Nonparametric Independence Screening in Sparse Ultra-High Dimensional Varying Coefficient Models

Jianqing Fan
Princeton University, USA
Yunbei Ma
Southwestern University of Finance and Economics, China
and Wei Dai
Princeton University, USA

Summary. The varying-coefficient model is an important nonparametric statistical model that allows us to examine how the effects of covariates vary with exposure variables. When the number of covariates is big, the issue of variable selection arrives. In this paper, we propose and investigate marginal nonparametric screening methods to screen variables in ultra-high dimensional sparse varying-coefficient models. The proposed nonparametric independence screening (NIS) selects variables by ranking a measure of the nonparametric marginal contributions of each covariate given the exposure variable. The sure independent screening property is established under some mild technical conditions when the dimensionality is of nonpolynomial order, and the dimensionality reduction of NIS is quantified. To enhance practical utility and the finite sample performance, two data-driven iterative NIS methods are proposed for selecting thresholding parameters and variables: conditional permutation and greedy methods, resulting in Conditional-INIS and Greedy-INIS. The effectiveness and flexibility of the proposed methods are further illustrated by simulation studies and real data applications.

Keywords: Sure independence screening; Variable selection; Sparsity; Conditional permutation; False positive rates

1. Introduction

The development of information and technology drives big data collections in many areas of advanced scientific research ranging from genomic and health science to machine learning and economics. The collected data frequently has an ultra-high dimensionality p that is allowed to diverge at nonpolynomial (NP) rate with the sample size n, namely $\log(p) = O(n^\rho)$ for some $\rho > 0$. For example, in biomedical research such as genomewide association studies for some mental diseases, millions of SNPs are potential covariates. Traditional statistical methods face significant challenges in dealing with such a high-dimensional problem with large sample sizes.

With the sparsity assumption, variable selection helps improve the accuracy of estimation and gain scientific insights. Many significant variable selection techniques have been developed, such as Bridge regression in Frank and Friedman (1993), Lasso in Tibshirani (1996), SCAD and folded concave penalty in Fan and Li (2001), the Elastic net in Zou and Hastie (2005), Adaptive Lasso (Zou, 2006), and the Dantzig selector in Candes and Tao (2007). Methods on the implementation of folded concave penalized least-squares include the local linear approximation algorithm in Zou and Li (2008) and the plus algorithm in Zhang (2010). However, due to the simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability, these methods do not perform well in ultra-high dimensional problems.

To tackle these problems, Fan and Lv (2008) introduced a sure independence screening (SIS) method to select important variables in ultra-high dimensional linear regression models via marginal correlation learning. Hall and Miller (2009) extended the method to the generalized correlation ranking, which was further extended by Fan, Feng and Song (2011) for ultra-high dimensional nonparametric additive models, resulting in nonparametric independence screening (NIS). On a different front, Fan and Song (2010)

Address for correspondence: Yunbei Ma, School of Statistics, Southwest University of Finance and Economics, Chengdu, China.
E-mail:myb@swufe.edu.cn

extended the SIS idea to ultra-high dimensional generalized linear models and devised a useful technical tool for establishing the sure screening results and bounding false selection rates. Other related methods include data-tilling method (Hall, Titterington and Xue, 2009), marginal partial likelihood method MPLE (Zhao and Li, 2010), and robust screening methods by rank correlation (Li, *et al.* , 2012) and distance correlation (Li, Zhong and Zhu, 2012). Inspired by these previous work, our study will focus on variable screening in nonparametric varying-coefficient models with NP dimensionality.

It is well known that nonparametric models are flexible enough to reduce modeing biases. However, they suffer from the so-called "curse of dimensionality". A remarkable simple and powerful nonparametric model for dimensionality reductions is the varying-coefficient model,

$$Y = \boldsymbol{\beta}^T(W)\mathbf{X} + \epsilon, \tag{1}$$

where $\mathbf{X} = (X_1, \cdots, X_p)^T$ is the vector of covariates, W is some observable exposure variables, Y is the response, and ε is the random noise with conditional mean 0 and finite conditional variance. An intercept term (i.e., $X_0 \equiv 1$) can be introduced if necessary. This model assumes that the variables in the covariate vector $\mathbf{X}$ enter the model linearly, meanwhile it allows regression coefficient functions to very smoothly with the exposure variable. The model retains general nonparametric characteristics and allows the nonlinear interactions between the exposure variable W and the covariates. It arises frequently from economics, finance, politics, epidemiology, medical science, ecology, among others. For an overview, see Fan and Zhang (2008).

When the dimensionality p is finite, Fan, Zhang and Zhang (2001) proposed the generalized likelihood ratio (GLR) test to select variables in the varying-coefficient model (1). For the time-varying coefficient model, a special case of (1) with the exposure variable W being the time t, Wang, Li and Huang (2008) applied the basis function approximations and the SCAD penalty to address the problem of variable selection. In the NP dimensional setting, Lian (2011) utilized the adaptive group Lasso penalty in time-varying coefficient models. These methods still face the aforementioned three challenges.

In this paper, we consider a nonparametric screening by ranking a measure of the marginal nonparametric contributions of each covariate given the exposure variable. For each given covariate, we fit marginal regressions of the response Y against the covariate X_j $(j = 1, \cdots, p)$ conditioning on W:

$$\min_{a_j, b_j} \mathrm{E}[(Y - a_j - b_j X_j)^2 | W] \tag{2}$$

Let $a_j(W)$ and $b_j(W)$ be the solution to (2) and $\widehat{a}_{nj}(W)$ and $\widehat{b}_{nj}(W)$ be their nonparametric estimates. Then, we rank the importance of each covariate in the joint model according to a measure of marginal utility (which is equivalent to the goodness of fit) in its marginal model. Under some reasonable conditions, the magnitude of these marginal contributions provides useful probes of the importance of variables in the joint varying-coefficient model. This is an important extension of SIS (Fan and Lv, 2008) to a more flexible class of varying coefficient models.

The sure screening property of NIS can be established under certain technical conditions. In some very specific cases, NIS can even be model selection consistent. In establishing this kind of results, three factors are related to the minimum distinguishable marginal signals: the stochastic error in estimating the nonparametric components, the approximation error in modeling nonparametric components, and the tail distributions of the covariates. Following Fan and Lv (2008) and Fan, Feng and Song (2011), we propose two nonparametric independence screening approaches in an iterative framework. One is called Greedy-INIS, in which we adopt a greedy method in the variable screening step. The other is called Conditional-INIS which is built on conditional random permutation to determine a data driven screening threshold. They both serve to effectively control the false positive rate and false negative rate with enhanced performance.

This article is organized as follows. In Section 2, we fit each marginal nonparametric regression model via B-spline basis approximation and screen variables by ranking a measure of these estimators. In Section 3, we establish the sure screening property and model selection consistency under certain technical conditions. Iterative NIS procedures (namely Greedy-INIS and Conditional-INIS) are developed in Section 4. In Section 5, a set of numerical studies are conducted to evaluate the performance of our proposed methods.

2. Models and Nonparametric Marginal Screening Method

In this section we study the varying-coefficient model with the conditional linear structure as in (1). Assume that the functional coefficient vector $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \cdots, \beta_p(\cdot))^T$ is sparse. Let $\mathcal{M}_* = \{j : \mathrm{E}[\beta_j^2(W)] > 0\}$ be the true sparse model with nonsparsity size $s_n = |\mathcal{M}_*|$. We allow p to grow with n and denote it by p_n whenever necessary.

2.1. Marginal Regression

For $j = 1, \cdots, p$, let $a_j(W)$ and $b_j(W)$ be the minimizer of the following marginal regression problem:

$$\min_{a_j(W), b_j(W) \in L_2(P)} \mathrm{E}[(Y - a_j(W) - b_j(W)X_j)^2 | W], \tag{3}$$

where P denotes the joint distribution of $(Y, W, \mathbf{X})$ and $L_2(P)$ is the class of square integrable functions under the measure P. By some algebra, we have that the minimizer of (3) is

$$b_j(W) = \frac{\mathrm{Cov}[X_j, Y|W]}{\mathrm{Var}[X_j|W]}, \ a_j(W) = \mathrm{E}[Y|W] - b_j(W)\mathrm{E}[X_j|W]. \tag{4}$$

Let $a_0(W) = \mathrm{E}[Y|W]$, we rank the marginal utility of covariates by

$$u_j = \|a_j(W) + b_j(W)X_j\|^2 - \|a_0(W)\|^2, \tag{5}$$

where $\|f\|^2 = \mathrm{E}f^2$. It can be seen that

$$u_j \quad = \quad \mathrm{E}[b_j^2(W)(X_j - \mathrm{E}[X_j|W])^2] = \mathrm{E}\left[\frac{(\mathrm{Cov}[X_j, Y|W])^2}{\mathrm{Var}[X_j|W]}\right]. \tag{6}$$

For each $j = 1, \cdots, p$, if $\mathrm{Var}[X_j|W] = 1$, then u_j has the same quantity as the measure of marginal functional coefficient $\|b_j(W)\|^2$. On the other hand, this marginal utility is closely related to the conditional correlation between $X_j's$ and Y, as $u_j = 0$ if and only if $\mathrm{Cov}[X_j, Y|W] = 0$ almost surely.

2.2. Marginal Regression Estimation with B-spline

To obtain an estimate of the marginal utility u_j, $j = 1, \cdots, p$, we approximate $a_j(W)$ and $b_j(W)$ by functions in $\mathcal{S}_n$, the space of polynomial splines of degree $l \geq 1$ on $\mathcal{W}$, a compact set. Let $\{B_k, k = 1, \cdots, L_n\}$ denote its normalized B-spline basis with $\|B_k\|_\infty \leq 1$, where $\|\cdot\|_\infty$ is the sup norm. Then

$$\begin{aligned} a_j(W) &\approx \sum_{k=1}^{L_n} \eta_{jk} B_k(W), \quad j = 0, \cdots, p, \\ b_j(W) &\approx \sum_{k=1}^{L_n} \theta_{jk} B_k(W), \quad j = 1, \cdots, p. \end{aligned}$$

where $\{\theta_{jk}\}_{k=1}^{L_n}$ and $\{\eta_{jk}\}_{k=1}^{L_n}$ are scalar coefficients.

We now consider the following sample version of the marginal regression problem:

$$\min_{\boldsymbol{\eta}_j, \boldsymbol{\theta}_j \in \mathbb{R}^{L_n}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{B}(W_i)\boldsymbol{\eta}_j - \mathbf{B}(W_i)\boldsymbol{\theta}_j X_{ji})^2, \tag{7}$$

where $\boldsymbol{\eta}_j = (\eta_{j1}, \cdots, \eta_{jL_n})^T$, $\boldsymbol{\theta}_j = (\theta_{j1}, \cdots, \theta_{jL_n})^T$ and $\mathbf{B}(\cdot) = (B_1(\cdot), \cdots, B_{L_n}(\cdot))$.

It is easy to show that the minimizers of (7) is given by

$$(\widehat{\boldsymbol{\eta}}_j^T, \ \widehat{\boldsymbol{\theta}}_j^T)^T = (\mathbf{Q}_{nj}^T \mathbf{Q}_{nj})^{-1} \mathbf{Q}_{nj}^T \mathbf{Y}, \tag{8}$$

where

$$\mathbf{Q}_{nj} = (\mathbf{B}_n, \ \boldsymbol{\Phi}_{nj}) = \begin{pmatrix} \mathbf{B}(W_1), & X_{j1}\mathbf{B}(W_1) \\ \vdots & \vdots \\ \mathbf{B}(W_n), & X_{jn}\mathbf{B}(W_n) \end{pmatrix}$$

is an $n \times 2L_n$ matrix. As a result, the estimates of a_j and b_j, $j = 1, \cdots, p$ are given by

$$\widehat{a}_{nj}(W) = \mathbf{B}(W)\widehat{\boldsymbol{\eta}}_j = (\mathbf{B}(W), \mathbf{0}_{L_n}^T)(\mathbf{Q}_{nj}^T\mathbf{Q}_{nj})^{-1}\mathbf{Q}_{nj}^T\mathbf{Y},$$
$$\widehat{b}_{nj}(W) = \mathbf{B}(W)\widehat{\boldsymbol{\theta}}_j = (\mathbf{0}_{L_n}^T, \mathbf{B}(W))(\mathbf{Q}_{nj}^T\mathbf{Q}_{nj})^{-1}\mathbf{Q}_{nj}^T\mathbf{Y}, \quad (9)$$

where $\mathbf{0}_{L_n}$ is an L_n-dimension vector with all entries 0. Similarly, we have the estimate of the intercept function a_0 by

$$\widehat{a}_{n0}(W) = \mathbf{B}(W)\widehat{\boldsymbol{\eta}}_0 = \mathbf{B}(W)(\mathbf{B}_n^T\mathbf{B}_n)^{-1}\mathbf{B}_n^T\mathbf{Y}, \quad (10)$$

where

$$\widehat{\boldsymbol{\eta}}_0 = \arg\min_{\boldsymbol{\eta}_0\in\mathbb{R}^{L_n}} \frac{1}{n}\sum_{i=1}^n (Y_i - \mathbf{B}(W_i)\boldsymbol{\eta}_0)^2. \quad (11)$$

We now define an estimate of the marginal utility u_j as

$$\begin{aligned}\widehat{u}_{nj} &= \|\widehat{a}_{nj}(\mathbf{W}) + \widehat{b}_{nj}(\mathbf{W})\mathbf{X}_j\|_n^2 - \|\widehat{a}_{n0}(\mathbf{W})\|_n^2 \\ &= \frac{1}{n}\sum_{i=1}^n (\widehat{a}_{nj}(W_i) + \widehat{b}_{nj}(W_i)X_{ji})^2 - \frac{1}{n}\sum_{i=1}^n (\widehat{a}_{n0}(W_i))^2, \quad (12)\end{aligned}$$

where $\mathbf{W} = (W_1, \cdots, W_n)^T$. Note that throughout this paper, whenever two vectors $\mathbf{a}$ and $\mathbf{b}$ are of the same length, $\mathbf{ab}$ denotes the componentwise product. Given a predefined threshold value τ_n, we select a set of variables as follows:

$$\mathcal{M}_{\tau_n} = \{1 \le j \le p : \widehat{u}_{nj} \ge \tau_n\}. \quad (13)$$

Alternatively, we can rank the covariates by the residual sum of squares of marginal nonparametric regressions, which is defined as

$$\widehat{v}_{nj} = \|\mathbf{Y} - \widehat{a}_{nj}(\mathbf{W}) - \widehat{b}_{nj}(\mathbf{W})\mathbf{X}_j\|_n^2, \quad (14)$$

and we select variables as follows,

$$\mathcal{M}_{\nu_n} = \{1 \le j \le p : \widehat{v}_{nj} \le \nu_n\}, \quad (15)$$

where ν_n is a predefined threshold value.

It is worth noting that ranking by marginal utility $\widehat{u}_{nj}$ is equivalent to ranking by the measure of goodness of fit $\widehat{v}_{nj}$. To see the equivalence, first note that

$$\|\widehat{a}_{nj}(\mathbf{W}) + \widehat{b}_{nj}(\mathbf{W})\mathbf{X}_j\|_n^2 = \frac{1}{n}\mathbf{Y}^T\mathbf{Q}_{nj}(\mathbf{Q}_{nj}^T\mathbf{Q}_{nj})^{-1}\mathbf{Q}_{nj}^T\mathbf{Y}, \quad (16)$$

and

$$\frac{1}{n}\sum_{i=1}^n Y_i(\widehat{a}_{nj}(W_i) + \widehat{b}_{nj}(W_i)X_{ji}) = \frac{1}{n}\mathbf{Y}^T\mathbf{Q}_{nj}(\mathbf{Q}_{nj}^T\mathbf{Q}_{nj})^{-1}\mathbf{Q}_{nj}^T\mathbf{Y}. \quad (17)$$

It follows from (16) and (17) that

$$\widehat{v}_{nj} = \|\mathbf{Y}\|_n^2 - \|\widehat{a}_{n0}(\mathbf{W})\|_n^2 - \widehat{u}_{nj}. \quad (18)$$

Since the first two terms on the right hand side of (18) do not vary in j, ranking by $\widehat{u}_{nj}$ is the same as that by $\widehat{v}_{nj}$. Therefore, selecting variables with large marginal utility is the same as picking those that yield small marginal residual sum of squares.

To bridge u_j and $\widehat{u}_{nj}$, we define the population version of the marginal regression using B-spline basis. From now on, we will omit the argument in $\mathbf{B}(W)$ and write $\mathbf{B}$ whenever the context is clear. Let $\tilde{a}_j(W) = \mathbf{B}\tilde{\boldsymbol{\eta}}_j$ and $\tilde{b}_j(W) = \mathbf{B}\tilde{\boldsymbol{\theta}}_j$, where $\tilde{\boldsymbol{\eta}}_j$ and $\tilde{\boldsymbol{\theta}}_j$ are the minimizer of

$$\min_{\boldsymbol{\eta}_j,\boldsymbol{\theta}_j\in\mathbb{R}^{L_n}} \mathrm{E}[(Y - \mathbf{B}\boldsymbol{\eta}_j - \mathbf{B}\boldsymbol{\theta}_j X_j)^2], \quad (19)$$

and $\tilde{a}_0(W) = \mathbf{B}\tilde{\boldsymbol{\eta}}_0$, where $\tilde{\boldsymbol{\eta}}_0$ is the minimizer of

$$\min_{\boldsymbol{\eta}_0 \in \mathbb{R}^{L_n}} \mathrm{E}[(Y - \mathbf{B}\boldsymbol{\eta}_0)^2]. \tag{20}$$

It can be seen that

$$\begin{aligned} (\tilde{a}_j(W), \tilde{b}_j(W))^T &= \mathrm{diag}(\mathbf{B}, \mathbf{B})(\mathrm{E}[\mathbf{Q}_j^T \mathbf{Q}_j])^{-1}\mathrm{E}[\mathbf{Q}_j^T Y], & (21) \\ \tilde{a}_0(W) &= \mathbf{B}(\mathrm{E}[\mathbf{B}^T\mathbf{B}])^{-1}\mathrm{E}[\mathbf{B}^T Y], & (22) \end{aligned}$$

where $\mathbf{Q}_j = (\mathbf{B}, X_j\mathbf{B})$

$$\begin{aligned} \tilde{u}_j &= \|\tilde{a}_j(W) + \tilde{b}_j(W)X_j\|^2 - \|\tilde{a}_0(W)\|^2 \\ &= \mathrm{E}[Y\mathbf{Q}_j](\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])^{-1}\mathrm{E}[\mathbf{Q}_j^T Y] - \mathrm{E}[Y\mathbf{B}](\mathrm{E}[\mathbf{B}^T\mathbf{B}])^{-1}\mathrm{E}[\mathbf{B}^T Y]. & (23) \end{aligned}$$

3. Sure Screening

In this section, we establish the sure screening properties of the proposed method for model (1). Recall that by (6) the population version of marginal utility quantifies the relationship between $X_j's$ and Y as follows:

$$u_j = \mathrm{E}\left[\frac{(\mathrm{Cov}[X_j, Y|W])^2}{\mathrm{Var}[X_j|W]}\right], \quad j = 1, \cdots, p. \tag{24}$$

Then the following two conditions guarantee that the marginal signal of the active components $\{u_j\}_{j\in\mathcal{M}_*}$ does not vanish.

(i) Suppose for $j = 1, \cdots, p$, $\mathrm{Var}[X_j|W]$ is uniformly bounded away from 0 and infinity on $\mathcal{W}$, where $\mathcal{W}$ is the compact support of W. That is, there exist some positive constants h_1 and h_2, such that $0 < h_1 \le \mathrm{Var}[X_j|W] \le h_2 < \infty$.

(ii) $\min_{j\in\mathcal{M}_*} \mathrm{E}[(\mathrm{Cov}[X_j, Y|W])^2] \ge c_1 L_n n^{-2\kappa}$, for some $\kappa > 0$ and $c_1 > 0$.

Then under conditions (i) and (ii),

$$\min_{j\in\mathcal{M}_*} u_j \ge c_1 L_n n^{-2\kappa}/h_2. \tag{25}$$

Note that in condition (ii), the number of basis functions L_n is not intrinsic. By the Remark 1 below, L_n should be chosen in correspondence to the smoothness condition of the nonparametric component. Therefore, condition (ii) depends only on κ and smoothness parameter d in condition (iii). We keep L_n here to make the relationship more explicit.

3.1. Sure Screening Properties

The following conditions (iii)-(vii) are required for the B-spline approximation in marginal regressions and establishing the sure screening properties.

(iii) The density function g of W is bounded away from zero and infinity on $\mathcal{W}$. That is, $0 < T_1 \le g(W) \le T_2 < \infty$ for some constants T_1 and T_2.

(iv) Functions $\{a_j\}_{j=0}^p$ and $\{b_j\}_{j=1}^p$ belong to a class of functions $\mathcal{B}$, whose rth derivative $f^{(r)}$ exists and is Lipschitz of order α. That is,

$$\mathcal{B} = \{f(\cdot) : |f^{(r)}(s) - f^{(r)}(t)| \le M|s - t|^\alpha \text{ for } s, t \in \mathcal{W}\},$$

for some positive constant M, where r is a nonnegative integer and $\alpha \in (0, 1]$ such that $d = r + \alpha > 0.5$.

(v) Suppose for all $j = 1, \cdots, p$, there exists a positive constant K_1 and $r_1 \geq 2$, such that

$$\mathrm{P}(|X_j| > t|W) \leq \exp(1 - (t/K_1)^{r_1}), \tag{26}$$

uniformly on $\mathcal{W}$, for any $t \geq 0$. Furthermore, let $m(\mathbf{X}^*) = \mathrm{E}[Y|\mathbf{X}, W]$, where $\mathbf{X}^* = (\mathbf{X}^T, W)^T$. Suppose there exists some positive constants K_2 and r_2 satisfying $r_1 r_2/(r_1 + r_2) \geq 1$, such that

$$\mathrm{P}(|m(\mathbf{X}^*)| > t|W) \leq \exp(1 - (t/K_2)^{r_2}). \tag{27}$$

uniformly on $\mathcal{W}$, for any $t \geq 0$.

(vi) The random errors $\{\varepsilon_i\}_{i=1}^n$ are i.i.d with conditional mean 0, and there exists some positive constants K_3 and r_3 satisfying $r_1 r_3/(r_1 + r_3) > 1$, such that

$$\mathrm{P}(|\varepsilon| > t|W) \leq \exp(1 - (t/K_3)^{r_3}), \tag{28}$$

uniformly on $\mathcal{W}$, for any $t \geq 0$.

(vii) There exists some constant $\xi \in (0, 1/h_2)$ such that $L_n^{-2d-1} \leq c_1(1/h_2 - \xi)n^{-2\kappa}/M_1$.

Proposition 1. *Under conditions (i)-(v), there exists a positive constant M_1 such that*

$$u_j - \tilde{u}_j \leq M_1 L_n^{-2d}. \tag{29}$$

In addition, when $L_n^{-2d-1} \leq c_1(1/h_2 - \xi)n^{-2\kappa}/M_1$ for some $\xi \in (0, 1/h_2)$, we have

$$\min_{j \in \mathcal{M}_*} \tilde{u}_j \geq c_1 \xi L_n n^{-2\kappa}. \tag{30}$$

Remark 1. *It follows from Proposition 1 that the minimum signal level of $\{\tilde{u}_j\}_{j\in\mathcal{M}_*}$ is approximately the same as $\{u_j\}_{j\in\mathcal{M}_*}$, provided that the approximation error is negligible. It also shows that the number of basis functions L_n should be chosen as*

$$L_n \geq C n^{2\kappa/(2d+1)},$$

for some positive constant C. In other words, the smoother the underlying function is (i.e., the larger d is), the smaller L_n we can take.

The following Theorem 1 provides the sure screening properties of the nonparametric independence screening method proposed in Section 2.2.

Theorem 1. *Suppose conditions (i)-(vi) hold.*

(i) If $n^{1-4\kappa}L_n^{-3} \to \infty$ as $n \to \infty$, then for any $c_2 > 0$, there exist some positive constants c_3 and c_4 such that

$$\begin{aligned} & P\left(\max_{1\leq j\leq p} |\widehat{u}_{nj} - \tilde{u}_j| \geq c_2 L_n n^{-2\kappa}\right) \\ \leq\ & 12 p_n L_n \{(2 + L_n)\exp(-c_3 n^{1-4\kappa} L_n^{-3}) + 3L_n \exp(-c_4 L_n^{-3} n)\}. \end{aligned} \tag{31}$$

(ii) If condition (vii) also holds, then by taking $\tau_n = c_5 L_n n^{-2\kappa}$ with $c_5 = c_1\xi/2$, there exist positive constants c_6 and c_7 such that

$$\begin{aligned} P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{\tau_n}\right) \ \geq\ & 1 - 12 s_n L_n \{(2 + L_n)\exp(-c_6 n^{1-4\kappa} L_n^{-3}) \\ & + 3L_n \exp(-c_7 L_n^{-3} n)\}. \end{aligned} \tag{32}$$

Remark 2. *According to Theorem 1 , we can handle NP dimensionality*

$$p = o(\exp\{n^{1-4\kappa} L_n^{-3}\}).$$

It shows that the number of spline bases L_n also affects the order of dimensionality: the smaller L_n is, the higher dimensionality we can handle. On the other hand, Remark 1 points out that it is required $L_n \geq C n^{2\kappa/(2d+1)}$ to have a good bias property. This means that the smoother the underlying function is (i.e. the larger d is), the smaller L_n we can take, and consequently higher dimensionality can be handled. The compatibility of these two requirements requires that $\kappa < (d + 0.5)/(4d + 5)$, which implies that $\kappa < 1/4$. We can take $L_n = O(n^{1/(2d+1)})$, which is the optimal convergence rate for nonparametric regression (Stone, 1982). In this case, the allowable dimensionality can be as high as

$$p = o(\exp\{n^{\frac{2(d-1)}{2d+1}}\}).$$

3.2. False Selection Rates

According to (30), the ideal case for vanishing false-positive rate is when

$$\max_{j\notin\mathcal{M}_*} \tilde{u}_j = o(L_n n^{-2\kappa})$$

so that there is a natural separation between important and unimportant variables. By Theorem 1(i), when (31) tends to zero, we have with probability tending to 1 that

$$\max_{j\notin\mathcal{M}_*} \widehat{u}_{nj} \le cL_n n^{-2\kappa}, \text{ for any } c > 0.$$

Consequently, by choosing τ_n as in Theorem 1(ii), NIS can achieve the model selection consistency under this ideal situation, i.e.,

$$P\left(\widehat{\mathcal{M}}_{\tau_n} = \mathcal{M}_*\right) = 1 - o(1).$$

In particular, this ideal situation occurs under the partial orthogonality condition, i.e., $\{X_j\}_{j\in\mathcal{M}_*}$ is independent of $\{X_i\}_{i\notin\mathcal{M}_*}$ given W, which implies $u_j = 0$ for $j \notin \mathcal{M}_*$

In general, the model selection consistency can not be achieved by a single step of marginal screening. The marginal probes can not separate important variables from unimportant variables. The following Theorem 2 quantifies how the size of selected models is related to the matrix of basis functions and the thresholding parameter τ_n.

THEOREM 2. *Under the same conditions in Theorem 1, for any $\tau_n = c_5 L_n n^{-2\kappa}$, there exist positive constants c_8 and c_9 such that*

$$P\left\{|\widehat{\mathcal{M}}_{\tau_n}| \le O(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma}))\right\} \ge 1 - 12p_n L_n\Big\{(2+L_n)\exp(-c_8 n^{1-4\kappa}L_n^{-3}) + 3L_n\exp(-c_9 n L_n^{-3})\Big\}, \quad (33)$$

where $\mathbf{\Sigma} = \mathrm{E}[\mathbf{Q}^T\mathbf{Q}]$, and $\mathbf{Q} = (\mathbf{Q}_1, \cdots, \mathbf{Q}_p)$ is a functional vector of $2p_nL_n$ dimension.

4. Iterative Nonparametric Independence Screening

As Fan and Lv (2008) points out, in practice the nonparametric independence screening (NIS) would still suffer from false negative (i.e., miss some important predictors that are marginally weakly correlated but jointly correlated with the response), and false positive (i.e., select some unimportant predictors which are highly correlated with the important ones). Therefore, we adopt an iterative framework to enhance the performance of this method. We repeatedly apply the large-scale variable screening (NIS) followed by a moderate-scale variable selection, where we use group-SCAD penalty as our selection strategy. In the NIS step, we propose two methods to determine a data-driven threshold for screening, which result in Conditional-INIS and Greedy-INIS, respectively.

4.1. Conditional-INIS Method

The conditional-INIS method builds upon *conditional* random permutation in determining the thresholding τ_n. Recall the random permutation used in Fan, Feng and Song (2011), which generalizes that Zhao and Li (2010). Randomly permute $\mathbf{Y}$ to get $\mathbf{Y}_{\boldsymbol{\pi}} = (Y_{\pi_1}, \cdots, Y_{\pi_n})^T$ and compute $\widehat{u}^{\boldsymbol{\pi}}_{nj}$, where $\boldsymbol{\pi}$ is a permutation of $\{1, \cdots, n\}$, based on the randomly coupled data $\{(Y_{\pi_i}, W_i, \mathbf{X}_i)\}_{i=1}^n$ that has no relationship between covariates and response. Thus, these estimates serve as the baseline of the marginal utilities under the null model (no relationship). To control the false selection rate at q/p under the null model, one would choose the screening threshold be τ_q, the qth-ranked magnitude of $\{\widehat{u}^{\boldsymbol{\pi}}_{nj},\ j = 1, \cdots, p\}$. Thus, the NIS step selects variables $\{j : \widehat{u}_{nj} \ge \tau_q\}$. In practice, one frequently uses $q = 1$, namely, the largest marginal utility under the null model.

When the correlations among covariates are large, there will be hardly any differentiability between the marginal utilities of the true variables and the false ones. This makes the selected variable set very large to begin with and hard to proceed the rest of iterations with limited false positives. For numerical illustrations, see section 5.2. Therefore, we propose a *conditional* permutation method to tackle this problem. Combining the other steps, our Conditional-INIS algorithm proceeds as follows.

0. For $j = 1, \cdots, p$, compute

$$\widehat{u}_{nj} = \|\widehat{a}_{nj}(\mathbf{W}) + \widehat{b}_{nj}(\mathbf{W})\mathbf{X}_j\|_n^2 - \|\widehat{a}_{n0}(\mathbf{W})\|_n^2,$$

where the estimates are defined in (9) and (10) using $\{(\mathbf{Y}, \mathbf{W}, \mathbf{X}_j), j = 1, \cdots, p\}$. Select the top K variables by ranking their marginal utilities $\widehat{u}_{nj}$, resulting in the index subset $\mathcal{M}_0$ to condition upon.

1. Regress $\mathbf{Y}$ on $\{(\mathbf{W}, \mathbf{X}_j), j \in \mathcal{M}_0\}$, and get intercept $\widehat{\beta}_{n0}(W)$ and their functional coefficients' estimators $\{\widehat{\beta}_{nj}(W), j \in \mathcal{M}_0\}$. Conditioning on $\mathcal{M}_0$, the n-dimensional partial residual is

$$\mathbf{Y}^* = \mathbf{Y} - \widehat{\beta}_{n0}(\mathbf{W}) - \sum_{j \in \mathcal{M}_0} \mathbf{X}_j \widehat{\beta}_{nj}(\mathbf{W}).$$

For all $j \in \mathcal{M}_0^c$, compute $\widehat{u}_{nj}^*$ using $\{(\mathbf{Y}^*, \mathbf{W}, \mathbf{X}_j), j \in \mathcal{M}_0^c\}$, which measures the additional utility of each covariate conditioning on the selected set $\mathcal{M}_0$.

To determine the threshold for NIS, we apply random permutation on the partial residual $\mathbf{Y}^*$, which yields $\mathbf{Y}^*_{\boldsymbol{\pi}}$. Compute $\widehat{u}_{nj}^{*\boldsymbol{\pi}}$ based on the decoupled data $\{(\mathbf{Y}^*_{\boldsymbol{\pi}}, \mathbf{W}, \mathbf{X}_j), j \in \mathcal{M}_0^c\}$. Let τ_q^* be the qth-ranked magnitude of $\{\widehat{u}_{nj}^{*\boldsymbol{\pi}}, j \in \mathcal{M}_0^c\}$. Then, the active variable set of variables is chosen as

$$\mathcal{A}_1 = \{j : \widehat{u}_{nj}^* \geq \tau_q^*, j \in \mathcal{M}_0^c\} \cup \mathcal{M}_0.$$

In our numerical studies, $q = 1$.

2. Apply the group-SCAD penalty on $\mathcal{A}_1$ to select a subset of variables $\mathcal{M}_1$. Details about the implementation of SCAD will be described later.

3. Repeat step 1-2, where we replace $\mathcal{M}_0$ in step 1 by $\mathcal{M}_l$, $l = 1, 2, \cdots$, and get $\mathcal{A}_{l+1}$ and $\mathcal{M}_{l+1}$ in step 2. Iterate until $\mathcal{M}_{l+1} = \mathcal{M}_k$ for some $k \leq l$ or $|\mathcal{M}_{l+1}| \geq \zeta_n$, for some prescribed positive integer ζ_n.

4.2. Greedy-INIS Method

Following Fan, Feng and Song (2011), we also implemented a greedy version of INIS method. We skip step 0 and start from step 1 in the algorithm above (i.e., take $\mathcal{M}_0 = \emptyset$), and select the top p_0 variables that have the largest marginal norms $\widehat{u}_{nj}$. This NIS step is followed by the same group-SCAD penalized regression as in step 2. We then iterate these steps until there are two identical subsets or the number of variables selected exceeds a prespecified ζ_n. In our simulation studies, p_0 is set as 1.

4.3. Implementation of SCAD

In the group-SCAD step, variables are selected as $\mathcal{M}_l = \{j \in \mathcal{A}_l : \widehat{\boldsymbol{\gamma}}_j^{(l)} \neq \mathbf{0}\}$ through minimizing the following objective function:

$$\min_{\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_j \in \mathbb{R}^{L_n}} \frac{1}{n} \sum_{i=1}^n \Big(Y_i - \mathbf{B}(W_i)\boldsymbol{\gamma}_0 - \sum_{j \in \mathcal{A}_l} \mathbf{B}(W_i) X_{ji} \boldsymbol{\gamma}_j\Big)^2 + \sum_{j \in \mathcal{A}_l} p_\lambda(||\boldsymbol{\gamma}_j||_B), \tag{34}$$

where $||\boldsymbol{\gamma}_j||_B = \sqrt{\frac{1}{n}\sum_{i=1}^n (\sum_{k=1}^{L_n} B_{jk}(W_i)\gamma_{jk})^2}$, and $p_\lambda(\cdot)$ is the SCAD penalty such that

$$p'_\lambda(|x|) \quad = \quad \lambda I(|x| \leq \lambda) + \frac{(a\lambda - |x|)_+}{a - 1} I(|x| > \lambda),$$

with $p_\lambda(0) = 0$. We set $a = 3.7$ as suggested and solve the optimization above via local quadratic approximations (Fan and Li, 2001). λ is chosen by BIC criteria $n \log(\widehat{\sigma}_\epsilon^2) + kL_n \log n$, where k is the number of covariates chosen. By Antoniadis and Fan (2001) and Yuan and Lin (2006), the norm-penalty in (34) encourages the group selection.

5. Numerical Studies

In this section, we carry out several simulation studies to assess the performance of our proposed methods. If not otherwise stated, the common setup for the following simulations are: cubic B-spline, $L_n = 7$, sample size $n = 400$, the number of variables $p = 1000$, and the number of simulations $N = 200$ for each example.

5.1. Comparison of Minimum Model Size

In this study, as in Fan and Song (2010), we illustrate the performance of NIS method in terms of the minimum model size (MMS) needed to include all the true variables, i.e., to possess sure screening property.

Example 1 Following Fan and Song (2010), we first consider a linear model as a special case of the varying coefficient model. Let $\{X_k\}_{k=1}^{950}$ be i.i.d. standard normal random variables and

$$X_k = \sum_{j=1}^{s}(-1)^{j+1}X_j/5 + \sqrt{1-\frac{s}{25}}\xi_k, \quad k = 951, \cdots, 1000,$$

where $\{\xi_k\}_{k=951}^{1000}$ are standard normal random variables. We construct the following model: $Y = \boldsymbol{\beta}^T\mathbf{X}+\epsilon$, where $\epsilon \sim \mathcal{N}(0, \sqrt{3}^2)$ and $\boldsymbol{\beta} = (1, -1, 1, -1, \cdots)^T$ has s nonzero components. To carry out NIS, we define an exposure W independently from the standard uniform distribution.

We compare NIS, Lasso and SIS (independence screening for linear models). The boxplots of minimum model size are presented in Figure 1. Note that when $s > 5$, the irrepresentable condition fails, and Lasso performs badly even in terms of pure screening. On the other hand, SIS performs better than NIS because the coefficients are indeed constant, and there are fewer parameters (p) involved in SIS than those of NIS (pL_n).

Example 2 For the second example, we illustrate that when the underlying model's coefficients are indeed varying, we do need nonparametric independence screening. Let $\{U_1, U_2, \cdots U_{p+2}\}$ be i.i.d. uniform random variables on $[0, 1]$, based on which we construct $\mathbf{X}$ and W as follows:

$$X_j = \frac{U_j + t_1U_{p+1}}{1+t_1}, \quad j = 1, \cdots, p, \qquad W = \frac{U_{p+2} + t_2U_{p+1}}{1+t_2},$$

where t_1 and t_2 controls the correlation among the covariates $\mathbf{X}$ and the correlation between $\mathbf{X}$ and W, respectively. When $t_1 = 0$, X_j's are uncorrelated, and when $t_1 = 1$ the correlation is 0.5. If $t_1 = t_2 = 1$, X_j's and W are also correlated with correlation coefficient 0.5.

For the varying coefficients part, we take coefficient functions

$$\beta_1(W) = W, \quad \beta_2(W) = (2W-1)^2, \quad \beta_3(W) = \sin(2\pi W).$$

The true data generation model is

$$Y = 5\beta_1(W)\cdot X_1 + 3\beta_2(W)\cdot X_2 + 4\beta_3(W)\cdot X_3 + \epsilon,$$

where ϵ's are i.i.d. standard Gaussian random variable.

Under different correlation settings, the comparison MMS between NIS and SIS methods are presented in Figure 2. When the correlation gets stronger, independence screening becomes harder.

5.2. Comparison of Permutation and Conditional Permutation

In this section, we illustrate the performance the conditional random permutation method.

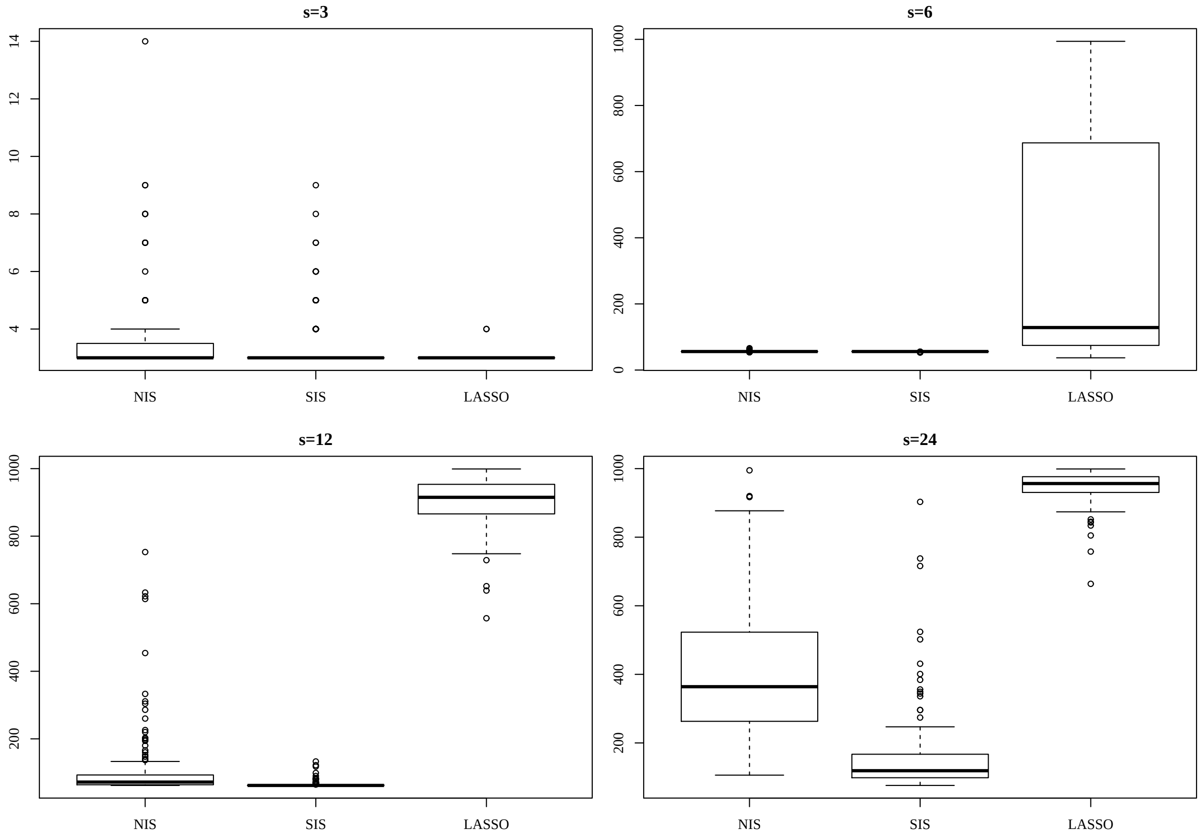


Fig. 1. Boxplots of minimum model sizes (left to right: NIS, Lasso and SIS) for Example 1 under different true models.

Example 3 Let $\{Z_1, \cdots, Z_p\}$ be i.i.d. standard normal, $\{U_1, U_2\}$ be i.i.d. standard uniformly distributed random variables, and the noise ϵ follows the standard normal distribution. We construct $\{W, \mathbf{X}\}$ and Y as follows:

$$\begin{aligned} X_j &= \frac{Z_j + t_1 U_1}{1 + t_1}, j = 1, \cdots, p, \qquad W = \frac{U_2 + t_2 U_1}{1 + t_2}, \\ Y &= 2X_1 + 3W \cdot X_2 + (W+1)^2 \cdot X_3 + \frac{4\sin(2\pi W)}{2 - \sin(2\pi W)} \cdot X_4 + \epsilon. \end{aligned}$$

We will take $t_1 = t_2 = 0$, resulting in uncorrelated case and $t_1 = 3$ and $t_2 = 1$, corresponding to $\text{corr}(X_j, X_k) = 0.43$ for all $j \neq k$ and $\text{corr}(X_j, W) = 0.46$. By taking $q = 1$ (i.e., take the maximum value of the marginal utility of the permuted estimates), we report the average of the true positive number (TP), model size, the lower bound of the marginal signal of true variables and the upper bound of the marginal signal of false variables for different correlation settings based on 200 simulations. Their robust standard deviations are also reported therein.

Based on Table 1, we see that when the correlation gets stronger, although sure screening properties can be achieved most of the time via unconditional ($K = 0$) random permutation thresholding, the model size becomes very large and therefore the false selection rate is high. The reason is that there is no differentiability between the marginal signals of the true variables and the false ones. This drawback makes the original random permutation not a feasible method to determine the screening threshold in practice.

We now applied the conditional permutation method, whose performance is illustrated in Table 1 for a few choices of tuning parameter K. The screening threshold is taken as τ_q with $q = 1$. Generally speaking, although the lower bound of the true positives' signals may be smaller than the upper bound of false variables' signals, the largest K norms still have a high possibility to contain at least some true variables. When conditioning on this small set of more relevant variables, the marginal contributions

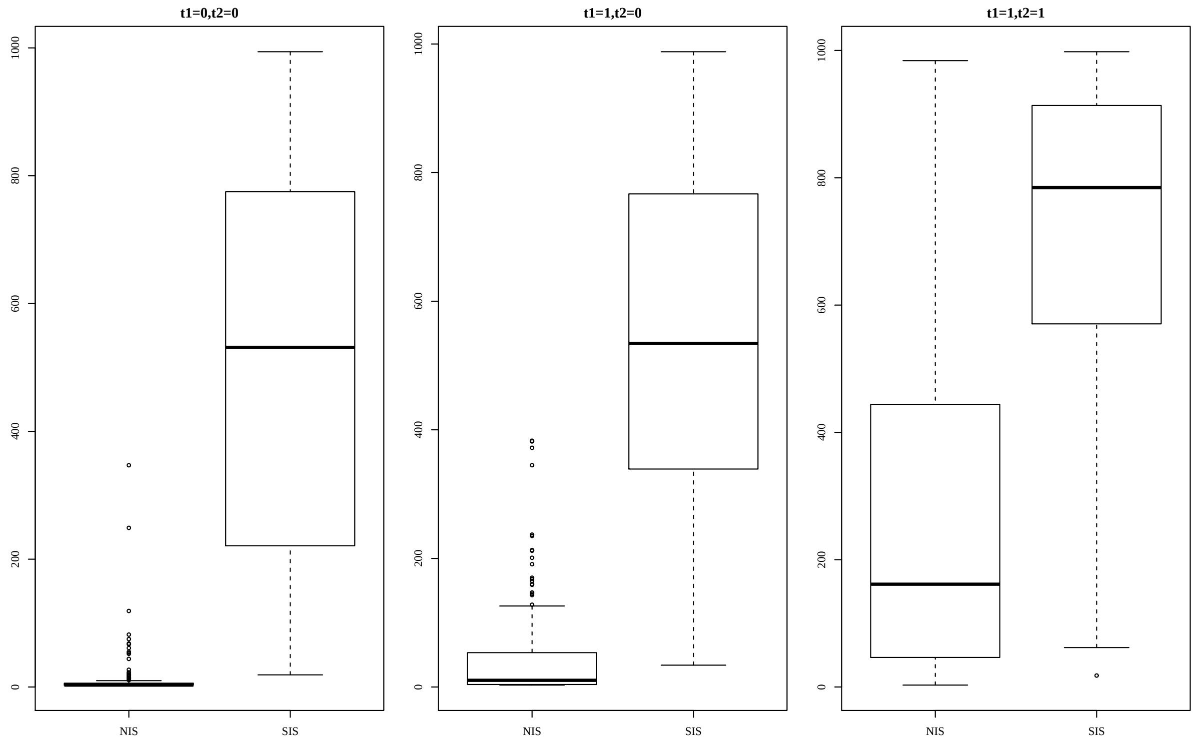


Fig. 2. Boxplots of minimum model sizes (left: NIS, right: SIS) for Example 2 under different correlation settings.

Table 1. Model size and marginal signals under different correlation settings (Example 3)

Model		TP	Size	$\min_{j\in\mathcal{M}^*\backslash\mathcal{M}_0} \widehat{u}^*_{nj}$	$\max_{j\in\mathcal{M}^{*c}\backslash\mathcal{M}_0} \widehat{u}^*_{nj}$	$\max_{j\in\{1,\cdots,p\}\backslash\mathcal{M}_0} \widehat{u}^{*\boldsymbol{\pi}}_{nj}$
K=0	$t_1 = 0, t_2 = 0$	4.00(0)	6.68(2.99)	2.96(0.72)	1.22(0.18)	1.12(0.15)
	$t_1 = 3, t_2 = 1$	4.00(0)	886.49(88.81)	0.61(0.10)	0.58(0.07)	0.22(0.03)
K=1	$t_1 = 0, t_2 = 0$	4.00(0)	5.70(1.49)	2.83(0.57)	0.75(0.10)	0.72(0.11)
	$t_1 = 3, t_2 = 1$	4.00(0)	202.50(154.85)	0.28(0.06)	0.20(0.03)	0.11(0.02)
K=4	$t_1 = 0, t_2 = 0$	4.00(0)	5.14(1.49)	NA	0.06(0.01)	0.06(0.01)
	$t_1 = 3, t_2 = 1$	4.00(0)	4.98(0.75)	0.16(0.05)	0.05(0.01)	0.06(0.01)
K=8	$t_1 = 0, t_2 = 0$	4.00(0)	8.92(0.75)	NA	0.05(0.01)	0.05(0.01)
	$t_1 = 3, t_2 = 1$	3.99(0)	8.43(0.75)	0.11(0.03)	0.04(0.01)	0.05(0.01)

of false positives get weaker. Note that in the absence of correlation, when $K \geq s$ (here $s = 4$), the first K variables have already included all the true variables (i.e., $\mathcal{M}^*\backslash\mathcal{M}_0 = \emptyset$), hence the minimum of true signal is not available. In other cases, we see that the gap between the marginal signals of true variables and false variables become large enough to differentiate them. Table 1 shows that by using the thresholding via the conditional permutation method, not only the sure screening properties are still maintained, but also the model sizes are dramatically reduced.

5.3. Comparison of Model Selection and Estimation

In this section we explore the performance of Conditional-INIS and Greedy-INIS method. In our iterative framework, conditional permutation serves as the initialization step (step 0) and we take $K = 5$ in the rest of the paper. For each method, we report the average number of true positive (TP), false positive (FP), prediction error (PE), and their robust standard deviations. Here the prediction error is the mean squared error calculated on the test dataset of size $n/2 = 200$ generated from the same model. As a measure of the complexity of the model, signal-to-noise-ratio (SNR), defined by $\mathrm{var}(\boldsymbol{\beta}^T(W)\mathbf{X})/\mathrm{var}(\epsilon)$, is computed. Table 2 reports the results using the simulated model specified in Example 3. We now illustrate the performance by using another example.

Table 2. Average values of the number of true positives (TP), false positives (FP), and prediction error (PE) for simulated model in Example 3. Robust standard deviations are given in parentheses.

Model	Correlation		Conditional-INIS			Greedy-INIS		
	X's	X's-W	TP	FP	PE	TP	FP	PE
$t_1 = 0, t_2 = 0$	0	0	4	0.54	1.10	4	13.01	1.41
(SNR $\approx$ 16.85)			(0)	(0.75)	(0.05)	(0)	(3.73)	(0.17)
$t_1 = 2, t_2 = 0$	0.25	0	4	0.20	0.78	4	0.41	1.10
(SNR $\approx$ 3.66)			(0)	(0)	(0.06)	(0)	(0)	(0.05)
$t_1 = 2, t_2 = 1$	0.25	0.36	3.97	0.26	1.27	3.90	0.14	1.63
(SNR $\approx$ 3.21)			(0)	(0)	(0.24)	(0)	(0)	(0.41)
$t_1 = 3, t_2 = 0$	0.43	0	4	0.19	1.03	3.99	0.57	1.22
(SNR $\approx$ 3.32)			(0)	(0)	(0.06)	(0)	(0)	(0.07)
$t_1 = 3, t_2 = 1$	0.43	0.46	3.95	0.31	1.30	3.77	0.27	1.29
(SNR $\approx$ 2.81)			(0)	(0.75)	(0.12)	(0)	(0)	(0.17)

Table 3. Average values of the number of true positives (TP), false positives (FP), and prediction error (PE) for the model in Example 4. Robust standard deviations are given in parentheses.

Model	Correlation		Conditional-INIS			Greedy-INIS		
	X's	X's-W	TP	FP	PE	TP	FP	PE
$t_1 = 0, t_2 = 0$	0	0	8	0.21	1.24	8	10.71	1.57
(SNR $\approx$ 47.68)			(0)	(0)	(0.09)	(0)	(3.73)	(0.20)
$t_1 = 2, t_2 = 0$	0.25	0	8	0.13	1.17	8	0.60	1.16
(SNR $\approx$ 9.40)			(0)	(0)	(0.09)	(0)	(0)	(0.10)
$t_1 = 2, t_2 = 1$	0.25	0.36	7.80	0.20	2.16	7.55	0.26	2.26
(SNR $\approx$ 8.62)			(0)	(0)	(0.58)	(0.75)	(0)	(0.70)
$t_1 = 3, t_2 = 0$	0.43	0	7.90	0.10	1.21	7.98	0.71	1.29
(SNR $\approx$ 8.18)			(0)	(0)	(0.12)	(0)	(0)	(0.10)
$t_1 = 3, t_2 = 1$	0.43	0.46	7.75	0.18	1.65	7.35	0.28	1.84
(SNR $\approx$ 7.61)			(0)	(0)	(0.26)	(0.75)	(0)	(0.42)

Example 4 Let $\{W, \mathbf{X}\}$, Y and ϵ be the same as in *Example 3*. We now introduce more complexities in the following model:

$$\begin{aligned} Y &= 3W \cdot X_1 + (W+1)^2 \cdot X_2 + (W-2)^3 \cdot X_3 + 3(\sin(2\pi W)) \cdot X_4 \\ &\quad + \exp(W) \cdot X_5 + 2 \cdot X_6 + 2 \cdot X_7 + 3\sqrt{W} \cdot X_8 + \epsilon. \end{aligned}$$

The results are present in Table 3.

Through the examples above, Conditional-INIS and Greedy-INIS show comparable performance in terms of TP, FP and PE. When the covariates are independent or weakly correlated, sure screening is easier to achieve and false positive is rare; as the correlation gets stronger, we see a decrease in TP and an increase in FP. It seems that Greedy-INIS selects slightly more false positives than Conditional-INIS, the reason being that in each step Greedy-INIS selects the top variable(s) by fitting the residuals conditional on previously chosen variable set and tends to overfit. However, the coefficient estimates for these false positives are fairly small, hence they do not affect prediction error very much. Regarding computation efficiency, Conditional-INIS performs better in our simulated examples, as it usually only requires two to three iterations, while Greedy-INIS would need at least s/p_0 iterations (here $p_0 = 1$ and $s = 4$ and 8 respectively for Examples 3 and 4).

5.4. *Real Data Analysis on Boston Housing Data*

In this section we illustrate the performance of our method through a real data analysis on Boston Housing Data (Harrison and Rubinfeld, 1978). This dataset contains housing data for 506 census tracts of Boston from the 1970 census. Most empirical results for the housing value equation are based on a common specification (Harrison and Rubinfeld, 1978),

$$\log(\text{MV}) = \beta_0 + \beta_1\ \text{RM}^2 + \beta_2\text{AGE} + \beta_3 \log(\text{DIS}) + \beta_4 \log(\text{RAD}) + \beta_5\text{TAX}$$

Table 4. Prediction error (PE) , model size and selected noise variables (SNV) over 100 repetitions and their robust standard deviations (in parentheses) for Conditional-INIS (p = 1000), Greedy-INIS (p = 1000), and SCAD fit (p=12).

method	PE	Size	SNV
Conditional-INIS (p = 1000)	0.046(0.020)	5.55(0.75)	0(0)
Greedy-INIS (p = 1000)	0.048(0.020)	4.80(1.49)	0.01(0)
SCAD fit (p=12)	0.052(0.019)	6.05(1.87)	*NA*

$$+\beta_6\text{PTRATIO} + \beta_7(\text{B} - 0.63)^2 + \beta_8 \log(\text{LSTAT}) + \beta_9\text{CRIM}$$
$$+\beta_{10}\text{ZN} + \beta_{11}\text{INDUS} + \beta_{12}\text{CHAS} + \beta_{13}\text{NOX}^2 + \epsilon,$$

where the dependent variable MV is the median value of owner-occupied homes, the independent variables are quantified measurement of its neighborhood whose description can be found in the manual of R package *mlbench*. The common specification uses RM^2 and NOX^2 to get a better fit, and for comparison we take these transformed variables as our input variables.

To exploit the power of varying coefficient model, we take the variable log(DIS), the weighted distances to five employment centers in the Boston region, as the exposure variable. This allows us to examine how the distance to the business hubs interact with other variables. It is reasonable to assume that the impact of other variables on housing price varies with the distance, which is an important characteristic of the neighborhood, i.e. the geographical accessibility to employment. Interestingly, Conditional-INIS selects the following submodel:

$$\begin{aligned}\log(\text{MV}) &= \beta_0(W) + \beta_1(W)\cdot\text{RM}^2 + \beta_2(W)\cdot\text{AGE} + \beta_5(W)\cdot\text{TAX} \\ &\quad +\beta_7(W)\cdot(\text{B} - 0.63)^2 + \beta_9(W)\cdot\text{CRIM} + \epsilon, \end{aligned} \tag{35}$$

where $W = \log(\text{DIS})$. The estimated functions $\widehat{\beta}_j(W)$'s are presented in Figure 3. This varying coefficient model shows very interesting aspects of housing valuation. The evidence of nonlinear interactions with the accessibility is clearly evidenced. For example, RM is the average number of rooms in owner units, which represents the size of a house. Therefore, the marginal cost of a big house is higher in employment centers where population is concentrated and supply of mansions is limited. The cost per room decreases as one moved away from the business centers and then gradually increases. CRIM is the crime rate in each township, which usually has a negative impact, and from its varying coefficient we see that it is a bigger concern near (demographically more complex) business centers. AGE is the proportion of owner units built prior to 1940, and its varying coefficient has a parabola shape: positive impact on housing values near employment centers and suburb areas, while negative effects in between. NOX (air pollution level) is generally a negative impact, and the impact is larger when the house is near employment centers where air is presumably more polluted than suburb area.

We now evaluate the performance of our INIS method in a high dimensional setting. To accomplish this, let $\{Z_1, \cdots, Z_p\}$ be i.i.d. the standard normal random variables and U follow the standard uniform distribution. We then expand the data set by adding the artificial predictors:

$$X_j = \frac{Z_j + tU}{1+t}, j = s+1, \cdots, p.$$

Note that $\{W, X_1, \cdots, X_s\}$ are the independent variables in original data set ($s = 13$ here) and the variables $\{X_j\}_{j=s+1}^p$ are known to be irrelevant to the housing price, though the maximum spurious correlation of these 987 artificial predictors to the housing price is now small. We take $p = 1000$, $t = 2$, and randomly select $n = 406$ samples as training set, and compute prediction mean squared error (PE) on the rest 100 samples. As a benchmark for comparison, we also do regression fit on $\{W, X_1, \cdots, X_s\}$ directly using SCAD penalty without screening procedure. We repeat $N = 100$ times and report the average prediction error and model size, and their robust standard deviation. Since $\{X_j\}_{j=s+1}^p$ are artificial variables, we also include the number of artificial variables selected by each method as a proxy for false positives. The results are presented in Table 4.

As seen from Table 4, our methods are very effective in filtering noise variables in a high dimensional setting, and can achieve comparable prediction error as if the noise were absent. In conclusion, the proposed INIS methodology is very useful in high-dimensional scientific discoveries, which can select a parsimonious close-to-truth model and reveal interesting relationship between variables, as illustrated in this section.

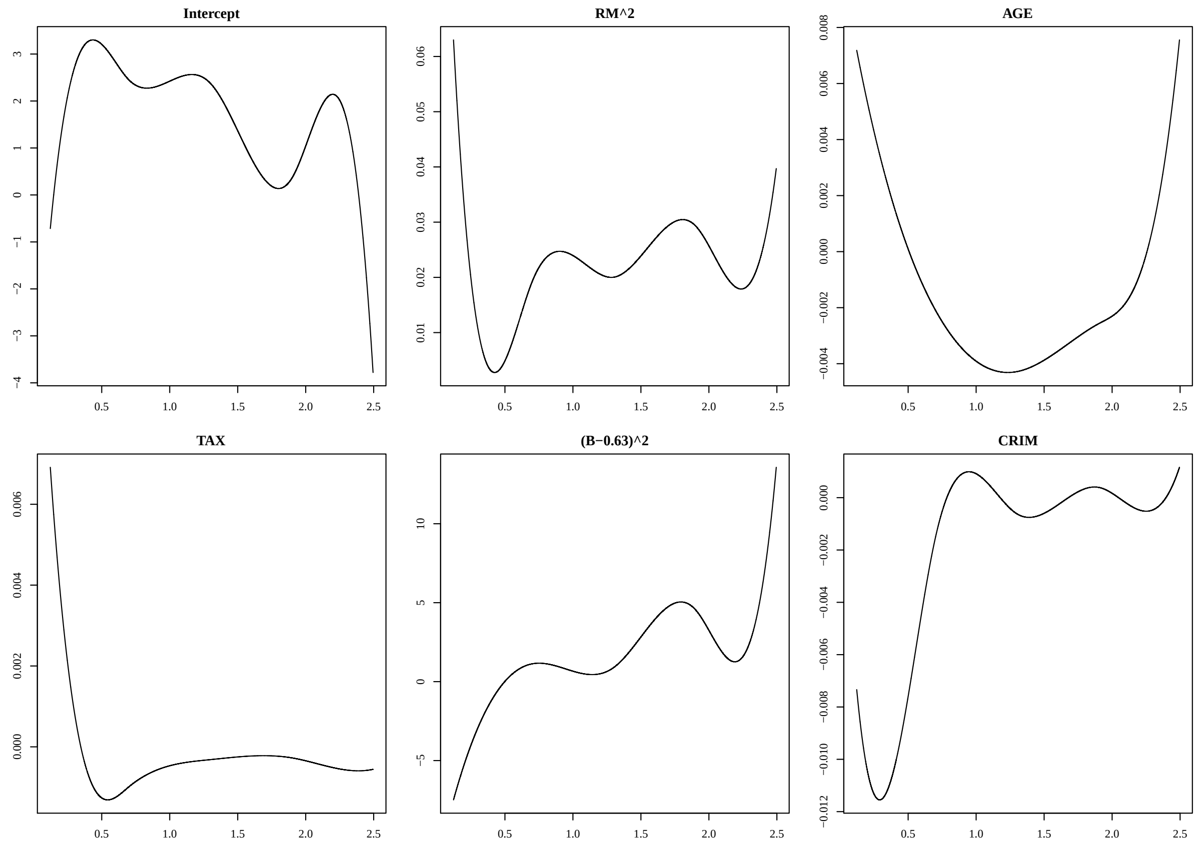


Fig. 3. Fitted functional estimates $\widehat{\beta}_j(W)'s$ selected by Conditional-INIS.

Acknowledgements

This project was supported by the National Institute of General Medical Sciences of the National Institutes of Health through Grant Numbers R01-GM072611 and R01-GMR01GM100474 and National Science Foundation grant DMS-1206464. The bulk of the research was carried while Yunbei Ma was a postdoctoral fellow at Princeton University.

Appendix

A.1. Properties of B-splines

Our estimation use the B-spline basis, which have the following properties (de Boor 1978): For each $j = 1, \cdots, p$ and $k = 1, \cdots, L_n$, $B_k(W) \geq 0$ and $\sum_{k=1}^{L_n} B_k(W) = 1$ for $W \in \mathcal{W}$. In addition, there exist positive constants T_3 and T_4 such that for any $\eta_k \in \mathbb{R}$, $k = 1, \cdots, L_n$,

$$L_n^{-1} T_3 \sum_{k=1}^{L_n} \eta_k^2 \leq \int \left(\sum_{k=1}^{L_n} \eta_k B_k(w) \right)^2 dw \leq L_n^{-1} T_4 \sum_{k=1}^{L_n} \eta_k^2. \tag{36}$$

Then under condition (iii), there exist positive constants C_1 and C_2 such that for $k = 1, \cdots, L_n$,

$$C_1 L_n^{-1} \leq \mathrm{E}[B_k^2(W)] \leq C_2 L_n^{-1}, \tag{37}$$

where $C_1 = T_1 T_3$ and $C_2 = T_2 T_4$.

Furthermore, under condition (iii), it follows from (36) that for any $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_{L_n})^T \in \mathbb{R}^{L_n}$ such that $\|\boldsymbol{\eta}\|_2^2 = 1$,

$$C_1 L_n^{-1} \leq \boldsymbol{\eta}^T \mathrm{E}[\mathbf{B}^T \mathbf{B}] \boldsymbol{\eta} \leq C_2 L_n^{-1}.$$

Or equivalently,

$$C_1 L_n^{-1} \le \lambda_{min}(\mathrm{E}[\mathbf{B}^T\mathbf{B}]) \le \lambda_{max}(\mathrm{E}[\mathbf{B}^T\mathbf{B}]) \le C_2 L_n^{-1}. \tag{38}$$

A.2. Technical Lemmas

Some technical lemmas needed for our main results are shown as follows. Lemma 1 and Lemma 2 give some characterization of exponential tails , which becomes handy in our proof. Lemma 3 and Lemma 4 is a Bernstein type inequality.

LEMMA 1. *Let X, W be random variables. Suppose X has a conditional exponential tail: $P(|X| > t|W) \le \exp(1-(t/K)^r)$ for all $t \ge 0$ and uniformly on the compact support of W, where $K > 0$ and $r \ge 1$. Then for all $m \ge 2$,*

$$\mathrm{E}(|X|^m|W) \le emK^m m!. \tag{39}$$

Proof. Recall that for any non-negative random variable Z, $\mathrm{E}[Z|W] = \int_0^\infty \mathrm{P}\{Z \ge t|W\}dt$. Then we have

$$\begin{aligned}\mathrm{E}(|X|^m|W) &= \int_0^\infty \mathrm{P}\{|X|^m \ge t|W\}dt\\ &\le \int_0^\infty \exp(1-(t^{1/m}/K)^r)dt\\ &= \frac{emK^m}{r}\Gamma(\frac{m}{r}).\end{aligned}$$

The lemma follows from the fact $r \ge 1$.

LEMMA 2. *Let Z_1, Z_2 and W be random variables. Suppose that there exist K_1, $K_2 > 0$ and r_1, $r_2 \ge 1$ such that $r_1 r_2/(r_1+r_2) \ge 1$, and*

$$P(|Z_i| > t|W) \le \exp(1-(t/K_i)^{r_i}), \quad i = 1,2$$

for all $t \ge 0$ and uniformly on $\mathcal{W}$. Then for some $r^* \ge 1$ and $K^* > 0$,

$$P(|Z_1 Z_2| > t|W) \le \exp(1-(t/K^*)^{r^*}) \tag{40}$$

for all $t \ge 0$ and uniformly on $\mathcal{W}$.

Proof. For any $t > 0$, let $M = (tK_2^{r_2/r_1}/K_1)^{\frac{r_1}{r_1+r_2}}$ and $r = r_1 r_2/(r_1+r_2)$. Then uniformly on $\mathcal{W}$, we have

$$\begin{aligned}\mathrm{P}(|Z_1 Z_2| > t|W) &\le \mathrm{P}(M|Z_1| > t|W) + \mathrm{P}(|Z_2| > M|W)\\ &\le \exp\{1-(t/K_1M)^{r_1}\} + \exp\{1-(M/K_2)^{r_2}\}\\ &= 2\exp\{1-(t/K_1K_2)^r\}.\end{aligned}$$

Let $r^* \in [1, r]$ and $K^* = \max\{(r^*/r)^{1/r}K_1K_2,\ (1+\log 2)^{1/r}K_1K_2\}$. It can be shown that $G(t) = (t/K_1K_2)^r - (t/K^*)^{r^*}$ is increasing when $t > K^*$. Hence $G(t) > G(K^*) \ge \log 2$ when $t > K^*$, which implies when $t > K^*$,

$$\mathrm{P}(|Z_1 Z_2| > t|W) \le 2\exp\{1-(t/K_1K_2)^{r_1}\} \le \exp\{1-(t/K^*)^{r^*}\}.$$

On the other hand, when $t \le K^*$,

$$\mathrm{P}(|Z_1 Z_2| > t|W) \le 1 \le \exp\{1-(t/K^*)^{r^*}\}.$$

Lemma 2 holds.

LEMMA 3. *(Bernstein inequality, lemma 2.2.11, van der Vaart and Wellner (1996)). For independent random variables $Y_1, \cdots, Y_n$ with mean zero such that $\mathrm{E}[|Y_i|^m] \le m!M^{m-2}\nu_i/2$ for every $m \ge 2$ (and all i) and some constants M and ν_i. Then*

$$P(|Y_1 + \cdots + Y_n| > x) \le 2\exp\{-x^2/(2(\nu + Mx))\},$$

for $v \ge \nu_1 + \cdots + \nu_n$.

LEMMA 4. *(Bernstein's inequality, lemma 2.2.9, van der Vaart and Wellner (1996)). For independent random variables $Y_1, \cdots, Y_n$ with bounded range $[-M, M]$ and mean zero,*

$$P(|Y_1 + \cdots + Y_n| > x) \leq 2\exp\{-x^2/(2(\nu + Mx/3))\},$$

for $\nu \geq var(Y_1 + \cdots + Y_n)$.

The following lemmas are needed for the proof of Theorem 1.

LEMMA 5. *Suppose conditions (i) and (iii)-(vi) hold. For any $\delta > 0$, there exist some positive constants b_1 and b_2 such that for $j = 1, \cdots, p$, $k = 1, \cdots, L_n$,*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_{ji}B_k(W_i)Y_i - \mathrm{E}[X_jB_kY]\right| \geq \frac{\delta}{n}\right) \leq 4\exp\left\{-\frac{\delta^2}{b_1L_n^{-1}n + b_2\delta)}\right\},$$

and

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} B_k(W_i)Y_i - \mathrm{E}[B_kY]\right| \geq \frac{\delta}{n}\right) \leq 4\exp\left\{-\frac{\delta^2}{b_1L_n^{-1}n + b_2\delta}\right\}.$$

Proof. Recall $m(\mathbf{X}_i^*) = E(Y_i|\mathbf{X}_i, W_i)$. Let $Z_{jki} = X_{ji}B_k(W_i)m(\mathbf{X}_i^*) - \mathrm{E}[X_jB_k(W)m(\mathbf{X}^*)]$ and $\xi_{jki} = X_{ji}B_k(W_i)\varepsilon_i$. Then

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} X_{ji}B_k(W_i)Y_i - \mathrm{E}[X_jB_k(W)Y]\right| \\
= &\left|\frac{1}{n}\sum_{i=1}^{n}\Big(X_{ji}B_k(W_i)m(\mathbf{X}_i^*) - \mathrm{E}[X_jB_k(W)m(\mathbf{X}^*)] + X_{ji}B_k(W_i)\varepsilon_i\Big)\right| \\
\leq &\left|\frac{1}{n}\sum_{i=1}^{n} Z_{jki}\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\xi_{jki}\right|.
\end{aligned}$$

We first bound $\frac{1}{n}\sum_{i=1}^{n} Z_{jki}$. Note that for each j and k, $\{Z_{jki}\}_{i=1}^{n}$ are a sequence of independent random variables with mean zero. By condition (v), (37), and Lemmas 1 and 2, we have for every $m \geq 2$, there exists a constant $K_4 > 0$, such that

$$\begin{aligned}
\mathrm{E}|Z_{jki}|^m &\leq 2^m\mathrm{E}|X_{ji}B_k(W_i)m(\mathbf{X}_i^*)|^m \\
&\leq 2^m\mathrm{E}[B_k^m(W_i)\mathrm{E}[|X_{ji}m(\mathbf{X}_i^*)|^m|W_i]] \\
&\leq 2^m\mathrm{E}[B_{jk}^2(W_i)emK_4^m m!] \\
&\leq m!(2K_4)^{m-2}(8emK_4^2C_2L_n^{-1})/2, \qquad (41)
\end{aligned}$$

where the first inequality comes from the Minkowski inequality. Hence, it follows from Lemma 3 that for any $\delta > 0$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_{jki}\right| \geq \frac{\delta}{2n}\right) \leq 2\exp\left\{-\frac{\delta^2}{64emK_4^2C_2L_n^{-1}n + 8K_4\delta}\right\} \qquad (42)$$

Next we bound $\frac{1}{n}\sum_{i=1}^{n}\xi_i$. Again ξ_i's are centered independent random variables. By conditions (v)-(vi), (37), and Lemmas 1 and 2, we have for every $m \geq 2$, there exists a constant $K_5 > 0$, such that

$$\begin{aligned}
\mathrm{E}|\xi_i|^m &= \mathrm{E}[B_k^m(W_i)\mathrm{E}[|X_{ji}\varepsilon_i|^m|W_i]] \\
&\leq m!K_5^{m-2}(2emK_5^2C_2L_n^{-1})/2.
\end{aligned}$$

Thus, according to Lemma 3,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right| \geq \frac{\delta}{2n}\right) \leq 2\exp\left\{-\frac{\delta^2}{16emK_5^2C_2L_n^{-1}n + 4K_5\delta}\right\}. \qquad (43)$$

Similarly, we can show that

$$P\left(\Big|\frac{1}{n}\sum_{i=1}^{n}B_k(W_i)m(\mathbf{X}_i^*)-\mathrm{E}[B_k(W)m(\mathbf{X}^*)]\Big|\geq\frac{\delta}{2n}\right)$$
$$\leq\ 2\exp\left\{-\frac{\delta^2}{64emK_2^2C_2L_n^{-1}n+8K_2\delta}\right\} \tag{44}$$

and

$$P\left(\Big|\frac{1}{n}\sum_{i=1}^{n}B_k(W_i)\varepsilon_i\Big|\geq\frac{\delta}{2n}\right)\leq 2\exp\left\{-\frac{\delta^2}{16emK_3^2C_2L_n^{-1}n+4K_3\delta}\right\}. \tag{45}$$

Let $b_1=16emC_2\max(4K_4^2,K_5^2,4K_2^2,K_3^2)$ and $b_2=\max(8K_4,4K_5,8K_2,4K_3)$. Then, the combination of (42) - (45) by union bound of probability yields the desired result. □

LEMMA 6. *Under conditions (i), (iii) and (v), there exist positive constants C_3 and C_4, such that for $j=1,\cdots,p$,*

$$C_3L_n^{-1}\leq\lambda_{min}(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])\leq\lambda_{max}(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])\leq C_4L_n^{-1}. \tag{46}$$

Proof. Recall that $\mathbf{Q}_j=(\mathbf{B},X_j\mathbf{B})$. For any $\boldsymbol{\eta}=(\boldsymbol{\eta}_1^T,\boldsymbol{\eta}_2^T)^T\in\mathbb{R}^{2L_n}$ such that $\|\boldsymbol{\eta}\|_2^2=1$,

$$\boldsymbol{\eta}^T\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]\boldsymbol{\eta}\ =\ \mathrm{E}\left[(\mathbf{B}\boldsymbol{\eta}_1,\mathbf{B}\boldsymbol{\eta}_2)\begin{pmatrix}1 & \mathrm{E}[X_j|W]\\ \mathrm{E}[X_j|W] & \mathrm{E}[X_j^2|W]\end{pmatrix}\begin{pmatrix}\mathbf{B}\boldsymbol{\eta}_1\\ \mathbf{B}\boldsymbol{\eta}_2\end{pmatrix}\right].$$

Consider eigenvalues λ_1 and λ_2 ($\lambda_1>\lambda_2$) of the 2×2 middle matrix on the right hand side of the equation above, we have $\lambda_1+\lambda_2=1+\mathrm{E}[X_j^2|W]$ (trace) and $\lambda_1\cdot\lambda_2=\mathrm{Var}[X_j|W]$ (determinant). Therefore, by Lemma 1

$$\lambda_1\leq 1+\mathrm{E}[X_j^2|W]\leq 1+4eK_1^2$$

and by assumption (i)

$$\lambda_2\geq\frac{\mathrm{Var}[X_j|W]}{\mathrm{E}[X_j^2|W]+1}\geq\frac{h_1}{1+4eK_1^2}.$$

Using the above two bounds on the minimum and maximum eigenvalues, we have

$$\frac{h_1}{1+4eK_1^2}\mathrm{E}[(\mathbf{B}\boldsymbol{\eta}_1)^2+(\mathbf{B}\boldsymbol{\eta}_2)^2]\leq\boldsymbol{\eta}^T\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]\boldsymbol{\eta}\leq(1+4eK_1^2)\mathrm{E}[(\mathbf{B}\boldsymbol{\eta}_1)^2+(\mathbf{B}\boldsymbol{\eta}_2)^2].$$

By (38), we have

$$\frac{h_1C_1}{1+4eK_1^2}L_n^{-1}\leq\boldsymbol{\eta}^T\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]\boldsymbol{\eta}\leq(1+4eK_1^2)C_2L_n^{-1}.$$

Take $C_3=h_1C_1L_n^{-1}/(1+4eK_1^2)$ and $C_4=(1+4eK_1^2)C_2L_n^{-1}$, result follows.

Throughout the rest of the proof, for any matrix $\mathbf{A}$, let $\|\mathbf{A}\|=\sqrt{\lambda_{\max}(\mathbf{A}^T\mathbf{A})}$ be the operator norm and $\|\mathbf{A}\|_\infty=\max_{i,j}|A_{ij}|$ be the infinity norm.

LEMMA 7. *Suppose conditions (i), (iii) and (v) hold. For any $\delta>0$ and $j=1,\cdots,p$, there exist some positive constants b_3 and b_4 such that*

$$P\left(\Big\|\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj}-\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]\Big\|\geq L_n\delta/n\right)\leq 6L_n^2\exp\left\{-\frac{\delta^2}{b_3L_n^{-1}n+b_4\delta}\right\},$$

and

$$P\left(\Big\|\frac{1}{n}\mathbf{B}_n^T\mathbf{B}_n-\mathrm{E}[\mathbf{B}^T\mathbf{B}]\Big\|\geq L_n\delta/n\right)\leq 6L_n^2\exp\left\{-\frac{\delta^2}{b_3L_n^{-1}n+b_4\delta}\right\}.$$

In addition, for any given positive constant b_5, there exists some positive constant b_6 such that

$$P\left(\left|\|(\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj})^{-1}\| - \|(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])^{-1}\|\right| \geq b_5\|(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])^{-1}\|\right) \leq 6L_n^2\exp\{-b_6L_n^{-3}n\},$$

and for any positive constant b_7, there exists some positive constant b_8 such that

$$P\left(\left|\left\|(\frac{1}{n}\mathbf{B}_n^T\mathbf{B}_n)^{-1}\right\| - \left\|(\mathrm{E}[\mathbf{B}^T\mathbf{B}])^{-1}\right\|\right| \geq b_7\|(\mathrm{E}[\mathbf{B}^T\mathbf{B}])^{-1}\|\right) \leq 6L_n^2\exp\{-b_8L_n^{-3}n\}.$$

Proof. Observe that for $j = 1, \cdots, p$,

$$\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj} - \mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j] = \begin{pmatrix} \mathbf{D}_1 & \mathbf{D}_{2j} \\ \mathbf{D}_{2j}^T & \mathbf{D}_{3j} \end{pmatrix},$$

where $\mathbf{D}_1 = \frac{1}{n}\sum_{i=1}^n \mathbf{B}^T(W_i)\mathbf{B}(W_i) - \mathrm{E}[\mathbf{B}^T\mathbf{B}]$, $\mathbf{D}_{2j} = \frac{1}{n}\sum_{i=1}^n X_{ji}\mathbf{B}^T(W_i)\mathbf{B}(W_i) - \mathrm{E}[X_j\mathbf{B}^T\mathbf{B}]$ and $\mathbf{D}_{3j} = \frac{1}{n}\sum_{i=1}^n X_{ji}^2\mathbf{B}^T(W_i)\mathbf{B}(W_i) - \mathrm{E}[X_j^2\mathbf{B}^T\mathbf{B}]$. Then

$$\begin{aligned} \|\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj} - \mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]\| &\leq 2L_n\|\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj} - \mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]\|_\infty \\ &= 2L_n\max(\|\mathbf{D}_1\|_\infty,\ \|\mathbf{D}_{2j}\|_\infty,\ \|\mathbf{D}_{3j}\|_\infty). \end{aligned} \quad (47)$$

We first bound $\|\mathbf{D}_1\|_\infty$. Recall that $0 \leq B_k(\cdot) \leq 1$ on $\mathcal{W}$, so

$$|B_k(W_i)B_l(W_i) - \mathrm{E}[B_k(W)B_l(W)]| \leq 2,$$

for all k and l By (37),

$$\mathrm{Var}\left(B_k(W_i)B_l(W_i) - \mathrm{E}[B_k(W)B_l(W)]\right) \leq \mathrm{E}[B_k^2(W)B_l^2(W)] \leq C_2L_n^{-1}.$$

By Lemma 4, we have

$$\begin{aligned} &P\left(|\frac{1}{n}\sum_{i=1}^n B_k(W_i)B_l(W_i) - \mathrm{E}[B_k(W)B_l(W)]| \geq \delta/6n\right) \\ \leq\ & 2\exp\{-\delta^2/(72C_2L_n^{-1}n + 24\delta)\}. \end{aligned}$$

It then follows from the union bound of probability that

$$P\left(\|\mathbf{D}_1\|_\infty \geq \delta/6n\right) \leq 2L_n^2\exp\{-\delta^2/(72C_2L_n^{-1}n + 24\delta)\}. \quad (48)$$

We next bound $\|\mathbf{D}_{2j}\|_\infty$. Note that for $k, l = 1, \cdots, L_n$,

$$\begin{aligned} &\mathrm{E}[|X_{ji}B_k(W_i)B_l(W_i) - \mathrm{E}[X_jB_k(W)B_l(W)]|^m] \\ \leq\ & 2^m\mathrm{E}[|X_{ji}B_k(W_i)B_l(W_i)|^m] \\ \leq\ & 2^m\mathrm{E}[|X_{ji}B_k(W_i)|^m] \\ =\ & 2^m\mathrm{E}[\mathrm{E}[|X_{ji}|^m|W_i]B_k^m(W_i)] \\ \leq\ & m!(2K_1)^{m-2}(8emK_1^2C_2L_n^{-1})/2, \end{aligned}$$

where Lemma 1 was used in the last inequality. By Lemma 3, we have

$$\begin{aligned} &P\left(|\frac{1}{n}\sum_{i=1}^n X_{ji}B_k(W_i)B_l(W_i) - \mathrm{E}[X_jB_k(W)B_l(W)]| \geq \delta/6n\right) \\ \leq\ & 2\exp\{-\delta^2/(576emK_1^2C_2L_n^{-1}n + 24K_1\delta)\}. \end{aligned}$$

It then follows from the union bound of probability that

$$P\left(\|\mathbf{D}_{2j}\|_\infty \geq \delta/6n\right) \leq 2L_n^2\exp\{-\delta^2/(576emK_1^2C_2L_n^{-1}n + 24K_1\delta)\}. \quad (49)$$

Similarly we can bound $\|\mathbf{D}_{3j}\|_\infty$. For every $m \geq 2$, for $k, l = 1, \cdots, L_n$, there exists a constant $K_6 > 0$ such that

$$\begin{aligned} & \mathrm{E}[|X_{ji}^2 B_k(W_i)B_l(W_i) - \mathrm{E}[X_j^2 B_k(W)B_l(W)]|^m] \\ \leq\ & 2^m \mathrm{E}[\mathrm{E}[|X_{ji}^2|^m|W_i]B_k^m(W_i)] \\ \leq\ & m!(2K_6)^{m-2}(8emK_6^2C_2L_n^{-1})/2. \end{aligned}$$

By Lemma 3, we have

$$\begin{aligned} & P\left(|X_{ji}^2 B_k(W_i)B_l(W_i) - \mathrm{E}[X_j^2 B_k(W)B_l(W)]| \geq \delta/6n\right) \\ \leq\ & 2\exp\{-\delta^2/(576emK_6^2C_2L_n^{-1}n + 24K_6\delta)\}. \end{aligned}$$

It then follows from the union bound of probability that

$$P(\|\mathbf{D}_{3j}\|_\infty \geq \delta/6n) \leq 2L_n^2 \exp\{-\delta^2/(576emK_6^2C_2L_n^{-1}n + 24K_6\delta)\}. \tag{50}$$

Let $b_3 = 72C_2\max\{1, 8emK_1^2, 8emK_6^2\}$ and $b_4 = 24\max\{1, K_1, K_6\}$, then combining (47)-(50) we have

$$P\left(\left\|\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj} - \mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]\right\| \geq L_n\delta/n\right) \leq 6L_n^2\exp\left\{-\frac{\delta^2}{b_3L_n^{-1}n + b_4\delta}\right\}. \tag{51}$$

Observe that $\|\frac{1}{n}\mathbf{B}_n^T\mathbf{B}_n - \mathrm{E}[\mathbf{B}^T\mathbf{B}]\| \leq 2L_n\|\mathbf{D}_1\|_\infty$. Thus, we have also proved that

$$P\left(\left\|\frac{1}{n}\mathbf{B}_n^T\mathbf{B}_n - \mathrm{E}[\mathbf{B}^T\mathbf{B}]\right\| \geq L_n\delta/n\right) \leq 6L_n^2\exp\left\{-\frac{\delta^2}{b_3L_n^{-1}n + b_4\delta}\right\}. \tag{52}$$

We next prove the second part of the lemma. Note that for any symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ (Fan, Feng and Song, 2011),

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \leq \max\{|\lambda_{\min}(\mathbf{A} - \mathbf{B})|, |\lambda_{\min}(\mathbf{B} - \mathbf{A})|\}. \tag{53}$$

It then follows from (53) that

$$\left|\lambda_{\min}(\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj}) - \lambda_{\min}(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])\right| \quad \leq \quad 2L_n\left\|\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj} - \mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]\right\|_\infty,$$

which implies that

$$\begin{aligned} & P\left(\left|\lambda_{\min}(\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj}) - \lambda_{\min}(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])\right| \geq L_n\delta/n\right) \\ \leq\ & 6L_n^2\exp\{-\delta^2/(b_3L_n^{-1}n + b_4\delta)\}. \end{aligned} \tag{54}$$

Let $\delta = b_9C_3L_n^{-2}n$ in (54) for $b_9 \in (0, 1)$. According to (46), we have

$$\begin{aligned} & P\left(\left|\lambda_{\min}(\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj}) - \lambda_{\min}(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])\right| \geq b_9\lambda_{\min}(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])\right) \\ \leq\ & 6L_n^2\exp(-b_6L_n^{-3}n), \end{aligned} \tag{55}$$

for some positive constant b_6. Next observe the fact that for $x, y > 0, a \in (0, 1)$ and $b = 1/(1-a) - 1$,

$$|x^{-1} - y^{-1}| \geq by^{-1} \text{ implies } |x - y| \geq ay.$$

This is because $x^{-1} - y^{-1} \geq by^{-1}$ is equivalent to $x^{-1} \geq \frac{1}{1-a}y^{-1}$, or $x - y \leq -ay$; on the other hand, $x^{-1} - y^{-1} \leq by^{-1}$ implies $x^{-1} \leq (1 - \frac{a}{1-a})y^{-1} \leq (1 - \frac{a}{1+a})y^{-1}$ as $a \in (0, 1)$, and therefore $x - y \geq ay$. Then let $b_5 = 1/(1 - b_9) - 1$, it follows from (55) that

$$\begin{aligned} & P\left(\left|(\lambda_{\min}(\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj}))^{-1} - (\lambda_{\min}(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]))^{-1}\right| \geq b_5(\lambda_{\min}(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]))^{-1}\right) \\ \leq\ & 6L_n^2\exp(-b_6L_n^{-3}n). \end{aligned} \tag{56}$$

Following the same proof, by (38) we also have for any positive constant b_7, there exists some positive constant b_8, such that

$$\begin{aligned} & P\left(\left|(\lambda_{\min}(\frac{1}{n}\mathbf{B}_n^T\mathbf{B}_n))^{-1} - (\lambda_{\min}(\mathrm{E}[\mathbf{B}^T\mathbf{B}]))^{-1}\right| \geq b_7(\lambda_{\min}(\mathrm{E}[\mathbf{B}^T\mathbf{B}]))^{-1}\right) \\ \leq \;& 6L_n^2\exp(-b_8L_n^{-3}n). \end{aligned} \quad (57)$$

The second part of the lemma then follows from the fact that for any symmetric matrix $\mathbf{A}$, $\lambda_{\min}(\mathbf{A})^{-1} = \lambda_{\max}(\mathbf{A}^{-1})$. □

A.3. Proof of Main Results

Proof of Proposition 1. Note that $\mathrm{E}[Y|W,X_j] = a_j(W) + b_j(W)X_j$. By Stone (1982), there exist $\{a_j^*\}_{j=0}^p$ and $\{b_j^*\}_{j=1}^p \in \mathcal{S}_n$ such that $\|a_j - a_j^*\|_\infty \leq M_2L_n^{-d}$ and $\|b_j - b_j^*\|_\infty \leq M_2L_n^{-d}$, where $\mathcal{S}_n$ is the space of polynomial splines of degree $l \geq 1$ with normalized B-spline basis $\{B_k, k = 1,\cdots,L_n\}$, and M_2 is some positive constant. Here $\|\cdot\|_\infty$ denotes the sup norm. Let $\boldsymbol{\eta}_j^*$ and $\boldsymbol{\theta}_j^*$ be L_n-dimensional vectors such that for $a_j^*(W) = \mathbf{B}(W)\boldsymbol{\eta}_j^*$ and $b_j^*(W) = \mathbf{B}_j(W)\boldsymbol{\theta}_j^*$.
Recall that $\tilde{a}_j(W) = \mathbf{B}(W)\tilde{\boldsymbol{\eta}}_j$ and $\tilde{b}_j(W) = \mathbf{B}(W)\tilde{\boldsymbol{\theta}}_j$. By definition of $\tilde{\boldsymbol{\eta}}_j$ and $\tilde{\boldsymbol{\theta}}_j$, we have

$$\begin{aligned} (\tilde{a}_j,\tilde{b}_j) &= \arg\min_{a_j,b_j\in\mathcal{S}_n} \mathrm{E}[(Y - a_j(W) - b_j(W)X_j)^2] \\ &= \arg\min_{a_j,b_j\in\mathcal{S}_n} \mathrm{E}[(\mathrm{E}[Y|W,X_j] - a_j(W) - b_j(W)X_j)^2], \end{aligned}$$

and therefore $\|\mathrm{E}[Y|W,X_j] - \tilde{a}_j - \tilde{b}_jX_j\|^2 \leq \|\mathrm{E}[Y|W,X_j] - a_j^* - b_j^*X_j\|^2$. In other words,

$$\begin{aligned} \|\tilde{a}_j + \tilde{b}_jX_j - (a_j + b_jX_j)\|^2 &\leq \|(a_j^* + b_j^*X_j) - (a_j + b_jX_j)\|^2 \\ &\leq 2\|a_j - a_j^*\|^2 + 2\|(b_j - b_j^*)X_j\|^2 \\ &\leq 2M_2^2L_n^{-2d}(1 + \mathrm{E}[X_j^2]). \end{aligned}$$

On the other hand, by the least-squares property,

$$\mathrm{E}[(Y - \tilde{a}_j - \tilde{b}_jX_j)(\tilde{a}_j + \tilde{b}_jX_j)] = 0,$$

and by conditioning in W_j and X_j, we have

$$\mathrm{E}[(Y - a_j - b_jX_j)(\tilde{a}_j + \tilde{b}_jX_j)] = 0.$$

The last two equalities imply that

$$\mathrm{E}[(a_j + b_jX_j - \tilde{a}_j - \tilde{b}_jX_j)(\tilde{a}_j + \tilde{b}_jX_j)] = 0$$

Thus, by the Pythagorean theorem, we have

$$\|a_j + b_jX_j\|^2 = \|\tilde{a}_j + \tilde{b}_jX_j\|^2 + \|\tilde{a}_j + \tilde{b}_jX_j - a_j - b_jX_j\|^2,$$

and

$$\|a_j + b_jX_j\|^2 - \|\tilde{a}_j + \tilde{b}_jX_j\|^2 \leq 2M_2^2L_n^{-2d}(1 + \mathrm{E}[X_j^2]). \quad (58)$$

Similary, we have

$$\|a_0\|^2 - \|\tilde{a}_0\|^2 \leq M_2^2L_n^{-2d}. \quad (59)$$

By taking $M_1 = M_2^2(8eK^2+3)$ (c.f. Lemma 1), the first part of Proposition 1 follows from (58) and (59):

$$\begin{aligned} u_j - \tilde{u}_j &= \|a_j + b_jX_j\|^2 - \|a_0\|^2 - (\|\tilde{a}_j + \tilde{b}_jX_j\|^2 - \|\tilde{a}_0\|^2) \\ &\leq M_1L_n^{-2d}. \end{aligned} \quad (60)$$

By (25) and (60), we have

$$\min_{j\in\mathcal{M}_*} \tilde{u}_j \geq c_1 L_n n^{-2\kappa}/h_2 - M_1 L_n^{-2d}.$$

Then the desired result follows from $L_n^{-2d-1} \leq c_1(1/h_2 - \xi)n^{-2\kappa}/M_1$ for some $\xi \in (0, 1/h_2)$. □

Proof of Theorem 1. We first prove part (1). Note that

$$\widehat{u}_{nj} - \tilde{u}_j \quad = \quad S_1 + S_2,$$

where

$$S_1 = \frac{1}{n}\mathbf{Y}^T\mathbf{Q}_{nj}(\mathbf{Q}_{nj}^T\mathbf{Q}_{nj})^{-1}\mathbf{Q}_{nj}^T\mathbf{Y} - \mathrm{E}[Y\mathbf{Q}_j](\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])^{-1}\mathrm{E}[\mathbf{Q}_j^T Y],$$

and

$$S_2 = \frac{1}{n}\mathbf{Y}^T\mathbf{B}_n(\mathbf{B}_n^T\mathbf{B}_n)^{-1}\mathbf{B}_n^T\mathbf{Y} - \mathrm{E}[Y\mathbf{B}](\mathrm{E}[\mathbf{B}^T\mathbf{B}])^{-1}\mathrm{E}[\mathbf{B}^T Y].$$

We first focus on S_1. Let $\mathbf{a}_n = \frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Y}$, $\mathbf{a} = \mathrm{E}[\mathbf{Q}_j^T Y]$, $\mathbf{U}_n = (\frac{1}{n}\mathbf{Q}_{nj}^T\mathbf{Q}_{nj})^{-1}$ and $\mathbf{U} = (\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])^{-1}$. Then

$$\begin{aligned} S_1 &= \mathbf{a}_n^T\mathbf{U}_n\mathbf{a}_n - \mathbf{a}^T\mathbf{U}\mathbf{a} \\ &= (\mathbf{a}_n - \mathbf{a})^T\mathbf{U}_n(\mathbf{a}_n - \mathbf{a}) + 2(\mathbf{a}_n - \mathbf{a})^T\mathbf{U}_n\mathbf{a} + \mathbf{a}^T(\mathbf{U}_n - \mathbf{U})\mathbf{a}. \end{aligned}$$

Denote the last three terms respectively by S_{11}, S_{12}, and S_{13}.

We first deal with S_{11}. Note that

$$|S_{11}| \leq \|\mathbf{U}_n\| \cdot \|\mathbf{a}_n - \mathbf{a}\|_2^2. \tag{61}$$

By Lemma 5 and the union bound of probability,

$$P(\|\mathbf{a}_n - \mathbf{a}\|_2^2 \geq 2L_n\delta^2/n^2) \quad \leq \quad 8L_n \exp\{-\delta^2/(b_1 L_n^{-1} n + b_2\delta)\}. \tag{62}$$

According to the second part of Lemma 7, for any given positive constant b_5, there exists a positive constant b_6 such that

$$P\left(|\|\mathbf{U}_n\| - \|\mathbf{U}\|| \geq b_5\|\mathbf{U}\|\right) \leq 6L_n^2 \exp\{-b_6 L_n^{-3} n\}.$$

Then it follows from (46) that

$$P\left(\|\mathbf{U}_n\| \geq (b_5 + 1)C_3^{-1}L_n\right) \leq 6L_n^2 \exp\{-b_6 L_n^{-3} n\}. \tag{63}$$

Combining (61)-(63) and based on the union bound of probability, we have

$$\begin{aligned} & P(|S_{11}| \geq 2(b_5 + 1)C_3^{-1}L_n^2\delta^2/n^2) \\ \leq \quad & 8L_n \exp\{-\delta^2/(b_1 L_n^{-1} n + b_2\delta)\} + 6L_n^2 \exp\{-b_6 L_n^{-3} n\}. \end{aligned} \tag{64}$$

We next bound S_{12}. Note that

$$|S_{12}| \leq 2\|\mathbf{a}_n - \mathbf{a}\|_2 \cdot \|\mathbf{U}_n\| \cdot \|\mathbf{a}\|_2 \tag{65}$$

By Lemma 1,

$$\begin{aligned} \|\mathbf{a}\|_2^2 &= \|\mathrm{E}[\mathbf{B}^T Y]\|_2^2 + \|\mathrm{E}[X_j\mathbf{B}^T Y]\|_2^2 \\ &= \sum_{k=1}^{L_n}(\mathrm{E}[B_k m(\mathbf{X}^*)])^2 + \sum_{k=1}^{L_n}(\mathrm{E}[X_j B_k m(\mathbf{X}^*)])^2 \\ &\leq \sum_{k=1}^{L_n}(\mathrm{E}[B_k^2 m^2(\mathbf{X}^*)] + \mathrm{E}[B_k^2 X_j^2 m^2(\mathbf{X}^*)]) \\ &\leq 4eC_2(K_2^2 + K_4^2), \end{aligned} \tag{66}$$

where the calculation as in (41) was used.

It follows from (62), (63), (65), (66) and the union bound of probability that

$$\begin{aligned} & P(|S_{12}| \geq 4\sqrt{2}(b_5+1)e^{1/2}C_2^{1/2}(K_2^2+K_4^2)^{1/2}C_3^{-1}L_n^{3/2}\delta/n) \\ \leq\ & 8L_n\exp\{-\delta^2/(b_1L_n^{-1}n+b_2\delta)\}+6L_n^2\exp\{-b_6L_n^{-3}n\}. \end{aligned} \quad (67)$$

To bound S_{13}, note that

$$|S_{13}| = \mathbf{a}^T\mathbf{U}_n(\mathbf{U}^{-1}-\mathbf{U}_n^{-1})\mathbf{U}\mathbf{a} \leq \|\mathbf{U}_n\|^2\cdot\|\mathbf{U}^{-1}-\mathbf{U}_n^{-1}\|\cdot\|\mathbf{a}\|_2^2. \quad (68)$$

Then it follows from Lemmas 6, Lemma 7, (63), (66), (68) and the union bound of probability that there exist b_3, b_4 and b_6 such that

$$\begin{aligned} & P(|S_{13}| \geq 4eC_2(K_2^2+K_4^2)(b_5+1)^2C_3^{-2}L_n^3\delta/n) \\ \leq\ & 6L_n^2\exp\{-\delta^2/(b_3L_n^{-1}n+b_4\delta)\}+6L_n^2\exp\{-b_6L_n^{-3}n\}. \end{aligned} \quad (69)$$

Hence, combining (64), (67) and (69), there exist some positive constants s_1, s_2 and s_3 such that

$$\begin{aligned} & P\left(|S_1| \geq s_1L_n^2\delta^2/n^2+s_2L_n^{3/2}\delta/n+s_3L_n^3\delta/n\right) \\ \leq\ & 16L_n\exp\{-\delta^2/(b_1L_n^{-1}n+b_2\delta)\}+6L_n^2\exp\{-\delta^2/(b_3L_n^{-1}n+b_4\delta)\} \\ & +18L_n^2\exp\{-b_6L_n^{-3}n\}. \end{aligned} \quad (70)$$

Similarly, we can prove that there exist positive constants s_4, s_5 and s_6 such that

$$\begin{aligned} & P\left(|S_2| \geq s_4L_n^2\delta^2/n^2+s_5L_n^{3/2}\delta/n+s_6L_n^3\delta/n\right) \\ \leq\ & 8L_n\exp\{-\delta^2/(b_1L_n^{-1}n+b_2\delta)\}+6L_n^2\exp\{-\delta^2/(b_3L_n^{-1}n+b_4\delta)\} \\ & +18L_n^2\exp\{-b_8L_n^{-3}n\}. \end{aligned} \quad (71)$$

Let $(s_1+s_4)L_n^2\delta^2/n^2+(s_2+s_5)L_n^{3/2}\delta/n+(s_3+s_6)L_n^3\delta/n = c_2L_nn^{-2\kappa}$ for any given $c_2>0$ (e.g., take $\delta = c_2L_n^{-2}n^{1-2\kappa}/(s_3+s_6)$). There exist some positive constants c_3 and c_4 such that

$$\begin{aligned} & P\left(|\widehat{u}_{nj}-\tilde{u}_j| \geq c_2L_nn^{-2\kappa}\right) \\ \leq\ & (24L_n+12L_n^2)\exp\{-c_3n^{1-4\kappa}L_n^{-3}\}+36L_n^2\exp\{-c_4L_n^{-3}n\}. \end{aligned} \quad (72)$$

Then Theorem 1(i) follows from the union bound of probability.

We now prove part (ii). Note that on the event

$$\mathcal{A}_n \equiv \left\{\max_{j\in\mathcal{M}_*}|\widehat{u}_{nj}-\tilde{u}_j| \leq c_1\xi L_nn^{-2\kappa}/2\right\},$$

by Proposition 1, we have

$$\widehat{u}_{nj} \geq c_1\xi L_nn^{-2\kappa}/2, \quad \text{for all } j\in\mathcal{M}_*. \quad (73)$$

Hence, by choosing $\tau_n = c_1\xi L_nn^{-2\kappa}/2$, we have $\mathcal{M}_* \subset \widehat{\mathcal{M}}_{\tau_n}$. On the other hand, by the union bound of probability, there exist positive constants c_6 and c_7, such that

$$P(\mathcal{A}_n^c) \leq s_n\left\{(24L_n+12L_n^2)\exp(-c_6n^{1-4\kappa}L_n^{-3})+36L_n^2\exp(-c_7L_n^{-3}n)\right\},$$

and Theorem 1(2) follows. □

Proof of Theorem 2. Let

$$\tilde{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \mathrm{E}[(Y-\mathbf{Q}\boldsymbol{\alpha})^2],$$

where $\mathbf{Q} = (\mathbf{Q}_1,\cdots,\mathbf{Q}_p)$ is a $2pL_n$-dimensional vector of functions. Then we have

$$\mathrm{E}[\mathbf{Q}^T(Y-\mathbf{Q}\tilde{\boldsymbol{\alpha}})] = \mathbf{0}_{2pL_n},$$

where $\mathbf{0}_{2pL_n}$ is a $2pL_n$-dimension vector with all entries 0. This implies

$$\|\mathrm{E}[\mathbf{Q}^TY]\|_2^2 = \tilde{\boldsymbol{\alpha}}^T\boldsymbol{\Sigma}^2\tilde{\boldsymbol{\alpha}} \leq \lambda_{\max}(\boldsymbol{\Sigma})\tilde{\boldsymbol{\alpha}}^T\boldsymbol{\Sigma}\tilde{\boldsymbol{\alpha}},$$

recalling $\mathbf{\Sigma} = \mathrm{E}[\mathbf{Q}^T\mathbf{Q}]$. It follows from orthogonal decomposition that $\mathrm{Var}(\mathbf{Q}\tilde{\boldsymbol{\alpha}}) \le \mathrm{Var}(Y)$ and $\mathrm{E}[\mathbf{Q}\tilde{\boldsymbol{\alpha}}] = \mathrm{E}[Y]$ (recall the inclusion of the intercept term). Therefore,

$$\tilde{\boldsymbol{\alpha}}^T\mathbf{\Sigma}\tilde{\boldsymbol{\alpha}} \le \mathrm{E}[Y^2] = O(1),$$

and

$$\|\mathrm{E}[\mathbf{Q}^T Y]\|_2^2 = O(\lambda_{\max}(\mathbf{\Sigma})). \tag{74}$$

Note that by the definition of $\tilde{u}_j$,

$$\begin{aligned}\sum_{j=1}^p \tilde{u}_j &= \sum_{j=1}^p \mathrm{E}[Y\mathbf{Q}_j]\left((\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])^{-1} - \begin{bmatrix} (\mathrm{E}[\mathbf{B}^T\mathbf{B}])^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right)\mathrm{E}[\mathbf{Q}_j^T Y] \\ &\le \max_{1\le j\le p}\lambda_{\max}\{(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])^{-1}\}\sum_{j=1}^p \|\mathrm{E}[\mathbf{Q}_j^T Y]\|_2^2 \\ &= \max_{1\le j\le p}\lambda_{\max}\{(\mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j])^{-1}\}\|\mathrm{E}[\mathbf{Q}^T Y]\|_2^2.\end{aligned}$$

By Lemma 6 and (74), the last term is of order $O(L_n\lambda_{\max}(\mathbf{\Sigma}))$. This implies that the number of $\{j : \tilde{u}_j > \delta L_n n^{-2\kappa}\}$ cannot exceed $O(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma}))$ for any $\delta > 0$.

On the set

$$\mathcal{B}_n = \left\{\max_{1\le j\le p}|\widehat{u}_{nj} - \tilde{u}_j| \le \delta L_n n^{-2\kappa}\right\},$$

the number of $\{j : \widehat{u}_{nj} > 2\delta L_n n^{-2\kappa}\}$ cannot exceed the number of $\{j : \tilde{u}_j > \delta L_n n^{-2\kappa}\}$, which is bounded by $O(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma}))$. By taking $\delta = c_5/2$, we have

$$P\left\{|\widehat{\mathcal{M}}_{\tau_n}| \le O(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma}))\right\} \ge P(\mathcal{B}_n).$$

Then the desired result follows from Theorem 1(i). □

References

Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations (with discussion). *Jour. Ameri. Statist. Assoc.*, **96**, 939-967.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion), *Ann. Statist.*, **35**, 2313-2404.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models, *J. Am. Statist. Assoc.*, **106**, 544-557.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Statist. Assoc.*, **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion), *J.Roy. Statisti. Soc. B.*, **70**, 849-911.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality, *Ann. Statist.*, **38**, 3567-3604.

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and wilks phenomenon, *Ann. Statist.*, **29**, 153-193.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Stat. Interface.*, **1**, 179-195.

Frank, I.E. and Friedman, J.H. (1993). A statistical view of some chemometrics regression tools (with discussion), *Technometrics*, **35**, 109-148.

Hall, P. and Miller, H. (2009). Using Generalised Correlation to Effect Variable Selection in Very High Dimensional Problems, *J. Comput. Graph. Stat.*, **18**, 533-550.

Hall, P., Titterington, D. M. and Xue, J. H. (2009). Tilting methods for assessing the influence of components in a classifier, *J.Roy. Statisti. Soc. B.*, **71**, 783-803.

Harrison, D. and Rubinfeld, D. (1978). Hedonic housing prices and the demand for clean air, *J. Environ. Econ. Manag.* **5**, 81-102.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, London: Chapman & Hall.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *J.Roy. Statisti. Soc. B.*, **55**, 757-796.

Li, G, Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *Ann. Statist.*, **40**, 1846-1877.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Am. Statist. Assoc.*, to appear.

Lian, H. (2011). Flexible shrinkage estimation in high-dimensional varying coefficient models, *manuscript*.

Stone, C.J. (1982). Optimal glaobal rates of convergence for nonparametric regression, *Ann. Statist.*, **10**, 1040-1053.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *J.Roy. Statisti. Soc. B.*, **58**, 267-288.

van der Vaart, A.W. and Wellner, J.A. (1996). Weak Convergence and Empirical Processes. Springer, New York.

Vershynin R. (2011). Introduction to the non-asymptotic analysis of random matrices, *manuscript*.

Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Am. Statist. Assoc.*, **103**, 1556-1569.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Jour. Roy. Statist. Soc. B*, **68**, 49-67.

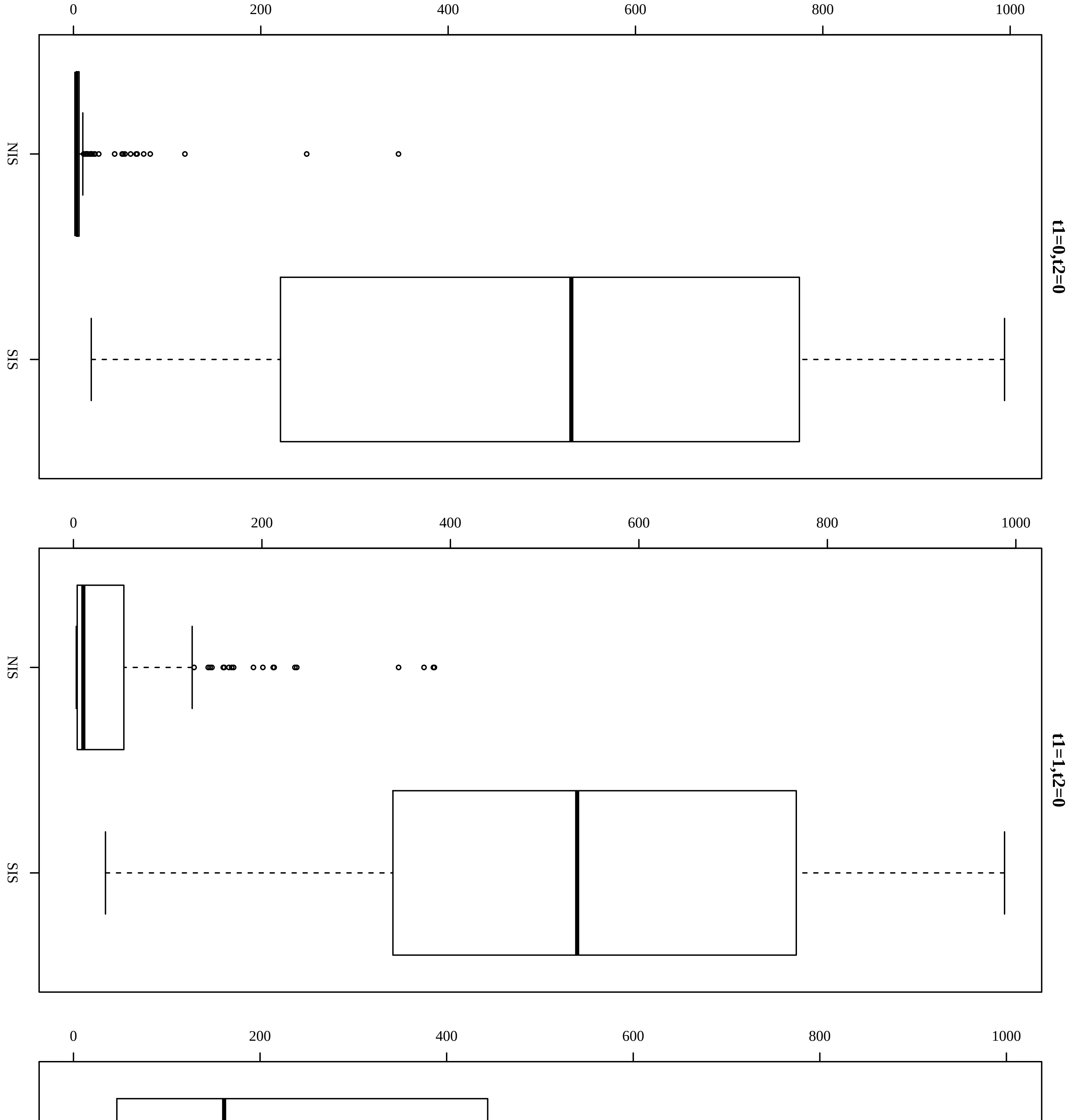
Zhang, C.-H. (2010), Nearly unbiased variable selection under minimax concave penalty, *Ann. Statist.*, **38**, 894-942.

Zhao, D. S., and Li, Y. (2010). Principled Sure Independence Screening for Cox Models With Ultra-High-Dimensional Covariates. *manuscript*, Harvard University.

Zou, H. (2006) The adaptive Lasso and its oracle properties, *J. Am. Statist. Assoc.*, **101**, 1418-1429.

Zou, H. and Hastie, T. (2005). Addendum: Regularization and variable selection via the Elastic net, *J.Roy. Statisti. Soc. B.*, **67**, 301-320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models, *Ann. Statist.*, **36**, 1509-1533.


0
200
400
600
800
1000
SIN
SIS
t1=0,t2=0
t1=1,t2=0