University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2007 *Paper* 20

Variable Selection for Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements

Lifeng Wang* Hongzhe Li†

*University of Pennsylvania, lifwang@mail.med.upenn.edu

†University of Pennsylvania, hli@mail.med.upenn.edu

Variable Selection for Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements

Lifeng Wang and Hongzhe Li

Abstract

Nonparametric varying-coefficient models are commonly used for analysis of data measured repeatedly over time, including longitudinal and functional responses data. While many procedures have been developed for estimating the varying-coefficients, the problem of variable selection for such models has not been addressed. In this article, we present a regularized estimation procedure for variable selection for such nonparametric varying-coefficient models using basis function approximations and a group smoothly clipped absolute deviation penalty (gSCAD). This gSCAD procedure simultaneously selects significant variables with time-varying effects and estimates unknown smooth functions using basis function approximations. With appropriate selection of the tuning parameters, we have established the oracle property of the procedure and the consistency of the function estimation. The methods are illustrated with simulations and an application to analysis of microarray time-course gene expression data to in order to identify the transcription factors that are related to yeast cell cycle process.

Variable Selection for Nonparametric Varying-Coefficient Models for Analysis of Repeated Measurements

Running Title: Variable Selection for Varying-Coefficient Models

BY LIFENG WANG AND HONGZHE LI

Department of Biostatistics and Epidemiology

University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

lifwang@mail.med.upenn.edu, hongzhe@mail.med.upenn.edu

SUMMARY:

Nonparametric varying-coefficient models are commonly used for analysis of data measured repeatedly over time, including longitudinal and functional responses data. While many procedures have been developed for estimating the varying-coefficients, the problem of variable selection for such models has not been addressed. In this article, we present a regularized estimation procedure for variable selection for such nonparametric varying-coefficient models using basis function approximations and a group smoothly clipped absolute deviation penalty (gSCAD). This gSCAD procedure simultaneously selects significant variables with time-varying effects and estimates unknown smooth functions using basis function approximations. With appropriate selection of the tuning parameters, we have established the oracle property of the procedure and the consistency of the function estimation. The methods are illustrated with simulations and an application to analysis of microarray time-course gene expression data to in order to identify the transcription factors that are related to yeast cell cycle process.

Key words and phrases: Regularized estimation, Functional response, Longitudinal data, Time course gene expression data.

1. Introduction

Varying-coefficient models (Hastie and Tibshirani, 1993) are commonly used for studying the time-dependent effects of covariates on responses measured repeatedly. Such models can be used for longitudinal data where subjects are often measured repeatedly over a given period of time, so that the measurements within each subject are possibly correlated with each other (Diggle *et al.*, 1994; Rice, 2004). Another setting is the functional response models (Rice, 2004), where the ith response is a smooth real function $y_i(t)$, $i = 1, \cdots, n, t \in \mathcal{T} = [0, T]$, although in practice only $y_i(t_{ij})$, $j = 1, \cdots, J_i$ are observed. For both settings, the response $Y(t)$ is a random process and the predictor $\boldsymbol{X}(t) = (X^{(1)}(t), \ldots, X^{(p)}(t))^T$ is a p-dimensional vector of random processes. In applications, observations for n randomly selected subjects are obtained as $(y_i(t_{ij}), \boldsymbol{x}_i(t_{ij}))$ for the ith subject at discrete time point t_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, J_i$. The linear varying-coefficient model can be written as

$$y_i(t_{ij}) = \boldsymbol{x}_i(t_{ij})^T \boldsymbol{\beta}(t_{ij}) + \varepsilon_i(t_{ij}), \tag{1.1}$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))$ is a p-dimensional vector of smooth functions of t, and $\varepsilon_i(t)$, $i = 1, \ldots, n$ are independently identically distributed random processes, independent of $\boldsymbol{x}_i(t)$. When

the number of covariates is small, depending on the designs of the studies, many methods have been developed for estimating the coefficients in Model (1.1), using parametric models (Diggle *et al.*, 1994) and nonparametric or semiparametric approaches (Zeger and Diggle, 1994; Lin and Carroll, 2000; Rice and Wu, 2001; Huang *et al.*, 2002). However, when the number of covariates in Model (1.1) is large, one important problem is to select the important variables in such models. The goal of this paper is to estimate $\boldsymbol{\beta}(t)$ nonparametrically and select relevant predictors $x_k(t)$ with non-zero functional coefficient $\beta_k(t)$, based on observations $\{(y_i(t_{ij}), \boldsymbol{x}_i(t_{ij})), i = 1, \ldots, n, j = 1, \ldots, J_i\}$.

Regularized estimation has received much attention as a way of performing variable selection for parametric regression models (see Bickel and Li, 2006 for a review). Important regularization procedures for variable selection include LASSO (Tibshirani, 1996) and smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and their recent extensions (Zou, 2006; Yuan and Lin, 2005; Zou and Hastie, 2005). However, these regularized estimation procedures were developed only for the parametric regression models where the model parameters belong to high-dimensional parametric space and cannot be applied directly to the nonparametric varying-coefficients models where the parameters are nonparametric smooth functions. Fan and Li (2004) proposed to use the SCAD penalty for model selection in longitudinal data analysis when $\beta(t)$ are assumed to be time-independent. Wang *et al.* (2007) proposed a group SCAD (gSCAD) procedure for model selection for varying-coefficient models with time-independent covariates and demonstrated its application in analysis of microarray time course gene expression data. The goal of this paper is to further develop the gSCAD procedure for general nonparametric varying coefficients models with possible time-dependent covariate processes and to provide theoretical justification for the gSCAD procedure for variable selection and estimation. This procedure simultaneously selects significant variables and estimates unknown smooth coefficient functions.

The rest of the paper is organized as follows. We first describe the general group SCAD regularized estimation procedure and the algorithm in Section 2. We then present theoretical properties of our estimators in Section 3, including the oracle property and the consistency of the estimates. Finally we present simulation results in Section 4 and application to analysis of microarray time-course gene expression data in Section 5. Proofs of the main results are presented in the Appendix.

2. Regularized Estimation Using gSCAD and Basis Function Expansion

In order to select the relevant covariates in Model (1.1), we propose a new method based on basis expansion of $\boldsymbol{\beta}(t)$ and penalized estimation using a grouped version of the SCAD penalty. In what follows, we assume $\beta_k(t) = \sum_{l=1}^{\infty} \gamma_{kl} B_{kl}(t) \in \mathcal{F}_k$, where $\{B_{kl}(t)\}_{l=1}^{\infty}$ are orthonormal basis functions of function space $\mathcal{F}_k$. Then $\beta_k(t)$ can be approximated by a truncated series $\beta_k(t) \approx \sum_{l=1}^{L_k} \gamma_{kl} B_{kl}(t)$, and Model (1.1) becomes

$$y_i(t_{ij}) \approx \sum_{k=1}^{p} \sum_{l=1}^{L_k} \gamma_{kl} x_i^{(k)}(t_{ij}) B_{kl}(t_{ij}) + \varepsilon_i(t_{ij}), \tag{2.1}$$

where L_k is the number of basis functions in approximating the function $\beta_k(t)$.

When the number of covariate p is small, Huang, Wu and Zhou (2002) proposed to estimate Model (2.1) via a weighted least square regression. In this article, we propose using penalized least square regression for the sake of variable selection. Note that under the truncated series approximations, each function $\beta_k(t)$ in Model (1.1) is characterized by a set of parameters $\gamma_k = (\gamma_{k1}, \cdots, \gamma_{kL_k})^T$ in Model (2.1). Instead of selecting nonzero γ_{kl} in this model, we should select nonzero γ_k. This motivates the following group version of the SCAD regularized estimation; we estimate $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$ by by minimizing

$$l(\boldsymbol{\gamma}) = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \left(y_i(t_{ij}) - \sum_{k=1}^{p} \sum_{l=1}^{L_k} \gamma_{kl} x_i^{(k)}(t_{ij}) B_{kl}(t_{ij}) \right)^2 + \sum_{k=1}^{p} p_\lambda(\|\boldsymbol{\gamma}_k\|_2), \tag{2.2}$$

where $N = \sum_{i=1}^{n} J_i$, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$, $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kL_k})^T$, $k = 1, \ldots, p$, $\|\boldsymbol{\gamma}_k\|_2 = (\sum_{l=1}^{L_k} \gamma_{kl}^2)^{1/2}$ is the l_2-norm of $\boldsymbol{\gamma}_k$, and $p_\lambda(\cdot)$ is the SCAD penalty function with λ as a tuning parameter, which is defined as

$$p_\lambda(|w|) = \begin{cases} \lambda |w| & \text{if } |w| \le \lambda, , \\ -\frac{(|w|^2 - 2a\lambda|w| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |w| < a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |w| > a\lambda. \end{cases} \tag{2.3}$$

The penalty function (2.3) is a quadratic spline function with two knots at λ and $a\lambda$, where a is another tuning parameter. Fan and Li (2001) showed that the Bayes risks are not sensitive to the choice of a and suggested using $a = 3.7$, which was also used in this paper. Through $\hat{\boldsymbol{\gamma}}$ which minimizes the objective function (2.2), an estimate of $\beta_k(t)$ can be obtained by $\hat{\beta}_k(t) = \sum_{l=1}^{L_k} \hat{\gamma}_{kl} B_{kl}(t)$.

To simply the expression of the objective function (2.2), we define

$$\boldsymbol{B}(t) = \begin{pmatrix} B_{11}(t) & \dots & B_{1L_1}(t) & 0 & \dots & 0 & 0 & \dots & 0 \\ & \vdots & & & \vdots & & & \vdots & 0 \\ 0 & \dots & 0 & 0 & \dots & 0 & B_{p1}(t) & \dots & B_{pL_p}(t) \end{pmatrix},$$

$\boldsymbol{U}_i(t_{ij}) = (\boldsymbol{x}_i(t_{ij})^T \boldsymbol{B}(t_{ij}))^T$, $\boldsymbol{U}_i = (\boldsymbol{U}_i(t_{i1}), \dots, \boldsymbol{U}_i(t_{iJ_i}))^T$, and $\boldsymbol{U} = (\boldsymbol{U}_1, \dots, U_n)$. We also define $\boldsymbol{y} = (y_1(t_{11}), \dots, y_n(t_{nJ_n}))^T$. The objective function (2.2) can then be written as

$$l(\boldsymbol{\gamma}) = N^{-1} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{U}_i \boldsymbol{\gamma})^T (\boldsymbol{y}_i - \boldsymbol{U}_i \boldsymbol{\gamma}) + \sum_{k=1}^{p} p_\lambda(\|\boldsymbol{\gamma}_k\|_2). \tag{2.4}$$

Remark 1. The requirement of the orthonormality of the basis $\{B_{kl}(t)\}_{l=1}^{\infty}$ is not essential. When non-orthonormal basis $\{B_{kl}(t)\}_{l=1}^{\infty}$ is used, the penalty $\sum_{k=1}^{p} p_\lambda(\|\boldsymbol{\gamma}_k\|_2)$ in (2.2) and (2.4) should be substituted accordingly by $\sum_{k=1}^{p} p_\lambda((\boldsymbol{\gamma}_k^T \boldsymbol{H}_k \boldsymbol{\gamma})^{1/2})$, where $\boldsymbol{H}_k = (h_{ij})_{L_k \times L_k}$ is a matrix with $h_{ij} = \int_{\mathcal{T}} B_{ki}(t) B_{kj}(t) dt$. The oracle property and convergence results in Section 3 still hold and the proofs just need slight modification.

2.1. Algorithm

Because of non-differentiability of the penalized loss $l(\beta)$, the commonly used gradient method is not applicable. In this section, we develop an iterative algorithm based on local quadratic approximation of the non-convex penalty $p_\lambda(\|\boldsymbol{\gamma}_k\|_2)$. Following Fan and Li (2001), in a neighborhood of a given positive $w_0 \in \mathbb{R}^+$,

$$p_\lambda(w) \approx p_\lambda(w_0) + 1/2\{p'_\lambda(w_0)/w_0\}(w^2 - w_0^2).$$

In our algorithm, a similar quadratic approximation is used by substituting $\boldsymbol{\gamma}$ with $\|\boldsymbol{\gamma}_k\|_2$, $k = 1, \dots, p$. Given an initial value of $\boldsymbol{\gamma}_k^0$ with $\|\boldsymbol{\gamma}_k^0\|_2 > 0$, $p_\lambda(\|\boldsymbol{\gamma}\|_2)$ can be approximated by a quadratic form

$$p_\lambda(\|\boldsymbol{\gamma}_k^0\|_2) + 1/2\{p'_\lambda(\|\boldsymbol{\gamma}_k^0\|_2)/\|\boldsymbol{\gamma}_k^0\|_2\}(\boldsymbol{\gamma}_k^T \boldsymbol{\gamma}_k - (\boldsymbol{\gamma}_k^0)^T \boldsymbol{\gamma}_k^0).$$

As a consequence, equation (2.4) becomes

$$l(\boldsymbol{\gamma}) = N^{-1}(\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma})^T (\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma}) + \boldsymbol{\gamma}^T \Sigma_\lambda(\boldsymbol{\gamma}^0) \boldsymbol{\gamma}, \tag{2.5}$$

where $\Sigma_\lambda(\boldsymbol{\gamma}^0) = diag\{(p'_\lambda(\|\boldsymbol{\gamma}_1^0\|_2)/\|\boldsymbol{\gamma}_1^0\|_2) I_{L_1}, \dots, (p'_\lambda(\|\boldsymbol{\gamma}_K^0\|_2)/\|\boldsymbol{\gamma}_K^0\|_2) I_{L_p}\}$ with I_{L_k} an L_k dimensional identity matrix. This is a quadratic form, and can be solved by

$$(N^{-1}\boldsymbol{U}^T\boldsymbol{U} + 1/2\Sigma_\lambda(\boldsymbol{\gamma}^0))\boldsymbol{\gamma} \quad = \quad \boldsymbol{U}\boldsymbol{y}. \tag{2.6}$$

We outline the algorithm as follows:

Step 1: Initialize $\boldsymbol{\gamma}^{(1)}$.

Step 2: Set $\boldsymbol{\gamma}^0 = \boldsymbol{\gamma}^{(m)}$, and solve $\boldsymbol{\gamma}^{(m+1)}$ by equation (2.6).

Step 3: Iterate Step 2 until convergence of γ.

In the initialization step, we obtain an initial estimation of $\boldsymbol{\gamma}$ using a ridge regression, which substitutes $p_\lambda(\|\boldsymbol{\gamma}_k\|_2)$ in equation (2.2) with a quadratic function $\|\boldsymbol{\gamma}_k\|_2^2$, and can be solved by matrix operations. At any iteration of step 2, if some $\|\boldsymbol{\gamma}_k^{(m)}\|_2$ is smaller than a cutoff value $\epsilon_1 > 0$, we set $\hat{\boldsymbol{\gamma}}_k = \mathbf{0}$ and treat $x^{(k)}(t)$ as irrelevant. If any matrix is singular when solving equation (2.6), a small perturbation ϵ_2 is added to the diagonal entry of the matrix. In our algorithm both ϵ_1 and ϵ_2 are set to 10^{-3}.

2.2. Selection of tuning parameters

To implement the proposed method, we need to choose the tuning parameters: L_k, $k = 1, \ldots, p$ and λ, where L_k controls the smoothness of $\hat{\beta}(t)$, while λ determines the sparsity. In section 3, we show that the oracle property holds, when these tuning parameters grow or decay at a proper rate with n. In practice, however, we need data-driven procedures to select the tuning parameters. In this paper, we only consider the situation when $L_k = L$ for all $\beta_k(t)$, $k = 1, \ldots, p$. To facilitate adaptive selection of L and λ, we propose using a closed form estimation of the generalized cross-validation error (GCV) or the "leave-one-subject-out" cross-validation (SJCV) for two different situations: independent or correlated errors.

If the errors $\varepsilon_i(t_{ij})$ are independent for different t_{ij}, $j = 1, \ldots, J_i$, an approximate GCV is applicable. Note that in the last step of our algorithm, due to the convergence of $\boldsymbol{\gamma}^k$, the nonzero components are estimated as $\hat{\boldsymbol{\gamma}} = (\boldsymbol{U}^T\boldsymbol{U} + N/2\Sigma_\lambda(\hat{\boldsymbol{\gamma}}))^{-1}\boldsymbol{U}Y$, which can be considered as the solution of a ridge regression as follows

$$\|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\gamma}\|_2^2 + N/2\boldsymbol{\gamma}^T\Sigma_\lambda(\hat{\boldsymbol{\gamma}})\boldsymbol{\gamma}. \tag{2.7}$$

Consequently, the optimal (L, λ) can be approximately selected by minimizing the GCV error for (2.7), which can be efficiently computed as

$$GCV(L, \lambda) = \frac{1}{N}\frac{\|y - M(L, \lambda)y\|_2^2}{(1 - tr[M(L, \lambda)]/N)^2},$$

where $M(L, \lambda) = \boldsymbol{U}(\boldsymbol{U}^T\boldsymbol{U} + N/2\Sigma_\lambda(\hat{\boldsymbol{\gamma}}))^{-1}\boldsymbol{U}^T$.

If the correlation structure of $\varepsilon(t)$ is unknown, the GCV is unsuitable. In such a situation, we choose SJCV in the spirit of Rice and Silverman (1991), Hoover *et al.* (1998) and Huang *et al.* (2002). Let $\hat{\boldsymbol{\gamma}}^{(-i)}$ be the solution of (2.2) after deleting the ith subject. The commonly used cross-validation error is then defined as

$$CV(L,\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{J_i}(y_i(t_{ij}) - \boldsymbol{U}^{(-i)}(t_{ij})\hat{\boldsymbol{\gamma}}^{(-i)})^2.$$

$CV(L,\lambda)$ is a good estimate of the true prediction error. However, its computation is very intensive, since it requires solving equation (2.2) n times. To overcome this difficulty, we propose using the following approximate cross-validation error

$$ACV(L,\lambda) = \sum_{i=1}^{n}\sum_{j=1}^{J_i}(y_i(t_{ij}) - \boldsymbol{U}^{(-i)}(t_{ij})\hat{\boldsymbol{\gamma}}^{\star(-i)})^2 = \sum_{i=1}^{N}\|\boldsymbol{y}_i - \boldsymbol{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)}\|_2^2,$$

where $\hat{\boldsymbol{\gamma}}^{\star(-i)}$ is obtained by solving (2.7) instead of (2.2), deleting the ith subject. We have the following "leave-one-subject-out" lemma (see Appendix A for the proof), which greatly facilitates the computation.

Lemma 1. *Define* $\tilde{\boldsymbol{y}}^{(i)} = (\boldsymbol{y}_1^T,\ldots,\boldsymbol{y}_{i-1}^T,\boldsymbol{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)},\boldsymbol{y}_{i+1}^T,\ldots,\boldsymbol{y}_n^T)$, *and let* $\tilde{\boldsymbol{\gamma}}^{(i)}$ *be the solution of (2.7) with* $\boldsymbol{y}$ *substituted by* $\tilde{\boldsymbol{y}}^{(i)}$. *Then,* $\boldsymbol{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)} = \boldsymbol{U}_i\tilde{\boldsymbol{\gamma}}^{(i)}$.

Note that $\hat{\boldsymbol{\gamma}} = \boldsymbol{A}\boldsymbol{y}$ and $\tilde{\boldsymbol{\gamma}}^{(i)} = \boldsymbol{A}\tilde{\boldsymbol{y}}^{(i)}$. As a consequence of Lemma 1,

$$\boldsymbol{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)} = \boldsymbol{U}_i\boldsymbol{A}\tilde{\boldsymbol{y}}^{(i)} = \boldsymbol{U}_i(\sum_{k\neq i}\boldsymbol{A}_k\boldsymbol{y}_k + \boldsymbol{A}_i\boldsymbol{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)}) = \boldsymbol{U}_i(\hat{\boldsymbol{\gamma}} - \boldsymbol{A}_i\boldsymbol{y}_i + \boldsymbol{A}_i\boldsymbol{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)}).$$

By some standard calculation, we have

$$\boldsymbol{y}_i - \boldsymbol{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)} = (\boldsymbol{I} - \boldsymbol{U}_iA_i)^{-1}(\boldsymbol{y}_i - \boldsymbol{U}_i\hat{\boldsymbol{\gamma}}) = (\boldsymbol{I} - \boldsymbol{M}_{ii}(L,\lambda))^{-1}(\boldsymbol{y}_i - \boldsymbol{U}_i\hat{\boldsymbol{\gamma}}).$$

Therefore,

$$ACV(L,\lambda) = \sum_{i=1}^{N}\|(\boldsymbol{I} - \boldsymbol{M}_{ii}(L,\lambda))^{-1}(\boldsymbol{y}_i - \boldsymbol{U}_i\hat{\boldsymbol{\gamma}})\|_2^2,$$

in which we only need to solve the inverse of the J_i-dimensional matrices $I - M_{ii}$. Then, we can choose the optimal (L,λ) by minimizing $ACV(L,\lambda)$.

3. Large-sample Properties

For the standard parametric linear regression models, Fan and Li (2001) established the oracle property of the SCAD penalized estimates, which indicates that the SCAD penalty enables consistent variable selection and parameter estimation simultaneously, as if the subset of relevant variables is already known. We show that this oracle property also holds for our proposed gSCAD method for varying coefficients models. In addition, we also establish the consistency and the convergence rate of our estimates of the smooth functions. Assume that only s predictors are relevant in the Model (1.1). Without loss of generality, let $\boldsymbol{\beta}_k(t)$, $k = 1, \ldots, s$ be the non-zero coefficients, and $\boldsymbol{\beta}_k(t) = 0$, $k = s+1, \ldots, p$. We made the following technical assumptions:

Assumption 1: The subjects $(y_i(t), \boldsymbol{x}_i(t))$, $i = 1, \ldots, n$, are i.i.d., and the observation time points t_{ij} are i.i.d. from an unknown density $f(t)$ on $[0, T]$, where $f(t)$ are uniformly bounded away from infinity and zero.

Assumption 2: The eigenvalues of the matrix $E\{X(t)X^T(t)\}$ are uniformly bound away from infinity and zero for all t.

Assumption 3: There exists a positive constant M_1 such that $|x_{ik}(t)| \leq M_1$ for all t.

Assumption 4: There exists a positive constant M_2 such that $E\varepsilon^2(t) \leq M_2$ for all t.

Define $\mathcal{G}_k(L_k) = \{g(t) = \sum_{l=1}^{L_k} \gamma_{kl} B_{kl}(t)\}$, $\rho_n = \sum_{k=1}^{p} \inf_{g \in \mathcal{G}_k} \|\beta_k - g\|_{L_\infty}$, $L_n = \max_{1 \leq k \leq p} L_k$. Here ρ_n is an approximation error of $\mathcal{G}_k(L_k)$ to $\beta_k(t)$, $k = 1, \ldots, p$, which approaches zero as L_k grows to infinity at a proper rate with sample size n. Furthermore, define

$$A_k(L_k) = \sup_{g \in \mathcal{G}_k(L_k), \|g\|_{L_2} \neq 0} ||g||_{L_\infty} / \|g\|_{L_2} \quad , \quad A_n = \max_k A_k(L_k),$$

where $\|g\|_{L_\infty} = \sup_{t \in [0,T]} |g(t)|$, and $\|g\|_{L_2} = (\int_{[0,T]} g(t)^2 \, dt)^{1/2}$. The following theorem, the proof of which is given in the Appendix B, shows the oracle property, the consistency and the convergence rates of the estimates.

Theorem 1. *Suppose the assumptions 1-4 listed above are satisfied,* $\lim_{n \to \infty} \rho_n = 0$ *and*

$$\lim_{n \to \infty} A_n^2 L_n \max\{N^{-1} \max_{1 \leq i \leq n} (J_i), N^{-2} \sum_i J_i^2\} = 0. \tag{3.1}$$

Then, with a choice of λ_n *such that* $\lambda_n \to 0$ *and* $\lambda_n / \max\{r_n, \rho_n\} \to \infty$*, we have*

(a.) $\hat{\boldsymbol{\beta}}_k = 0$, $k = s+1, \ldots, p$, *with probability approaching 1.*

(b.) $\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_{L_2} = O_p(\max(r_n, \rho_n))$, $k = 1, \ldots, s$, *with* $r_n = (N^{-2} L_n \sum_{i=1}^{n} J_i^2)^{1/2}$.

Note that Theorem 1 gives the oracle property (1(a)) and the consistency (1(b)) for general basis choices for basis function approximations. The following corollary gives a specific convergence rates for a class of spline estimators (see Appendix C for the proof).

Corollary 1. *Suppose that the assumptions in Theorem 1 hold,* $J_i = J$, $i = 1, \ldots, n$, *and that* $\beta_k(t)$ *have bounded second derivatives,* $k = 1, \ldots, s$, $\beta_k(t) = 0$, $k = s+1, \ldots, p$. *Let* $\lambda_n \to 0$, $n^{2/5}\lambda_n \to \infty$, *and let* $\{B_{kl}(t)\}_{l=1}^{L_k+4}$ *be the cubic spline basis with* L_k *equally spaced interior knots, where* $L_k = O(n^{1/5})$, $k = 1, \ldots, p$. *Then,*

(a.) $\hat{\boldsymbol{\beta}}_k = 0$, $k = s+1, \ldots, p$, *with probability approaching 1.*

(b.) $\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_{L_2} = O_p(n^{-2/5})$, $k = 1, \ldots, s$.

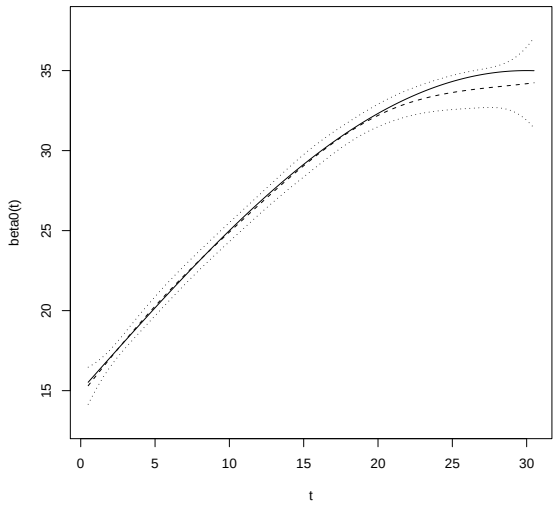
Note that the rate of convergence is the optimal rate for nonparametric regression with independent, identically distributed data under the same smoothness assumptions (Stone, 1982).

4. Monte Carlo Simulation

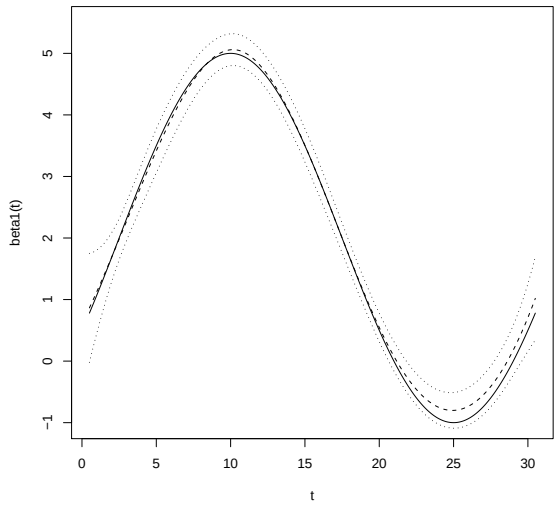
We conducted simulation studies to assess the performance of the proposed procedure. In each simulation run, we generated a simple random sample of 200 subjects according to the model used in Huang *et al.* (2002), which assumes

$$Y(t_{ij}) = \beta_0(t_{ij}) + \sum_{k=1}^{23} \beta_k(t_{ij}) x_k(t_{ij}) + \varepsilon(t_{ij}), \quad i = 1, \ldots, 200, j = 1, \ldots, J_i.$$

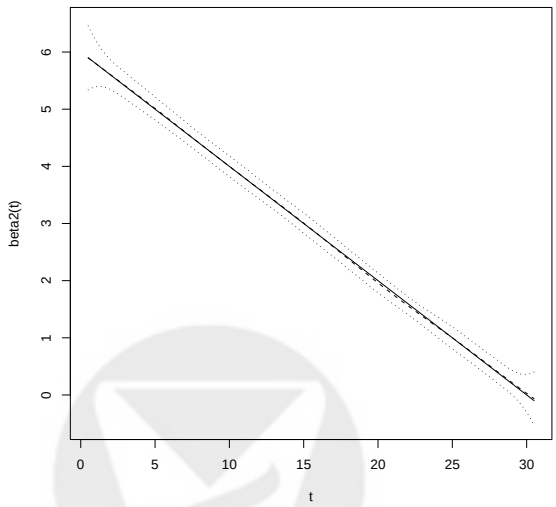
The first three variables $x_i(t)$, $i = 1, \ldots, 3$, are the true relevant covariates, which are simulated the same way as in Huang *et al.* (2002): $x_1(t)$ is sampled uniformly from $[t/10, 2 + t/10]$ at any given time point t; $x_2(t)$, conditioning on $x_1(t)$, is gaussian with mean zero and variance $(1+x_1(t))/(2+x_1(t))$; $x_3(t)$, independent of x_1 and x_2, is a Bernoulli random variable with success rate 0.6. In addition to x_k, $k = 1, 2, 3$, 20 redundant variables $x_k(t)$, $k = 4, \ldots, 23$, are simulated to demonstrate the performance of variable selection, where each $x_k(t)$, independent of each other, is a random realization of a gaussian process with covariance structure $\mathrm{cov}(x_k(t), x_k(s)) = 4\exp(-|t-s|)$. The random error $\varepsilon(t)$ is given by $Z(t) + E(t)$, where $Z(t)$ has the same distribution as $x_k(t)$, $k = 4, \ldots, 23$, and $E(t)$ are independent measurement errors from $N(0, 4)$ distribution at each time t. The coefficients $\beta_k(t)$, $k = 0, \ldots, 3$, corresponding to the constant



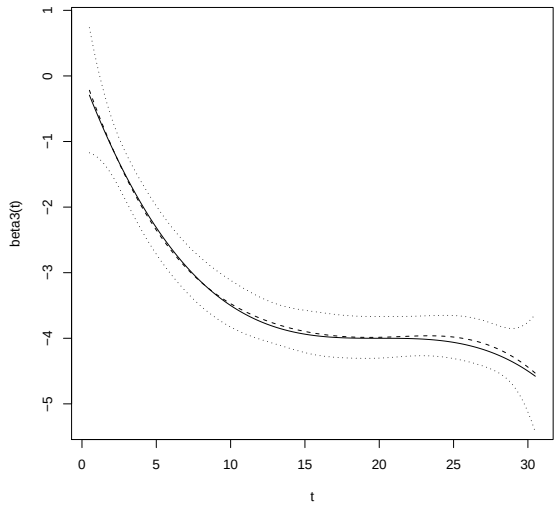
(a) $\beta_0(t)$



(b) $\beta_1(t)$



(c) $\beta_2(t)$



(d) $\beta_3(t)$

Figure 4.1: *True (solid lines) and average of estimated (dashed lines) time-varying coefficients $\beta_0(t)$ (a), $\beta_1(t)$ (b), $\beta_2(t)$ (c) and $\beta_3(t)$ (+/- 1 point-wise SE) over 100 replications.*

term and the first three variables, are given by

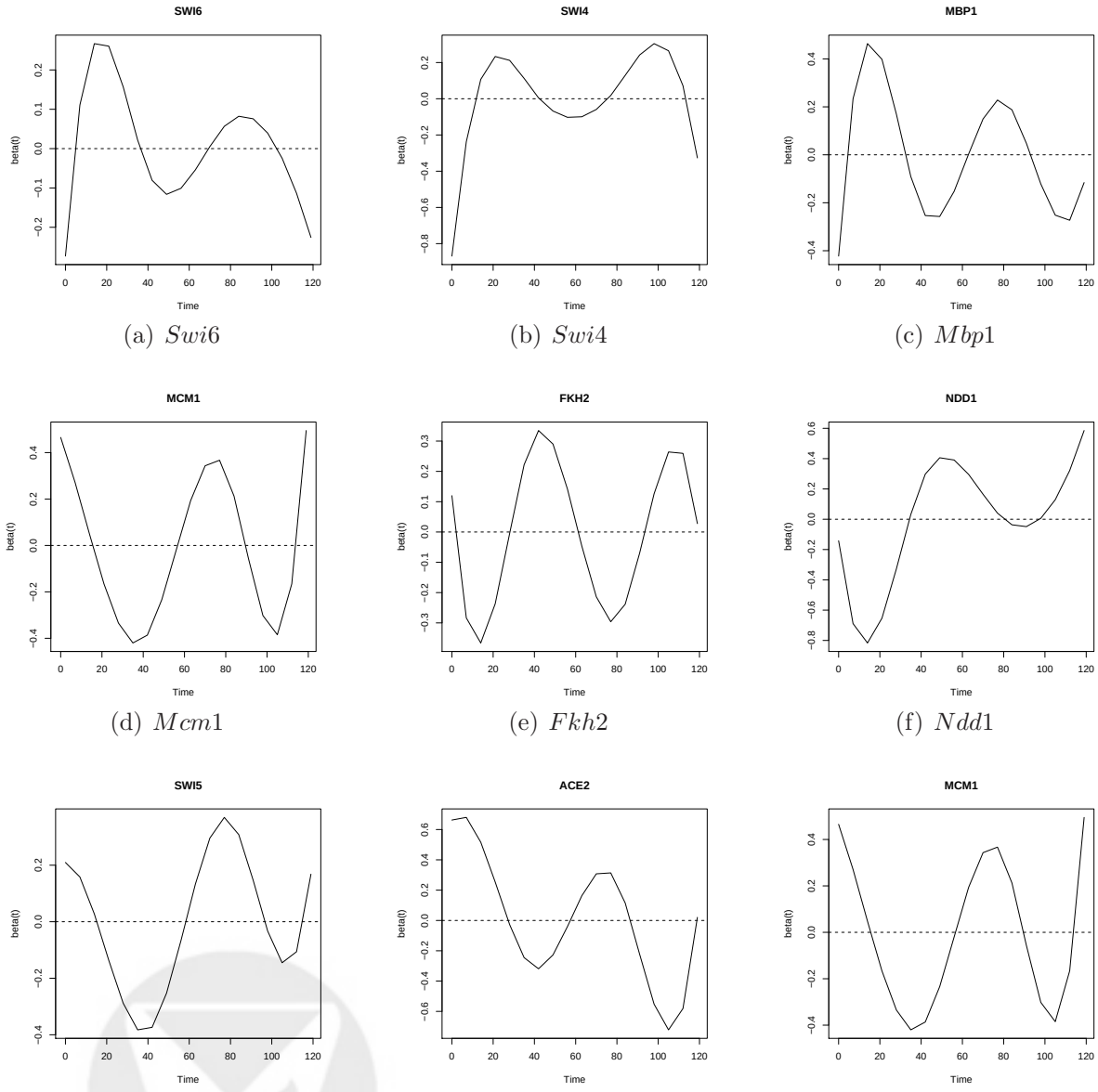
$$\beta_0(t) = 15 + 20\sin(\frac{\pi t}{60}), \qquad \beta_1(t) = 2 - 3\cos(\frac{\pi(t-25)}{15}),$$
$$\beta_2(t) = 6 - 0.2t, \qquad \beta_3(t) = -4 + \frac{(20-t)^3}{2000},$$

(see solid lines of Figure 4.1) while the remaining coefficients, corresponding to the irrelevant variables, are given by $\beta_k(t) = 0$, $k = 4, \ldots, 23$. The observation time points t_{ij} are generated following the same scheme as in Huang *et al.* (2002), where each subject has a set of "scheduled" time points $\{1, \ldots, 30\}$, and each scheduled time has a probability of 60% of being skipped. Then, the actual observation time t_{ij} is obtained by adding a random perturbation from $Uniform(-0.5, 0.5)$ to the nonskipped scheduled time.

We repeated the simulations 100 times. Out of these 100 replications, the variables 1-3 were selected in each of the runs. Figure 4.1 shows the estimates of the time-varying coefficients of $\beta_k(t), k = 0, 1, 2, 3$, indicating that the estimates fit the true function very well. As a comparison, among the 20 irrelevant variables, 10 were selected 1 time, 3 were selected 2 times, 5 were selected three times and 2 were selected 5 times. The simulations indicate that the proposed procedure indeed provides an effective method for selecting variables with time-varying coefficients and for estimating the coefficient functions.

5. Application to Yeast Cell Cycle Gene Expression Data

We present results from our analysis of the yeast cell cycle microarray gene expression data set of Spellman *et al.* (1998). They monitored genome-wide mRNA levels for 6178 yeast ORFs simultaneously using several different methods of synchronization including an α-factor-mediated G_1 arrest, which covers approximately two cell-cycle periods with measurements at 7-min intervals for 119 mins with a total of 18 time points. Using a model-based approach, Luan and Li (2003) identified 297 cell-cycle regulated genes based on the α-factor synchronization experiments. In addition, we applied the mixture model approach (Chen *et al.*, 2007) using the ChIP data of Lee *et al.* (2002) to derive the binding probabilities x_{ik} for these 297 cell-cycle-regulated genes for a total of 96 transcriptional factors with at least one nonzero binding probability in the 297 genes. Our goal is to identify the transcriptional factors that might be related to the expression patterns of these 297 cell-cycle-regulated genes. Since different transcriptional factors may regulate the gene expression at different time points during the cell-cycle process, their effects on gene expression are expected to be time-dependent.



(a) $Swi6$ (b) $Swi4$ (c) $Mbp1$

(d) $Mcm1$ (e) $Fkh2$ (f) $Ndd1$

(g) $Swi5$ (h) $Ace2$ (i) $Mcm1$

Figure 5.2: *Estimated time-varying coefficients for selected transcription factors (TFs): (a)-(c): TFs that regulate genes expressed at the G1 phase; (d)-(f): TFs that regulate genes expressed at the G2 phase; (g)-(i) TFs that regulate genes expressed at the M phase.*

We applied the gSCAD procedure using the GCV for selecting the tuning parameter in order to identify the TFs that affect the expression changes over time for these 297 cell-cycle-regulated genes in the α-factor synchronization experiment. The gSCAD procedure identified a total of 71 TFs that are related to yeast cell-cycle processes, including 19 of the 21 known and experimentally verified cell-cycle related TFs, all showing time-dependent effects of these TFs on gene expression levels. In addition, the effects followed similar trends between the two cell-cycle periods. The other two TFs, CBF1 and GCN4 were not selected by the gSCAD procedure; it was not clear why CBF1 and GCN4 were not selected by the gSCAD. The minimum p-values over 18 time points from simple linear regressions are 0.06 and 0.14, respectively, also indicating that CBF1 and GCN4 were not related to expression variation over time. Overall, the model can explain 43% of the total variations of the gene expression levels.

Figure 5.2 shows that estimated time-dependent transcriptional effects of nine of the experimentally verified TFs. The top panel shows the transcriptional effects of three TFs, Swi4, Swi6 and MBP1, that regulate gene expression at the G1 phase (Simon *et al.*, 2001). The estimated transcriptional effects of these three TFs are quite similar with peak effects obtained at the time points corresponding to the G1 phase of the cell cycle process. The middle panel shows the transcriptional effects of three TFs, Mcm1, Fkh2 and Ndd1, that regulate gene expression at the G2 phase (Simon *et al.*, 2001). Again, the estimated transcriptional effects of these three TFs are quite similar with peak effects obtained at the time points corresponding to the G2 phase of the cell cycle process. Finally, the bottom panel shows the transcriptional effects of three TFs, Swi5, Ace2 and Mcm1, that regulate gene expression at the M phase (Simon *et al.*, 2001), indicating similar transcriptional effects of these three TFs with peak effects at the point points corresponding to the M phase of the cell cycle.

The 52 additional TFs that were selected by the gSCAD procedure almost all showed estimated periodic transcriptional effects. The identified TFs include many pairs of cooperative or synergistic pairs of TFs involved in the yeast cell cycle process reported in the literature (Banerjee and Zhang, 2003; Tsai *et al.*, 2005). Of these 52 TFs, 34 of them belong to the cooperative pairs of the TFs identified by Banerjee and Zhang (2003).

Finally, to assess false identifications of the TFs that are related to a dynamic biological procedure, we randomly permuted the gene expression values across genes and time points and applied the gSCAD procedure again to the permuted data sets. We repeated this procedure 50

times. Among the 50 runs, 5 runs selected 4 TFs, 1 run selected 3 TFs, 16 runs selected 2 TFs and the rest of the 28 runs did not select any of the TFs, indicating that our procedure indeed selects the relevant TFs with few false positives.

6. Discussion

We have proposed a regularized estimation procedure variable selection for nonparametric varying-coefficient models. Such a procedure can simultaneously perform variable selection and estimation of the smooth functions and can be applied to both the longitudinal setting and the functional responses setting. The proposed gSCAD estimator have the oracle properties and is easy to solve using a local quadratic approximation algorithm. Simulation studies indicated that this procedure is very effective in selecting the relevant groups of variables and in estimating the regression coefficients. Results from application to the yeast cell cycle data set indicate that the procedure can be effective in selecting the transcriptional factors that potentially play important roles in regulation of gene expressions during the cell cycle process.

This paper focuses on linear varying-coefficient models; however, the proposed estimation procedure can be extended to more general regression models, such as the varying coefficients Cox regression model or the generalized linear models (Hastie and Tibshirani, 1993). Another possibility of extending the proposed work is to use smoothing splines for estimating the varying coefficients, with nodes chosen at the observed time points and a smoothing parameter to control the smoothness of the coefficients. We are currently pursuing these extensions.

Acknowledgments

This research was supported by NIH grant CA127334 and a grant from the Pennsylvania Department of Health. We thank Mr. Edmund Weisberg, MS at Penn CCEB for editorial assistance.

References

Banerjee N and Zhang MQ (2003): Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Research*, 31: 7024-7031.

Bickel P and Li B (2006): Regularization in Statistics (with discussion). *Test*, 15: 271-344.

Chen G, Jensen S, and Stoeckert C (2007): Clustering of genes into regulons using integrated modeling (CORIM). *Genome Biology*, 8, 1, R4.

Conlon EM, Liu XS, Lieb JD and Liu JS (2003): Integrating regulatory motif discovery and genome-wide expression analysis *Proceedings of National Academy of Sciences*, 100: 3339-3344;

Diggle PJ, Liang KY and Zeger SL (1994): *Analysis of longitudinal data.* Oxford: Oxford University Press.

Efron B, Hastie T, Johnstone I and Tibshirani R (2004): Least angle regression *Annals of Statistics*, 32, 407499.

Fan J and Li R (2001): Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association*, 96: 1348-1360.

Fan J and Li R (2004): New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of American Statistical Association*, 99: 710-723.

Hastie T and Tibshirani R (1993): Varying-coefficient models. *Journal of Royal Statistical Society, Ser B*, 55: 757-796.

Hoover DR, Rice JA, Wu CO and Yang L (1998): Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85: 809-822.

Huang JZ, Wu CO and Zhou L (2002): Varying-coefficient models and basis function approximation for the analysis of repeated measurements. *Biometrika*, 89: 111-128.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, *et al.*: Transcriptional regulatory networks in *S. cerevisiae*. *Science*, 298: 799-804.

Lin X and Carroll RJ (2000): Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of American Statistical Association*, 95: 520-534.

Luan Y and Li H (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, 19, 474-482.

Rice RA (2004): Functional and longitudinal data analysis: perspectives on smoothing. *Statistica Sinica*, 14: 631-647.

Rice JA and Silverman BW (1991): Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of Royal Statistical Society B*, 53: 233-243.

Rice JA and Wu CO (2001): Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, 57: 253-259.

Schumaker LL (1981): *Spline Functions: Basic Theory*. Wiley, New York.

Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakola TS and Young RA (2001): Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106: 697-708.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D and Futcher B (1998): Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of Cell*, 9, 3273-3297

Stone CJ (1982): Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, 10: 1348-1360.

Tibshirani RJ (1996): Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*, 58: 267-288.

Tsai HK, Lu SHH, and Li WH (2005): Statistical methods for identifying yeast cell cycle transcription factors. *Proceedings of National Academy of Sciences*, 102(38): 13532 - 13537.

Wang L, Chen G and Li H (2007): Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, in press.

Yuan M and Lin Y (2006): Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society B*, 68: 49-67.

Zou H (2006): The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.

Zou H and Hastie T (2005): Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society SER B-STAT MET*, 67:301-320.

Zeger SL and Diggle PJ (1994): Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, 50: 689-699.

Appendix

Appendix A. Lemmas and proof

We first prove Lemma 1 that is related to leave-on-out cross-validation analysis. We then present and prove the other lemmas used in the proof of Theorem 1.

Proof of Lemma 1: Use proof by contradiction. Suppose $\boldsymbol{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)} \neq \boldsymbol{U}_i\tilde{\boldsymbol{\gamma}}^{(i)}$. Denote (2.7) as $l_{ridge}(\boldsymbol{\gamma}, \boldsymbol{y})$. Then,

$$\begin{aligned} l_{ridge}(\tilde{\boldsymbol{\gamma}}^{(i)}, \tilde{\boldsymbol{y}}^{(i)}) &= \sum_{k=1}^{n} \|\tilde{\boldsymbol{y}}_k^{(i)} - \boldsymbol{U}_k\tilde{\boldsymbol{\gamma}}_k^{(i)}\|_2^2 + N/2(\tilde{\boldsymbol{\gamma}}^{(i)})^T\Sigma_\lambda(\hat{\boldsymbol{\gamma}})\tilde{\boldsymbol{\gamma}}^{(i)} \\ &> \sum_{k\neq i} \|\boldsymbol{y}_k - \boldsymbol{U}_k\tilde{\boldsymbol{\gamma}}_k^{(i)}\|_2^2 + N/2(\tilde{\boldsymbol{\gamma}}^{(i)})^T\Sigma_\lambda(\hat{\boldsymbol{\gamma}})\tilde{\boldsymbol{\gamma}}^{(i)} \\ &\geq \sum_{k\neq i} \|\boldsymbol{y}_k - \boldsymbol{U}_k\hat{\boldsymbol{\gamma}}_k^{\star(-i)}\|_2^2 + N/2(\hat{\boldsymbol{\gamma}}^{\star(-i)})^T\Sigma_\lambda(\hat{\boldsymbol{\gamma}})\hat{\boldsymbol{\gamma}}^{\star(-i)} \\ &= l_{ridge}(\hat{\boldsymbol{\gamma}}^{\star(-i)}, \tilde{\boldsymbol{y}}^{(i)}). \end{aligned}$$

This contradicts the fact that $\tilde{\boldsymbol{\gamma}}^{(i)}$ minimizes $l_{ridge}(\tilde{\boldsymbol{\gamma}}^{(i)}, \tilde{\boldsymbol{y}}^{(i)})$, which proves the result. □

Define $\tilde{\boldsymbol{y}}_i = E(\boldsymbol{y}_i|\boldsymbol{x}_i)$, $\tilde{\boldsymbol{\gamma}} = (\sum_{i=1}^n \boldsymbol{U}_i^T\boldsymbol{U}_i)^{-1}(\sum_{i=1}^n \boldsymbol{U}_i^T\tilde{\boldsymbol{y}}_i)$, and $\tilde{\beta}(t) = \boldsymbol{B}(t)\tilde{\boldsymbol{\gamma}}$. Here $\tilde{\beta}(t)$ can be regarded as the conditional mean of $\hat{\boldsymbol{\beta}}(t)$.

To prove Theorem 1, we use the scheme as follows. First, using Lemma 2-3, we quantify the convergence rate of $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L_2}$, which is established in Lemma 4. Then by Lemma 4, we prove the consistency of variable selection in part (a) of Theorem 1. Finally, we improve the rate in Lemma 4 to obtain the rate in part (b) of Theorem 1.

Lemma 2. *Suppose that (3.1) holds. There exists an interval $[M_3, M_4]$, $0 < M_3 < M_4 \leq \infty$, such that all the eigenvalues of $N^{-1}\sum_{i=1}^n \boldsymbol{U}_i^T\boldsymbol{U}_i$ fall in $[M_3, M_4]$ with probability approaching 1 as $n \to \infty$.*

Lemma 3. *Suppose that (3.1) holds. Then, $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{L_2} = O_P(\rho_n)$.*

Lemma 2 and 3 are from Lemma A2 and A3 in Huang *et al.* (2002), respectively. The proofs are omitted.

Lemma 4. *Suppose that (3.1) holds and that* $\lambda \equiv \lambda_n$, r_n *and* ρ_n *approach 0 as* $n \to \infty$ *satisfying* $\lambda_n / \max\{r_n, \rho_n\} \to \infty$. *Then,* $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L_2} = O_P(r_n + (\lambda_n \rho_n)^{1/2})$.

Proof of Lemma 4: Because of the orthonormality of the basis, $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L_2}^2 = \sum_{k=1}^{p} \|\hat{\boldsymbol{\gamma}}_k - \tilde{\boldsymbol{\gamma}}_k\|_2^2 = (\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})^T (\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}})$. Let $\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}} = \delta \boldsymbol{u}$ with δ a scaler and $\boldsymbol{u}$ a vector satisfying $\|\boldsymbol{u}\|_2 = 1$. We first prove that $\|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\|_2 = \delta = O_p(r_n + \lambda_n)$.

Note that

$$\begin{aligned} l(\hat{\boldsymbol{\gamma}}) - l(\tilde{\boldsymbol{\gamma}}) &= N^{-1} \sum_{i=1}^{n} (\|\boldsymbol{y}_i - \boldsymbol{U}_i(\tilde{\boldsymbol{\gamma}} + \delta \boldsymbol{u})\|_2^2 - \|\boldsymbol{y}_i - \boldsymbol{U}_i \tilde{\boldsymbol{\gamma}}\|_2^2) \\ &\qquad + \sum_{k=1}^{p} (p_\lambda(\|\tilde{\boldsymbol{\gamma}}_k + \delta \boldsymbol{u}_k\|_2) - p_\lambda(\|\tilde{\boldsymbol{\gamma}}_k\|_2)) \\ &= (-2N^{-1} \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i{}^T \boldsymbol{U}_i \boldsymbol{u}) \delta + (N^{-1} \boldsymbol{u}^T \sum_{i=1}^{n} \boldsymbol{U}_i^T \boldsymbol{U}_i \boldsymbol{u}) \delta^2 \\ &\qquad + \sum_{k=1}^{p} (p_\lambda(\|\tilde{\boldsymbol{\gamma}}_k + \delta \boldsymbol{u}_k\|_2) - p_\lambda(\|\tilde{\boldsymbol{\gamma}}_k\|_2)), \end{aligned} \tag{7.1}$$

where $\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_{i1}), \ldots, \varepsilon_i(t_{iJ_i}))^T$. For the first term in (7.1), note that

$$E_t(B_{kl}(t)^2) = \int_0^T B_{kl}(t)^2 f(t) dt \le \sup_{t \in [0,T]} f(t) \int_0^T B_{kl}(t)^2 dt = \sup_{t \in [0,T]} f(t),$$

Then,

$$\begin{aligned} E(\varepsilon_i \boldsymbol{U}_i \boldsymbol{u})^2 &= E(\sum_{j=1}^{J_i} \boldsymbol{\varepsilon}_i(t_{ij}) \boldsymbol{x}_i^T(t_{ij}) \boldsymbol{B}(t_{ij}) \boldsymbol{u})^2 \le E[\sum_j \varepsilon_i(t_{ij})^2 \sum_j (\boldsymbol{x}_i^T(t_{ij}) \boldsymbol{B}(t_{ij}) \boldsymbol{u})^2] \\ &\le J_i M_2 E[\sum_j (\boldsymbol{x}_i^T(t_{ij}) \boldsymbol{B}(t_{ij}) \boldsymbol{u})^2] \\ &\le J_i M_2 \sum_j E[\|\boldsymbol{x}_i(t_{ij})\|_2^2 \|\boldsymbol{u}\|_2^2 \mathrm{tr}\{\boldsymbol{B}(t_{ij}) \boldsymbol{B}^T(t_{ij})\}] \\ &\le J_i M_2 p M_3^2 \sum_j E_{\boldsymbol{t}}(\sum_k \sum_l B_{kl}(t_{ij})^2) = O(J_i^2 L_n). \end{aligned}$$

As a result,

$$N^{-1} \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i{}^T \boldsymbol{U}_i \boldsymbol{u} = O_P(E[N^{-1} \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i{}^T \boldsymbol{U}_i \boldsymbol{u}]^2)^{1/2} = O_P(N^{-2} L_n \sum_{i=1}^{n} J_i^2)^{1/2} = O_P(r_n). \tag{7.2}$$

For the second term in (7.1), by Lemma 2,

$$(N^{-1}\boldsymbol{u}^T\sum_{i=1}^{n}\boldsymbol{U}_i^T\boldsymbol{U}_i\boldsymbol{u}) \geq M_3 \tag{7.3}$$

with probability approaching 1.

For the third term in (7.1), note that $|p_\lambda(a) - p_\lambda(b)| \leq \lambda|a-b|$. Therefore,

$$\sum_{k=1}^{p}(p_\lambda(\|\tilde{\boldsymbol{\gamma}}_k + \delta\boldsymbol{u}_k\|_2) - p_\lambda(\|\tilde{\boldsymbol{\gamma}}_k\|_2) \geq -\lambda\|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\|_2 = -\lambda\delta. \tag{7.4}$$

Combining (7.2), (7.3), (7.4) and the fact that $l(\hat{\boldsymbol{\gamma}}) - l(\tilde{\boldsymbol{\gamma}}) \leq 0$, (7.1) becomes $0 \geq -O_p(r_n)\delta + M_3\delta^2 - \lambda\delta$ with probability approaching 1, which implies $\|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\|_2 = \delta = O_p(r_n + \lambda_n)$.

Now we proceed to improve the obtained rate and show that $\delta = O_P(r_n + (\lambda_n\rho_n)^{1/2})$. Note that $|\|\hat{\boldsymbol{\gamma}}_k\|_2 - \|\tilde{\boldsymbol{\gamma}}_k\|_2| = o_p(1)$, and $|\|\tilde{\boldsymbol{\gamma}}_k\|_2 - \|\beta_k\|_{L_2}| \leq \|\tilde{\boldsymbol{\beta}}_k - \beta_k\|_{L_2} = O_p(\rho_n)$, $k = 1, \ldots, p$. We have $\|\hat{\boldsymbol{\gamma}}_k\|_2 \to \|\beta_k\|_{L_2} > a\lambda_n$ in probability, $\|\tilde{\boldsymbol{\gamma}}_k\|_2 \to \|\beta_k\|_{L_2} > a\lambda_n$ in probability, $k = 1, \ldots, s$, and $\|\tilde{\boldsymbol{\gamma}}_k\|_2 = O_p(\rho_n) < \lambda_n$ in probability, $k = s+1, \ldots, p$, because $\lambda_n \to 0$ and $\lambda_n/\rho_n \to \infty$. By the definition of $p_\lambda(\cdot)$, it follows that $P\{p_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_2) = p_{\lambda_n}(\|\hat{\boldsymbol{\gamma}}_k\|_2)\} \to 1$, $k = 1, \ldots, s$, and $P\{p_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_2) = \lambda_n\|\tilde{\boldsymbol{\gamma}}_k\|_2\} \to 1$, $k = s+1, \ldots, p$. Combining with (7.2) and (7.3), with probability approaching 1, we have,

$$\begin{aligned} l(\hat{\boldsymbol{\gamma}}) - l(\tilde{\boldsymbol{\gamma}}) \quad &\geq \quad N^{-1}\sum_{i=1}^{n}(\|\boldsymbol{y}_i - \boldsymbol{U}_i(\tilde{\boldsymbol{\gamma}} + \delta\boldsymbol{u})\|_2^2 - \|\boldsymbol{y}_i - \boldsymbol{U}_i\tilde{\boldsymbol{\gamma}}\|_2^2) \\ &\qquad + \sum_{k=1}^{s}(p_\lambda(\|\tilde{\boldsymbol{\gamma}}_k + \delta\boldsymbol{u}_k\|_2) - p_\lambda(\|\tilde{\boldsymbol{\gamma}}_k\|_2)) - \sum_{k=s+1}^{p} p_\lambda(\|\tilde{\boldsymbol{\gamma}}_k\|_2) \\ &\geq \quad -O_p(r_n)\delta + M_3\delta^2 - O_p(\lambda_n\rho_n), \end{aligned}$$

which implies $\|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\|_2 = \delta = O_P(r_n + (\lambda_n\rho_n)^{1/2})$. This proves the desired result. □

Appendix B. Proof of Theorem 1

To prove part (a) of Theorem 1, we use proof by contradiction. Suppose that there exists a constant $\delta > 0$ such that with probability at least δ, there exist a large n and a $k_0 > s$ such that $\hat{\beta}_{k_0}(t) \neq 0$. Then $\|\hat{\boldsymbol{\gamma}}_{k_0}\|_2 = \|\hat{\beta}_{k_0}(t)\|_{L_2} > 0$. Let $\boldsymbol{\gamma}^*$ be a vector constructed by replacing $\hat{\boldsymbol{\gamma}}_{k_0}$ with $\mathbf{0}$ in $\hat{\boldsymbol{\gamma}}$. Note that $\lambda_n/\max(r_n, \rho_n) \to \infty$. Then, $\lambda_n > \|\hat{\boldsymbol{\gamma}}_{k_0}\|_2 = O_p(r_n + (\lambda_n\rho_n)^{1/2})$ with

probability approaching 1, and

$$\begin{aligned} l(\hat{\boldsymbol{\gamma}}) - l(\boldsymbol{\gamma}^*) &= N^{-1}\sum_{i=1}^{n}(\|\boldsymbol{y}_i - \boldsymbol{U}_i\hat{\boldsymbol{\gamma}}\|_2^2 - \|\boldsymbol{y}_i - \boldsymbol{U}_i\boldsymbol{\gamma}^*\|_2^2) + p_\lambda(\|\hat{\boldsymbol{\gamma}}_{k_0}\|_2) \\ &= N^{-1}\sum_{i=1}^{n}(-2\boldsymbol{y}_i^T\boldsymbol{U}_i(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*) + (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)^T\boldsymbol{U}_i^T\boldsymbol{U}_i(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)) + \lambda\|\hat{\boldsymbol{\gamma}}_{k_0}\|_2 \\ &\geq -O_p(r_n)\|\hat{\boldsymbol{\gamma}}_{k_0}\|_2 + M_3\|\hat{\boldsymbol{\gamma}}_{k_0}\|_2^2 + \lambda_n\|\hat{\boldsymbol{\gamma}}_{k_0}\|_2. \end{aligned} \tag{7.5}$$

Note that in (7.5), the third term dominates both the first term and the second term. This contradicts the fact that $l(\hat{\boldsymbol{\gamma}}) - l(\boldsymbol{\gamma}^*) \leq 0$, which proves part (a).

To prove part (b), first write $\boldsymbol{\gamma} = ((\boldsymbol{\gamma}^{(1)})^T, (\boldsymbol{\gamma}^{(2)})^T)^T$, where $\boldsymbol{\gamma}^{(1)} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_s^T)^T$ and $\boldsymbol{\gamma}^{(2)} = (\boldsymbol{\gamma}_{s+1}^T, \ldots, \gamma_p^T)^T$. Similarly, write $\boldsymbol{\beta}(t) = (\boldsymbol{\beta}^{(1)^T}, \boldsymbol{\beta}^{(2)^T})^T$ and $\boldsymbol{U}_i = (\boldsymbol{U}_i^{(1)}, \boldsymbol{U}_i^{(2)})$. Then define the oracle version of $\tilde{\boldsymbol{\gamma}}$,

$$\begin{aligned} \tilde{\boldsymbol{\gamma}}_{oracle} &= \arg\min_{\boldsymbol{\gamma}=(\boldsymbol{\gamma}^{(1)^T}, \mathbf{0}^T)^T} N^{-1}\sum_{i=1}^{n}(\tilde{\boldsymbol{y}}_i - \boldsymbol{U}_i\boldsymbol{\gamma})^T(\tilde{\boldsymbol{y}}_i - \boldsymbol{U}_i\boldsymbol{\gamma}) \\ &= \begin{pmatrix} (\sum_i \boldsymbol{U}_i^{(1)^T}\boldsymbol{U}_i^{(1)})^{-1}(\sum_i \boldsymbol{U}_i^{(1)^T}\tilde{\boldsymbol{y}}_i) \\ \mathbf{0} \end{pmatrix}, \end{aligned}$$

which is obtained as if the information of the nonzero components were given. Note that the true $\boldsymbol{\beta}(t) = (\boldsymbol{\beta}^{(1)^T}, \mathbf{0}^T)^T$. Then by Lemma 3, $\|\tilde{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\|_{L_2} = O_P(\rho_n)$. To quantify $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{oracle}\|_{L_2} = \|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_{oracle}\|_2$, note that by part (a) of Theorem 1, with probability approaching 1, $\hat{\boldsymbol{\gamma}} = ((\hat{\boldsymbol{\gamma}}^{(1)})^T, \mathbf{0}^T)^T$. Let $\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_{oracle} = \delta\boldsymbol{u}$ with $\boldsymbol{u} = ((\boldsymbol{u}^{(1)})^T, \mathbf{0}^T)^T$ and $\|\boldsymbol{u}^{(1)}\|_2 = 1$. Then, with probability approaching 1,

$$\begin{aligned} l(\hat{\boldsymbol{\gamma}}) - l(\tilde{\boldsymbol{\gamma}}_{oracle}) &= N^{-1}\sum_{i=1}^{n}(\|\boldsymbol{y}_i - \boldsymbol{U}_i(\tilde{\boldsymbol{\gamma}}_{oracle} + \delta\boldsymbol{u})\|_2^2 - \|\boldsymbol{y}_i - \boldsymbol{U}_i\tilde{\boldsymbol{\gamma}}_{oracle}\|_2^2) \\ &= (-2N^{-1}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i{}^T\boldsymbol{U}_i^{(1)}\boldsymbol{u}^{(1)})\delta + (N^{-1}(\boldsymbol{u}^{(1)})^T\sum_{i=1}^{n}(\boldsymbol{U}_i^{(1)})^T\boldsymbol{U}_i^{(1)}\boldsymbol{u}^{(1)})\delta^2 \\ &\geq -O_p(r_n)\delta + M_3\delta^2, \end{aligned}$$

which implies $\|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_{oracle}\|_2 = \delta = O_p(r_n)$. By triangle inequality, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{L^2} = O_p(\rho_n + r_n)$, and the result follows. □

Appendix C. Proof of Corollary 1

By Theorem 6.27 in Schumaker (1981), $\inf_{g\in\mathcal{G}_k}\|\beta_k - g\|_{L_\infty} = O(L_k^{-2})$. Thus $\rho_n = O(\sum_{k=1}^{p} L_k^{-2})$, and $r_n = (L_n/n)^{1/2}$ because $J_i = J$. Use Theorem 1 and solve $\rho_n = r_n$. We obtain that $\rho_n = r_n = n^{-2/5}$, when $L_k = n^{1/5}$. This completes the proof. □