Journal of the American Statistical Association

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.

To cite this article: Jianqing Fan, Yunbei Ma & Wei Dai (2014) Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Varying Coefficient Models, Journal of the American Statistical Association, 109:507, 1270-1284, DOI: 10.1080/01621459.2013.879828

To link to this article: http://dx.doi.org/10.1080/01621459.2013.879828

PLEASE SCROLL DOWN FOR ARTICLE
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-tailoring 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,

\[
Y = \beta^T(W)X + \epsilon, \tag{1}
\]

where \( Y \) is the response, \( W \) is some univariate observable exposure variable, \( X = (X_1, \ldots, X_p)^T \) is the vector of covariates, and \( \epsilon \) 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 \( X \) 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, 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...
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 dimen-
sional setting, Lian (2011) used the adaptive group lasso penalty
to address the problem of variable selection. In the NP dimen-
sional settings, Fan, Ma, and Dai (2011) used the adaptive group lasso penalty
applied the basis function approximations and the SCAD penalty
to address the problem of variable selection. In the NP dimen-
sional setting, Lian (2011) used the adaptive group lasso penalty
applied the basis function approximations and the SCAD penalty
to address the problem of variable selection. In the NP dimen-
sional setting, Lian (2011) used the adaptive group lasso penalty
applied the basis function approximations and the SCAD penalty
under some reason-

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 func-
tional coefficient vector \( \beta(\cdot) = (\beta_1(\cdot), \ldots, \beta_p(\cdot))^T \) is sparse. Let
\( \mathcal{M}_s = \{ j : \mathbb{E}[\beta_j^2(W)] > 0 \} \) be the true sparse model with non-
sparcity size \( s_n = |\mathcal{M}_s| \). We allow \( p \) to grow with \( n \) and denote by \( p_n \)
whenever necessary.

2.1 Marginal Regression
For \( j = 1, \ldots, p \), let \( a_j(W) \) and \( b_j(W) \) be the minimizer of the follow-
ing marginal regression problem:
\[
\min_{a_j(W), b_j(W) \in L_2(\mathcal{P})} \mathbb{E}[(Y - a_j(W) - b_j(W)X_j)^2 | W],
\]
where \( P \) denotes the joint distribution of \((Y, W, X)\) and \( L_2(\mathcal{P}) \)
is the class of square integrable functions under the measure \( P \).

Let \( a_0(W) = \mathbb{E}[Y|W] \), \( b_0(W) = \mathbb{E}[Y|W] - b_j(W)\mathbb{E}[X_j|W] \),
where \( \mathbb{E}[X_j|W] = \mathbb{E}[X_j|Y,W,W] \). We rank the marginal utility of covariates
by
\[
u_j = \mathbb{E}[b_j(W)X_j - \mathbb{E}[X_j|W]^2] \approx \mathbb{E}\left( \frac{\text{cov}[X_j, Y|W]}{\text{var}[X_j|W]} \right).
\]

For each \( j = 1, \ldots, p \), 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 \( u_j, j = 1, \ldots, p \),
we approximate \( a_j(W) \) and \( b_j(W) \) by functions in \( \mathcal{S}_n \), the space of
polynomial splines of degree \( d \geq 1 \) on \( \mathcal{V} \), a compact set. Let \( \{B_k, k = 1, \ldots, L_n\} \)
denote its normalized B-spline basis, where \( L_n \) is the number of basis functions. Note that \( \| B_k \| \leq 1 \),
where \( \| \cdot \|_\infty \) is the sup norm. Then
\[
a_j(W) \approx \sum_{k=1}^{L_n} \eta_jk B_k(W), \quad j = 0, \ldots, p,
\]
\[
b_j(W) \approx \sum_{k=1}^{L_n} \theta_jk B_k(W), \quad j = 1, \ldots, p,
\]
where \( \{\eta_jk\}_{k=1}^{L_n} \) and \( \{\theta_jk\}_{k=1}^{L_n} \) are scalar coefficients.

We now consider the following sample version of the marginal
regression problem:
\[
\min_{\eta, \theta \in \mathbb{R}^{L_n}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - B(W_i)\eta_j - B(W_i)\theta_j X_{ji})^2,
\]
where \( \eta_j = (\eta_{j1}, \ldots, \eta_{jL_n})^T, \theta_j = (\theta_{j1}, \ldots, \theta_{jL_n})^T \),
and \( B(\cdot) = (B_1(\cdot), \ldots, B_{L_n}(\cdot)) \).
It is easy to show that the minimizers of (5) are given by
\[
(\hat{\eta}_j^T, \hat{\theta}_j^T)^T = (Q_{nj}^T Q_{nj})^{-1} Q_{nj}^T Y,
\]
where
\[
Q_{nj} = (B_n, \Phi_{nj}) = \begin{pmatrix}
B(W_1), & X_{j}B(W_1) \\
\vdots & \vdots \\
B(W_n), & X_{jn}B(W_n)
\end{pmatrix}
\]
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
\[
\hat{a}_{nj}(W) = B(W)\hat{\eta}_j = (B(W), 0_{L_n}^T)(Q_{nj}^T Q_{nj})^{-1} Q_{nj}^T Y,
\]
\[
\hat{b}_{nj}(W) = B(W)\hat{\theta}_j = (0_{L_n}^T, B(W))(Q_{nj}^T Q_{nj})^{-1} Q_{nj}^T Y,
\]
(6)
where \(0_{L_n}\) is an \(L_n\)-dimension vector with all entries 0. Similarly, we have the estimate of the intercept function \(a_0\) by
\[
\hat{a}_{0}(W) = B(W)\hat{\eta}_0 = B(W)(B_n^T B_n)^{-1} B_n^T Y,
\]
(7)
where
\[
\hat{\eta}_0 = \arg\min_{\eta_0 \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n (Y_i - B(W_i)\eta_0)^2.
\]
We now define an estimate of the marginal utility \(u_j\) as
\[
\hat{u}_{nj} = \|\hat{a}_{nj}(W) + \hat{b}_{nj}(W)X_j\|_n^2 - \|\hat{a}_0(W)\|_n^2
\]
\[
= \frac{1}{n} \sum_{i=1}^n (\hat{a}_{nj}(W_i) + \hat{b}_{nj}(W_i)X_{ji})^2 - \frac{1}{n} \sum_{i=1}^n (\hat{a}_0(W_i))^2,
\]
where \(W = (W_1, \ldots, W_n)^T\). Note that throughout this article, whenever two vectors \(a\) and \(b\) are of the same length, \(ab\) denotes the componentwise product. Given a predefined threshold value \(\tau_n\), 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
\[
\hat{v}_{nj} = \|Y - \hat{a}_{nj}(W) - \hat{b}_{nj}(W)X_j\|_n^2,
\]
and we select variables as follows:
\[
\mathcal{M}_{\tau_n} = \{1 \leq j \leq p : \hat{v}_{nj} \geq v_n\},
\]
where \(v_n\) 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
\[
\|\hat{a}_{nj}(W) + \hat{b}_{nj}(W)X_j\|_n^2 = \frac{1}{n} Y^T Q_{nj} (Q_{nj}^T Q_{nj})^{-1} Q_{nj}^T Y,
\]
(8)
and
\[
\frac{1}{n} \sum_{i=1}^n Y_i (\hat{a}_{nj}(W_i) + \hat{b}_{nj}(W_i)X_{ji})^2
\]
\[
= \frac{1}{n} Y^T Q_{nj} (Q_{nj}^T Q_{nj})^{-1} Q_{nj}^T Y.
\]
(9)
It follows from (8) and (9) that
\[
\hat{v}_{nj} = \|Y\|_n^2 - \|\hat{a}_0(W)\|_n^2 - \hat{u}_{nj},
\]
(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 \(B(W)\) and write \(B\) whenever the context is clear. Let \(\hat{a}_j(W) = B\hat{\eta}_j\) and \(\hat{b}_j(W) = B\hat{\theta}_j\), where \(\hat{\eta}_j\) and \(\hat{\theta}_j\) are the minimizer of
\[
\min_{\eta_j, \theta_j \in \mathbb{R}^{L_n}} E[(Y - B\eta_j - B\theta_j X_j)^2],
\]
and \(\hat{a}_0(W) = B\hat{\eta}_0\), where \(\hat{\eta}_0\) is the minimizer of
\[
\min_{\eta \in \mathbb{R}^n} E[(Y - B\eta)^2].
\]
It can be seen that
\[
(\hat{a}_j(W), \hat{b}_j(W))^T = \text{diag}(B, B)(E[Q_j^T Q_j])^{-1}E[Q_j^T Y],
\]
\[
\hat{a}_0(W) = B(E[B^T B])^{-1}E[B^T Y],
\]
and \(Q_j = (B, X_jB)\). Then we can define
\[
\hat{u}_j = ||\hat{a}_j(W) + \hat{b}_j(W)X_j||^2 - ||\hat{a}_0(W)||^2
\]
\[
= E[Y Q_j E[Q_j^T Q_j]]^{-1}E[Q_j^T Y]
\]
\[
- E[Y B][E[B^T B]]^{-1}E[B^T Y].
\]

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:
\[
u_j = E \left[ \frac{(\text{cov}[X_j, Y|W])^2}{\text{var}[X_j|W]} \right], \quad j = 1, \ldots, p.
\]
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 \(j = 1, \ldots, p\), \(\text{var}[X_j|W]\) is uniformly bounded away from 0 and infinity on \(W\), where \(W\) 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) \(\min_{j \in \mathcal{M}} E[\text{cov}[X_j, Y|W]^2] \geq c_1 L_n^{-2\kappa}, \text{ for some } \kappa > 0 \text{ and } c_1 > 0\).

Then under conditions (i) and (ii),
\[
\min_{j \in \mathcal{M}} u_j \geq c_1 L_n^{-2\kappa}/h_2.
\]
(11)
Note that in condition (ii), the number of basis functions \(\lambda_n\) is not intrinsic. By the Remark 1, \(\lambda_n\) should be chosen in correspondence to the smoothness condition of the nonparametric component. Therefore, condition (ii) depends only on \(\kappa\) and smoothness parameter \(d\) in condition (iv). We keep \(\lambda_n\) 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 $\alpha$. That is,

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

for some positive constant $M$, where $r$ is a non-negative 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

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

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

$$P(\{|X_j| > t | W\}) \leq \exp(1 - (t/K_2)^r_2),$$

uniformly on $W$, 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 $r_1 r_3/(r_1 + r_3) \geq 1$, such that

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

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

(vii) There exists some positive 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 $\varepsilon$, 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

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

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

$$\min_{j \notin M_*} \bar{u}_j \geq c_1 \xi L_n^{-2\kappa}. \quad (12)$$

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

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

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

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

(ii) If condition (vii) also holds, then by taking $\tau_n = c_5 L_n n^{-2\kappa}$, we have existence positive constants $c_6$ and $c_7$ such that

$$P \left( M_* \subseteq \widehat{M}_n \right) \geq 1 - 12 \tau_n L_n \left\{ (2 + L_n) \exp \left( -c_6 n^{1-4\kappa} L_n^{-3} \right) + 3L_n \exp \left( -c_7 L_n^{-3} n \right) \right\}.$$

**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 $L_n$, the higher dimensionality we can handle. On the other hand, Remark 1 points out that if $L_n \geq Cn \kappa/(2d+1)$ 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 $L_n = O(n^{\kappa/(2d+1)})$, which is the optimal convergence rate for nonparametric regression (Stone 1982). In this case, the allowable dimensionality can be as high as $p = o(\exp(n^{\kappa/(2d+1)}))$.

### 3.2 False Selection Rates

According to (12), the ideal case for vanishing FP rate is when

$$\max_{j \notin M_*} \bar{u}_j = o(L_n n^{-2\kappa})$$

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

$$\max_{j \notin M_n} \bar{u}_{nj} \leq c L_n n^{-2\kappa},$$

for any $c > 0$.

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

$$P(\widehat{M}_n = M_*) = 1 - o(1).$$

In particular, this ideal situation occurs under the partial orthogonality condition, that is, $\{X_j\}_{j \notin M_*}$ is independent of $\{X_j\}_{j \notin M_*}$, given $W$, which implies $u_j = 0$ for $j \notin M_*$. 
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 \( t_n \).

**Theorem 2.** Under the same conditions in Theorem 1, for any \( t_n = c_5 L_n n^{-2k} \), there exist positive constants \( c_5 \) and \( c_9 \) such that

\[
P(\|\tilde{M}_n\| \leq O(n^{2k} \lambda_{\max}(\Sigma))) \geq 1 - 12p_n L_n \left( 2 + L_n \right) \exp \left( -c_9 n^{1-4k} L_n^{-3} \right) + 3L_n \exp \left( -c_9 n L_n^{-3} \right),
\]

where \( \Sigma = E(\mathbf{Q}^T \mathbf{Q}) \), and \( \mathbf{Q} = (\mathbf{Q}_1, \ldots, \mathbf{Q}_p) \) 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 \( \Sigma \) 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 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-SAC 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-NIS and greedy-NIS, respectively.

#### 4.1 Conditional-INIS Method

The conditional-INIS method builds upon conditional random permutation in determining the threshold \( t_n \). Recall the random permutation used in Fan, Feng, and Song (2011), which generalizes that in Zhao and Li (2012). Randomly permute \( Y \) to get \( Y^{(n)} = (Y_{n1}, \ldots, Y_{nL})^T \) and compute \( \hat{u}_{n1}^{(n)} \), where \( n \) is a permutation of \( \{1, \ldots, n\} \), based on the randomly coupled data \( \{(Y_{ni}, W_i, X_i)\}_{i=1}^n \) 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 as \( t_{jq} \), the \( q \)-th ranked magnitude of \( \{\hat{u}_{nj}^{(n)}, j = 1, \ldots, p\} \). Thus, the NIS step selects variables \( \{j : \hat{u}_{nj}^{(n)} \geq t_{jq}\} \). 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.

1. For \( j = 1, \ldots, p \), compute

\[
\hat{u}_{nj} = \|\tilde{u}_{nj}(\mathbf{W}) + \tilde{b}_{nj}(\mathbf{W})\mathbf{X}_j\|_2^2 - \|\tilde{u}_{nj}(\mathbf{W})\|_2^2,
\]

where the estimates are defined in (6) and (7) using \( \{(Y, W, X_j), j = 1, \ldots, p\} \). Select the top \( K \) variables by ranking their marginal utilities \( \hat{u}_{nj} \), resulting in the index subset \( M_0 \) to condition upon.

2. Regress \( Y \) on \( \{(W, X_j), j \in M_0\} \), and get intercept \( \tilde{b}_{0j}(W) \) and their functional coefficients’ estimators \( \{\tilde{b}_{nj}(W), j \in M_0\} \). Conditioning on \( M_0 \), the \( n \)-dimensional partial residual is

\[
Y^* = Y - \tilde{b}_{0j}(W) - \sum_{j \in M_0} X_j \tilde{b}_{nj}(W).
\]

For all \( j \in M_0 \), compute \( \hat{u}_{nj}^* \) using \( \{Y^*, W, X_j\}, j \in M_0\} \), which measures the additional utility of each covariate conditioning on the selected set \( M_0 \).

3. Repeat Steps 1–2, where we replace \( M_0 \) in Step 1 by \( M_k \), \( l = 1, 2, \ldots \), and get \( A_{l+1} \) and \( M_{l+1} \) in Step 2. Iterate until \( M_{l+1} = M_k \) for some \( k \leq l \) or \( |M_{l+1}| \geq \zeta_n \), for some prescribed positive integer \( \zeta_n \) (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 \( M_0 = \emptyset \)), 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-SAC penalized regression as in Step 2. We then iterate these two steps (screening top \( p_0 \) variables and group-SAC) until there are two identical subsets or the number of variables selected exceeds a prespecified \( \zeta_n \). In our simulation studies, \( q = 1 \).

#### 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 \( M_l = \{j \in A_l : \hat{\phi}_j^{(l)} \neq 0\} \) through minimizing the following objective
function:
\[
\min_{\mathbf{y}} \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \mathbf{B}(W_i) \mathbf{y}_0 - \sum_{j \in A_i} \mathbf{B}(W_i) X_{ji} \mathbf{y}_j \right)^2 \\
+ \sum_{j \in A_i} p_j(\| \mathbf{y}_j \|_2),
\]
where \( \| \mathbf{y}_j \|_2 = \sqrt{\sum_{k=1}^{n} b_{jk}(W_i)|\mathbf{y}_k|^2} \), and \( p_j(\cdot) \) is the SCAD penalty such that
\[
\text{p}_j^\alpha(|x|) = \lambda I(|x| < \lambda) + \frac{(a - |x|)^+}{a - 1} I(|x| > \lambda),
\]
with \( p_j(0) = 0 \). We set \( a = 3.7 \) as suggested and solve the optimization above via local quadratic approximations (Fan and Li 2001). The penalization parameter \( \lambda \) is chosen by Bayesian information criteria (BIC) \( n \log(\hat{\sigma}^2) + k \log n \), where \( \hat{\sigma}^2 \) is the residual variance and \( k \) is the number of covariates chosen.

By Antoniadis and Fan (2001) and Yuan and Lin (2006), the common setup for the following simulations is:

- We compare NIS, lasso, and SIS (independence screening for linear models). The boxplots of MMS are presented in Figure 1.
- We study two settings: 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:

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

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 \( \epsilon \)'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 \( \epsilon \) follows the standard normal distribution. We construct \( \{W, \mathbf{X}\} \) and \( Y \) as follows:

\[ X_j = Z_j + t_1 U_1, \quad j = 1, \ldots, p, \quad W = U_2 + t_2 U_1 + \frac{4 \sin(2\pi W)}{2 - \sin(2\pi W)}, \]

\[ Y = 2X_1 + 3W \cdot X_2 + (W + 1)^2 \cdot X_3 + \epsilon. \]

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 (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...
tuning parameter $K$. Generally speaking, although the lower bound of the TP’s 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., $\mathcal{M}^* \setminus \mathcal{M}_0 = \emptyset$), 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 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>
<thead>
<tr>
<th>Model</th>
<th>TP $\hat{\alpha}_{M_j \setminus M_0}$</th>
<th>Size $\mathcal{M}^* \setminus \mathcal{M}_0$</th>
<th>$\min_{j \in \mathcal{M}^* \setminus \mathcal{M}<em>0} \hat{\alpha}</em>{M_j \setminus M_0}$</th>
<th>$\max_{j \in \mathcal{M}^* \setminus \mathcal{M}<em>0} \hat{\alpha}</em>{M_j \setminus M_0}$</th>
<th>$\max_{j \in {1,...,p} \setminus \mathcal{M}<em>0} \hat{\alpha}</em>{M_j \setminus M_0}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K = 0$</td>
<td>$t_1 = 0, t_2 = 0$</td>
<td>4.00(0.00)</td>
<td>6.68(2.99)</td>
<td>2.96(0.72)</td>
<td>1.22(0.18)</td>
</tr>
<tr>
<td></td>
<td>$t_1 = 3, t_2 = 1$</td>
<td>4.00(0.00)</td>
<td>886.49(88.81)</td>
<td>0.61(0.10)</td>
<td>0.58(0.07)</td>
</tr>
<tr>
<td>$K = 1$</td>
<td>$t_1 = 0, t_2 = 0$</td>
<td>4.00(0.00)</td>
<td>5.70(1.49)</td>
<td>2.83(0.57)</td>
<td>0.75(0.10)</td>
</tr>
<tr>
<td></td>
<td>$t_1 = 3, t_2 = 1$</td>
<td>4.00(0.00)</td>
<td>202.50(154.85)</td>
<td>0.28(0.06)</td>
<td>0.20(0.03)</td>
</tr>
<tr>
<td>$K = 4$</td>
<td>$t_1 = 0, t_2 = 0$</td>
<td>4.00(0.00)</td>
<td>5.14(1.49)</td>
<td>NA</td>
<td>0.06(0.01)</td>
</tr>
<tr>
<td></td>
<td>$t_1 = 3, t_2 = 1$</td>
<td>4.00(0.00)</td>
<td>4.98(0.75)</td>
<td>0.16(0.05)</td>
<td>0.05(0.01)</td>
</tr>
<tr>
<td>$K = 8$</td>
<td>$t_1 = 0, t_2 = 0$</td>
<td>4.00(0.00)</td>
<td>8.92(0.75)</td>
<td>NA</td>
<td>0.05(0.01)</td>
</tr>
<tr>
<td></td>
<td>$t_1 = 3, t_2 = 1$</td>
<td>3.99(0.00)</td>
<td>8.43(0.75)</td>
<td>0.11(0.03)</td>
<td>0.04(0.01)</td>
</tr>
</tbody>
</table>
the same model. As a measure of the complexity of the model, signal-to-noise ratio (SNR), defined by $\frac{\text{var}(\beta^T (W)X)}{\text{var}(\epsilon)}$, 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>TP</th>
<th>FP</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K = 3$</td>
<td>$t_1 = 0, t_2 = 0$</td>
<td>4.00(0.00)</td>
<td>1.27(1.49)</td>
</tr>
<tr>
<td></td>
<td>$t_1 = 3, t_2 = 1$</td>
<td>3.96(0.00)</td>
<td>0.13(0.00)</td>
</tr>
<tr>
<td>$K = 5$</td>
<td>$t_1 = 0, t_2 = 0$</td>
<td>4.00(0.00)</td>
<td>1.57(1.49)</td>
</tr>
<tr>
<td></td>
<td>$t_1 = 3, t_2 = 1$</td>
<td>3.97(0.00)</td>
<td>0.05(0.00)</td>
</tr>
<tr>
<td>$K = 10$</td>
<td>$t_1 = 0, t_2 = 0$</td>
<td>4.00(0.00)</td>
<td>1.20(1.49)</td>
</tr>
<tr>
<td></td>
<td>$t_1 = 3, t_2 = 1$</td>
<td>3.99(0.00)</td>
<td>0.20(0.00)</td>
</tr>
</tbody>
</table>

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 \( \{W, X\} , Y, \) and $\epsilon$ be the same as in Example 3. We now introduce more complexities in the following model:

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

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
The results for the housing value equation are based on a common census tracts of Boston from the 1970 census. Most empirical studies for the housing value equation are based on a common specification (Harrison and Rubinfeld 1978),

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

where the dependent variable MV is the median value of owner-occupied homes, the independent variables are quantified measurement of its neighborhood whose description can be found in the manual of R 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:

\[
\log(MV) = \beta_0(W) + \beta_1(W) \cdot RM^2 + \beta_2(W) \cdot AGE + \beta_3(W) \cdot TAX + \beta_4(W) \cdot (B - 0.63)^2 + \beta_5(W) \cdot CRIM + \epsilon,
\]

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 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 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Correlation</th>
<th>Conditional-INIS</th>
<th>Greedy-INIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X's</td>
<td>X's-W</td>
<td></td>
</tr>
<tr>
<td>( t_1 = 0, t_2 = 0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{SNR} \approx 16.85 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( t_1 = 0, t_2 = 0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{SNR} \approx 3.66 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( t_1 = 3, t_2 = 0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{SNR} \approx 3.21 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( t_1 = 3, t_2 = 1 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{SNR} \approx 3.32 )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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

<table>
<thead>
<tr>
<th>Model</th>
<th>Correlation</th>
<th>Conditional-INIS</th>
<th>Greedy-INIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X's</td>
<td>X's-W</td>
<td></td>
</tr>
<tr>
<td>( t_1 = 0, t_2 = 0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{SNR} \approx 47.68 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( t_1 = 0, t_2 = 0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{SNR} \approx 9.40 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( t_1 = 0, t_2 = 0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{SNR} \approx 8.62 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( t_1 = 0, t_2 = 0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{SNR} \approx 8.18 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( t_1 = 0, t_2 = 0 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \text{SNR} \approx 7.61 )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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:

\[
X_j = Z_j + tU_{1 + j - s}, \quad j = s + 1, \ldots, p.
\]

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.

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 \(k = 1, \ldots, L_n\),

\[
B_k(W) \geq 0 \quad \text{and} \quad \sum_{k=1}^{L_n} B_k(W) = 1 \quad \text{for} \quad W \in \mathcal{W}.
\]

In addition, there exist positive constants \(T_3\) and \(T_4\) such that for any \(\eta_k \in \mathbb{R}, k = 1, \ldots, L_n\),

\[
L_n^{-1}T_3 \sum_{k=1}^{L_n} \eta_k^2 \leq \int \left( \sum_{k=1}^{L_n} \eta_k B_k(w) \right)^2 dw \leq L_n^{-1}T_4 \sum_{k=1}^{L_n} \eta_k^2. \quad (A.1)
\]
Then under condition (iii), for $C_1 = T_1T_2$ and $C_2 = T_1T_3$,

$$C_1L_n^{-1} \leq \mathbb{E}[B_2^2(W)] \leq C_2L_n^{-1}, \quad \text{for } k = 1, \ldots, L_w. \quad (A.2)$$

Furthermore, under condition (iii), it follows from (A.1) that for any $\eta = (\eta_1, \ldots, \eta_L)^T \in \mathbb{R}^L$ such that $\|\eta\|^2 = 1$, $C_1L_n^{-1} \leq \eta^T \mathbb{E}[B_i^T B_j] \eta \leq C_2L_n^{-1}$. Or equivalently,

$$C_1L_n^{-1} \leq \lambda_{\min}(\mathbb{E}[B_i^T B_j]) \leq \lambda_{\max}(\mathbb{E}[B_i^T B_j]) \leq C_2L_n^{-1}. \quad (A.3)$$

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: $P(|X| > t|W) \leq \exp(1 - (t/K)\gamma)$ 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 nonnegative random variable $Z$, $E[Z^m | W] = \int_0^\infty P[Z \geq t | W] dt$. Then we have

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

The last inequality follows from the fact $r \geq 1$, 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 $r_1, r_2 \geq 1$ such that $r_1r_2/(r_1 + r_2) \geq 1$, and

$$P(|Z_i| > t | W) \leq \exp(1 - (t/K_i)\gamma), \quad i = 1, 2$$

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

$$P(|Z_1Z_2| > t | W) \leq \exp(1 - (t/K^*)\gamma)$$

for all $t \geq 0$ and uniformly on $W$.

**Proof.** For any $t > 0$, let $M = (r^*K^*)^{1/(r^* + 1)}$ and $r = r_1r_2/(r_1 + r_2)$. Then uniformly on $W$, we have

$$P(|Z_1Z_2| > t | W) \leq P(M | Z_1 > t | W) + P(|Z_2| > M | W)$$

$$\leq \exp(1 - (t/K^*)\gamma) + \exp(1 - (M/K^*)\gamma)$$

$$= \exp(1 - (t/K^*)\gamma).$$

Let $r^* \in [1, r]$ and $K^* = \max(r^*K, K_1K_2, (1 + \log 2)^{1/r} K_1 K_2)$. It can be shown that $G(t) = (t/K_1K_2)^{1/r} - (t/K^*)^r$ is increasing when $t > K^*$. Hence $G(t) = G(K^*) \geq \log 2$ when $t > K^*$, which implies when $t > K^*$,

$$P(|Z_1Z_2| > t | W) \leq 2\exp(1 - (t/K^*)\gamma).$$

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

**Lemma A.3.** (Bernstein’s inequality, Lemma 2.2.11, van der Vaart and Wellner 1996). For independent random variables $Y_1, \ldots, Y_n$ 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 $\nu_i$. Then

$$P(|Y_1 + \cdots + Y_n| > x) \leq 2 \exp\left[-x^2/(2(\nu + Mx/3))\right],$$

for $\nu \geq \nu_i + \cdots + \nu_n$.

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

$$P(|Y_1 + \cdots + Y_n| > x) \leq 2 \exp\left[-x^2/(2(\nu + Mx/3))\right].$$

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 $j = 1, \ldots, p, k = 1, \ldots, L_i$,

$$P\left(\frac{1}{n} \sum_{i=1}^n X_{ji} B_i(W_j) - E[X_iB_i(W)] \geq \frac{\delta}{n}\right) \leq 4 \exp\left\{-\frac{\delta^2}{b_1L_n^{-1}n + b_2}\right\}.$$

**Proof.** Recall $m(X^i) = E[Y_i|X_i, W_i]$. Let $Z_{kji} = X_{ji} B_i(W_j)$, $m(X^{i'}) = E[Y_iB_i(W) - m(X^i)]$ and $\xi_{kji} = X_{ji} B_i(W_i)$. Then

$$\frac{1}{n} \sum_{i=1}^n X_{ji} B_i(W_j) - E[X_iB_i(W)] = \frac{1}{n} \sum_{i=1}^n (X_{ji} B_i(W_j)m(X^{i'}) + X_{ji} B_i(W_i)\xi_{kji})$$

$$\leq \frac{1}{n} \sum_{i=1}^n Z_{kji} + \frac{1}{n} \sum_{i=1}^n \xi_{kji}.$$
and
\[ P \left( \frac{1}{n} \sum_{i=1}^{n} B_k(W_i) s_i \geq \frac{\delta}{2n} \right) \leq 2 \exp \left\{ -\frac{\delta^2}{16eK_1^2 C_2 L_n^{-1}n + 4K_0\delta} \right\}. \]  
(A.8)

Let \( b_1 = 16eC_2 \max(4K_1^2, K_3^2, 4K_0^2, K_2^2) \) 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 \( \Sigma_j = E[Q_j^T Q_j] \),
\[ C_3 L_n^{-1} \leq \lambda_{\min}(\Sigma_j) \leq \lambda_{\max}(\Sigma_j) \leq C_4 L_n^{-1}. \]  
(A.9)

**Proof.** Recall that \( Q_j = (B, X_j) \). For any \( \eta = (\eta_1^T, \eta_2^T) \in \mathbb{R}^{2L_n} \) such that \( \|\eta\|_2^2 = 1 \),
\[ \eta^T \Sigma_j \eta = E \left[ \left( \begin{array}{cc} B_{\eta_1} & B_{\eta_2} \\ E[X_j\|W] & E[X_j\|W] \end{array} \right) \left( \begin{array}{c} B_{\eta_1} \\ B_{\eta_2} \end{array} \right) \right]. \]

Consider eigenvalues \( \lambda_1 \) and \( \lambda_2 (\lambda_1 > \lambda_2) \) of the \( 2 \times 2 \) matrix on the right-hand side of the equation above, we have \( \lambda_1 + \lambda_2 = 1 + E[X_j\]W] (trace) and \( \lambda_1 \cdot \lambda_2 = \text{var}[X_j\|W] \) (determinant). Therefore, by Lemma A.1 \( \lambda_1 \leq 1 + E[X_j\|W] \leq 1 + 4eK_1^2 \) and by assumption (i)
\[ \lambda_2 \geq \frac{\text{var}[X_j\|W]}{E[X_j\|W] + 1} \geq \frac{h_1}{1 + 4eK_1^2}. \]
Using the above two bounds on the minimum and maximum eigenvalues, we have
\[ \frac{h_1}{1 + 4eK_1^2} E[B_{\eta_1}^T + (B_{\eta_2}^T \eta_2^T)] \]
\[ \leq \eta^T \Sigma_j \eta \leq \left( 1 + 4eK_1^2 \right) E[B_{\eta_1}^T + (B_{\eta_2}^T \eta_2^T)]. \]
By (A.3), we have
\[ \frac{h_1 C_1}{1 + 4eK_1^2} L_n^{-1} \leq \eta^T \Sigma_j \eta \leq \left( 1 + 4eK_1^2 \right) C_2 L_n^{-1}. \]
Take \( C_1 = h_1 C_1 L_n^{-1}/(1 + 4eK_1^2) \) and \( C_4 = (1 + 4eK_1^2) C_2 L_n^{-1} \), result follows.

Throughout the rest of the proof, for any matrix \( A \), let \( \|A\| = \sqrt{\lambda_{\max}(A^T A)} \) be the operator norm and \( \|A\|_\infty = \max_j |A_{ij}| \) 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
\[ P(\|\Sigma_{aj} - \Sigma_j\| \geq L_n \delta/n) \leq 6L_n^2 \exp \left\{ -\frac{\delta^2}{b_3 L_n^{-1}n + b_3 \delta} \right\}, \]
\[ P \left( \frac{1}{n} B_j^T B_j - E[B_j^T B_j] \geq L_n \delta/n \right) \leq 6L_n^2 \exp \left\{ -\frac{\delta^2}{b_3 L_n^{-1}n + b_3 \delta} \right\}, \]
where \( \Sigma_{aj} = \frac{1}{n} Q_j^T Q_j \). In addition, for any given positive constant \( b_3 \), there exists some positive constant \( b_4 \) such that
\[ P(\|\Sigma_{aj}\|^{-1} - \|\Sigma_j\|^{-1} \geq b_3 \|\Sigma_j\|^{-1}) \leq 6L_n^2 \exp \left\{ -b_4 L_n^{-3}n \right\}, \]
and for any positive constant \( b_3 \), there exists some positive constant \( b_4 \) such that
\[ P \left( \left\| \frac{1}{n} B_j^T B_j \right\|^{-1} - \left\| E[B_j^T B_j] \right\|^{-1} \geq b_4 \left\| E[B_j^T B_j] \right\|^{-1} \right\} \leq 6L_n^2 \exp \left\{ -b_4 L_n^{-3}n \right\}. \]

**Proof.** Observe that for \( j = 1, \ldots, p \),
\[ \Sigma_{aj} - \Sigma_j = \left( \begin{array}{cc} D_{1j} & D_{2j} \\ D_{2j}^T & D_{3j} \end{array} \right), \]
where \( D_{1j} = \frac{1}{n} \sum_{i=1}^{n} B_j(W_i)B_i) - E[B_j^T B_j], \) \( D_{2j} = \frac{1}{n} \sum_{i=1}^{n} X_jB_j(W_i) - E[X_jB_j^T B_j], \) and \( D_{3j} = \frac{1}{n} \sum_{i=1}^{n} X_j^2 B_j^T B_j(W_i) - E[X_j^2 B_j^T B_j]. \) Then
\[ \|\Sigma_{aj} - \Sigma_j\| \leq 2L_n \|\Sigma_j\| - \Sigma_j\|_\infty \]
\[ = 2L_n \max(\|D_{1j}\|_\infty, \|D_{2j}\|_\infty, \|D_{3j}\|_\infty). \] (A.10)

We first bound \( \|D_{1j}\|_\infty \). Recall that \( 0 \leq B_k(\cdot) \leq 1 \) on \( \mathbb{W} \), so
\[ \|B_k(W_i)B_i(W_j) - E[B_k(W_i)B_j(W_j)]\| \leq 2, \]
for all \( k \) and \( l \). By (A.2),
\[ \text{var}(B_k(W_i)B_i(W_j) - E[B_k(W_i)B_j(W_j)]) \leq E[B_k^2(W_i)B_k^2(W_j)] \leq C_2 L_n^{-1}. \]
By Lemma A.4, we have
\[ P \left( \frac{1}{n} \sum_{i=1}^{n} B_k(W_i)B_i(W_j) - E[B_k(W_i)B_j(W_j)] \geq \delta/6n \right) \]
\[ \leq 2 \exp \left\{ -\delta^2/(72C_2 L_n^{-1}n + 24\delta) \right\}. \] (A.11)
We next bound \( \|D_{2j}\|_\infty \). Note that for \( k, l = 1, \ldots, L_n \),
\[ E[X_jB_k(W_i)B_j(W_l) - E[X_jB_k(W_i)B_j(W_l)]^m] \]
\[ \leq 2^m E[X_jB_k(W_i)B_j(W_l)]^m \leq m! \|2K_j\|^{2m}(8eK_1^2 C_2 L_n^{-1})/2, \]
where Lemma A.1 was used in the last inequality. By Lemma A.3, we have
\[ P \left( \frac{1}{n} \sum_{i=1}^{n} X_jB_k(W_i)B_j(W_l) - E[X_jB_k(W_i)B_j(W_l)] \geq \delta/6n \right) \]
\[ \leq 2 \exp \left\{ -\delta^2/(576eK_1^2 C_2 L_n^{-1}n + 24K_1\delta) \right\}. \] (A.12)
Similarly, we can bound \( \|D_{3j}\|_\infty \). For every \( m \geq 2 \), for \( k, l = 1, \ldots, L_n \), there exists a constant \( K_0 > 0 \) such that
\[ E[X_j^2 B_k(W_i)B_j(W_l) - E[X_j^2 B_k(W_i)B_j(W_l)]^m] \]
\[ \leq m! (2K_j)^{m-2}(8eK_1^2 C_2 L_n^{-1})/2. \]
By Lemma A.3, we have
\[ P \left( \frac{1}{n} \sum_{i=1}^{n} X_jB_k(W_i)B_j(W_l) - E[X_j^2 B_k(W_i)B_j(W_l)] \geq \delta/6n \right) \]
\[ \leq 2 \exp \left\{ -\delta^2/(576eK_1^2 C_2 L_n^{-1}n + 24K_1\delta) \right\}. \] (A.13)

Let \( b_3 = 72C_2 \max(1, 8eK_1^2, 8eK_0^2) \) and \( b_4 = 24 \max(1, K_1, K_0) \), then combining (A.10)–(A.13) we have
\[ P(\|\Sigma_{aj} - \Sigma_j\| \geq L_n \delta/n) \leq 6L_n^2 \exp \left\{ -\frac{\delta^2}{b_4 L_n^{-3}n + b_4 \delta} \right\}. \]
Observe that \( \| \frac{1}{n} B_n^* B_n - E[B^T B] \| \leq 2 L_n \| D \|_\infty \). Thus, we have also proved that
\[
P \left( \left\| \frac{1}{n} B_n^* B_n - E[B^T B] \right\| \geq L_n \delta / n \right) \leq 6L_n^2 \exp \left( -\frac{\delta^2}{5b_3 L_n^4 n + b_4 \delta} \right).
\]

We next prove the second part of the lemma. Note that for any symmetric matrices \( A \) and \( B \) (Fan, Feng, and Song 2011),
\[
|\lambda_{\text{max}}(A) - \lambda_{\text{min}}(B)| \leq \max(|\lambda_{\text{max}}(A - B)|, |\lambda_{\text{max}}(B - A)|).
\]
It then follows from (A.14) that
\[
|\lambda_{\text{min}}(\Sigma_{aj}) - \lambda_{\text{min}}(\Sigma_j)| \leq 2L_n \| \Sigma_{aj} - \Sigma_j \|_\infty,
\]
which implies that
\[
P(|\lambda_{\text{min}}(\Sigma_{aj}) - \lambda_{\text{min}}(\Sigma_j)| \geq b_5 \lambda_{\text{min}}(\Sigma_j)) \leq 6L_n^2 \exp \left( -b_5 L_n^{-3} n \right).
\]
(A.15)
Let \( \delta = b_6 C_3 L_n^{-3} n \) in (A.15) for \( b_6 \in (0, 1) \). According to (A.9), we have
\[
P(|\lambda_{\text{min}}(\Sigma_{aj}) - \lambda_{\text{min}}(\Sigma_j)| \geq b_5 \lambda_{\text{min}}(\Sigma_j)) \leq 6L_n^2 \exp \left( -b_5 L_n^{-3} n \right).
\]
(A.16)
for some positive constant \( b_5 \). Next observe the fact that for \( x, y > 0 \), \( a \in (0, 1) \), and \( b = 1/(1 - a) - 1 \), \( x^{-1} - y^{-1} \geq b y^{-1} \), which implies \( |x - y| \geq ay \). This is because \( x^{-1} - y^{-1} \geq b y^{-1} \) is equivalent to \( x^{-1} \geq b x^{-1} + 1 \), or \( x - y \leq -ay \); on the other hand, \( x^{-1} - y^{-1} \leq b y^{-1} \) implies \( x^{-1} \leq (1 - \frac{a}{b} x^{-1})^{-1} \leq (1 - \frac{a}{b} y^{-1})^{-1} \) as \( a \in (0, 1) \), and therefore \( x - y \geq ay \). Then let \( b_6 = 1/(1 - b_6) - 1 \), it follows from (A.16) that
\[
P\left( |\lambda_{\text{min}}(\Sigma_{aj})| - |\lambda_{\text{min}}(\Sigma_j)| \geq b_5 \lambda_{\text{min}}(\Sigma_j) \right) \leq 6L_n^2 \exp \left( -b_5 L_n^{-3} n \right).
\]
(A.17)
Following the same procedure, by (A.3) we also have for any positive constant \( b_7 \), there exists some positive constant \( b_8 \), such that
\[
P \left( \left| \frac{1}{n} B_n^* B_n - E[B^T B] \right| \geq b_7 \lambda_{\text{min}}(E[B^T B]) \right) \leq 6L_n^2 \exp \left( -b_7 \lambda_{\text{min}}(E[B^T B]) \right).
\]
(A.18)
for some positive constant \( b_7 \). Note that \( |\lambda_{\text{min}}(A) - \lambda_{\text{max}}(A)| = \lambda_{\text{max}}(A) - \lambda_{\text{min}}(A) \).

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 \( \alpha_j^{(r)} \) and \( \beta_j^{(r)} \) such that \( |a_j - a_j^{(r)}|_\infty \leq M_1 L_n^{-d} \) and \( |b_j - b_j^{(r)}|_\infty \leq M_1 L_n^{-d} \), where \( S_n \) is the space of polynomial splines of degree \( d \geq 1 \) 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 \( \eta_j \) be \( L_n \)-dimensional vectors such that \( \alpha_j^{(r)} = B(W) \eta_j^{(r)} \) and \( \beta_j^{(r)} = B(W) \beta_j^{(r)} \). Recall that \( \tilde{a}_j(W) = B(W) \tilde{\eta}_j \) and \( \tilde{b}_j(W) = B(W) \tilde{\beta}_j \). By definition of \( \tilde{\eta}_j \) and \( \tilde{\beta}_j \),
\[
(\tilde{a}_j, \tilde{b}_j) = \arg \min_{a_j, b_j} \text{E}[E[Y | W, X_j] - a_j(W) - b_j(W) X_j]^2,
\]
and therefore \( \| \text{E}[E[Y | W, X_j] - a_j(W) - b_j(W) X_j]^2 \| \leq 2 \| \text{E}[E[Y | W, X_j] - a_j^{(r)}(W) - b_j^{(r)}(W) X_j]^2 \| \). On the other hand, \( \| \tilde{a}_j + \tilde{b}_j X_j - (a_j + b_j X_j) \| \leq 2 \| a_j - a_j^{(r)} \| + 2 \| b_j - b_j^{(r)} \| X_j \| \leq 2M_2 L_n^{-d} (1 + \text{E}[X_j^2]) \). Similarly, we have
\[
\| a_0^2 - \| a_0 \| \leq M_2 L_n^{-2d}.
\]
(A.19)
By taking \( M_1 = M_2(SeK^2 + 3) \) (see Lemma A.1), the first part of Proposition 1 follows from (A.17) and (A.18): \( u_j - \hat{u}_j = \| a_j + b_j X_j \|^2 - \| a_0 \|^2 - (\| a_j + b_j X_j \|^2 - \| \hat{a}_j \|^2) \leq M_1 L_n^{-2d} \). (A.20)
We first focus on \( S_1 \). Let \( a_0 = \frac{1}{n} Q_{yj}^T Y \) and \( a = E[Q_{yj}^T Y] \). Then \( S_1 = (a_0 - a)^T \Sigma_{aj}^{-1} (a_0 - a) + 2(a_0 - a)^T \Sigma_{aj}^{-1} a + a^T \Sigma_{aj}^{-1} a \). Denote the last three terms, respectively, by \( S_{11}, S_{12}, \) and \( S_{13} \). We first deal with \( S_{11} \). Note that
\[
\| S_{11} \| \leq \| \Sigma_{aj}^{-1} \| \| a_0 - a \|^2.
\]
(A.21)
By Lemma A.5 and the union bound of probability,
\[
P(\| a_0 - a \|^2 \geq 2L_n \delta^2 / n^2) \leq 8L_n \exp \left( -\delta^2 / (b_1 L_n^{-1} n + b_2 \delta) \right).
\]
(A.22)
According to the second part of Lemma A.7, for any given positive constant \( b_3 \), there exists a positive constant \( b_4 \) such that
\[
P \left( \| \Sigma_{aj}^{-1} \| \geq b_3 \| \Sigma_j^{-1} \| \right) \leq 6L_n^2 \exp \left( -b_3 L_n^{-3} n \right)
\]
Then it follows from (A.9) that
\[
P \left( \| \Sigma_{aj} \| \geq b_3 + b_4 \right) \leq 6L_n^2 \exp \left( -b_4 L_n^{-3} n \right)
\]
(A.23)
Combining (A.20) and (A.22) and based on the union bound of probability, we have
\[
P \left( |S_{11}| \geq 2(b_3 + b_4) L_n^{-2d} \delta^2 / n^2 \right) \leq 8L_n \exp \left( -\delta^2 / (b_1 L_n^{-1} n + b_2 \delta) \right) + 6L_n \exp \left( -b_4 L_n^{-3} n \right)
\]
We next bound \( S_{12} \). Note that
\[
|S_{12}| \leq 2 |a_0 - a| \cdot \| \Sigma_{aj}^{-1} \| \cdot \| a \|^2.
\]
(A.24)
By Lemma A.1,
\[
\| a \|^2 = \| E[B^T Y] \|^2 + \| E[X^T B^T Y] \|^2 \leq \sum_{k=1}^{L_n} \left( E[B_k^T m(X_k')] + E[B_k^T X_k^T m(X_k')] \right) \leq 4c_1(K_2^2 + K_3^2)
\]
(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

\[
P(|S_1| \geq 4\sqrt{2}(b_1 + 1)e^{1/2}c_2^{1/2}(K_2^2 + K_3^2)^{1/2}C_1^{-1}L_n^2/n^2) \\
\leq 8L_n \exp \left\{ -\delta^2/(b_1L_n^2 + b_2\delta) \right\} + 6L_n^3 \exp \left\{ -b_2L_n^{-3}n \right\}.
\]

(A.26)

To bound $S_{13}$, note that

\[
|S_{13}| = a^T \Sigma_j (\Sigma_j - \Sigma_n) a \leq \|\Sigma_j - \Sigma_n\|^2 \cdot \|\Sigma_j - \Sigma_n\| \cdot \|a\|^2.
\]

(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 $b_1$, $b_2$, and $b_3$ such that

\[
P(|S_{13}| \geq 4eC_2(K_2^2 + K_3^2)(b_1 + 1)^2C_2^{-2}L_n^2/n) \\
\leq 6L_n^2 \exp \left\{ -\delta^2/(b_1L_n^2 + b_2\delta) \right\} + 18L_n^2 \exp \left\{ -b_2L_n^{-3}n \right\}.
\]

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

\[
P(|S_1| \geq s_4L_n^{2\delta}/n^2 + s_5L_n^{3/2}\delta/n + s_6L_n^2/\delta/n) \\
\leq 8L_n \exp \left\{ -\delta^2/(b_1L_n^2 + b_2\delta) \right\} + 18L_n^2 \exp \left\{ -b_2L_n^{-3}n \right\}.
\]

Let $(s_1 + s_2)L_n^{2\delta}/n^2 + (s_3 + s_4)L_n^{3/2}\delta/n + (s_5 + s_6)L_n^2/\delta/n = c_2L_n^{2\delta}/n^2$ for any given $c_2 > 0$ (e.g., take $\delta = c_2L_n^{-2\delta}/(s_3 + s_6)$). There exist some positive constants $c_5$ and $c_6$ such that

\[
P(|\tilde{u}_{nj} - \tilde{u}_j| \geq c_2L_n^{-2\delta}/n^2) \\
\leq (24L_n + 12L_n^2) \exp \left\{ -c_5n^{1-4\delta}L_n^{-3}\right\} + 36L_n^2 \exp \left\{ -c_5L_n^{-3}n \right\}.
\]

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

We now prove part (ii). Note that on the event $A_n = \left\{ \tilde{u}_{nj} - \tilde{u}_j \leq c_1\xi L_n^{-2\delta}/2 \right\}$ by Proposition 1, we have $\tilde{u}_{nj} \geq c_1\xi L_n^{-2\delta}/2$, for all $j \in M_n$. Hence, by choosing $\tau_n = c_1\xi L_n^{-2\delta}/2$, we have $M_n \subset \tilde{M}_n$. On the other hand, the union bound of probability, there exist positive constants $c_6$ and $c_7$, such that

\[
P(A_n^c) \leq s_4\left\{ (24L_n + 12L_n^2) \exp \left\{ -c_6n^{1-4\delta}L_n^{-3}\right\} + 36L_n^2 \exp \left\{ -c_6L_n^{-3}n \right\} \right\},
\]

and Theorem 1(2) follows.

Proof of Theorem 2. Let $\tilde{a} = \arg \min Q = \arg \min_{\tilde{Q} \in 2P(n)} E[(Y - \tilde{Q}\tilde{a})^2]$, where $Q = (Q_1, \ldots, Q_p)$ is a $pL_n$-dimensional vector of functions. Then we have

\[
E[Q^T(Y - \tilde{Q}\tilde{a})^2] = \theta_{2pL_n},
\]

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

\[
\|E[Q^TY]\|_F^2 = \tilde{a}^T \Sigma \tilde{a} \leq \lambda_{\max}(\Sigma) \tilde{a}^T \Sigma \tilde{a}.
\]

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

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

\[
\sum_{j=1}^p \tilde{u}_j \leq \max_{1 \leq j \leq p} \lambda_{\max}(\Sigma_j^{-1}) \sum_{j=1}^p \|E[Q_j^TY]\|_F^2 = \max_{1 \leq j \leq p} \lambda_{\max}(\Sigma_j^{-1}) \|E[Q_j^TY]\|_F^2.
\]

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

\[
\sum_{j=1}^p \tilde{u}_j \leq \max_{1 \leq j \leq p} \lambda_{\max}(\Sigma_j^{-1}) \|E[Q_j^TY]\|_F^2.
\]

By Lemma A.6 and (A.29), the last term is of order $O(L_n\lambda_{\max}(\Sigma))$. This implies that the number of $\{j : \tilde{u}_j > \delta L_n^{-2\delta}\}$ cannot exceed $O(n^2\lambda_{\max}(\Sigma))$ for any $\delta > 0$. On the set $B_n = \{\tilde{u}_j > \delta L_n^{-2\delta}\}$, the number of $\{j : \tilde{u}_j > \delta L_n^{-2\delta}\}$ cannot exceed the number of $\{j : \tilde{u}_j > \delta L_n^{-2\delta}\}$, which is bounded by $O(n^2\lambda_{\max}(\Sigma))$. By taking $\delta = c_5/2$, we have

\[
P(|\tilde{a}_j| \leq O(n^2\lambda_{\max}(\Sigma))) \geq P(B_n).
\]

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

\[\text{[Received March 2013. Revised October 2013.]}\]

REFERENCES


\[\text{[1270,1275]}\]

\[\text{[1271]}\]


