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

Lifeng WANG, Hongzhe LI, and Jianhua Z. HUANG

Nonparametric varying-coefficient models are commonly used for analyzing data measured repeatedly over time, including longitudinal and functional response data. Although many procedures have been developed for estimating varying coefficients, the problem of variable selection for such models has not been addressed to date. In this article we present a regularized estimation procedure for variable selection that combines basis function approximations and the smoothly clipped absolute deviation penalty. The proposed procedure simultaneously selects significant variables with time-varying effects and estimates the nonzero smooth coefficient functions. Under suitable conditions, we establish the theoretical properties of our procedure, including consistency in variable selection and the oracle property in estimation. Here the oracle property means that the asymptotic distribution of an estimated coefficient function is the same as that when it is known a priori which variables are in the model. The method is illustrated with simulations and two real data examples, one for identifying risk factors in the study of AIDS and one using microarray time-course gene expression data to identify the transcription factors related to the yeast cell-cycle process.

KEY WORDS: Functional response; Longitudinal data; Nonparametric function estimation; Oracle property; Regularized estimation.

1. INTRODUCTION

Varying-coefficient (VC) 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, so that the measurements within each subject are possibly correlated with one another (Diggle, Liang, and Zeger 1994; Rice 2004). Another setting is that of functional response models (Rice 2004), where the ith response is a smooth real function $y_i(t), i = 1, \ldots, n, t \in \mathcal{T} = [0, T]$, although in practice only $y_i(t_{ij})$, $j = 1, \ldots, J_i$ are observed. For both settings, the response $Y(t)$ is a random process, and the predictor $\mathbf{X}(t) = (X^{(1)}(t), \ldots, X^{(p)}(t))^T$ is a p-dimensional random process. In applications, observations for n randomly selected subjects are obtained as $(y_i(t_{ij}), \mathbf{x}_i(t_{ij}))$ for the ith subject at discrete time point t_{ij}, $i = 1, \ldots, n$ and $j = 1, \ldots, J_i$. The linear VC model can be written as

$$y_i(t_{ij}) = \mathbf{x}_i(t_{ij})^T \boldsymbol{\beta}(t_{ij}) + \varepsilon_i(t_{ij}), \tag{1}$$

where $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$ is a p-dimensional vector of smooth functions of t and $\varepsilon_i(t), i = 1, \ldots, n$, are iid random processes, independent of $\mathbf{x}_i(t)$. This model has the simplicity of linear models but also has the flexibility of allowing time-varying covariate effects. Since its introduction in the longitudinal data setting by Hoover, Rice, Wu, and Yang (1998), many methods for estimation and inference of model (1) have been developed (see, e.g., Wu and Chiang 2000 and Fan and Zhang 2000 for the local polynomial kernel method; Huang, Wu, and Zhou 2002, 2004 and Qu and Li 2006 for basis expansion and the spline methods; and Chiang, Rice, and Wu

2001 for the smoothing spline method). One important problem not well studied in the literature is how to select significant variables in model (1). The goal of this article is to estimate $\boldsymbol{\beta}(t)$ nonparametrically and select relevant predictors $x_k(t)$ with nonzero functional coefficient $\beta_k(t)$, based on observations $\{(y_i(t_{ij}), \mathbf{x}_i(t_{ij})), i = 1, \ldots, n, j = 1, \ldots, J_i\}$.

Traditional methods for variable selection include hypothesis testing and using information criteria, such as the Akaike information criterion and the Bayes information criterion. Recently, regularized estimation has received much attention as a way to perform variable selection for parametric regression models (see Bickel and Li 2006 for a review). Existing regularization procedures for variable selection include LASSO (Tibshirani 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li 2001), and their extensions (Yuan and Lin 2006; Zou and Hastie 2005; Zou 2006). But these regularized estimation procedures were developed for regression models, where parameters are Euclidean and cannot be applied directly to the nonparametric VC models where parameters are nonparametric smooth functions. Moreover, these procedures are studied mainly for independent data; an exception is the study of Fan and Li (2004) in which the SCAD penalty was used for variable selection in longitudinal data analysis. On the other hand, some recent results on regularization methods with diverging numbers of parameters (e.g., Fan and Peng 2004; Huang, Horowitz, and Ma 2008) are related to nonparametric settings, but the model error due to function approximation was not of concern in those studies.

Regularized estimation for selecting relevant variables in nonparametric settings has been developed in the general context of smoothing spline analysis of variance (ANOVA) models. Lin and Zhang (2006) developed COSSO for component selection and smoothing in smoothing spline ANOVA. Zhang and Lin (2006) extended the COSSO method to nonparametric regression in exponential families, Zhang (2006) extended it to support vector machines, and Leng and Zhang (2007) extended it to hazard regression. Lin and Zhang (2006) studied

rates of convergence of COSSO estimators; the other authors focused on computational algorithms and did not provide theoretical results. Asymptotic distribution results have not been developed for COSSO and its extensions. Moreover, these approaches have focused on independent data, not on longitudinal data, as we studied in the present work.

In this article we use the method of basis expansion to estimate smooth functions in VC models because of the simplicity of the basis expansion approach, as illustrated by Huang et al. (2002). We extend the application of the SCAD penalty to a nonparametric setting. Besides developing a computational algorithm, we study the asymptotic property of the estimator. We show that our procedure is consistent in variable selection; that is, the probability that it correctly selects the true model tends to 1. Furthermore, we show that our procedure has an oracle property; that is, the asymptotic distribution of an estimated coefficient function is the same as that when it is known a priori which variables are in the model. Such results are new in nonparametric settings.

The rest of the article is organized as follows. We describe the regularized estimation procedure using basis expansion and the SCAD penalty in Section 2, and present a computational algorithm and a method for selecting the tuning parameter in Section 3. We present theoretical results, including the consistency in variable selection and the oracle property in Section 4, and then some simulation results in Section 5. In Section 6 we illustrate the proposed method using two real data examples, a CD4 data set and a microarray time-course gene expression data set. We gather technical proofs in the Appendix.

2. BASIS FUNCTION EXPANSION AND REGULARIZED ESTIMATION USING THE SCAD PENALTY

Huang et al. (2002) proposed estimating the unknown time-varying coefficient functions using basis expansion. Suppose that the coefficient $\beta_k(t)$ can be approximated by a basis expansion $\beta_k(t) \approx \sum_{l=1}^{L_k} \gamma_{kl} B_{kl}(t)$, where L_k is the number of basis functions in approximating the function $\beta_k(t)$. Model (1) then becomes

$$y_i(t_{ij}) \approx \sum_{k=1}^{p} \sum_{l=1}^{L_k} \gamma_{kl} x_i^{(k)}(t_{ij}) B_{kl}(t_{ij}) + \varepsilon_i(t_{ij}). \quad (2)$$

The parameters γ_{kl} in the basis expansion can be estimated by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} w_i \sum_{j=1}^{J_i} \left\{ y_i(t_{ij}) - \sum_{k=1}^{p} \sum_{l=1}^{L_k} \gamma_{kl} x_i^{(k)}(t_{ij}) B_{kl}(t_{ij}) \right\}^2, \quad (3)$$

where the w_i's are weights taking value 1 if we treat all observations equally or $1/n_i$ if we treat each subject equally. An estimate of $\beta_k(t)$ is obtained by $\hat{\beta}_k(t) = \sum_{l=1}^{L_k} \hat{\gamma}_{kl} B_{kl}(t)$, where the $\hat{\gamma}_{kl}$'s are minimizers of (3). Various basis systems, including Fourier bases, polynomial bases, and B-spline bases, can be used in the basis expansion. Huang et al. (2002) studied consistency and rates of convergence of such estimators for general basis choices, and also studied asymptotic normality of the estimators when the basis functions B_{kl} are splines (2004).

Now suppose that some variables are not relevant in the regression, so that the corresponding coefficient functions are zero functions. We introduce a regularization penalty to (3) so that these zero coefficient functions are estimated as identically 0. Toward that end, it is convenient to rewrite (3) using function space notation. Let $\mathcal{G}_k$ denote all functions that have the form $\sum_{l=1}^{L_k} \gamma_{kl} B_{kl}(t)$ for a given basis system $\{B_{kl}(t)\}$. Then (3) can be written as $(1/n) \sum_i w_i \sum_j \{y_i(t_{ij}) - \sum_k g_k(t_{ij}) x_i^{(k)}(t_{ij})\}^2$, where $g_k(t) = \sum_{l=1}^{L_k} \gamma_{kl} B_{kl}(t) \in \mathcal{G}_k$. Let $p_\lambda(u), u \geq 0$, be a nonnegative penalty function that depends on a penalty parameter λ. We assume that $p_\lambda(0) = 0$ and that $p_\lambda(u)$ is nondecreasing as a function of u. Let $\|g_k\|$ denote the L_2-norm of the function g_k. Over $g_k \in \mathcal{G}_k$, we minimize the penalized criterion

$$\frac{1}{n} \sum_{i=1}^{n} w_i \sum_{j=1}^{J_i} \left\{ y_i(t_{ij}) - \sum_{k=1}^{p} g_k(t_{ij}) x_i^{(k)}(t_{ij}) \right\}^2$$
$$+ \sum_{k=1}^{p} p_\lambda(\|g_k\|). \quad (4)$$

There are two sets of tuning parameters. The L_k's control the smoothness of the coefficient functions, and the λ governs variable selection or sparsity of the model. Selection of these parameters is discussed in Section 4. In our implementation of the method, we use B-splines as the basis functions. Thus $L_k = n_k + d + 1$, where n_k is the number of interior knots for g_k and d is the degree of the spline. The interior knots of the splines can be either equally spaced or placed on the sample quantiles of the data, so that there are about the same number of observations between any two adjacent knots. We use equally spaced knots for all numerical examples in this article.

There are many ways to specify the penalty function $p_\lambda(\cdot)$. If $p_\lambda(u) = \lambda u$, then the penalty term in (4) is similar to that in COSSO. But here we use the SCAD penalty function of Fan and Li (2001), defined as

$$p_\lambda(u) = \begin{cases} \lambda u & \text{if } 0 \leq u \leq \lambda \\ -\dfrac{(u^2 - 2a\lambda u + \lambda^2)}{2(a-1)} & \text{if } \lambda < u < a\lambda \\ \dfrac{(a+1)\lambda^2}{2} & \text{if } u \geq a\lambda. \end{cases} \quad (5)$$

The penalty function (5) is a quadratic spline function with two knots at λ and $a\lambda$, where a is another tuning parameter. Fan and Li (2001) suggested that $a = 3.7$ is a reasonable choice, which we adopt in this article. Using the SCAD penalty allows us to obtain nice theoretical properties, such as consistent variable selection and the oracle property, for the proposed method.

For $g_k(t) = \sum_l \gamma_{kl} B_{kl}(t) \in \mathcal{G}_k$, the squared L_2-norm can be written as a quadratic form in $\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kL_k})^T$. Let $\mathbf{R}_k = (r_{ij})_{L_k \times L_k}$ be a matrix with entries $r_{ij} = \int_{\mathcal{T}} B_{ki}(t) B_{kj}(t) \, dt$. Then $\|g_k\|^2 = \boldsymbol{\gamma}_k^T \mathbf{R}_k \boldsymbol{\gamma}_k \triangleq \|\boldsymbol{\gamma}_k\|_k^2$. Set $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$. The penalized weighted least squares criterion (4) can be written as

$$pl(\boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} w_i \sum_{j=1}^{J_i} \left\{ (y_i(t_{ij}) - \sum_{k=1}^{p} \sum_{l=1}^{L_k} \gamma_{kl} x_i^{(k)}(t_{ij}) B_{kl}(t_{ij}) \right\}^2$$
$$+ \sum_{k=1}^{p} p_\lambda(\|\boldsymbol{\gamma}_k\|_k). \quad (6)$$

To express the criterion function (6) using vectors and matrixes, we define

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

$\mathbf{U}_i(t_{ij}) = (\mathbf{x}_i(t_{ij})^T \mathbf{B}(t_{ij}))^T$, $\mathbf{U}_i = (\mathbf{U}_i(t_{i1}), \dots, \mathbf{U}_i(t_{iJ_i}))^T$, and $\mathbf{U} = (\mathbf{U}_1^T, \dots, \mathbf{U}_n^T)^T$. We also define $\mathbf{y}_i = (y_i(t_{i1}), \dots, y_i(t_{iJ_i}))^T$ and $\mathbf{y} = (\mathbf{y}_1^T, \dots, \mathbf{y}_n^T)^T$. The criterion function (6) can be rewritten as

$$pl(\boldsymbol{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} w_i (\mathbf{y}_i - \mathbf{U}_i \boldsymbol{\gamma})^T (\mathbf{y}_i - \mathbf{U}_i \boldsymbol{\gamma}) + \sum_{k=1}^{p} p_\lambda(\|\boldsymbol{\gamma}_k\|_k)$$

$$= (\mathbf{y} - \mathbf{U}\boldsymbol{\gamma})^T \mathbf{W}(\mathbf{y} - \mathbf{U}\boldsymbol{\gamma}) + \sum_{k=1}^{p} p_\lambda(\|\boldsymbol{\gamma}_k\|_k), \qquad (7)$$

where $\mathbf{W}$ is a diagonal weight matrix with w_i repeated n_i times.

Remark. The penalized weighted least squares criterion here does not take into account the within-subject correlation typically present in the longitudinal data, because the correlation structure is usually unknown a priori. A popular method for dealing with within-subject correlation is to use a working correlation structure as in the generalized estimating equations (GEEs; Liang and Zeger 1986). The concept of GEEs can be easily adapted to our methodolgy. Specifically, the weight matrix $\mathbf{W}$ in (7) can be constructed using a working correlation structure, such as AR or compound symmetry. Actually, the criteria (4) and (6) correspond to a working independence correlation structure. Using a working independence correlation will not affect the consistency of variable selection (Sec. 5). On the other hand, using an appropriate working correlation structure may yield more efficient function estimation if the structure is specified correctly.

3. COMPUTATIONAL ALGORITHM

Because of nondifferentiability of the penalized loss (7), the commonly used gradient-based optimization method is not applicable here. In this section we develop an iterative algorithm using local quadratic approximation of the nonconvex penalty $p_\lambda(\|\boldsymbol{\gamma}_k\|_2)$. Following Fan and Li (2001), in a neighborhood of a given positive $u_0 \in \mathbb{R}^+$,

$$p_\lambda(u) \approx p_\lambda(u_0) + \frac{1}{2}\{p'_\lambda(u_0)/u_0\}(u^2 - u_0^2). \qquad (8)$$

In our algorithm a similar quadratic approximation is used by substituting u with $\|\boldsymbol{\gamma}_k\|_k$ in (8), $k = 1, \dots, p$. Given an initial value of $\boldsymbol{\gamma}_k^0$ with $\|\boldsymbol{\gamma}_k^0\|_k > 0$, $p_\lambda(\|\boldsymbol{\gamma}_k\|_k)$ can be approximated by a quadratic form,

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

Consequently, removing an irrelevant constant, the penalized loss (7) becomes

$$pl(\boldsymbol{\gamma}) = \frac{1}{n}(\mathbf{y} - \mathbf{U}\boldsymbol{\gamma})^T \mathbf{W}(\mathbf{y} - \mathbf{U}\boldsymbol{\gamma}) + \frac{1}{2}\boldsymbol{\gamma}^T \boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0)\boldsymbol{\gamma}, \qquad (9)$$

where $\boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0) = \text{diag}\{(p'_\lambda(\|\boldsymbol{\gamma}_1^0\|_2)/\|\boldsymbol{\gamma}_1^0\|_2)\mathbf{R}_1, \dots, (p'_\lambda(\|\boldsymbol{\gamma}_K^0\|_2)/\|\boldsymbol{\gamma}_K^0\|_2)\mathbf{R}_p\}$, where the $\mathbf{R}_k$'s are as defined before (6). This is a quadratic form with a minimizer satisfying

$$\left\{\mathbf{U}^T \mathbf{W} \mathbf{U} + \frac{n}{2}\boldsymbol{\Omega}_\lambda(\boldsymbol{\gamma}^0)\right\}\boldsymbol{\gamma} = \mathbf{U}^T \mathbf{W} \mathbf{y}. \qquad (10)$$

The foregoing discussion leads to the following algorithm:

Step 1: Initialize $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(1)}$.
Step 2: Given $\boldsymbol{\gamma}^{(m)}$, update $\boldsymbol{\gamma}$ to $\boldsymbol{\gamma}^{(m+1)}$ by solving (10), where $\boldsymbol{\gamma}^0$ is set to be $\boldsymbol{\gamma}^{(m)}$.
Step 3: Iterate step 2 until convergence of $\boldsymbol{\gamma}$ is achieved.

The initial estimation of $\boldsymbol{\gamma}$ in step 1 can be obtained using a ridge regression, which substitutes $p_\lambda(\|\boldsymbol{\gamma}_k\|_2)$ in the penalized loss (7) with the quadratic function $\|\boldsymbol{\gamma}_k\|_2^2$. The ridge regression has a closed-form solution. At any iteration of step 2, if some $\|\boldsymbol{\gamma}_k^{(m)}\|_2$ is smaller than a cutoff value $\epsilon > 0$, then we set $\hat{\boldsymbol{\gamma}}_k = \mathbf{0}$ and treat $x^{(k)}(t)$ as irrelevant. In our implementation of the algorithm, ϵ is set to 10^{-3}.

Remark. Due to the nonconvexity of $pl(\boldsymbol{\gamma})$, the algorithm sometimes may fail to converge, and the search falls into an infinite loop surrounding several local minimizers. In this situation, we choose the local minimizer that gives the smallest value of $pl(\boldsymbol{\gamma})$.

4. SELECTION OF TUNING PARAMETERS

To implement the proposed method, we need to choose the tuning parameters, L_k, $k = 1, \dots, p$, and λ, where L_k controls the smoothness of $\hat{\beta}_k(t)$ and λ determines the variable selection. In Section 5 we show that our method is consistent in variable selection when these tuning parameters grow or decay at a proper rate with n. In practice, however, we need a data-driven procedure to select the tuning parameters. We propose using "leave-one-observation-out" cross-validation (CV) when the errors are independent and using "leave-one-subject-out" CV for correlated errors.

In this section we develop some computational shortcuts using suitable approximations, to prevent actual computation of leave-one-out estimates. Although the derivation of model selection criteria generally holds, we focus our presentation on the situation when $L_k = L$ for all $\beta_k(t)$, $k = 1, \dots, p$. To simplify our presentation, we also set $\mathbf{W} = \mathbf{I}$; extension to general $\mathbf{W}$ is straightforward by simply replacing $\mathbf{y}$ with $\mathbf{W}^{1/2}\mathbf{y}$ and $\mathbf{U}$ with $\mathbf{W}^{1/2}\mathbf{U}$ in the derivation that follows. At the convergence of our algorithm, the nonzero components are estimated as $\hat{\boldsymbol{\gamma}} = \{\mathbf{U}^T\mathbf{U} + (n/2)\boldsymbol{\Omega}_\lambda(\hat{\boldsymbol{\gamma}})\}^{-1}\mathbf{U}^T\mathbf{y}$, which can be considered the solution of the ridge regression

$$\|\mathbf{y} - \mathbf{U}\boldsymbol{\gamma}\|_2^2 + \frac{n}{2}\boldsymbol{\gamma}^T \boldsymbol{\Omega}_\lambda(\hat{\boldsymbol{\gamma}})\boldsymbol{\gamma}. \qquad (11)$$

The hat matrix of this ridge regression is denoted by $\mathbf{H}(L, \lambda) = \mathbf{U}\{\mathbf{U}^T\mathbf{U} + (n/2)\boldsymbol{\Omega}_\lambda(\hat{\boldsymbol{\gamma}})\}^{-1}\mathbf{U}^T$. The fitted $\mathbf{y}$ from this ridge regression satisfies $\hat{\mathbf{y}} = \mathbf{H}(L, \lambda)\mathbf{y}$.

When the errors, ε_{ij}, are independent for different t_{ij}, $j = 1, \dots, J_i$, leave-one-observation-out CV is a reasonable choice for selecting tuning parameters. For the ridge regression (11), computation of leave-one-out estimates for CV can be avoided

(Hastie and Tibshirani 1993). The CV criterion has a closed-form expression,

$$CV_1(L, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \frac{\{(y_i(t_{ij}) - \mathbf{U}_i^T(t_{ij})\hat{\boldsymbol{\gamma}}\}^2}{[1 - \{\mathbf{H}(L, J)\}_{(ij),(ij)}]^2}, \quad (12)$$

where $\{\mathbf{H}(L, \lambda)\}_{(ij),(ij)}$ is the diagonal element of $H(L, \lambda)$ corresponding to the observation at time t_{ij}. The corresponding generalized CV (GCV) criterion, which replaces the diagonal elements in the foregoing CV formula by their average, is

$$GCV_1(L, \lambda) = \frac{1}{n} \frac{\|\mathbf{y} - \mathbf{H}(L, \lambda)\mathbf{y}\|_2^2}{[1 - \text{tr}\{\mathbf{H}(L, \lambda)\}/n]^2}. \quad (13)$$

When the errors ε_{ij} are not independent and the correlation structure of $\varepsilon_i(t)$ is unknown, leave-one-observation-out CV is unsuitable, and leave-one-subject-out CV is needed (Rice and Silverman 1991; Hoover et al. 1998; Huang et al. 2002). Let $\hat{\boldsymbol{\gamma}}^{(-i)}$ be the solution of (6) after the ith subject is deleted. The leave-one-subject-out CV can be written as

$$CV_2(L, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \{y_i(t_{ij}) - \mathbf{U}_i^T(t_{ij})\hat{\boldsymbol{\gamma}}^{(-i)}\}^2. \quad (14)$$

The $CV_2(L, \lambda)$ can be viewed as an estimate of the true prediction error. But computation of CV_2 is intensive, because it requires minimization of (6) n times. To overcome this difficulty, we propose to minimize a delete-one version of (11) instead of (6) when computing CV_2, where $\hat{\boldsymbol{\gamma}}$ in $\boldsymbol{\Omega}_\lambda(\hat{\boldsymbol{\gamma}})$ of (11) is the estimate based on the whole data set. The approximated CV (ACV) error is

$$ACV_2(L, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J_i} \{y_i(t_{ij}) - \mathbf{U}_i^T(t_{ij})\hat{\boldsymbol{\gamma}}^{\star(-i)}\}^2$$

$$= \sum_{i=1}^{n} \|\mathbf{y}_i - \mathbf{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)}\|_2^2, \quad (15)$$

where $\hat{\boldsymbol{\gamma}}^{\star(-i)}$ is obtained by solving (11) instead of (6), deleting the ith subject. It turns out that a computational shortcut is available for (15), as we discuss next. This computational shortcut relies on the following "leave-one-subject-out" formula, the proof of which is given in Appendix A.

Lemma 1. Define $\tilde{\mathbf{y}}^{(i)} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_{i-1}^T, \mathbf{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)}, \mathbf{y}_{i+1}^T, \ldots, \mathbf{y}_n^T)$, and let $\tilde{\boldsymbol{\gamma}}^{(i)}$ be the minimizer of (11) with $\mathbf{y}$ substituted by $\tilde{\mathbf{y}}^{(i)}$. Then $\mathbf{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)} = \mathbf{U}_i\tilde{\boldsymbol{\gamma}}^{(i)}$.

Let $\mathbf{A} = \{\mathbf{U}^T\mathbf{U} + (n/2)\boldsymbol{\Omega}_\lambda(\hat{\boldsymbol{\gamma}})\}^{-1}\mathbf{U}^T$. Partition $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_n)$ into blocks of columns, with each block corresponding to one subject. Then

$$\hat{\boldsymbol{\gamma}} = \mathbf{A}\mathbf{y} = \sum_{i=1}^{n} \mathbf{A}_i\mathbf{y}_i$$

and

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

$$= \hat{\boldsymbol{\gamma}} - \mathbf{A}_i\mathbf{y}_i + \mathbf{A}_i\mathbf{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)}.$$

As a consequence of Lemma 1,

$$\mathbf{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)} = \mathbf{U}_i\tilde{\boldsymbol{\gamma}}^{(i)} = \mathbf{U}_i\hat{\boldsymbol{\gamma}} - \mathbf{U}_i\mathbf{A}_i(\mathbf{y}_i - \mathbf{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)}).$$

It follows that

$$\mathbf{y}_i - \mathbf{U}_i\hat{\boldsymbol{\gamma}}^{\star(-i)} = (\mathbf{I} - \mathbf{U}_i\mathbf{A}_i)^{-1}(\mathbf{y}_i - \mathbf{U}_i\hat{\boldsymbol{\gamma}})$$

$$= \{\mathbf{I} - \mathbf{H}_{ii}(L, \lambda)\}^{-1}(\mathbf{y}_i - \mathbf{U}_i\hat{\boldsymbol{\gamma}}).$$

Therefore,

$$ACV_2(L, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \|\{\mathbf{I} - \mathbf{H}_{ii}(L, \lambda)\}^{-1}(\mathbf{y}_i - \mathbf{U}_i\hat{\boldsymbol{\gamma}})\|_2^2, \quad (16)$$

where $\mathbf{H}_{ii}(L, \lambda)$ is the (i, i)th diagonal block corresponding to observations for the ith subject. Because J_i usually is not a big number, inversion of the J_i-dimensional matrixes $\mathbf{I} - \mathbf{H}_{ii}$ will not create a computational burden in the evaluation of (16). Using (16) as an approximation of the leave-one-subject-out CV helps us avoid actual computation of many delete-one estimates. The quality of the approximation is illustrated in Figure 1 using a data set from the simulation study reported in Section 6.

5. LARGE–SAMPLE PROPERTIES

Fan and Li (2001) established the large-sample properties of the SCAD penalized estimates for parametric regression models. They showed that the SCAD penalty enables consistent variable selection and has an oracle property for parameter estimation; the estimates of the nonzero regression coefficients behave as if the subset of relevant variables were already known. We show that similar large-sample properties hold for our proposed method for the VC models. Note that some care must be taken in developing an asymptotic theory for a nonparametric setting; for example, the rate of convergence of a nonparametric estimate is not root-n. The approximation error due to the use of basis expansion also must be studied carefully. For simplicity, we present our asymptotic results only for the working independence case. Technical proofs of all asymptotic results are given in the Appendix.

We focus our asymptotic analysis for $\mathcal{G}_k$ being a space of spline functions. Extension to general basis expansions can be obtained by combining the technical arguments in this article with those of Huang et al. (2002). Define $L_n = \max_{1 \leq k \leq p} L_k$ and $\rho_n = \max_{1 \leq k \leq p} \inf_{g \in \mathcal{G}_k} \|\beta_k - g\|_{L_\infty}$. Thus ρ_n characterizes the approximation error due to spline approximation. Assume that only s predictors are relevant in model (1). Without loss of generality, let $\boldsymbol{\beta}_k(t)$, $k = 1, \ldots, s$, be the nonzero coefficient functions and let $\boldsymbol{\beta}_k(t) \equiv 0$, $k = s + 1, \ldots, p$.

We make the following technical assumptions:

(C1) The response and covariate processes, $(y_i(t), \mathbf{x}_i(t))$, $i = 1, \ldots, n$, are iid as $(y(t), \mathbf{X}(t))$, and the observation time points, t_{ij}, are iid from an unknown density, $f(t)$, on $[0, T]$, where $f(t)$ are uniformly bounded away from infinity and 0.

(C2) The eigenvalues of the matrix $E\{\mathbf{X}(t)\mathbf{X}^T(t)\}$ are uniformly bound away from 0 and infinity for all t.

(C3) There exists a positive constant M_1 such that $|X_k(t)| \leq M_1$ for all t and $1 \leq k \leq p$.

(C4) There exists a positive constant M_2 such that $E\{\varepsilon^2(t)\} \leq M_2$ for all t.
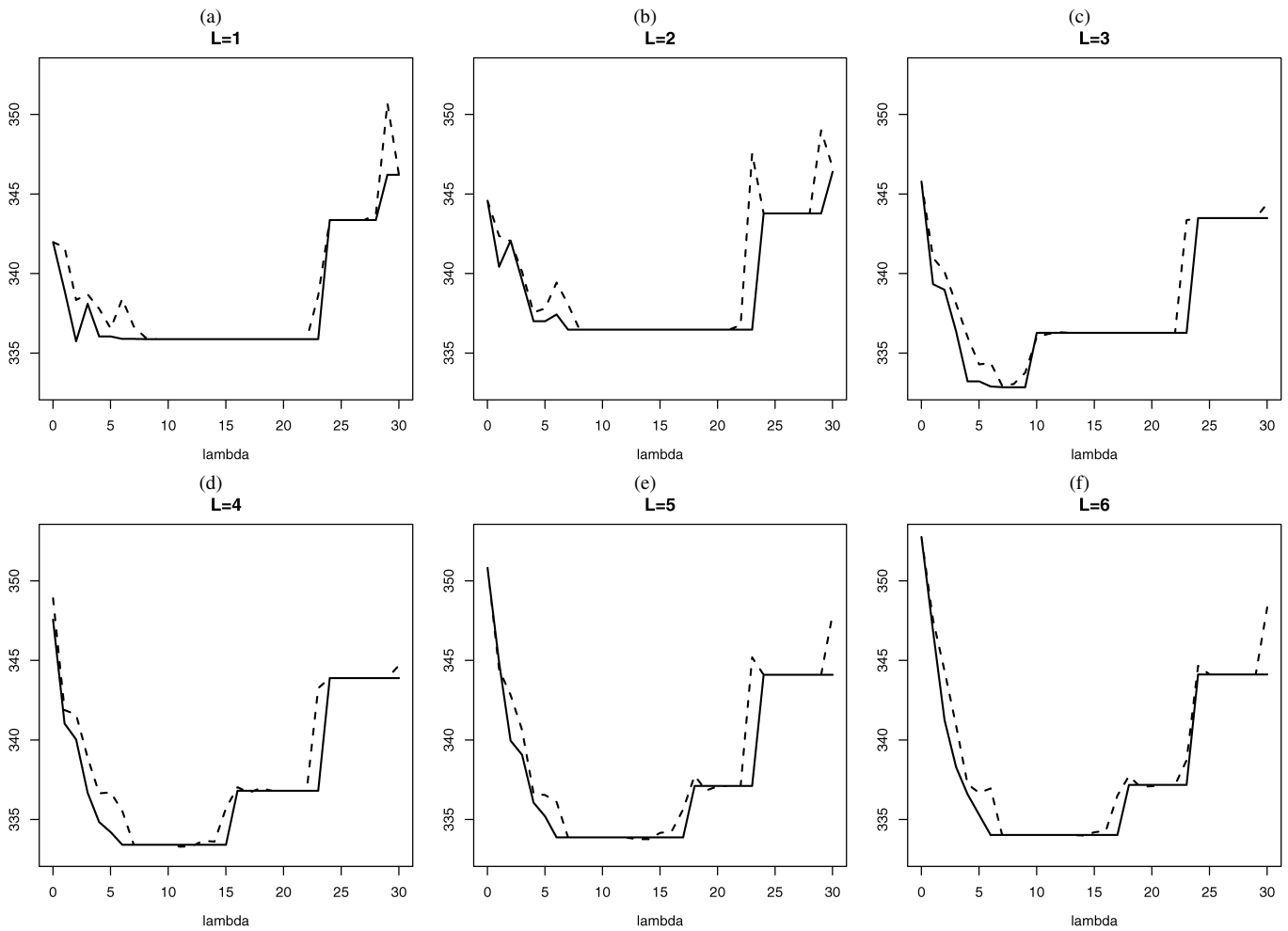
Figure 1. The approximation of *CV* scores for one simulated data set for different values of *L* and λ. The *ACV* given in (16) (——) closely approximates the leave-one-subject-out *CV* in (14) (- -).

(C5) $\limsup_n (\max_k L_k / \min_k L_k) < \infty$.

(C6) The process $\epsilon(t)$ can be decomposed as the sum of two independent stochastic processes, that is, $\epsilon(t) = \epsilon^{(1)}(t) + \epsilon^{(2)}(t)$, where $\epsilon^{(1)}$ is an arbitrary mean 0 process and $\epsilon^{(2)}$ is a process of measurement errors that are independent at different time points and have mean 0 and constant variance σ^2.

The same set of assumptions has been used by Huang et al. (2004). Our results can be extended to cases with deterministic observation times; the assumption on independent observation times in (C1) also can be relaxed (see remarks 3.1 and 3.2 of Huang et al. 2004).

Define $r_n = n^{-1}[\sum_{i=1}^n w_i^{-2}\{L_n J_i + J_i(J_i - 1)\}]^{1/2}$. When $w_i = 1/J_i$ or $w_i = 1/J_i$ with J_i bounded uniformly, we can show that $r_n \asymp (L_n/n)^{1/2}$.

Theorem 1. Suppose that conditions (C1)–(C5) hold, $\lim_{n\to\infty} \rho_n = 0$, and

$$\lim_{n\to\infty} n^{-1} L_n \log(L_n) = 0. \qquad (17)$$

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 the following:

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

b. $\|\hat{\beta}_k - \beta_k\|_{L_2} = O_p(\max(r_n, \rho_n))$, $k = 1, \ldots, s$.

Part a of Theorem 1 says that the proposed penalized least squares method is consistent in variable selection; that is, it can identify the zero coefficient functions with probability tending to 1. Part b provides the rate of convergence in estimating the nonzero coefficient functions.

Corollary 1. Suppose that conditions (C1)–(C4) hold and that $\beta_k(t)$ have bounded second derivatives, $k = 1, \ldots, s$, and $\beta_k(t) = 0$, $k = s + 1, \ldots, p$. Let $\mathcal{G}_k$ be a space of splines with degree no less than 1 and with L_k equally spaced interior knots, where $L_k \asymp n^{1/5}$, $k = 1, \ldots, p$. If $\lambda_n \to 0$ and $n^{2/5}\lambda_n \to \infty$, then the following hold:

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

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

Note that the rate of convergence given in part b of Corollary 1 is the optimal rate for nonparametric regression with iid data under the same smoothness assumption on the unknown function (Stone 1982).

Now we consider the asymptotic variance of proposed estimate. Let $\boldsymbol{\beta}^{(1)} = (\beta_1, \ldots, \beta_s)^T$ denote the vector of nonzero coefficient functions, and let $\hat{\boldsymbol{\beta}}^{(1)} = (\hat{\beta}_1, \ldots, \hat{\beta}_s)^T$ denote its estimate obtained by minimizing (6). Let $\hat{\boldsymbol{\gamma}}^{(1)} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_s^T)^T$,

let $\mathbf{U}^{(1)}$ denote the selected columns of $\mathbf{U}$ corresponding to $\boldsymbol{\beta}^{(1)}$, and, similarly, let $\boldsymbol{\Omega}_\lambda^{(1)}$ denote the selected diagonal blocks of $\boldsymbol{\Omega}_\lambda$. By part a of Theorem 1, with probability tending to 1, $\hat{\gamma}_k$, $k = s + 1, \ldots, p$, are vectors of 0's. Thus the quadratic approximation (9) yields

$$\hat{\boldsymbol{\gamma}}^{(1)} = \left\{ (\mathbf{U}^{(1)})^T \mathbf{W} \mathbf{U}^{(1)} + \frac{n}{2} \boldsymbol{\Omega}_{\lambda_n}^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)}) \right\}^{-1} (\mathbf{U}^{(1)})^T \mathbf{W} \mathbf{y}. \quad (18)$$

Let $\mathbf{U}_i^{(1)}$ be the rows of $\mathbf{U}^{(1)}$ corresponding to subject i, and let

$$\mathbf{H}^{(1)} = (\mathbf{U}^{(1)})^T \mathbf{W} \mathbf{U}^{(1)} + \frac{n}{2} \boldsymbol{\Omega}_\lambda^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)})$$

$$= \sum_{i=1}^n (\mathbf{U}_i^{(1)})^T \mathbf{W}_i \mathbf{U}_i^{(1)} + \frac{n}{2} \boldsymbol{\Omega}_\lambda^{(1)}(\hat{\boldsymbol{\gamma}}^{(1)}). \quad (19)$$

The asymptotic variance of $\hat{\boldsymbol{\gamma}}^{(1)}$ is

$$\operatorname{avar}(\hat{\boldsymbol{\gamma}}^{(1)})$$

$$= (\mathbf{H}^{(1)})^{-1} (\mathbf{U}^{(1)})^T \mathbf{W} \operatorname{cov}(\mathbf{y}) \mathbf{W} \mathbf{U}^{(1)} (\mathbf{H}^{(1)})^{-1}$$

$$= (\mathbf{H}^{(1)})^{-1} \sum_{i=1}^n (\mathbf{U}_i^{(1)})^T \mathbf{W}_i \operatorname{cov}(\mathbf{y}_i) \mathbf{W}_i \mathbf{U}_i^{(1)} (\mathbf{H}^{(1)})^{-1}, \quad (20)$$

the diagonal blocks of which give $\operatorname{avar}(\hat{\boldsymbol{\gamma}}_k)$, $k = 1, \ldots, s$. Because $\hat{\boldsymbol{\beta}}^{(1)}(t) = (\mathbf{B}^{(1)})^T(t)\hat{\boldsymbol{\gamma}}^{(1)}$, where $\mathbf{B}^{(1)}(t)$ are the first s rows of $\mathbf{B}(t)$, we have that $\operatorname{avar}\{\hat{\boldsymbol{\beta}}^{(1)}(t)\} = (\mathbf{B}^{(1)})^T(t) \times \operatorname{avar}(\hat{\boldsymbol{\gamma}}^{(1)}) \mathbf{B}^{(1)}(t)$, the diagonal elements of which are $\operatorname{avar}\{\hat{\beta}_k(t)\}$, $k = 1, \ldots, s$.

Let $\tilde{\beta}_k(t) = E\{\hat{\beta}_k(t)\}$ be the conditional mean given all observed X's. Let $\tilde{\boldsymbol{\beta}}^{(1)}(t) = (\tilde{\beta}_1(t), \ldots, \tilde{\beta}_s(t))$. For positive definite matrixes A, let $A^{1/2}$ denote the unique square root of A, and let $A^{-1/2} = (A^{-1})^{1/2}$. Let $\operatorname{avar}^*\{\hat{\boldsymbol{\beta}}^{(1)}(t)\}$ denote a modification of $\operatorname{avar}\{\hat{\boldsymbol{\beta}}^{(1)}(t)\}$ by replacing $\boldsymbol{\Omega}_{\lambda_n}$ in (19) with 0.

Theorem 2. Suppose that conditions (C1)–(C6) hold, $\lim_{n \to \infty} \rho_n = 0$, $\lim_{n \to \infty} L_n \max_i J_i / n = 0$, and (17) holds. Then, as $n \to \infty$, $\{\operatorname{avar}^*(\hat{\boldsymbol{\beta}}^{(1)}(t))\}^{-1/2} \{\hat{\boldsymbol{\beta}}^{(1)}(t) - \tilde{\boldsymbol{\beta}}^{(1)}(t)\} \to$ N$(0, I)$ in distribution, and in particular, $\{\operatorname{avar}^*(\hat{\beta}_k(t))\}^{-1/2} \times \{\hat{\beta}_k(t) - \tilde{\beta}_k(t)\} \to$ N$(0, 1)$, $k = 1, \ldots, s$.

Note that $\operatorname{avar}^*\{\hat{\boldsymbol{\beta}}^{(1)}(t)\}$ is exactly the same asymptotic variance of the nonpenalized weighted least squares estimate using only those covariates corresponding to nonzero coefficient functions (see thm. 3 of Huang et al. 2004). Theorem 2 implies that our penalized least squares estimate has the so-called "oracle property"—the joint asymptotic distribution of estimates of nonzero coefficient functions are the same as that of the nonpenalized least squares estimate that uses the information about zero coefficient functions.

Theorem 2 can be used to construct pointwise asymptotic confidence intervals for $\tilde{\beta}_k(t)$, $k = 1, \ldots, s$. Theorem 4 and corollary 1 of Huang et al. (2004) discuss how to bound the biases $\tilde{\beta}_k(t) - \beta_k(t)$ and under what conditions the biases are negligible relative to variance. Those results are applicable to the current context. When constructing the confidence intervals, to improve finite-sample performance, we still use (19) and the sandwich formula (20) to calculate the variance. In actual calculation, we need to replace $\operatorname{cov}(\mathbf{y}_i)$ in (19) by its estimate $\mathbf{e}_i \mathbf{e}_i^T$, where $\mathbf{e}_i$ is the residual vector for subject i.

6. 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 by Huang et al. (2002), which assumes that

$$Y(t_{ij}) = \beta_0(t_{ij}) + \sum_{k=1}^{23} \beta_k(t_{ij}) x_k(t_{ij}) + \tau \varepsilon(t_{ij}),$$

$$i = 1, \ldots, 200, \ j = 1, \ldots, J_i.$$

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 0 and variance $(1 + x_1(t))/(2 + x_1(t))$; and $x_3(t)$, independent of x_1 and x_2, is a Bernoulli random variable with success rate .6. In addition to x_k, $k = 1, 2, 3, 20$, redundant variables $x_k(t)$, $k = 4, \ldots, 23$, were simulated to demonstrate the performance of variable selection, where each $x_k(t)$, independent of the others, is a random realization of a Gaussian process with covariance structure $\operatorname{cov}(x_k(t), x_k(s)) = 4 \exp(-|t - s|)$. 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 the $E(t)$'s are independent measurement errors from the N$(0, 4)$ distribution at each time t. The parameter τ in the model was used to control the model's signal-to-noise ratio. We considered models with $\tau = 1.00, 4.00, 5.66$, and 8.00. The coefficients $\beta_k(t)$, $k = 0, \ldots, 3$, corresponding to the constant term and the first three variables, are given by

$$\beta_0(t) = 15 + 20 \sin\left(\frac{\pi t}{60}\right),$$

$$\beta_1(t) = 2 - 3 \cos\left(\frac{\pi(t - 25)}{15}\right),$$

$$\beta_2(t) = 6 - .2t,$$

and

$$\beta_3(t) = -4 + \frac{(20 - t)^3}{2,000}$$

(Fig. 2, solid lines) and 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}, were generated following the same scheme as that of Huang et al. (2002), with each subject having a set of "scheduled" time points $\{1, \ldots, 30\}$ and each scheduled time having a probability of 60% of being skipped. Then the actual observation time t_{ij} is obtained by adding a random perturbation from Uniform$(-.5, .5)$ to the nonskipped scheduled time.

For each simulated data set, we minimized the penalized least squares criterion (4) over a space of cubic splines with equally spaced knots. We selected the number of knots and the tuning parameter λ by minimizing the ACV_2 criterion in (15) as described in Section 4 (see Fig. 1 for an example). We repeated the simulations 500 times. Table 1 gives the results of variable selection of our proposed procedure for four different models with different levels of noise variance. Clearly, the noise level had a significant effect on how well the proposed method selected the exact correct models. When the noise level was low
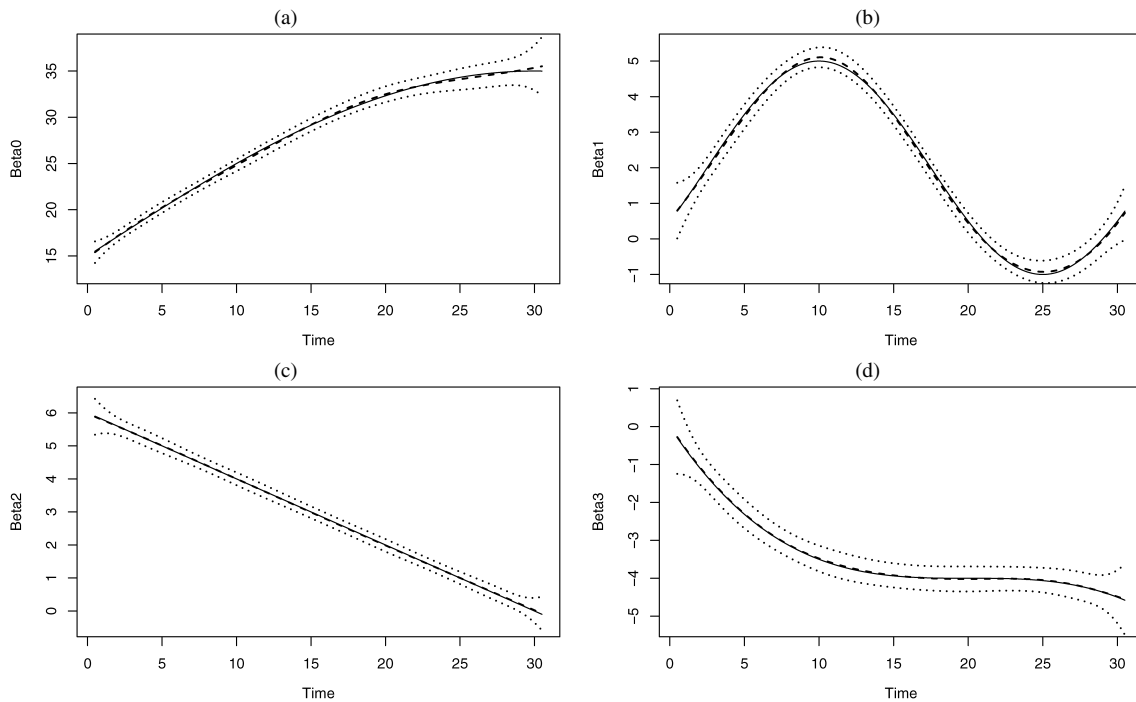
Figure 2. Simulation study for $\tau = 1.00$. True (——) and average of estimated (- -) time-varying coefficients (a) $\beta_0(t)$, (b) $\beta_1(t)$, (c) $\beta_2(t)$, and (d) $\beta_3(t)$ (± 1 pointwise standard deviation) over 500 replications.

($\tau = 1.00$), the method selected the exact correct model in 79% of the replications. In addition, out of these 500 replications, variables 1–3 were selected in each of the runs. In contrast, the number of times that each of the 20 irrelevant variables was selected among the 500 simulation runs ranged from 3 to 12, with an average of 7.3 times (i.e., 1.46%). But when the noise level was high ($\tau = 8.00$), the method selected the exact true model in only 26% of the replications, and the average number of times that the three relevant variables were selected was reduced to 2.24. Interestingly, however, the average num-

Table 1. Simulation results for models with different noise levels (τ) based on 500 replications

	Noise level (τ)			
	1.00	4.00	5.66	8.00
Variable selection				
Perc.T	.79	.68	.48	.26
Aver.R	3.00	2.94	2.73	2.24
Aver.I	.29	.47	.55	.54
Aver.Bias^2				
$\beta_0(t)$	.021 (.023)	.39 (.37)	1.60 (.42)	4.55 (.42)
$\beta_1(t)$	.003 (.004)	.067 (.067)	.29 (.076)	.97 (.073)
$\beta_2(t)$	.000089 (.0002)	.0022 (.0022)	.0047 (.0043)	.026 (.0089)
$\beta_3(t)$	.00023 (.0004)	.0079 (.0060)	.061 (.010)	.88 (.02)
Aver.Var.				
$\beta_0(t)$	.98 (.95)	14.11 (12.53)	28.29 (23.98)	49.09 (47.25)
$\beta_1(t)$	.12 (.11)	1.95 (1.52)	4.08 (2.84)	7.15 (5.52)
$\beta_2(t)$	.049 (.047)	.70 (.67)	1.37 (1.31)	3.06 (2.63)
$\beta_3(t)$	.15 (.15)	2.13 (2.01)	4.70 (3.96)	9.98 (7.88)
95%Cov.Prob.				
$\beta_0(t)$	.929	.935	.937	.926
$\beta_1(t)$	.924	.929	.930	.929
$\beta_2(t)$	.936	.940	.946	.942
$\beta_3(t)$	.944	.949	.947	.949

NOTE: Perc.T represents percentage of replications that the exact true model was selected; Aver.R, average of the numbers of the three relevant variables selected in the model; Aver.I, average of the numbers of the irrelevant variables selected in the model; Aver.Bias^2, average of the squared biases; Aver.Var., average of the empirical variances; and 95%Cov.Prob., 95% coverage probability based on the replications when the true models were selected. For Aver.Bias^2, Aver.Var., and 95%Cov.Prob., all values were averaged over a grid of 300 points and 500 replications. Numbers in parentheses are results of the oracle estimators that assume the true models are known.

ber of times that the irrelevant variables were selected remained low (see the column "Ave.I" in Table 1). The simulations indicate that the proposed procedure indeed provides an effective method for selecting variables with time-varying coefficients.

To examine how well the method estimates the coefficient functions, Figure 2 shows the estimates of the time-varying coefficients of $\beta_k(t)$, $k = 0, 1, 2, 3$, for the model with $\tau = 1.00$, which indicate that the estimates fit the true function very well. Table 1 gives the averages of the squared biases and the empirical variances of the estimates of $\beta_k(t)$, $k = 0, 1, 2, 3$, averaged over a grid of 300 points. In general, the biases were small, except for the estimates of $\beta_0(t)$ when the noise level was high, in which case the proposed method did not select the variables well. Similarly, the variances of the estimates of $\beta_0(t)$ also were relatively larger, due to the fact that the magnitude of true $\beta_0(t)$ was larger than that of the other three functions. For comparison, we also calculated the biases and the empirical variances of the estimates based on the true models (i.e., the oracle estimators; see Table 1). Again, we observed that the both biases and empirical variances were very comparable to the oracle estimators when the noise level was low. When the noise level was high, the biases and the empirical variances became large, because the method tended to select fewer relevant variables.

Finally, we examined the coverage probabilities of the theoretical pointwise confidence intervals based on Theorem 2. For each model, the empirical coverage probabilities averaged over a grid of 300 points are presented in Table 1 for $\beta_0(t)$, $\beta_1(t)$, $\beta_2(t)$, and $\beta_3(t)$, for nominal 95% pointwise confidence intervals. The slight undercoverage is caused by bias; the GCV criterion attempted to balance the bias and variance. When we used a number of knots larger than that suggested by GCV (undersmoothing), the empirical coverage probabilities became closer to the nominal level; for example, for the model with $\tau = 1$, the average empirical coverage probabilities became .946, .949, .942, and .947 when the number of interior knots was fixed at 5.

7. APPLICATIONS TO REAL DATA SETS

To demonstrate the effectiveness of the proposed methods in selecting the variables with time-varying effects and in estimating these time-varying effects, in this section we present results from the analysis of two real data sets, a longitudinal AIDS data set (Kaslow et al. 1987) and a repeated-measurements microarray time-course gene expression data set (Spellman et al. 1998).

7.1 Application to AIDS Data

Kaslow et al. (1987) reported a Multicenter AIDS Cohort Study conducted obtain repeated measurements of physical examinations, laboratory results, and CD4 cell counts and percentages of homosexual men who became human immunodeficiency virus (HIV)-positive during 1984 and 1991. All individuals were scheduled to undergo measurements at semiannual visits, but because many individuals missed some of their scheduled visits and the HIV infections occured randomly during the study, there were unequal numbers of repeated measurements and different measurement times for each individual. As a subset of the cohort, our analysis focused on the 283 homosexual men who become HIV-positive and aimed to evaluate the effects of cigarette smoking, pre-HIV infection CD4

cell percentage, and age at HIV infection on the mean CD4 percentage after infection. In this subset, the number of repeated measurements per subject ranged from 1 to 14, with a median of 6 and a mean of 6.57. The number of distinct measurement time points was 59. Let t_{ij} be the time in years of the jth measurement of the ith individual after HIV infection, let y_{ij} be the ith individual's CD4 percentage at time t_{ij}, and let x_{i1} be the smoking status of the ith individual, taking a value of 1 or 0 for smoker or nonsmoker. In addition, let x_{i2} be the centered age at HIV infection for the ith individual and let x_{i3} be the centered preinfection CD4 percentage. We assume the following VC model for y_{ij}:

$$y_{ij} = \beta_0(t_{ij}) + \sum_{k=1}^{3} x_{ik}\beta_k(t_{ik}) + \epsilon_{ij},$$

where $\beta_0(t)$ is the baseline CD4 percentage, representing the mean CD4 percentage t years after HIV infection for a nonsmoker with average preinfection CD4 percentage and average age at HIV infection, and $\beta_1(t)$, $\beta_2(t)$, and $\beta_3(t)$ measure the time-varying effects for cigarette smoking, age at HIV infection, and preinfection CD4 percentage, on the postinfection CD4 percentage at time t.

Our procedure identified two nonzero coefficients, $\beta_0(t)$ and $\beta_3(t)$, indicating that cigarette smoking and age at HIV infection have no effect on the postinfection CD4 percentage. Figure 3 shows the fitted coefficient functions (solid curves) and their 95% pointwise confidence intervals, indicating that the baseline CD4 percentage of the population decreases with time and preinfection CD4 percentage appears to be positively associated with high postinfection CD4 percentage. These results agree with those obtained by the local smoothing methods of Wu and Chiang (2000) and Fan and Zhang (2000) and by the basis function approximation method of Huang et al. (2002).

To assess our method in more challenging cases, we added 30 redundant variables, generated by randomly permuting the values of each predictor variable 10 times, and applied our method to the augmented data set. We repeated this procedure 50 times. Of the 3 observed covariates, smoking status was never selected, age was selected twice, and preinfection CD4 percentage was always selected, whereas of the 30 redundant variables, 2 were selected twice, 4 were selected once, and the rest were never selected.

7.2 Application to Yeast Cell-Cycle Gene Expression Data

The cell cycle is a tightly regulated set of processes by which cells grow, replicate their DNA, segregate their chromosomes, and divide into daughter cells. The cell-cycle process is commonly divided into G1–S–G2–M stages, where the G1 stage stands for "GAP 1"; the S stage stands for "synthesis," during which DNA replication occurs; the G2 stage stands for "GAP 2"; and the M stage stands for "mitosis," which is when nuclear (chromosome separation) and cytoplasmic (cytokinesis) division occur. Coordinate expression of genes whose products contribute to stage-specific functions is a key feature of the cell cycle (Simon et al. 2001). Transcription factors (TFs) have been identified that play roles in regulating transcription of a small set of yeast cell-cycle–regulated genes; these include
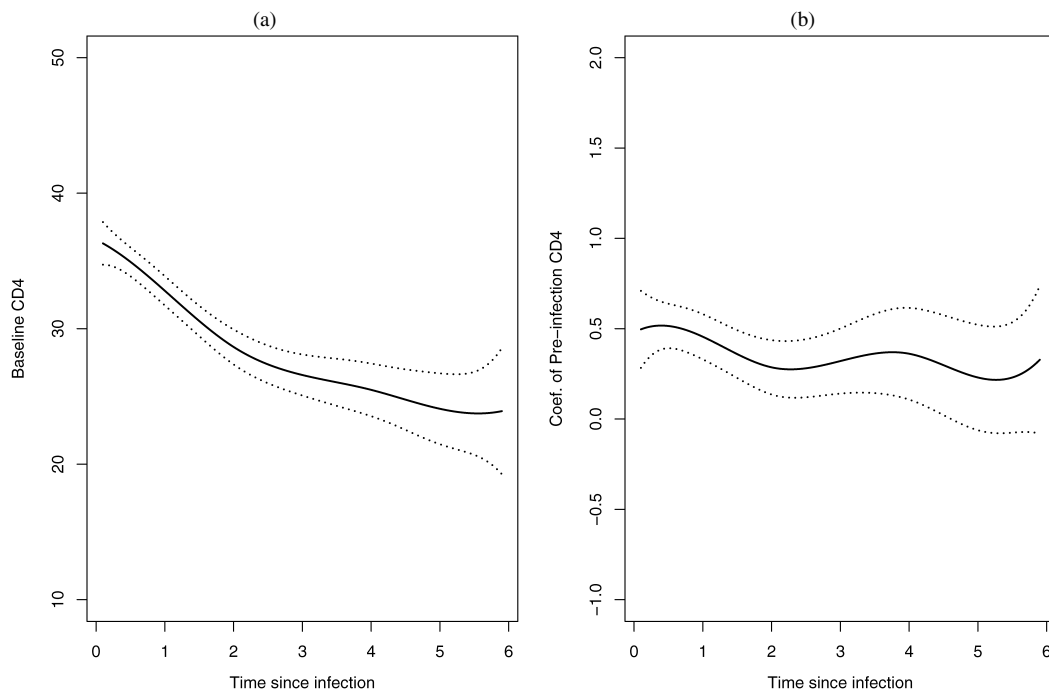
Figure 3. Application to AIDS data. Estimated coefficient functions for intercept (a) and preinfection CD4 percentages (b) (——) and their pointwise 95% confidence intervals (- -).

Mbp1, Swi4, Swi6, Mcm1, Fkh1, Fkh2, Ndd1, Swi5, and Ace2. Among these, Swi6, Swi4, and Mbp1 have been reported to function during the G1 stage; McM1, Ndd1, and Fkh1/2 have been reported to function during the G2 stage; and Ace2, Swi5, and Mcm1 have been reported to function during the M stage (Simon et al. 2001). It is not clear where these TFs regulate all cell cycle genes, however.

We applied our methods to identify the key TFs that play roles in the network of regulations from a set of gene expression measurements captured at different time points during the cell cycle. The data set that we use comes from the α-factor synchronized cultures of Spellman et al. (1998). These authors monitored genome-wide mRNA levels for 6,178 yeast ORFs simultaneously using several different methods of synchronization, including an α-factor-mediated G1 arrest, which covers approximately two cell-cycle periods with measurements at 7-minute intervals for 119 minutes, for 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. We let y_{it} denote the log-expression level for gene i at time point t during the cell-cycle process, for $i = 1, \ldots, 297$ and $t = 1, \ldots, 18$. We then applied the mixture model approach (Chen, Jensen, and Stoeckert 2007; Wang, Chen, and Li 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 TFs with at least 1 nonzero binding probability in the 297 genes. We assumed the following VC model to link the binding probabilities to the gene expression levels:

$$y_{it} = \beta_0(t) + \sum_{k=1}^{96} \beta_k(t) x_{ik} + \epsilon_{it},$$

where $\beta_k(t)$ models the transcription effect of the kth TF on gene expression at time t during the cell-cycle process. In this model we assumed that gene expression levels were independent, conditioning on effects of the relevant TFs. In addition, for a given gene i, we also assumed that the ϵ_{it}'s were independent over different time points due to the cross-sectional nature of the experiments. Our goal was to identify the TFs that might be related to the expression patterns of these 297 cell-cycle–regulated genes. Because different TFs may regulate the gene expression at different time points during the cell-cycle process, their effects on gene expression would be expected to be time-dependent.

We applied our method using the GCV defined in (13) to select the tuning parameter, to identify the TFs that affect the expression changes over time for these 297 cell-cycle–regulated genes in the α-factor synchronization experiment. Our procedure identified a total of 71 TFs related to yeast cell-cycle processes, including 19 of the 21 known and experimentally verified cell-cycle–related TFs, all of which exhibited time-dependent effects on gene expression levels. These effects followed similar trends in the two cell cycle periods. The other two TFs (CBF1 and GCN4) were not selected by our procedure. The minimum p values over 18 time points from simple linear regressions were .06 and .14, also indicating that CBF1 and GCN4 were not related to expression variation over time. Almost all of the 52 additional TFs selected by our procedure showed estimated periodic transcriptional effects. The identified TFs included 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, Lu, and Li 2005). Of these 52 TFs, 34 belong to the cooperative pairs of the TFs identified by Banerjee and Zhang (2003). Overall, the model
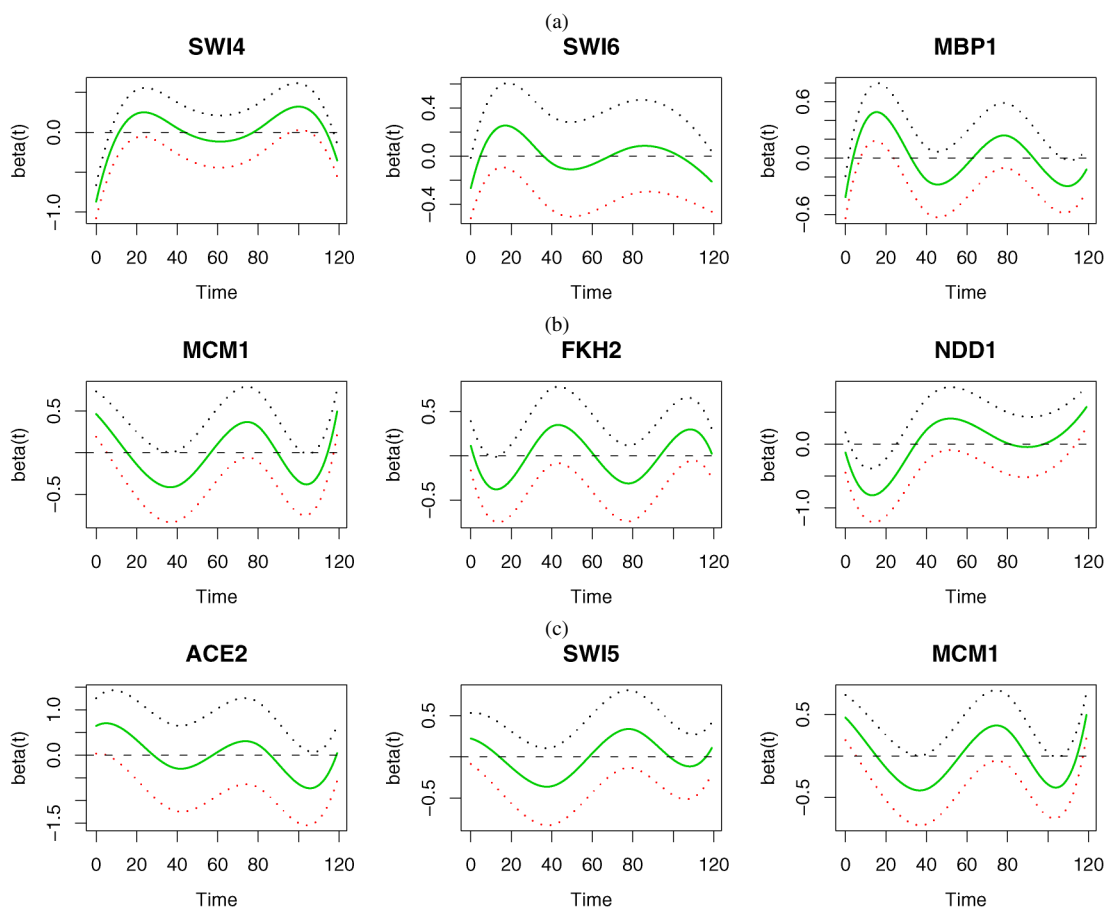
Figure 4. Application to yeast cell-cycle gene expression data. Estimated time-varying coefficients for selected TFs (——) and pointwise 90% confidence intervals (···). (a) TFs regulating genes expressed at the G1 phase. (b) TFs regulating genes expressed at the G2 phase. (c) TFs regulating genes expressed at the M phase.

can explain 43% of the total variations of the gene expression levels.

Figure 4 shows the estimated time-dependent transcriptional effects of nine of the experimentally verified TFs known to regulate cell-cycle–regulated genes at different stages. The panel's in (a) show the transcriptional effects of three TFs, Swi4, Swi6, and Mbp1, which 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 approximately the time points corresponding to the G1 phase of the cell-cycle process. Figure 4(b) shows the transcriptional effects of three TFs, Mcm1, Fkh2, and Ndd1, which regulate gene expression at the G2 phase (Simon et al. 2001). The estimated transcriptional effects of two of these three TFs, Fkh2 and Ndd1, are similar, with peak effects obtained at approximately the time points corresponding to the G2 phase of the cell-cycle process. The estimated effect of Mcm1 was somewhat different, however. Mcm1 is also known to regulate gene expression at the M phase (Simon et al. 2001). The panels in Figure 4(c) show the transcriptional effects of three TFs, Swi5, Ace2, and Mcm1, which regulate gene expression at the M phase (Simon et al. 2001), indicating similar transcriptional effects of these three TFs, with peak effects at approximately the time points corresponding to the M phase of the cell cycle. These plots demonstrate that our procedure indeed

can identify the known cell-cycle–related TFs and point to the stages at which these TFs function to regulate the expressions of cell-cycle genes. In comparison, only two of these nine TFs (SWI4 and FKH2) showed certain periodic patterns in the observed gene expressions over the two cell-cycle periods, indicating that by examining only the expression profiles, we could miss many relevant TFs. Our method incorporates both time-course gene expression and ChIP-chip binding data and enables us to identify more TFs related to the cell-cycle process.

Finally, to assess the false identification of the TFs related to a dynamic biological procedure, we randomly assigned the observed gene expression profiles to the 297 genes while keeping the values and the orders of the expression values for a given profile unchanged, and applied our procedure to the new data sets. The goal of such permutations was to scramble the response data and to create new expression profiles that do not depend on the motif-binding probabilities, to see to what extent the proposed procedure identifies the spurious variables. 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.

8. DISCUSSION

We have proposed a regularized estimation procedure for nonparametric varying-coefficient models that can simultaneously perform variable selection and estimation of smooth coefficient functions. The procedure is applicable to both longitudinal and functional responses. Our method extends the SCAD penalty of Fan and Li (2001) from parametric settings to a nonparametric setting. We have shown that the proposed method is consistent in variable selection and has the oracle property that the asymptotic distribution of the nonparametric estimate of a smooth function with variable selection is the same as that when the variables that are included in the model are known. This oracle property has been obtained for parametric settings; however, our result seems to be the first of its kind in a nonparametric setting. A computation algorithm was developed based on local quadratic approximation to the criterion function. Simulation studies indicated that the proposed procedure was very effective in selecting relevant variables and estimating the smooth regression coefficient functions. Results from application to the yeast cell-cycle data set indicate that the procedure can be effective in selecting the TFs that may play important roles in regulating gene expression during the cell-cycle process.

In our method, we need regularization for two purposes: control of the smoothness of the function estimates and variable selection. In our approach, the number of basis functions/number of knots of splines is used to control the smoothness, whereas the SCAD parameter is used to set the threshold for variable selection. Thus the two tasks of regularization are disentangled. To achieve finer control of smoothness, however, it is desirable to use the roughness penalty for introducing smoothness in function estimation, as in penalized splines or smoothing splines. Following the idea of COSSO, we could replace $\sum_k p_\lambda(\|g_k\|)$ in (4) by $\lambda \sum_k \|g_k\|_W$, where $\|\cdot\|_W$ denotes a Sobolev norm, and then minimize the modified criterion over the corresponding Sobolev space. But such an extension of COSSO is not very attractive compared with our method, because a single tuning parameter λ serves two different purposes of regularization. It is an interesting open question to develop a unified penalization method that does both smoothing and variable selection but also decouples the two tasks in some manner. Recent results on regularization methods with a diverging number of parameters (Fan and Peng 2004; Huang et al. 2008) may provide theoretical insight into such a procedure.

As we explained in Section 2, our methodology applies when a working correlation structure is used to take into account within-subject correlation for longitudinal data. We have presented asymptotic results only for the working independence case, however. Extension of these results to parametrically estimated correlation structures is relatively straightforward, using the arguments of Huang, Zhang, and Zhou (2007); the ratio of the smallest and largest eigenvalues of the weighting matrix $\mathbf{W}$ must be assumed to be bounded. Extension to nonparametrically estimated correlation structures (Wu and Pourahmadi 2003; Huang et al. 2006; Huang, Liu, and Liu 2007) will be more challenging. A careful study of the asymptotics of our methodology for general working correlation structures is left for future research.

APPENDIX: PROOFS

A.1 Proof of Lemma 1

Here we provide proof by contradiction. Suppose that $\mathbf{U}_i \hat{\boldsymbol{\gamma}}^{\star(-i)} \neq \mathbf{U}_i \tilde{\boldsymbol{\gamma}}^{(i)}$. Denote (11) by $pl_{ridge}(\boldsymbol{\gamma}, \mathbf{y})$. Then

$$pl_{ridge}(\tilde{\boldsymbol{\gamma}}^{(i)}, \tilde{\mathbf{y}}^{(i)})$$
$$= \frac{1}{n} \sum_{k=1}^n \|\tilde{\mathbf{y}}_k^{(i)} - \mathbf{U}_k \tilde{\boldsymbol{\gamma}}_k^{(i)}\|_2^2 + \frac{1}{2}(\tilde{\boldsymbol{\gamma}}^{(i)})^T \Sigma_\lambda(\hat{\boldsymbol{\gamma}}) \tilde{\boldsymbol{\gamma}}^{(i)}$$
$$> \frac{1}{n} \sum_{k \neq i} \|\mathbf{y}_k - \mathbf{U}_k \tilde{\boldsymbol{\gamma}}_k^{(i)}\|_2^2 + \frac{1}{2}(\tilde{\boldsymbol{\gamma}}^{(i)})^T \Sigma_\lambda(\hat{\boldsymbol{\gamma}}) \tilde{\boldsymbol{\gamma}}^{(i)}$$
$$\geq \frac{1}{n} \sum_{k \neq i} \|\mathbf{y}_k - \mathbf{U}_k \hat{\boldsymbol{\gamma}}_k^{\star(-i)}\|_2^2 + \frac{1}{2}(\hat{\boldsymbol{\gamma}}^{\star(-i)})^T \Sigma_\lambda(\hat{\boldsymbol{\gamma}}) \hat{\boldsymbol{\gamma}}^{\star(-i)}$$
$$= pl_{ridge}(\hat{\boldsymbol{\gamma}}^{\star(-i)}, \tilde{\mathbf{y}}^{(i)}).$$

This contradicts the fact that $\tilde{\boldsymbol{\gamma}}^{(i)}$ minimizes $pl_{ridge}(\tilde{\boldsymbol{\gamma}}^{(i)}, \tilde{\mathbf{y}}^{(i)})$.

A.2 Useful Lemmas

This section provides some lemmas that facilitate the proofs of Theorems 1 and 2. Define $\tilde{\mathbf{y}}_i = E(\mathbf{y}_i | \mathbf{x}_i)$, $\tilde{\boldsymbol{\gamma}} = (\sum_{i=1}^n \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i)^{-1} \times (\sum_{i=1}^n \mathbf{U}_i^T \mathbf{W}_i \tilde{\mathbf{y}}_i)$, and $\tilde{\boldsymbol{\beta}}(t) = \mathbf{B}(t)\tilde{\boldsymbol{\gamma}}$. Here $\tilde{\boldsymbol{\beta}}(t)$ can be considered the conditional mean of $\hat{\boldsymbol{\beta}}(t)$. We have a bias-variance decomposition, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Lemmas A.2 and A.4 quantify the rates of convergence of $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{L_2}$ and $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L_2}$. The rate given in Lemma A.4 for the variance term is refined in part b of Theorem 1.

Lemma A.1. Suppose that (17) holds. Then there exists an interval $[M_3, M_4]$, $0 < M_3 < M_4 < \infty$, such that all of the eigenvalues of $n^{-1} L_n \sum_{i=1}^n \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i$ fall in $[M_3, M_4]$ with probability approaching 1 as $n \to \infty$.

Lemma A.2. Suppose that (17) holds. Then $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{L_2} = O_P(\rho_n)$.

Lemmas A.1 and A.2 are from lemmas A.3 and A.7 of Huang et al. (2004), and their proofs are omitted here.

Define $r_n = n^{-1}[\sum_{i=1}^n w_i^{-2}\{L_n J_i + J_i(J_i - 1)\}]^{1/2}$. When $w_i = 1/J_i$ or $w_i = 1/J_i$, with J_i bounded uniformly, we can show that $r_n \asymp (L_n/n)^{1/2}$.

Lemma A.3. $n^{-1} L_n^{1/2} \sup_{\mathbf{v}}(|\boldsymbol{\varepsilon}^T \mathbf{W}\mathbf{U}\mathbf{v}|/|\mathbf{v}|) = O_P(r_n)$, where $|v| = \|v\|_{l_2}$.

Proof of Lemma A.3. By the Cauchy–Schwarz inequality,

$$\sup_{\mathbf{v}} \left\{ \left(\frac{|\boldsymbol{\varepsilon}^T \mathbf{W}\mathbf{U}\mathbf{v}|}{|\mathbf{v}|} \right)^2 \right\} \leq \boldsymbol{\varepsilon}^T \mathbf{W}\mathbf{U}\mathbf{U}^T \mathbf{W}\boldsymbol{\varepsilon}. \qquad (A.1)$$

The expected value of the right side equals

$$E(\boldsymbol{\varepsilon}^T \mathbf{W}\mathbf{U}\mathbf{U}^T \mathbf{W}\boldsymbol{\varepsilon}) = E\left\{ \left(\sum_{i=1}^n \boldsymbol{\varepsilon}_i^T \mathbf{W}_i \mathbf{U}_i \right) \left(\sum_{i=1}^n \mathbf{U}_i^T \mathbf{W}_i \boldsymbol{\varepsilon}_i \right) \right\}$$
$$= \sum_{i=1}^n E(\boldsymbol{\varepsilon}_i^T \mathbf{W}_i \mathbf{U}_i \mathbf{U}_i^T \mathbf{W}_i \boldsymbol{\varepsilon}_i). \qquad (A.2)$$

Note that

$$E(\boldsymbol{\varepsilon}_i^T \mathbf{W}_i \mathbf{U}_i \mathbf{U}_i^T \mathbf{W}_i \boldsymbol{\varepsilon}_i)$$
$$= w_i^{-2} E\left[\sum_{k,l} \left\{ \sum_j \varepsilon_i(t_{ij}) X_{ik}(t_{ij}) B_{kl}(t_{ij}) \right\}^2 \right]$$
$$= w_i^{-2} \sum_{k,l} \left[\sum_j E\{\varepsilon_i(t_{ij})^2 X_{ik}(t_{ij})^2 B_{kl}(t_{ij})^2\} \right.$$

$$+ \sum_{j \neq j'} E\{\varepsilon_i(t_{ij}) X_{ik}(t_{ij}) B_{kl}(t_{ij}) \varepsilon_i(t_{ij'}) X_{ik}(t_{ij'}) B_{kl}(t_{ij'})\}\Bigg].$$

On the other hand,

$$E\{B_{kl}(t_{ij}) B_{kl}(t_{ij'})\} = E\{B_{kl}(t_{ij})\} E\{B_{kl}(t_{ij'})\}$$

$$\leq \left\{ \sup_{t \in [0,T]} f(t) \int_0^T B_{kl}(t)\, dt \right\}^2$$

$$\leq C_1^2 L_k^{-2}$$

and $E\{B_{kl}(t)^2\} \leq E\{B_{kl}(t)\} \leq C_1 L_k^{-1}$; therefore,

$$E(\boldsymbol{\varepsilon}_i^T \mathbf{W}_i \mathbf{U}_i \mathbf{U}_i^T \mathbf{W}_i \boldsymbol{\varepsilon}_i) \leq C_2 w_i^{-2}\{J_i + J_i(J_i - 1) L_n^{-1}\}. \quad (A.3)$$

Combining (A.2) and (A.3), and using the Markov inequality, we obtain

$$\boldsymbol{\varepsilon}^T \mathbf{W} \mathbf{U} \mathbf{U}^T \mathbf{W} \boldsymbol{\varepsilon} = O_P\left[\sum_{i=1}^n w_i^{-2}\{J_i + J_i(J_i - 1) L_n^{-1}\} \right],$$

which, together with (A.1), yields the desired result.

Lemma A.4. Suppose that (17) holds and that $\lambda_n, r_n, \rho_n \to 0$ and $\lambda_n/\rho_n \to \infty$ as $n \to \infty$. Then $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L_2} = O_P\{r_n + (\lambda_n \rho_n)^{1/2}\}$.

Proof of Lemma A.4. Using properties of B-spline basis functions (see sec. A.2 of Huang et al. 2004), we have

$$\|\hat{\boldsymbol{\beta}}_k - \tilde{\boldsymbol{\beta}}_k\|_{L_2}^2 = \|\hat{\boldsymbol{\gamma}}_k - \tilde{\boldsymbol{\gamma}}_k\|_k^2 \asymp L_k^{-1} \|\hat{\boldsymbol{\gamma}}_k - \tilde{\boldsymbol{\gamma}}_k\|_{l_2}^2.$$

Sum over k to obtain

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L_2}^2 = \sum_{k=1}^p \|\hat{\boldsymbol{\gamma}}_k - \tilde{\boldsymbol{\gamma}}_k\|_k^2 \asymp L_n^{-1} \|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}\|_{l_2}^2.$$

Let $\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}} = \delta_n L_n^{1/2} \mathbf{v}$, with δ_n a scalar and $\mathbf{v}$ a vector satisfying $\|\mathbf{v}\|_{l_2} = 1$. It follows that $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L_2} \asymp \delta_n$. We first show that $\delta_n = O_p(r_n + \lambda_n)$, and then refine this rate of convergence in the second step of the proof.

Let $\boldsymbol{\varepsilon}_i = (\varepsilon_i(t_{i1}), \ldots, \varepsilon_i(t_{iJ_i}))^T$, $i = 1, \ldots, n$, and $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, \ldots, \boldsymbol{\varepsilon}_n^T)^T$. By the definition of $\tilde{\boldsymbol{\gamma}}$, $\mathbf{U}^T \mathbf{W}(\mathbf{y} - \boldsymbol{\varepsilon} - \mathbf{U}\tilde{\boldsymbol{\gamma}}) = 0$. Because $\hat{\boldsymbol{\gamma}} = \tilde{\boldsymbol{\gamma}} + \delta_n L_n^{1/2} \mathbf{v}$,

$$(\mathbf{y} - \mathbf{U}\hat{\boldsymbol{\gamma}})^T \mathbf{W}(\mathbf{y} - \mathbf{U}\hat{\boldsymbol{\gamma}}) - (\mathbf{y} - \mathbf{U}\tilde{\boldsymbol{\gamma}})^T \mathbf{W}(\mathbf{y} - \mathbf{U}\tilde{\boldsymbol{\gamma}})$$

$$= -2\delta_n L_n^{1/2} \mathbf{v}^T \mathbf{U}^T \mathbf{W}(\mathbf{y} - \mathbf{U}\tilde{\boldsymbol{\gamma}}) + \delta_n^2 L_n \mathbf{v}^T \mathbf{U}^T \mathbf{W} \mathbf{U} \mathbf{v}$$

$$= -2\delta_n L_n^{1/2} \mathbf{v}^T \mathbf{U}^T \mathbf{W} \boldsymbol{\varepsilon} + \delta_n^2 L_n \mathbf{v}^T \mathbf{U}^T \mathbf{W} \mathbf{U} \mathbf{v}.$$

Thus

$$pl(\hat{\boldsymbol{\gamma}}) - pl(\tilde{\boldsymbol{\gamma}}) = -\frac{2\delta_n L_n^{1/2}}{n} \mathbf{v}^T \mathbf{U}^T \mathbf{W} \boldsymbol{\varepsilon} + \frac{\delta_n^2 L_n}{n} \mathbf{v}^T \mathbf{U}^T \mathbf{W} \mathbf{U} \mathbf{v}$$

$$+ \sum_{k=1}^p \{p_{\lambda_n}(\|\hat{\boldsymbol{\gamma}}_k\|_k) - p_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_k)\}, \quad (A.4)$$

where the $\mathbf{v}_k$'s are components of $\mathbf{v}$ when the latter is partitioned as $\tilde{\boldsymbol{\gamma}}$.

By the definition of $\hat{\boldsymbol{\gamma}}$, $pl(\hat{\boldsymbol{\gamma}}) - pl(\tilde{\boldsymbol{\gamma}}) \leq 0$. According to Lemma A.3,

$$\frac{L_n^{1/2}}{n} \mathbf{v}^T \mathbf{U}^T \mathbf{W} \boldsymbol{\varepsilon} = \frac{L_n^{1/2}}{n} \sum_{i=1}^n \boldsymbol{\varepsilon}_i^T \mathbf{W}_i \mathbf{U}_i \mathbf{v} = O_P(r_n). \quad (A.5)$$

By Lemma A.1,

$$\frac{L_n}{n} \mathbf{v}^T \mathbf{U}^T \mathbf{W} \mathbf{U} \mathbf{v} = \frac{L_n}{n} \sum_{i=1}^n \mathbf{v}^T \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i \mathbf{v} \geq M_3, \quad (A.6)$$

with probability approaching 1. Using the inequality $|p_\lambda(a) - p_\lambda(b)| \leq \lambda|a - b|$, we obtain

$$\sum_{k=1}^p \{p_{\lambda_n}(\|\hat{\boldsymbol{\gamma}}_k\|_k) - p_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_k)\} \geq \sum_{k=1}^p -\lambda_n \|\hat{\boldsymbol{\gamma}}_k - \tilde{\boldsymbol{\gamma}}_k\|_k$$

$$\asymp -\lambda_n \delta_n. \quad (A.7)$$

Therefore, $0 \geq -O_p(r_n)\delta_n + M_3 \delta_n^2 - \lambda_n \delta_n$ with probability approaching 1, which implies that $\delta_n = O_p(r_n + \lambda_n)$.

Now we proceed to improve the obtained rate and show that $\delta_n = O_P\{r_n + (\lambda_n \rho_n)^{1/2}\}$. For $k = 1, \ldots, p$, we have $|\|\hat{\boldsymbol{\gamma}}_k\|_k - \|\tilde{\boldsymbol{\gamma}}_k\|_k| \leq \|\hat{\boldsymbol{\gamma}}_k - \tilde{\boldsymbol{\gamma}}_k\|_k = o_p(1)$ and

$$\left| \|\tilde{\boldsymbol{\gamma}}_k\|_k - \|\boldsymbol{\beta}_k\|_{L_2} \right| \leq \left| \|\tilde{\boldsymbol{\beta}}_k\|_{L_2} - \|\boldsymbol{\beta}_k\|_{L_2} \right|$$

$$\leq \|\tilde{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_{L_2} = O_p(\rho_n) = o_P(1). \quad (A.8)$$

It follows that $\|\hat{\boldsymbol{\gamma}}_k\|_k \to \|\beta_k\|_{L_2}$ and $\|\tilde{\boldsymbol{\gamma}}_k\|_k \to \|\beta_k\|_{L_2}$ with probability. Because $\|\beta_k\|_{L_2} > 0$ for $k = 1, \ldots, s$ and $\lambda_n \to 0$, we have that, with probability approaching 1, $\|\hat{\boldsymbol{\gamma}}_k\|_k > a\lambda_n$, $\|\tilde{\boldsymbol{\gamma}}_k\|_k > a\lambda_n$, $k = 1, \ldots, s$. On the other hand, $\|\beta_k\|_{L_2} = 0$ for $k = s+1, \ldots, p$, so (A.8) implies that $\|\tilde{\boldsymbol{\gamma}}_k\|_k = O_P(\rho_n)$. Because $\lambda_n/\rho_n \to \infty$, we have that with probability approaching 1, $\|\tilde{\boldsymbol{\gamma}}_k\|_k < \lambda_n$, $k = s+1, \ldots, p$. Consequently, by the definition of $p_\lambda(\cdot)$, $P\{p_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_{H_k}) = p_{\lambda_n}(\|\hat{\boldsymbol{\gamma}}_k\|_{H_k})\} \to 1$, $k = 1, \ldots, s$, and $P\{p_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_{H_k}) = \lambda_n \times \|\tilde{\boldsymbol{\gamma}}_k\|_{H_k}\} \to 1$, $k = s+1, \ldots, p$; therefore,

$$\sum_{k=1}^p \{p_{\lambda_n}(\|\hat{\boldsymbol{\gamma}}_k\|_k) - p_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_k)\} = \lambda_n \sum_{k=s+1}^p \|\tilde{\boldsymbol{\gamma}}_k\|_k \geq -O_P(\lambda_n \rho_n).$$

This result, combined with (A.4), (A.5) and (A.6), implies that, with probability approaching 1,

$$pl(\hat{\boldsymbol{\gamma}}) - pl(\tilde{\boldsymbol{\gamma}}) \geq -O_p(r_n)\delta_n + M_3 \delta_n^2 - O_p(\lambda_n \rho_n),$$

which in turn implies that $\delta_n = O_P\{r_n + (\lambda_n \rho_n)^{1/2}\}$. The proof is complete.

A.3 Proof of Theorem 1

To prove part a of Theorem 1, we use proof by contradiction. Suppose that for n sufficiently large, there exists a constant $\eta > 0$ such that with probability at least η, there exists a $k_0 > s$ such that $\hat{\beta}_{k_0}(t) \neq 0$. Then $\|\hat{\boldsymbol{\gamma}}_{k_0}\|_{k_0} = \|\hat{\beta}_{k_0}(\cdot)\|_{L_2} > 0$. Let $\hat{\boldsymbol{\gamma}}^*$ be a vector constructed by replacing $\hat{\boldsymbol{\gamma}}_{k_0}$ with $\mathbf{0}$ in $\hat{\boldsymbol{\gamma}}$. Then

$$pl(\hat{\boldsymbol{\gamma}}) - pl(\hat{\boldsymbol{\gamma}}^*)$$

$$= \frac{1}{n} \left(\|\mathbf{y} - \mathbf{U}\hat{\boldsymbol{\gamma}}\|^2 - \|\mathbf{y} - \mathbf{U}\hat{\boldsymbol{\gamma}}^*\|^2 \right) + p_{\lambda_n}\left(\|\hat{\boldsymbol{\gamma}}_{k_0}\|_{k_0} \right). \quad (A.9)$$

By Lemma A.4 and the fact that $\beta_{k_0}(t) = 0$, $\|\hat{\boldsymbol{\gamma}}_{k_0}\|_{k_0} = \|\hat{\beta}_{k_0}\| = O_p\{r_n + (\lambda_n \rho_n)^{1/2}\}$. Because $\lambda_n/\max(r_n, \rho_n) \to \infty$, we have that $\lambda_n > \|\hat{\boldsymbol{\gamma}}_{k_0}\|_{k_0}$ and thus $p_{\lambda_n}(\|\hat{\boldsymbol{\gamma}}_{k_0}\|_{k_0}) = \lambda_n \|\hat{\boldsymbol{\gamma}}_{k_0}\|_{k_0}$ with probability approaching 1. To analyze the first term on the right side of (A.9), note that

$$\|\mathbf{y} - \mathbf{U}\hat{\boldsymbol{\gamma}}\|^2 - \|\mathbf{y} - \mathbf{U}\hat{\boldsymbol{\gamma}}^*\|^2$$

$$\geq -(\mathbf{y} - \mathbf{U}\hat{\boldsymbol{\gamma}}^*)^T \mathbf{W} \mathbf{U}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)$$

$$= -2(\mathbf{y} - \mathbf{U}\tilde{\boldsymbol{\gamma}})^T \mathbf{W} \mathbf{U}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)$$

$$- 2(\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)^T \mathbf{U}^T \mathbf{W} \mathbf{U}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*). \quad (A.10)$$

By the Cauchy–Schwarz inequality and Lemma A.1,

$$(\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)^T (\mathbf{U}^T \mathbf{W} \mathbf{U})(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)$$

$$\leq \{(\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)^T (\mathbf{U}^T \mathbf{W} \mathbf{U})(\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)\}^{1/2}$$

$$\times \{(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)^T (\mathbf{U}^T \mathbf{W} \mathbf{U})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*)\}^{1/2}$$

$$\leq M_4 \frac{n}{L_n} \|\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*\|_{l_2} \|\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*\|_{l_2}.$$

From the triangle inequality and Lemmas A.2 and A.4, it follows that

$$\|\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*\|_{l_2} \le \|\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}\|_{l_2} + \|\hat{\boldsymbol{\gamma}}_{k_0}\|_{l_2}$$

$$= O_p\big[L_n^{1/2}\{r_n + (\lambda_n\rho_n)^{1/2} + \rho_n\}\big];$$

thus

$$(\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)^T(\mathbf{U}^T\mathbf{W}\mathbf{U})(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)$$

$$= O_p\big[L_n^{1/2}\{r_n + (\lambda_n\rho_n)^{1/2} + \rho_n\}\big]\frac{n}{L_n}\|\hat{\boldsymbol{\gamma}}_{k_0}\|_{l_2}. \quad \text{(A.11)}$$

Lemma A.3 implies that

$$|(\mathbf{y} - \mathbf{U}\tilde{\boldsymbol{\gamma}})^T\mathbf{W}\mathbf{U}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)| = |\boldsymbol{\varepsilon}^T\mathbf{W}\mathbf{U}(\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}^*)|$$

$$= O_P\left(\frac{nr_n}{L_n^{1/2}}\right)\|\hat{\boldsymbol{\gamma}}_{k_0}\|_{l_2}. \quad \text{(A.12)}$$

Combining (A.9)–(A.12), we get

$$pl(\hat{\boldsymbol{\gamma}}) - pl(\hat{\boldsymbol{\gamma}}^*) \ge -O_p\left(\frac{r_n}{L_n^{1/2}}\right)\|\hat{\boldsymbol{\gamma}}_{k_0}\|_{l_2}$$

$$- O_p\left(\frac{r_n + (\lambda_n\rho_n)^{1/2} + \rho_n}{L_n^{1/2}}\right)\|\hat{\boldsymbol{\gamma}}_{k_0}\|_{l_2}$$

$$+ \frac{\lambda_n}{L_n^{1/2}}\|\hat{\boldsymbol{\gamma}}_{k_0}\|_{l_2}. \quad \text{(A.13)}$$

Note that on the right side of (A.13), the third term dominates both the first and second terms, because $\lambda_n/\max(r_n, \rho_n) \to \infty$. This contradicts the fact that $pl(\hat{\boldsymbol{\gamma}}) - pl(\boldsymbol{\gamma}^*) \le 0$. We thus have proved part a.

To prove part b, write $\boldsymbol{\beta} = ((\boldsymbol{\beta}^{(1)})^T, (\boldsymbol{\beta}^{(2)})^T)^T$, where $\boldsymbol{\beta}^{(1)} = (\beta_1, \ldots, \beta_s)^T$ and $\boldsymbol{\beta}^{(2)} = (\beta_{s+1}, \ldots, \beta_p)^T$, and 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 $\boldsymbol{\gamma}^{(2)} = (\boldsymbol{\gamma}_{s+1}^T, \ldots, \boldsymbol{\gamma}_p^T)^T$. Similarly, write $\mathbf{U}_i = (\mathbf{U}_i^{(1)}, \mathbf{U}_i^{(2)})$ and $\mathbf{U} = (\mathbf{U}^{(1)}, \mathbf{U}^{(2)})$. Define the oracle version of $\tilde{\boldsymbol{\gamma}}$,

$$\tilde{\boldsymbol{\gamma}}_{oracle} = \arg \min_{\boldsymbol{\gamma}=(\boldsymbol{\gamma}^{(1)T},\mathbf{0}^T)^T} \frac{1}{n}\sum_{i=1}^n w_i(\tilde{\mathbf{y}}_i - \mathbf{U}_i\boldsymbol{\gamma})^T(\tilde{\mathbf{y}}_i - \mathbf{U}_i\boldsymbol{\gamma})$$

$$= \begin{pmatrix} (\sum_i \mathbf{U}_i^{(1)T}\mathbf{W}_i\mathbf{U}_i^{(1)})^{-1}(\sum_i \mathbf{U}_i^{(1)T}\mathbf{W}_i\tilde{\mathbf{y}}_i) \\ \mathbf{0} \end{pmatrix}, \quad \text{(A.14)}$$

which is obtained as if the information of the nonzero components were given; the corresponding vector of coefficient functions is designated $\tilde{\boldsymbol{\beta}}_{oracle}$. Note that the true vector of coefficient functions is $\boldsymbol{\beta} = ((\boldsymbol{\beta}^{(1)})^T, \mathbf{0}^T)^T$. By Lemma A.2, $\|\tilde{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\|_{L_2} = O_P(\rho_n) = o_P(1)$. By Lemmas A.2 and A.4, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{L_2} = o_P(1)$. Thus, with probability approaching 1, $\|\tilde{\beta}_{k,oracle}\|_{L_2} \to \|\beta_k\|_{L_2} > 0$ and $\|\hat{\beta}_k\|_{L_2} \to \|\beta_k\|_{L_2} > 0$, for $k = 1, \ldots, s$. On the other hand, by the definition of $\tilde{\boldsymbol{\beta}}_{oracle}$, $\|\tilde{\beta}_{k,oracle}\|_{L_2} = 0$, and by part a of the theorem, with probability approaching 1, $\|\hat{\beta}_k\|_{L_2} = 0$, for $k = s+1, \ldots, p$; consequently,

$$\sum_{k=1}^p p_{\lambda_n}(\|\tilde{\beta}_{k,oracle}\|_{L_2}) = \sum_{k=1}^p p_{\lambda_n}(\|\hat{\beta}_k\|_{L_2}), \quad \text{(A.15)}$$

with probability approaching 1. From part a of the theorem, with probability approaching 1, $\hat{\boldsymbol{\gamma}} = ((\hat{\boldsymbol{\gamma}}^{(1)})^T, \mathbf{0}^T)^T$. Let $\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_{oracle} = \delta_n L_n^{1/2}\mathbf{v}$ with $\mathbf{v} = ((\mathbf{v}^{(1)})^T, \mathbf{0}^T)^T$ and $\|\mathbf{v}^{(1)}\|_{l_2} = 1$. Then $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}_{oracle}\|_{L_2} \asymp L_n^{-1}\|\hat{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_{oracle}\|_{l_2} = \delta_n$. By (A.4) and (A.15) and the argument following (A.4),

$$0 \ge pl(\hat{\boldsymbol{\gamma}}) - pl(\tilde{\boldsymbol{\gamma}}_{oracle})$$

$$= -\frac{2\delta_n L_n^{1/2}}{n}\mathbf{v}^T(\mathbf{U}^{(1)})^T\mathbf{W}\boldsymbol{\varepsilon} + \frac{\delta_n^2 L_n}{n}\mathbf{v}^T(\mathbf{U}^{(1)})^T\mathbf{W}\mathbf{U}^{(1)}\mathbf{v}$$

$$\ge -O_p(r_n)\delta_n + M_3\delta_n^2.$$

Thus $\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|_{L_2} \asymp \delta_n = O_p(r_n)$, which, together with $\|\tilde{\boldsymbol{\beta}}_{oracle} - \boldsymbol{\beta}\|_{L_2} = o_P(\rho_n)$, implies that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{L^2} = O_p(\rho_n + r_n)$. The desired result follows.

A.4 Proof of Corollary 1

By theorem 6.27 of Schumaker (1981), $\inf_{g\in\mathcal{G}_k} \|\beta_k - g\|_{L_\infty} = O(L_k^{-2})$, and thus $\rho_n = O(\sum_{k=1}^p L_k^{-2}) = O(L_n^{-2})$. Moreover, $r_n \asymp (L_n/n)^{1/2}$. Thus the convergence rate is $\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\|_{L_2} = O_P\{(L_n/n)^{1/2} + L_n^{-2}\}$. When $L_n \asymp n^{1/5}$, $(L_n/n)^{1/2} + L_n^{-2} \asymp n^{-2/5}$. The proof is complete.

A.5 Proof of Theorem 2

According to the proof of Lemma A.4, with probability approaching 1, $\|\tilde{\boldsymbol{\gamma}}_k\|_k > a\lambda_n$, $\|\hat{\boldsymbol{\gamma}}_k\|_k > a\lambda_n$, and thus $p_{\lambda_n}(\|\tilde{\boldsymbol{\gamma}}_k\|_k) = p_{\lambda_n}(\|\hat{\boldsymbol{\gamma}}_k\|_k)$, $k = 1, \ldots, s$. By Theorem 1, with probability approaching 1, $\hat{\boldsymbol{\gamma}} = ((\hat{\boldsymbol{\gamma}}^{(1)})^T, \mathbf{0}^T)^T$ is a local minimizer of $pl(\boldsymbol{\gamma})$. Note that $pl(\boldsymbol{\gamma}^{(1)})$ is quadratic in $\boldsymbol{\gamma}^{(1)} = (\gamma_1, \ldots, \gamma_s)^T$ when $\|\boldsymbol{\gamma}_k\|_k > a\lambda_n$, $k = 1, \ldots, s$; therefore, $\partial pl(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^{(1)}|_{\boldsymbol{\gamma}^{(1)}=\hat{\boldsymbol{\gamma}}^{(1)},\boldsymbol{\gamma}^{(2)}=\mathbf{0}} = \mathbf{0}$, which implies that

$$\hat{\boldsymbol{\gamma}}^{(1)} = \left(\sum_{i=1}^n (\mathbf{U}_i^{(1)})^T\mathbf{W}_i\mathbf{U}_i^{(1)}\right)^{-1}\left(\sum_{i=1}^n (\mathbf{U}_i^{(1)})^T\mathbf{W}_i\mathbf{Y}_i\right).$$

Let

$$\tilde{\boldsymbol{\gamma}}^{(1)} = \left(\sum_{i=1}^n (\mathbf{U}_i^{(1)})^T\mathbf{W}_i\mathbf{U}_i^{(1)}\right)^{-1}\left(\sum_{i=1}^n (\mathbf{U}_i^{(1)})^T\mathbf{W}_i E(\mathbf{Y}_i|\mathbf{x}_i)\right).$$

Let avar*$(\hat{\boldsymbol{\gamma}}^{(1)})$ be a modification of avar$(\hat{\boldsymbol{\gamma}}^{(1)})$ given in (20) by replacing $\boldsymbol{\Omega}_{\lambda_n}$ in (19) with 0. Applying theorem V.8 of Petrov (1975) as in the proof of theorem 4.1 of Huang (2003) and using the arguments in the proof of lemma A.8 of Huang et al. (2004), we obtain that for any vector $\mathbf{c}_n$ with dimension $\sum_{k=1}^s L_s$ and components not all 0,

$$\{\mathbf{c}_n^T \text{avar}^*(\hat{\boldsymbol{\gamma}}^{(1)})\mathbf{c}_n\}^{-1/2}\mathbf{c}_n^T(\hat{\boldsymbol{\gamma}}^{(1)} - \tilde{\boldsymbol{\gamma}}^{(1)})$$

$$\to N(0, 1) \qquad \text{in distribution.}$$

For any s-vector $\mathbf{a}_n$ whose components are not all 0, choosing $\mathbf{c}_n = (\mathbf{B}^{(1)}(t))^T\mathbf{a}_n$ yields

$$[\mathbf{a}_n^T \text{avar}^*\{\hat{\boldsymbol{\beta}}^{(1)}(t)\}\mathbf{a}_n]^{-1/2}\mathbf{a}_n^T\{\hat{\boldsymbol{\beta}}^{(1)}(t) - \tilde{\boldsymbol{\beta}}^{(1)}(t)\}$$

$$\to N(0, 1) \qquad \text{in distribution,}$$

which in turn yields the desired result.

REFERENCES

Banerjee, N., and Zhang, M. Q. (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.

Chiang, C., Rice, J. A., and Wu, C. O. (2001), "Smoothing Spine Estimation for Varying Coeficients Models With Repeated Measured Dependent Variables," *Journal of the American Statistical Association*, 96, 605–619.

Conlon, E. M., Liu, X. S., Lieb, J. D., and Liu, J. S. (2003), "Integrating Regulatory Motif Discovery and Genome-Wide Expression Analysis," *Proceedings of National Academy of Sciences*, 100, 3339–3344.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford, U.K.: Oxford University Press.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Like-lihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

—— (2004), "New Estimation and Model Selection Procedures for Semi-parametric Modeling in Longitudinal Data Analysis," *Journal of the American Statistical Association*, 99, 710–723.

Fan, J., and Peng, H. (2004), "On Non-Concave Penalized Likelihood With Diverging Number of Parameters," *The Annals of Statistics*, 32, 928–961.

Fan, J., and Zhang, J.-T. (2000), "Functional Linear Models for Longitudinal Data," *Journal of the Royal Statisitcal Society*, Ser. B, 62, 303–322.

Hastie, T., and Tibshirani, R. (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society*, Ser. B, 55, 757–796.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. (1998), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data," *Biometrika*, 85, 809–822.

Huang, J., Horowitz, J. L., and Ma, S. (2008), "Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models," *The Annals of Statistics*, 36, 587–613.

Huang, J. Z. (2003), "Local Asymptotics for Polynomial Spline Regression," *The Annals of Statistics*, 31, 1600–1635.

Huang, J. Z., Liu, L., and Liu, N. (2007), "Estimation of Large Covariance Matrices of Longitudinal Data With Basis Function Approximations," *Journal of Computational and Graphical Statistics*, 16, 189–209.

Huang, J. Z., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance Selection and Estimation via Penalised Normal Likelihood," *Biometrika*, 93, 85–98.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002), "Varying-Coefficient Models and Basis Function Approximation for the Analysis of Repeated Measurements," *Biometrika*, 89, 111–128.

—— (2004), "Polynomial Spline Estimation and Inference for Varying Co-efficient Models With Longitudinal Data," *Statistica Sinica*, 14, 763–788.

Huang, J. Z., Zhang, L., and Zhou, L. (2007), "Efficient Estimation in Marginal Partially Linear Models for Longitudinal/Clustered Data Using Splines," *Scandinavian Journal of Statistics*, 34, 451–477.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R. (1987), "The Multicenter AIDS Cohort Study: Rationale, Organization and Selected Characteristics of the Participants," *American Journal of Epidemiology*, 126, 310–318.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., et al. (2002), "Transcriptional Regulatory Networks in *S. cerevisiae*," *Science*, 298, 799–804.

Leng, C., and Zhang, H. H. (2007), "Nonparametric Model Selection in Hazard Regression," *Journal of Nonparametric Statistics*, 18, 417–429.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lin, X., and Carroll, R. J. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor Is Measured Without/With Error," *Journal of the American Statistical Association*, 95, 520–534.

Lin, Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models—COSSO," *The Annals of Statistics*, 34, 2272–2297.

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.

Petrov, V. (1975), *Sums of Independence Random Variables*, New York: Springer-Verlag.

Qu, A., and Li, R. (2006), "Quadratic Interference Functions for Varyibg Coefficient Models With Longitudinal Data," *Biometrics*, 62, 379–391.

Rice, J. A. (2004), "Functional and Longitudinal Data Analysis: Perspectives on Smoothing," *Statistica Sinica*, 14, 631–647.

Rice, J. A., and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves," *Journal of the Royal Statistical Society*, Ser. B, 53, 233–243.

Rice, J. A., and Wu, C. O. (2001), "Nonparametric Mixed Effects Models for Unequally Sampled Noisy Curves," *Biometrics*, 57, 253–259.

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

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakola, T. S., and Young, R. A. (2001), "Serial Regulation of Transcriptional Regulators in the Yeast Cell Cycle," *Cell*, 106, 697–708.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998), "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization," *Molecular Biology of Cell*, 9, 3273–3297.

Stone, C. J. (1982), "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10, 1348–1360.

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Tsai, H. K., Lu, S. H. H., and Li, W. H. (2005), "Statistical Methods for Identifying Yeast Cell Cycle Transcription Factors," *Proceedings of the National Academy of Sciences*, 102, 13532–13537.

Wang, L., Chen, G., and Li, H. (2007), "Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data," *Bioinformatics*, 23, 1486–1494.

Wu, C. O., and Chiang, C.-T. (2000), "Kernel Smoothing on Varying Coeficient Models With Longitudinal Dependent Variable," *Statistica Sinica*, 10, 433–456.

Wu, W. B., and Pourahmadi, M. (2003), "Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data," *Biometrika*, 90, 831–844.

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

Zeger, S. L., and Diggle, P. J. (1994), "Semiparametric Models for Longitudinal Data With Application to CD4 Cell Numbers in HIV Seroconverters," *Biometrics*, 50, 689–699.

Zhang, H. (2006), "Variable Selection for Support Vector Machines via Smoothing Spline ANOVA," *Statistica Sinica*, 16, 659–674.

Zhang, H., and Lin, Y. (2006), "Component Selection and Smoothing for Nonparametric Regression in Exponential Families," *Statistica Sinica*, 16, 1021–1042.

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

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Ser. B, 67, 301–320.