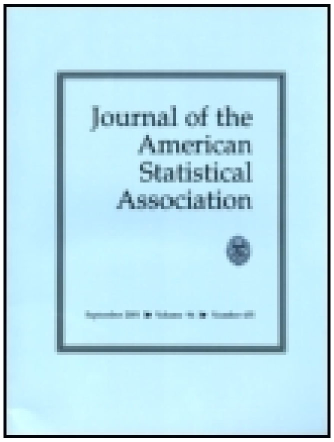

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

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

Jianqing Fan, Yunbei Ma & Wei Dai

Accepted author version posted online: 14 Jan 2014.Published online: 02 Oct 2014.

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

Jianqing FAN, Yunbei MA, and Wei DAI

The varying coefficient model is an important class of nonparametric statistical model, which 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 article, 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 (INIS) 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.

KEY WORDS: 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 single nucleotide polymorphisms (SNPs) are potential covariates. Traditional statistical methods face significant challenges when dealing with such high-dimensional problems.

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 (Frank and Friedman 1993), lasso (Tibshirani 1996), smoothly clipped absolute deviation (SCAD) and folded concave penalty (Fan and Li 2001), the elastic net (Zou and Hastie 2005), adaptive lasso (Zou 2006), and the Dantzig selector (Candes and Tao 2007). Methods on the implementation of folded concave penalized least square 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, Titterington, and Xue 2009), marginal partial likelihood method (MPLE; Zhao and Li 2012), 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,

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

where Y is the response, W is some univariate observable exposure variable, $\mathbf{X} = (X_1, \ldots, X_p)^T$ 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 $\boldsymbol{\beta}(\cdot)$ 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, and 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

Jianqing Fan (E-mail: *jqfan@princeton.edu*) is Director, Center for Statistical Research, Academy of Mathematics and Systems Science, The Chinese Academy of Sciences, Beijing 100080, China, and the current Frederick L. Moore '18 Professor of Finance, Professor of Statistics, and Chairman; and Wei Dai (E-mail: *weidai@princeton.edu*) is graduate student, Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544. Yunbei Ma is Assistant Professor, School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China (E-mail: *myb@swufe.edu.cn*). 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.

Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2013.879828

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) used the adaptive group lasso penalty in time-varying coefficient models. These methods still face the aforementioned challenges of designing a robust algorithm with reasonable computational cost while achieving statistical precision. While existing theory and estimation methods are not directly applicable in the ultra-high-dimensional regime, we aim to address this need, both theoretically and practically.

In this article, 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:

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

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 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 toward (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 will be shown later, our assumptions are much weaker than 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 NIS 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 that adopts a greedy approach in the variable screening step. They both serve to effectively control the false positive (FP) and false negative (FN) 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 $\boldsymbol{\beta}(\cdot) = (\beta_1(\cdot), \dots, \beta_p(\cdot))^T$ is sparse. Let $\mathcal{M}_* = \{j : \mathrm{E}[\beta_j^2(W)] > 0\}$ be the true sparse model with non-sparsity 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:

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

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

$$b_j(W) = \frac{\mathrm{cov}[X_j, Y|W]}{\mathrm{var}[X_j|W]},\ a_j(W) = \mathrm{E}[Y|W] - b_j(W)\mathrm{E}[X_j|W].$$

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

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

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

$$u_j = \mathrm{E}\big[b_j^2(W)(X_j - \mathrm{E}[X_j|W])^2\big] = \mathrm{E}\left[\frac{(\mathrm{cov}[X_j, Y|W])^2}{\mathrm{var}[X_j|W]}\right]. \tag{4}$$

For each $j = 1, \dots, p$, 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, $j = 1, \dots, p$, 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

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

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

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:

$$\min_{\boldsymbol{\eta}_j, \boldsymbol{\theta}_j \in \mathbb{R}^{L_n}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{B}(W_i)\boldsymbol{\eta}_j - \mathbf{B}(W_i)\boldsymbol{\theta}_j X_{ji})^2, \tag{5}$$

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

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

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

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, $j = 1, \ldots, p$, are given by

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

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

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

where

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

We now define an estimate of the marginal utility u_j as

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

where $\mathbf{W} = (W_1, \ldots, W_n)^T$. Note that throughout this article, 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:

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

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

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

and we select variables as follows:

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

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

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

and

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

It follows from (8) and (9) that

$$\widehat{v}_{nj} = \|\mathbf{Y}\|_n^2 - \|\widehat{a}_{n0}(\mathbf{W})\|_n^2 - \widehat{u}_{nj}. \quad (10)$$

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

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

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

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

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

It can be seen that

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

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

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

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:

$$u_j = \mathrm{E}\left[\frac{(\mathrm{cov}[X_j, Y|W])^2}{\mathrm{var}[X_j|W]}\right], \quad j = 1, \ldots, p.$$

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 $j = 1, \ldots, p$, $\mathrm{var}[X_j|W]$ 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),

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

Note that in condition (ii), the number of basis functions L_n is not intrinsic. By the Remark 1, 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 rth derivative $f^{(r)}$ exists and is Lipschitz of order α. That is,

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

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

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

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

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

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

(vi) The random errors $\{\varepsilon_i\}_{i=1}^n$ are iid 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

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

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 $m(X^*)$, and the noise ε, 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

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

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 NIS 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

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

(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

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

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

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

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 (12), the ideal case for vanishing FP 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

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

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

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

In particular, this ideal situation occurs under the partial orthogonality condition, that is, $\{X_j\}_{j\in\mathcal{M}_*}$ 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 cannot be achieved by a single step of marginal screening. The marginal probes cannot 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

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

where $\boldsymbol{\Sigma} = \mathrm{E}[\mathbf{Q}^T\mathbf{Q}]$, and $\mathbf{Q} = (\mathbf{Q}_1, \ldots, \mathbf{Q}_p)$ is a functional vector of $2p_nL_n$ 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 $\lambda_{\max}(\boldsymbol{\Sigma})$ 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 $\boldsymbol{\Sigma}$ is block diagonal with *j*th block $E[\mathbf{Q}_j^T\mathbf{Q}_j]$, and therefore $\lambda_{\max}(\boldsymbol{\Sigma}) = O(L_n^{-1})$.

4. ITERATIVE NONPARAMETRIC INDEPENDENCE SCREENING

As Fan and Lv (2008) pointed out, in practice the NIS would still suffer from FN (i.e., miss some important predictors that are marginally weakly correlated but jointly correlated with the response) and FP (i.e., select some unimportant predictors that 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 (2012). Randomly permute $\mathbf{Y}$ to get $\mathbf{Y}_{\boldsymbol{\pi}} = (Y_{\pi_1}, \ldots, Y_{\pi_n})^T$ and compute $\widehat{u}_{nj}^{\boldsymbol{\pi}}$, where $\boldsymbol{\pi}$ is a permutation of $\{1, \ldots, n\}$, 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 qth-ranked magnitude of $\{\widehat{u}_{nj}^{\boldsymbol{\pi}}, j = 1, \ldots, p\}$. Thus, the NIS step selects variables $\{j : \widehat{u}_{nj} \ge \tau_{jq}\}$. 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 FPs. 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 $j = 1, \ldots, p$, 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 (6) and (7) using $\{(\mathbf{Y}, \mathbf{W}, \mathbf{X}_j), j = 1, \ldots, p\}$. 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_{jq}^* 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
$$\mathcal{A}_1 = \left\{j : \widehat{u}_{nj}^* \ge \tau_{jq}^*, j \in \mathcal{M}_0^c\right\} \cup \mathcal{M}_0.$$
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 Steps 1–2, where we replace $\mathcal{M}_0$ in Step 1 by $\mathcal{M}_l$, $l = 1, 2, \ldots$, 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 $k \le l$ or $|\mathcal{M}_{l+1}| \ge \zeta_n$, for some prescribed positive integer ζ_n (such as $[n/\log(n)]$).

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 $\widehat{u}_{nj}$. 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 $\mathcal{M}_l = \{j \in \mathcal{A}_l : \widehat{\boldsymbol{\gamma}}_j^{(l)} \ne \mathbf{0}\}$ through minimizing the following objective function:

$$\min_{\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_j \in \mathbb{R}^{L_n}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \mathbf{B}(W_i)\boldsymbol{\gamma}_0 - \sum_{j \in \mathcal{A}_l} \mathbf{B}(W_i) X_{ji} \boldsymbol{\gamma}_j \right)^2 + \sum_{j \in \mathcal{A}_l} p_\lambda(||\boldsymbol{\gamma}_j||_B), \tag{14}$$

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

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

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 Bayesian information criteria (BIC) $n \log(\widehat{\sigma}^2_\epsilon) + kL_n \log n$, where $\widehat{\sigma}^2_\epsilon$ 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 is: 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 the L_n is, the larger the 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 $L_n = \lfloor 2n^{1/5} \rceil$, where $\lfloor \cdot \rceil$ 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, that is, 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 iid standard normal random variables and

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

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

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

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

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

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 iid standard Gaussian random variable.

Under different correlation settings, the comparison of MMS between NIS and SIS methods is 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 of the conditional random permutation method.

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

$$X_j = \frac{Z_j + t_1 U_1}{1 + t_1}, \ j = 1, \ldots, p, \qquad W = \frac{U_2 + t_2 U_1}{1 + t_2},$$

$$Y = 2X_1 + 3W \cdot X_2 + (W+1)^2 \cdot X_3 + \frac{4\sin(2\pi W)}{2 - \sin(2\pi W)} \cdot X_4 + \epsilon.$$

We study two settings: $t_1 = t_2 = 0$, resulting in uncorrelated case and $t_1 = 3$ and $t_2 = 1$, corresponding to $\mathrm{corr}(X_j, X_k) = 0.43$ for all $j \ne k$ and $\mathrm{corr}(X_j, W) = 0.46$. We report the average of the number of true positives (TPs), 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

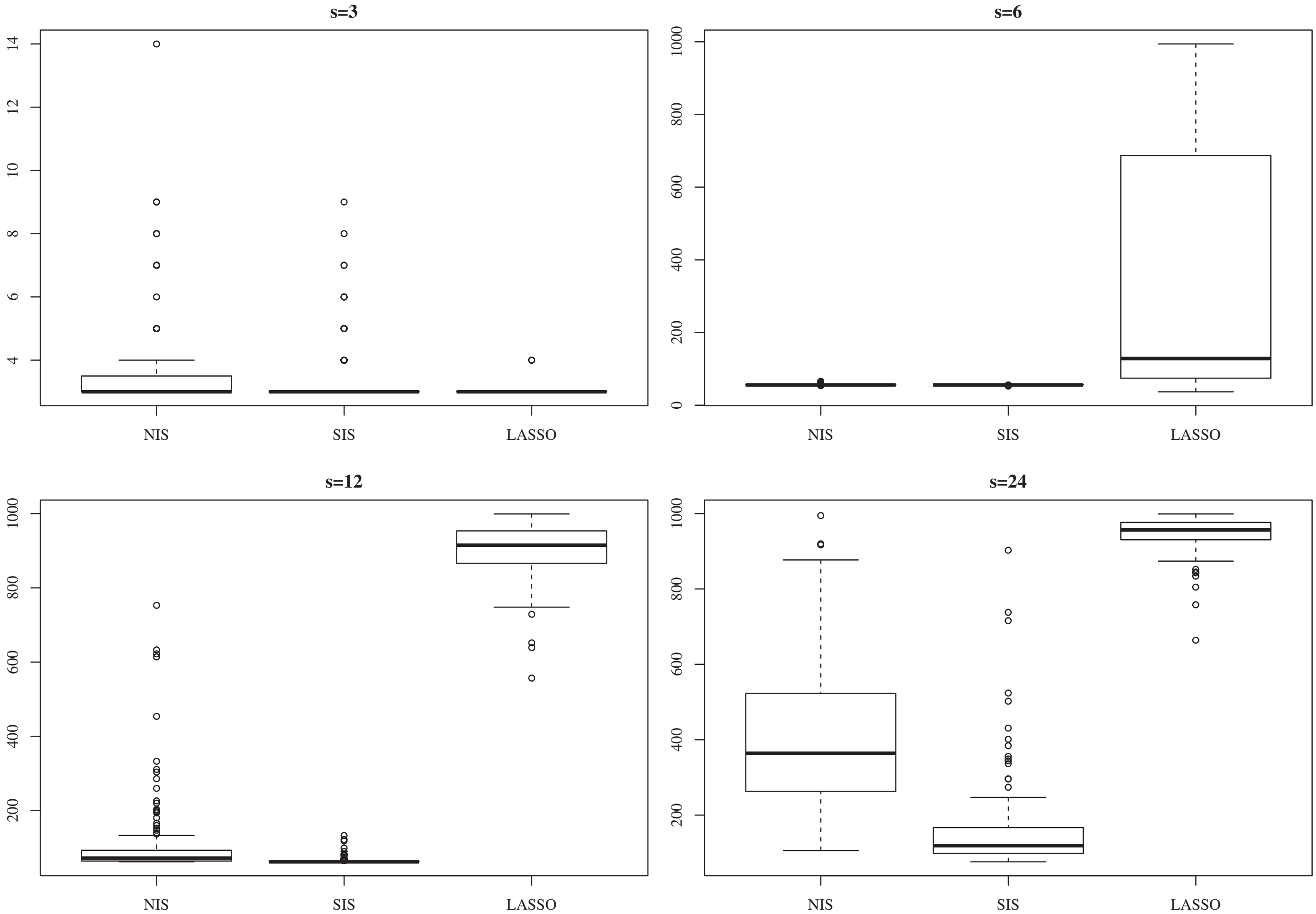

Figure 1. Boxplots of minimum model size (left to right: NIS, lasso, and SIS) for Example 1 under different true models.

tuning parameter K. Generally speaking, although the lower bound of the TPs' 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 FPs 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., $\mathcal{M}^*\backslash\mathcal{M}_0 = \emptyset$), 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 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 TP, FP, prediction error (PE), and their robust standard deviations. Here, the PE is the mean squared error calculated on the test dataset of size $n/2 = 200$ generated from

Table 1. Average values of the number of true positives (TPs), model size, the minimum true signal, the maximum 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

Model		TP	Size	$\min_{j\in\mathcal{M}^*\backslash\mathcal{M}_0} \widehat{u}^*_{nj}$	$\max_{j\in\mathcal{M}^{*c}\backslash\mathcal{M}_0} \widehat{u}^*_{nj}$	$\max_{j\in\{1,\dots,p\}\backslash\mathcal{M}_0} \widehat{u}^{*\boldsymbol{\pi}}_{nj}$
$K = 0$	$t_1 = 0, t_2 = 0$	4.00(0.00)	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)	0.58(0.07)	0.22(0.03)
$K = 1$	$t_1 = 0, t_2 = 0$	4.00(0.00)	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)	0.20(0.03)	0.11(0.02)
$K = 4$	$t_1 = 0, t_2 = 0$	4.00(0.00)	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)	0.05(0.01)	0.06(0.01)
$K = 8$	$t_1 = 0, t_2 = 0$	4.00(0.00)	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)	0.04(0.01)	0.05(0.01)

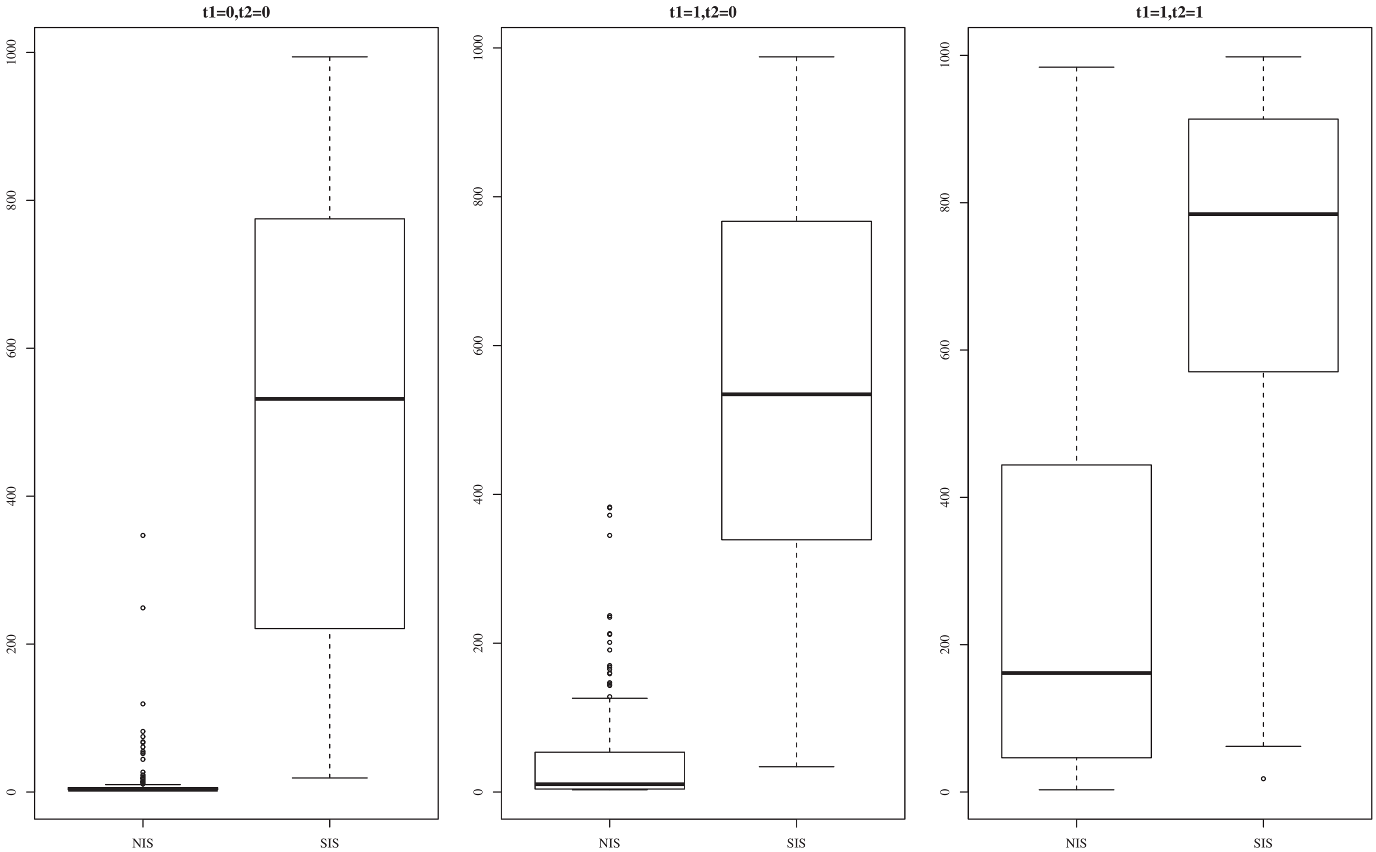

Figure 2. Boxplots of minimum model size (left: NIS, right: SIS) for Example 2 under different correlation settings.

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.

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 article.

Table 2. Average values of the number of true positives (TPs), false positives (FPs), 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 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:

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

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 FPs are fairly small, hence they do not affect PE very much. Regarding computational 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

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

	Correlation		Conditional-INIS			Greedy-INIS		
Model	X's	X's–W	TP	FP	PE	TP	FP	PE
$t_1 = 0, t_2 = 0$	0	0	4.00	1.57	1.34	4.00	0.06	1.17
(SNR $\approx$ 16.85)			(0.00)	(1.49)	(0.14)	(0.00)	(0.00)	(0.06)
$t_1 = 2, t_2 = 0$	0.25	0	4.00	0.15	0.89	4.00	0.00	0.99
(SNR $\approx$ 3.66)			(0.00)	(0.00)	(0.05)	(0.00)	(0.00)	(0.05)
$t_1 = 2, t_2 = 1$	0.25	0.36	4.00	0.12	1.24	3.99	0.02	1.32
(SNR $\approx$ 3.21)			(0.00)	(0.00)	(0.10)	(0.00)	(0.00)	(0.13)
$t_1 = 3, t_2 = 0$	0.43	0	4.00	0.01	1.13	3.98	0.00	1.19
(SNR $\approx$ 3.32)			(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.06)
$t_1 = 3, t_2 = 1$	0.43	0.46	3.97	0.05	1.31	3.96	0.03	1.17
(SNR $\approx$ 2.81)			(0.00)	(0.00)	(0.15)	(0.00)	(0.00)	(0.10)

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),

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

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 packages *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, that is, the geographical accessibility to employment. Interestingly, conditional-INIS (with $L_n = 7$ and $K = 5$) selects the following submodel:

$$\begin{aligned}\log(\mathrm{MV}) = \beta_0(W) &+ \beta_1(W)\cdot\mathrm{RM}^2 + \beta_2(W)\cdot\mathrm{AGE}\\ &+ \beta_5(W)\cdot\mathrm{TAX} + \beta_7(W)\cdot(\mathrm{B}-0.63)^2 + \beta_9(W)\\ &\cdot\mathrm{CRIM} + \epsilon,\end{aligned} \tag{15}$$

where $W = \log(\mathrm{DIS})$. 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 nonlinear interactions with the accessibility are 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.

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

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

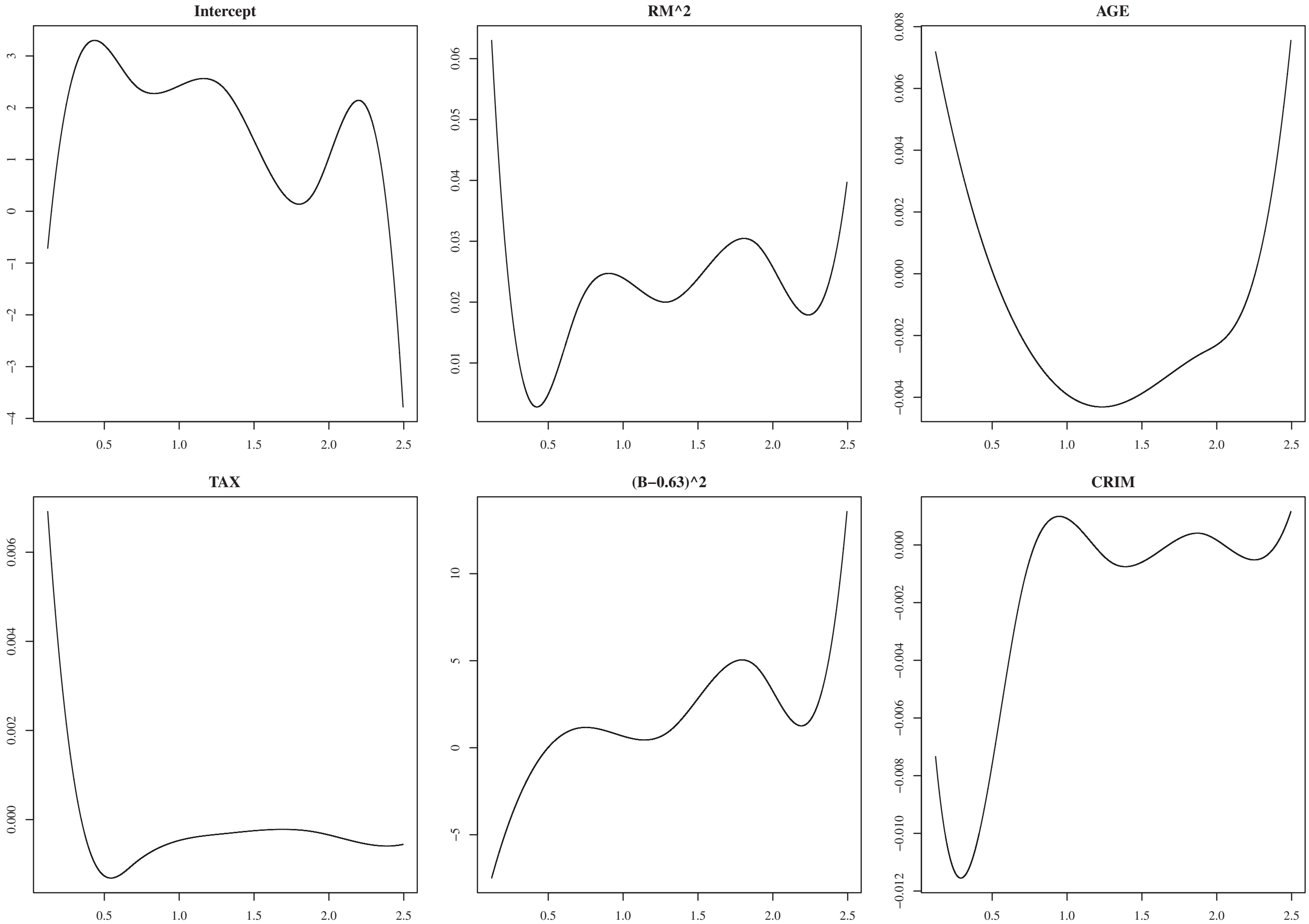

Figure 3. Fitted functional estimates $\widehat{\beta}_j(W)$'s selected by conditional-INIS.

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

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

Note that $\{W, X_1, \ldots, X_s\}$ are the independent variables in original dataset ($s = 13$ 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 $\{W, X_1, \ldots, X_s\}$ directly using SCAD penalty without screening procedure. We repeat $N = 100$ times and report the average PE 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 FPs. 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 PE 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.

Table 5. Average values of prediction error (PE), model size, and the number of selected noise variables (SNVs) 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

APPENDIX A: PROOFS

A.1 Properties of B-Splines

Our estimation use the B-spline basis, which has the following properties (de Boor 1978): for each $j = 1, \ldots, p$ and $k = 1, \ldots, L_n$, $B_k(W) \geq 0$ 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 $\eta_k \in \mathbb{R}$, $k = 1, \ldots, L_n$,

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

Then under condition (iii), for $C_1 = T_1T_3$ and $C_2 = T_2T_4$,

$$C_1L_n^{-1} \le \mathrm{E}\big[B_k^2(W)\big] \le C_2L_n^{-1}, \quad \text{for all } k = 1, \dots, L_n. \quad \text{(A.2)}$$

Furthermore, under condition (iii), it follows from (A.1) that for any $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{L_n})^T \in \mathbb{R}^{L_n}$ such that $\|\boldsymbol{\eta}\|_2^2 = 1$, $C_1L_n^{-1} \le \boldsymbol{\eta}^T\mathrm{E}[\mathbf{B}^T\mathbf{B}]\boldsymbol{\eta} \le C_2L_n^{-1}$. Or equivalently,

$$C_1L_n^{-1} \le \lambda_{\min}(\mathrm{E}[\mathbf{B}^T\mathbf{B}]) \le \lambda_{\max}(\mathrm{E}[\mathbf{B}^T\mathbf{B}]) \le C_2L_n^{-1}. \quad \text{(A.3)}$$

A.2 Technical Lemmas

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

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

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

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

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

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

Lemma A.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 $r_1r_2/(r_1 + r_2) \ge 1$, and

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

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

$$\mathrm{P}(|Z_1Z_2| > t|W) \le \exp(1 - (t/K^*)^{r^*})$$

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 $r = r_1r_2/(r_1 + r_2)$. Then uniformly on $\mathcal{W}$, we have

$$\mathrm{P}(|Z_1Z_2| > t|W) \le \mathrm{P}(M|Z_1| > t|W) + \mathrm{P}(|Z_2| > M|W)$$
$$\le \exp\{1 - (t/K_1M)^{r_1}\} + \exp\{1 - (M/K_2)^{r_2}\}$$
$$= 2\exp\{1 - (t/K_1K_2)^r\}.$$
□

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 $t > K^*$,

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

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

Lemma A.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 $\mathrm{E}[|Y_i|^m] \le m!M^{m-2}\nu_i/2$ 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 $\nu \ge \nu_1 + \cdots + \nu_n$.

Lemma A.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 $\nu \ge \mathrm{var}(Y_1 + \cdots + Y_n)$, then

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

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

Lemma A.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 $j = 1, \dots, p$, $k = 1, \dots, L_n$,

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

and

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

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

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

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

$$\mathrm{E}|Z_{jki}|^m \le 2^m\mathrm{E}|X_{ji}B_k(W_i)m(\mathbf{X}_i^*)|^m$$
$$\le 2^m\mathrm{E}\big[B_{jk}^2(W_i)eK_4^m m!\big] \le m!(2K_4)^{m-2}\big(8eK_4^2C_2L_n^{-1}\big)/2, \quad \text{(A.4)}$$

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

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n Z_{jki}\right| \ge \frac{\delta}{2n}\right) \le 2\exp\left\{-\frac{\delta^2}{64eK_4^2C_2L_n^{-1}n + 8K_4\delta}\right\}. \quad \text{(A.5)}$$

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

$$\mathrm{E}|\xi_i|^m = \mathrm{E}\big[B_k^m(W_i)\mathrm{E}[|X_{ji}\varepsilon_i|^m|W_i]\big] \le m!K_5^{m-2}\big(2eK_5^2C_2L_n^{-1}\big)/2.$$

Thus, according to Lemma A.3,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n \xi_i\right| \ge \frac{\delta}{2n}\right) \le 2\exp\left\{-\frac{\delta^2}{16eK_5^2C_2L_n^{-1}n + 4K_5\delta}\right\}. \quad \text{(A.6)}$$

Similarly, we can show that

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n B_k(W_i)m(\mathbf{X}_i^*) - \mathrm{E}[B_k(W)m(\mathbf{X}^*)]\right| \ge \frac{\delta}{2n}\right)$$
$$\le 2\exp\left\{-\frac{\delta^2}{64eK_2^2C_2L_n^{-1}n + 8K_2\delta}\right\} \quad \text{(A.7)}$$

and

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}B_k(W_i)\varepsilon_i\right| \geq \frac{\delta}{2n}\right) \leq 2\exp\left\{-\frac{\delta^2}{16eK_3^2C_2L_n^{-1}n+4K_3\delta}\right\}. \tag{A.8}$$

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 (A.5)–(A.8) by union bound of probability yields the desired result.

Lemma A.6. Under conditions (i), (iii), and (v), there exist positive constants C_3 and C_4, such that for $j = 1, \ldots, p$ and $\mathbf{\Sigma}_j = \mathrm{E}[\mathbf{Q}_j^T\mathbf{Q}_j]$,

$$C_3L_n^{-1} \leq \lambda_{\min}(\mathbf{\Sigma}_j) \leq \lambda_{\max}(\mathbf{\Sigma}_j) \leq C_4L_n^{-1}. \tag{A.9}$$

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 $\|\boldsymbol{\eta}\|_2^2 = 1$,

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

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 A.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

$$\begin{aligned}&\frac{h_1}{1+4eK_1^2}\mathrm{E}[(\mathbf{B}\boldsymbol{\eta}_1)^2 + (\mathbf{B}\boldsymbol{\eta}_2)^2] \\ &\quad \leq \boldsymbol{\eta}^T\mathbf{\Sigma}_j\boldsymbol{\eta} \leq \left(1+4eK_1^2\right)\mathrm{E}[(\mathbf{B}\boldsymbol{\eta}_1)^2 + (\mathbf{B}\boldsymbol{\eta}_2)^2].\end{aligned}$$

By (A.3), we have

$$\frac{h_1C_1}{1+4eK_1^2}L_n^{-1} \leq \boldsymbol{\eta}^T\mathbf{\Sigma}_j\boldsymbol{\eta} \leq \left(1+4eK_1^2\right)C_2L_n^{-1}.$$

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

$$P(\|\mathbf{\Sigma}_{nj} - \mathbf{\Sigma}_j\| \geq L_n\delta/n) \leq 6L_n^2\exp\left\{-\frac{\delta^2}{b_3L_n^{-1}n+b_4\delta}\right\},$$

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

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

$$P(|\|(\mathbf{\Sigma}_{nj})^{-1}\| - \|(\mathbf{\Sigma}_j)^{-1}\|| \geq b_5\|(\mathbf{\Sigma}_j)^{-1}\|) \leq 6L_n^2\exp\left\{-b_6L_n^{-3}n\right\},$$

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

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

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

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

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

$$\begin{aligned}\|\mathbf{\Sigma}_{nj} - \mathbf{\Sigma}_j\| &\leq 2L_n\|\mathbf{\Sigma}_{nj} - \mathbf{\Sigma}_j\|_\infty \\ &= 2L_n\max(\|\mathbf{D}_1\|_\infty, \|\mathbf{D}_{2j}\|_\infty, \|\mathbf{D}_{3j}\|_\infty).\end{aligned} \tag{A.10}$$

$\square$

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 (A.2),

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

By Lemma A.4, we have

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

It then follows from the union bound of probability that

$$P\left(\|\mathbf{D}_1\|_\infty \geq \delta/6n\right) \leq 2L_n^2\exp\left\{-\delta^2/\left(72C_2L_n^{-1}n+24\delta\right)\right\}. \tag{A.11}$$

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

$$\begin{aligned}&\mathrm{E}[|X_{ji}B_k(W_i)B_l(W_i) - \mathrm{E}[X_jB_k(W)B_l(W)]|^m] \\ &\quad \leq 2^m\mathrm{E}[|X_{ji}B_k(W_i)B_l(W_i)|^m] \leq m!(2K_1)^{m-2}\left(8eK_1^2C_2L_n^{-1}\right)/2,\end{aligned}$$

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

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

It then follows from the union bound of probability that

$$P(\|\mathbf{D}_{2j}\|_\infty \geq \delta/6n) \leq 2L_n^2\exp\left\{-\delta^2/\left(576eK_1^2C_2L_n^{-1}n+24K_1\delta\right)\right\}. \tag{A.12}$$

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

$$\begin{aligned}&\mathrm{E}\left[\left|X_{ji}^2B_k(W_i)B_l(W_i) - \mathrm{E}\left[X_j^2B_k(W)B_l(W)\right]\right|^m\right] \\ &\quad \leq m!(2K_6)^{m-2}\left(8eK_6^2C_2L_n^{-1}\right)/2.\end{aligned}$$

By Lemma A.3, we have

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

It then follows from the union bound of probability that

$$P(\|\mathbf{D}_{3j}\|_\infty \geq \delta/6n) \leq 2L_n^2\exp\left\{-\delta^2/\left(576eK_6^2C_2L_n^{-1}n+24K_6\delta\right)\right\}. \tag{A.13}$$

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 (A.10)–(A.13) we have

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

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

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

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),

$$|\lambda_{\min}(\mathbf{A}) - \lambda_{\min}(\mathbf{B})| \le \max\{|\lambda_{\min}(\mathbf{A}-\mathbf{B})|, |\lambda_{\min}(\mathbf{B}-\mathbf{A})|\}. \quad (A.14)$$

It then follows from (A.14) that

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

which implies that

$$\begin{aligned} &P(|\lambda_{\min}(\boldsymbol{\Sigma}_{nj}) - \lambda_{\min}(\boldsymbol{\Sigma}_j)| \ge L_n\delta/n) \\ &\quad\le 6L_n^2\exp\{-\delta^2/(b_3L_n^{-1}n + b_4\delta)\}. \end{aligned} \quad (A.15)$$

Let $\delta = b_9C_3L_n^{-2}n$ in (A.15) for $b_9 \in (0,1)$. According to (A.9), we have

$$P(|\lambda_{\min}(\boldsymbol{\Sigma}_{nj}) - \lambda_{\min}(\boldsymbol{\Sigma}_j)| \ge b_9\lambda_{\min}(\boldsymbol{\Sigma}_j)) \le 6L_n^2\exp(-b_6L_n^{-3}n), \quad (A.16)$$

for some positive constant b_6. Next observe the fact that for $x, y > 0$, $a \in (0,1)$ and $b = 1/(1-a) - 1$, $|x^{-1} - y^{-1}| \ge by^{-1}$, which implies $|x - y| \ge ay$. This is because $x^{-1} - y^{-1} \ge by^{-1}$ is equivalent to $x^{-1} \ge \frac{1}{1-a}y^{-1}$, or $x - y \le -ay$; on the other hand, $x^{-1} - y^{-1} \le by^{-1}$ implies $x^{-1} \le (1 - \frac{a}{1-a})y^{-1} \le (1 - \frac{a}{1+a})y^{-1}$ as $a \in (0,1)$, and therefore $x - y \ge ay$. Then let $b_5 = 1/(1-b_9) - 1$, it follows from (A.16) that

$$\begin{aligned} &P(|(\lambda_{\min}(\boldsymbol{\Sigma}_{nj}))^{-1} - (\lambda_{\min}(\boldsymbol{\Sigma}_j)^{-1}| \ge b_5(\lambda_{\min}(\boldsymbol{\Sigma}_j))^{-1}) \\ &\quad\le 6L_n^2\exp(-b_6L_n^{-3}n). \end{aligned}$$

Following the same proof, by (A.3) we also have for any positive constant b_7, there exists some positive constant b_8, such that

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

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 $\|a_j - a_j^*\|_\infty \le M_2L_n^{-d}$ and $\|b_j - b_j^*\|_\infty \le M_2L_n^{-d}$, where $\mathcal{S}_n$ is the space of polynomial splines of degree $l \ge 1$ with normalized B-spline basis $\{B_k, k = 1, \ldots, L_n\}$, 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

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

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

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

□

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

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

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

$$\begin{aligned} &\|a_j + b_jX_j\|^2 = \|\tilde{a}_j + \tilde{b}_jX_j\|^2 + \|\tilde{a}_j + \tilde{b}_jX_j - a_j - b_jX_j\|^2, \\ &\|a_j + b_jX_j\|^2 - \|\tilde{a}_j + \tilde{b}_jX_j\|^2 \le 2M_2^2L_n^{-2d}\left(1 + \mathrm{E}\left[X_j^2\right]\right). \end{aligned} \quad (A.17)$$

Similarly, we have

$$\|a_0\|^2 - \|\tilde{a}_0\|^2 \le M_2^2L_n^{-2d}. \quad (A.18)$$

By taking $M_1 = M_2^2(8eK^2 + 3)$ (see Lemma A.1), the first part of Proposition 1 follows from (A.17) and (A.18):

$$\begin{aligned} u_j - \tilde{u}_j &= \|a_j + b_jX_j\|^2 - \|a_0\|^2 - (\|\tilde{a}_j + \tilde{b}_jX_j\|^2 - \|\tilde{a}_0\|^2) \\ &\le M_1L_n^{-2d}. \end{aligned} \quad (A.19)$$

By (11) and (A.19), we have $\min_{j\in\mathcal{M}_*}\tilde{u}_j \ge c_1L_nn^{-2\kappa}/h_2 - M_1L_n^{-2d}$. The desired result follows from $L_n^{-2d-1} \le c_1(1/h_2 - \xi)n^{-2\kappa}/M_1$ for some $\xi \in (0, 1/h_2)$.

Proof of Theorem 1. We first prove part (1). Note that $\widehat{u}_{nj} - \tilde{u}_j = S_1 + S_2$, where

$$\begin{aligned} S_1 &= \frac{1}{n^2}\mathbf{Y}^T\mathbf{Q}_{nj}\boldsymbol{\Sigma}_{nj}^{-1}\mathbf{Q}_{nj}^T\mathbf{Y} - \mathrm{E}[Y\mathbf{Q}_j]\boldsymbol{\Sigma}_j^{-1}\mathrm{E}\left[\mathbf{Q}_j^TY\right], \quad \text{and} \\ S_2 &= \frac{1}{n}\mathbf{Y}^T\mathbf{B}_n\left(\mathbf{B}_n^T\mathbf{B}_n\right)^{-1}\mathbf{B}_n^T\mathbf{Y} - \mathrm{E}[Y\mathbf{B}](\mathrm{E}[\mathbf{B}^T\mathbf{B}])^{-1}\mathrm{E}[\mathbf{B}^TY]. \end{aligned}$$

□

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

$$S_1 = (\mathbf{a}_n - \mathbf{a})^T\boldsymbol{\Sigma}_{nj}^{-1}(\mathbf{a}_n - \mathbf{a}) + 2(\mathbf{a}_n - \mathbf{a})^T\boldsymbol{\Sigma}_{nj}^{-1}\mathbf{a} + \mathbf{a}^T(\boldsymbol{\Sigma}_{nj}^{-1} - \boldsymbol{\Sigma}_j^{-1})\mathbf{a}.$$

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

We first deal with S_{11}. Note that

$$|S_{11}| \le \|\boldsymbol{\Sigma}_{nj}^{-1}\| \cdot \|\mathbf{a}_n - \mathbf{a}\|_2^2. \quad (A.20)$$

By Lemma A.5 and the union bound of probability,

$$P\left(\|\mathbf{a}_n - \mathbf{a}\|_2^2 \ge 2L_n\delta^2/n^2\right) \le 8L_n\exp\{-\delta^2/(b_1L_n^{-1}n + b_2\delta)\}. \quad (A.21)$$

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

$$P\left(\left|\left\|\boldsymbol{\Sigma}_{nj}^{-1}\right\| - \left\|\boldsymbol{\Sigma}_j^{-1}\right\|\right| \ge b_5\left\|\boldsymbol{\Sigma}_j^{-1}\right\|\right) \le 6L_n^2\exp\{-b_6L_n^{-3}n\}.$$

Then it follows from (A.9) that

$$P\left(\left\|\boldsymbol{\Sigma}_{nj}^{-1}\right\| \ge (b_5 + 1)C_3^{-1}L_n\right) \le 6L_n^2\exp\{-b_6L_n^{-3}n\}. \quad (A.22)$$

Combining (A.20)–(A.22) and based on the union bound of probability, we have

$$\begin{aligned} &P\left(|S_{11}| \ge 2(b_5 + 1)C_3^{-1}L_n^2\delta^2/n^2\right) \\ &\quad\le 8L_n\exp\{-\delta^2/(b_1L_n^{-1}n + b_2\delta)\} + 6L_n^2\exp\{-b_6L_n^{-3}n\}. \end{aligned} \quad (A.23)$$

We next bound S_{12}. Note that

$$|S_{12}| \le 2\|\mathbf{a}_n - \mathbf{a}\|_2 \cdot \left\|\boldsymbol{\Sigma}_{nj}^{-1}\right\| \cdot \|\mathbf{a}\|_2. \quad (A.24)$$

By Lemma A.1,

$$\begin{aligned} \|\mathbf{a}\|_2^2 &= \|\mathrm{E}[\mathbf{B}^TY]\|_2^2 + \|\mathrm{E}[X_j\mathbf{B}^TY]\|_2^2 \\ &\le \sum_{k=1}^{L_n}\left(\mathrm{E}\left[B_k^2m^2(\mathbf{X}^*)\right] + \mathrm{E}\left[B_k^2X_j^2m^2(\mathbf{X}^*)\right]\right) \\ &\le 4eC_2\left(K_2^2 + K_4^2\right), \end{aligned} \quad (A.25)$$

where the calculation as in (A.4) was used.

It follows from (A.21), (A.22), (A.24), (A.25), and the union bound of probability that

$$P\big(|S_{12}| \ge 4\sqrt{2}(b_5+1)e^{1/2}C_2^{1/2}\big(K_2^2+K_4^2\big)^{1/2}C_3^{-1}L_n^{3/2}\delta/n\big) \le 8L_n \exp\{-\delta^2/\big(b_1L_n^{-1}n+b_2\delta\big)\} + 6L_n^2\exp\{-b_6L_n^{-3}n\}. \tag{A.26}$$

To bound S_{13}, note that

$$|S_{13}| = \mathbf{a}^T\mathbf{\Sigma}_{nj}^{-1}(\mathbf{\Sigma}_j - \mathbf{\Sigma}_{nj})\mathbf{\Sigma}_j^{-1}\mathbf{a} \le \|\mathbf{\Sigma}_{nj}^{-1}\|^2 \cdot \|\mathbf{\Sigma}_j - \mathbf{\Sigma}_{nj}\| \cdot \|\mathbf{a}\|_2^2. \tag{A.27}$$

Then it follows from Lemmas A.6, A.7, (A.22), (A.25), (A.27), and the union bound of probability that there exist b_3, b_4, and b_6 such that

$$P\big(|S_{13}| \ge 4eC_2\big(K_2^2+K_4^2\big)(b_5+1)^2C_3^{-2}L_n^3\delta/n\big) \le 6L_n^2\exp\{-\delta^2/\big(b_3L_n^{-1}n+b_4\delta\big)\} + 6L_n^2\exp\{-b_6L_n^{-3}n\}. \tag{A.28}$$

Hence, combining (A.23), (A.26), and (A.28), there exist some positive constants s_1, s_2, and s_3 such that

$$P\left(|S_1| \ge s_1L_n^2\delta^2/n^2 + s_2L_n^{3/2}\delta/n + s_3L_n^3\delta/n\right) \le 16L_n\exp\{-\delta^2/\big(b_1L_n^{-1}n+b_2\delta\big)\} + 6L_n^2\exp\{-\delta^2/\big(b_3L_n^{-1}n+b_4\delta\big)\} + 18L_n^2\exp\{-b_6L_n^{-3}n\}.$$

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

$$P\left(|S_2| \ge s_4L_n^2\delta^2/n^2 + s_5L_n^{3/2}\delta/n + s_6L_n^3\delta/n\right) \le 8L_n\exp\{-\delta^2/\big(b_1L_n^{-1}n+b_2\delta\big)\} + 6L_n^2\exp\{-\delta^2/\big(b_3L_n^{-1}n+b_4\delta\big)\} + 18L_n^2\exp\{-b_8L_n^{-3}n\}.$$

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

$$P\left(|\widehat{u}_{nj} - \tilde{u}_j| \ge c_2L_nn^{-2\kappa}\right) \le \big(24L_n + 12L_n^2\big)\exp\{-c_3n^{1-4\kappa}L_n^{-3}\} + 36L_n^2\exp\{-c_4L_n^{-3}n\}.$$

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

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

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

by Proposition 1, we have $\widehat{u}_{nj} \ge c_1\xi L_nn^{-2\kappa}/2$, for all $j \in \mathcal{M}_*$. 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

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

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 $\mathbf{Q} = (\mathbf{Q}_1, \ldots, \mathbf{Q}_p)$ 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

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

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

$$\|\mathrm{E}[\mathbf{Q}^TY]\|_2^2 = O(\lambda_{\max}(\mathbf{\Sigma})). \tag{A.29}$$

□

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

$$\sum_{j=1}^p \tilde{u}_j \le \max_{1\le j\le p}\lambda_{\max}\{(\mathbf{\Sigma}_j)^{-1}\}\sum_{j=1}^p \big\|\mathrm{E}\big[\mathbf{Q}_j^TY\big]\big\|_2^2 = \max_{1\le j\le p}\lambda_{\max}\{(\mathbf{\Sigma}_j)^{-1}\}\|\mathrm{E}[\mathbf{Q}^TY]\|_2^2.$$

By Lemma A.6 and (A.29), the last term is of order $O(L_n\lambda_{\max}(\mathbf{\Sigma}))$. This implies that the number of $\{j : \tilde{u}_j > \delta L_nn^{-2\kappa}\}$ cannot exceed $O(n^{2\kappa}\lambda_{\max}(\mathbf{\Sigma}))$ for any $\delta > 0$.

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

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

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

[Received March 2013. Revised October 2013.]

REFERENCES

Antoniadis, A., and Fan, J. (2001), "Regularized Wavelet Approximations" (with discussion), *Journal American Statistical Association*, 96, 939–967. [1274,1275]

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When *p* is Much Larger than *n*" (with discussion), *The Annals of Statistics*, 35, 2313–2404. [1270]

de Boor, C. (1978), *A Practical Guide to Splines*, New York: Springer-Verlag. [1279]

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models," *Journal of the American Statistical Association*, 106, 544–557. [1270,1271,1273,1274,1282]

Fan, J., and Li, R. (2001), "Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1270,1275]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society,* Series B, 70, 849–911. [1270,1271,1274]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [1270,1275]

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29, 153–193. [1270]

Fan, J., and Zhang, W. (2008), "Statistical Methods With Varying Coefficient Models," *Statistics and its Interface*, 1, 179–195. [1270]

Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools" (with discussion), *Technometrics*, 35, 109–148. [1270]

Hall, P., and Miller, H. (2009), "Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems," *Journal of Computational and Graphical Statistics*, 18, 533–550. [1270]

Hall, P., Titterington, D. M., and Xue, J. H. (2009), "Tilting Methods for Assessing the Influence of Components in a Classifier," *Journal of the Royal Statistical Society,* Series B, 71, 783–803. [1270]

Harrison, D., and Rubinfeld, D. (1978), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81–102. [1277]

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, 40, 1846–1877. [1270]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening Via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [1270]

Lian, H. (2011), "Flexible Shrinkage Estimation in High-Dimensional Varying Coefficient Models," arXiv preprint arXiv:1008.2271. [1271]

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1040–1053. [1273,1282]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [1270]

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer. [1280]

Wang, L., Li, H., and Huang, J. Z. (2008), "Variable Selection in Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements," *Journal of the American Statistical Association*, 103, 1556–1569. [1271,1273,1274]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society,* Series B, 68, 49–67. [1274,1275]

Zhang, C.-H (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [1270]

Zhao, S. D., and Li, Y. (2012), "Principled Sure Independence Screening for Cox Models With Ultra-High-Dimensional Covariates," *Journal of Multivariate Analysis*, 105, 397–411. [1270,1274]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1270]

Zou, H., and Hastie, T. (2005), "Addendum: Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society,* Series B, 67, 301–320. [1270]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, 36, 1509–1533. [1270]