.97	0.06
		BCor-SIS	0.83	0.93	0.64	0.28	0.87	0.10
		DT-SIR	0.46	0.73	0.30	0.07	0.61	0.00
		DC-SIS	0.64	0.69	0.35	0.37	0.56	0.01
		CDC-SIS	0.05	0.06	0.04	0.07	0.05	0.00
		CC-SIS	0.39	0.43	0.56	0.81	0.48	0.01
		C-SIRS	0.97	0.98	0.92	1.00	0.95	0.83

with other conditional screening methods according to the Table 8.

In short, thanks to the model-free start point of C-SIRS, it is applicable for both generalized varying coefficient models and any nonlinear dependence structure of the predictor effects on the exposure variables, as shown in Example 1 and Case 1 and 3 of Example 2. As it does not impose strong distributional assumptions on the error term, the heavy-tailed error distribution and outlier issues can be well addressed in Case 1, 2 and 3 of Example 2 when ε is generated from $t(1)$. Furthermore, as illustrated in Case 4 of Example 2, it is also powerful to detect significant predictor effects on the variance of response, in addition to the mean, since we utilize the entire distribution function of the response to construct the screening score.

4.2 Real data analysis

Breast cancer is the one of most common malignancy among women, with a high lethality rate. There is a urgent need for the early diagnosis of breast cancer and simultaneously monitoring the disease progression. In this section, we evaluate the performance of proposed C-SIRS procedure through the breast cancer data reported by Chin et al. (2006). The dataset, available from the R package PMA, consists of $p = 19672$ gene expressions over $n = 88$ cancer patient samples. Our interest is to detect the most influential genes to the 136 comparative genomic hybridization (CGH) measurements, as in some previous research (Witten et al., 2009; Wen et al., 2018). CGH is known as an indicator of the genome copy number variation and cancer diagnosis. In addition, recent researches (McPherson et al., 2000; DeSantis et al., 2017) found that age is a risk factor that influences the incidence of breast cancer. And the effects of gene expression levels to CGH might depend on age. However, as a typical challenge of most ultrahigh dimensional problems, it

Table 7. The proportions of $\mathcal{P}_a$ and $\mathcal{P}_k$ given the model size d for Example 2

	d	Method	$\mathcal{P}_2$	$\mathcal{P}_{100}$	$\mathcal{P}_{400}$	$\mathcal{P}_{600}$	$\mathcal{P}_{1000}$	$\mathcal{P}_a$
		SIRS	0.89	1.00	0.72	0.01	0.93	0.01
		BCor-SIS	0.86	0.99	0.72	0.08	0.80	0.03
		DT-SIR	0.49	0.91	0.25	0.02	0.46	0.00
	16	DC-SIS	0.84	0.95	0.65	0.01	0.76	0.01
		CDC-SIS	0.85	0.96	0.65	0.01	0.77	0.01
		CC-SIS	0.50	0.84	0.42	0.23	0.28	0.03
		C-SIRS	0.93	1.00	0.92	0.78	0.93	0.63
		SIRS	0.94	1.00	0.81	0.04	0.96	0.03
		BCor-SIS	0.92	0.99	0.80	0.12	0.86	0.07
		DT-SIR	0.59	0.94	0.35	0.04	0.58	0.00
Case3	32	DC-SIS	0.90	0.96	0.75	0.04	0.85	0.03
		CDC-SIS	0.90	0.97	0.73	0.05	0.85	0.03
		CC-SIS	0.58	0.86	0.51	0.32	0.40	0.09
		C-SIRS	0.97	1.00	0.96	0.88	0.97	0.79
		SIRS	0.96	1.00	0.86	0.07	0.97	0.06
		BCor-SIS	0.94	0.99	0.86	0.17	0.89	0.12
		DT-SIR	0.65	0.95	0.44	0.07	0.65	0.01
	48	DC-SIS	0.92	0.97	0.79	0.06	0.88	0.05
		CDC-SIS	0.93	0.98	0.79	0.08	0.87	0.06
		CC-SIS	0.62	0.88	0.55	0.38	0.46	0.12
		C-SIRS	0.98	1.00	0.97	0.91	0.98	0.85
		SIRS	0.78	0.97	0.58	0.02	0.93	0.01
		BCor-SIS	0.69	0.89	0.50	0.88	0.79	0.20
		DT-SIR	0.19	0.40	0.11	0.02	0.33	0.00
	16	DC-SIS	0.13	0.22	0.08	0.92	0.15	0.00
		CDC-SIS	0.12	0.20	0.05	0.95	0.16	0.00
		CC-SIS	0.03	0.02	0.02	0.59	0.02	0.00
		C-SIRS	0.91	0.95	0.80	0.87	0.87	0.50
		SIRS	0.87	0.98	0.72	0.04	0.96	0.02
		BCor-SIS	0.79	0.94	0.62	0.94	0.87	0.37
		DT-SIR	0.29	0.52	0.19	0.04	0.43	0.00
Case4	32	DC-SIS	0.20	0.32	0.13	0.97	0.24	0.01
		CDC-SIS	0.20	0.29	0.10	0.98	0.23	0.01
		CC-SIS	0.05	0.05	0.04	0.68	0.05	0.00
		C-SIRS	0.95	0.97	0.88	0.96	0.92	0.72
		SIRS	0.90	0.99	0.78	0.06	0.97	0.04
		BCor-SIS	0.85	0.96	0.67	0.96	0.91	0.48
		DT-SIR	0.36	0.58	0.25	0.06	0.50	0.00
	48	DC-SIS	0.26	0.40	0.16	0.98	0.31	0.03
		CDC-SIS	0.26	0.35	0.14	0.99	0.28	0.01
		CC-SIS	0.07	0.07	0.06	0.73	0.06	0.00
		C-SIRS	0.97	0.97	0.91	0.99	0.94	0.81

Table 8. Run times (CPU seconds) for each experiment

Method	Mean	SD
SIRS	0.052	0.008
BCor-SIS	4.076	0.020
DT-SIR	0.085	0.008
DC-SIS	3.038	0.060
CDC-SIS	68.498	0.487
CC-SIS	0.344	0.007
C-SIRS	5.894	0.043

Table 9. Results for breast cancer analysis

Method	Size	MSPE
SIRS	10	3.694
BCor-SIS	15	4.378
DT-SIR	12	4.202
DC-SIS	9	4.837
CDC-SIS	9	3.626
CC-SIS	10	4.067
C-SIRS	9	3.533

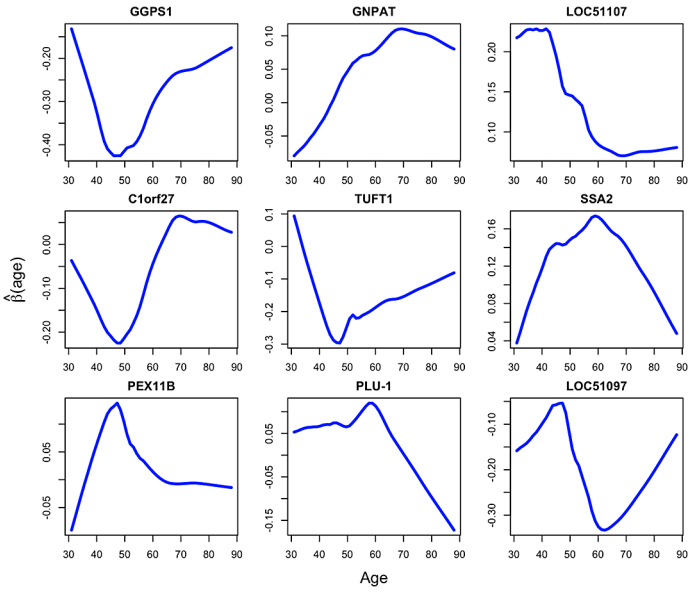

Figure 1. The estimated coefficient functions of the selected nine genes.

is hard to determine the functional dependent-forms. Thus, treating age as the exposure variable u, we apply the C-SIRS method as a model-free screening technique. Furthermore, we obtain the first principal component of 136 CGH measurements as the response, and the 19672 gene expressions as predictors.

In our analysis, we first use the seven screening methods SIRS, BCor-SIS, DT-SIR, DC-SIS, CDC-SIS, CC-SIS and the proposed C-SIRS to select the most of relevant genes with top $d = 2[n^{4/5}/\log(n^{4/5})] = 20$ included, and then conduct a second-stage selection to fit a varying coefficient model with SCAD penalty. Table 9 summarizes the number of genes selected by each method (Size) and mean squared prediction error (MSPE) by five-fold cross-validation. The proposed C-SIRS achieves the best performance with the sparsest model size 9 and the smallest MSPE 3.533, and five of nine identified genes by C-SIRS are consistent with Wen et al. (2018). We also show the estimated coefficient functions of the selected nine genes by C-SIRS in Figure 1, which imply the age-dependent effects of genes. For instance, the first selected gene GGPS1 has a small negative effect on CGH when people are in their 30s, and the negative effect

becomes most severe at about 50s, and recovers gradually thereafter. The third selected gene LOC51107 has a large positive effect at people's early stage, but gets less important as people get order. Other functions can be interpreted in the similar fashion.

5. A BRIEF DISCUSSION

In this paper, we utilize the conditional correlation between predictors and the empirical distribution function of response given exposure variables to develop a model-free sure independence screening procedure C-SIRS. It is resistant to the heavy-tailed distribution, outliers and extreme values of the response. In addition, C-SIRS is applicable for any model involving exposure variables, including generalized varying coefficient model, and any nonlinear dependent structure. It is also powerful to detect those variables with significant effects on variance components. The ranking consistency and sure screening property of C-SIRS are rendered in both theoretical properties and simulation studies. The breast cancer dataset is systematically analyzed by C-SIRS, along with the comparison with other related methods.

6. APPENDIX: PROOF OF THEOREMS

6.1 Appendix A: Proof of Theorem 1

For easy presentation, we introduce some notations first. $\lambda_{\max}(\mathbf{C})$ and $\lambda_{\min}(\mathbf{C})$ denote the largest and smallest eigenvalues of a matrix $\mathbf{C}$. We denote $\mathbf{v}\mathbf{v}^{\mathrm{T}}$ by $\mathbf{v}^2$ for a vector $\mathbf{v}$ and $\Omega_{\mathcal{A}}(u,y)=\mathrm{corr}\{\mathbf{x}_{\mathcal{A}}, I(Y\le y)|u\}$. If we say that a_n-b_n uniformly in n, then it means $\liminf_{n\longrightarrow\infty} a_n-b_n>0$.

Without loss of generality, we assume that $E(X_k\mid u)=0$, $\mathrm{var}(X_k\mid u)=1$ for $k=1,\cdots,p$ and $\mathrm{var}\{I(Y\le y)\mid u\}=1$. Meanwhile, $\boldsymbol{\beta}_{\mathcal{A}}(u)$ satisfies $\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathrm{cov}(\mathbf{x}_{\mathcal{A}}\mid u)\boldsymbol{\beta}_{\mathcal{A}}(u)=1$. Then the conditional linearity condition (A1) is simplified as $E\{X_k\mid \mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}(u),u\}=\mathrm{cov}\{X_k,\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}(u)\mid u\}\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathbf{x}_{\mathcal{A}}$. Following law of iterated expectations,

$$
\begin{aligned}
&E\{X_kI(Y\le y)|u,y\}\\
&\quad=E\left[E\left\{X_kI(Y\le y)|\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathbf{x}_{\mathcal{A}},u,y\right\}|u,y\right]\\
&\quad=E\left[E\left\{X_k\mid\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathbf{x}_{\mathcal{A}},u\right\}\right.\\
&\qquad\left.\cdot E\left\{I(Y\le y)\mid\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathbf{x}_{\mathcal{A}},u,y\right\}\mid u,y\right]\\
&\quad=\mathrm{cov}\left\{X_k,\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}(u)\mid u\right\}E\left\{\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathbf{x}_{\mathcal{A}}I(Y<y)\mid u,y\right\},
\end{aligned}
$$

where the second equation holds because of the conditional independence of $\mathbf{x}$ and Y given $\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathbf{x}_{\mathcal{A}}$ and u in (A3) and the third equation is due to the simplified conditional linearity condition.

$$
\begin{aligned}
(2)\qquad \max_{k\in\mathcal{A}^c}\Omega_k^2(u,y)&=\max_{k\in\mathcal{A}^c}E^2\{X_kI(Y\le y)\mid u,y\}\\
&=\max_{k\in\mathcal{A}^c}\left[\mathrm{cov}^2\left\{X_k,\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}(u)\mid u\right\}\right]\\
&\quad\cdot E^2\left\{\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathbf{x}_{\mathcal{A}}I(Y\le y)\mid u,y\right\}.
\end{aligned}
$$

Then we start to deal with the first term of (2).

$$
\begin{aligned}
&(3)\\
&\max_{k\in\mathcal{A}^c}\left[\mathrm{cov}^2\left\{X_k,\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\boldsymbol{\beta}_{\mathcal{A}}(u)\mid u\right\}\right]\\
&\quad=\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\left[\max_{k\in\mathcal{A}^c}\{\mathrm{cov}\left(\mathbf{x}_{\mathcal{A}},X_k\mid u\right)\right.\\
&\qquad\left.\cdot\,\mathrm{cov}(X_k,\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\}\right]\boldsymbol{\beta}_{\mathcal{A}}(u)\\
&\quad\le\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathrm{cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}^c}^{\mathrm{T}}\mid u)\mathrm{cov}(\mathbf{x}_{\mathcal{A}^c},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\boldsymbol{\beta}_{\mathcal{A}}(u)\\
&\quad=\left\{\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathrm{cov}^{1/2}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\right\}\mathrm{cov}^{-1/2}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\\
&\qquad\cdot\,\mathrm{cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}^c}^{\mathrm{T}}\mid u)\mathrm{cov}(\mathbf{x}_{\mathcal{A}^c},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\mathrm{cov}^{-1/2}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\\
&\qquad\cdot\left\{\mathrm{cov}^{1/2}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\boldsymbol{\beta}_{\mathcal{A}}(u)\right\}\\
&\quad\le\lambda_{\max}\left\{\mathrm{cov}^{-1/2}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\mathrm{cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}^c}^{\mathrm{T}}\mid u)\right.\\
&\qquad\left.\cdot\,\mathrm{cov}(\mathbf{x}_{\mathcal{A}^c},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\mathrm{cov}^{-1/2}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\right\}\\
&\quad\le\lambda_{\max}\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}^c}^{\mathrm{T}}\mid u)\mathrm{cov}(\mathbf{x}_{\mathcal{A}^c},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\right\}\\
&\qquad\cdot\,\lambda_{\max}\left\{\mathrm{cov}^{-1}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\right\}\\
&\quad=\lambda_{\max}\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}^c}^{\mathrm{T}}\mid u)\mathrm{cov}(\mathbf{x}_{\mathcal{A}^c},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\right\}\\
&\qquad/\lambda_{\min}\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\right\},
\end{aligned}
$$

where (3) holds because of the fact $\lambda_{\max}(\mathbf{C}^{\mathrm{T}}\mathbf{B}\mathbf{C})\le\lambda_{\max}(\mathbf{B})\lambda_{\max}(\mathbf{C}^{\mathrm{T}}\mathbf{C})$, where the matrix $\mathbf{B}\ge 0$. Similarly, we can verify that

$$
\begin{aligned}
&E^2\left\{\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}(u)\mathbf{x}_{\mathcal{A}}I(Y\le y)\mid u,y\right\}\\
&\quad\le\lambda_{\max}\left[E^2\left\{\mathbf{x}_{\mathcal{A}}I(Y\le y)\mid u,y\right\}\right]\\
&\qquad/\lambda_{\min}\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\right\}.
\end{aligned}
$$

These two inequations yield that

$$
\begin{aligned}
\max_{k\in\mathcal{A}^c}\omega_k&=\max_{k\in\mathcal{A}^c}E\{\Omega_k^2(y,u)\}\\
&\le E\{\max_{k\in\mathcal{A}^c}\ \Omega_k^2(y,u)\}\\
&\le E\left[\lambda_{\max}\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}^c}^{\mathrm{T}}\mid u)\mathrm{cov}(\mathbf{x}_{\mathcal{A}^c},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\right\}\right.\\
&\qquad\left.\cdot\,\lambda_{\max}\{\Omega_{\mathcal{A}}^2(y,u)\}/\lambda_{\min}^2\left\{\mathrm{cov}(\mathbf{x}_{\mathcal{A}},\mathbf{x}_{\mathcal{A}}^{\mathrm{T}}\mid u)\right\}\right],
\end{aligned}
$$

which completes the proof. □

6.2 Appendix A: Proof of Theorem 2

Define $E(\cdot\mid u)=g(\cdot\mid u)/f(u)$, and

$$
\begin{aligned}
\widehat{E}(X_k\mid u)&=\widehat{g}(X_k\mid u)/\widehat{f}(u),\\
\widehat{E}\{I(Y\le y)\mid u\}&=\widehat{g}(y\mid u)/\widehat{f}(u),
\end{aligned}
$$

where $\widehat{g}(X_k \mid u) = n^{-1}\sum_{i=1}^n K_h(u_i-u)X_{ik}$, $\widehat{g}(y \mid u) = n^{-1}\sum_{i=1}^n K_h(u_i-u)I(Y_i \le y)$, $\widehat{f}(u) = n^{-1}\sum_{i=1}^n K_h(u_i-u)$ and $K_h(u) = K(u/h)/h$.

Lemma 1. *(Hoeffding's Inequality (Hoeffding, 1963)) Let $X_1,\ldots,X_n$ be independent random variables, and $Pr(a_i \le X_i \le b_i) = 1$ for $i = 1,\ldots,n$. Then for any $t > 0$,*

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

Lemma 2. *(Liu et al., 2014) Suppose that X is a random variable with $E(e^{a|X|}) < \infty$ for some $a > 0$. Then for any $M > 0$, there exist positive constants b and c such that*

$$Pr(|X| \ge M) \le b\exp(-cM).$$

The following lemma is a slight modified version of Theorem 3.1 of Zhu (1993) and Lemma 3.2 of Zhu and Fang (1996).

Lemma 3. *Suppose that conditions (C1) to (C5) are fulfilled, and $\sup_{\mathbf{x}} |X_k| \le M$, then for any $\varepsilon_n > 0$,*

$$\begin{aligned} &Pr\left\{\sup_u |\widehat{g}(X_k \mid u) - E\widehat{g}(X_k \mid u)| \ge 8MM_Kh^{-1}n^{-1/2}\varepsilon_n\right\} \\ &\quad \le 3c\left(n^{1/2}/\varepsilon_n\right)^4 \exp\left(-\varepsilon_n^2/128\delta^2\right) + 4c\delta^{-8}\exp\left(-n\delta^2\right), \end{aligned}$$

where $\delta \ge \sup_u[\mathrm{var}\{K(\frac{\widetilde{u}-u}{h})\}]^{1/2}$.

Lemma 4. *Suppose that conditions (C1) to (C5) are fulfilled, for any $0 < \gamma + \theta < 1/4$ and $0 < \gamma \le m\theta$, then we have*

$$\begin{aligned} &Pr\left\{\sup_u |\widehat{g}(X_k \mid u) - g(X_k \mid u)| \ge cn^{-\gamma}\right\} \\ &\quad \le a_1n^2\exp(-b_1n^{1/2-2\theta-2\gamma}), \end{aligned}$$

where a_1, b_1 and c are some positive constants.

Proof of Lemma 4. For any $M > 0$,

$$\begin{aligned} &\Pr\left\{\sup_u |\widehat{g}(X_k \mid u) - g(X_k \mid u)| \ge cn^{-\gamma}\right\} \\ &= \left\{\sup_u |\widehat{g}(X_k \mid u) - g(X_k \mid u)| \ge cn^{-\gamma}, \max|X_{ik}| \le M\right\} \\ &\quad + \Pr\left\{\sup_u |\widehat{g}(X_k \mid u) - g(X_k \mid u)| \ge cn^{-\gamma}, \max|X_{ik}| \ge M\right\} \\ &\le \Pr\left\{\sup_u |\widehat{g}(X_k \mid u) - g(X_k \mid u)| \ge cn^{-\gamma}, \max|X_{ik}| \le M\right\} \\ &\quad + \Pr\{\max|X_{ik}| \ge M \text{ for some } i\} = P_1(u) + P_2, \end{aligned}$$

where $P_1(u)$ and P_2 are defined in the obvious way. We derive the first term $P_1(u)$. The following arguments are all under the condition $\max|X_{ik}| \le M$, which is omitted for notation simplicity. We expand $g(X_k \mid u)$ in a Taylor series with the Lagrange remainder term under Condition (C3). There exists a positive constant c_1 such that

$$\begin{aligned} &\sup_u |E\widehat{g}(X_k \mid u) - g(X_k \mid u)| \\ &= \sup_u \left| \int K_h(\widetilde{u}-u)\{E(X_k \mid \widetilde{u})f(\widetilde{u}) - E(X_k \mid u)f(u)\}d\widetilde{u} \right| \\ &= \sup_u \Big| \int K(t)\{E(X_k \mid u+ht)f(u+ht) \\ &\qquad - E(X_k \mid u)f(u)\}\,dt \Big| \le c_1\nu_m h^m. \end{aligned}$$

Since the kernel function K is uniformly continuous on its compact support, we have

$$\begin{aligned} &\sup_u \left[\mathrm{var}\left\{K\left(\frac{\widetilde{u}-u}{h}\right)\right\}\right]^{1/2} \\ &\quad \le \sup_u \left\{\int K^2\left(\frac{\widetilde{u}-u}{h}\right) f(\widetilde{u})d\widetilde{u}\right\}^{1/2} \le M_K \end{aligned}$$

Following the Lemma 3, for $0 < \gamma \le m\theta$,

(4)
$$\begin{aligned} &\Pr\left\{\sup_u |\widehat{g}(X_k \mid u) - g(X_k \mid u)| \ge cn^{-\gamma}, \max|X_{ik}| \le M\right\} \\ &= \Pr\Big\{\sup_u \big|\widehat{g}(X_k \mid u) - E\widehat{g}(X_k \mid u) + E\widehat{g}(X_k \mid u) \\ &\qquad - g(X_k \mid u)\big| \ge cn^{-\gamma}, \max|X_{ik}| \le M\Big\} \\ &\le \Pr\Big\{\sup_u \big|\widehat{g}(X_k \mid u) - E\widehat{g}(X_k \mid u)\big| \\ &\qquad \ge cn^{-\gamma}/2, \max|X_{ik}| \le M\Big\} \\ &= O\left\{n^2\exp(-b_1n^{1-2\theta-2\gamma}/M^2)\right\}, \end{aligned}$$

for some positive constant b_1.

Then we deal with $\Pr\{\max|X_{ik}| \ge M \text{ for some } i\}$. Following the assumption (C5) and Lemma 2, there exist some positive constants t_1 and t_2 such that for any $M > 0$, $\Pr(|X_k| \ge M) \le t_1\exp(-t_2M)$. Then,

$$\begin{aligned} \Pr\{\max|X_{ik}| \ge M \text{ for some } i\} &\le n\Pr(|X_k| \ge M) \\ &\le nt_1\exp(-t_2M) \end{aligned}$$

Thus, together with (4), we have

(5)
$$\begin{aligned} &\Pr\left\{\sup_u |\widehat{g}(X_k \mid u) - g(X_k \mid u)| \ge cn^{-\gamma}\right\} \\ &\quad \le c_1n^2\exp(-b_1n^{1-2\theta-2\gamma}/M^2) + nt_1\exp(-t_2M) \\ &\quad \le c_1n^2\exp\{-b_1n^{1/2-2\theta-2\gamma}\cdot(n^{1/2}/M^2)\} \\ &\qquad + nt_1\exp\{-t_2n^{1/2-2\theta-2\gamma}\cdot Mn^{2\theta+2\gamma-1/2}\} \end{aligned}$$

Following Liu et al. (2014), we take $M = O(n^\tau)$, where $1/2 - 2\theta - 2\gamma < \tau < 1/4$. For large n, we can see that $n^{1/2}/M^2 =$

$n^{1/2-2\tau} > 1$, and $Mn^{2\theta+2\gamma-1/2} = n^{\tau+2\theta+2\gamma-1/2} > 1$. Now (5) becomes

$$\Pr\left\{\sup_u |\widehat{g}(X_k \mid u) - g(X_k \mid u)| \geq cn^{-\gamma}\right\} \leq a_1 n^2 \exp(-b_1 n^{1/2-2\theta-2\gamma}),$$

where a_1, b_1 and c are some positive constants. The proof is complete. □

Lemma 5. *Suppose that conditions (C1) to (C5) are fulfilled, for any $0 < \gamma + \theta < 1/2$ and $0 < \gamma \leq m\theta$, then we have*

$$Pr\left\{\sup_u \left|\widehat{f}(u) - f(u)\right| \geq cn^{-\gamma}\right\} \leq a_2 n^2 \exp(-b_2 n^{1-2\theta-2\gamma}),$$

$$Pr\left\{\sup_{u,y} |\widehat{g}(y \mid u) - g(y \mid u)| \geq cn^{-\gamma}\right\} \leq a_3 n^3 \exp(-b_3 n^{1-2\theta-2\gamma}),$$

where a_2, b_2, a_3, b_3 and c are some positive constants.

Proof of Lemma 5. We use the same technique as the proof for Lemma 4. According to Theorem 3.1 of Zhu (1993) and Lemma 3.1 of Zhu and Fang (1996), for any $\varepsilon_n > 0$,

(6)
$$\Pr\left\{\sup_u \left|\widehat{f}(u) - E\widehat{f}(u)\right| \geq 8M_K h^{-1} n^{-1/2}\varepsilon_n\right\} \leq 3A\left(n^{1/2}/\varepsilon_n\right)^4 \exp\left(-\varepsilon_n^2/128\delta_1^2\right) + 4A\delta_1^{-8}\exp\left(-n\delta_1^2\right),$$

(7)
$$\Pr\left\{\sup_{u,y} |\widehat{g}(y \mid u) - E\widehat{g}(y \mid u)| \geq 8M_K h^{-1} n^{-1/2}\varepsilon_n\right\} \leq 3A\left(n^{1/2}/\varepsilon_n\right)^6 \exp\left(-\varepsilon_n^2/128\delta_2^2\right) + 4A\delta_2^{-12}\exp\left(-n\delta_2^2\right),$$

where $\delta_1 \geq \sup_u \left[\mathrm{var}\left\{K\left(\frac{\widetilde{u}-u}{h}\right)\right\}\right]^{1/2}$ and $\delta_2 \geq \sup_{u,y}\left[\mathrm{var}\left\{K\left(\frac{\widetilde{u}-u}{h}\right) I(Y \leq y)\right\}\right]^{1/2}$. And A is a generic constant and may take different values at different places. We expand $f(u)$ and $g(y \mid u)$ in a Taylor series with the Lagrange remainder term under Condition (C3). There exists some positive constants c_2 and c_3 such that

$$\begin{aligned}
&\sup_u \left|E\widehat{f}(u) - f(u)\right| \\
&= \sup_u \left|\int K_h(\widetilde{u}-u)\{f(\widetilde{u}) - f(u)\}\, d\widetilde{u}\right| \\
&= \sup_u \left|\int K(t)\{f(u+ht) - f(u)\}\, dt\right| \leq c_2\nu_m h^m, \\
&\sup_{u,y} |E\widehat{g}(y \mid u) - g(y \mid u)| \\
&= \sup_u \left|\int K_h(\widetilde{u}-u)\{F(y \mid \widetilde{u})f(\widetilde{u}) - F(y \mid u)f(u)\} d\widetilde{u}\right| \\
&= \sup_{u,y} \left|\int K(t)\{F(y \mid u+ht)f(u+ht) - F(y \mid u)f(u)\}\, dt\right| \\
&\leq c_3\nu_m h^m.
\end{aligned}$$

Since the kernel function K is uniformly continuous on its compact support, we have

$$\sup_{u,y}\left[\mathrm{var}\left\{K\left(\frac{\widetilde{u}-u}{h}\right) I(Y \leq y)\right\}\right]^{1/2} \leq \sup_u \left\{\int K^2\left(\frac{\widetilde{u}-u}{h}\right) f(\widetilde{u}) d\widetilde{u}\right\}^{1/2} \leq M_K$$

Thus for $0 < \gamma \leq m\theta$,

(8)
$$\begin{aligned}
&\Pr\left\{\sup_u \left|\widehat{f}(u) - f(u)\right| \geq cn^{-\gamma}\right\} \\
&= \Pr\left\{\sup_u \left|\widehat{f}(u) - E\widehat{f}(u) + E\widehat{f}(u) - f(u)\right| \geq cn^{-\gamma}\right\} \\
&\leq \Pr\left\{\sup_u \left|\widehat{f}(u) - E\widehat{f}(u)\right| \geq cn^{-\gamma}/2\right\} \\
&= O\left\{n^2\exp(-b_2 n^{1-2\theta-2\gamma})\right\},
\end{aligned}$$

(9)
$$\begin{aligned}
&\Pr\left\{\sup_{u,y} |\widehat{g}(y \mid u) - g(y \mid u)| \geq cn^{-\gamma}\right\} \\
&= \Pr\left\{\sup_{u,y} \left|\widehat{g}(y \mid u) - E\widehat{g}(y \mid u) + E\widehat{g}(y \mid u) - g(y \mid u)\right| \geq cn^{-\gamma}\right\} \leq \Pr\left\{\sup_{u,y} \left|\widehat{g}(y \mid u) - E\widehat{g}(y \mid u)\right| \geq cn^{-\gamma}/2\right\} \\
&= O\left\{n^3\exp(-b_3 n^{1-2\theta-2\gamma})\right\},
\end{aligned}$$

where b_2, b_3 and c are some positive constants. The proof is complete. □

Lemma 6. *(Liu et al., 2014) Suppose $T(u,y)$ and $S(u,y)$ are two uniformly-bounded functions of u and y. For any given u and y, $\widehat{T}(u,y)$ and $\widehat{S}(u,y)$ are estimates of $T(u,y)$ and $S(u,y)$ based on a sample with size n. For any $0 < \gamma + \theta < 1/4$ and $0 < \gamma \leq m\theta$, suppose that*

$$Pr\left\{\sup_{u,y}\left|\widehat{T}(u,y) - T(u,y)\right| \geq cn^{-\gamma}\right\} \leq a_4 n^3\exp(-b_4 n^{1/2-2\theta-2\gamma}),$$

$$Pr\left\{\sup_{u,y}\left|\widehat{S}(u,y) - S(u,y)\right| \geq cn^{-\gamma}\right\} \leq a_5 n^3\exp(-b_5 n^{1/2-2\theta-2\gamma}).$$

where a_4, b_4, a_5, b_5 and c are some positive constants. Then we have

$$Pr\left\{\sup_{u,y}\left|\widehat{T}(u,y)\widehat{S}(u,y) - T(u,y)S(u,y)\right| \geq cn^{-\gamma}\right\}$$

$$\leq a_6 n^3 \exp(-b_6 n^{1/2-2\theta-2\gamma}),$$

$$Pr\left\{\sup_{u,y}\left|\widehat{T}(u,y)/\widehat{S}(u,y) - T(u,y)/S(u,y)\right| \geq cn^{-\gamma}\right\}$$
$$\leq a_7 n^3 \exp(-b_7 n^{1/2-2\theta-2\gamma}),$$

$$Pr\left\{\sup_{u,y}\left|\{\widehat{T}(u,y)\widehat{S}(u,y)\} - \{T(u,y) - S(u,y)\}\right| \geq cn^{-\gamma}\right\}$$
$$\leq a_8 n^3 \exp(-b_8 n^{1/2-2\theta-2\gamma}),$$

where a_i, b_i for $i = 6, 7, 8$ and c are some positive constants.

Proof of Theorem 2. We divide the proof into two steps.

Step 1. We first prove that, under conditions (C1)-(C5), for any $0 < \gamma + \theta \leq 1/4$ and $0 < \gamma \leq m\theta$, there exists some positive constant c such that

$$\Pr\left(|\widehat{\omega}_k - \omega_k| \geq cn^{-\gamma}\right) \leq cn^3 \exp(-cn^{1/2-2\gamma-2\theta}).$$

We define Δ_1 and Δ_2 as follows,

$$\Delta_1 = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[\widehat{\text{corr}}^2\{X_k, I(Y \leq Y_j) \mid u_i\} - \text{corr}^2\{X_k, I(Y \leq Y_j) \mid u_i\}\Big];$$

$$\Delta_2 = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\text{corr}^2\{X_k, I(Y \leq Y_j) \mid u_i\} - \omega_k\right].$$

Then we have

$$\Pr\left(|\widehat{\omega}_k - \omega_k| \geq cn^{-\gamma}\right) \leq \Pr\left(|\Delta_1| \geq cn^{-\gamma}/2\right) + \Pr\left(|\Delta_2| \geq cn^{-\gamma}/2\right).$$

We deal with the first part of the summation.

(10)

$$\Pr\left(|\Delta_1| \geq cn^{-\gamma}/2\right)$$
$$\leq \Pr\left(n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\widehat{\text{corr}}^2\{X_k, I(Y \leq Y_j) \mid u_i\} - \text{corr}^2\{X_k, I(Y \leq Y_j) \mid u_i\}\right| \geq cn^{-\gamma}/2\right)$$
$$\leq \Pr\left(\sup_{u,y}\left|\widehat{\text{corr}}^2\{X_k, I(Y \leq y) \mid u\} - \text{corr}^2\{X_k, I(Y \leq y) \mid u\}\right| \geq cn^{-\gamma}/2\right).$$

For notation clarity, we define

$$\widehat{g}(X_k, y|u) = n^{-1}\sum_{i=1}^{n} K_h(u_i - u)X_{ik}I(Y_i \leq y), \text{ and}$$
$$\widehat{g}(X_k^2|u) = n^{-1}\sum_{i=1}^{n} K_h(u_i - u)X_{ik}^2.$$

Then $\text{corr}^2\{X_k, I(Y \leq y) \mid u\}$ can be written as

$$\frac{\left\{\widehat{g}(X_k, y \mid u)\widehat{f}(u) - \widehat{g}(X_k \mid u)\widehat{g}(y \mid u)\right\}^2}{\left\{\widehat{g}(X_k^2 \mid u)\widehat{f}(u) - \widehat{g}^2(X_k \mid u)\right\}\left\{\widehat{g}(y \mid u)\widehat{f}(u) - \widehat{g}^2(y \mid u)\right\}}$$

Similar to the proof of Lemma 4, we can obtain

$$\Pr\left\{\sup_u |\widehat{g}(X_k \mid u) - g(X_k \mid u)| \geq cn^{-\gamma}\right\} \leq a_9 n^2 \exp(-b_9 n^{1/2-2\gamma-2\theta}).$$

Together with Lemmas 4, 5 and 6, it is clear that

$$\Pr\Big(\sup_{u,y}\left|\widehat{\text{corr}}^2\{X_k, I(Y \leq y) \mid u\} - \text{corr}^2\{X_k, I(Y \leq y) \mid u\}\right| \geq cn^{-\gamma}\Big) \leq cn^3 \exp(-cn^{1/2-2\gamma-2\theta}).$$

Thus (10) becomes

$$\Pr\left(|\Delta_1| \geq cn^{-\gamma}/2\right) \leq cn^3 \exp(-cn^{1/2-2\gamma-2\theta}).$$

With Hoeffding's inequality in Lemma 1, we can show that $\Pr\left(|\Delta_2| \geq cn^{-\gamma}/2\right) \leq 2\exp(-c^2 n^{1-2\gamma}/2)$. Thus, there exists some positive constants c such that

$$\Pr\left(|\widehat{\omega}_k - \omega_k| \geq cn^{-\gamma}\right) \leq cn^3 \exp(-cn^{1/2-2\gamma-2\theta}).$$

Step 2. Assume the condition $\min_{k\in\mathcal{A}} \omega_k \geq 2cn^{-\gamma}$. We prove that

$$\Pr(\mathcal{A} \subseteq \widehat{\mathcal{A}}) \quad \geq \quad 1 - O\left\{n^3|\mathcal{A}|\exp(-cn^{1/2-2\gamma-2\theta})\right\}.$$

If $\mathcal{A} \nsubseteq \widehat{\mathcal{A}}$, there must exist some $j \in \mathcal{A}$ such that $\widehat{\omega}_j < cn^{-\gamma}$. Under the condition $\min_{k\in\mathcal{A}} \omega_k \geq 2cn^{-\gamma}$, we have $|\widehat{\omega}_j - \omega_j| \geq cn^{-\gamma}$ for this particular j, which implies

$$\{\mathcal{A} \nsubseteq \widehat{\mathcal{A}}\} \subseteq \{|\widehat{\omega}_j - \omega_j| \geq cn^{-\gamma}, \text{ for some } j \in \mathcal{A}\}.$$

Then, it is clear that

$$\Pr(\mathcal{A} \subseteq \widehat{\mathcal{A}}) \geq 1 - \Pr\left\{|\widehat{\omega}_j - \omega_j| \geq cn^{-\gamma}, \text{ for some } j \in \mathcal{A}\right\}$$
$$\geq 1 - |\mathcal{A}| \max_{j\in\mathcal{A}} \Pr\{|\widehat{\omega}_j - \omega_j| \geq cn^{-\gamma}\}$$
$$\geq 1 - O\left\{n^3|\mathcal{A}|\exp(-cn^{1/2-2\gamma-2\theta})\right\},$$

which implies the desired conclusion. □

6.3 Appendix C: Proof of Theorem 3

The conditions (A1)-(A3) illustrate that $\min_{k\in\mathcal{A}} \omega_k - \max_{k\in\mathcal{A}^c} \omega_k \geq 0$.Thus, there exists some $\delta > 0$ such that $\min_{k\in\mathcal{A}} \omega_k - \max_{k\in\mathcal{A}^c} \omega_k = \delta$. Then we have

$$\Pr\left\{\min_{k\in\mathcal{A}} \widehat{\omega}_k \leq \max_{k\in\mathcal{A}^c} \widehat{\omega}_k\right\}$$

$$
\begin{aligned}
&\leq \Pr\left\{\min_{k\in\mathcal{A}}\widehat{\omega}_k - \min_{k\in\mathcal{A}}\omega_k + \delta \leq \max_{k\in\mathcal{A}^c}\widehat{\omega}_k - \max_{k\in\mathcal{A}^c}\omega_k\right\}\\
&\leq \Pr\left\{\left|(\min_{k\in\mathcal{A}}\widehat{\omega}_k - \max_{k\in\mathcal{A}^c}\widehat{\omega}_k) - (\min_{k\in\mathcal{A}}\omega_k - \max_{k\in\mathcal{A}^c}\omega_k)\right| \geq \delta\right\}\\
&\leq \Pr\left\{2\max_{1\leq k\leq p}|\widehat{\omega}_k - \omega_k| \geq \delta\right\}\\
&\leq O\left\{p\,n^3\exp\left(-cn^{1/2-2\theta}\right)\right\},
\end{aligned}
$$

Then by Fatou's Lemma,

$$
\begin{aligned}
&\Pr\left\{\liminf_{n\longrightarrow\infty}\left(\min_{k\in\mathcal{A}}\widehat{\omega}_k - \max_{k\in\mathcal{A}^c}\widehat{\omega}_k\right) \leq 0\right\}\\
&\quad\leq \lim_{n\to\infty}\Pr\left(\min_{k\in\mathcal{A}}\widehat{\omega}_k - \max_{k\in\mathcal{A}^c}\widehat{\omega}_k \leq 0\right) = 0
\end{aligned}
$$

In other words, we have

$$
\Pr\left\{\liminf_{n\longrightarrow\infty}\left(\min_{k\in\mathcal{A}}\widehat{\omega}_k - \max_{k\in\mathcal{A}^c}\widehat{\omega}_k\right) > 0\right\} = 1.
$$

□

Received 10 April 2018

REFERENCES

Chin, K., DeVries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., et al. (2006). Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell*, 10(6):529–541.

DeSantis, C. E., Ma, J., Goding Sauer, A., Newman, L. A., and Jemal, A. (2017). Breast cancer statistics, 2017, racial disparity in mortality by state. *CA: A Cancer Journal for Clinicians*, 67(6):439–448.

Fan, J., Ma, Y., and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association*, 109(507):1270–1284. MR3265696

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(494):544–557. MR2847969

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B*, 70(5):849–911. MR2530322

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6):3567–3604. MR2766861

He, X., Wang, L., Hong, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41(1):342–369. MR3059421

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30. MR0144363

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012a). Robust rank correlation based screening. *The Annals of Statistics*, 40(3):1846–1877. MR3015046

Li, R., Zhong, W., and Zhu, L. (2012b). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139. MR3010900

Lin, Q., Zhao, Z., Liu, J. S., et al. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics*, 46(2):580–610. MR3782378

Liu, J., Li, R., and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *Journal of the American Statistical Association*, 109(505):266–274. MR3180562

Liu, J., Zhong, W. and Li, R. (2015). A selective overview of feature screening for ultrahigh-dimensional data. *Science China Mathematics*, 58(10):2033–2054. MR3400642

McPherson, K., Steel, C., and Dixon, J. (2000). Abc of breast diseases: breast cancer—epidemiology, risk factors, and genetics. *BMJ: British Medical Journal*, 321(7261):624-628.

Pan, W., Wang, X., Xiao, W., and Zhu, H. (2018). A generic sure independence screening procedure. *Journal of the American Statistical Association*, To appear.

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734. MR3449068

Wen, C., Pan, W., Huang, M., and Wang, X. (2018). Sure independence screening adjusted for confounding covariates with ultrahigh dimensional data. *Statistica Sinica*, 28:293–317. MR3752262

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Witten, D. M., Tibshirani, R., Gross, S. and Narasimhan, B. PMA: Penalized multivariate analysis. R package version 1.0.9, 2013. Available from: http://CRAN.R-project.org/package=PMA.

Zhu, L. P., Li, L., Li, R., and Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475. MR2896849

Zhu, L. X. (1993). Convergence rates of the empirical processes indexed by the classes of functions with applications. *Journal of Systems Science and Mathematical Sciences*, 13(1):33–41. MR1261313

Zhu, L. X. and Fang, K. T. (2007). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics*, 24(3):1053–1068. MR1401836

Yeqing Zhou
School of Statistics and Management
Shanghai University of Finance and Economics
Shanghai
China
E-mail address: yqzhou1991@hotmail.com

Jingyuan Liu
Department of Statistics in School of Economics
Wang Yanan Institute for Studies in Economics
Fujian Key Laboratory of Statistical Science
Xiamen University
Xiamen
China
E-mail address: jingyuan@xmu.edu.cn

Zhihui Hao
Wang Yanan Institute for Studies in Economics
Xiamen University
Xiamen
China
E-mail address: 1320744040@qq.com

Liping Zhu
Research Center for Applied Statistical Science
Institute of Statistics and Big Data
Renmin University of China
Beijing
China
E-mail address: zhu.liping@ruc.edu.cn