NONPARAMETRIC INDEPENDENCE SCREENING AND STRUCTURAL IDENTIFICATION FOR ULTRA-HIGH DIMENSIONAL LONGITUDINAL DATA

BY MING-YEN CHENG *, TOSHIRO HONDA, JIALIANG LI,, and HENG PENG

National Taiwan University*, Hitotsubashi University†, National University of Singapore‡, and Hong Kong Baptist University§

Ultra-high dimensional longitudinal data are increasingly common and the analysis is challenging both theoretically and methodologically. We offer a new automatic procedure in hunting for a sparse semivarying coefficient model, which has been widely accepted for modeling longitudinal data. Our proposed method first reduces the number of covariates to a moderate order by employing a screening procedure, and then identifies both the varying and constant coefficients using a group SCAD estimator, which is then refined by accounting for the within-subject correlation. The screening procedure is based on working independence and B-spline marginal models. Under weaker conditions than those in the literature, we show that with high probability only irrelevant variables will be screened out, and the number of selected variables can be bounded by a moderate order. This allows the desirable sparsity and oracle properties of the subsequent structure identification step. Note that existing methods require some kind of iterative screening in order to achieve this, thus they demand heavy computational effort and consistency is not guaranteed. The refined semivarying coefficient model employs profile least squares, local linear smoothing and nonparametric covariance estimation, and is semiparametric efficient. We also suggest ways to implement the proposed methods, and to select the tuning parameters. An extensive simulation study is summarized to demonstrate its finite sample performance and the yeast cell cycle data is analyzed.

*Research was supported in part by the National Science Council grants NSC97-2118-M-002-001-MY3 and NSC101-2118-M-002-001-MY3, and the Mathematics Division, National Center of Theoretical Sciences (Taipei Office)
†Research was supported by the JSPS Grant-in-Aids for Scientific Research (A) 24243031 and (C) 25400197.
‡Research was supported by grants AcRF R-155-000-130-112 and NMRC/CRG/0014/2012.
§Research was supported by CERG grants from the Hong Kong Research Grants Council (HKBU 20610 and HKBU 202012), FRG grants from Hong Kong Baptist University (FRG2/11-12/130 and FRG2/12-13/077), and a grant from NSFC (11271094).
MSC 2010 subject classifications: Primary 62G08
Keywords and phrases: Independence screening, longitudinal data, B-spline, SCAD, sparsity, oracle property.
1. Introduction. Longitudinal data arise in many modern scientific fields, including finance, genetics, medicine, and so on. Specifically, we consider observing independent realizations of a scalar response process \( y(t) \) and a \( p \)-dimensional covariate process \( x(t) = (x^{(1)}(t), \ldots, x^{(p)}(t))^T \) at \( t = t_1, \ldots, t_m \), where \( t_1, \ldots, t_m \), independent of \( x(t) \), are i.i.d. with density \( f_T(t) \) satisfying \( C_1 \leq f_T(t) \leq C_2 \). In this paper, \( C, C_1, C_2, \ldots \) are positive generic constants. There exist various parametric, nonparametric and semiparametric models for regressing \( y(t) \) on \( x(t) \), [11, 39]. Among the three categories, the semiparametric approach, in particular varying coefficient models, is in general preferred to the other two. Parametric models are efficient if correctly specified, but can be seriously biased otherwise. While nonparametric approach avoids this problem, the curse of dimensionality issue arises.

Consider the varying coefficient model, which can capture the dynamical impacts of the covariates on the response variable, given as:

\[
y(t) = \beta_0(t) + \sum_{k=1}^p x^{(k)}(t) \beta_k(t) + \epsilon(t), \quad t \in [0, 1],
\]

where \( \beta_0(t), \beta_1(t), \ldots, \beta_p(t) \) are the unknown varying coefficients and \( \epsilon(t) \) is an error process with \( \mathbb{E}\{\epsilon(t)\} = 0 \). For a generic real-valued function \( g \), write \( g(t) = (g(t_1), \ldots, g(t_m)) \), where \( t = (t_1, \ldots, t_m) \). Suppose we are given \( n \) independent observations on \((y(t), x(t))\): for the \( i \)-th subject we observe \( y_i(t) \) and \( x_i(t) = (x_i^{(1)}(t), \ldots, x_i^{(p)}(t))^T \) at \( t = t_{i1}, \ldots, t_{im_i} \). Here, \( m_i \) can be random, but is uniformly bounded and independent of \( x_i(t) \). Writing \( t_i = (t_{i1}, \ldots, t_{im_i})^T \), we have \((y_i(t_i), x_i(t_i))\) where \( x_i(t_i) = (x_i(t_{i1}), \ldots, x_i(t_{im_i})) \) is a \( p \times m_i \) random matrix. Based on model (1.1), we have

\[
y_i(t_i) = \beta_0(t_i) + \sum_{k=1}^p x_i^{(k)}(t_i) \beta_k(t_i)^T + \epsilon_i(t_i), \quad i = 1, \ldots, n.
\]

where \( \epsilon_i(t_i) = (\epsilon_i(t_{i1}), \ldots, \epsilon_i(t_{im_i})) \) is the error process in the \( i \)-th subject.

As technology evolves rapidly over the recent decades, high-dimensional longitudinal data have become common, and they pose new challenges in the analysis from methodological, theoretical, and computational aspects. When \( p \) is large, it is often the case that many of the covariates are irrelevant. Then, including the irrelevant variables in the model would create undesirable identifiability and estimation instability problems, and variable selection is a natural way to address the challenges. In parametric regression for i.i.d data, popular tools for this purpose include the SCAD [6], Lasso [27], adaptive Lasso [40], and group Lasso [22, 37] estimators. These ideas have been adopted to select important variables in varying coefficient models for
i.i.d. data [35]. For longitudinal data, when $p$ is fixed, group SCAD penalized B-spline methods were studied in [30] and [23], and regularized P-spline methods were considered in [2]. When $p$ diverges and $p = o(n^{2/5})$, where $n$ is the sample size, [33] and [1] examined adaptive group Lasso estimators.

However, it occurs often in today’s longitudinal studies that $p$ is very large. An example we will investigate in Section 5.2 is the famous yeast cell cycle data set, which consists of gene expression measurements at different time points during the cell cycle [25]. In this dataset, $p = 96$ and $n = 297$, thus $p$ is much larger than $n^{2/5} \approx 10$. Under such circumstances, there is no guarantee that existing variable selection procedures can find the relevant variables consistently. More generally, we consider the ultra-high dimensional case where $p$ can be larger than $n$. Our idea is first reducing the dimensionality to a moderate order by employing some screening procedure, and then selecting variables using a group SCAD estimator. In parametric settings, existing screening methods include the sure independence screening procedures [7, 10], the rank correlation screening procedure [17], and others. In semiparametric settings, screening procedures have been considered for additive and varying coefficient models [3, 9, 20] when the data are i.i.d.

In the present setup, we adopt the nonparametric independence screening (NIS) idea in [3]. However, the covariance structure of $\epsilon(t)$ is unknown and it is infeasible to estimate it at this stage, so we base our NIS procedure on a working independence. Intuitively, this approach is expected to work since the coefficient estimators based on working independence achieve the same convergence rate as that based on the true covariance function.

Under weaker conditions than those in the literature, our NIS step can effectively cut the dimensionality down to a moderate order, $q$ say. Then, writing as $x^{(1)}, \ldots, x^{(q)}$ the remaining variables, we now reduce the full varying coefficient model (1.2) to the following lower-dimensional one:

\[(1.3) \quad y_i(t_i) = \beta_0(t_i) + \sum_{k=1}^{q} x^{(k)}_i(t_i)\beta_k(t_i)^T + \epsilon_i(t_i) , \ i = 1, \ldots, n.\]

Under appropriate smoothness assumptions, we can estimate the unknown coefficient functions in model (1.3) using B-spline smoothing [24]. However, the dimensionality $q$ may be still too large for the modeling purpose, thus it is preferable to further select among these $q$ variables the significant ones. Noticeably, we can proceed directly with variable selection as $q$ can be controlled at $o(n^{2/5}/\sqrt{\log n})$ after the NIS step, while existing methods require some sort of iterative screening to achieve similar goals [3, 9]. We choose the SCAD penalty in both steps because it enjoys a faster convergence rate than the Lasso $L_1$ penalty when the dimension is very large [12].
Besides, we pay attention to the structure identification problem. That is, some of the important variables may simply have constant effects. Identifying the non-zero constant coefficients is an important issue because treating a constant coefficient as varying will yield a slower convergence rate than $\sqrt{n}$. When $p$ is fixed, significant effort has been devoted to address this problem in varying coefficient models for both i.i.d. and longitudinal data [34, 39, 26]. In addition, structure identification was considered for partially linear additive models by [38], and Cox proportional hazard models with varying coefficients by [36, 19]. To achieve simultaneously variable selection and structure identification, we construct a group SCAD penalty to penalize both spurious non-constant effects and spurious non-zero effects. After this step, we further reduce the varying coefficient model (1.3) to the following semivarying coefficient model, for $i = 1, \ldots, n$,

\begin{equation}
(1.4) \quad y_i(t_i) = \beta_0(t_i) + \sum_{k=1}^{s_1} x_i^{(k)}(t_i) \beta_k + \sum_{k=s_1+1}^{s} x_i^{(k)}(t_i) \beta_k(t_i)^T + \epsilon_i(t_i),
\end{equation}

where $s_1$ and $s$ satisfy $0 \leq s_1 \leq s < q$, $\beta_1, \ldots, \beta_{s_1}$ are the constant coefficients, and $\beta_{s_1+1}(t), \ldots, \beta_s(t)$ are the functional coefficients. We treat (1.4) as the final model, and estimate both the constant and varying coefficient functions with the covariance structure of $\epsilon(t)$ taken into account.

To the best of our knowledge, for the present setup, both screening and simultaneous variable selection and structure identification have not been studied before, and the estimation methods are new. Note that, in [39], $p$ is fixed and the structure identification method is a model selection approach. We show both theoretically and numerically that, for $p$ as any exponential order of $n$, using working independence only, the proposed NIS procedure can keep the relevant variables with high probability. In addition, we relax the conditions on the threshold parameter as compared to those in the literature [3, 9, 20], such that the dimensionality after the NIS step can be controlled at a moderate order fulfilling the conditions on the dimensionality in the subsequent group SCAD step. This result provides the theoretical ground for our new sequential screening and variable selection approach. In addition, we discuss the computational and tuning parameter selection issues.

In Section 2 our NIS procedure is introduced and its theoretical properties are studied. The group SCAD procedure for simultaneous variable selection and structure identification, and its consistency, sparsity and oracle properties are given in Section 3. The refined estimation procedure for estimating the constant and varying coefficients in the final model (1.4) is detailed in Section 4. Results of a simulation study and an application to the yeast
cell cycle data are reported and discussed in Section 5. Proofs of all the theorems, and some lemmas, are placed in Appendix A and Supplement A.

2. Nonparametric independence screening. We introduce some notation first. Denote the sup norm and the $L_2$ norm of a function $g$ on $[0,1]$ by $\|g\|_\infty$ and $\|g\|_{L_2}$, respectively. In order to describe and examine our procedures, we define respectively the empirical and theoretical inner products of two vector-valued stochastic processes $u(t) \in \mathbb{R}^k$ and $v(t) \in \mathbb{R}^l$ by

$$\langle u, v^T \rangle_n = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} (u_i(t_{i1}), \ldots, u_i(t_{im_i}))(v_i(t_{i1}), \ldots, v_i(t_{im_i}))^T$$

and

$$\langle u, v^T \rangle = E\{\langle u, v^T \rangle_n\},$$

where $u_i(t)$ and $v_i(t)$ are i.i.d. samples of $u(t)$ and $v(t)$. When $u(t)$ is not stochastic, we should take $u_i(t) = u(t)$. When $k = 1$, we define respectively $\|u\|_n$ and $\|u\|$ by $\|u\|_n = \langle u, u \rangle_n$ and $\|u\|^2 = \langle u, u \rangle$. Note that for any square integrable function $g$ on $[0,1]$, $C_1\|g\|_{L_2} \leq \|g\| \leq C_2\|g\|_{L_2}$ uniformly in $g$.

2.1. Nonparametric Independence Screening Algorithm. Consider the full model (1.2). Define the set of indices of relevant covariates by

$$M_\kappa = \{k \geq 1 | \|\beta_k\|^2 \geq C_\kappa n^{-2\kappa} L\},$$

for some positive constant $\kappa$. Here $L$ is the dimension of the B-spline basis. Under the sparsity assumption M2(2) given in Section 2.2, we can carry out the nonparametric independence screening (NIS) prescribed in the following.

Similar to (3) of [9], we consider for each $k = 1, \ldots, p$ a marginal model for $y(t)$ and $x^{(k)}(t)$ defined by

$$y(t) = a_k(t) + b_k(t)x^{(k)}(t) + \eta_k(t),$$

where $a_k(t)$ and $b_k(t)$ are given by $\arg\min_{a, b \in L^2[0,1]} \|y - a - bx^{(k)}\|^2$. Alternatively, [20] employed a conditional correlation approach. Let $B(t) = (B_1(t), \ldots, B_L(t))^T$ be an equi-spaced B-spline basis of order 3 on $[0,1]$, where $L$ is the dimension of the basis. Write $B(t_i) = (B(t_{i1}), \ldots, B(t_{im_i}))$. Then, under the smoothness conditions specified in Assumption M1, $b_k(t)$ in (2.1) can be approximated by some linear combination of $B(t)$. Thus, we can estimate $b_k$ by minimizing the following objective function:

$$\|y - \gamma_1 T B - \gamma_2 T x^{(k)}(t) \|_n^2 = \|y - \gamma_1 T B - \gamma_2 T W_k\|_n^2, \quad \gamma_1, \gamma_2 \in \mathbb{R}^L.$$
Note that the regressors $W_k(t)$ and $W_k(t)$, and their sample versions $W_{ik}(t)$ and $W_{ik}(t_i)$, for the above B-spline estimation of $b_k$ are given by:

$$W_k(t) = w^{(k)}(t)B(t) = (w_{k1}(t), \ldots, w_{kL}(t))^T \in \mathbb{R}^L,$$

$$W_k(t) = (W_k(t_1), \ldots, W_k(t_m)) \in \mathbb{R}^{L \times m},$$

$$W_{ik}(t) = (W_{ik1}(t), \ldots, W_{ikL}(t))^T = x_{i}^{(k)}(t)B(t) \in \mathbb{R}^L, i = 1, \ldots, n,$$

$$W_{ik}(t_i) = (W_{ik}(t_{i1}), \ldots, W_{ik}(t_{im_i})) \in \mathbb{R}^{L \times m_i}, i = 1, \ldots, n.$$

Writing $\hat{\gamma}_{1k}$ and $\hat{\gamma}_{2k}$ for the minimizer of (2.2), we define the B-spline estimator of $b_k$ by

$$\hat{b}_k(t) = \hat{\gamma}_{2k}^T B(t).$$

Given $\hat{b}_k, k = 1, \ldots, p$, we carry out the nonparametric independence screening and define the index set of selected covariates, denoted as $\hat{M}_\kappa$, by

$$\hat{M}_\kappa = \{k \geq 1 | \|\hat{b}_k\|_2^2 \geq C_{\kappa 3}n^{-2\kappa}L\}$$

for some sufficiently small positive constant $C_{\kappa 3}$ satisfying $C_{\kappa 3} < C_{\kappa 2}/2$.

Intuitively, we may still have too many irrelevant variables left in the analysis if the threshold parameter $\kappa$ in (2.5) is chosen too small. On the other hand, we run into the danger of screening out some of the relevant variables if it is too large. In the literature, the screening step is immediately followed by the model fitting step, thus usually some iterative screening procedure is employed to control the false selection rate [7, 21, 10, 3]. We avoid such time-consuming iterations by adding between these two steps a variable selection step, given in Section 3.1. The theory given in Sections 2.2 and 3.2 guarantees that, with proper choices of $\kappa$ and $L$, by the first screening step we can reduce the dimensionality to a moderate order, while controlling the false negative rate, and this allows the sparsity and oracle properties of the next variable selection step to hold. In practice, we sort the $\|\hat{b}_k\|_2^2$’s in the descending order and keep the first $[n^{\alpha}/\log n]$ variables, for some $2/5 \leq \alpha \leq 1$. In the numerical sections, $\alpha$ was taken as 1.

2.2. Theory of the proposed NIS procedure. Here we collect the technical assumptions on the marginal models given in (2.1). Let $I_d$ denote the identity matrix of order $d$ and $\#A$ be the number of the elements in a set $A$.

Assumption M1 There are positive constants $C_{M0}$ and $C_{M2}$ satisfying (1)-(3) in the following, for $k = 1, \ldots, p$.

1. $a_k$ and $b_k$ are twice continuously differentiable.
(2) \( \|a_k\|_\infty \leq C_{M0} \) and \( \|b_k\|_\infty \leq C_{M0} \).
(3) \( \|a_k''\|_\infty \leq C_{M2} \) and \( \|b_k''\|_\infty \leq C_{M2} \).

**Assumption M2** For \( \kappa \) given in \( M_k \), there are positive constants \( C_{\kappa_1} \) and \( C_{\kappa_2} \) such that (1)-(2) in the following hold and we also have (3) given below.

(1) If \( \|\beta_k\|^2 \geq C_{\kappa_1} n^{-2\kappa} L \), we have \( \|b_k\|^2 \geq C_{\kappa_2} n^{-2\kappa} L \).
(2) If \( \|\beta_k\|^2 < C_{\kappa_1} n^{-2\kappa} L \), we have \( \|b_k\|^2 = 0 \).
(3) \( n^{1-4\kappa} L/ \log n \to \infty, n^{-2\kappa} L = o(1) \), and \( L^{-3} = o(n^{-2\kappa}) \).

Assumption M1 is necessary in order to bound the approximation error to \( a_k \) and \( b_k \) by B-spline bases. Assumption M2 requires that the marginal models (2.1) still reflect the significance of relevant covariates, and similar assumptions are assumed in the NIS literature [3]. We mention that we have imposed Assumption M2(2) merely for simplicity of presentation.

**Assumption T**

(1) For some positive constant \( C_{T1} \), we have \( C_{T1} \leq E[\{\tilde{x}^{(k)}(t)\}]^2 \) for \( t \in [0, 1] \) and \( k = 1, \ldots, p \), where \( \tilde{x}^{(k)}(t) = x^{(k)}(t) - E\{x^{(k)}(t)\} \).
(2) For any positive constant \( C_1 \), there is a positive constant \( C_2 \) such that

\[
E[\exp\{C_1|\xi_i(t_i)|/\sqrt{m_i}\} | x_i(t_i), t_i] < C_2.
\]

(3) Let \( x_M(t) \) be the covariate vector consisting of 1 and all the covariates in \( M_k \). Then there is a positive constant \( C_{T2} \) such that

\[
E\{x_M(t)x_M(t)^T\} \geq C_{T2} I_{\#M_k+1} \text{ for any } t \in [0, 1].
\]

(4) For some positive constant \( C_{T3} \), \( \sup_{t \in [0, 1]} |x^{(k)}(t)| \leq C_{T3} \), \( k = 1, \ldots, p \).

(5) For some positive constant \( C_{T4} \), \( \sum_{k \in A} \sup_{t \in [0, 1]} |\beta_k(t)| \leq C_{T4} \), where \( A = \{k \mid 0 \leq k \leq p \text{ and } \sup_{t \in [0, 1]} |\beta_k(t)| > 0\} \).

(6) The functions \( \beta_k(t) \), \( k = 0, \ldots, p \), are twice continuously differentiable. Besides, \( \sum_{k \in A} \sup_t |\beta_k''(t)| \leq C_{T5} \) for some positive constant \( C_{T5} \).

Assumptions T(1) and T(4) imply that for some positive constants \( C_1 \) and \( C_2 \), we have

\[
(2.6) \quad C_1 I_2 \leq E \left( \frac{1}{x^{(k)}(t)} \right) x^{(k)}(t) \{x^{(k)}(t)\}^2 \leq C_2 I_2, \quad t \in [0, 1] \text{ and } 1 \leq k \leq p.
\]

Assumptions T(2), T(4) and T(5) are technical assumptions needed in order to apply the exponential inequalities. Note that \( \#A \) may increase, but we
assume that the signal $\sum_{k \in A} x_i^{(k)}(t) \beta_k(t)$ should not diverge by imposing Assumptions T(5) and T(6). Similar conditions are made in Assumption D of [3]. We can relax T(4) and T(5) slightly, for example we can replace $C_{T3}$ and $C_{T4}$ with $C_{T3} \log n$ and $C_{T4} \log n$ at the expense of multiplying the present convergence rate by $(\log n)^c$ for some positive $c$. We can also relax Assumption T(6) similarly with conformable changes made in the approximation error of the B-spline basis. We need Assumption T(1) and (2.6) for identifiability and estimation of the marginal models. Assumption T(3) is the identifiability condition of the coefficient functions in model (1.1). We use T(5) and T(6) to evaluate the approximation error of the B-spline basis when we consider the group SCAD variable selection discussed in Section 3.

**Theorem 2.1.** Suppose Assumptions M1, M2, and T(1)-(5) hold. Then,

$$P(\mathcal{M}_\kappa \subset \hat{\mathcal{M}}_\kappa) \geq 1 - C_{p1}pL \exp(-C_{p2}n^{1-4\kappa}L),$$

where $C_{p1}$ and $C_{p2}$ are positive constants and depend on $C_{\kappa j}$, $j = 1, 2, 3$, and the constants specified in the above mentioned assumptions.

Theorem 2.1 implies that all the relevant covariates will be selected with high probability, due to Assumption M2(1) and the uniform consistency. Specifically, when $p = O(n^{c_p})$ for any positive $c_p$ we have $P(\mathcal{M}_\kappa \subset \hat{\mathcal{M}}_\kappa) \to 1$, if $\kappa$ satisfies Assumption M2(3). Under the smoothness assumption M1(1), the optimal rate of $L$ is $L = c_L n^{1/5}$ for some positive $c_L$. In this case, Assumption M2(3) reduces to $n^{6/5-4\kappa}/\log n \to \infty$, $n^{1/5-2\kappa} = o(1)$, and $n^{2\kappa-3/5} = o(1)$. Then a sufficient condition on $\kappa$ is that

$$1/10 < \kappa < 3/10. \tag{2.7}$$

Thus, the proposed screening procedure may reduce the number of covariates drastically. However, $\# \hat{\mathcal{M}}_\kappa$ may be still too large to apply any variable selection procedures with consistency property. Fortunately we have succeeded in giving an upper bound on $\# \hat{\mathcal{M}}_\kappa$, as given in Theorem 2.2, which circumvents such situations. We emphasize that condition (2.7) is weaker than those in the literature: Theorem 1 of [9] requires that $n^{1-4\kappa}L^{-3} \to \infty$ which reduces to $\kappa < 1/10$ when $L$ is of the order $n^{1/5}$, and in Theorem 2 of [20] the condition on $\kappa$ implies $\kappa < 1/10$ as well. This improvement is crucial for us to obtain a tighter upper bound in Theorem 2.2, as compared to that in [9] (no upper bound is provided in [20]), which leads to (2.10). We succeeded in achieving this improvement by exploiting the band diagonal property of $\langle \mathbf{B}, \mathbf{B}^T \rangle_n$, $\langle \mathbf{B}, \mathbf{B}^T \rangle$, $\langle \mathbf{W}_k, \mathbf{W}_l \rangle_n$, $\langle \mathbf{W}_k, \mathbf{W}_l \rangle$, and so on.
In order to state Theorem 2.2, we need a little more notation. Define

\[ w_k = w_k - a_k b \quad \text{and} \quad w = (w_1^T, \ldots, w_p^T)^T. \]

where \( a_k = \langle w_k, b^T \rangle (b, b^T)^{-1} \). Note that \( w_k \) and \( w \) are respectively \( L \)- and \( pL \)-dimensional stochastic processes on \([0,1]\). Besides we define

\[ \Sigma = \langle w, w^T \rangle, \]

which is a \( pL \times pL \) matrix. We write \( \lambda_{\text{max}}(A) \) and \( \lambda_{\text{min}}(A) \) for the maximum and minimum eigenvalue of a symmetric matrix \( A \), respectively.

**Theorem 2.2.** Under the same assumptions as in Theorem 2.1 except \( M2(2) \), we have for some positive \( C_{\kappa 4} \),

\[ P(\#M_\kappa \leq C_{\kappa 4} n^{2\kappa} \lambda_{\text{max}}(\Sigma)) \geq 1 - C_{p1} pL \exp(-C_{p2} n^{1-4\kappa} L), \]

where \( C_{p1} \) and \( C_{p2} \) are the same constants as in Theorem 2.1.

Theorem 2.2 implies that, with high probability, the number of variables selected by our screening procedure will not be large. Note that it does not require Assumption \( M2(2) \). This means that, although some of the irrelevant covariates (with \( \|\beta_k\| \) small) may be included in \( \hat{M}_\kappa \) merely because they have large values of \( \|b_k\| \), the number of such selected irrelevant covariates is limited. Furthermore, they will be subsequently removed by the variable selection procedure given in Section 3.

Define \( \tilde{w}_k(t), \tilde{w}_k(t), \tilde{w}_{ik}(t_i) \) and \( \tilde{w}_{ik}(t_i) \) by replacing \( x^{(k)}(t) \) and \( x_i^{(k)}(t) \) in the definitions of \( w_k(t), w_k(t), w_{ik}(t) \) and \( w_{ik}(t_i) \) given in (2.3) with

\[ \tilde{x}^{(k)}(t) = x^{(k)}(t) - E \{ x^{(k)}(t) \} \quad \text{and} \quad \tilde{x}_i^{(k)}(t) = x_i^{(k)}(t) - E \{ x_i^{(k)}(t) \}, \]

respectively. It is easy to see, by properties of orthogonal projection, that

\[ \langle \tilde{w}, \tilde{w}^T \rangle \leq \Sigma \leq \langle w, w^T \rangle, \]

where \( \tilde{w} = (\tilde{w}_1^T, \ldots, \tilde{w}_p^T)^T \) and \( w = (w_1^T, \ldots, w_p^T)^T \). The maximum eigenvalue of \( \langle \tilde{w}, \tilde{w}^T \rangle \) may tend to infinity very quickly with \( p \). However, since we do a kind of centerization to \( w \) and obtain \( \tilde{w} \) as in (2.8), we conjecture that \( \Sigma \) is very close to \( \langle \tilde{w}, \tilde{w}^T \rangle \) under some regularity conditions.

If the maximum eigenvalues of the two matrices have the same order, and for some positive \( K_n, \lambda_{\text{max}}(E \{ \tilde{x}(t) \tilde{x}(t)^T \}) \leq K_n \) uniformly in \( t \), then

\[ \lambda_{\text{max}}(\Sigma) \leq C_1 L^{-1} K_n \quad \text{and} \quad \# M_\kappa \leq C_2 n^{2\kappa} L^{-1} K_n. \]
with probability tending to 1. Suppose $L$ is chosen to be of the optimal order $n^{1/5}$. Then, for $\kappa$ satisfying condition (2.7), this implies that

$\# \hat{M} = O_p(n^{2/5-\eta})K_n$ for some $0 < \eta < 2/5$.

Thus, when $K_n$ is bounded, $\# \hat{M}$ fulfills the requirement (3.4) on $q$ in the subsequent variable selection step given in Section 3. In addition, if we choose a smaller value of $\kappa$ we can further allow a moderately increasing $K_n$.

3. Variable selection and structure identification. We can remove a lot of irrelevant covariates by the NIS procedure given in Section 2. However, it does not have the consistency property in selecting the important variables. In this section, we propose a group SCAD estimator for variable selection and structure identification, and establish its consistency, sparsity and oracle properties. Here, we denote the number of covariates by $q$, instead of $p$ as in Section 2. This distinction is necessary. When the dimensionality $p$ is very large, we have to employ some screening procedure before we can carry out any variable selection procedure. In this case, $p$ and $q$ are respectively the number of variables before and after the screening procedure is applied. For simplicity of notation, we still denote as $x^{(1)}, \ldots, x^{(q)}$ the variables selected by the NIS algorithm. When the $p$ is not very large, we can simply take $q = p$ and proceed directly with the group SCAD procedure.

3.1. Group SCAD procedure. Suppose we are given $y_i(t), x_i^{(k)}(t), i = 1, \ldots, n, k = 1, \ldots, q$ and consider the varying coefficient model (1.3). To estimate the coefficient functions $\beta_k(t), k = 0, 1, \ldots, q$, first we define

$$l_q(\gamma) = \|y - \gamma^T_0 B - \sum_{k=1}^{q} \gamma^T_k W_k\|^2_n,$$

where $\gamma = (\gamma^T_0, \ldots, \gamma^T_q)^T \in \mathbb{R}^{(q+1)L}$. When $q$ is fixed and sufficiently small, based on working independence, we can estimate $\beta_k(t)$ by minimizing the objective function $l_q(\gamma)$. Denoting the minimizer by $\tilde{\gamma} = (\tilde{\gamma}^T_0, \ldots, \tilde{\gamma}^T_q)^T$, for $k = 0, 1, \ldots, q$, we estimate $\beta_k(t)$ with

$$\tilde{\beta}_k(t) = \tilde{\gamma}^T_k B(t).$$

Suppose $q$ satisfies

$$q = o\left(\sqrt{n/(L \log n)}\right).$$

The restriction is necessary since the Hessian matrix of the objective function $l_q(\gamma)$ given in (3.1) must be positive definite. Note that [30] imposed similar
conditions in the case where \( q \) is a fixed constant. Recall that \( L \) is the dimension of the B-spline basis. When it is taken as the optimal order \( L = c_L n^{1/5} \) for some positive \( c_L \), condition (3.3) reduces to

\[
q = o(n^{2/5}(\log n)^{-1/2}).
\]

When \( q \) is relatively large and a lot of the covariates seem to be irrelevant, we would add a penalty term to \( l_q(\gamma) \) given in (3.1), such as the group SCAD or the adaptive group Lasso penalty, and then conduct variable selection and estimation simultaneously. After the variable selection step, if necessary, we can estimate the coefficient functions of the selected variables again without the penalty term. Besides, we are also interested in structure identification of the model. That is, some of the coefficient functions \( \beta_k \) associated with the relevant variables may be constant while the others are time-varying. As mentioned in Section 1, when there is no a priori knowledge on which of the coefficient functions are indeed constant, treating the constant coefficients as time-varying would result in a loss in the convergence rate. Thus an important issue is to identify them based on data. To this end, we can add another penalty term to regularize the estimated coefficient functions. A similar kind of penalty term was used in [36] for Cox proportional hazard models with time-varying coefficients.

Now we define our group SCAD penalty term for simultaneous variable selection and structure identification. First, we introduce an orthogonal decomposition of \( g_k(t) \equiv \gamma_k B(t) \) with respect to the \( L_2 \) norm by

\[
g_k(t) = (g_k)_c + (g_k)_f(t),
\]

where \( (g_k)_c = \int_0^1 g_k(t)dt \) and \( (g_k)_f(t) = g_k(t) - (g_k)_c \). Then, we have \( \|g_k\|_{L_2}^2 = \|(g_k)_c\|^2 + \|(g_k)_f\|_{L_2}^2 \). Let \( p_\lambda(\cdot) \) be the SCAD function given by

\[
p_\lambda(u) = \begin{cases} 
\lambda u & \text{if } 0 \leq u \leq \lambda \\
-(u^2 - 2a_0\lambda u + \lambda^2)/\{2(a_0 - 1)\} & \text{if } \lambda < u \leq a_0\lambda \\
(a_0 + 1)\lambda^2/2 & \text{if } u > a_0\lambda
\end{cases},
\]

where \( a_0 \) is a constant larger than 1. We take \( a_0 = 3.7 \) as suggested by [6]. Our group SCAD penalty is defined by \( \sum_{k=1}^q \{p_{\lambda_1}(\|(g_k)_c\|) + p_{\lambda_2}(\|(g_k)_f\|_{L_2})\} \), where \( g_k = \gamma_k B, k = 0, \ldots, q \). We specify \( \lambda_1 \) and \( \lambda_2 \) later. Our objective function for our group SCAD estimator is then given by

\[
Q_q(\gamma) = l_q(\gamma) + \sum_{k=1}^q \{p_{\lambda_1}(\|(g_k)_c\|) + p_{\lambda_2}(\|(g_k)_f\|_{L_2})\}.
\]
Based on $Q_q(\gamma)$, we can carry out variable selection, structure identification, and estimation simultaneously by the following procedure:

$$(3.8) \quad \hat{\gamma} = (\hat{\gamma}_0^T, \ldots, \hat{\gamma}_q^T)^T = \arg\min_{\gamma \in \mathbb{R}^{(q+1)L}} Q_p(\gamma) \quad \text{and} \quad \hat{\beta}_k = \hat{\gamma}_k^T B, \ k = 0, \ldots, q.$$ 

Then we can choose the significant covariates as those $x^{(k)}$ with $\|\hat{\beta}_k\|_{L_2} > 0$ and identify among the selected variables the constant components consistently by the criterion $\|\hat{\beta}_k\|_{L_2} = 0$. We call $\hat{\beta}_k$ the group SCAD estimator.

To compute the group SCAD estimator, we use the approximation to the SCAD function suggested in [6]: $p_\lambda(u) \approx p_\lambda(u_0) + \frac{1}{2}(p'_\lambda(u_0)/u_0)(u^2 - u_0^2)$, for $u$ in a neighborhood of any given $u_0 \in \mathbb{R}^+$. Define $\tau_j = \frac{1}{2} \int_0^1 B_j(t)dt$, and $\overline{B}_j = \sqrt{n}(B_j - \tau_j B_1)$, $j = 0, 1, \ldots, L$, where $\{B_0(t), B_1(t), \ldots, B_L(t)\}$ is the B-spline basis on $[0, 1]$. Then it will be convenient to use the new basis $(1, \overline{B}_2, \ldots, \overline{B}_L)$ when we calculate the penalty term. The number of covariates $q$ after the NIS step should be small enough to calculate the least squares estimates, which can be used as the initial estimates in the iterative algorithm for finding $\hat{\gamma}$. To select the tuning parameters $\lambda_1$ and $\lambda_2$ in (3.7), we treat them as equal $\lambda_1 = \lambda_2 = \lambda$ and then use the BIC criterion to select $\lambda$: $\text{BIC}(\lambda) = \|y - \hat{\beta}_0(\lambda) - \sum_{k=1}^q \hat{\beta}_k(\lambda)x^{(k)}\|^2_n + K \log N$, where $\hat{\beta}_k(\lambda)$ is the group SCAD estimate based on $\lambda$, $K$ is the number of parameters in the fitted model, and $N = \sum_{i=1}^n m_i$. A similar BIC criterion was suggested by [28], and a generalized information criterion was considered by [13] for tuning parameter selection in penalized likelihood models.

3.2. Asymptotic properties of the group SCAD procedure. In this section we state the consistency, sparsity, and oracle properties of the proposed group SCAD estimator. The proofs of the theorems are deferred to Appendix A. First, we state the sparsity assumption. We can relax Assumption S(2) in some sense. See Remark 1 for more details.

**Assumption S**

1. There is a positive integer $s < q$ such that the following holds:
   - for $k = 1, \ldots, s$, $|(\beta_k)_c|/\lambda_1 \to \infty$ if $|\beta_k| > 0$ and $\|\beta_k\|_{L_2}/\lambda_2 \to \infty$ if $\|\beta_k\|_{L_2} > 0$;
   - for $k = s + 1, \ldots, q$, $\|\beta_k\|_{L_2} = 0$;
   - the above convergence is uniform in $k = 0, \ldots, s$.

2. $\lambda_1/r_{qn} \to \infty$ and $\lambda_2/r_{qn} \to \infty$, where $r_{qn}$ is defined in (3.9).

We define the spline estimation space, denoted as $\mathbf{G}$, by

$$\mathbf{G} = \{\mathbf{g} = (g_0, \ldots, g_q)^T | g_k = \gamma_k^T B, \ k = 0, 1, \ldots, q\}$$
and $G_0$, which we may call the oracle space under Assumption S, by

$$G_0 = \{ g \in G \mid (g_k)_c = 0 \text{ if } |(\beta_k)_c| = 0 \} \text{ and } (g_k)_f = 0 \text{ if } |(\beta_k)_f|_{L_2} = 0, \ k = 1, \ldots, q \}.$$  

We introduce two norms on $G$ here. For $g = (g_0, g_1, \ldots, g_q)^T \in G$, define $\|g\|_{L_2}^2 = \sum_{k=0}^q \|g_k\|_{L_2}^2$ and $\|g\|_{\infty} = \sum_{k=0}^q \|g_k\|_{\infty}$. The approximation error of spline functions to $\beta = (\beta_0, \ldots, \beta_q)^T$, denoted as $\rho_{\text{qn}}$, affects the convergence rates of the least squares and the group SCAD estimators, and we define it by $\rho_{\text{qn}} = \sup_\beta \inf_{g \in G} \|\beta - g\|_{\infty}$, where $\beta$ is over the functions satisfying Assumption T(5)-(6). Corollary 6.26 of [24] and Assumption T(5)-(6) imply that $\rho_{\text{qn}} \leq C_\rho L^{-2}$ for some positive constant $C_\rho$. Before we state Theorems 3.1-3.3, we define the convergence rates of the least squares and the group SCAD estimators, respectively denoted as $r_{\text{qn}}$ and $r_{\text{sn}}$, by

$$(3.9) \quad r_{\text{qn}} = \max\{ (qL/n)^{1/2}, \rho_{\text{qn}} \} \quad \text{and} \quad r_{\text{sn}} = \max\{ (sL/n)^{1/2}, \rho_{\text{qn}} \} ,$$

where $s$, defined in Assumption S, is the number of relevant variables.

We state two technical assumptions here. Set

$$(3.10) \quad \Sigma_n = (B^T W^T)^T, (B^T W^T)_{n} \text{ and } \Sigma = E\{ \Sigma_n \},$$

where $W = (W_1^T, \ldots, W_q^T)$. A sufficient condition for Assumption V is $\lambda_{\min} E\{x(t)x(t)^T\} \geq C$ uniformly in $t$ for some positive $C$.

**Assumption E** There is a positive constant $C_E$ such that uniformly in $t$, $E\{\epsilon_i(t_i)\epsilon_i(t_i)^T \mid x_i(t_i), t_i\} \leq C_E I_{m_i}$.

**Assumption V** There is a positive constant $C_V$ such that $\lambda_{\min}(\Sigma) \geq C_V / L$.

In Theorem 3.2, we derive the $L_2$ convergence rate and establish some desirable properties of the group SCAD estimator given in (3.8), under the sparsity assumption (Assumption S). Before that, in Theorem 3.1 we deal with the $L_2$ convergence of the B-spline estimator given in (3.2).

**Theorem 3.1.** Suppose that Assumptions T(4)-(6), V, and E hold. Then

$$(\hat{\gamma}_0^T B, \ldots, \hat{\gamma}_q^T B)^T - (\beta_0, \ldots, \beta_q)^T \parallel^2_{L_2} = \sum_{k=0}^q \|\hat{\gamma}_k^T B - \beta_k\|^2_{L_2} = O_p(r_{\text{qn}}^2).$$

**Theorem 3.2.** Suppose that Assumptions T(4)-(6), V, E, and S hold. Then with probability tending to 1, there exists a local minimizer of $Q_q(\gamma)$ on $\mathbb{R}^{(q+1)L}$, denoted by $\hat{\gamma} = (\hat{\gamma}_0^T, \ldots, \hat{\gamma}_q^T)^T$, such that

$$(\hat{\gamma}_0^T B, \ldots, \hat{\gamma}_q^T B)^T - (\beta_0, \ldots, \beta_q)^T \parallel^2_{L_2} = \sum_{k=0}^q \|\hat{\gamma}_k^T B - \beta_k\|^2_{L_2} = O_p(r_{\text{qn}}^2).$$
Next, we define the sparsity and the oracle properties of estimators.

**Sparsity Property.** Suppose that Assumption S(1) holds. Then if an estimator $\mathbf{g} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_q)^T$ of $(\beta_0, \ldots, \beta_q)^T$ satisfies the conditions below with probability tending to 1, we say that $\mathbf{g}$ has the sparsity property.

1. For $k = 0, \ldots, s$: $|\widehat{\beta}_k| > 0$ if and only if $|\beta_k| > 0$, and $\|\mathbf{g}_k\|_{L_2} > 0$ if and only if $\|\beta_k\|_{L_2} > 0$.
2. For $k = s + 1, \ldots, p$: $\|\mathbf{g}_k\|_{L_2} = 0$.

**Oracle Property.** If we knew the value of $s$ in Assumption S(1), we would use the knowledge and minimize $l_q(\gamma)$ on the subspace of $\mathbb{R}^{q+1}L$ corresponding to $G_0$. We call this imaginary estimator the oracle estimator. We say that an estimator has the oracle property if it is asymptotically equivalent to this oracle estimator.

Theorem 3.3 is about the sparsity property and the oracle property of the group SCAD estimator defined in (3.8). Note that the existence of the local solution in Theorem 3.3 is established in Theorem 3.2.

**Theorem 3.3.** Suppose that Assumptions T(4)-(6), V, E, and S hold. Let $\{\eta_n\}$ be a sequence of positive numbers satisfying $\eta_n \to \infty$, $\lambda_1/(\eta_n r_{qn}) \to \infty$, and $\lambda_2/(\eta_n r_{qn}) \to \infty$. Then, with probability tending to 1, any local minimizer $\mathbf{g} = (\widehat{\gamma}_0^T, \ldots, \widehat{\gamma}_q^T)^T$ of $Q_q(\gamma)$ satisfying $\|Q_q(\beta_0, \ldots, \beta_q)^T - (\beta_0, \ldots, \beta_q)^T\|_{L_2} \leq \eta_n r_{qn}$ is equal to the oracle estimator. We also have

$$
(\widehat{\gamma}_0^T, \ldots, \widehat{\gamma}_q^T)^T \in G_0 \quad \text{and} \quad \sum_{k=0}^s \|\widehat{\gamma}_k^T - \beta_k\|_{L_2}^2 = O_p(r_{sn}^2).
$$

Since $Q_q(\gamma)$ may not be concave, there may be another local minimizer of $Q_q(\gamma)$ outside $\{\gamma \in \mathbb{R}^{q+1}L \mid \|Q_q(\beta_0, \ldots, \beta_q)^T - (\beta_0, \ldots, \beta_q)^T\|_{L_2} \leq \eta_n r_{qn}\}$.

**Remark 1.** Assumption S(2) may be restrictive when $q$ is large compared to $s$, for example, $q = c_n n^{2/5}/\sqrt{\log n}$ with $c_n \to 0$ slowly, $L = c_n n^{1/5}$, and $s$ bounded. Thus it would be desirable if we could replace $r_{qn}$ in the denominators with some quantity independent of $q$. This is possible in some sense and here we give an example. Consider only variable selection, not structure identification. Then the penalty term is given by $\sum_{j=0}^q p_\lambda(\|g_j\|_{L_2})$ and we assume that $\lambda/\max\{\sqrt{s}r_{sn}, L^{-3/2}\} \to \infty$. We also need Assumption T(2) to employ exponential inequalities and denote the global minimizer of $l_q(\gamma)$ on $\mathbb{R}^{q+1}L$ by $\widehat{\gamma} \in \mathbb{R}^{q+1}L$. Then $(\widehat{\gamma}^T, 0^T)^T \in \mathbb{R}^{q+1}L$ is a local minimizer of $Q_q(\overline{\gamma})$, where $\overline{\gamma} \in \mathbb{R}^{q+1}L$, with probability tending to 1. This implies that some flexibility will be allowed in tuning parameter selection when $s$ is bounded. Its proof is outlined in Supplement A.
4. Refinement of the group SCAD estimator. To ease the notation, without loss of generality, denote respectively the constant coefficients and the corresponding variables by $\beta_1 \in \mathbb{R}^{s_1}$ and $x_1(t) \in \mathbb{R}^{s_1}$, and denote respectively the coefficients and the corresponding variables by $\beta_2(t) \in \mathbb{R}^{s_2}$ and $x_2(t) \in \mathbb{R}^{s_2}$. Then, we can rewrite model (1.4) as the following:

\begin{align}
(4.1) \quad y_i(t_{ij}) = \beta_0(t_{ij}) + x_{1i}(t_{ij})^T \beta_1 + x_{2i}(t_{ij})^T \beta_2(t_{ij}) + \epsilon_i(t_{ij}), \quad j = 1, \ldots, m_i,
\end{align}

where $x_{1i}(t_{ij})$ and $x_{2i}(t_{ij})$ denote respectively the observations on $x_1(t)$ and $x_2(t)$ in the $i$-th subject at time $t_{ij}$. When $x_1$ and $x_2$ are given, and $s_1$ and $s_2$ are fixed and small, this estimation problem has been extensively studied in the literature [39, 18]. We revisit this problem to provide a practical procedure when we encounter ultra-high or large dimensionality and we do not have a priori knowledge of the relevant variables, nor which of them have constant coefficients. First, there is room to improve the group SCAD estimator given in (3.8). One reason is that it uses working independence, which does not hold for longitudinal data in general. Another reason is that B-spline smoothing suffers from boundary effects. In model (3.8), the selected variables are divided into two groups. The first group $x_1$ are those having constant coefficients with $|\hat{\beta}_k| > 0$ and $\|\hat{\beta}_k\|_2 = 0$; the second group $x_2$ are those having time-varying coefficients with $\|\hat{\beta}_k\|_2 > 0$. Note that when $|\hat{\beta}_k| = 0$ and $\|\hat{\beta}_k\|_2 > 0$ the constant part is zero, but we still include the variable in $x_2$ without such a constraint on $\beta_k(t)$.

Our estimation procedure for the coefficients in (4.1) consists of three steps: (i) constructing initial estimators, (ii) estimating the covariance function of the error process, and (iii) estimating the coefficients based on the covariance estimate, which are detailed in the following sections. Alternatively, after the initial coefficient estimates in Section 4.1 are obtained, we may also iterate between the covariance function estimation in Section 4.2 and the coefficients estimation in Section 4.3 until convergence.

4.1. Initial coefficients estimation. We could use the group SCAD estimator (3.8) as initial estimator for the coefficients in model (4.1). However, it may suffer from boundary effects, and the following profile least squares estimator is preferred [4, 16]. Recall that $t_i = (t_{i1}, \ldots, t_{im_i})$, $y_i(t_i) = (y_i(t_{i1}), \ldots, y_i(t_{im_i}))$, and $\epsilon_i(t_i) = (\epsilon_i(t_{i1}), \ldots, \epsilon_i(t_{im_i}))$. Let $K$ denote a kernel function, which is usually taken as a symmetric pdf, and take a bandwidth $h_1 > 0$. Given $\beta_1 \in \mathbb{R}^{s_1}$, we can estimate $\beta_0(t)$ and $\beta_2(t)$ in model
by minimizing the following local sum of squares:

(4.2) \[ \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{ y_i(t_{ij}) - x_{1i}(t_{ij})^T \beta_1 - (1, x_{2i}(t_{ij})^T) (\alpha_0 + \alpha_1(t_{ij} - t)) \}^2 \times K_{h_1}(t_{ij} - t) \]

= \sum_{i=1}^{n} \{ y_i(t_i)^T - x_{1i}(t_i)^T \beta_1 - (1, x_{2i}(t_i)^T, T_i(t)(1_{m_i}, x_{2i}(t_i)^T))^T = \sum_{i=1}^{n} \{ y_i(t_i)^T - x_{1i}(t_i)^T \beta_1 - (1, x_{2i}(t_i)^T, T_i(t)(1_{m_i}, x_{2i}(t_i)^T))^T \alpha \}

where \( K_{h_1}(\cdot) = K(\cdot/h_1) / h_1, \alpha_0, \alpha_1 \in \mathbb{R}^{s_2 + 1}, 1_{m_i} \) is the \( m_i \)-dimensional one-vector, \( T_i = \text{diag}\{ t_1, \ldots, t_{im_i} - t \}, W_{dh_i}(t) = \text{diag}\{ K_{h_1}(t_1 - t), \ldots, K_{h_1}(t_{im_i} - t) \} \), and \( \alpha = (\alpha_0^T, \alpha_1^T)^T \). Let \( 0_{k \times l} \) be the \( k \times l \) dimensional zero matrix. For any given \( \beta_1 \in \mathbb{R}^{s_1} \), denote the minimizer of (4.2) by \( \tilde{\alpha}(t, \beta_1) \). Then, an estimator of \( (\beta_0(t), \beta_2(t)^T)^T \) is

(4.3) \[ (\tilde{\beta}_0(t, \beta_1), \tilde{\beta}_2(t, \beta_1)^T)^T = (I_{s_2 + 1}, 0_{(s_2 + 1) \times (s_2 + 1)}) (\tilde{\alpha}(t, \beta_1) = (I_{s_2 + 1}, 0_{(s_2 + 1) \times (s_2 + 1)}) (V(t)^T W_{h_1}(t)(V(t)))^{-1} V(t)^T W_{h_1}(t)(Y - X_1 \beta_1), \]

where \( Y = (y_1(t_1), \ldots, y_1(t_n))^T, X_1 = (x_{11}(t_1), \ldots, x_{1n}(t_n))^T \), and \( V(t) = (V_{11}(t), \ldots, V_{1n}(t))^T \) with \( V_{ih} = (1_{m_i}, x_{2i}(t_i)^T, T_i(t)(1_{m_i}, x_{2i}(t_i)^T), i = 1, \ldots, n. \)

Based on working independence, the initial profile least squares estimator for the constant coefficients \( \beta_1 \) in model (4.1) is defined as

\[ \tilde{\beta}_1^{PLS} = \arg\min_{\beta_1 \in \mathbb{R}^{s_1}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{ y_i(t_{ij}) - x_{1i}(t_{ij})^T \beta_1 - \tilde{\beta}_0(t_{ij}, \beta_1) - x_{2i}(t_{ij})^T \tilde{\beta}_2(t_{ij}, \beta_1) \}^2, \]

and the initial estimator for \( (\beta_0(t), \beta_2(t)^T)^T \) is defined as \( (\tilde{\beta}_0^{PLS}(t), \tilde{\beta}_2^{PLS}(t)^T)^T = (\beta_0(t), \tilde{\beta}_1^{PLS}) \), \( \beta_2(t, \beta_1^{PLS})^T \). Note that \( \beta_1^{PLS} \) can be written as

\[ \tilde{\beta}_1^{PLS} = \arg\min_{\beta_1 \in \mathbb{R}^{s_1}} \left\{ (T - S)(Y - X_1 \beta_1) \right\}^T (T - S)(Y - X_1 \beta_1), \]

where \( S = (S_{11}, \ldots, S_{1m_1}, \ldots, S_{n1}, \ldots, S_{nm_n})^T, S_{ij} = (1, x_{2i}(t_i)^T, 0_{1 \times (s_2 + 1)}) \)

\( (V(t_{ij})^T W_{h_1}(t)(V(t_{ij}))^{-1} V(t_{ij})^T W_{h_1}(t_{ij}), j = 1, \ldots, m_i, i = 1, \ldots, n, \) and \( T = I_{m_i} \ldots + m_n \). Thus, we have

(4.4) \[ \tilde{\beta}_1^{PLS} = \left\{ X_1^T (T - S)^T (T - S) X_1 \right\}^{-1} X_1^T (T - S)^T (I - S)(Y, \]

and, from the definition of \( (\tilde{\beta}_0^{PLS}(t), \tilde{\beta}_2^{PLS}(t)^T)^T \) and (4.3), we have

(4.5) \[ (\tilde{\beta}_0^{PLS}(t), \tilde{\beta}_2^{PLS}(t)^T)^T = (I_{s_2 + 1}, 0_{(s_2 + 1) \times (s_2 + 1)}) (V(t)^T W_{h_1}(t)(V(t)))^{-1} V(t)^T W_{h_1}(t)(Y - X_1 \tilde{\beta}_1^{PLS}). \]
To select the bandwidth $h_1$ in (4.4) and (4.5), we choose the value of $h_1$ that minimizes the leave-one-subject-out cross-validation function.

It is well known that the working independence estimator $\tilde{\beta}_1^{PLS}$ is not semiparametric efficient when the error process is dependent [32]. In the following sections we estimate the covariance function of the error process using residuals obtained from the initial estimators $\hat{\beta}_1^{PLS}$, $\hat{\beta}_0^{PLS}(t)$ and $\hat{\beta}_2^{PLS}(t)$, and then construct semiparametric efficient estimators. The semiparametric efficiency results in [18] concern generalized partially linear models and carry over to the considered seminarying coefficient models.

4.2. Estimation of covariance function of the error process. Denote the covariance function of the error process $\epsilon(t)$ by $\phi(u,v) = \text{Cov}(\epsilon(u),\epsilon(v))$, $u,v \in [0,1]$, and assume that $\epsilon(t)$ consists of two independent components:

$$\epsilon(t) = \epsilon_1(t) + \epsilon_2(t),$$

where $\epsilon_1(t)$ has a smooth covariance function $\psi(s,t)$ and $\epsilon_2(t)$ models the measurement error process. Write the residuals obtained from the initial profile least squares estimators $\hat{\beta}_1^{PLS}$ and $\hat{\beta}_2^{PLS}(t)$ given in Section 4.1 as:

$$\hat{\epsilon}_{ij} = y(t_{ij}) - x_1i(t_{ij})^T \hat{\beta}_1^{PLS} - \hat{\beta}_0^{PLS}(t_{ij}) - x_2i(t_{ij})^T \hat{\beta}_2^{PLS}(t_{ij}),$$

$i = 1,\ldots,n$, $j = 1,\ldots,m_i$. We can estimate $\phi$ based on these residuals. There exist (semi-)parametric approaches to covariance estimation for longitudinal data [5, 11]. Such methods will be efficient when the parametric assumptions hold, but can suffer from large biases otherwise. In general, we may not have knowledge about the complicated covariance structure and we can use nonparametric methods to avoid this problem.

Specifically, we use the nonparametric method of [14] to estimate the covariance function $\phi$ based on the residuals. First, noting that

$$\phi(t_{ij},t_{ik}) = \psi(t_{ij},t_{ik}) + \text{Var}(\epsilon_2(t_{ij}))[I(t_{ij} = t_{ik}), i = 1,\ldots,n, j,k = 1,\ldots,m_i],$$

we can estimate $\psi(u,v)$ by $\tilde{\psi}(u,v) = \tilde{a}(u,v)$ where $\tilde{a}(u,v)$ is the first element of $\tilde{a}(u,v)$ which minimizes the following local sum of squares:

$$\sum_{i=1}^{n} \sum_{j \neq k} \{\epsilon_{ij}\epsilon_{ik} - a - b(t_{ij} - u) - c(t_{ik} - v)\}^2 K_{h_2}(t_{ij} - u)K_{h_2}(t_{ik} - v),$$

with a bandwidth $h_2 > 0$. An explicit formulae for $\tilde{\psi}(u,v)$ is available [14]. The covariance function estimate $\tilde{\psi}(u,v)$ is not positive semidefinite.
in general. We can modify this estimate by truncating the negative components in its spectral decomposition \( \hat{\psi}(u, v) = \sum_{k=1}^{\infty} \hat{\omega}_k \hat{\psi}_k(u) \hat{\psi}_k(v) \), where \( \hat{\omega}_1 \geq \hat{\omega}_2 \geq \ldots \) are the eigenvalues of the operator \( \hat{\psi} \), given by \( (\hat{\psi} \alpha)(u) = \int_{[0,1]} \alpha(v) \hat{\psi}(u, v) dv \) for \( \alpha \in L_2([0,1]) \). \( \hat{\psi}_j \) is the eigenfunction corresponding to \( \hat{\omega}_j \), \( j = 1, 2, \ldots \), and \( \hat{\omega}_n = \max \{ k : \hat{\omega}_j > 0, j = 1, \ldots, k \} \).

Next, we estimate the variance function of the error process \( \epsilon(t) \): \( \sigma^2(t) \equiv \text{Var}(\epsilon_2(t)) \) by \( \hat{\sigma}^2(t) \equiv \hat{a} \) where \( \hat{a} \) is defined by

\[
(4.8) \quad (\hat{a}, \hat{b})^T = \arg \min_{(a, b)^T \in \mathbb{R}^2} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (\hat{\epsilon}_{ij}^2 - a - b(t_{ij} - t))^2 \hat{K}_{h_3}(t_{ij} - t),
\]

with \( h_3 > 0 \). Then an estimator for \( \phi(u, v) \) is defined as

\[
(4.9) \quad \hat{\phi}(u, v) = \hat{\psi}(u, v) I(u \neq v) + \hat{\sigma}^2(u) I(u = v).
\]

To select the bandwidths \( h_2 \) and \( h_3 \) in (4.7) and (4.8) we employ the leave-one-subject-out cross-validation.

4.3. Model estimation accounting for dependent errors. In this section we construct semiparametric efficient estimators for the constant and varying coefficient functions \( \beta_1 \) and \( \beta_2(t) \) using \( \hat{\phi}(u, v) \) given in (4.9). Let \( \hat{A}_i = (\hat{\phi}(t_{ij}, t_{ik}))_{j,k=1,\ldots,m_i}, i = 1, \ldots, n \). For any \( \beta_1 \in \mathbb{R}^{s_1} \), define \( \hat{\alpha}(t, \beta_1) \) as the minimizer of the following objective function of \( \alpha \in \mathbb{R}^{2(s_2+1)} \):

\[
\sum_{i=1}^{n} \{ y_i(t_i)^T - x_{1i}(t_{ij})^T \beta_1 - (1_m, x_{2i}(t_i)^T, T_i(t)(1_m, x_{2i}(t_i)^T)) \alpha \}^T \hat{A}_i^{-1/2} W_{ih_1}(t) \hat{A}_i^{-1/2} \{ y_i(t_i) - x_{1i}(t_{ij})^T \beta_1 - (1_m, x_{2i}(t_i)^T, T_i(t)(1_m, x_{2i}(t_i)^T)) \alpha \}.
\]

Then, given \( \beta_1 \in \mathbb{R}^{s_1} \), an estimator for \( (\beta_0(t), \beta_2(t)^T) \) is taken as

\[
(4.10) \quad ((\hat{\beta}_0(t, \beta_1), \hat{\beta}_2(t, \beta_1)^T)^T = (1_{s_2+1}, 0_{(s_2+1) \times (s_2+1)}) \hat{\alpha}(t, \beta_1) = (1_{s_2+1}, 0_{(s_2+1) \times (s_2+1)}) (\hat{V}(t)^T \hat{W}_{h_1}(t) \hat{V}(t))^{-1} \hat{V}(t)^T \hat{W}_{h_1}(t) \hat{A}_i^{-1/2} (y - X_1 \beta_1),
\]

where \( \hat{A}_i^{-1/2} = \text{diag}\{ \hat{A}_1^{-1/2}, \ldots, \hat{A}_n^{-1/2} \} \), and \( \hat{V}(t) = (\hat{V}_{1t}, \ldots, \hat{V}_{nt})^T \) with \( \hat{V}_{it} = (\hat{A}_1^{-1/2}(1_m, x_{2i}(t_i)^T), \hat{A}_i^{-1/2} T_i(t)(1_m, x_{2i}(t_i)^T)) \), \( i = 1, \ldots, n \).

The profile least squares estimator for \( \beta_1 \), denoted by \( \hat{\beta}_1^{PLS} \), accounting for within-subject correlation, is defined as the minimizer of the following objective function of \( \beta_1 \in \mathbb{R}^{s_1} \):

\[
\sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=1}^{m_i} \{ y_i(t_{ij}) - x_{1i}(t_{ij})^T \beta_1 - \hat{\beta}_0(t_{ij}, \beta_1) - x_{2i}(t_{ij})^T \hat{\beta}_2(t_{ij}, \beta_1) \}^T \hat{A}_i^{-1}(j, k) \{ y_i(t_{ik}) - x_{1i}(t_{ik})^T \beta_1 - \hat{\beta}_0(t_{ij}, \beta_1) - x_{2i}(t_{ij})^T \hat{\beta}_2(t_{ij}, \beta_1) \}.
\]
Then the corresponding estimator for $(\beta_0(t), \beta_2(t))^T$ is as in (4.10) with $\beta_1$ replaced by $\hat{\beta}^{PLS}_1$. We call these the refined estimators. Rewrite $\hat{\beta}^{PLS}_1$ as

$$\hat{\beta}^{PLS}_1 = \arg\min_{\beta_1 \in \mathbb{R}^n} \{(\mathcal{I} - \hat{\mathcal{S}})(Y - X_1\beta_1)\}^T \hat{\Lambda}^{-1} \{(\mathcal{I} - \hat{\mathcal{S}})(Y - X_1\beta_1)\},$$

where $\hat{\Lambda}^{-1} = \text{diag}\{\hat{\Lambda}_1^{-1}, \ldots, \hat{\Lambda}_n^{-1}\}$, $\hat{\mathcal{S}} = (\hat{S}_{11}, \ldots, \hat{S}_{im_1}, \ldots, \hat{S}_{n1}, \ldots, \hat{S}_{nnm_n})^T$, $\hat{S}_{ij} = (1, x_{2i}(t_{ij})^T, 0_{1 \times (s_2 + 1)})(\hat{V}(t_{ij})^T W_{h_1}(t_{ij}) \hat{V}(t_{ij}))^{-1} \hat{V}(t_{ij})^T W_{h_1}(t_{ij}) \hat{\Lambda}^{-1/2}$, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. Then, we have

$$\hat{\beta}^{PLS}_1 = \{X_1^T(\mathcal{I} - \hat{\mathcal{S}})^T \hat{\Lambda}^{-1}(\mathcal{I} - \hat{\mathcal{S}})X_1\}^{-1} X_1^T(\mathcal{I} - \hat{\mathcal{S}})^T \hat{\Lambda}^{-1}(\mathcal{I} - \hat{\mathcal{S}})Y,$$

and it follows from the definition of $(\hat{\beta}^{PLS}_0(t), \hat{\beta}^{PLS}_2(t))^T$ and (4.10) that

$$\begin{align*}
(\hat{\beta}^{PLS}_0(t), \hat{\beta}^{PLS}_2(t))^T &= (I_{s_2 + 1}, 0_{(s_2 + 1) \times (s_2 + 1)}) \\
&\times (\hat{V}(t)^T W_{h_1}(t) \hat{V}(t))^{-1} \hat{V}(t)^T W_{h_1}(t) \hat{\Lambda}^{-1/2}(Y - X_1\hat{\beta}^{PLS}_1).
\end{align*}$$

To select the bandwidth $h_1$ in (4.11) and (4.12), we choose the value of $h_1$ that minimizes the leave-one-subject-out cross-validation function.

5. Numerical studies.

5.1. Simulations. In our simulation study summarized in this section, the data were generated from model (4.1), where each $t_i$ is a vector of $m$ equally-spaced grid points on $[0, 1]$. We considered three types of coefficients settings:

Case I: $\beta_0(t) = 3.5 \sin(2\pi t)$, $s_1 = 2$, $s_2 = 3$, $\beta_1 = (5, -5)^T$ and $\beta_2(t) = (5(1-t)^2, 3.5(\exp(-(3t-1)^2)+\exp(-(4t-3)^2))-1.5, 3.5t^{1/2})^T$.

Case II: $\beta_0(t) = 3.5 \sin(2\pi t)$, $s_1 = 0$, $s_2 = 5$, and $\beta_2(t) = (5(1-t)^2, 3.5(\exp(-(3t-1)^2)+\exp(-(4t-3)^2))-1.5, 3.5t^{1/2}, 6-2t, 2-3\cos(4\pi t))^T$.

Case III: $\beta_0(t) = 3.5 \sin(2\pi t)$, $s_1 = 5$, $s_2 = 0$, $\beta_1 = (5, -5, 2.5, -2.5, 1)^T$.

We generated covariates $x_i(t)$ from a $p$-dimensional Gaussian process whose component processes each has mean zero and covariance function $\text{cov}(x_{ik}(s), x_{ik}(t)) = 2\sin(2\pi s)\sin(2\pi t)$. The correlation between components is specified as follows. The first $s_1 + s_2 + s_0$ elements of $x_i(t)$ are correlated with a constant correlation $\rho$ and thus follow a compound-symmetry covariance structure. The remaining $p - (s_1 + s_2 + s_0)$ elements of $x_i(t)$ are uncorrelated with each other and the first $s_1 + s_2 + s_0$ elements. The first $s_1$ and $s_2$ elements of $x_i(t)$ are used as $x_{1i}(t)$ and $x_{2i}(t)$ in the model, respectively. The next $s_0$ elements of $x_i(t)$ are spurious variables that are not
related to \( y_i(t) \) but correlated to \( x_{1i}(t) \) and \( x_{2i}(t) \). The random error \( \epsilon_i(t) \) was simulated from an ARMA(1,1) Gaussian process with mean zero and covariance function \( \text{cov}(\epsilon_i(s), \epsilon_i(t)) = \omega |s-t| \), with \( \omega = 0.85 \) and \( r = 0.5 \).

We considered two more cases with different covariate and error distributions.

Case IV: The same as Case I except that the covariance matrix of the first \( s_1 + s_2 + s_0 \) elements of \( x_i(t) \) is an AR(1) matrix with \((j, j')\)th element equal to \( \rho |j-j'| \). We set \( \omega = 0.85 \) and \( r = 0.6 \) for the error process.

Case V: The same as Case I except that the covariance matrix of the first \( s_1 + s_2 + s_0 \) elements of \( x_i(t) \) is a symmetric matrix with \((j, j')\)th element equal to \( |j-j'|/\{2(s_1 + s_2 + s_0)\} + \rho |j-j'| \). We set \( \omega = 0.95 \) and \( r = 0.5 \) for the error process.

After 500 simulations, we summarized the performance of our variable selection and structure identification procedures in Table 1 for the five considered cases. The results for larger values of \( \rho \) under cases II and III are very similar to that under case I, and thus are not reported here. Both varying-coefficients and constant non-zero coefficients were selected correctly with high probability over the simulations, especially when the spurious correlation \( \rho \) is low. True positive fractions for the overall model, the varying-coefficient part and the constant coefficient part were close to the true number of variables while false positive fractions were small in general. When spurious correlation was moderate or high, under-selection and over-selection were observed more often. In general, the selection accuracy improved as we increased the sample size \( n \).

The estimation results for various model components are summarized in Table 2 for the three cases I–III with \( n = 100, m = 20 \) and \( \rho = 0.1 \). The results for moderate or higher correlation values are similar, and are thus skipped here. For estimates of the parametric components, we report the estimation mean absolute error (MAE) and the root mean square er-

![Fig 1. Estimated varying coefficients with median MISE for Case I. From left to right are \( \beta_0(t) \), \( \beta_{21}(t) \), \( \beta_{22}(t) \), and \( \beta_{23}(t) \). The lines with “+” symbols are the true functions; the solid and dashed lines are respectively the refined and initial estimates.](image-url)
Table 1. Variable selection results from 500 simulations. Cvar (Cfix) is the proportion of variables with varying-coefficients (constant non-zero coefficients) that are selected; Size is the average model size; U (O) is the proportion of underselection (overselection); TP (FP) is the average number of true positive (false positive); TPvar (FPvar) is the average number of true positive (false positive) for the varying-coefficients; TPfix (FPfix) is the average number of true positive (false positive) for the constant coefficients; MMMS is the median of the minimum model size to contain all true non-zeros in the screening step. Here, \( m = 20, s_0 = 10, p = 500 \), \( n \) is the sample size and \( \rho \) is the spurious correlation. The values in the parentheses are the robust standard deviations.

| Case | I | II | III | IV | V |
|------|---|----|-----|----|---|
| \( n \) | 100 | 100 | 200 | 100 | 200 | 100 | 200 | 100 | 200 | 100 | 200 |
| \( \rho \) | 0.1 | 0.5 | 0.5 | 0.3 | 0.3 | 0.4 | 0.4 | 0.4 | 0.5 | 0.4 | 0.5 |
| Cvar | 0.958 | 0.904 | 0.996 | 0.952 | 0.986 | - | - | 0.976 | 0.992 | 0.925 | 0.998 |
| Cfix | 0.926 | 0.812 | 0.912 | - | - | 0.938 | 0.969 | 0.854 | 0.936 | 0.810 | 0.914 |
| Size | 5.01 (.79) | 6.43 (1.23) | 5.02 (.54) | 5.12 (.13) | 5.04 (.01) | 5.25 (.13) | 5.01 (.01) | 4.95 (.63) | 4.98 (.40) | 4.87 (.96) | 4.99 (.22) |
| U | 0.00 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| O | 0.01 | 0.19 | 0.02 | 0.08 | 0.01 | 0.11 | 0.03 | 0.03 | 0.00 | 0.04 | 0.00 |
| TP | 5.00 (.79) | 4.99 (1.18) | 5.00 (.54) | 4.97 (.13) | 5.00 (.01) | 4.96 (.12) | 5.00 (.01) | 4.93 (.59) | 4.97 (.40) | 4.81 (.93) | 4.99 (.22) |
| FP | 0.01 (.06) | 0.81 (.25) | 0.02 (.01) | 0.09 (.01) | 0.04 (.00) | 0.15 (.01) | 0.05 (.01) | 0.03 (.19) | 0.00 (.00) | 0.05 (.19) | 0.00 (.00) |
| TPvar | 2.93 (.54) | 2.87 (.75) | 2.97 (.42) | 4.98 (.11) | 5.00 (.04) | - | - | 2.82 (.48) | 2.92 (.34) | 2.79 (.65) | 2.97 (.31) |
| FPvar | 0.10 (.04) | 0.15 (.09) | 0.09 (.06) | 0.03 (.08) | 0.01 (.00) | 0.03 (.08) | 0.01 (.01) | 0.01 (.08) | 0.00 (.00) | 0.04 (.21) | 0.01 (.05) |
| TPfix | 1.92 (.33) | 1.84 (.48) | 1.92 (.22) | - | - | 4.93 (.15) | 5.00 (.04) | 1.96 (.26) | 1.98 (.17) | 1.88 (.43) | 1.99 (.09) |
| FPfix | 0.04 (.28) | 0.80 (.42) | 0.14 (.21) | 0.06 (.14) | 0.03 (.04) | 0.10 (.12) | 0.02 (.01) | 0.16 (.40) | 0.06 (.23) | 0.20 (.42) | 0.08 (.27) |
| MMMS | 5 (0) | 5 (0) | 5 (0) | 5 (0) | 5 (0) | 5 (0) | 5 (0) | 5 (0) | 5 (0) | 5 (0) | 5 (0) |
ror (RMSE). For estimates of the nonparametric components, we report the mean integrated absolute error (MIAE) and root mean integrated squared error (RMISE). The practical estimates performed almost as well as the respective oracle estimates. The refined estimates in general performed better than the respective initial estimates. Typical estimates for the nonparametric functions in Case I with median MISE were depicted in Figure 1.

Table 2

| Case I Parameters | Oracle estimate | Practical estimate |
|-------------------|-----------------|--------------------|
|                   | Initial $\beta_k^{PLS}$ | Refined $\beta_k^{PLS}$ | Initial $\hat{\beta}_k^{PLS}$ | Refined $\hat{\beta}_k^{PLS}$ |
| $\beta_{11}$      | 0.0374          | 0.0623             | 0.0387          | 0.0806          | 0.0266          | 0.0458          |
| $\beta_{12}$      | 0.0507          | 0.0642             | 0.0296          | 0.0768          | 0.0330          | 0.0500          |
| $\beta_{13}$      | 0.0256          | 0.0332             | 0.0352          | 0.0538          | 0.0286          | 0.0469          |
| $\beta_{14}$      | 0.0206          | 0.0360             | 0.0175          | 0.0565          | 0.0249          | 0.0489          |
| $\beta_{15}$      | 0.0300          | 0.0400             | 0.0265          | 0.0648          | 0.0430          | 0.0608          |

| Case II Functions | Oracle estimate | Practical estimate |
|-------------------|-----------------|--------------------|
| $\beta_0(t)$      | 0.1678          | 0.2410             | 0.0872          | 0.1526          | 0.1922          | 0.2863          | 0.1020          | 0.1902          |
| $\beta_{21}(t)$   | 0.1697          | 0.2497             | 0.1098          | 0.1805          | 0.2066          | 0.2819          | 0.1243          | 0.2111          |
| $\beta_{22}(t)$   | 0.1526          | 0.2433             | 0.1151          | 0.1568          | 0.1815          | 0.2760          | 0.1261          | 0.1957          |
| $\beta_{23}(t)$   | 0.1804          | 0.2819             | 0.1241          | 0.2007          | 0.2119          | 0.2939          | 0.1317          | 0.2293          |

| Case III Functions | Oracle estimate | Practical estimate |
|-------------------|-----------------|--------------------|
| $\beta_{11}$      | 0.0136          | 0.0264             | 0.0109          | 0.0223          | 0.0142          | 0.0402          | 0.0136          | 0.0387          |
| $\beta_{12}$      | 0.0125          | 0.0200             | 0.0118          | 0.0141          | 0.0144          | 0.0458          | 0.0138          | 0.0374          |
| $\beta_{13}$      | 0.0256          | 0.0400             | 0.0162          | 0.0332          | 0.0352          | 0.0538          | 0.0286          | 0.0469          |
| $\beta_{14}$      | 0.0206          | 0.0360             | 0.0175          | 0.0282          | 0.0327          | 0.0565          | 0.0249          | 0.0489          |
| $\beta_{15}$      | 0.0300          | 0.0400             | 0.0265          | 0.0346          | 0.0525          | 0.0648          | 0.0430          | 0.0608          |

5.2. Real Data Analysis. We analyzed the well-known Yeast Cycle gene expression data set, originally studied by [25]. There are $n = 297$ cell-cycle-regularized genes whose expression levels were measured at $m = 18$ time points covering two cell-cycle periods. We aimed at identifying important transcription factors (TF) that affect the gene expression. Using the same subset of the original data as [29], we included totally $p = 96$ TFs as covariates in the following analysis.

We first applied the nonparametric screening procedure and 51 TFs were kept after the screening. The nonparametric estimates of varying-coefficients
The names of the TFs are given in the following list:

HIR2 HIR1 MET4 FKH2 NDD1 SWI4 SWI5 SKN7 FKH1 MCM1
SMP1 PHD1 SWI6 PUT3 ACE2 MBP1 CIN5 ABF1 RLM1 GRF10.Pho2.
MSN1 RTG1 STE12 SOK2 RGM1 MTH1 CBF1 RTG3 STB1 INO4
DOT6 GAT3 SIP4 REB1 STP1 YAP6 HAL9 DAL81 GAL4 YAP5
PDR HAP4 MSN4 RAP1 DIG1 CUP9 NRG1 INO2 HAP5 FHL1 RFX1

The above list include all the TFs mentioned in [30]. Comparing to the 21 known yeast TFs related to cell cycle process included in Figure 2 of [29], our list do not include BAS1, GCN4, GCR1, GCR2, LEU3 and MET31 and include all the other 15 TFs. Comparing to the additional TFs identified in Table 2 of [29], the above list do not include 23 of their total 52 TFs.

We then applied our proposed variable selection and structure identification procedure and identified 11 TFs, among which 9 TFs have varying-coefficients and the other 2 TFs have constant coefficients. The refined estimates for the 9 varying-coefficients are plotted in Figures 3, along with the 95% confidence intervals computed from bootstrap based on 500 resamples. For the sake of comparison, in each panel we also displayed the corresponding initial estimate and its 95% bootstrap confidence interval. The two estimates are similar but have distinctions. The confidence intervals for the initial estimates are always slightly wider than those for the respective refined estimates. The estimated constant coefficients are given in Table 3 where the standard errors (SE) were computed based on the bootstrap.

Our results are comparable to previous publications. The estimated varying-coefficients almost all show periodic transcriptional effects, as was evidenced in earlier publications. Of the 9 TFs with varying coefficients, SWI6, FKH2, NDD1 and SWI5 were also identified as important TFs in [29] and [30]; ABF1, HIR1, HIR2, MET4, and SMP1 were also identified as important TFs in [29]. Of the 2 TFs with constant coefficients, MCM1 was identified
Table 3

Estimated constant coefficients $\beta$ for two transcription factors in the Yeast Cell Cycle data. In the parentheses are the bootstrap standard errors.

| TF    | Initial $\beta^{PLS}_k$         | Refined $\beta^{PLS}_k$  |
|-------|--------------------------------|--------------------------|
| MCM1  | 0.0129 (0.0103)                | 0.0220 (0.0101)          |
| RLM1  | -0.0032 (0.0095)               | -0.0097 (0.0094)         |

before in [29] and [30] but its effect was estimated as a varying coefficient instead of constant coefficient; RLM1 was also included in the list of important TFs reported in [29]. [31] used a penalized estimating equation approach to analyze this data and identified similar number of TFs although the authors did not report the name of the TFs.

For two typical individuals selected from the data set, we displayed their observed and fitted time-varying responses in Figure 4. The prediction from the fitted model resembles the true functional response closely and provides a more natural and smooth interpretation for the cell cycle process. These results may serve as useful tools for biologists to study molecular events with large variability.

Finally, we remark that the cell cycle is a complicated biological process and the between-gene heterogeneity may prohibit investigators from making the identical distribution assumption. Our endeavor in this work is to model a collective time-varying effect of the TF that remains relatively fixed among genes. A more refined analysis for individual phases of the cycle may be carried out to reduce the variability. Caution must be exercised to generalize the results, especially to a set of genes with entirely different regulatory

![Figure 3](image-url)
mechanisms.

Fig 4. Observed and fitted response curves for subjects 25 (left) and 177 (right) in the sample. Solid lines are the observed response curves; broken lines are the fitted curves from refined estimates; Dotted lines are the fitted curves from initial estimates.

APPENDIX A: PROOFS OF THEOREMS 3.1-3.3

Denote the Euclidean norm and the sup norm of a vector $v$ by $|v|$ and $|v|_\infty$, respectively. Also, for a matrix $A = (a_{ij})$, define $|A| = \sup_{|x|=1} |Ax|$ and $|A|_\infty = \sup_{i,j} |a_{ij}|$. We describe some important facts first. It is easy to see that we have for some positive $C_{N1}$ and $C_{N2}$,

$$C_{N1} L^{-1/2} |\gamma| \leq \| (\gamma^T_0 \mathbf{B}, \ldots, \gamma^T_q \mathbf{B})^T \|_{L^2} \leq C_{N2} L^{-1/2} |\gamma|$$
uniformly in $\gamma = (\gamma_0, \ldots, \gamma_q)^T$. Lemmas 3 and 4 imply that there are positive constants $C_{B1}$ and $C_{B2}$ such that, with probability tending to 1,

$$C_{B1}/L \leq \lambda_{\min}(\langle \mathbf{B}, \mathbf{B}^T \rangle_n) \leq \lambda_{\max}(\langle \mathbf{B}, \mathbf{B}^T \rangle_n) \leq C_{B2}/L.$$

A.1. Proof of Theorem 3.1. Lemma 5 imply that $l_q(\gamma)$ is strictly concave with probability tending to 1. Thus by Lemma 6, we have only to demonstrate that there is a local minimizer of $l_q(\gamma)$ on $R^{(q+1)L}$, $\bar{\gamma} = (\bar{\gamma}^T_0, \ldots, \bar{\gamma}^T_q)^T$, satisfying

$$\| (\bar{\gamma}^T_0 \mathbf{B}, \ldots, \bar{\gamma}^T_q \mathbf{B})^T - (\gamma^T_{0m} \mathbf{B}, \ldots, \gamma^T_{qm} \mathbf{B})^T \|_{L^2}^2 = O_p(r_{qn}^2).$$

Recalling the definition of $r_{qn}$ and (A.1), we define $\Gamma_M$ for a positive $M$ by

$$\Gamma_M = \{ \gamma \in R^{(q+1)L} | |\gamma - \gamma_m| = M(qL/n)^{1/2} L^{1/2} \}$$

and represent $l_q(\gamma)$ as

$$(A.2) \quad l_q(\gamma) = l_q(\gamma_m) - 2(\gamma - \gamma_m)^T (\mathbf{B}^T \mathbf{W}^T)^T, \epsilon_n + (\gamma - \gamma_m)^T \Sigma_n (\gamma - \gamma_m)$$

By Lemma 7, we have uniformly on $\Gamma_M$, the first term in the RHS of (A.2) is $MqLn^{-1}O_p(1)$. By Lemma 5, we have the second term in the RHS of
\( (A.2) \) is at least \( CM^2 q_L n^{-1} \) uniformly on \( \Gamma_M \) with probability tending to 1, where \( C \) does not depend on \( M \). Thus, we have
\[
\lim_{M \to \infty} \limsup_{n \to \infty} P \left( \inf_{\gamma \in \Gamma_M} l_q(\gamma) > l_q(\gamma_m) \right) = 1.
\]
From the above result, the strict concavity of \( l_q(\gamma) \), and Lemma 6, there is a unique minimizer \( \gamma \) of \( l_q(\gamma) \) on \( R^{(q+1)L} \) giving the desired convergence rate.

**A.2. Proof of Theorem 3.2.** Define \( \gamma = (\gamma_0^T, \ldots, \gamma_q^T)^T \in R^{(q+1)L} \) by
\[
(\text{A.3}) \quad \text{argmin} \| (\beta_0, \ldots, \beta_q)^T - (\gamma_0^T B, \ldots, \gamma_q^T B)^T \|_{L_2}^2.
\]
Then \( (\gamma_0^T B, \ldots, \gamma_q^T B)^T \in G_0 \) due to Assumption S(1) and the minimum in (A.3) is not larger than that of \( \gamma_m \) in Lemma 6. Thus Lemma 6 and (A.1) imply that with probability tending to 1, \( |\gamma - \gamma_m|^2 \leq C\rho_{\epsilon / n}^2 \) for some positive \( C \). The desired result follows if we demonstrate that
\[
(\text{A.4}) \quad \lim_{M \to \infty} \limsup_{n \to \infty} P \left( \inf_{\gamma \in \Gamma_M} Q_q(\gamma) > Q_q(\gamma) \right) = 1,
\]
where \( \Gamma_M = \left\{ \gamma \in R^{(q+1)L} \mid |\gamma - \gamma| = ML^{1/2} r_{\epsilon / n} \right\} \).

Write
\[
(\text{A.5}) \quad Q_p(\gamma) - Q_p(\gamma) = \{ l_q(\gamma) - l_q(\gamma) \}
\]
\[
+ \left[ \sum_{k=0}^q \{ p_\lambda(\| (g_k)_c \|) + p_\lambda(\| (g_k)_f \|_{L_2}) \} - \sum_{k=0}^q \{ p_\lambda(\| (g_k)_c \|) + p_\lambda(\| (g_k)_f \|_{L_2}) \} \right],
\]
where \( g_k = \gamma_k^T B \) and \( \gamma_k = \gamma_k^T B \). We have
\[
(\text{A.6}) \quad l_q(\gamma) - l_q(\gamma)
\]
\[
= 2(\gamma - \gamma)^T \left\{ -\langle B^T W^T \rangle, \epsilon/n + \Sigma_{\epsilon}(\gamma - \gamma_m) \right\} + (\gamma - \gamma)^T \Sigma_{\epsilon}(\gamma - \gamma)
\]
Lemmas 5 and 7 imply that uniformly in \( \gamma \in \Gamma_M \), the first term in the RHS of (A.6) equals \( ML^{1/2} r_{\epsilon / n} O_p((q/n)^{1/2}) + ML r_{\epsilon / n} O_p(L^{-1}) \rho_n = M O_p r_{\epsilon / n}^2 \). By Lemma 5, there is a positive constant \( C \) such that the second term in the RHS of (A.6) is no less than \( CM^2 r_{\epsilon / n}^2 \), uniformly in \( \gamma \in \Gamma_M \) with probability tending to 1. Thus, we have
\[
(\text{A.7}) \quad \lim_{M \to \infty} \limsup_{n \to \infty} P \left( \inf_{\gamma \in \Gamma_M} \{ l_q(\gamma) - l_q(\gamma) \} \geq CM^2 r_{\epsilon / n}^2 / 2 \right) = 1.
\]
Next we consider the difference of the penalty terms. Recall that \( a_0 \) appearing below comes from the SCAD function in (3.6). When \( |\gamma - \gamma| = ML^{1/2} r_{\epsilon / n} \), we have for \( k = 0, \ldots, q \) and sufficiently large \( M \),
\[
| (\gamma_k^T B)_c |, | (\gamma_k^T B)_c | > a_0 \lambda_1 \quad \text{or} \quad | (\gamma_k^T B)_c | = a(\lambda_1) \quad & | (\gamma_k^T B)_c | = 0,
\]
3.2 and 3.1. \( G \) is completed.
3.3 \( G 
\)

\[\| (\gamma_k^T B)_f \|_{L_2}, \| (\tilde{\gamma}_k^T B)_f \|_{L_2} > a_0 \lambda_2 \text{ or } \| (\gamma_k^T B)_f \|_{L_2} = o(\lambda_2) \& \| (\tilde{\gamma}_k^T B)_f \|_{L_2} = 0.\]

The above relations and the properties of the SCAD function imply the second term of the RHS of (A.5) is nonnegative on \( \Gamma_M \). Thus (A.4) follows from (A.6) and (A.7). Hence the proof of Theorem 3.2 is completed.

### A.3. Proof of Theorem 3.3

We prove the sparsity property in a way similar to the former half of the proof of Theorem 1 in [30]. First we prove that any local minimizer of \( Q_q(\gamma) \) on \( R^{(q+1)L} \), \( \tilde{\gamma} = (\tilde{\gamma}_0^T \ldots \tilde{\gamma}_q^T) \), satisfying (A.8) and (A.9) below is the unique minimizer of \( l_q(\gamma) \) on \( \mathcal{G}_0 \), where \( \mathcal{G}_0 \) is the subspace of \( R^{(q+1)L} \) corresponding to the oracle space \( \mathcal{G}_0 \).

\[
(A.8) \quad (\tilde{\gamma}_0^T B, \ldots, \tilde{\gamma}_q^T B)^T \in \mathcal{G}_0
\]

\[
(A.9) \quad \| (\tilde{\gamma}_0^T B, \ldots, \tilde{\gamma}_q^T B)^T - (\beta_0, \ldots, \beta_q)^T \|_{L_2} \leq \eta_n r_{qn}
\]

Next we derive the \( L_2 \) convergence rate of the oracle estimator. Finally we verify that any local minimizer satisfying (A.9) has the property of (A.8) with probability tending to 1.

Let \( \gamma \) be a local minimizer of \( Q_q(\gamma) \) on \( R^{(q+1)L} \), \( \tilde{\gamma} \), satisfying (A.8) and (A.9). Then, consider \( Q_q(\tilde{\gamma} + \delta) \) for \( \tilde{\gamma} + \delta \in \mathcal{G}_0 \). When \(|\delta|\) is small enough, we have the same value of the penalty term as that for \( \tilde{\gamma} \) due to the flatness of the SCAD function. On the other hand, the local optimality of \( \tilde{\gamma} \) implies that \( Q_q(\tilde{\gamma} + \delta) \geq Q_q(\tilde{\gamma}) \). Thus there is a small neighborhood of \( \tilde{\gamma} \) in \( \mathcal{G}_0 \), \( \Gamma_h \), such that \( \inf_{\gamma \in \Gamma_h} l_q(\gamma) \geq l_q(\tilde{\gamma}) \). This shows that \( \tilde{\gamma} \) is a local minimizer of \( l_q(\gamma) \) on \( \mathcal{G}_0 \). Since \( l_q(\gamma) \) is strictly concave on \( R^{(q+1)L} \) with probability tending to 1, this \( \tilde{\gamma} \) must be the unique minimizer of \( l_q(\gamma) \) on \( \mathcal{G}_0 \). A similar argument can be found in [8].

Next we deal with the unique minimizer of \( l_q(\gamma) \) on \( \mathcal{G}_0 \). We neglect \( x^{(s+1)}(t), \ldots, x^{(q)}(t) \) and restrict \( \mathcal{G}_0 \) to \( R^{(s+1)L} \). Besides, we can define \( \gamma_m \in R^{(s+1)L} \) and \( \gamma \in \mathcal{G}_0 \subset R^{(s+1)L} \) as in the proofs of Theorems 3.1 and 3.2. Then we have for \( \gamma \in R^{(s+1)L} \),

\[
l_s(\gamma) - l_s(\bar{\gamma}) = 2(\gamma - \bar{\gamma})^T \{-((B^T W^T)^T, \epsilon)_n + \Sigma_n(\gamma - \gamma_m)\} + (\gamma - \bar{\gamma})^T \Sigma_n(\gamma - \bar{\gamma}),
\]

where \( W \) is defined with \( x^{(s+1)}(t), \ldots, x^{(q)}(t) \) removed. Proceeding in the same way as in the proof of Theorem 3.2, we obtain

\[
\lim_{M \to \infty} \limsup_{n \to \infty} P(\inf_{\gamma \in \Gamma_M} l_s(\gamma) > l_s(\bar{\gamma})), \text{ where } \Gamma_M = \{\gamma \in \mathcal{G}_0 \mid |\gamma - \bar{\gamma}| = M L^{1/2 r_{sn}}\}. \text{ Thus the unique minimizer of } l_s(\gamma) \text{ on } \mathcal{G}_0, \tilde{\gamma}, \text{ satisfies } |\tilde{\gamma} - \bar{\gamma}| = O_p(L^{1/2 r_{sn}}).
\]
Finally we consider the local minimizer satisfying (A.9) and prove that it also satisfies (A.8). Take a local minimizer \( \hat{\gamma} \) satisfying \( \| (\hat{\gamma}_0, \ldots, \hat{\gamma}_q, \ldots, \hat{\gamma}_q^T B) - (\beta_0, \ldots, \beta_q) \|_{L_2} \leq \eta_n r_q n \) and suppose that \( (\hat{\beta}_j)_c \neq 0 \) and \( (\beta_j)_c = 0 \) for some \( j \), where \( \hat{\beta}_j = \hat{\gamma}_j^T B \). Then we have that \( 0 < |(\hat{\beta}_j)_c| = o(\lambda_1) \). Define \( \hat{\beta}_t \) for \( t \in [0, 1/2] \) by

\[
\hat{\beta}_t = \hat{\beta} + t(\hat{\beta}_{-c_j} - \hat{\beta}) = (1 - t)\hat{\beta} + t\hat{\beta}_{-c_j},
\]

where \( \hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_q)^T \) and we define \( \hat{\beta}_{-c_j} \) by replacing \( \hat{\beta}_j \) of \( \hat{\beta} \) with \( (\hat{\beta}_j)_f \).

Defining \( \hat{\gamma}_t = (\hat{\gamma}_{0t}, \ldots, \hat{\gamma}_{qt})^T \) by \( \hat{\beta}_t = (\hat{\gamma}_{0t}, \ldots, \hat{\gamma}_{qt}^T B)^T \), we evaluate

\[
Q_q(\hat{\gamma}_t) - Q_q(\hat{\gamma}) = \{l_q(\hat{\gamma}_t) - l_q(\hat{\gamma})\} + \{p_{\lambda_1}((1 - t)|(\hat{\beta}_j)_c|) - p_{\lambda_1}(|(\hat{\beta}_j)_c|)\} = J_5 + J_6.
\]

It is easy to see that for some \( \tilde{t} \in [0, t] \), \( J_6 = -t|(\hat{\beta}_j)_c| p'_{\lambda_1}((1 - \tilde{t})|(\hat{\beta}_j)_c|) \). We can represent \( J_5 \) as

\[
J_5 = -2(\hat{\gamma}_t - \hat{\gamma})^T (B^T W^T)^T, y - \hat{\gamma}_0^T B - \sum_{k=1}^q \hat{\gamma}_k^T W_k ) n + (\hat{\gamma}_t - \hat{\gamma})^T \Sigma_n (\hat{\gamma}_t - \hat{\gamma})
\]

Lemmas 5 and 7 imply that the two term in \( J_5 \) can be expressed as \( -2(\hat{\gamma}_t - \hat{\gamma})^T (B^T W^T)^T, (B^T W^T) ) n (\gamma_m - \hat{\gamma}) - 2(\hat{\gamma}_t - \hat{\gamma})^T (B^T W^T)^T, (\gamma_m - \hat{\gamma})_n \) and \( t|(\hat{\beta}_j)_c|O_p(\eta_n r_q n) \). Hence we get \( J_5 = t|(\hat{\beta}_j)_c|O_p(\eta_n r_q n) \). From the above results, property of the SCAD function and Assumption S(2),

\[
Q_p(\hat{\gamma}_t) - Q_p(\hat{\gamma}) = t|(\hat{\beta}_j)_c| \{O_p(\eta_n r_q n) - p'_{\lambda_1}((1 - \tilde{t})|(\hat{\beta}_j)_c|)\} < 0
\]

uniformly in \( t \in (0, 1/2) \) with probability tending to 1 and the probability does not depend on the specific \( j \). This contradicts with the local optimality of \( \hat{\gamma} \) and \( (\beta_j)_c \) must be equal to 0 if \( (\beta_j)_c = 0 \). We can treat the other cases in the same way. Hence (A.8) is established for the local minimizer \( \hat{\gamma} \).

REFERENCES

[1] Antoniadis, A., Gijbels, I. and Lambert-Lacroix, S. (2013). Penalized estimation in additive varying coefficient models using grouped regularization. Statistical Papers in press.

[2] Antoniadis, A., Gijbels, I. and Verhaege, A. (2012). Variable selection in varying-coefficient models using P-splines. J. Comput. Graph. Stat. 21 638–661.

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

[4] Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. Bernoulli 11 1031–1057.
[5] Fan, J., Huang, T. and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Amer. Statist. Assoc.* **102** 632–641.

[6] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348 – 1360.

[7] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. Royal Statist. Soc. Ser. B* **70** 849–911.

[8] Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transaction on Information Theory* **57** 5467–5484.

[9] Fan, J., Ma, Y. and Dai, W. (2013). Nonparametric independence screening in sparse ultra-high dimensional feature space. *J. Royal Statist. Soc. Ser. B* **70** 849–911.

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

[11] Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *J. Amer. Statist. Assoc.* **103** 1520–1533.

[12] Fan, Y. and Lv, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *J. Amer. Statist. Assoc.* **108** 1044–1061.

[13] Fan, Y. and Tang, C. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. Royal Statist. Soc. Ser. B* **75** 531–552.

[14] Hall, P., Müller, H. G. and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Stat.* **34** 1943–1517.

[15] Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Biometrika* **91** 763–788.

[16] Lam, C. and Fan, J. (2008). Profile-Kernel likelihood inference with diverging number of parameters. *Ann. Stat.* **36** 2232–2260.

[17] Li, G. R., Peng, H., Zhang, J. and Zhu, L. X. (2012). Robust rank correlation based screening. *Ann. Stat.* **40** 1846–1877.

[18] Li, Y. (2001). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika* **98** 355–370.

[19] Lian, H., Lai, P. and Liang, H. (2013). Partially linear structure selection in Cox models with varying coefficients. *Biometrika* **69** 348–357.

[20] Liu, J., Li, R. and Wu, R. (2013). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *Technical Report*.

[21] Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Stat.* **37** 3498–3528.

[22] Meier, L., Van de Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *J. Royal Statist. Soc. Ser. B* **70** 53–71.

[23] Noh, H. S. and Park, B. U. (2010). Sparse variable coefficient models for longitudinal data. *Statist. Sinica* **20** 1183–1292.

[24] Schumaker, L. L. (2007). *Spline Functions: Basic Theory 3rd ed.* Cambridge University Press, Cambridge.

[25] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of Cell* **9** 3273–3297.

[26] Tao, H. and Xia, Y. (2011). Adaptive Semi-varying Coefficient Model Selection. *Statist. Sinica* **22** 575–599.

[27] Tibshirani, R. J. (1996). Regression shrinkage and selection via the Lasso. *J. Royal Statist. Soc. Ser. B* **58** 267–288.
[28] Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568.

[29] Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23** 1486-1494.

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

[31] Wang, L., Zhou, J. and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68** 353-360.

[32] Wang, N., Carroll, R. J. and Lin, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Amer. Statist. Assoc.* **100** 147–157.

[33] Wei, F., Huang, J. and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica* **21** 1515-1540.

[34] Xia, Y., Zhang, W. and Tong, H. (2004). Efficient estimation for semivarying-coefficient models. *Biometrika* **91** 661–681.

[35] Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *J. Machine Learning Research* **13** 1973-1998.

[36] Yan, J. and Huang, J. (2012). Model selection for Cox models with time-varying Coefficients. *Biometrics* **68** 419–428.

[37] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Royal Statist. Soc. Ser. B* **68** 49–67.

[38] Zhang, H. H., Cheng, G. and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *J. Amer. Statist. Assoc.* **106** 1099–1112.

[39] Zhang, W., Fan, J. and Sun, Y. (2009). A semiparametric model for cluster data. *Ann. Stat.* **37** 2377–2408.

[40] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429.
Supplement A: Lemmas, and proofs of Theorems 2.1-2.2 and Remark 1
(doi: 10.1214/00-AOSXXXXSUPP).

First, in order to prove Theorem 2.2, we need to define $b_{kn}$, which is the theoretical version of the B-spline estimator $\hat{b}_k$, by minimizing

$$\|y - \gamma_1^T B - \gamma_2^T x^{(k)} B\|^2 = \|y - \gamma_1^T B - \gamma_2^T W_k\|^2$$

with respect to $\gamma_1 \in \mathbb{R}^L$ and $\gamma_2 \in \mathbb{R}^L$. Denoting the minimizer of the above quantity by $\gamma_{1k}$ and $\gamma_{2k}$, we define $b_{kn}$ by

(S.1) $b_{kn}(t) = \gamma_{2k}^T B(t)$.

We compare $\|b_k\|^2$ and $\|b_{kn}\|^2$ and evaluate the difference between $\|\hat{b}_{kn}\|^2$ and $\|b_{kn}\|^2$ in the proof of Theorem 2.2.

Some lemmas. Next, we state some lemmas. Lemmas 1-4 are needed in the proofs of Theorems 2.1 and 2.2. Lemmas 2 and 3 are a version of Lemmas 4 and 5 of [3], respectively. Lemmas 5-7 are used in the proofs of Theorem 3.1 - 3.3. The proofs of Lemmas 1-7 are given later.

Lemma 1. Suppose that Assumptions M1 and T(1)-(5) hold. Set $\rho_n = L^{-2}$. Then there is a positive constant $C_\rho$ such that we have for $k = 1, \ldots, p$,

$$\|b_k - b_{kn}\|^2 \leq C_\rho \rho_n^2 = C_\rho L^{-4}.$$

Before we state Lemma 2, recall that $B_l$ is the $l$th element of $B$ and that $x^{(k)} B_l$ is the $l$th element of $W_k$.

Lemma 2. Under Assumption T(1)-(5), we have for $l = 1, \ldots, L$ and $k = 1, \ldots, p$,

$$\mathbb{P}(|\langle B_l; y \rangle_n - \langle B_l; y \rangle_n| \geq \delta n^{-1}) \leq c_1 \exp\{-\delta^2/(c_2 n L^{-1} + c_3 \delta)\},$$
$$\mathbb{P}(|\langle x^{(k)} B_l; y \rangle_n - \langle x^{(k)} B_l; y \rangle_n| \geq \delta n^{-1}) \leq c_1 \exp\{-\delta^2/(c_2 n L^{-1} + c_3 \delta)\},$$

where $c_1$, $c_2$, and $c_3$ are positive constants depending on Assumption T.

Before we state Lemma 3, we set

$$D_{kn} = \langle (B^T W_k^T)^T, (B^T W_k^T) \rangle_n \quad \text{and} \quad D_k = \mathbb{E}(D_{kn}).$$
Lemma 3. Under Assumption T(1)-(5), we have for \( k = 1, \ldots, p \),
\[
P(|D_{kn} - D_k| \geq \delta_n) \leq c_4 L \exp\{-\delta^2/(c_5 n L^{-1} + c_6 \delta)\},
\]
where \( c_4, c_5, \) and \( c_6 \) are positive constants depending on Assumption T. This inequality also holds for \( \langle B, B^T \rangle_n - \langle B, B^T \rangle \).

We evaluate eigenvalues of \( D_k \) and \( \langle B, B^T \rangle \) in Lemma 4 below.

Lemma 4. Suppose that Assumption T(1)-(5) hold. Then there exist positive constants \( C_{e1}, \ldots, C_{e6} \) such that
\[
C_{e1} L^{-1} I_L \leq \langle B, B^T \rangle \leq C_{e2} L^{-1} I_L,
\]
\[
C_{e3} L^{-1} I_{2L} \leq D_k \leq C_{e4} L^{-1} I_{2L}, \quad \text{for } k = 1, \ldots, p,
\]
\[
C_{e5} L^{-1} I_L \leq \langle W_k, W_k^T \rangle \leq C_{e6} L^{-1} I_L, \quad \text{for } k = 1, \ldots, p.
\]

Lemma 5. Suppose that Assumptions T(4) and V hold. Then there is a positive constants \( C_W \) such that
\[
\lambda_{\min}(\Sigma_n) \geq C_W / L
\]
with probability tending to 1. See (3.10) for the definition of \( \Sigma_n \).

We define \( \gamma_m = (\gamma_{0m}^T, \ldots, \gamma_{qm}^T)^T \in R^{(q+1)L} \) which has some desirable properties:

\[
(S.2) \quad \gamma_m = \arg\min_{(\gamma_{0m}^T, \ldots, \gamma_{qm}^T)^T \in R^{(q+1)L}} \left\| \beta_0 + \sum_{k=1}^{q} \beta_k x^{(k)} - \gamma_{0m}^T B - \sum_{k=1}^{q} \gamma_k^T W_k \right\|^2_n.
\]

Then it is easy to see that \( \gamma_m \) satisfies

\[
(S.3) \quad \langle \beta_0 + \sum_{k=1}^{q} \beta_k x^{(k)} - \gamma_{0m}^T B - \sum_{k=1}^{q} \gamma_k^T W_k, W \rangle_n = 0,
\]
\[
(S.4) \quad \langle \beta_0 + \sum_{k=1}^{q} \beta_k x^{(k)} - \gamma_{0m}^T B - \sum_{k=1}^{q} \gamma_k^T W_k, B \rangle_n = 0.
\]

The approximation property will be established in Lemma 6 below.

Lemma 6. Suppose that Assumptions T(4)-(6) and V hold. Then there exists a positive constant \( C_m \) such that
\[
\| (\beta_0, \ldots, \beta_q)^T - (\gamma_{0m}^T B, \ldots, \gamma_{qm}^T B)^T \|^2_{L_2} \leq C_m \rho_{qn}^2.
\]
with probability tending to 1.
Lemma 7. Suppose Assumptions T(4)-(6), V, and E hold. Then we have
\[
\langle (B^T W^T)^T, \epsilon \rangle_n = O_p((q/n)^{1/2}).
\]

Proof of Theorem 2.1. We prove Theorem 2.1 in a similar manner as the proof of Theorem 1 in [3]. Define 2L-dimensional vectors \(d_{kn}\) and \(d_k\) by
\[
d_{kn} = \langle (B^T W_k^T)^T, y \rangle_n \quad \text{and} \quad d_k = E(d_{kn}).
\]
Note that we have \(|d_k| \leq CL^{-1/2}\) by the properties of B-spline bases and Assumption T(4)-(5). Then we have
\[
\hat{b}_k(t) = (0^T B(t)^T)D_k^{-1}d_{kn} \quad \text{and} \quad b_{kn}(t) = (0^T B(t)^T)D_k^{-1}d_k.
\]
Lemma 1 and the relation of \(n^{2\kappa}\) and \(L\) imply that we have only to evaluate
\[
\langle B_l, y \rangle_n - \langle B_l, y \rangle \leq \delta/n, \quad l = 1, \ldots, L,
\]
\[
\langle x^{(k)} B_l, y \rangle_n - \langle x^{(k)} B_l, y \rangle \leq \delta/n, \quad k = 1, \ldots, p \quad \text{and} \quad l = 1, \ldots, L,
\]
\[
|D_{kn} - D_k| \leq \delta/n, \quad k = 1, \ldots, p,
\]
\[
|\langle B, B^T \rangle_n - \langle B, B^T \rangle| \leq \delta/n.
\]

Lemmas 2 and 3 imply that
\[
P(\Omega_\delta) \geq 1 - (p + 1)Lc_1 \exp\{-\delta^2/(c_2nL^{-1} + c_3\delta)\}
\]
\[
- (p + 1)Lc_4 \exp\{-\delta^2/(c_5nL^{-1} + c_6\delta)\}.
\]

We rewrite the RHS of (S.5) as
\[
d_{kn}^T D_{kn}^{-1}Q_n D_{kn}^{-1}d_{kn} - d_k^T D_k^{-1}Q D_k^{-1}d_k
\]
\[
= (d_{kn}^T D_{kn}^{-1} - d_k^T D_k^{-1})Q_n (D_{kn}^{-1}d_{kn} - D_k^{-1}d_k)
\]
\[
+ 2(d_{kn}^T D_{kn}^{-1} - d_k^T D_k^{-1})Q_n D_k^{-1}d_k
\]
\[
+ d_k^T D_k^{-1}(Q_n - Q) D_k^{-1}d_k
\]
\[
= R_1 + R_2 + R_3 \quad \text{(say)}.
\]
Before we consider \((S.7)-(S.9)\), we evaluate on \(\Omega_\delta\),

\[
(D_k^{-1}d_k - D_k^{-1}d_{kn}) = \sum (D_k^{-1} - D_k^{-1})d_k.
\]

Notice that we have on \(\Omega_\delta\),

\[
|d_{kn} - d_k| \leq CL^{1/2}\delta/n, \quad |D_k^{-1}| \leq CL, \quad \text{and} \quad |D_{kn} - D_k| \leq C\delta/n.
\]

Here we used Assumption M2(3) and Lemma 4. Thus we obtain

\[
|D_k^{-1}d_k - d_{kn}| \leq CL^{3/2}\delta/n,
\]

\[
(D_k^{-1} - D_k^{-1})d_k \leq (D_k^{-1}||d_k - D_{kn}||d_k) \leq C\delta/n.
\]

We used the fact that \(|d_k| \leq CL^{1/2}\delta/n\). By combining the above inequalities, we get for \(k = 1, \ldots, p\),

\[
|D_k^{-1}d_k - D_k^{-1}d_{kn}| \leq CL^{3/2}\delta/n.
\]

We also have

\[
|Q_n| \leq CL^{-1} \quad \text{and} \quad |Q - Q_n| \leq C\delta/n.
\]

Hence we have

\[
|R_1| \leq C(L^{3/2}\delta/n)^2L^{-1} = CL^2\delta^2/n^2,
\]

\[
|R_2| \leq C(L^{3/2}\delta/n)L^{-1}LL^{-1/2} = CL\delta/n,
\]

\[
|R_3| \leq C(L^{-1/2}L)^2\delta/n = CL\delta/n.
\]

The above inequalities yield

\[
|d_k^T D_k^{-1} Q_n D_k^{-1} d_{kn} - d_k^T D_k^{-1} Q D_k^{-1} d_k| \leq C(L^2\delta^2/n^2 + L\delta/n) \leq CC\delta Ln^{-2\kappa}.
\]

We choose \(C_\delta\) such that \(C_\delta C \leq \min\{C\kappa2/3, C\kappa3/2\}\) in \((S.11)\). In addition, Lemma 1 and Assumption M2(3) imply that \(||b_k||^2 - ||b_{kn}||^2 = o(n^{-2\kappa}L)\). Then if \(n\) is sufficiently large, we have on \(\Omega_\delta\),

\[
||b_k||^2 \geq C\kappa2n^{-2\kappa}L \quad \Rightarrow \quad ||\hat{b}_k||^2 \geq C\kappa2n^{-2\kappa}L/2 \quad \Rightarrow \quad k \in \hat{M}_\kappa.
\]

The above relations and \((S.6)\) yield the desired result.

**Proof of Theorem 2.2.** We prove Theorem 2.2 by following closely the proof of Theorem 2 in [3]. We use \(\Omega_\delta\) defined in the proof of Theorem 2.1. On \(\Omega_\delta\) with the same \(\delta\) as in the proof of Theorem 2.1, we have

\[
||\hat{b}_k||^2 \geq C\kappa3n^{-2\kappa}L \quad \Rightarrow \quad ||b_{kn}||^2 \geq C\kappa3n^{-2\kappa}L/2.
\]
Since
\[ \sum_{k \in \hat{M}_n} \|b_{kn}\|^2 \geq \frac{C_{\kappa}^3}{2} n^{-2\kappa} L \# \hat{M}_n, \]
we have
\[ (# \hat{M}_n \leq C n^{2\kappa} L^{-1} \sum_{k=1}^{p} \|b_{kn}\|^2). \]

Thus we have only to evaluate \( \sum_{k=1}^{p} \|b_{kn}\|^2 \). It is easy to see by the definition of \( b_{kn} \) in (S.1) that
\[ \sum_{k=1}^{p} \|b_{kn}\|^2 = \sum_{k=1}^{p} \langle y, \bar{W}_k^T \rangle \langle \bar{W}_k, \bar{W}_k^T \rangle^{-1} \langle \bar{W}_k, y \rangle \]
\[ \leq C L \sum_{k=1}^{p} |\langle \bar{W}_k, y \rangle|^2 \leq C L |\langle \bar{W}, y \rangle|^2. \]

We used Lemmas 3 and 4 here. To evaluate \( |\langle \bar{W}, y \rangle|^2 \), we define \( \gamma_n \) by
\[ \gamma_n = \text{argmin} \|y - \bar{W}^T \gamma\|^2, \]
as in [3]. Then we have
\[ \langle \bar{W}, y - \bar{W}^T \gamma_n \rangle = 0 \in R^{pL}. \]

This implies that
\[ |\langle \bar{W}, y \rangle|^2 \leq \gamma_n^T \Sigma \gamma_n \leq \lambda_{\text{max}}(\Sigma) \gamma_n^T \Sigma \gamma_n. \]

Since we have by the definition of \( \gamma_n \),
\[ \|y\|^2 = \|y - \bar{W}^T \gamma_n\|^2 + \|\bar{W}^T \gamma_n\|^2, \]
we obtain
\[ \gamma_n^T \Sigma \gamma_n \leq \|y\|^2 = O(1). \]

Hence the desired result follows from (S.12)-(S.15).

\[ \blacksquare \]

Proofs of Lemmas 1-7

Proof of Lemma 1. Assumption T(1) and (2.6) implies that there are positive constants \( C_1 \) and \( C_2 \) such that
\[ C_1(\|a\|_{L_2}^2 + \|b\|_{L_2}^2) \leq \|a + bx^{(k)}\|^2 \leq C_2(\|a\|_{L_2}^2 + \|b\|_{L_2}^2), \]
where $a$ and $b$ are square integrable functions. By Assumption M1 and Corollary 6.26 in Shumaker (2007), there exist $\tilde{\gamma}_{k1} \in R^L$ and $\tilde{\gamma}_{k2} \in R^L$ such that

$$\|a_k - \tilde{\gamma}_{k1}^T B\|_\infty \leq C\rho_n \quad \text{and} \quad \|b_k - \tilde{\gamma}_{k2}^T B\|_\infty \leq C\rho_n$$

for $k = 1, \ldots, p$. Then we have

(S.17)  
$$\|a_k + b_k x^{(k)} - \tilde{\gamma}_{k1}^T B - (\tilde{\gamma}_{k2}^T B)x^{(k)}\|_2^2 \leq C\rho_n^2.$$ 

By the definition of $a_{kn}$ and $b_{kn}$, we have

(S.18)  
$$\|a_k + b_k x^{(k)} - a_{kn} - b_{kn} x^{(k)}\|_2^2 \leq \|a_k + b_k x^{(k)} - \tilde{\gamma}_{k1}^T B - (\tilde{\gamma}_{k2}^T B)x^{(k)}\|_2^2 \leq C\rho_n^2.$$ 

The desired result follows from (S.16), (S.17), and (S.18). \hfill \Box

We can prove Lemmas 2 and 3 in a similar way to the proofs of Lemmas 4 and 5 in [3]. We just outline the proofs here.

**Proof of Lemmas 2 and 3.** We have only to note the following three facts. The desired results follow from the proofs of Lemmas 4 and 5 in [3]. However, we have exploited the fact that the matrices in Lemma 6.26 in Shumaker (2007), there exist $\tilde{\gamma}_{i} \in R^L$ and $\tilde{\gamma}_{j} \in R^L$ such that the matrices in Lemma 3 have a kind of band diagonal property and we have improved the rate.

1. There are positive constants $C_1$ and $C_2$ such that for $r \geq 2$,

$$\frac{1}{m_i} |\epsilon_i(t_i)B_i(t_i)^T|^r \leq C_1 C_2 \rho_i^{-1} |B_i(t_i)|^2 (m_i^{-1}/|\epsilon_i(t_i)|)^r.$$ 

Recall that $B_i(t_i) = (B_i(t_{i1}), \ldots, B_i(t_{im_i}))$ and $\epsilon_i(t_i) = (\epsilon_i(t_{i1}), \ldots, \epsilon_i(t_{im_i}))$ and that they are $m_i$-dimensional row vectors. By Assumption T(2), we have for some positive $C_3$,

$$\mathbb{E}\left\{ \frac{1}{m_i} |\epsilon_i(t_i)B_i(t_i)^T|^r \right\} \leq C_3 C_2 \rho_i^r L^{-1}.$$ 

2. Let $f_i(x_i, t_i)$ be an $m_i$-dimensional random row vector satisfying $|f_i(x_i, t_i)| \leq C_1$. Then we have for some positive $C_2$ and $C_3$,

$$m_i^{-1} f_i(x_i, t_i) B_i(t_i)^T |^2 \leq \frac{C_2}{m_i} |B_i(t_i)|^2 \text{ and } \mathbb{E}\{ m_i^{-1} f_i(x_i, t_i) B_i(t_i)^T |^2 \} \leq \frac{C_3}{L}.$$ 

3. There are positive constants $C_1$ and $C_2$ such that

$$m_i^{-1} B_j(t_i) B_i(t_i)^T |^2 \leq C_1 (m_i^{-1} |B_j(t_i)|^2) (m_i^{-1} |B_i(t_i)|^2) \leq \frac{C_1}{m_i} |B_j(t_i)|^2, \text{ and } \mathbb{E}\{ m_i^{-1} B_j(t_i) B_i(t_i)^T |^2 \} \leq C_2 L^{-1}.$$
Proof of Lemma 4. By the definition of $\langle \cdot, \cdot \rangle$, we have
\[
\langle B, B^T \rangle = E \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} B(t_i) B(t_i)^T \right\}.
\]

The desired result for $\langle B, B^T \rangle$ follows from a well-known result on B-spline bases and the assumption on $t_i$. For example, see A.2 in [15]. Similarly,
\[
D_k = E \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} \left( \frac{1}{x_i^T(t_{ij})} \{ x_i^T(t_{ij}) \}^2 \right) \langle B(t_{ij}) B(t_{ij})^T \rangle \right\}
\]

Thus the properties of B-spline bases and (2.6) yield the desired result for $D_k$. The result for $\hat{W}_k$ follows from Assumption T(1) and the following inequalities: $\langle \hat{W}_k, \hat{W}_k^T \rangle \leq \langle W_k, W_k^T \rangle$. 

Here we verify Lemmas 5-7 for the group SCAD estimator. Hereafter the number of covariates is $q$.

Proof of Lemma 5. We exploit the band diagonal property of $\langle W_k, B^T \rangle_n$, $\langle W_k, B^T \rangle$, $\langle W_k, W^T \rangle_n$, $\langle W_k, W^T \rangle$, $\langle B, B^T \rangle_n$, and $\langle B, B^T \rangle$. The band diagonal property is due to the equi-spaced B-spline basis. Besides, the proof of Lemmas 2 and 3 implies that each element of $\langle W_k, B^T \rangle_n - \langle W_k, B^T \rangle$, $\langle W_k, W^T \rangle_n - \langle W_k, W^T \rangle$, and $\langle B, B^T \rangle_n - \langle B, B^T \rangle$ is uniformly $O_p(\sqrt{\log n/(nL)})$.

We evaluate the difference between $\Sigma_n$ and $\Sigma$ and use Assumption V. Set
\[
\Delta_1 = \Sigma_n - \Sigma = \langle W, W^T \rangle_n - \langle W, W^T \rangle.
\]

It is easy to see that
\[
\max \{|\lambda_{\min}(\Delta_1)|, |\lambda_{\max}(\Delta_1)|\} \leq \max_{i=1}^{qL} \sum_{j=1}^{qL} |\Delta_{1ij}| = qO_p(\sqrt{\log n/(nL)}),
\]
where $\Delta_{1ij}$ is the $(i, j)$th element of $\Delta_1$. The desired result follows from (3.3) and (S.19). 

Proof of Lemma 6. The following arguments hold with probability tending to 1 by Lemmas 3-5. By Assumption T(5)-(6), there is a $(q+1)L$-dimensional vector $\gamma_\infty = (\gamma_{0\infty}^T, \ldots, \gamma_{q\infty}^T)^T$ such that
\[
\| (\beta_0, \ldots, \beta_q)^T - (\gamma_{0\infty}^T, \ldots, \gamma_{q\infty}^T)^T \|_\infty \leq \rho_q n.
\]

Notice that we also have
\[
\| (\beta_0, \ldots, \beta_q)^T - (\gamma_{0\infty}^T, \ldots, \gamma_{q\infty}^T)^T \|_2 \leq \rho_q n.
\]
Then using Assumption T(4) and (S.20), we have for some positive $C$,

\[(S.22) \quad \left\| \beta_0 + \sum_{k=1}^{q} \beta_k x^{(k)} - \gamma_{0\infty}^T B - \sum_{j=1}^{q} \gamma_{jm}^T W_j \right\| \leq C \rho q n.\]

The definition of $\gamma_m$ and (S.22) imply that

\[(S.23) \quad \left\| \beta_0 + \sum_{k=1}^{q} \beta_k x^{(k)} - \gamma_{0m}^T B - \sum_{j=1}^{q} \gamma_{jm}^T W_j \right\| \leq C \rho q n.\]

By combining (S.22) and (S.23), we obtain

\[(S.24) \quad \left\| (\gamma_{0\infty} - \gamma_{0m})^T B + \sum_{j=1}^{q} (\gamma_{jm} - \gamma_{jm})^T W_j \right\| \leq C \rho q n.\]

We can represent (S.24) as

\[(S.25) \quad (\gamma_{\infty} - \gamma_m)^T \Sigma_n (\gamma_{\infty} - \gamma_m) \leq C \rho^2 q n\]

and we have by Lemma 5 that

\[(S.26) \quad (\gamma_{\infty} - \gamma_m)^T (\gamma_{\infty} - \gamma_m) \leq C \rho^2 q n L.\]

Now, (A.1), (S.21) and (S.26) yield

\[\left\| (\beta_0, \ldots, \beta_q)^T - (\gamma_{0m}^T B, \ldots, \gamma_{qm}^T B)^T \right\|_{L^2} \leq C \rho q n.\]

Proof of Lemma 7. Assumption E implies that, for some positive $C_1$ and $C_2$,

\[E\left[ \mathbb{E}\left\{ \langle (B^T W^T), \epsilon \rangle_n (\epsilon, (B^T W^T)^T)\rangle_n | X, T \right\} \right] \leq \frac{C_1}{n} E\left\{ \text{tr}((B^T W^T)^T, (B^T W^T))\rangle_n \right\} \leq \frac{C_2}{n} (q + 1).\]

Note that $E\{\cdot | X, T\}$ is a conditional probability on all the covariates and the observation times. The desired result follows from the above inequalities. Hence the proof of Lemma 7 is completed.

Outlined proof of Remark 1. The proof consists of four steps. First we consider $p_\lambda(\|\gamma_j\|/\sqrt{L})$ instead of $p_\lambda(\|g_j\|_{L^2})$. We note in step 1 that we also have Theorems 3.1-3.3 for $R(s+1)L$. In step 2, we adapt the proof of Theorem 1 of [21] to the present setup. Then we prove a version of (19) of [21] in step 3. Finally in step 4, we describe how to deal with $p_\lambda(\|g_j\|_{L^2})$. 

\[\square\]
Step 1: Notice that Theorems 3.1-3.3 hold with \( q = s \) if \( \lambda / r_{sn} \to \infty \). Then we consider \( Q_s(\gamma) \), where \( \gamma \in R^{(s+1)L} \), on \( R^{(s+1)L} \) and we can define \( \gamma_{m} \in R^{(s+1)L} \) as in (S.2). Then by Theorem 3.3, any local minimizer of \( Q_s(\gamma) \) on \( N_s = \{ \gamma \in R^{(s+1)L} \mid |\gamma - \gamma_{m}| \leq r_{sn}\sqrt{L} \} \) is equal to \( \hat{\gamma} \) in the remark with probability tending to 1.

Step 2: (18) of [21] is satisfied for \( \hat{\gamma} \) here. Suppose that a version of their (19) given below holds. Define \( b_j \in R^L, j = s + 1, \ldots, q \), by \( b_j = \langle W_j, y - W_a^T \hat{\gamma} \rangle_n \), where \( W_a = (B^T, W_1^T, \ldots, W_s^T)^T \). Our version of (19) therein is

\[
\max_{s+1 \leq j \leq q} |b_j|/(\lambda/\sqrt{L}) \to 0
\]

in probability if \( \lambda / \max\{\sqrt{r_{sn}}, L^{-3/2}\} \to \infty \).

Let \( N_q \subset R^{(q+1)L} \) be a sufficiently small neighborhood of \( (\gamma_{m}^T, 0)^T \in R^{(q+1)L} \) such that the projection of \( N_q \) into \( R^{(s+1)L} \) is included in \( N_s \). Note that \( N_q \) can depend on the observations. Taking \( \tilde{\gamma} = (\gamma_1^T, \gamma_2^T)^T \in N_q \), where \( \gamma_1 \in R^{(s+1)L} \) and \( \gamma_2 \in R^{(q-s)L} \), and proceeding as in the proof of Theorem 1 of [21], we have with probability tending to 1,

\[
Q_q(\tilde{\gamma}) > Q_s(\gamma_1) \quad \text{on} \quad N_q.
\]

Besides, Theorem 3.3 implies that with probability tending to 1,

\[
Q_s(\gamma_1) > Q_s(\tilde{\gamma}) = l_s(\tilde{\gamma}).
\]

Hence \( (\tilde{\gamma}^T, 0^T)^T \in R^{(q+1)L} \) is a local minimizer of \( Q_s(\tilde{\gamma}) \) on \( R^{(q+1)L} \). Note that (20) of [21] is not necessary since we have Theorem 3.3.

Step 3: We prove (S.27) here. First write for \( j = s + 1, \ldots, q \),

\[
\langle W_j, y - W_a^T \hat{\gamma} \rangle_n = \langle W_j, \epsilon + x_a^T \beta_a - x_a^T \beta_m + W_a^T \gamma_m - W_a^T \hat{\gamma} \rangle_n = \langle W_j, \epsilon \rangle_n + \langle W_j, x_a^T (\beta_a - \beta_m) \rangle_n + \langle W_j, W_a^T \gamma_m - W_a^T \hat{\gamma} \rangle_n = J_{ja} + J_{jb} + J_{jc} \quad \text{(say)},
\]

where \( x_a = (1, x^{(1)}, \ldots, x^{(s)})^T, \beta_a = (\beta_0, \beta_1, \ldots, \beta_s)^T, \beta_m = (\gamma_{0m}^T B, \gamma_{1m}^T B, \ldots, \gamma_{sm}^T B)^T \). We evaluate \( J_{ja}, J_{jb}, \) and \( J_{jc} \) separately. We begin with \( J_{ja} \). By applying exponential inequalities as in the proofs for the NIS, we have uniformly in \( j \) and \( l \), \( \langle (W_j, \epsilon) \rangle_n = O_p(\sqrt{\log n/(nL)}) \), where \( \langle (W_j, \epsilon) \rangle \) is the \( l \)th element of \( (W_j, \epsilon)_n \). Then \( |(W_j, \epsilon)_n|^2 = O_p(\log n/n) \).

This yields that, uniformly in \( j \),

\[
\sqrt{L}|J_{ja}| = O_p(\sqrt{L \log n/n}).
\]
Next we deal with $J_{jb}$. Notice that Lemma 6 implies that $\|x_a^T(\beta - \beta_m)\|_n^2 = O_p(L^{-4})$. Then uniformly in $j$ and $l$, we have

$$\|(W_j, x_a^T(\beta - \beta_m))_n\|_l \leq \|x^{(j)} B_l\|_n \|x_a^T(\beta - \beta_m)\|_n = O_p(L^{-5/2}).$$

The above inequalities yield that uniformly in $j$, $\|(W_j, x_a^T(\beta - \beta_m))_n\| = O_p(L^{-2})$ and we have uniformly in $j$,

$$\sqrt{L}|J_{jb}| = O_p(L^{-3/2}).$$

Finally we evaluate $J_{jc}$. Recall that $|\hat{\gamma} - \gamma_m| = O_p(\sqrt{r_{sn}})$. We should evaluate $\|(W_j, W_a^T)_n\|$, where $(W_j, W_a^T)_n$ is an $L \times (s + 1)L$ random matrix. We evaluate the maximum eigenvalues of an $L \times L$ matrix $(W_j, W_a^T)_n(W_a^T, W_j)_n$ by exploiting the band diagonal property and find that the maximum eigenvalues of the $L \times L$ matrix is bounded from above by $s/L^2$ uniformly in $j$ with probability tending to 1. Then we have for some positive $C$,

$$\|(W_j, W_a^T)_n\| \leq C \sqrt{s/L}$$

uniformly in $j$ with probability tending to 1. Hence,

$$\|(W_j, W_a^T)_n(\hat{\gamma} - \gamma_m)\| = O_p(r_{sn} \sqrt{L/s})$$

uniformly in $j$, with probability tending to 1. Thus we have uniformly in $j$,

$$\sqrt{L}|J_{jc}| = O_p(\sqrt{sr_{sn}}).$$

By combining (S.28)-(S.30), we obtain (S.27).

Step 4: We describe necessary modifications for $\sum_{j=0}^q p_\lambda(|\eta_j|_{l_2})$, where $g_j = \gamma_j^T B$. We denote $\int_0^1 B(t)B(t)^T dt$ by $\Delta_B$ and note that

$$C_1 L^{-1} \leq \lambda_{\min}(\Delta_B) \leq \lambda_{\max}(\Delta_B) \leq C_2 L^{-1}$$

for some positive $C_1$ and $C_2$. By making make a change of variables $\eta_j = \Delta_B^{1/2} \gamma_j$, $j = 0, \ldots, q$, and setting $\overline{\eta} = (\eta_0^T, \ldots, \eta_q^T)^T$, we have

$$Q_q(\overline{\eta}) = l_q(\overline{\eta}) + \sum_{j=0}^q p_\lambda(|\eta_j|), \quad \partial Q_q(\overline{\eta}) = \Delta_B^{-1/2} \frac{\partial l_q}{\partial \gamma_j} (\overline{\eta}) + \sum_{j=0}^q p_\lambda'(|\eta_j|) \frac{\eta_j}{|\eta_j|}.$$  

Then our version of (19) of [21] reduces to

$$\left(\frac{\partial l_q}{\partial \gamma_j}\right)^T \Delta_B^{-1} \left(\frac{\partial l_q}{\partial \gamma_j}\right) \lambda^{-2} \leq C L \left(\frac{\partial l_q}{\partial \gamma_j}\right)^T \left(\frac{\partial l_q}{\partial \gamma_j}\right) \lambda^{-2} \to 0$$

(S.31)
for some positive $C$. Steps 1-3 and (S.31) imply that we have the same result for $\sum_{j=0}^{q} p_{\lambda}(\|g_j\|_{L^2})$. Hence the proof of the remark is completed. \hfill \Box

Department of Mathematics  
National Taiwan University  
Taipei 106, Taiwan  
E-mail: cheng@math.ntu.edu.tw

Graduate School of Economics  
Hitotsubashi University  
Kunitachi, Tokyo 186-8601, Japan  
E-mail: t.honda@r.hit-u.ac.jp

Department of Statistics & Applied Probability  
National University of Singapore  
Singapore  
E-mail: stalj@nus.edu.sg

Department of Mathematics  
Hong Kong Baptist University  
Kowloon, Hong Kong  
E-mail: hpeng@math.hkbu.edu.hk