The Annals of Statistics
2014, Vol. 42, No. 5, 1819–1849
DOI: 10.1214/14-AOS1236

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

BY MING-YEN CHENG[1], TOSHIO HONDA[2], JIALIANG LI[3] AND HENG PENG[4]

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 for finding a sparse semivarying coefficient model, which is widely accepted for longitudinal data analysis. 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 subsequently 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.

1. Introduction. Longitudinal data arise in many modern scientific fields, including finance, genetics, medicine and so on. Specifically, we consider observing

Received August 2013; revised May 2014.

[1]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).

[2]Supported by the JSPS Grant-in-Aids for Scientific Research (A) 24243031 and (C) 25400197.

[3]Supported by Grants AcRF R-155-000-130-112 and NMRC/CBRG/0014/2012.

[4]Supported by CERG Grants from the Hong Kong Research Grants Council (HKBU 201610 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).

MSC2010 subject classifications. 62G08.

Key words and phrases. Independence screening, longitudinal data, B-spline, SCAD, sparsity, oracle property.

independent realizations of a scalar response process $y(t)$ and a p-dimensional covariate process $\mathbf{x}(t) = (x^{(1)}(t), \ldots, x^{(p)}(t))^T$ at $t = t_1, \ldots, t_m$, where $t_1, \ldots, t_m$, independent of $\mathbf{x}(t)$, are i.i.d. with density $f_T(t)$ satisfying $C_1 \le f_T(t) \le C_2$. 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 $\mathbf{x}(t)$ [12, 38]. 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

$$y(t) = \beta_0(t) + \sum_{k=1}^{p} x^{(k)}(t)\beta_k(t) + \varepsilon(t), \qquad t \in [0, 1], \tag{1.1}$$

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

$$y_i(\mathbf{t}_i) = \beta_0(\mathbf{t}_i) + \sum_{k=1}^{p} x_i^{(k)}(\mathbf{t}_i)\beta_k(\mathbf{t}_i)^T + \varepsilon_i(\mathbf{t}_i), \qquad i = 1, \ldots, n, \tag{1.2}$$

where $\varepsilon_i(\mathbf{t}_i) = (\varepsilon_i(t_{i1}), \ldots, \varepsilon_i(t_{im_i}))$ is the error process in the ith subject.

As technology evolves rapidly over the recent decades, high-dimensional longitudinal data have become commonly encountered, and the analysis poses new challenges from methodological, theoretical and computational aspects. When p is large, it is often the case that many of the covariates are irrelevant. Under such circumstances, 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 [7], Lasso [26], adaptive Lasso [39] and group Lasso [22, 36] estimators. These ideas have been adopted to select important variables in varying coefficient models for i.i.d. data, that is, $m_1 = \cdots m_n = 1$, [34]. For longitudinal data, when p is fixed, group SCAD penalized B-spline methods were studied in [29] 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, [32] 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. We consider the more general ultra-high dimensional case where p can be lager 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, which possesses the desirable sparsity and oracle properties. In parametric settings, existing screening methods include the sure independence screening procedures [8, 11], the rank correlation screening procedure [18] and others. In semiparametric settings, screening procedures have been considered for additive and varying coefficient models when the data are i.i.d. [4, 10, 21]. In the present setup, we adopt the nonparametric independence screening (NIS) idea in [4]. Moreover, the covariance structure of $\varepsilon(t)$ is unknown in general, and it is infeasible to estimate it at this stage. We base our NIS procedure on a working independence structure. 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 structure.

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

$$y_i(\mathbf{t}_i) = \beta_0(\mathbf{t}_i) + \sum_{k=1}^{q} x_i^{(k)}(\mathbf{t}_i)\beta_k(\mathbf{t}_i)^T + \varepsilon_i(\mathbf{t}_i), \qquad i = 1, \ldots, n. \tag{1.3}$$

Under appropriate smoothness assumptions, we can estimate the unknown coefficient functions in model (1.3) using B-spline smoothing [24]. However, the dimension 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 we show that 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 [8, 11]. We choose the SCAD penalty in both of the variable screening and selection steps because it enjoys a faster convergence rate than the Lasso L_1 penalty when the dimension is very large [13].

Besides variable screening and variable selection, we pay attention to the structure identification problem. That is, some of the important variables may simply have constant effects. Identifying the nonzero 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

[16, 33, 38]. In addition, structure identification was considered for partially linear additive models by [37] and for Cox proportional hazard models with varying coefficients by [20, 35]. To achieve simultaneous variable selection and structure identification, we construct a group SCAD penalty to penalize both spurious nonconstant effects and spurious nonzero effects. After this step, we further reduce the varying coefficient model (1.3) to the following semivarying coefficient model:

$$y_i(\mathbf{t}_i) = \beta_0(\mathbf{t}_i) + \sum_{k=1}^{s_1} x_i^{(k)}(\mathbf{t}_i)\beta_k + \sum_{k=s_1+1}^{s} x_i^{(k)}(\mathbf{t}_i)\beta_k(\mathbf{t}_i)^T + \varepsilon_i(\mathbf{t}_i), \tag{1.4}$$

$i = 1, \ldots, n$, where s_1 and s satisfy $0 \le s_1 \le s \ll q$, $\beta_1, \ldots, \beta_{s_1}$ 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 $\varepsilon(t)$ 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 p is fixed and the structure identification method is a model selection approach in [38]. We show both theoretically and numerically that, for p of any exponential order of n, based on working independence, 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 [4, 10, 21]. A consequence is that the dimension after the NIS step can be controlled at a moderate order which fulfills the conditions on the dimensionality in the subsequent group SCAD step. This provides the theoretical ground for our new sequential screening and variable selection approach. In addition, we discuss the computation 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 application to the yeast cell cycle data are reported and discussed in Section 5. Proofs of the theorems and some lemmas are placed in Appendix and the supplementary material [3].

2. Nonparametric independence screening. 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 $|A|_\infty = \sup_{i,j} |a_{ij}|$. 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 $\mathbf{u}(t) \in \mathbb{R}^k$ and $\mathbf{v}(t) \in \mathbb{R}^l$ by

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

and

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

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

2.1. *Nonparametric independence screening algorithm.* Consider the full model (1.2). Define the set of indices of relevant covariates by

$$\mathcal{M}_\kappa = \{k \ge 1 | \|\beta_k\|^2 \ge C_{\kappa 1} n^{-2\kappa} L\},$$

for some positive constant κ. 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 [10], we consider for each $k = 1, \ldots, p$ a marginal model for $y(t)$ and $x^{(k)}(t)$ defined by

$$y(t) = a_k(t) + b_k(t)x^{(k)}(t) + \eta_k(t), \tag{2.1}$$

where $a_k(t)$ and $b_k(t)$ are given by $\arg\min_{a,b \in \mathcal{L}^2[0,1]} \|y - a - bx^{(k)}\|^2$. Alternatively, [21] employed a conditional correlation approach. Let $\mathbf{B}(t) = (B_1(t), \ldots, B_L(t))^T$ be an equispaced B-spline basis of order 3 on $[0, 1]$, where L is the dimension of the basis. Write $\mathbf{B}(\mathbf{t}_i) = (\mathbf{B}(t_{i1}), \ldots, \mathbf{B}(t_{im_i}))$. Then, under the smoothness conditions specified in Assumption M1, $b_k(t)$ in (2.1) can be approximated by some linear combination of $\mathbf{B}(t)$. Thus, we can estimate b_k by minimizing the following objective function:

$$\|y - \gamma_1^T \mathbf{B} - \gamma_2^T x^{(k)} \mathbf{B}\|_n^2 \equiv \|y - \gamma_1^T \mathbf{B} - \gamma_2^T \mathbf{W}_k\|_n^2, \qquad \gamma_1, \gamma_2 \in \mathbb{R}^L. \tag{2.2}$$

Note that the regressors $\mathbf{W}_k(t)$ and $\mathbf{W}_k(\mathbf{t})$, and their sample versions $\mathbf{W}_{ik}(t)$ and $\mathbf{W}_{ik}(\mathbf{t}_i)$, for the above B-spline estimation of b_k are given by

$$\begin{aligned}
\mathbf{W}_k(t) &= x^{(k)}(t)\mathbf{B}(t) = \big(W_{k1}(t), \ldots, W_{kL}(t)\big)^T \in \mathbb{R}^L, \\
\mathbf{W}_k(\mathbf{t}) &= \big(\mathbf{W}_k(t_1), \ldots, \mathbf{W}_k(t_m)\big) \in \mathbb{R}^{L \times m}, \\
\mathbf{W}_{ik}(t) &= \big(W_{ik1}(t), \ldots, W_{ikL}(t)\big)^T = x_i^{(k)}(t)\mathbf{B}(t) \in \mathbb{R}^L, \qquad i = 1, \ldots, n, \\
\mathbf{W}_{ik}(\mathbf{t}_i) &= \big(\mathbf{W}_{ik}(t_{i1}), \ldots, \mathbf{W}_{ik}(t_{im_i})\big) \in \mathbb{R}^{L \times m_i}, \qquad i = 1, \ldots, n.
\end{aligned} \tag{2.3}$$

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

$$\widehat{b}_k(t) = \widehat{\gamma}_{2k}^T \mathbf{B}(t). \tag{2.4}$$

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

$$\widehat{\mathcal{M}}_\kappa = \{k \geq 1 | \|\widehat{b}_k\|_n^2 \geq C_{\kappa 3} n^{-2\kappa} L\} \tag{2.5}$$

for some sufficiently small positive constant $C_{\kappa 3}$ satisfying $C_{\kappa 3} < C_{\kappa 2}/2$, where $C_{\kappa 2}$ is given in Assumption M2(1).

Intuitively, we may still have too many irrelevant variables kept in the analysis if the threshold parameter κ in (2.5) is chosen too large. On the other hand, we may run into the danger of screening out some of the relevant variables if it is chosen too small. In the literature, the screening step is immediately followed by the model fitting step, thus some iterative screening procedure is employed to control the false selection rate [8, 11]. 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 Section 2.2 guarantees that, with proper choices of κ and L, by the first screening step we can reduce the dimensionality to a moderate order with the false negative rate under control. This allows the next variable selection step to possess the sparsity and oracle properties given in Section 3.2. In practice, we sort the $\|\widehat{b}_k\|_n^2$'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, we took $\alpha = 1$.

2.2. *Theory of the proposed NIS procedure.* Here, we collect the technical assumptions on the marginal models given in (2.1). Let $\mathbf{I}_d$ denote the identity matrix of dimension d and let $\#A$ 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 the κ in the definition of $\mathcal{M}_\kappa$, 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 $\|\beta_k\|^2 = 0$.
(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 [4]. We mention that, in Assumption M2(2), we require that $\|\beta_k\|^2 = 0$ merely for simplicity of presentation, and it is sufficient to replace it with $\|\beta_k\|^2 = o(n^{-2\kappa} L)$.

ASSUMPTION T. (1) For some positive constant C_{T1}, we have $C_{T1} \le \mathrm{E}[\{\widetilde{x}^{(k)}(t)\}^2]$ for $t \in [0,1]$ and $k = 1, \ldots, p$, where $\widetilde{x}^{(k)}(t) = x^{(k)}(t) - \mathrm{E}\{x^{(k)}(t)\}$.

(2) For any positive constant C_1, there is a positive constant C_2 such that

$$\mathrm{E}\big[\exp\{C_1|\varepsilon_i(\mathbf{t}_i)|/\sqrt{m_i}\}|\mathbf{x}_i(\mathbf{t}_i), \mathbf{t}_i\big] < C_2.$$

(3) Let $\mathbf{x}_{\mathcal{M}}(t)$ be the covariate vector consisting of 1 and all the covariates in $\mathcal{M}_\kappa$. Then there is a positive constant C_{T2} such that

$$\mathrm{E}\{\mathbf{x}_{\mathcal{M}}(t)\mathbf{x}_{\mathcal{M}}(t)^T\} \ge C_{T2}\mathbf{I}_{\#\mathcal{M}_\kappa+1} \qquad \text{for any } t \in [0,1].$$

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

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

(6) The functions $\beta_k(t)$, $k = 0, \ldots, p$, are twice continuously differentiable. Besides, $\sum_{k\in\mathcal{A}}\sup_{t\in[0,1]}|\beta_k''(t)| \le C_{T5}$ for some positive constant C_{T5}.

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

$$C_1\mathbf{I}_2 \le \mathrm{E}\begin{pmatrix} 1 & x^{(k)}(t) \\ x^{(k)}(t) & \{x^{(k)}(t)\}^2 \end{pmatrix} \le C_2\mathbf{I}_2, \qquad t \in [0,1] \text{ and } 1 \le k \le p. \tag{2.6}$$

Assumption T(2), (4) and (5) are technical assumptions needed in order to apply the exponential inequalities. Note that $\#\mathcal{A}$ may increase, but we assume that the signal $\sum_{k\in\mathcal{A}}x^{(k)}(t)\beta_k(t)$ should not diverge by imposing Assumption T(5) and (6). Similar conditions are made in Assumption D of [4]. 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 Assumption T(5) and (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*

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

where C_{p1} and C_{p2} are positive constants depending 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 $\mathrm{P}(\mathcal{M}_\kappa \subset \widehat{\mathcal{M}}_\kappa) \to 1$, if κ 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 κ is that

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

Thus, the proposed screening procedure may reduce the number of covariates drastically. However, $\#\widehat{\mathcal{M}}_\kappa$ may be still too large to apply any variable selection procedures with consistency property. Fortunately, we have succeeded in giving an upper bound on $\#\widehat{\mathcal{M}}_\kappa$, 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 [10] 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 [21] the condition on κ 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 [10] (no upper bound is provided in [21]), which leads to (2.10). We succeed 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

$$\overline{\mathbf{W}}_k = \mathbf{W}_k - A_k \mathbf{B} \quad \text{and} \quad \overline{\mathbf{W}} = \bigl(\overline{\mathbf{W}}_1^T, \ldots, \overline{\mathbf{W}}_p^T\bigr)^T, \tag{2.8}$$

where $A_k = \langle \mathbf{W}_k, \mathbf{B}^T \rangle \langle \mathbf{B}, \mathbf{B}^T \rangle^{-1}$. Note that $\overline{\mathbf{W}}_k$ and $\overline{\mathbf{W}}$ are, respectively, L- and pL-dimensional stochastic processes on $[0, 1]$. Besides we define

$$\overline{\Sigma} = \langle \overline{\mathbf{W}}, \overline{\mathbf{W}}^T \rangle, \tag{2.9}$$

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

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

$$\mathrm{P}\bigl(\#\widehat{\mathcal{M}}_\kappa \le C_{\kappa 4} n^{2\kappa} \lambda_{\max}(\overline{\Sigma})\bigr) \ge 1 - C_{p1} pL \exp\bigl(-C_{p2} n^{1-4\kappa} L\bigr),$$

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 $\widehat{\mathcal{M}}_\kappa$ merely because they have large values of $\|b_k\|$, the number of such variables is limited. Furthermore, they will be removed by the subsequent variable selection procedure given in Section 3.

Define $\widetilde{\mathbf{W}}_k(t)$, $\widetilde{\mathbf{W}}_k(\mathbf{t})$, $\widetilde{\mathbf{W}}_{ik}(t)$ and $\widetilde{\mathbf{W}}_{ik}(\mathbf{t}_i)$ by replacing $x^{(k)}(t)$ and $x_i^{(k)}(t)$ in the definitions of $\mathbf{W}_k(t)$, $\mathbf{W}_k(\mathbf{t})$, $\mathbf{W}_{ik}(t)$ and $\mathbf{W}_{ik}(\mathbf{t}_i)$ given in (2.3) with

$$\widetilde{x}^{(k)}(t) = x^{(k)}(t) - \mathrm{E}\{x^{(k)}(t)\} \quad \text{and} \quad \widetilde{x}_i^{(k)}(t) = x_i^{(k)}(t) - \mathrm{E}\{x_i^{(k)}(t)\},$$

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

$$\langle \widetilde{\mathbf{W}}, \widetilde{\mathbf{W}}^T \rangle \le \overline{\Sigma} \le \langle \mathbf{W}, \mathbf{W}^T \rangle,$$

where $\widetilde{\mathbf{W}} = (\widetilde{\mathbf{W}}_1^T, \ldots, \widetilde{\mathbf{W}}_p^T)^T$ and $\mathbf{W} = (\mathbf{W}_1^T, \ldots, \mathbf{W}_p^T)^T$. The maximum eigenvalue of $\langle \mathbf{W}, \mathbf{W}^T \rangle$ may tend to infinity very quickly with p. However, since we do a kind of centerization to $\mathbf{W}$ and obtain $\overline{\mathbf{W}}$ as in (2.8), we conjecture that $\overline{\Sigma}$ is very close to $\langle \widetilde{\mathbf{W}}, \widetilde{\mathbf{W}}^T \rangle$ under some regularity conditions. If the maximum eigenvalues of the two matrices have the same order, and for some positive K_n, $\lambda_{\max}(\mathrm{E}\{\widetilde{\mathbf{x}}(t)\widetilde{\mathbf{x}}(t)^T\}) \le K_n$ uniformly in t, then

$$\lambda_{\max}(\overline{\Sigma}) \le C_1 L^{-1} K_n \quad \text{and} \quad \#\widehat{M}_\kappa \le C_2 n^{2\kappa} L^{-1} K_n$$

with probability tending to 1. Suppose L is chosen to be of the optimal order $n^{1/5}$. Then, for κ satisfying condition (2.7), this implies that

$$\#\widehat{M} = O_p\big(n^{2/5-\eta}\big) K_n \qquad \text{for some } 0 < \eta < 2/5. \tag{2.10}$$

Thus, when K_n is bounded, $\#\widehat{M}$ 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 κ, 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 $y_i(\mathbf{t}_i), x_i^{(k)}(\mathbf{t}_i)$, $i = 1, \ldots, n$, $k = 1, \ldots, q$ and consider the varying coefficient model (1.3). To estimate the coefficient functions $\beta_k(t), k = 0, 1, \ldots, q$, first we define

$$l_q(\boldsymbol{\gamma}) = \Big\| y - \gamma_0^T \mathbf{B} - \sum_{k=1}^{q} \gamma_k^T \mathbf{W}_k \Big\|_n^2, \tag{3.1}$$

where $\boldsymbol{\gamma} = (\gamma_0^T, \ldots, \gamma_q^T)^T \in \mathbb{R}^{(q+1)L}$. When q is fixed and sufficiently small, based on working independence, we can estimate $\beta_k(t)$ by minimizing the objective function $l_q(\boldsymbol{\gamma})$. Denoting the minimizer by $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\gamma}_0^T, \ldots, \widetilde{\gamma}_q^T)^T$, for $k = 0, 1, \ldots, q$, we can estimate $\beta_k(t)$ by

$$\widetilde{\beta}_k(t) = \widetilde{\gamma}_k^T \mathbf{B}(t). \tag{3.2}$$

Recall that L is the dimension of the B-spline basis. Suppose q satisfies

$$q = o\bigl(\sqrt{n/(L \log n)}\bigr). \tag{3.3}$$

This restriction is necessary since the Hessian matrix of the objective function $l_q(\boldsymbol{\gamma})$ given in (3.1) must be positive definite. Note that [29] imposed similar conditions in the case where q is a fixed constant. When L is taken as the optimal order $n^{1/5}$, condition (3.3) reduces to

$$q = o\bigl(n^{2/5}(\log n)^{-1/2}\bigr). \tag{3.4}$$

When q is relatively large and a lot of the covariates seem to be irrelevant, we would add a penalty term to $l_q(\boldsymbol{\gamma})$ 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. That is, some of the coefficient functions 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 [35] 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 $g_k(t) \equiv \gamma_k^T \mathbf{B}(t)$ with respect to the L_2 norm by

$$g_k(t) = (g_k)_c + (g_k)_f(t), \tag{3.5}$$

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

$$p_\lambda(u) = \begin{cases} \lambda u, & \text{if } 0 \le u \le \lambda, \\ -\bigl(u^2 - 2a_0\lambda u + \lambda^2\bigr)/\bigl\{2(a_0 - 1)\bigr\}, & \text{if } \lambda < u \le a_0\lambda, \\ (a_0 + 1)\lambda^2/2, & \text{if } u > a_0\lambda, \end{cases} \tag{3.6}$$

where a_0 is a constant larger than 1. We take $a_0 = 3.7$ as suggested by [7]. 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

$g_k = \gamma_k^T \mathbf{B}$, $k = 0, \ldots, q$. We specify the values of λ_1 and λ_2 later. Our objective function for our group SCAD estimator is then given by

$$Q_q(\boldsymbol{\gamma}) = l_q(\boldsymbol{\gamma}) + \sum_{k=1}^{q} \{p_{\lambda_1}(|(g_k)_c|) + p_{\lambda_2}(\|(g_k)_f\|_{L_2})\}. \tag{3.7}$$

Based on $Q_q(\boldsymbol{\gamma})$, we can carry out variable selection, structure identification, and estimation simultaneously by the following procedure:

$$\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_0^T, \ldots, \widehat{\gamma}_q^T)^T = \underset{\boldsymbol{\gamma} \in \mathbb{R}^{(q+1)L}}{\arg\min}\ Q_q(\boldsymbol{\gamma}) \quad \text{and} \quad \widehat{\beta}_k = \widehat{\gamma}_k^T \mathbf{B}, k = 0, \ldots, q. \tag{3.8}$$

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

To compute the group SCAD estimator given in (3.8), we use the approximation to the SCAD function suggested in [7]: $p_\lambda(u) \approx p_\lambda(u_0) + \frac{1}{2}(p'_\lambda(u_0)/u_0)(u^2 - u_0^2)$, for u in a neighborhood of any given $u_0 \in \mathbb{R}^+$. Define $\tau_j = \tau^{-1} \int_0^1 B_j(t)\,dt$, and $\overline{B}_j = \sqrt{L}(B_j - \tau_1^{-1}\tau_j B_1)$, $j = 0, 1, \ldots, L$, 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 group SCAD 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 $\widehat{\boldsymbol{\gamma}}$. To select the tuning parameters λ_1 and λ_2 in (3.7), we treat them as equal $\lambda_1 = \lambda_2 = \lambda$ and use the BIC criterion to select λ: $\mathrm{BIC}(\lambda) = \|y - \widehat{\beta}_0(\lambda) - \sum_{k=1}^q \widehat{\beta}_k(\lambda) x^{(k)}\|_n^2 + \mathcal{K} \log \mathcal{N}$, where $\widehat{\beta}_k(\lambda)$ is the group SCAD estimate based on λ, $\mathcal{K}$ is the number of parameters in the fitted model, and $\mathcal{N} = \sum_{i=1}^n m_i$. A similar BIC criterion was suggested by [27], and a generalized information criterion was considered by [14] 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 given in (3.8). The proofs of the theorems are deferred to Appendix. 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 hold:

for $k = 1, \ldots, s$, $|(\beta_k)_c|/\lambda_1 \to \infty$ if $|(\beta_k)_c| > 0$ and $\|(\beta_k)_f\|_{L_2}/\lambda_2 \to \infty$ if $\|(\beta_k)_f\|_{L_2} > 0$; for $k = s+1, \ldots, q$, $\|\beta_k\|_{L_2} = 0$; the above divergence is uniform in $k = 0, 1, \ldots, s$.

(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

$$\mathbf{G} = \{\mathbf{g} = (g_0, \ldots, g_q)^T | g_k = \gamma_k^T \mathbf{B}, k = 0, 1, \ldots, q\}$$

and $\mathbf{G}_0$, which we may call the oracle space under Assumption S, by

$$\mathbf{G}_0 = \{\mathbf{g} \in \mathbf{G} | (g_k)_c = 0 \text{ if } |(\beta_k)_c| = 0 \text{ and } (g_k)_f = 0 \text{ if } \|(\beta_k)_f\|_{L_2} = 0, \\ k = 1, \ldots, q\}.$$

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

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

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

We state two technical assumptions here. Set

$$\Sigma_n = \langle (\mathbf{B}^T \mathbf{W}^T)^T, (\mathbf{B}^T \mathbf{W}^T) \rangle_n \quad \text{and} \quad \Sigma = \mathrm{E}\{\Sigma_n\}, \tag{3.10}$$

where $\mathbf{W} = (\mathbf{W}_1^T, \ldots, \mathbf{W}_q^T)^T$. A sufficient condition for Assumption V is $\lambda_{\min} \mathrm{E}\{\mathbf{x}(t)\mathbf{x}(t)^T\} \ge C$ uniformly in t for some positive C.

ASSUMPTION E. There is a positive constant C_E such that uniformly in t, $\mathrm{E}\{\varepsilon_i(\mathbf{t}_i)\varepsilon_i(\mathbf{t}_i)^T | \mathbf{x}_i(\mathbf{t}_i), \mathbf{t}_i\} \le C_E \mathbf{I}_{m_i}$.

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

In Theorem 3.2, we derive the L_2 convergence rate of the group SCAD estimator given in (3.8). 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*

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

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* $Q_q(\boldsymbol{\gamma})$ *on* $\mathbb{R}^{(q+1)L}$, *denoted by* $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_0^T, \ldots, \widehat{\gamma}_q^T)^T$, *such that*

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

Next, we define the sparsity and the oracle properties of estimators.

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

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

ORACLE PROPERTY. If we knew the value of s in Assumption S(1), we would use the knowledge and minimize $l_q(\boldsymbol{\gamma})$ on the subspace of $\mathbb{R}^{(q+1)L}$ corresponding to $\mathbf{G}_0$. 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* $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_0^T, \ldots, \widehat{\gamma}_q^T)^T$ *of* $Q_q(\boldsymbol{\gamma})$ *satisfying* $\|(\widehat{\gamma}_0^T\mathbf{B}, \ldots, \widehat{\gamma}_q^T\mathbf{B})^T - (\beta_0, \ldots, \beta_q)^T\|_{L_2} \le \eta_n r_{qn}$ *is equal to the oracle estimator. We also have*

$$\big(\widehat{\gamma}_0^T\mathbf{B}, \ldots, \widehat{\gamma}_q^T\mathbf{B}\big)^T \in \mathbf{G}_0$$

and

$$\sum_{k=0}^{s} \big\|\widehat{\gamma}_k^T\mathbf{B} - \beta_k\big\|_{L_2}^2 = O_p\big(r_{sn}^2\big).$$

Since $Q_q(\gamma)$ may not be concave, there may be another local minimizer of $Q_q(\boldsymbol{\gamma})$ outside $\{\boldsymbol{\gamma} \in \mathbb{R}^{(q+1)L} | \|(\gamma_0^T\mathbf{B}, \ldots, \gamma_q^T\mathbf{B})^T - (\beta_0, \ldots, \beta_q)^T\|_{L_2} \le \eta_n r_{qn}\}$.

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, $L = c_L n^{1/5}$ 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, and no structure identification. Then the penalty term in the objective function $Q_q(\boldsymbol{\gamma})$ 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 $l_s(\boldsymbol{\gamma}_1) = \|y - \gamma_0^T\mathbf{B} - \sum_{k=1}^{s} \gamma_k^T\mathbf{W}_k\|_n^2$ on $\mathbb{R}^{(s+1)L}$ by $\widehat{\boldsymbol{\gamma}}_1 \in \mathbb{R}^{(s+1)L}$. Then, with probability tending to 1, $(\widehat{\boldsymbol{\gamma}}_1^T, \mathbf{0}^T)^T \in \mathbb{R}^{(q+1)L}$ is a local minimizer of $Q_q(\overline{\boldsymbol{\gamma}})$,

where $\overline{\boldsymbol{\gamma}} \in \mathbb{R}^{(q+1)L}$. Thus, some flexibility will be allowed in the tuning parameter selection when s is bounded. The proof of this result is outlined in the supplementary material [3].

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 $\boldsymbol{\beta}_1 \in \mathbb{R}^{s_1}$ and $\mathbf{x}_1$, and denote, respectively, the functional coefficients and the corresponding variables by $\boldsymbol{\beta}_2(t) \in \mathbb{R}^{s_2}$ and $\mathbf{x}_2$. Then we can rewrite model (1.4) as the following:

$$y_i(t_{ij}) = \beta_0(t_{ij}) + \mathbf{x}_{1i}(t_{ij})^T \boldsymbol{\beta}_1 + \mathbf{x}_{2i}(t_{ij})^T \boldsymbol{\beta}_2(t_{ij}) + \varepsilon_i(t_{ij}), \quad j = 1, \ldots, m_i, \tag{4.1}$$

where $\mathbf{x}_{1i}(t_{ij})$ and $\mathbf{x}_{2i}(t_{ij})$ denote, respectively, the observations on $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ in the ith subject at time t_{ij}. When $\mathbf{x}_1$ and $\mathbf{x}_2$ are given, and s_1 and s_2 are fixed and small, this estimation problem has been extensively studied in the literature [19, 38]. 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 variables in $\mathbf{x}_1$ have constant coefficients with $|(\widehat{\beta}_k)_c| > 0$ and $\|(\widehat{\beta}_k)_f\|_{L_2} = 0$, and those in $\mathbf{x}_2$ have time-varying coefficients with $\|(\widehat{\beta}_k)_f\|_{L_2} > 0$. Note that when $|(\widehat{\beta}_k)_c| = 0$ and $\|(\widehat{\beta}_k)_f\|_{L_2} > 0$ the constant part is zero, but we still include the variable in $\mathbf{x}_2$ 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 given in Section 4.1 are obtained, we may also iterate between the covariance function estimation step and the coefficient estimation step 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 [5, 17]. Recall that $\mathbf{t}_i = (t_{i1}, \ldots, t_{im_i})$, $y_i(\mathbf{t}_i) = (y_i(t_{i1}), \ldots, y_i(t_{im_i}))$, and $\varepsilon_i(\mathbf{t}_i) = (\varepsilon_i(t_{i1}), \ldots, \varepsilon_i(t_{im_i}))$. Let K denote a kernel function, which is usually taken as a symmetric p.d.f., and take a bandwidth $h_1 > 0$. For any given

$\boldsymbol{\beta}_1 \in \mathbb{R}^{s_1}$, we can estimate $\beta_0(t)$ and $\boldsymbol{\beta}_2(t)$ in model (4.1) by minimizing the following local sum of squares:

$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{m_i}\bigl\{y_i(t_{ij}) - \mathbf{x}_{1i}(t_{ij})^T\boldsymbol{\beta}_1 - \bigl(1, \mathbf{x}_{2i}(t_{ij})^T\bigr)\bigl(\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1(t_{ij}-t)\bigr)\bigr\}^2 K_{h_1}(t_{ij}-t)\\
&\quad= \sum_{i=1}^{n}\bigl\{y_i(\mathbf{t}_i)^T - \mathbf{x}_{1i}(t_{ij})^T\boldsymbol{\beta}_1 - \bigl(\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T, \mathbf{T}_i(t)\bigl(\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T\bigr)\bigr)\boldsymbol{\alpha}\bigr\}^T\\
&\qquad\times \mathbf{W}_{ih_1}(t)\bigl\{y_i(\mathbf{t}_i)^T - \mathbf{x}_{1i}(t_{ij})^T\boldsymbol{\beta}_1\\
&\qquad\qquad - \bigl(\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T, \mathbf{T}_i(t)\bigl(\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T\bigr)\bigr)\boldsymbol{\alpha}\bigr\},
\end{aligned}
\tag{4.2}
$$

where $K_{h_1}(\cdot) = K(\cdot/h_1)/h_1$, $\boldsymbol{\alpha}_0, \boldsymbol{\alpha}_1 \in \mathbb{R}^{s_2+1}$, $\mathbf{1}_{m_i}$ is the m_i-dimensional one-vector, $\mathbf{T}_i = \operatorname{diag}\{t_{i1}-t, \ldots, t_{im_i}-t\}$, $\mathbf{W}_{ih_1}(t) = \operatorname{diag}\{K_{h_1}(t_{i1}-t), \ldots, K_{h_1} \times (t_{im_i}-t)\}$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^T, \boldsymbol{\alpha}_1^T)^T$. Let $\mathbf{0}_{k\times l}$ be the $k \times l$ dimensional zero matrix. For any given $\boldsymbol{\beta}_1 \in \mathbb{R}^{s_1}$, denote the minimizer of (4.2) by $\widetilde{\boldsymbol{\alpha}}(t, \boldsymbol{\beta}_1)$. Then an estimator of $(\beta_0(t), \boldsymbol{\beta}_2(t)^T)^T$ is

$$
\begin{aligned}
&\bigl(\widetilde{\beta}_0(t, \boldsymbol{\beta}_1), \widetilde{\boldsymbol{\beta}}_2(t, \boldsymbol{\beta}_1)^T\bigr)^T\\
&\quad= (\mathbf{I}_{s_2+1}, \mathbf{0}_{(s_2+1)\times(s_2+1)})\widetilde{\boldsymbol{\alpha}}(t, \boldsymbol{\beta}_1)\\
&\quad= (\mathbf{I}_{s_2+1}, \mathbf{0}_{(s_2+1)\times(s_2+1)})\bigl(\mathbf{V}(t)^T\mathbf{W}_{h_1}(t)\mathbf{V}(t)\bigr)^{-1}\mathbf{V}(t)^T\mathbf{W}_{h_1}(t)\\
&\qquad\times (\mathbf{Y} - \mathbf{X}_1\boldsymbol{\beta}_1),
\end{aligned}
\tag{4.3}
$$

where $\mathbf{Y} = (y_1(\mathbf{t}_1), \ldots, y_1(\mathbf{t}_n))^T$, $\mathbf{X}_1 = (\mathbf{x}_{11}(\mathbf{t}_1), \ldots, \mathbf{x}_{1n}(\mathbf{t}_n))^T$, and $\mathbf{V}(t) = (\mathbf{V}_{1t}^T, \ldots, \mathbf{V}_{nt}^T)^T$ with $\mathbf{V}_{it} = (\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T, \mathbf{T}_i(t)(\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T))$, $i = 1, \ldots, n$.

Then, based on working independence, the initial profile least squares estimator for the constant coefficients $\boldsymbol{\beta}_1$ in model (4.1) is defined as

$$
\widetilde{\boldsymbol{\beta}}_1^{\mathrm{PLS}} = \arg\min_{\boldsymbol{\beta}_1\in\mathbb{R}^{s_1}} \sum_{i=1}^{n}\sum_{j=1}^{m_i}\bigl\{y_i(t_{ij}) - \mathbf{x}_{1i}(t_{ij})^T\boldsymbol{\beta}_1 - \widetilde{\beta}_0(t_{ij}, \boldsymbol{\beta}_1) - \mathbf{x}_{2i}(t_{ij})^T\widetilde{\boldsymbol{\beta}}_2(t_{ij}, \boldsymbol{\beta}_1)\bigr\}^2,
$$

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

$$
\widetilde{\boldsymbol{\beta}}_1^{\mathrm{PLS}} = \arg\min_{\boldsymbol{\beta}_1\in\mathbb{R}^{s_1}}\bigl\{(\mathcal{I} - \mathcal{S})(\mathbf{Y} - \mathbf{X}_1\boldsymbol{\beta}_1)\bigr\}^T\bigl\{(\mathcal{I} - \mathcal{S})(\mathbf{Y} - \mathbf{X}_1\boldsymbol{\beta}_1)\bigr\},
$$

where $\mathcal{S} = (\mathbf{S}_{11}^T, \ldots, \mathbf{S}_{1m_1}^T, \ldots, \mathbf{S}_{n1}^T, \ldots, \mathbf{S}_{nm_n}^T)^T$, $\mathbf{S}_{ij} = (1, \mathbf{x}_{2i}(t_{ij})^T, \mathbf{0}_{1\times(s_2+1)}) \times (\mathbf{V}(t_{ij})^T\mathbf{W}_{h_1}(t_{ij})\mathbf{V}(t_{ij}))^{-1}\mathbf{V}(t_{ij})^T\mathbf{W}_{h_1}(t_{ij})$, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$, and $\mathcal{I} = \mathbf{I}_{m_1+\cdots+m_n}$. Thus, we have

$$
\widetilde{\boldsymbol{\beta}}_1^{\mathrm{PLS}} = \bigl\{\mathbf{X}_1^T(\mathcal{I} - \mathcal{S})^T(\mathcal{I} - \mathcal{S})\mathbf{X}_1\bigr\}^{-1}\mathbf{X}_1^T(\mathcal{I} - \mathcal{S})^T(\mathcal{I} - \mathcal{S})\mathbf{Y},
\tag{4.4}
$$

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

$$\begin{aligned}&\big(\widetilde{\beta}_0^{\mathrm{PLS}}(t), \widetilde{\boldsymbol{\beta}}_2^{\mathrm{PLS}}(t)^T\big)^T\\ &\quad= (\mathbf{I}_{s_2+1}, \mathbf{0}_{(s_2+1)\times(s_2+1)})\big(\mathbf{V}(t)^T\mathbf{W}_{h_1}(t)\mathbf{V}(t)\big)^{-1}\mathbf{V}(t)^T\mathbf{W}_{h_1}(t)\\ &\qquad\times\big(\mathbf{Y}-\mathbf{X}_1\widetilde{\boldsymbol{\beta}}_1^{\mathrm{PLS}}\big).\end{aligned}\tag{4.5}$$

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 $\widetilde{\boldsymbol{\beta}}_1^{\mathrm{PLS}}$ is not semiparametric efficient when the error process is indeed dependent [31]. In the following sections, we estimate the covariance function of the error process using residuals obtained from the initial estimators $\widetilde{\boldsymbol{\beta}}_1^{\mathrm{PLS}}$, $\widetilde{\beta}_0^{\mathrm{PLS}}(t)$ and $\widetilde{\boldsymbol{\beta}}_2^{\mathrm{PLS}}(t)$, and then construct semiparametric efficient estimators. The semiparametric efficiency results in [19] concern generalized partially linear models and carry over to the considered semivarying coefficient models.

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

$$\varepsilon(t) = \varepsilon_1(t) + \varepsilon_2(t),$$

where $\varepsilon_1(t)$ has a smooth covariance function $\psi(s,t)$ and $\varepsilon_2(t)$ models the measurement error process. Write the residuals obtained from the initial profile least squares estimators $\widetilde{\boldsymbol{\beta}}_1^{\mathrm{PLS}}$ and $\widetilde{\boldsymbol{\beta}}_2^{\mathrm{PLS}}(t_{ij})$ given in Section 4.1 as

$$\widehat{\varepsilon}_{ij} = y(t_{ij}) - \mathbf{x}_{1i}(t_{ij})^T\widetilde{\boldsymbol{\beta}}_1^{\mathrm{PLS}} - \widetilde{\beta}_0^{\mathrm{PLS}}(t_{ij}) - \mathbf{x}_{2i}(t_{ij})^T\widetilde{\boldsymbol{\beta}}_2^{\mathrm{PLS}}(t_{ij}),\tag{4.6}$$

$i = 1, \ldots, n$, $j = 1, \ldots, m_i$. We can estimate ϕ based on these residuals. There exist (semi)parametric approaches to covariance estimation for longitudinal data [6, 12]. 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 [15] to estimate the covariance function ϕ based on the residuals. First, noting that

$$\begin{aligned}\phi(t_{ij}, t_{ik}) = \psi(t_{ij}, t_{ik}) + \mathrm{Var}\big(\varepsilon_2(t_{ij})\big)I(t_{ij} = t_{ik}),\\ i = 1, \ldots, n,\ j, k = 1, \ldots, m_i,\end{aligned}$$

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

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{j\neq k}\big\{\widehat{\varepsilon}_{ij}\widehat{\varepsilon}_{ik} - a - b(t_{ij}-u) - c(t_{ik}-v)\big\}^2\\ &\qquad\times K_{h_2}(t_{ij}-u)K_{h_2}(t_{ik}-v),\end{aligned}\tag{4.7}$$

with a bandwidth $h_2 > 0$. An explicit formulae for $\widetilde{\psi}(u,v)$ is available [15]. The covariance function estimate $\widetilde{\psi}(u,v)$ is not positive semidefinite in general. We can modify this estimate by truncating the negative components in its spectral decomposition $\widehat{\psi}(u,v) = \sum_{k=1}^{\zeta_n} \widetilde{\omega}_k \widetilde{\psi}_k(u)\widetilde{\psi}_k(v)$, where $\widetilde{\omega}_1 \geq \widetilde{\omega}_2 \geq \cdots$ are the eigenvalues of the operator $\widetilde{\psi}$, given by $(\widetilde{\psi}\alpha)(u) = \int_{[0,1]} \alpha(v)\widetilde{\psi}(u,v)\,dv$ for $\alpha \in L_2([0,1])$, $\widetilde{\psi}_j$ is the eigenfunction corresponding to $\widetilde{\omega}_j$, $j = 1,2,\ldots,$ and $\zeta_n = \max\{k : \widetilde{\omega}_j > 0, j = 1,\ldots,k\}$.

Next, we estimate the variance function of the error process $\varepsilon(t)$: $\sigma^2(t) \equiv \mathrm{Var}(\varepsilon_2(t))$ by $\widehat{\sigma}^2(t) \equiv \widehat{a}$ where $\widehat{a}$ is defined by

$$(\widehat{a},\widehat{b})^T = \underset{(a,b)^T \in \mathbb{R}^2}{\arg\min} \sum_{i=1}^{n}\sum_{j=1}^{m_i} \{\widehat{\varepsilon}_{ij}^2 - a - b(t_{ij}-t)\}^2 K_{h_3}(t_{ij}-t), \tag{4.8}$$

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

$$\widehat{\phi}(u,v) = \widehat{\psi}(u,v)I(u \neq v) + \widehat{\sigma}^2(u)I(u = v). \tag{4.9}$$

To select the bandwidths h_2 and h_3 in (4.7) and (4.8) we can 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 $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2(t)$ using $\widehat{\phi}(u,v)$ given in (4.9). Let $\widehat{\boldsymbol{\Lambda}}_i = (\widehat{\phi}(t_{ij},t_{ik}))_{j,k=1,\ldots,m_i}$, $i = 1,\ldots,n$. For any $\boldsymbol{\beta}_1 \in \mathbb{R}^{s_1}$, define $\widehat{\boldsymbol{\alpha}}(t,\boldsymbol{\beta}_1)$ as the minimizer of the following objective function of $\boldsymbol{\alpha} \in \mathbb{R}^{2(s_2+1)}$:

$$\begin{aligned}
&\sum_{i=1}^{n}\{y_i(\mathbf{t}_i)^T - \mathbf{x}_{1i}(t_{ij})^T\boldsymbol{\beta}_1 - (\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T, \mathbf{T}_i(t)(\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T))\boldsymbol{\alpha}\}^T \widehat{\boldsymbol{\Lambda}}_i^{-1/2} \\
&\quad \times \mathbf{W}_{ih_1}(t)\widehat{\boldsymbol{\Lambda}}_i^{-1/2}\{y_i(\mathbf{t}_i)^T - \mathbf{x}_{1i}(t_{ij})^T\boldsymbol{\beta}_1 \\
&\qquad\qquad - (\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T, \mathbf{T}_i(t)(\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T))\boldsymbol{\alpha}\}.
\end{aligned}$$

Then, given $\boldsymbol{\beta}_1 \in \mathbb{R}^{s_1}$, an estimator for $(\beta_0(t), \boldsymbol{\beta}_2(t)^T)^T$ is taken as

$$\begin{aligned}
&(\widehat{\beta}_0(t,\boldsymbol{\beta}_1), \widehat{\boldsymbol{\beta}}_2(t,\boldsymbol{\beta}_1)^T)^T \\
&\quad = (\mathbf{I}_{s_2+1}, \mathbf{0}_{(s_2+1)\times(s_2+1)})\widehat{\boldsymbol{\alpha}}(t,\boldsymbol{\beta}_1) \\
&\quad = (\mathbf{I}_{s_2+1}, \mathbf{0}_{(s_2+1)\times(s_2+1)})(\widehat{\mathbf{V}}(t)^T\mathbf{W}_{h_1}(t)\widehat{\mathbf{V}}(t))^{-1}\widehat{\mathbf{V}}(t)^T\mathbf{W}_{h_1}(t)\widehat{\boldsymbol{\Lambda}}^{-1/2} \\
&\qquad \times (\mathbf{Y} - \mathbf{X}_1\boldsymbol{\beta}_1),
\end{aligned} \tag{4.10}$$

where $\widehat{\boldsymbol{\Lambda}}^{-1/2} = \mathrm{diag}\{\widehat{\boldsymbol{\Lambda}}_1^{-1/2}, \ldots, \widehat{\boldsymbol{\Lambda}}_n^{-1/2}\}$, and $\widehat{\mathbf{V}}(t) = (\widehat{\mathbf{V}}_{1t}^T, \ldots, \widehat{\mathbf{V}}_{nt}^T)^T$ with $\widehat{\mathbf{V}}_{it} = (\widehat{\boldsymbol{\Lambda}}_i^{-1/2}(\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T), \widehat{\boldsymbol{\Lambda}}_i^{-1/2}\mathbf{T}_i(t)(\mathbf{1}_{m_i}, \mathbf{x}_{2i}(\mathbf{t}_i)^T))$, $i = 1,\ldots,n$.

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

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i}\{y_i(t_{ij}) - \mathbf{x}_{1i}(t_{ij})^T\boldsymbol{\beta}_1 - \widehat{\beta}_0(t_{ij},\boldsymbol{\beta}_1) - \mathbf{x}_{2i}(t_{ij})^T\widehat{\boldsymbol{\beta}}_2(t_{ij},\boldsymbol{\beta}_1)\}^T\widehat{\boldsymbol{\Lambda}}_i^{-1}(j,k)$$
$$\times\{y_i(t_{ik}) - \mathbf{x}_{1i}(t_{ik})^T\boldsymbol{\beta}_1 - \widehat{\beta}_0(t_{ij},\boldsymbol{\beta}_1) - \mathbf{x}_{2i}(t_{ij})^T\widehat{\boldsymbol{\beta}}_2(t_{ij},\boldsymbol{\beta}_1)\}.$$

Then the corresponding estimator for $(\beta_0(t), \boldsymbol{\beta}_2(t)^T)^T$ is as in (4.10) with $\boldsymbol{\beta}_1$ replaced by $\widehat{\boldsymbol{\beta}}_1^{\mathrm{PLS}}$. We call these the refined estimators. Rewrite $\widehat{\boldsymbol{\beta}}_1^{\mathrm{PLS}}$ as

$$\widehat{\boldsymbol{\beta}}_1^{\mathrm{PLS}} = \arg\min_{\boldsymbol{\beta}_1\in\mathbb{R}^{s_1}}\{(\mathcal{I}-\widehat{\mathcal{S}})(\mathbf{Y}-\mathbf{X}_1\boldsymbol{\beta}_1)\}^T\widehat{\boldsymbol{\Lambda}}^{-1}\{(\mathcal{I}-\widehat{\mathcal{S}})(\mathbf{Y}-\mathbf{X}_1\boldsymbol{\beta}_1)\},$$

where $\widehat{\boldsymbol{\Lambda}}^{-1} = \mathrm{diag}\{\widehat{\boldsymbol{\Lambda}}_1^{-1},\ldots,\widehat{\boldsymbol{\Lambda}}_n^{-1}\}$, $\widehat{\mathcal{S}} = (\widehat{\mathbf{S}}_{11}^T,\ldots,\widehat{\mathbf{S}}_{1m_1}^T,\ldots,\widehat{\mathbf{S}}_{n1}^T,\ldots,\widehat{\mathbf{S}}_{nm_n}^T)^T$, $\widehat{\mathbf{S}}_{ij} = (1,\mathbf{x}_{2i}(t_{ij})^T,\mathbf{0}_{1\times(s_2+1)})(\widehat{\mathbf{V}}(t_{ij})^T\mathbf{W}_{h_1}(t_{ij})\widehat{\mathbf{V}}(t_{ij}))^{-1}\widehat{\mathbf{V}}(t_{ij})^T\mathbf{W}_{h_1}(t_{ij})\widehat{\boldsymbol{\Lambda}}^{-1/2}$, $j=1,\ldots,m_i$, $i=1,\ldots,n$. Then we have

$$\widehat{\boldsymbol{\beta}}_1^{\mathrm{PLS}} = \{\mathbf{X}_1^T(\mathcal{I}-\widehat{\mathcal{S}})^T\widehat{\boldsymbol{\Lambda}}^{-1}(\mathcal{I}-\widehat{\mathcal{S}})\mathbf{X}_1\}^{-1}\mathbf{X}_1^T(\mathcal{I}-\widehat{\mathcal{S}})^T\widehat{\boldsymbol{\Lambda}}^{-1}(\mathcal{I}-\widehat{\mathcal{S}})\mathbf{Y}, \tag{4.11}$$

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

$$\begin{aligned}&\big(\widehat{\beta}_0^{\mathrm{PLS}}(t), \widehat{\boldsymbol{\beta}}_2^{\mathrm{PLS}}(t)^T\big)^T\\ &\quad= (\mathbf{I}_{s_2+1},\mathbf{0}_{(s_2+1)\times(s_2+1)})\\ &\qquad\times\big(\widehat{\mathbf{V}}(t)^T\mathbf{W}_{h_1}(t)\widehat{\mathbf{V}}(t)\big)^{-1}\widehat{\mathbf{V}}(t)^T\mathbf{W}_{h_1}(t)\widehat{\boldsymbol{\Lambda}}^{-1/2}\big(\mathbf{Y}-\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1^{\mathrm{PLS}}\big).\end{aligned} \tag{4.12}$$

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 $\mathbf{t}_i$ is a vector of m equally-spaced grid points on [0, 1]. We considered three coefficients settings:

Case I. $\beta_0(t) = 3.5\sin(2\pi t)$, $s_1 = 2$, $s_2 = 3$, $\boldsymbol{\beta}_1 = (5,-5)^T$ and $\boldsymbol{\beta}_2(t) = (5(1-t)^2, 3.5(\exp(-(3t-1)^2)+\exp(-(4t-3)^2))-1.5, 3.5t^{1/2})^T$.

Case II. $\beta_0(t) = 3.5\sin(2\pi t)$, $s_1 = 0$, $s_2 = 5$, and $\boldsymbol{\beta}_2(t) = (5(1-t)^2, 3.5(\exp(-(3t-1)^2)+\exp(-(4t-3)^2))-1.5, 3.5t^{1/2}, 6-2t, 2-3\cos(4\pi t))^T$.

Case III. $\beta_0(t) = 3.5\sin(2\pi t)$, $s_1 = 5$, $s_2 = 0$, $\boldsymbol{\beta}_1 = (5,-5,2.5,-2.5,1)^T$.

We generated covariates $\mathbf{x}(t)$ from a p-dimensional Gaussian process whose component processes each has mean zero and covariance function $\mathrm{Cov}(x_k(s),$

$x_k(t)) = 2\sin(2\pi s)\sin(2\pi t)$. The correlation between components is specified as follows. The first $s_1 + s_2 + s_0$ elements of $\mathbf{x}(t)$ are correlated with a constant correlation ρ, and thus follow a compound-symmetry covariance structure. The remaining $p - (s_1 + s_2 + s_0)$ elements of $\mathbf{x}(t)$ are uncorrelated with each other and the first $s_1 + s_2 + s_0$ elements. The first s_1 and s_2 elements of $\mathbf{x}(t)$ are used as $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$ in the model, respectively. The next s_0 elements of $\mathbf{x}(t)$ are spurious variables that are not related to $y(t)$ but correlated to $\mathbf{x}_1(t)$ and $\mathbf{x}_2(t)$. The random error $\varepsilon(t)$ was simulated from an ARMA(1, 1) Gaussian process with mean zero and covariance function $\mathrm{Cov}(\varepsilon(s), \varepsilon(t)) = \omega \mathrm{r}^{|s-t|}$, with $\omega = 0.85$ and $\mathrm{r} = 0.5$.

In addition, 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 $\mathbf{x}(t)$ is an AR(1) matrix with (j, j')th element equal to $\rho^{|j-j'|}$. We set $\omega = 0.85$ and $\mathrm{r} = 0.6$ 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 $\mathbf{x}(t)$ 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 $\mathrm{r} = 0.5$ 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 ρ under cases II and III are very similar to that under case I, and thus are not reported here. Both varying-coefficients and constant nonzero coefficients were selected correctly with high probability over the simulations, especially when the spurious correlation ρ 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 was 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 error (RMSE). For estimates of the nonparametric components, we report the mean integrated absolute error (MIAE) and the 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 coefficient functions in case I with median MISE are depicted in Figure 1.

TABLE 1

Variable selection results from 500 *simulations. Cvar* (*Cfix*) *is the proportion of variables with varying-coefficients* (*constant nonzero 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 nonzeros in the screening step. Here*, $m = 20$, $s_0 = 10$, $p = 500$, *n is the sample size and* ρ *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**
ρ	**0.1**	**0.5**	**0.5**	**0.3**	**0.3**	**0.4**	**0.4**	**0.4**	**0.5**	**0.4**	**0.5**
Cvar	0.965	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 (0.79)	6.43 (1.23)	5.02 (0.54)	5.12 (0.13)	5.04 (0.01)	5.25 (0.13)	5.01 (0.01)	4.95 (0.63)	4.98 (0.40)	4.87 (0.96)	4.99 (0.22)
U	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.02	0.01	0.03	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 (0.79)	4.99 (1.18)	5.00 (0.54)	4.97 (0.13)	5.00 (0.01)	4.96 (0.12)	5.00 (0.01)	4.93 (0.59)	4.97 (0.40)	4.81 (0.93)	4.99 (0.22)
FP	0.01 (0.06)	0.81 (0.25)	0.02 (0.01)	0.09 (0.01)	0.04 (0.00)	0.15 (0.01)	0.05 (0.01)	0.03 (0.19)	0.00 (0.00)	0.05 (0.19)	0.00 (0.00)
TPvar	2.93 (0.54)	2.87 (0.75)	2.97 (0.42)	4.98 (0.11)	5.00 (0.04)	–	–	2.82 (0.48)	2.92 (0.34)	2.79 (0.65)	2.97 (0.31)
FPvar	0.10 (0.04)	0.15 (0.09)	0.09 (0.06)	0.03 (0.08)	0.01 (0.00)	0.03 (0.08)	0.01 (0.01)	0.01 (0.08)	0.00 (0.00)	0.04 (0.21)	0.01 (0.05)
TPfix	1.92 (0.33)	1.84 (0.48)	1.92 (0.22)	–	–	4.93 (0.15)	5.00 (0.04)	1.96 (0.26)	1.98 (0.17)	1.88 (0.43)	1.99 (0.09)
FPfix	0.04 (0.28)	0.80 (0.42)	0.14 (0.21)	0.06 (0.14)	0.03 (0.04)	0.10 (0.12)	0.02 (0.01)	0.16 (0.40)	0.06 (0.23)	0.20 (0.42)	0.08 (0.27)
MMMS	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)	5 (0)

TABLE 2
Estimation results of 500 *simulations. Oracle* (*Practical*) *estimate refers to estimation with known model* (*after screening and model selection*)

	Oracle estimate				**Practical estimate**			
	Initial $\widetilde{\beta}_k^{\mathrm{PLS}}$		**Refined $\widehat{\beta}_k^{\mathrm{PLS}}$**		**Initial $\widetilde{\beta}_k^{\mathrm{PLS}}$**		**Refined $\widehat{\beta}_k^{\mathrm{PLS}}$**	
Case I Parameters	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
β_{11}	0.0374	0.0623	0.0253	0.0387	0.0572	0.0806	0.0266	0.0458
β_{12}	0.0507	0.0642	0.0296	0.0417	0.0662	0.0768	0.0330	0.0500
Functions	**MIAE**	**RMISE**	**MIAE**	**RMISE**	**MIAE**	**RMISE**	**MIAE**	**RMISE**
$\beta_0(t)$	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 II Functions	**MIAE**	**RMISE**	**MIAE**	**RMISE**	**MIAE**	**RMISE**	**MIAE**	**RMISE**
$\beta_0(t)$	0.1748	0.2571	0.1042	0.1794	0.2254	0.3179	0.1220	0.2535
$\beta_{21}(t)$	0.1939	0.2703	0.0859	0.1389	0.2316	0.3315	0.1015	0.2220
$\beta_{22}(t)$	0.1532	0.2357	0.1029	0.1473	0.1964	0.3003	0.1221	0.2088
$\beta_{23}(t)$	0.1710	0.2381	0.1019	0.1462	0.2156	0.2901	0.1215	0.2383
$\beta_{24}(t)$	0.2074	0.3352	0.1181	0.1889	0.2473	0.3880	0.1243	0.2565
$\beta_{25}(t)$	0.2425	0.3562	0.1252	0.2441	0.2680	0.4055	0.1362	0.2590
Case III Parameters	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
β_{11}	0.0136	0.0264	0.0109	0.0223	0.0142	0.0402	0.0136	0.0387
β_{12}	0.0125	0.0200	0.0118	0.0141	0.0144	0.0458	0.0138	0.0374
β_{13}	0.0256	0.0400	0.0162	0.0332	0.0352	0.0538	0.0286	0.0469
β_{14}	0.0206	0.0360	0.0175	0.0282	0.0327	0.0565	0.0249	0.0489
β_{15}	0.0300	0.0400	0.0265	0.0346	0.0525	0.0648	0.0430	0.0608
Functions	**MIAE**	**RMISE**	**MIAE**	**RMISE**	**MIAE**	**RMISE**	**MIAE**	**RMISE**
$\beta_0(t)$	0.1067	0.0985	0.1038	0.0938	0.1215	0.1126	0.1156	0.1101

5.2. *Real data analysis*. We analyzed the well-known Yeast Cell 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 aim at identifying important transcription factors (TF) that affect the gene expression. Using the same subset of the original data as in [28], we included totally $p = 96$ TFs as covariates in the following analysis.

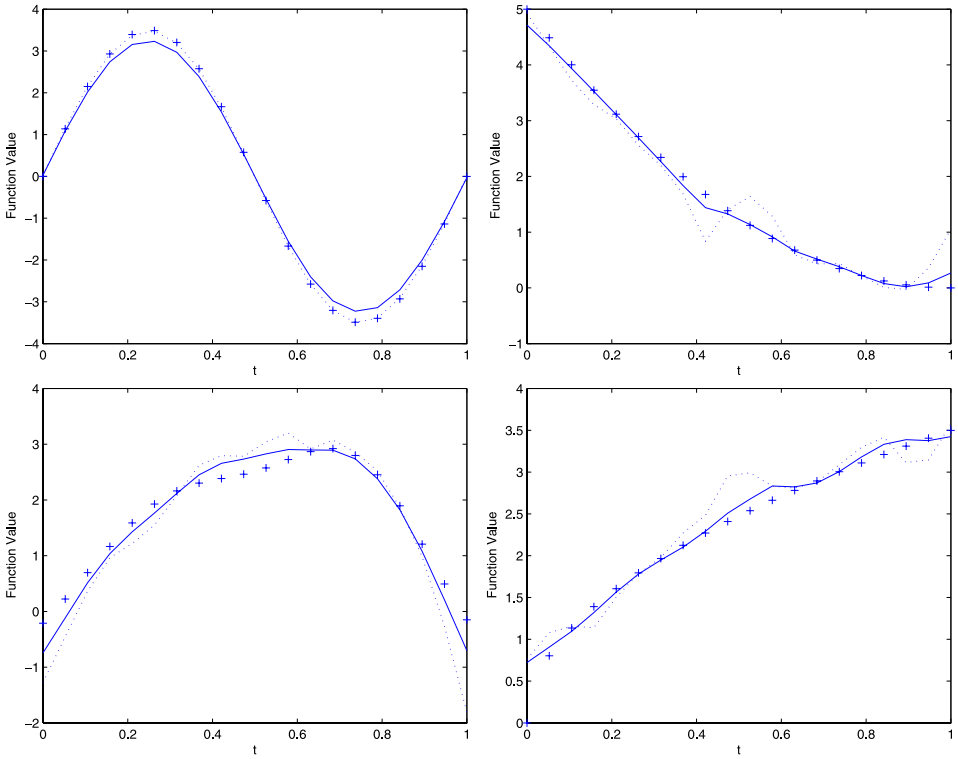

FIG. 1. *Estimated varying coefficients with median MISE for case I. From left to right are* (*top*) $\beta_0(t)$, $\beta_{21}(t)$, (*bottom*) $\beta_{22}(t)$ *and* $\beta_{23}(t)$. *The lines with "+" symbols are the true functions*; *the solid and dashed lines are*, *respectively*, *the refined and initial estimates*.

We first applied the nonparametric screening procedure and 51 TFs were kept after the screening. The nonparametric estimates of varying-coefficients for the 51 TFs from the inital screening are plotted in Figure 2. The names of these 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 includes all of the TFs mentioned in [29]. Comparing to the 21 known yeast TFs related to the cell cycle process included in Figure 2 of [28], our list does not include BAS1, GCN4, GCR1, GCR2, LEU3 and MET31 and includes all of the other 15 TFs. Comparing to the additional TFs identified in Table 2 of [28], the above list does not include 23 of their totally 52 TFs.

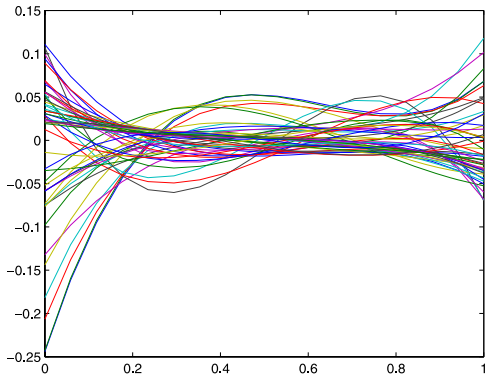

FIG. 2. *Estimated varying coefficients of the* 51 *TFs obtained from NIS.*

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 Figure 3, along with the 95% confidence intervals computed from bootstrap based on 500 resamples. For the sake of comparison, in each panel we also display 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 [28] and [29]; ABF1, HIR1, HIR2, MET4 and SMP1 were also identified as important TFs in [28]. Of the 2 TFs with constant coefficients, MCM1 was identified before in [28] and [29] but its effect was estimated as a varying coefficient instead of a constant one; RLM1 was also included in the list of important TFs reported in [28]. Furthermore, [30] used a penalized estimating equation approach to analyze this data set and identified similar number of TFs although the authors did not report the names of the identified 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.

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 here is to model a collective time-varying

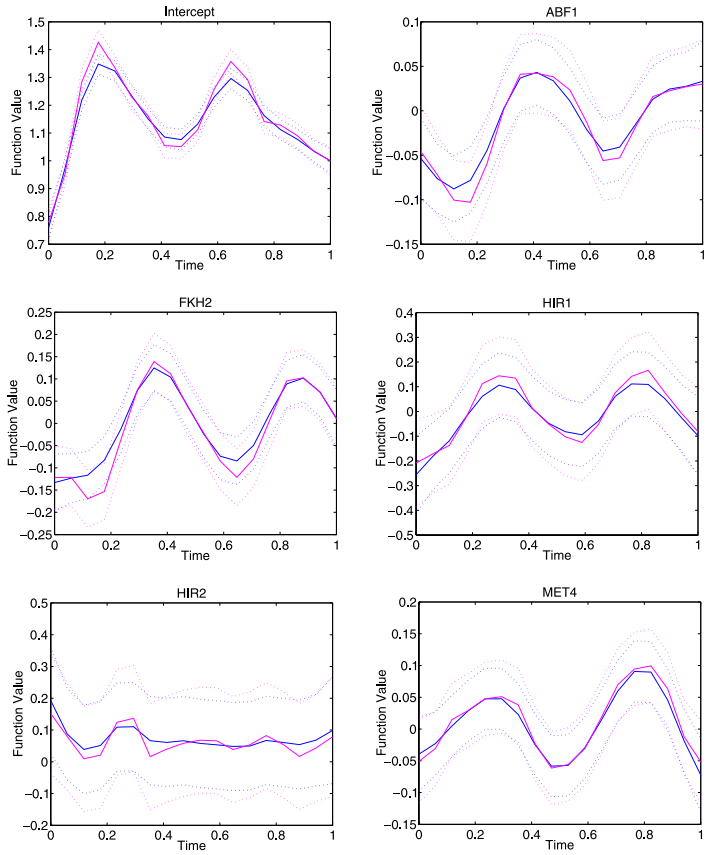

FIG. 3. *Estimated varying coefficients for Yeast Cycle data. Blue curves display the refined estimates and red curves display the initial estimates. Solid curves are the estimated functions while dashed curves are the* 95% *confidence intervals.*

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 mechanisms.

APPENDIX: PROOFS OF THEOREMS 3.1–3.3

We describe some important facts first. Recall that $|\boldsymbol{\gamma}|$ denotes the Euclidean norm of $\boldsymbol{\gamma}$. It is easy to see that, for some positive C_{N1} and C_{N2},

$$C_{N1}L^{-1/2}|\boldsymbol{\gamma}| \le \left\|\left(\gamma_0^T\mathbf{B},\ldots,\gamma_q^T\mathbf{B}\right)^T\right\|_{L_2} \le C_{N2}L^{-1/2}|\boldsymbol{\gamma}| \tag{A.1}$$

uniformly in $\boldsymbol{\gamma} = (\gamma_0^T,\ldots,\gamma_q^T)^T$. Lemmas 3 and 4 of the supplementary material [3] imply that there are positive constants C_{B1} and C_{B2} such that, with proba-

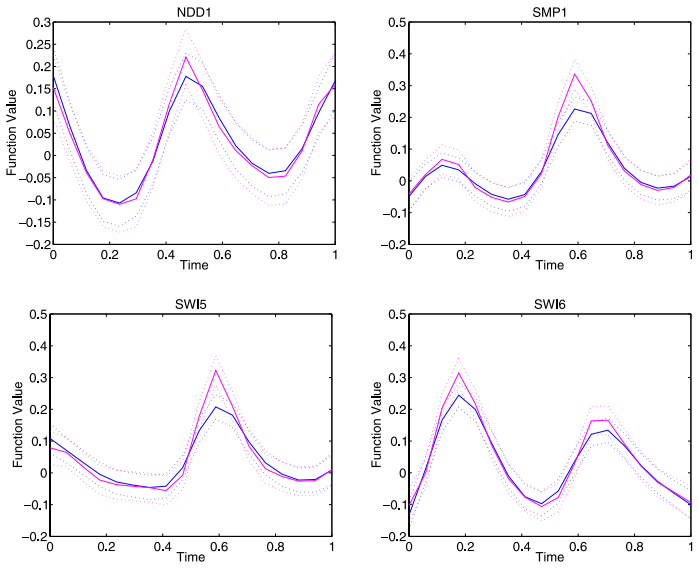

FIG. 3. (*Continued*).

TABLE 3

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

TF	Initial $\widetilde{\boldsymbol{\beta}}_k^{\mathbf{PLS}}$	Refined $\widehat{\boldsymbol{\beta}}_k^{\mathbf{PLS}}$
MCM1	0.0129 (0.0103)	0.0220 (0.0101)
RLM1	−0.0032 (0.0095)	−0.0097 (0.0094)

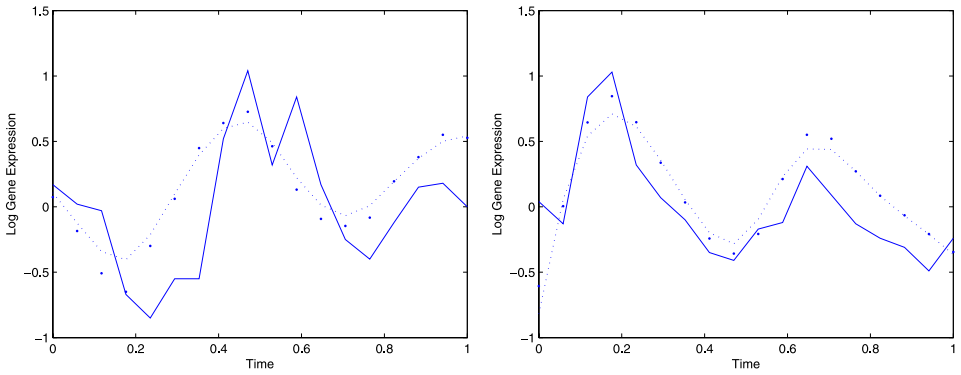

FIG. 4. *Observed and fitted response curves for subjects* 25 (*left*) *and* 177 (*right*) *in the sample. Solid lines are the observed response curves*; *broken lines are the fitted curves from refined estimates*; *Dotted lines are the fitted curves from the initial estimates.*

bility tending to 1,

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

A.1. Proof of Theorem 3.1. Define $\boldsymbol{\gamma}_m = (\gamma_{0m}^T, \ldots, \gamma_{qm}^T)^T \in R^{(q+1)L}$ by

$$\boldsymbol{\gamma}_m = \underset{(\gamma_0^T,\ldots,\gamma_q^T)^T \in R^{(q+1)L}}{\arg\min} \Big\| \beta_0 + \sum_{k=1}^{q} \beta_k x^{(k)} - \gamma_0^T \mathbf{B} - \sum_{k=1}^{q} \gamma_k^T \mathbf{W}_k \Big\|_n^2. \tag{A.2}$$

Lemma 5 of the supplementary material [3] implies that $l_q(\boldsymbol{\gamma})$ is strictly concave with probability tending to 1. Thus, by Lemma 6 of the supplementary material [3], we have only to demonstrate that there is a local minimizer of $l_q(\boldsymbol{\gamma})$ on $R^{(q+1)L}$, denoted by $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\gamma}_0^T, \ldots, \widetilde{\gamma}_q^T)^T$, satisfying

$$\big\|(\widetilde{\gamma}_0^T \mathbf{B}, \ldots, \widetilde{\gamma}_q^T \mathbf{B})^T - (\gamma_{0m}^T \mathbf{B}, \ldots, \gamma_{qm}^T \mathbf{B})^T\big\|_{L_2}^2 = O_p(r_{qn}^2).$$

Recalling the definition of r_{qn}, given in (3.9), and (A.1), we define Γ_M by

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

for a positive M, and represent $l_q(\boldsymbol{\gamma})$ as

$$\begin{aligned} l_q(\boldsymbol{\gamma}) = l_q(\boldsymbol{\gamma}_m) &- 2(\boldsymbol{\gamma} - \boldsymbol{\gamma}_m)^T \big\langle (\mathbf{B}^T \mathbf{W}^T)^T, \varepsilon \big\rangle_n \\ &+ (\boldsymbol{\gamma} - \boldsymbol{\gamma}_m)^T \Sigma_n (\boldsymbol{\gamma} - \boldsymbol{\gamma}_m). \end{aligned} \tag{A.3}$$

By Lemma 7 of the supplementary material [3], we have uniformly on Γ_M, the first term in the right-hand side of (A.3) is $MqLn^{-1}O_p(1)$. By Lemma 5 of the supplementary material [3], we have the second term in the right-hand side of (A.3) is at least $CM^2 qLn^{-1}$ uniformly on Γ_M with probability tending to 1, where C does not depend on M. Thus, we have

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathrm{P}\Big(\inf_{\boldsymbol{\gamma}\in\Gamma_M} l_q(\boldsymbol{\gamma}) > l_q(\boldsymbol{\gamma}_m) \Big) = 1.$$

It follows from the above result, the strict concavity of $l_q(\boldsymbol{\gamma})$, and Lemma 6 of the supplementary material [3], that there is a unique minimizer $\widetilde{\boldsymbol{\gamma}}$ of $l_q(\boldsymbol{\gamma})$ on $R^{(q+1)L}$ giving the desired convergence rate.

A.2. Proof of Theorem 3.2. Define $\overline{\boldsymbol{\gamma}} = (\overline{\gamma}_0^T, \ldots, \overline{\gamma}_q^T)^T \in R^{(q+1)L}$ by

$$\arg\min \big\|(\beta_0, \ldots, \beta_q)^T - (\gamma_0^T \mathbf{B}, \ldots, \gamma_q^T \mathbf{B})^T\big\|_{L_2}^2. \tag{A.4}$$

Then, $(\overline{\gamma}_0^T \mathbf{B}, \ldots, \overline{\gamma}_q^T \mathbf{B})^T \in \mathbf{G}_0$ due to Assumption S(1), and the minimum in (A.4) is no larger than that of $\boldsymbol{\gamma}_m$ in Lemma 6 of the supplementary material [3]. Thus, Lemma 6 of the supplementary material [3] and (A.1) together imply that, with

probability tending to 1, $|\overline{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_m|^2 \le C\rho_{qn}^2 L$ for some positive C. The desired result follows if we show that

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathrm{P}\Big(\inf_{\boldsymbol{\gamma}\in\overline{\Gamma}_M} Q_q(\boldsymbol{\gamma}) > Q_q(\overline{\boldsymbol{\gamma}}) \Big) = 1, \tag{A.5}$$

where $\overline{\Gamma}_M = \{\boldsymbol{\gamma} \in R^{(q+1)L} | |\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}| = ML^{1/2} r_{qn}\}$. Write

$$\begin{aligned} Q_q(\boldsymbol{\gamma}) - Q_q(\overline{\boldsymbol{\gamma}}) &= \{l_q(\boldsymbol{\gamma}) - l_q(\overline{\boldsymbol{\gamma}})\} \\ &\quad + \Bigg[\sum_{k=0}^{q} \{ p_{\lambda_1}(|(g_k)_c|) + p_{\lambda_2}(\|(g_k)_f\|_{L_2}) \} \\ &\qquad - \sum_{k=0}^{q} \{ p_{\lambda_1}(|(\overline{g}_k)_c|) + p_{\lambda_2}(\|(\overline{g}_k)_f\|_{L_2}) \} \Bigg], \end{aligned} \tag{A.6}$$

where $g_k = \gamma_k^T \mathbf{B}$ and $\overline{g}_k = \overline{\gamma}_k^T \mathbf{B}$, $k = 0, 1, \ldots, q$. We have

$$\begin{aligned} &l_q(\boldsymbol{\gamma}) - l_q(\overline{\boldsymbol{\gamma}}) \\ &\quad = 2(\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}})^T \{ -\langle (\mathbf{B}^T \mathbf{W}^T)^T, \varepsilon \rangle_n + \Sigma_n(\overline{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_m) \} \qquad \text{(A.7)} \\ &\qquad + (\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}})^T \Sigma_n (\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}). \qquad \text{(A.8)} \end{aligned}$$

Lemmas 5 and 7 of the supplementary material [3] imply that uniformly in $\boldsymbol{\gamma} \in \overline{\Gamma}_M$, the first term in the right-hand side of (A.7) equals $ML^{1/2} r_{qn} \times O_p((q/n)^{1/2}) + MLr_{qn} O_p(L^{-1})\rho_n = MO_p(r_{qn}^2)$. By Lemma 5 of the supplementary material [3], there is a positive constant C such that the second term in the right-hand side of (A.7) is no less than $CM^2 r_{qn}^2$, uniformly in $\boldsymbol{\gamma} \in \overline{\Gamma}_M$ and with probability tending to 1. Thus, we have

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathrm{P}\Big(\inf_{\boldsymbol{\gamma}\in\overline{\Gamma}_M} \{l_q(\boldsymbol{\gamma}) - l_q(\overline{\boldsymbol{\gamma}})\} \ge CM^2 r_{qn}^2/2 \Big) = 1. \tag{A.9}$$

Next, we consider the difference between the penalty terms. Recall that a_0 appearing below comes from the SCAD function in (3.6). When $|\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma}}| = ML^{1/2} r_{pn}$ we have, for $k = 0, 1, \ldots, q$ and sufficiently large M,

$$\begin{aligned} &|(\gamma_k^T \mathbf{B})_c|, |(\overline{\gamma}_k^T \mathbf{B})_c| > a_0\lambda_1 \quad \text{or} \quad |(\gamma_k^T \mathbf{B})_c| = o(\lambda_1) \quad \text{and} \quad |(\overline{\gamma}_k^T \mathbf{B})_c| = 0, \\ &\|(\gamma_k^T \mathbf{B})_f\|_{L_2}, \|(\overline{\gamma}_k^T \mathbf{B})_f\|_{L_2} > a_0\lambda_2 \quad \text{or} \quad \|(\gamma_k^T \mathbf{B})_f\|_{L_2} = o(\lambda_2) \quad \text{and} \\ &\qquad\qquad\qquad\qquad\qquad\qquad \|(\overline{\gamma}_k^T \mathbf{B})_f\|_{L_2} = 0. \end{aligned}$$

The above relations and the properties of the SCAD function imply the second term of the right-hand side of (A.6) is nonnegative on $\overline{\Gamma}_M$. Thus, (A.5) follows from (A.7) and (A.9), and the proof of Theorem 3.2 is complete.

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 [29].

First, let $\widehat{\boldsymbol{\gamma}}$ be a local minimizer of $Q_q(\boldsymbol{\gamma})$ on $R^{(q+1)L}$ satisfying

$$\bigl(\widehat{\gamma}_0^T\mathbf{B},\ldots,\widehat{\gamma}_q^T\mathbf{B}\bigr)^T \in \mathbf{G}_0, \tag{A.10}$$

$$\bigl\|\bigl(\widehat{\gamma}_0^T\mathbf{B},\ldots,\widehat{\gamma}_q^T\mathbf{B}\bigr)^T-(\beta_0,\ldots,\beta_q)^T\bigr\|_{L_2}\le\eta_n r_{qn}. \tag{A.11}$$

Then consider $Q_q(\widehat{\boldsymbol{\gamma}}+\boldsymbol{\delta})$ for $\widehat{\boldsymbol{\gamma}}+\boldsymbol{\delta}\in\overline{\mathbf{G}}_0$, where $\overline{\mathbf{G}}_0$ is the subspace of $R^{(q+1)L}$ corresponding to the oracle space $\mathbf{G}_0$. When $|\boldsymbol{\delta}|$ is small enough, we have the same value of the penalty term as that for $\widehat{\boldsymbol{\gamma}}$ due to the flatness of the SCAD function. On the other hand, the local optimality of $\widehat{\boldsymbol{\gamma}}$ implies that $Q_q(\widehat{\boldsymbol{\gamma}}+\boldsymbol{\delta})\ge Q_q(\widehat{\boldsymbol{\gamma}})$. Thus, there is a small neighborhood of $\widehat{\boldsymbol{\gamma}}$ in $\overline{\mathbf{G}}_0$, Γ_h, such that $\inf_{\boldsymbol{\gamma}\in\Gamma_h} l_q(\boldsymbol{\gamma})\ge l_q(\widehat{\boldsymbol{\gamma}})$. This shows that $\widehat{\boldsymbol{\gamma}}$ is a local minimizer of $l_q(\boldsymbol{\gamma})$ on $\overline{\mathbf{G}}_0$. Since $l_q(\boldsymbol{\gamma})$ is strictly concave on $R^{(q+1)L}$ with probability tending to 1, this $\widehat{\boldsymbol{\gamma}}$ must be the unique minimizer of $l_q(\boldsymbol{\gamma})$ on $\overline{\mathbf{G}}_0$, denoted by $\widehat{\boldsymbol{\gamma}}_0$. A similar argument can be found in [9].

Next, we deal with the oracle estimator $\widehat{\boldsymbol{\gamma}}_0$. We neglect $x^{(s+1)}(t),\ldots,x^{(q)}(t)$ and restrict $\overline{\mathbf{G}}_0$ to $R^{(s+1)L}$. Besides, we define $\boldsymbol{\gamma}_m\in R^{(s+1)L}$ and $\overline{\boldsymbol{\gamma}}\in\overline{\mathbf{G}}_0\subset R^{(s+1)L}$ similarly as in the proofs of Theorems 3.1 and 3.2. Then we have

$$\begin{aligned} &l_s(\boldsymbol{\gamma})-l_s(\overline{\boldsymbol{\gamma}})\\ &\quad=2(\boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}})^T\bigl\{-\bigl\langle\bigl(\mathbf{B}^T\mathbf{W}^T\bigr)^T,\varepsilon\bigr\rangle_n+\Sigma_n(\overline{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_m)\bigr\}+(\boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}})^T\Sigma_n(\boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}}), \end{aligned}$$

for $\boldsymbol{\gamma}\in R^{(s+1)L}$, where $\mathbf{W}$ 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 $\lim_{M\to\infty}\limsup_{n\to\infty}\mathrm{P}(\inf_{\boldsymbol{\gamma}\in\widetilde{\Gamma}_M} l_s(\boldsymbol{\gamma})>l_s(\overline{\boldsymbol{\gamma}}))$, where $\widetilde{\Gamma}_M=\{\boldsymbol{\gamma}\in\overline{\mathbf{G}}_0||\boldsymbol{\gamma}-\overline{\boldsymbol{\gamma}}|=ML^{1/2}r_{sn}\}$. Thus, $\widehat{\boldsymbol{\gamma}}_0$ satisfies $|\widehat{\boldsymbol{\gamma}}_0-\overline{\boldsymbol{\gamma}}|=O_p(L^{1/2}r_{sn})$.

Finally, we consider a local minimizer $\widehat{\boldsymbol{\gamma}}$ satisfying (A.11) and prove that it also satisfies (A.10). For this $\widehat{\boldsymbol{\gamma}}$, suppose that $(\widehat{\beta}_j)_c\neq 0$ and $(\beta_j)_c=0$ for some j, where $\widehat{\beta}_j=\widehat{\gamma}_j^T\mathbf{B}$. Then we have that $0<|(\widehat{\beta}_j)_c|=o(\lambda_1)$. Define $\widehat{\boldsymbol{\beta}}_t$ for $t\in[0,1/2]$ by

$$\widehat{\boldsymbol{\beta}}_t=\widehat{\boldsymbol{\beta}}+t(\widehat{\boldsymbol{\beta}}_{-cj}-\widehat{\boldsymbol{\beta}})=(1-t)\widehat{\boldsymbol{\beta}}+t\widehat{\boldsymbol{\beta}}_{-cj},$$

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

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

$$\begin{aligned} Q_q(\widehat{\boldsymbol{\gamma}}_t)-Q_q(\widehat{\boldsymbol{\gamma}})&=\bigl\{l_q(\widehat{\boldsymbol{\gamma}}_t)-l_q(\widehat{\boldsymbol{\gamma}})\bigr\}+\bigl\{p_{\lambda_1}\bigl((1-t)|(\widehat{\beta}_j)_c|\bigr)-p_{\lambda_1}\bigl(|(\widehat{\beta}_j)_c|\bigr)\bigr\}\\ &=J_5+J_6. \end{aligned}$$

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

$$J_5=-2(\widehat{\boldsymbol{\gamma}}_t-\widehat{\boldsymbol{\gamma}})^T\Biggl\langle\bigl(\mathbf{B}^T\mathbf{W}^T\bigr)^T,y-\widehat{\gamma}_0^T\mathbf{B}-\sum_{k=1}^{q}\widehat{\gamma}_k^T\mathbf{W}_k\Biggr\rangle_n+(\widehat{\boldsymbol{\gamma}}_t-\widehat{\boldsymbol{\gamma}})^T\Sigma_n(\widehat{\boldsymbol{\gamma}}_t-\widehat{\boldsymbol{\gamma}}).$$

Lemmas 5 and 7 of the supplementary material [3] imply that the two terms in J_5 can be expressed as $-2(\widehat{\boldsymbol{\gamma}}_t - \widehat{\boldsymbol{\gamma}})^T \langle (\mathbf{B}^T\mathbf{W}^T)^T, (\mathbf{B}^T\mathbf{W}^T)\rangle_n (\boldsymbol{\gamma}_m - \widehat{\boldsymbol{\gamma}}) - 2(\widehat{\boldsymbol{\gamma}}_t - \widehat{\boldsymbol{\gamma}})^T \langle (\mathbf{B}^T\mathbf{W}^T)^T, \varepsilon\rangle_n$ and $t|(\widehat{\beta}_j)_c| O_p(\eta_n r_{qn})$. Hence we get $J_5 = t|(\widehat{\beta}_j)_c| O_p \times (\eta_n r_{qn})$. From the above results, property of the SCAD function and Assumption S(2),

$$Q_q(\widehat{\boldsymbol{\gamma}}_t) - Q_q(\widehat{\boldsymbol{\gamma}}) = t|(\widehat{\beta}_j)_c|\{O_p(\eta_n r_{qn}) - p'_{\lambda_1}((1-\bar{t})|(\widehat{\beta}_j)_c|)\} < 0$$

uniformly in $t \in (0, 1/2)$ with probability tending to 1, and the probability does not depend on the specific value of j. This contradicts with the local optimality of $\widehat{\boldsymbol{\gamma}}$, and thus $(\widehat{\beta}_j)_c$ must be equal to 0 if $(\beta_j)_c = 0$. We can treat the other cases in the same way. Hence, (A.10) is established for the local minimizer $\widehat{\boldsymbol{\gamma}}$, and the proof is complete.

SUPPLEMENTARY MATERIAL

Some technical material: Supplement to "Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data" (DOI: 10.1214/14-AOS1236SUPP; .pdf). Some lemmas, and proofs of Theorems 2.1–2.2 and Remark 1.

REFERENCES

[1] ANTONIADIS, A., GIJBELS, I. and LAMBERT-LACROIX, S. (2014). Penalized estimation in additive varying coefficient models using grouped regularization. *Statist. Papers* **55** 727–750.

[2] ANTONIADIS, A., GIJBELS, I. and VERHASSELT, A. (2012). Variable selection in varying-coefficient models using P-splines. *J. Comput. Graph. Statist.* **21** 638–661. MR2970912

[3] CHENG, M. Y., HONDA, T., LI, J. and PENG, H. (2014). Supplement to "Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal data." DOI:10.1214/14-AOS1236SUPP.

[4] 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. MR2847969

[5] FAN, J. and HUANG, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11** 1031–1057. MR2189080

[6] FAN, J., HUANG, T. and LI, R. (2007). Analysis of longitudinal data with semiparametric estimation of convariance function. *J. Amer. Statist. Assoc.* **102** 632–641. MR2370857

[7] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

[8] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

[9] FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. MR2849368

[10] FAN, J., MA, Y. and DAI, W. (2014). Nonparametric independence screening in sparse ultra-high dimensional varying coefficient models. *J. Amer. Statist. Assoc.* To appear.

[11] FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

[12] FAN, J. and WU, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *J. Amer. Statist. Assoc.* **103** 1520–1533. MR2504201

[13] FAN, Y. and LV, J. (2013). Asymptotic equivalence of regularization methods in thresholded parameter space. *J. Amer. Statist. Assoc.* **108** 1044–1061. MR3174683

[14] FAN, Y. and TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **75** 531–552. MR3065478

[15] HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. MR2278365

[16] HU, T. and XIA, Y. (2012). Adaptive semi-varying coefficient model selection. *Statist. Sinica* **22** 575–599. MR2954353

[17] LAM, C. and FAN, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *Ann. Statist.* **36** 2232–2260. MR2458186

[18] LI, G., PENG, H., ZHANG, J. and ZHU, L. (2012). Robust rank correlation based screening. *Ann. Statist.* **40** 1846–1877. MR3015046

[19] LI, Y. (2011). Efficient semiparametric regression for longitudinal data with nonparametric covariance estimation. *Biometrika* **98** 355–370. MR2806433

[20] LIAN, H., LAI, P. and LIANG, H. (2013). Partially linear structure selection in Cox models with varying coefficients. *Biometrics* **69** 348–357. MR3071053

[21] LIU, J., LI, R. and WU, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *J. Amer. Statist. Assoc.* **109** 266–274. MR3180562

[22] MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 53–71. MR2412631

[23] NOH, H. S. and PARK, B. U. (2010). Sparse varying coefficient models for longitudinal data. *Statist. Sinica* **20** 1183–1202. MR2730179

[24] SCHUMAKER, L. L. (2007). *Spline Functions*: *Basic Theory*, 3rd ed. Cambridge Univ. Press, Cambridge. MR2348176

[25] SPELLMAN, P. T., SHERLOCK, G., ZHANG, M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., BROWN, P. O., BOTSTIEN, D. and FUTCHER, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast sacchromyces cerevisiae by microarray hybridization. *Mol. Biol. Cell* **9** 3273–3297.

[26] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[27] WANG, H., LI, R. and TSAI, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94** 553–568. MR2410008

[28] WANG, L., CHEN, G. and LI, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23** 1486–1494.

[29] 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. MR2504204

[30] WANG, L., ZHOU, J. and QU, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68** 353–360. MR2959601

[31] WANG, N., CARROLL, R. J. and LIN, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Amer. Statist. Assoc.* **100** 147–157. MR2156825

[32] WEI, F., HUANG, J. and LI, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica* **21** 1515–1540. MR2895107

[33] XIA, Y., ZHANG, W. and TONG, H. (2004). Efficient estimation for semivarying-coefficient models. *Biometrika* **91** 661–681. MR2090629

[34] XUE, L. and QU, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *J. Mach. Learn. Res.* **13** 1973–1998. MR2956349

[35] YAN, J. and HUANG, J. (2012). Model selection for Cox models with time-varying coefficients. *Biometrics* **68** 419–428. MR2959608

[36] YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

[37] 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. MR2894767

[38] ZHANG, W., FAN, J. and SUN, Y. (2009). A semiparametric model for cluster data. *Ann. Statist.* **37** 2377–2408. MR2543696

[39] ZOU, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418–1429. MR2279469

M.-Y. CHENG
DEPARTMENT OF MATHEMATICS
NATIONAL TAIWAN UNIVERSITY
TAIPEI 106
TAIWAN
E-MAIL: cheng@math.ntu.edu.tw

J. LI
DEPARTMENT OF STATISTICS & APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 117546
E-MAIL: stalj@nus.edu.sg

T. HONDA
GRADUATE SCHOOL OF ECONOMICS
HITOTSUBASHI UNIVERSITY
KUNITACHI, TOKYO 186-8601
JAPAN
E-MAIL: t.honda@r.hit-u.ac.jp

H. PENG
DEPARTMENT OF MATHEMATICS
HONG KONG BAPTIST UNIVERSITY
KOWLOON
HONG KONG
CHINA
E-MAIL: hpeng@math.hkbu.edu.hk