



Published in final edited form as:

J Am Stat Assoc. 2014 ; 109(507): 1270–1284. doi:10.1080/01621459.2013.879828.

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

Jianqing Fan¹, Yunbei Ma², and Wei Dai¹

¹Princeton University, Princeton, New Jersey 08544 USA

²Southwestern University of Finance and Economics, Chengdu, Sichuan 611130 China

Abstract

The varying-coefficient model is an important class of nonparametric statistical model that allows us to examine how the effects of covariates vary with exposure variables. When the number of covariates is large, the issue of variable selection arises. In this paper, we propose and investigate marginal nonparametric screening methods to screen variables in sparse ultra-high dimensional 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 the practical utility and 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

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

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 have an ultra-high dimensionality p that can 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 when dealing with such high-dimensional problems.

⁰Address for correspondence: Yunbei Ma, School of Statistics, Southwest University of Finance and Economics, Chengdu, China. myb@swufe.edu.cn.

With the sparsity assumption, variable selection helps improve the accuracy of estimation and gain scientific insights. Many 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 settings.

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) to ultra-high dimensional additive models, resulting in nonparametric independence screening (NIS). On a different front, Fan and Song (2010) 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, Titterton and Xue, 2009), marginal partial likelihood method MPLE (Zhao and Li, 2010), 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 modeling biases, but suffer from the so-called “curse of dimensionality”. A remarkably simple and powerful nonparametric model for dimensionality reduction is the varying-coefficient model,

$$Y = \beta^T(W)\mathbf{X} + \varepsilon, \quad (1)$$

where Y is the response, W is some univariate observable exposure variable, $\mathbf{X} = (X_1, \dots, X_p)^T$ is the vector of covariates 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. The covariates \mathbf{X} enter the model linearly, and the regression coefficient functions $\beta(\cdot)$ vary smoothly with the exposure variable W . The model retains general nonparametric characteristics and allows nonlinear interactions between the exposure variable and the covariates. It arises frequently in economics, finance, 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 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 challenges of designing robust algorithm with reasonable computational cost while achieving statistical precision. While existing theory and estimation methods are not directly applicable in ultra-high dimensional regime, we aim to address such need, both theoretically and practically.

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

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

Let $a_j(W)$ and $b_j(W)$ be the solution to (2) and $\hat{a}_{nj}(W)$ and $\hat{b}_{nj}(W)$ be their nonparametric estimates as defined later in (6). 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. Along with previously established understanding towards (generalized) linear models and additive models, this work is another useful building block of the universality of the sure screening framework.

The sure screening property and false selection rate of NIS can be established under certain technical conditions. As to be shown later, our assumptions are much weaker than that in previous work. In some special cases, NIS can even be model selection consistent. In establishing these results, three factors play important roles: the approximation error in modeling nonparametric components, the stochastic error in estimating the nonparametric components, and the tail distributions of the variables. We also propose two nonparametric independence screening methods in an iterative framework, following Fan and Lv (2008) and Fan, Feng and Song (2011). One is the Conditional-INIS, in which we propose the novel conditional random permutation to determine a data driven screening threshold. It is worth mentioning that this conditional permutation idea is not limited to varying coefficient models, and is applicable to other settings. The other is called Greedy-INIS, which adopts a greedy approach in the variable screening step. They both serve to effectively control the false positive 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 Conditional-INIS and Greedy-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 $\beta(\cdot) = (\beta_1(\cdot), \dots, \beta_p(\cdot))^T$ is sparse. Let $\mathcal{M}_* = \{j: E[\beta_j^2(W)] > 0\}$ 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 $j = 1, \dots, p$, let $a_j(W)$ and $b_j(W)$ be the minimizer of the following marginal regression problem:

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

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

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

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

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

where $\|f\|^2 = E[f^2]$. It can be seen that

$$u_j = E[b_j^2(W)(X_j - E[X_j | W])^2] = E\left[\frac{(\text{Cov}[X_j, Y | W])^2}{\text{Var}[X_j | W]}\right]. \quad (4)$$

For each $j = 1, \dots, p$, if $\text{Var}[X_j | W] = 1$, 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 $\text{Cov}[X_j, Y | W] = 0$ almost surely.

2.2 Marginal Regression Estimation with B-spline

To obtain an estimate of the marginal utility u_j , $j = 1, \dots, p$, 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 $\{B_k, k = 1, \dots, L_n\}$ denote its normalized B-spline basis, where L_n is the number of basis functions. Note that $\|B_k\|_\infty \leq 1$, where $\|\cdot\|_\infty$ is the sup norm. Then

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

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:

$$\min_{\eta_j, \theta_j \in \mathbb{R}^{L_n}} \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{B}(W_i)\eta_j - \mathbf{B}(W_i)\theta_j X_{ji})^2, \quad (5)$$

where $\eta_j = (\eta_{j1}, \dots, \eta_{jL_n})^T$, $\theta_j = (\theta_{j1}, \dots, \theta_{jL_n})^T$ and $\mathbf{B}(\cdot) = (B_1(\cdot), \dots, B_{L_n}(\cdot))$.

It is easy to show that the minimizers of (5) is given by

$$(\hat{\eta}_j^T, \hat{\theta}_j^T)^T = (\mathbf{Q}_{nj}^T \mathbf{Q}_{nj})^{-1} \mathbf{Q}_{nj}^T \mathbf{Y}, \text{ where } \mathbf{Q}_{nj} = (\mathbf{B}_n, \Phi_{nj}) = \begin{pmatrix} \mathbf{B}(W_1), & X_{j1}\mathbf{B}(W_1) \\ \vdots & \vdots \\ \mathbf{B}(W_n), & X_{jn}\mathbf{B}(W_n) \end{pmatrix}$$

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

$$\hat{a}_{nj}(W) = \mathbf{B}(W)\hat{\eta}_j = (\mathbf{B}(W), \mathbf{0}_{L_n}^T)(\mathbf{Q}_{nj}^T \mathbf{Q}_{nj})^{-1} \mathbf{Q}_{nj}^T \mathbf{Y}, \quad \hat{b}_{nj}(W) = \mathbf{B}(W)\hat{\theta}_j = (\mathbf{0}_{L_n}^T, \mathbf{B}(W))(\mathbf{Q}_{nj}^T \mathbf{Q}_{nj})^{-1} \mathbf{Q}_{nj}^T \mathbf{Y}, \quad (6)$$

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

$$\hat{a}_{n0}(W) = \mathbf{B}(W)\hat{\eta}_0 = \mathbf{B}(W)(\mathbf{B}_n^T \mathbf{B}_n)^{-1} \mathbf{B}_n^T \mathbf{Y}, \quad (7)$$

where

$$\hat{\eta}_0 = \arg \min_{\eta_0 \in \mathbb{R}^{L_n}} \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{B}(W_i)\eta_0)^2.$$

We now define an estimate of the marginal utility u_j as

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

where $\mathbf{W} = (W_1, \dots, W_n)^T$. 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:

$$\mathcal{M}_{\tau_n} = \{1 \leq j \leq p: \hat{u}_{nj} \geq \tau_n\}.$$

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

$$\hat{v}_{nj} = \|\mathbf{Y} - \hat{a}_{nj}(\mathbf{W}) - \hat{b}_{nj}(\mathbf{W})\mathbf{X}_j\|_n^2.$$

and we select variables as follows,

$$\mathcal{M}_{\nu_n} = \{1 \leq j \leq p: \hat{v}_{nj} \leq \nu_n\},$$

where ν_n is a predefined threshold value.

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

$$\|\hat{a}_{nj}(\mathbf{W}) + \hat{b}_{nj}(\mathbf{W})\mathbf{X}_j\|_n^2 = \frac{1}{n} \mathbf{Y}^T \mathbf{Q}_{nj} (\mathbf{Q}_{nj}^T \mathbf{Q}_{nj})^{-1} \mathbf{Q}_{nj}^T \mathbf{Y}, \quad (8)$$

and

$$\frac{1}{n} \sum_{i=1}^n Y_i (\hat{a}_{nj}(W_i) + \hat{b}_{nj}(W_i) X_{ji}) = \frac{1}{n} \mathbf{Y}^T \mathbf{Q}_{nj} (\mathbf{Q}_{nj}^T \mathbf{Q}_{nj})^{-1} \mathbf{Q}_{nj}^T \mathbf{Y}. \quad (9)$$

It follows from (8) and (9) that

$$\hat{\mathcal{L}}_{nj} = \|\mathbf{Y}\|_n^2 - \|\hat{a}_{nj}(\mathbf{W})\|_n^2 - \hat{u}_{nj}. \quad (10)$$

Since the first two terms on the right hand side of (10) do not vary in j , selecting variables with large marginal utility \hat{u}_{nj} is the same as picking those that yield small marginal residual sum of squares \hat{v}_{nj} .

To bridge u_j and \hat{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 $\tilde{a}_j(W) = \mathbf{B}\tilde{\eta}_j$ and $b_j(W) = \mathbf{B}\tilde{\theta}_j$, where $\tilde{\eta}_j$ and $\tilde{\theta}_j$ are the minimizer of

$$\min_{\eta_j, \theta_j \in \mathbb{R}^{L_n}} E[(Y - \mathbf{B}\eta_j - \mathbf{B}\theta_j X_j)^2],$$

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

$$\min_{\eta_0 \in \mathbb{R}^{L_n}} E[(Y - \mathbf{B}\eta_0)^2].$$

It can be seen that

$$(\tilde{a}_j(W), \tilde{b}_j(W))^T = \text{diag}(\mathbf{B}, \mathbf{B})(E[\mathbf{Q}_j^T \mathbf{Q}_j])^{-1} E[\mathbf{Q}_j^T Y], \quad \tilde{a}_0(W) = \mathbf{B}(E[\mathbf{B}^T \mathbf{B}])^{-1} E[\mathbf{B}^T Y],$$

where $\mathbf{Q}_j = (\mathbf{B}, X_j \mathbf{B})$. Then we can define

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

3 Sure Screening

In this section, we establish the sure screening properties of the proposed method for model (1). Recall that by (4) the population version of marginal utility quantifies the relationship between X_j 's and Y as follows:

$$u_j = E \left[\frac{(\text{Cov}[X_j, Y|W])^2}{\text{Var}[X_j|W]} \right], j=1, \dots, p.$$

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 $j = 1, \dots, p$, $\text{Var}[X_j|W]$ 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 $0 < h_1 \leq \text{Var}[X_j|W] \leq h_2 < \infty$.
- (ii) $\min_{j \in \mathcal{M}_*} E[(\text{Cov}[X_j, Y|W])^2] \geq c_1 L_n n^{-2\kappa}$, for some $\kappa > 0$ and $c_1 > 0$.

Then under conditions (i) and (ii),

$$\min_{j \in \mathcal{M}_*} u_j \geq c_1 L_n n^{-2\kappa} / h_2. \quad (11)$$

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 (iv). 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, $0 < T_1 \leq g(W) \leq T_2 < \infty$ 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 r th derivative $f^{(r)}$ exists and is Lipschitz of order α . That is,

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

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 $j = 1, \dots, p$, there exists positive constants K_1 and $r_1 \geq 2$, such that

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

uniformly on \mathcal{W} , for any $t \geq 0$. Furthermore, let $m(\mathbf{X}^*) = E[Y|\mathbf{X}, W]$, 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

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

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

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

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

- (vii) There exists some constant $\xi \in (0, 1/h_2)$ such that

$$L_n^{-2d-1} \leq c_1(1/h_2 - \xi)n^{-2\kappa}/M_1.$$

Conditions (v) and (vi) are requirements for the tail distribution of each covariate X_j , the conditional mean function $m(\mathbf{X}^*)$ and the noise ε , in order to establish the sure screening property. Our assumptions are much weaker in comparison with previous work on high dimensional varying coefficient models by Wang, Li and Huang (2008), which assumes all the covariates to be uniformly bounded. It is also weaker than NIS in Fan, Feng and Song (2011), which assumes the conditional mean function to be bounded. Condition (vii) is to make sure that the marginal signal level of important variables is of the same rate as that of their B-spline approximations.

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

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

In addition, when condition (vii) also holds, we have

$$\min_{j \in \mathcal{M}_*} \tilde{u}_j \geq c_1 \xi L_n n^{-2\kappa}. \quad (12)$$

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

$$P \left(\max_{1 \leq j \leq p} |\hat{u}_{nj} - \tilde{u}_j| \geq c_2 L_n n^{-2\kappa} \right) \leq 12p_n L_n \{ (2+L_n) \exp(-c_3 n^{1-4\kappa} L_n^{-3}) + 3L_n \exp(-c_4 L_n^{-3} n) \}. \quad (13)$$

- (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

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

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 $L_n \geq Cn^{2\kappa/(2d+1)}$ 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 (12), 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 (13) tends to zero, we have with probability tending to 1 that

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

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

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

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*

$$P\left\{|\hat{\mathcal{M}}_{\tau_n}| \leq O(n^{2\kappa} \lambda_{\max}(\Sigma))\right\} \geq 1 - 12p_n L_n \left\{(2 + L_n) \exp(-c_8 n^{1-4\kappa} L_n^{-3}) + 3L_n \exp(-c_9 n L_n^{-3})\right\},$$

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

According to Theorem 2, the number of selected variables and thus the false selection rate are related to the correlation structure of the covariance matrix. As long as $\lambda_{\max}(\Sigma)$ is of polynomial order, the number of selected variables is also of polynomial order. In the special case where all the covariates are independent, the matrix Σ is block diagonal with j th block $E[\mathbf{Q}_j^T \mathbf{Q}_j]$, and therefore $\lambda_{\max}(\Sigma) = O(L_n^{-1})$.

4 Iterative Nonparametric Independence Screening

As Fan and Lv (2008) pointed 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 NIS. We repeatedly apply the large-scale variable screening (NIS) followed by a moderate-scale variable selection, where we use group-SCAD penalty in Wang, Li and Huang (2008) 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 threshold τ_n . Recall the random permutation used in Fan, Feng and Song (2011), which generalizes that in Zhao and Li (2010). Randomly permute \mathbf{Y} to get $\mathbf{Y}_\pi = (Y_{\pi_1}, \dots, Y_{\pi_n})^T$ and compute $\hat{u}_{n,j}^\pi$, where π is a permutation of $\{1, \dots, n\}$, 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. So 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_{jq} , the q th-ranked magnitude of $\{\hat{u}_{nj}^{\pi}, j=1, \dots, p\}$. Thus, the NIS step selects variables $\{j : \hat{u}_{nj} \geq \tau_{jq}\}$. 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 $j = 1, \dots, p$, compute

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

where the estimates are defined in (6) and (7) using $\{(\mathbf{Y}, \mathbf{W}, \mathbf{X}_j), j = 1, \dots, p\}$. Select the top K variables by ranking their marginal utilities \hat{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 $\hat{\beta}_{n0}(W)$ and their functional coefficients' estimators $\{\hat{\beta}_{nj}(W), j \in \mathcal{M}_0\}$. Conditioning on \mathcal{M}_0 , the n -dimensional partial residual is

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

For all $j \in \mathcal{M}_0^c$, compute \hat{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}_{π}^* . Compute $\hat{u}_{nj}^{*\pi}$ based on the decoupled data $\{(\mathbf{Y}_{\pi}^*, \mathbf{W}, \mathbf{X}_j), j \in \mathcal{M}_0^c\}$. Let τ_{jq}^* be the q th-ranked magnitude of $\{\hat{u}_{nj}^{*\pi}, j \in \mathcal{M}_0^c\}$. Then, the active variable set of variables is chosen as

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

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 is described in section 4.3.
- 3 Repeat step 1–2, where we replace \mathcal{M}_0 in step 1 by \mathcal{M}_l , $l = 1, 2, \dots$, 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 (such as $\lceil n/\log(n) \rceil$).

4.2 Greedy-INIS Method

Following Fan, Feng and Song (2011), we also implement a greedy version of the INIS procedure. 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 \hat{u}_{nj} . This NIS step is followed by the same group-SCAD penalized regression as in step 2. We then iterate these two steps (screening top p_0 variables and group-SCAD) 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

As the varying-coefficient functions are expanded in a spline basis, an estimated coefficient function vanishes if and only if all of its coefficients in the spline expansion are zero. Therefore, group penalty is needed (Antoniadis and Fan, 2001; Yuan and Lin, 2006).

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

$$\min_{\gamma_0, \gamma_j \in \mathbb{R}^{L_n}} \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{B}(W_i)\gamma_0 - \sum_{j \in \mathcal{A}_l} \mathbf{B}(W_i)X_{ji}\gamma_j)^2 + \sum_{j \in \mathcal{A}_l} p_\lambda(\|\gamma_j\|_B). \quad (14)$$

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

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

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). The penalization parameter λ is chosen by BIC criteria $n \log(\hat{\sigma}_\varepsilon^2) + k L_n \log n$, where $\hat{\sigma}_\varepsilon^2$ is the residual variance and k is the number of covariates chosen. By Antoniadis and Fan (2001) and Yuan and Lin (2006), the norm-penalty in (14) 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: sample size $n = 400$, the number of covariates $p = 1000$, cubic B-spline, $L_n = 7$ and the number of simulations $N = 200$ for each example. Note that L_n should not be too large since the larger L_n is, the larger estimation variance is, and the more difficult it is to distinguish important variables from unimportant ones. On the other hand, L_n should not be too small to create probing biases. Here we choose $L_n = \lfloor 2n^{1/5} \rfloor$, where $\lfloor \cdot \rfloor$ denotes rounding to the nearest integer.

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

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

where $\{\xi_k\}_{k=951}^{1000}$ are standard normal random variables. We construct the following model: $Y = \beta^T \mathbf{X} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 3)$ and $\beta = (1, -1, 1, -1, \dots)^T$ 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 in 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 $\{U_1, U_2, \dots, U_{p+2}\}$ be i.i.d. uniform random variables on $[0, 1]$, based on which we construct \mathbf{X} and W as follows:

$$X_j = \frac{U_j + t_1 U_{p+1}}{1 + t_1}, j=1, \dots, p, W = \frac{U_{p+2} + t_2 U_{p+1}}{1 + t_2},$$

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. We define coefficient functions

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

The true data generation model is

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

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

Under different correlation settings, the comparison of 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.

Example 3 Let $\{Z_1, \dots, Z_p\}$ 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:

$$X_j = \frac{Z_j + t_1 U_1}{1 + t_1}, j=1, \dots, p, W = \frac{U_2 + t_2 U_1}{1 + t_2}, Y = 2X_1 + 3W \cdot X_2 + (W+1)^2 \cdot X_3 + \frac{4\sin(2\pi W)}{2 - \sin(2\pi W)} \cdot X_4 + \varepsilon.$$

We study two settings: $t_1 = t_2 = 0$, 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 $\text{corr}(X_j, W) = 0.46$. We report the average of the number of true positives (TP), model size, the minimum true signal, the maximum false signal and the maximum null signal based on 200 simulations. Their robust standard deviations are also reported therein.

Based on the first row of 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, 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 apply the conditional permutation method, whose performance is also illustrated in Table 1 for a few choices of tuning parameter K . 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 of false positives are weakened. 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., $\mathcal{M}^* \setminus \mathcal{M}_0 = \emptyset$), 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 ones become large enough to differentiate them. Therefore 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. 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 $\text{var}(\beta^T(W)\mathbf{X})/\text{var}(\varepsilon)$, is computed.

We first explore the performance of Conditional-INIS procedure using different K 's for the simulated model specified in Example 3. Table 2 shows that under both uncorrelated and highly correlated settings, the model selection and estimation results are rather robust to the choice of K . This is not surprising since the conditional permutation mainly serves as the initialization step (step 0) in our iterative framework. We recommend using a small K as long as the conditional permutation can select a set of variables of a reasonable size to continue. We take $K = 5$ in the rest of the paper.

Table 3 reports the results for Conditional-INIS and Greedy-INIS using the simulated model specified in Example 3 under different correlation settings. We now illustrate the performance by using another example.

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:

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

The results are present in Table 4.

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; as the correlation gets stronger, we see a decrease in TP and an increase in FP. 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 needs at least and usually more than 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),

$$\begin{aligned}
\log(MV) = & \beta_0 + \beta_1 RM^2 \\
& + \beta_2 AGE + \beta_3 \log(DIS) + \beta_4 \log(RAD) \\
& + \beta_5 TAX + \beta_6 PTRATIO + \beta_7 (B - 0.63)^2 \\
& + \beta_8 \log(LSTAT) \\
& + \beta_9 CRIM + \beta_{10} ZN \\
& + \beta_{11} INDUS \\
& + \beta_{12} CHAS \\
& + \beta_{13} NOX^2 + \varepsilon,
\end{aligned}$$

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 (with $L_n = 7$ and $K = 5$) selects the following submodel:

$$\log(MV) = \beta_0(W) + \beta_1(W) \cdot RM^2 + \beta_2(W) \cdot AGE + \beta_5(W) \cdot TAX + \beta_7(W) \cdot (B - 0.63)^2 + \beta_9(W) \cdot CRIM + \varepsilon, \quad (15)$$

where $W = \log(DIS)$. The estimated functions $\hat{\beta}_j(W)$'s are presented in Figure 3. This varying coefficient model shows very interesting aspects of housing valuation. The 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 $\{Z_1, \dots, Z_p\}$ 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:

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

Note that $\{W, X_1, \dots, X_s\}$ 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 $\{W, X_1, \dots, X_s\}$ 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 5.

As seen from Table 5, 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.

Acknowledgments

Fan was supported by National Institutes of Health grants R01-GM072611 and R01-GM100474, and National Science Foundation grant DMS-1206464, and Ma was supported by National Natural Science Foundation of China (Grant No.11301424). The authors thank the editor, the associate editor, and referees for their constructive comments.

Appendix

A.1. Properties of B-splines

Our estimation use the B-spline basis, which have the following properties (de Boor 1978):

For each $j = 1, \dots, p$ and $k = 1, \dots, L_n$, $B_k(W) \geq 0$ 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 $\eta_k \in \mathbb{R}$, $k = 1, \dots, L_n$,

$$L_n^{-1} T_3 \sum_{k=1}^{L_n} \eta_k^2 \leq \int \left(\sum_{k=1}^{L_n} \eta_k B_k(w) \right)^2 dw \leq L_n^{-1} T_4 \sum_{k=1}^{L_n} \eta_k^2. \quad (16)$$

Then under condition (iii), for $C_1 = T_1 T_3$ and $C_2 = T_2 T_4$,

$$C_1 L_n^{-1} \leq E[B_k^2(W)] \leq C_2 L_n^{-1}, \text{ for } \text{all } k=1, \dots, L_n \quad (17)$$

Furthermore, under condition (iii), it follows from (16) that for any $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{L_n})^T \in \mathbb{R}^{L_n}$ such that $\|\boldsymbol{\eta}\|_2^2 = 1$, $C_1 L_n^{-1} \leq \boldsymbol{\eta}^T E[\mathbf{B}^T \mathbf{B}] \boldsymbol{\eta} \leq C_2 L_n^{-1}$. Or equivalently,

$$C_1 L_n^{-1} \leq \lambda_{\min}(E[\mathbf{B}^T \mathbf{B}]) \leq \lambda_{\max}(E[\mathbf{B}^T \mathbf{B}]) \leq C_2 L_n^{-1}. \quad (18)$$

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: $P(|X| > t|W) \leq \exp(1 - (t/K)^r)$ for all $t \geq 0$ and uniformly on the compact support of W , where $K > 0$ and $r \geq 1$. Then for all $m \geq 2$,*

$$E(|X|^m|W) \leq eK^m m!.$$

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

$$E(|X|^m|W) = \int_0^\infty P\{|X|^m \geq t|W\} dt \leq \int_0^\infty \exp(1 - (t^{1/m}/K)^r) dt = \frac{emK^m}{r} \Gamma(\frac{m}{r}) \leq eK^m \Gamma(m+1).$$

The last inequality follows from the fact $r \geq 1$, thus Lemma 1 holds.

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

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

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

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

for all $t \geq 0$ 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

$$\begin{aligned} P(|Z_1 Z_2| > t|W) &\leq P(M|Z_1| > t|W) \\ &+ P(|Z_2| > M|W) \leq \exp\{1 \\ &- (t/K_1 M)^{r_1}\} + \exp\{1 - (M/K_2)^{r_2}\} = 2\exp\{1 \\ &- (t/K_1 K_2)^r\}. \end{aligned}$$

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 $G(t) > G(K^*) \geq \log 2$ when $t > K^*$, which implies when $t > K^*$,

$$P(|Z_1Z_2| > t|W) \leq 2\exp\{1 - (t/K_1K_2)^{r_1}\} \leq \exp\{1 - (t/K^*)^{r^*}\}.$$

Also, when $t \leq K^*$, $P(|Z_1Z_2| > t|W) \leq 1 \leq \exp\{1 - (t/K^*)^{r^*}\}$. Lemma 2 holds.

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

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

for $\nu \geq \nu_1 + \dots + \nu_n$.

Lemma 4 (Bernstein's inequality, lemma 2.2.9, van der Vaart and Wellner (1996)). For independent random variables Y_1, \dots, Y_n with bounded range $[-M, M]$ and mean zero, let $\nu \geq \text{var}(Y_1 + \dots + Y_n)$, then

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

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 $j = 1, \dots, p, k = 1, \dots, L_n$,

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

and

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

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

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

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), (17), and Lemmas 1 and 2, we have for every $m \geq 2$, there exists a constant $K_4 > 0$, such that

$$E|Z_{jki}|^m \leq 2^m E|X_{ji} B_k(W_i) m(\mathbf{X}_i^*)|^m \leq 2^m E[B_{jk}^2(W_i) e K_4^m m!] \leq m!(2K_4)^{m-2} (8eK_4^2 C_2 L_n^{-1})/2. \quad (19)$$

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

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n Z_{jki}\right| \geq \frac{\delta}{2n}\right) \leq 2 \exp\left\{-\frac{\delta^2}{64eK_4^2 C_2 L_n^{-1} n + 8K_4 \delta}\right\} \quad (20)$$

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), (17), and Lemmas 1 and 2, we have for every $m \geq 2$, there exists a constant $K_5 > 0$, such that

$$E|\xi_i|^m = E[B_k^m(W_i) E[|X_{ji} \varepsilon_i|^m | W_i]] \leq m! K_5^{m-2} (2eK_5^2 C_2 L_n^{-1})/2.$$

Thus, according to Lemma 3,

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n \xi_i\right| \geq \frac{\delta}{2n}\right) \leq 2 \exp\left\{-\frac{\delta^2}{16eK_5^2 C_2 L_n^{-1} n + 4K_5 \delta}\right\}. \quad (21)$$

Similarly, we can show that

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n B_k(W_i) m(\mathbf{X}_i^*) - E[B_k(W) m(\mathbf{X}^*)]\right| \geq \frac{\delta}{2n}\right) \leq 2 \exp\left\{-\frac{\delta^2}{64eK_2^2 C_2 L_n^{-1} n + 8K_2 \delta}\right\} \quad (22)$$

and

$$P \left(\left| \frac{1}{n} \sum_{i=1}^n B_k(W_i) \varepsilon_i \right| \geq \frac{\delta}{2n} \right) \leq 2 \exp \left\{ - \frac{\delta^2}{16eK_3^2 C_2 L_n^{-1} n + 4K_3 \delta} \right\}. \quad (23)$$

Let $b_1 = 16eC_2 \max(4K_4^2, K_5^2, 4K_2^2, K_3^2)$ and $b_2 = \max(8K_4, 4K_5, 8K_2, 4K_3)$. Then, the combination of (20) – (23) 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 $j = 1, \dots, p$ and $\Sigma_j = E[\mathbf{Q}_j^T \mathbf{Q}_j]$,*

$$C_3 L_n^{-1} \leq \lambda_{\min}(\Sigma_j) \leq \lambda_{\max}(\Sigma_j) \leq C_4 L_n^{-1}, \quad (24)$$

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 $\|\boldsymbol{\eta}\|_2 = 1$,

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

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 $\lambda_1 + \lambda_2 = 1 + E[X_j^2 | W]$ (trace) and $\lambda_1 \cdot \lambda_2 = \text{Var}[X_j | W]$ (determinant). Therefore, by Lemma 1 $\lambda_1 \leq 1 + E[X_j^2 | W] \leq 1 + 4eK_1^2$ and by assumption (i)

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

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

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

By (18), we have

$$\frac{h_1 C_1}{1 + 4eK_1^2} L_n^{-1} \leq \boldsymbol{\eta}^T \Sigma_j \boldsymbol{\eta} \leq (1 + 4eK_1^2) C_2 L_n^{-1}.$$

Take $C_3 = h_1 C_1 L_n^{-1} / (1 + 4eK_1^2)$ and $C_4 = (1 + 4eK_1^2) C_2 L_n^{-1}$, 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 $j = 1, \dots, p$, there exist some positive constants b_3 and b_4 such that*

$$P\left(\left\|\Sigma_{nj} - \Sigma_j\right\| \geq L_n \delta/n\right) \leq 6L_n^2 \exp\left\{-\frac{\delta^2}{b_3 L_n^{-1} n + b_4 \delta}\right\},$$

$$P\left(\left\|\frac{1}{n} \mathbf{B}_n^T \mathbf{B}_n - \mathbb{E}[\mathbf{B}^T \mathbf{B}]\right\| \geq L_n \delta/n \leq 6L_n^2 \exp\left\{-\frac{\delta^2}{b_3 L_n^{-1} n + b_4 \delta}\right\},$$

where $\Sigma_{nj} = \frac{1}{n} \mathbf{Q}_{nj}^T \mathbf{Q}_{nj}$. In addition, for any given positive constant b_5 , there exists some positive constant b_6 such that

$$P\left(\left\|(\Sigma_{nj})^{-1}\right\| - \left\|(\Sigma_j)^{-1}\right\| \geq b_5 \left\|(\Sigma_j)^{-1}\right\|\right) \leq 6L_n^2 \exp\{-b_6 L_n^{-3} n\},$$

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

$$P\left(\left\|\left(\frac{1}{n} \mathbf{B}_n^T \mathbf{B}_n\right)^{-1}\right\| - \left\|\mathbb{E}[\mathbf{B}^T \mathbf{B}]\right\|^{-1}\right| \geq b_7 \left\|\mathbb{E}[\mathbf{B}^T \mathbf{B}]\right\|^{-1}\right) \leq 6L_n^2 \exp\{-b_8 L_n^{-3} n\}.$$

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

$$\Sigma_{nj} - \Sigma_j = \begin{pmatrix} \mathbf{D}_1 & \mathbf{D}_{2j} \\ \mathbf{D}_{2j}^T & \mathbf{D}_{3j} \end{pmatrix},$$

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

$$\left\|\Sigma_{nj} - \Sigma_j\right\| \leq 2L_n \left\|\Sigma_{nj} - \Sigma_j\right\|_{\infty} = 2L_n \max(\left\|\mathbf{D}_1\right\|_{\infty}, \left\|\mathbf{D}_{2j}\right\|_{\infty}, \left\|\mathbf{D}_{3j}\right\|_{\infty}). \quad (25)$$

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

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

for all k and l . By (17),

$$\text{Var}(B_k(W_i) B_l(W_i) - \mathbb{E}[B_k(W) B_l(W)]) \leq \mathbb{E}[B_k^2(W) B_l^2(W)] \leq C_2 L_n^{-1}.$$

By Lemma 4, we have

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n B_k(W_i) B_l(W_i) - \mathbb{E}[B_k(W) B_l(W)]\right| \geq \delta/6n\right) \leq 2 \exp\{-\delta^2 / (72 C_2 L_n^{-1} n + 24\delta)\}.$$

It then follows from the union bound of probability that

$$P(\|\mathbf{D}_1\|_\infty \geq \delta/6n) \leq 2L_n^2 \exp\{-\delta^2/(72C_2L_n^{-1}n+24\delta)\}. \quad (26)$$

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

$$E[|X_{ji}B_k(W_i)B_l(W_i) - E[X_jB_k(W)B_l(W)]|^m] \leq 2^m E[|X_{ji}B_k(W_i)B_l(W_i)|^m] \leq m!(2K_1)^{m-2}(8eK_1^2C_2L_n^{-1})/2.$$

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

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_{ji}B_k(W_i)B_l(W_i) - E[X_jB_k(W)B_l(W)]\right| \geq \delta/6n\right) \leq 2\exp\{-\delta^2/(576eK_1^2C_2L_n^{-1}n+24K_1\delta)\}.$$

It then follows from the union bound of probability that

$$P(\|\mathbf{D}_{2j}\|_\infty \geq \delta/6n) \leq 2L_n^2 \exp\{-\delta^2/(576eK_1^2C_2L_n^{-1}n+24K_1\delta)\}. \quad (27)$$

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

$$E[|X_{ji}^2B_k(W_i)B_l(W_i) - E[X_j^2B_k(W)B_l(W)]|^m] \leq m!(2K_6)^{m-2}(8eK_6^2C_2L_n^{-1})/2.$$

By Lemma 3, we have

$$P(|X_{ji}^2B_k(W_i)B_l(W_i) - E[X_j^2B_k(W)B_l(W)]| \geq \delta/6n) \leq 2\exp\{-\delta^2/(576eK_6^2C_2L_n^{-1}n+24K_6\delta)\}.$$

It then follows from the union bound of probability that

$$P(\|\mathbf{D}_{3j}\|_\infty \geq \delta/6n) \leq 2L_n^2 \exp\{-\delta^2/(576eK_6^2C_2L_n^{-1}n+24K_6\delta)\}. \quad (28)$$

Let $b_3 = 72C_2 \max\{1, 8eK_1^2, 8eK_6^2\}$ and $b_4 = 24 \max\{1, K_1, K_6\}$, then combining (25)–(28) we have

$$P(\|\Sigma_{nj} - \Sigma_j\| \geq L_n\delta/n) \leq 6L_n^2 \exp\left\{-\frac{\delta^2}{b_3L_n^{-1}n+b_4\delta}\right\}.$$

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

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

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),

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \leq \max\{|\lambda_{\min}(\mathbf{A} - \mathbf{B})|, |\lambda_{\min}(\mathbf{B} - \mathbf{A})|\}. \quad (29)$$

It then follows from (29) that

$$|\lambda_{\min}(\boldsymbol{\Sigma}_{nj}) - \lambda_{\min}(\boldsymbol{\Sigma}_j)| \leq 2L_n \|\boldsymbol{\Sigma}_{nj} - \boldsymbol{\Sigma}_j\|_{\infty},$$

which implies that

$$P(|\lambda_{\min}(\boldsymbol{\Sigma}_{nj}) - \lambda_{\min}(\boldsymbol{\Sigma}_j)| \geq L_n \delta / n) \leq 6L_n^2 \exp\{-\delta^2 / (b_3 L_n^{-1} n + b_4 \delta)\}. \quad (30)$$

Let $\delta = b_9 C_3 L_n^{-2} n$ in (30) for $b_9 \in (0, 1)$. According to (24), we have

$$P(|\lambda_{\min}(\boldsymbol{\Sigma}_{nj}) - \lambda_{\min}(\boldsymbol{\Sigma}_j)| \geq b_9 \lambda_{\min}(\boldsymbol{\Sigma}_j)) \leq 6L_n^2 \exp(-b_6 L_n^{-3} n). \quad (31)$$

for some positive constant b_6 . Next observe the fact that for $x, y > 0$, $a \in (0, 1)$ and $b = 1/(1 - a) - 1$, $|x^{-1} - y^{-1}| \geq by^{-1}$, which 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 (31) that

$$P(|(\lambda_{\min}(\boldsymbol{\Sigma}_{nj}))^{-1} - (\lambda_{\min}(\boldsymbol{\Sigma}_j))^{-1}| \geq b_5 (\lambda_{\min}(\boldsymbol{\Sigma}_j))^{-1}) \leq 6L_n^2 \exp(-b_6 L_n^{-3} n).$$

Following the same proof, by (18) we also have for any positive constant b_7 , there exists some positive constant b_8 , such that

$$P\left(\left|(\lambda_{\min}(\frac{1}{n} \mathbf{B}_n^T \mathbf{B}_n))^{-1} - (\lambda_{\min}(E[\mathbf{B}^T \mathbf{B}]))^{-1}\right| \geq b_7 (\lambda_{\min}(E[\mathbf{B}^T \mathbf{B}]))^{-1}\right) \leq 6L_n^2 \exp(-b_8 L_n^{-3} n).$$

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 $E[Y|W, X_j] = a_j(W) + b_j(W)X_j$. By Stone (1982), there exist $\{a_j^*\}_{j=0}^p$ and $\{b_j^*\}_{j=1}^p \in \mathcal{S}_n$ such that $\|a_j - a_j^*\|_{\infty} \leq M_2 L_n^{-d}$ and $\|b_j - b_j^*\|_{\infty} \leq M_2 L_n^{-d}$, where \mathcal{S}_n is the space of polynomial splines of degree $l \geq 1$ with normalized B-spline basis $\{B_k, k = 1, \dots, L_n\}$, 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 $\tilde{a}_j(W) = \mathbf{B}(W)\boldsymbol{\eta}_j$ and $\tilde{b}_j(W) = \mathbf{B}(W)\boldsymbol{\theta}_j$. By definition of $\boldsymbol{\eta}_j$ and $\boldsymbol{\theta}_j$, we have

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

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

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

On the other hand, by the least-squares property and conditioning in W_j and X_j ,

$$E[(Y - \tilde{a}_j - \tilde{b}_j X_j)(\tilde{a}_j + \tilde{b}_j X_j)] = 0, \quad E[(Y - a_j - b_j X_j)(\tilde{a}_j + \tilde{b}_j X_j)] = 0.$$

The last two equalities imply that $E[(a_j + b_j X_j - \tilde{a}_j - \tilde{b}_j X_j)(\tilde{a}_j + \tilde{b}_j X_j)] = 0$. Thus, by the Pythagorean theorem, we have

$$\|a_j + b_j X_j\|^2 = \|\tilde{a}_j + \tilde{b}_j X_j\|^2 + \|\tilde{a}_j + \tilde{b}_j X_j - a_j - b_j X_j\|^2, \quad \|a_j + b_j X_j\|^2 - \|\tilde{a}_j + \tilde{b}_j X_j\|^2 \leq 2M_2^2 L_n^{-2d}(1 + E[X_j^2]). \quad (32)$$

Similarity, we have

$$\|a_0\|^2 - \|\tilde{a}_0\|^2 \leq M_2^2 L_n^{-2d}. \quad (33)$$

By taking $M_1 = M_2^2(8eK^2 + 3)$ (c.f. Lemma 1), the first part of Proposition 1 follows from (32) and (33):

$$u_j - \tilde{u}_j = \|a_j + b_j X_j\|^2 - \|a_0\|^2 - (\|\tilde{a}_j + \tilde{b}_j X_j\|^2 - \|\tilde{a}_0\|^2) \leq M_1 L_n^{-2d}. \quad (34)$$

By (11) and (34), we have $\min_{j \in \mathcal{M}_*} \tilde{u}_j \geq c_1 L_n n^{-2\kappa} / h_2 - M_1 L_n^{-2d}$. 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 $\hat{u}_{nj} - \tilde{u}_j = S_1 + S_2$, where

$$S_1 = \frac{1}{n^2} \mathbf{Y}^T \mathbf{Q}_{nj} \Sigma_{nj}^{-1} \mathbf{Q}_{nj}^T \mathbf{Y} - E[\mathbf{Y} \mathbf{Q}_j] \Sigma_j^{-1} E[\mathbf{Q}_j^T \mathbf{Y}], \quad \text{and} \quad S_2 = \frac{1}{n} \mathbf{Y}^T \mathbf{B}_n (\mathbf{B}_n^T \mathbf{B}_n)^{-1} \mathbf{B}_n^T \mathbf{Y} - E[\mathbf{Y} \mathbf{B}] (E[\mathbf{B}^T \mathbf{B}])^{-1} E[\mathbf{B}^T \mathbf{Y}].$$

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

$$S_1 = (\mathbf{a}_n - \mathbf{a})^T \Sigma_{nj}^{-1} (\mathbf{a}_n - \mathbf{a}) + 2(\mathbf{a}_n - \mathbf{a})^T \Sigma_{nj}^{-1} \mathbf{a} + \mathbf{a}^T (\Sigma_{nj}^{-1} - \Sigma_j^{-1}) \mathbf{a}.$$

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

We first deal with S_{11} . Note that

$$|S_{11}| \leq \|\boldsymbol{\Sigma}_{n_j}^{-1}\| \cdot \|\mathbf{a}_n - \mathbf{a}\|_2^2. \quad (35)$$

By Lemma 5 and the union bound of probability,

$$P(\|\mathbf{a}_n - \mathbf{a}\|_2^2 \geq 2L_n\delta^2/n^2) \leq 8L_n \exp\{-\delta^2/(b_1L_n^{-1}n + b_2\delta)\}. \quad (36)$$

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

$$P(\|\boldsymbol{\Sigma}_{n_j}^{-1}\| - \|\boldsymbol{\Sigma}_j^{-1}\| \geq b_5\|\boldsymbol{\Sigma}_j^{-1}\|) \leq 6L_n^2 \exp\{-b_6L_n^{-3}n\}.$$

Then it follows from (24) that

$$P(\|\boldsymbol{\Sigma}_{n_j}^{-1}\| \geq (b_5+1)C_3^{-1}L_n) \leq 6L_n^2 \exp\{-b_6L_n^{-3}n\}. \quad (37)$$

Combining (35)–(37) and based on the union bound of probability, we have

$$P(|S_{11}| \geq 2(b_5+1)C_3^{-1}L_n^2\delta^2/n^2) \leq 8L_n \exp\{-\delta^2/(b_1L_n^{-1}n + b_2\delta)\} + 6L_n^2 \exp\{-b_6L_n^{-3}n\}. \quad (38)$$

We next bound S_{12} . Note that

$$|S_{12}| \leq 2\|\mathbf{a}_n - \mathbf{a}\|_2 \cdot \|\boldsymbol{\Sigma}_{n_j}^{-1}\| \cdot \|\mathbf{a}\|_2 \quad (39)$$

By Lemma 1,

$$\|\mathbf{a}\|_2^2 = \|\mathbf{E}[\mathbf{B}^T Y]\|_2^2 + \|\mathbf{E}[X_j \mathbf{B}^T Y]\|_2^2 \leq \sum_{k=1}^{L_n} (\mathbf{E}[B_k^2 m^2(\mathbf{X}^*)] + \mathbf{E}[B_k^2 X_j^2 m^2(\mathbf{X}^*)]) \leq 4eC_2(K_2^2 + K_4^2), \quad (40)$$

where the calculation as in (19) was used.

It follows from (36), (37), (39), (40) and the union bound of probability that

$$P(|S_{12}| \geq 4\sqrt{2}(b_5+1)e^{1/2}C_2^{1/2}(K_2^2 + K_4^2)^{1/2}C_3^{-1}L_n^{3/2}\delta/n) \leq 8L_n \exp\{-\delta^2/(b_1L_n^{-1}n + b_2\delta)\} + 6L_n^2 \exp\{-b_6L_n^{-3}n\}. \quad (41)$$

To bound S_{13} , note that

$$|S_{13}| = \mathbf{a}^T \boldsymbol{\Sigma}_{n_j}^{-1} (\boldsymbol{\Sigma}_j - \boldsymbol{\Sigma}_{n_j}) \boldsymbol{\Sigma}_j^{-1} \mathbf{a} \leq \|\boldsymbol{\Sigma}_{n_j}^{-1}\|^2 \cdot \|\boldsymbol{\Sigma}_j - \boldsymbol{\Sigma}_{n_j}\| \cdot \|\mathbf{a}\|_2^2. \quad (42)$$

Then it follows from Lemmas 6, Lemma 7, (37), (40), (42) and the union bound of probability that there exist b_3 , b_4 and b_6 such that

$$P(|S_{13}| \geq 4eC_2(K_2^2 + K_4^2)(b_5+1)^2C_3^{-2}L_n^3\delta/n) \leq 6L_n^2 \exp\{-\delta^2/(b_3L_n^{-1}n + b_4\delta)\} + 6L_n^2 \exp\{-b_6L_n^{-3}n\}. \quad (43)$$

Hence, combining (38), (41) and (43), there exist some positive constants s_1, s_2 and s_3 such that

$$\begin{aligned} P(|S_1| \geq s_1 L_n^2 \delta^2 / n^2 &+ s_2 L_n^{3/2} \delta / n \\ &+ s_3 L_n^3 \delta / n) \leq 16 L_n \exp\{ \\ &- \delta^2 / (b_1 L_n^{-1} n + b_2 \delta)\} \\ &+ 6 L_n^2 \exp\{-\delta^2 / (b_3 L_n^{-1} n + b_4 \delta)\} \\ &+ 18 L_n^2 \exp\{ \\ &- b_6 L_n^{-3} n\}. \end{aligned}$$

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

$$\begin{aligned} P(|S_2|) \geq s_4 L_n^2 \delta^2 / n^2 &+ s_5 L_n^{3/2} \delta / n \\ &+ s_6 L_n^3 \delta / n) \leq 8 L_n \exp\{ \\ &- \delta^2 / (b_1 L_n^{-1} n + b_2 \delta)\} \\ &+ 6 L_n^2 \exp\{-\delta^2 / (b_3 L_n^{-1} n + b_4 \delta)\} \\ &+ 18 L_n^2 \exp\{ \\ &- b_8 L_n^{-3} n\}. \end{aligned}$$

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

$$P(|\hat{u}_{nj} - \tilde{u}_j| \geq c_2 L_n n^{-2\kappa}) \leq (24 L_n + 12 L_n^2) \exp\{-c_3 n^{1-4\kappa} L_n^{-3}\} + 36 L_n^2 \exp\{-c_4 L_n^{-3} n\}.$$

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

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

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

by Proposition 1, we have $\hat{u}_{nj} \geq c_1 \xi L_n n^{-2\kappa} / 2$, for all $j \in \mathcal{M}_*$. Hence, by choosing $\tau_n = c_1 \xi L_n n^{-2\kappa} / 2$, we have $\mathcal{M}_* \subset \hat{\mathcal{M}}_{\tau_n}$. On the other hand, by the union bound of probability, there exist positive constants c_6 and c_7 , such that and

$$P(\mathcal{A}_n^c) \leq s_n \{(24 L_n + 12 L_n^2) \exp(-c_6 n^{1-4\kappa} L_n^{-3}) + 36 L_n^2 \exp(-c_7 L_n^{-3} n)\},$$

Theorem 1(2) follows.

Proof of Theorem 2. Let $\tilde{\alpha} = \underset{\alpha}{\operatorname{argmin}} E[(Y - \mathbf{Q}\alpha)^2]$, where $\mathbf{Q} = (\mathbf{Q}_1, \dots, \mathbf{Q}_p)$ is a $2pL_n$ -dimensional vector of functions. Then we have

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

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

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

recalling $\Sigma = E[\mathbf{Q}^T \mathbf{Q}]$. It follows from orthogonal decomposition that $\operatorname{Var}(\mathbf{Q}\tilde{\alpha}) \leq \operatorname{Var}(Y)$ and $E[\mathbf{Q}\tilde{\alpha}] = E[Y]$ (recall the inclusion of the intercept term). Therefore, $\tilde{\alpha}^T \Sigma \tilde{\alpha} \leq E[Y^2] = O(1)$, and

$$\|E[\mathbf{Q}^T Y]\|_2^2 = O(\lambda_{\max}(\Sigma)). \quad (44)$$

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

$$\sum_{j=1}^p \tilde{u}_j \leq \max_{1 \leq j \leq p} \lambda_{\max}\{(\Sigma_j)^{-1}\} \sum_{j=1}^p \|E[\mathbf{Q}_j^T Y]\|_2^2 = \max_{1 \leq j \leq p} \lambda_{\max}\{(\Sigma_j)^{-1}\} \|E[\mathbf{Q}^T Y]\|_2^2.$$

By Lemma 6 and (44), the last term is of order $O(L_n \lambda_{\max}(\Sigma))$. This implies that the number of $\{j : \tilde{u}_j > \delta L_n n^{-2\kappa}\}$ cannot exceed $O(n^{2\kappa} \lambda_{\max}(\Sigma))$ for any $\delta > 0$.

On the set $\mathcal{B}_n = \{\max_{1 \leq j \leq p} |\hat{u}_{nj} - \tilde{u}_j| \leq \delta L_n n^{-2\kappa}\}$, the number of $\{j : \hat{u}_{nj} > 2\delta L_n n^{-2\kappa}\}$ cannot exceed the number of $\{j : \tilde{u}_j > \delta L_n n^{-2\kappa}\}$, which is bounded by $O(n^{2\kappa} \lambda_{\max}(\Sigma))$. By taking $\delta = c_5/2$, we have

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

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

References

- Antoniadis A, Fan J. Regularized wavelet approximations (with discussion). *Jour. Amer. Statist. Assoc.* 2001; 96:939–967.
- Candes E, Tao T. The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* 2007; 35:2313–2404.
- Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Am. Statist. Assoc.* 2011; 106:544–557.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* 2001; 96:1348–1360.
- Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. B.* 2008; 70:849–911.
- Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* 2010; 38:3567–3604.

- Fan J, Zhang C, Zhang J. Generalized likelihood ratio statistics and wilks phenomenon. *Ann. Statist.* 2001; 29:153–193.
- Fan J, Zhang W. Statistical methods with varying coefficient models. *Stat. Interface.* 2008; 1:179–195. [PubMed: 18978950]
- Frank IE, Friedman JH. A statistical view of some chemometrics regression tools (with discussion). *Technometrics.* 1993; 35:109–148.
- Hall P, Miller H. Using Generalised Correlation to Effect Variable Selection in Very High Dimensional Problems. *Journal of Computational and Graphical Statistics.* 2009; 18:533–550.
- Hall P, Titterton DM, Xue JH. Tilting methods for assessing the influence of components in a classifier. *J. Roy. Statist. Soc. B.* 2009; 71:783–803.
- Harrison D, Rubinfeld D. Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.* 1978; 5:81–102.
- Hastie T, Tibshirani R. *Generalized Additive Models.* London: Chapman & Hall; 1990.
- Hastie T, Tibshirani R. Varying-coefficient models. *J. Roy. Statist. Soc. B.* 1993; 55:757–796.
- Li G, Peng H, Zhang J, Zhu L. Robust rank correlation based screening. *Ann. Statist.* 2012; 40:1846–1877.
- Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. *J. Am. Statist. Assoc.* 2012 to appear.
- Lian H. Flexible shrinkage estimation in high-dimensional varying coefficient models. manuscript. 2011
- Stone CJ. Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* 1982; 10:1040–1053.
- Tibshirani R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B.* 1996; 58:267–288.
- van der Vaart, AW.; Wellner, JA. *Weak Convergence and Empirical Processes.* New York: Springer; 1996.
- Vershynin R. Introduction to the non-asymptotic analysis of random matrices. manuscript. 2011
- Wang L, Li H, Huang JZ. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Am. Statist. Assoc.* 2008; 103:1556–1569.
- Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Jour. Roy. Statist. Soc. B.* 2006; 68:49–67.
- Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* 2010; 38:894–942.
- Zhao, DS.; Li, Y. Principled Sure Independence Screening for Cox Models With Ultra-High-Dimensional Covariates. Harvard University; 2010. *manuscript*
- Zou H. The adaptive Lasso and its oracle properties. *J. Am. Statist. Assoc.* 2006; 101:1418–1429.
- Zou H, Hastie T. Addendum: Regularization and variable selection via the Elastic net. *J. Roy. Statist. Soc. B.* 2005; 67:301–320.
- Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* 2008; 36:1509–1533.

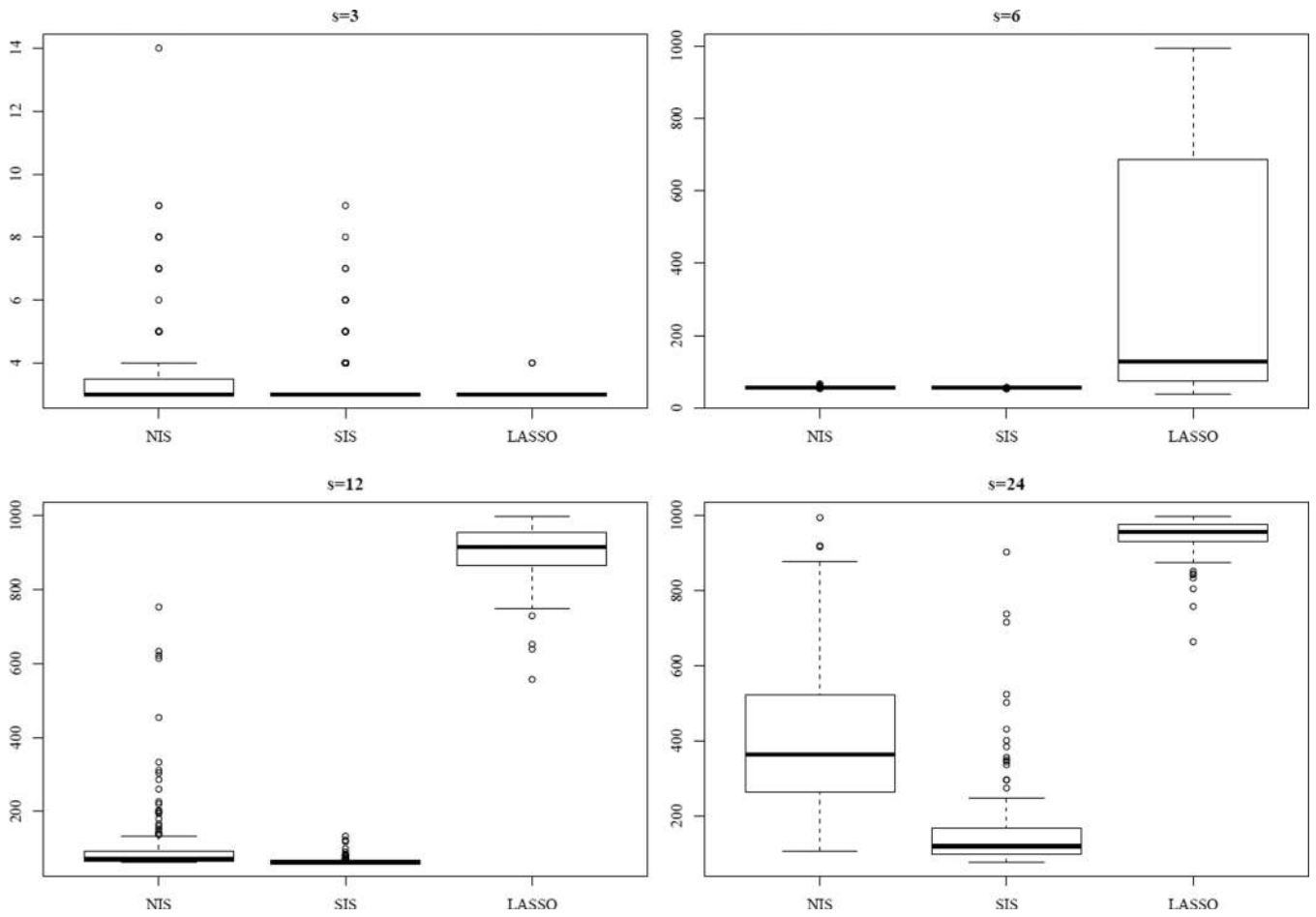


Figure 1. Boxplots of minimum model size (left to right: NIS, Lasso and SIS) for Example 1 under different true models.

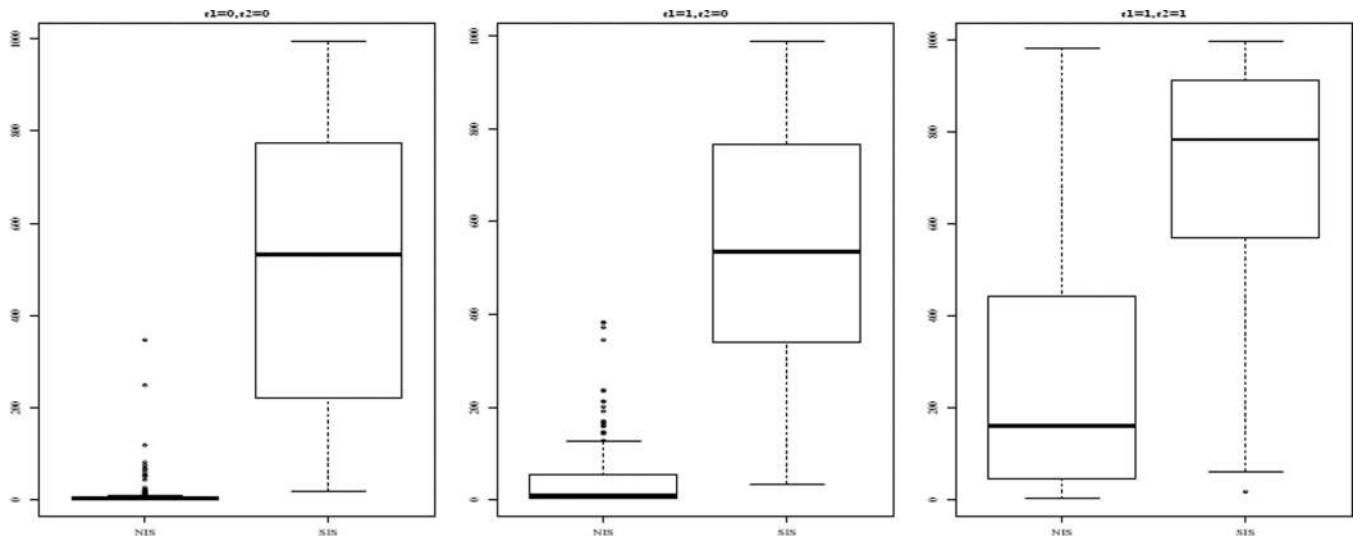


Figure 2. Boxplots of minimum model size (left: NIS, right: SIS) for Example 2 under different correlation settings.

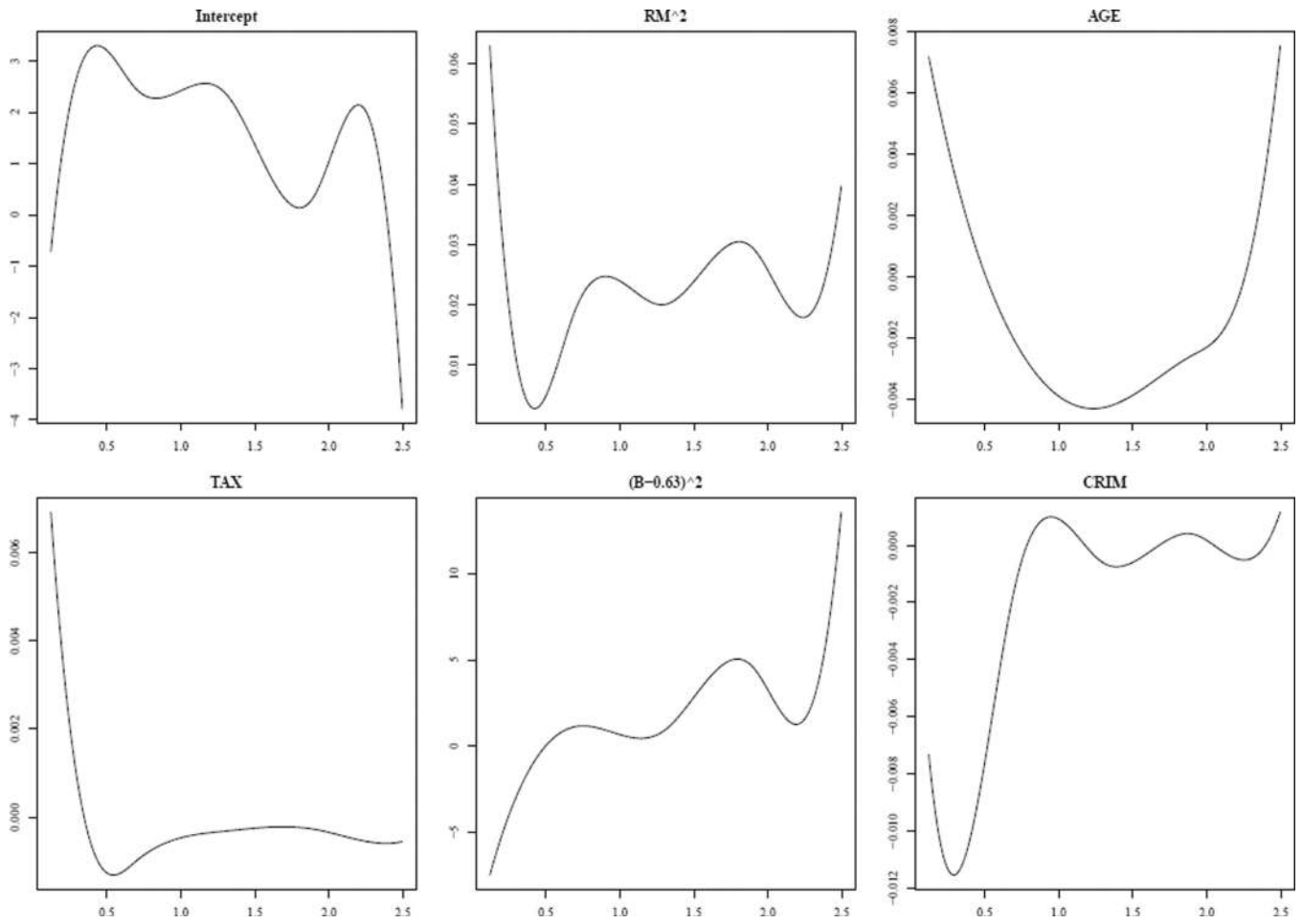


Figure 3. Fitted functional estimates $\hat{\beta}_j(W)$'s selected by Conditional-INIS.

Average values of the number of true positives (TP), model size, the minimum true signal, the minimum false signal and the maximum null signal using conditional permutation with different K 's for simulated model in Example 3 under different correlation settings. Robust standard deviations are given in parentheses.

Table 1

Model	TP	Size	$\min_{j \in \mathcal{M}^* \setminus \mathcal{M}_0} \hat{u}_{nj}^*$	$\max_{j \in \mathcal{M}^* \setminus \mathcal{M}_0} \hat{u}_{nj}^*$	$\max_{j \in \{1, \dots, p\} \setminus \mathcal{M}_0} \hat{u}_{nj}^{*\pi}$
$K=0$	$t_1=0, t_2=0$	6.68(2.99)	2.96(0.72)	1.22(0.18)	1.12(0.15)
	$t_1=3, t_2=1$	4.00(0.00)	886.49(88.81)	0.61(0.10)	0.22(0.03)
$K=1$	$t_1=0, t_2=0$	5.70(1.49)	2.83(0.57)	0.75(0.10)	0.72(0.11)
	$t_1=3, t_2=1$	4.00(0.00)	202.50(154.85)	0.28(0.06)	0.11(0.02)
$K=4$	$t_1=0, t_2=0$	5.14(1.49)	NA	0.06(0.01)	0.06(0.01)
	$t_1=3, t_2=1$	4.00(0.00)	4.98(0.75)	0.16(0.05)	0.06(0.01)
$K=8$	$t_1=0, t_2=0$	8.92(0.75)	NA	0.05(0.01)	0.05(0.01)
	$t_1=3, t_2=1$	3.99(0.00)	8.43(0.75)	0.11(0.03)	0.05(0.01)

Table 2

Average values of the number of true positives (TP), false positives (FP), and prediction error (PE) using Conditional-INIS with different K 's for simulated model in Example 3 under different correlation settings. Robust standard deviations are given in parentheses.

Model	TP	FP	PE	
$K = 3$	$t_1 = 0, t_2 = 0$	4.00(0.00)	1.27(1.49)	1.16(0.13)
	$t_1 = 3, t_2 = 1$	3.96(0.00)	0.13(0.00)	1.32(0.10)
$K = 5$	$t_1 = 0, t_2 = 0$	4.00(0.00)	1.57(1.49)	1.34(0.14)
	$t_1 = 3, t_2 = 1$	3.97(0.00)	0.05(0.00)	1.31(0.15)
$K = 10$	$t_1 = 0, t_2 = 0$	4.00(0.00)	1.20(1.49)	1.20(0.14)
	$t_1 = 3, t_2 = 1$	3.99(0.00)	0.20(0.00)	1.39(0.07)

Table 3

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
$t_1 = 0, t_2 = 0$ (SNR ≈ 16.85)	0	0	4.00 (0.00)	1.57 (1.49)	1.34 (0.14)	4.00 (0.00)	0.06 (0.00)	1.17 (0.06)
$t_1 = 2, t_2 = 0$ (SNR ≈ 3.66)	0.25	0	4.00 (0.00)	0.15 (0.00)	0.89 (0.05)	4.00 (0.00)	0.00 (0.00)	0.99 (0.05)
$t_1 = 2, t_2 = 1$ (SNR ≈ 3.21)	0.25	0.36	4.00 (0.00)	0.12 (0.00)	1.24 (0.10)	3.99 (0.00)	0.02 (0.00)	1.32 (0.13)
$t_1 = 3, t_2 = 0$ (SNR ≈ 3.32)	0.43	0	4.00 (0.00)	0.01 (0.00)	1.13 (0.06)	3.98 (0.00)	0.00 (0.00)	1.19 (0.06)
$t_1 = 3, t_2 = 1$ (SNR ≈ 2.81)	0.43	0.46	3.97 (0.00)	0.05 (0.00)	1.31 (0.15)	3.96 (0.00)	0.03 (0.00)	1.17 (0.10)

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

Table 4

Model	Correlation			Conditional-INIS			Greedy-INIS		
	X's	X's-W	TP	FP	PE	TP	FP	PE	
$t_1 = 0, t_2 = 0$ (SNR ≈ 47.68)	0	0	8.00 (0.00)	1.26 (1.49)	1.46 (0.15)	8.00 (0.00)	0.08 (0.00)	1.25 (0.12)	
$t_1 = 2, t_2 = 0$ (SNR ≈ 9.40)	0.25	0	8.00 (0.00)	0.08 (0.00)	1.14 (0.10)	8.00 (0.00)	0.00 (0.00)	1.34 (0.12)	
$t_1 = 2, t_2 = 1$ (SNR ≈ 8.62)	0.25	0.36	8.00 (0.00)	0.30 (0.00)	1.60 (0.34)	8.00 (0.00)	0.10 (0.00)	1.98 (0.47)	
$t_1 = 3, t_2 = 0$ (SNR ≈ 8.18)	0.43	0	7.99 (0.00)	0.02 (0.00)	1.30 (0.08)	7.98 (0.00)	0.02 (0.00)	1.17 (0.10)	
$t_1 = 3, t_2 = 1$ (SNR ≈ 7.61)	0.43	0.46	7.98 (0.00)	0.26 (0.00)	1.70 (0.19)	7.90 (0.00)	0.25 (0.00)	1.76 (0.48)	

Table 5

Average values of prediction error (PE), model size and the number of selected noise variables (SNV) over 100 repetitions for Conditional-INIS ($p = 1000$), Greedy-INIS ($p = 1000$), and SCAD fit ($p=12$). Robust standard deviations are given in parentheses.

method	PE	Size	SNV
Conditional-INIS ($p = 1000$)	0.046(0.020)	5.55(0.75)	0.00(0.00)
Greedy-INIS ($p = 1000$)	0.048(0.020)	4.80(1.49)	0.01(0.00)
SCAD fit ($p=12$)	0.052(0.019)	6.05(1.87)	NA