New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis

Jianqing FAN and Runze LI

Semiparametric regression models are very useful for longitudinal data analysis. The complexity of semiparametric models and the structure of longitudinal data pose new challenges to parametric inferences and model selection that frequently arise from longitudinal data analysis. In this article, two new approaches are proposed for estimating the regression coefficients in a semiparametric model. The asymptotic normality of the resulting estimators is established. An innovative class of variable selection procedures is proposed to select significant variables in the semiparametric models. The proposed procedures are distinguished from others in that they simultaneously select significant variables and estimate unknown parameters. Rates of convergence of the resulting estimators are established. With a proper choice of regularization parameters and penalty functions, the proposed variable selection procedures are shown to perform as well as an oracle estimator. A robust standard error formula is derived using a sandwich formula and is empirically tested. Local polynomial regression techniques are used to estimate the baseline function in the semiparametric model.

KEY WORDS: Local polynomial regression; Partial linear model; Penalized least squares; Profile least squares; Smoothly clipped absolute deviation.

1. INTRODUCTION

Longitudinal data are often highly unbalanced because the data were collected at irregular and possibly subject-specific time points. Due to their unbalanced nature, it is difficult to directly apply traditional multivariate regression techniques for analyzing such data. Various parametric models and statistical tools have been developed for longitudinal data analysis (see, e.g., Diggle, Heagerty, Liang, and Zeger 2002; Verbeke and Molenberghs 2000). Parametric models are very useful for analyzing longitudinal data and for providing a parsimonious description of the relationship between the response variable and its covariates. But they are used at the risk of introducing modeling biases. To relax the assumptions on parametric forms, various nonparametric models, including varying coefficient models, functional linear models, and their extensions, have been proposed for longitudinal data analysis (see, e.g., Hoover, Rice, Wu, and Yang 1998; Wu, Chiang, and Hoover 1998; Fan and Zhang 2000; Martinussen and Scheike 2001; Chiang, Rice, and Wu 2001; Huang, Wu, and Zhou 2002; references therein). Although parametric models may be restrictive for some applications, nonparametric models may be too flexible to make concise conclusions in comparison with parsimonious parametric models. Semiparametric models are good compromises and retain nice features of both the parametric and nonparametric models. In this article, we study the semiparametric model

$$y(t) = \alpha(t) + \boldsymbol{\beta}^T \mathbf{x}(t) + \varepsilon(t), \tag{1}$$

where $y(t)$ is the response variable and $\mathbf{x}(t)$ is the $d \times 1$ covariate vector at time t, $\alpha(t)$ is an unspecified baseline function of t, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients, and $\varepsilon(t)$ is a mean-0 stochastic process. Model (1) does not require data analysts to parameterize the baseline function which may be difficult in practice. It keeps the flexibility of the nonparametric models for the baseline function, while maintaining the explanatory power of parametric models. Therefore, model (1) and its variations have been receiving increasing attention recently. Zeger and Diggle (1994) proposed an iterative algorithm to estimate $\alpha(t)$ and $\boldsymbol{\beta}$ by the backfitting method. Extending the idea of partial residual (Speckman 1988) for partial linear models, Moyeed and Diggle (1994) proposed an estimation procedure that improves Zeger and Diggle's procedure. Lin and Carroll (2001b) considered a marginal model for clustered data through specification of the mean and variance function and suggested using a kernel generalized estimation equation (GEE) for their marginal model. Martinussen and Scheike (1999, 2001), Cheng and Wei (2000), and Lin and Ying (2001) proposed estimation procedures under the formation of point processes. All aforementioned works focus on estimation of the baseline function and the regression coefficients. Issues related to model selection, including selection smoothing parameter for the baseline function and variable selection for x-variables, have not been addressed to date.

In this article, we first propose two simple, reliable, and effective estimation procedures for regression coefficients. The difference-based estimator (DBE) of $\boldsymbol{\beta}$ provides a simple and good initial estimate of $\boldsymbol{\beta}$ and does not rely on any smoothing techniques. The estimator is then refined by the newly proposed profile least squares estimator, which depends on a choice of smoothing parameter. This can be selected relatively easily. With a good initial estimate of $\boldsymbol{\beta}$, such as the DBE, model (1) becomes a univariate nonparametric regression problem. Thus a wealth of bandwidth selection techniques for univariate nonparametric regression can be used. The asymptotic normality of the profile least squares is established and a consistent standard error (SE) formula is derived using the sandwich formula. Our methods are very effective in the class of working independent estimators. Our study shows that our newly proposed estimators, including the DBE, outperforms the proposal of Lin and Ying. In this article, we propose an estimation procedure for

Jianqing Fan is Professor of Statistics, Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544 (E-mail: *jqfan@Princeton.edu*). Runze Li is Assistant Professor, Department of Statistics, The Pennsylvania State University, University Park, PA 16802-2111 (E-mail: *rli@stat.psu.edu*). Fan's research was supported in part by National Institute of Health grant R01 HL69720. Li's research was supported by National Science Foundation grants DMS-01-02505 and DMS-03-48869 and a National Institute on Drug Abuse grant 1-P50-DA10075. The authors thank the associate editor and the referees for constructive comments that substantially improved an earlier draft, and the MACS study for the data used in Section 4.4.

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/016214504000001060

the baseline function, $\alpha(t)$, in model (1) using local polynomial regression.

Like parametric regression models, variable selection is important in the semiparametric model (1). The number of variables in (1) can easily be large when nonlinear terms and interactions between covariates are introduced to reduce possible modeling biases. It is common in practice to include only important variables in the model, to enhance predictability and to give a parsimonious description between the response and the covariates. Stepwise deletion and best-subset variable selection may be extended to semiparametric regression analysis, but pose greater challenges for implementation, such as the choice of smoothing parameter for each submodel. Furthermore, as analyzed by Breiman (1996), they suffer from several drawbacks, including a lack of stability. Although they are useful in practice, the stepwise deletion and the best-subset methods ignore stochastic errors inherited in the stage of variable selection. Hence their theoretical properties are somewhat hard to understand, and the sampling properties of the resulting estimates are difficult to establish, even in the classical linear model. Consequently, the confidence intervals based on these methods may not necessarily be valid.

Nonconcave penalized likelihood approaches to select significant variables for parametric regression models have been proposed (Fan and Li 2001). These are useful extensions of the work by Tibshirani (1996). With a suitable choice of penalty functions (Fan and Li 2001), the resulting estimates of the nonconcave penalized likelihood approaches have an oracle property. This encourages us to extend the methodology to semiparametric regression analysis for longitudinal data. Semiparametric structure poses new challenges for the procedure. Because the baseline function has not yet been parameterized, we introduce a new quadratic loss between the observed data and the theoretical model that involves only the unknown parameter $\boldsymbol{\beta}$. This permits us to extend the penalized least squares technique to the semiparametric model (1). The simultaneous selection of variables and estimation of unknown parameters allows us to construct a confidence interval for the coefficients. It also enables us to establish rates of convergence for the resulting estimator. Further, we show that, with a proper choice of regularization parameters and penalty functions, the proposed procedure performs as well as an oracle estimator. The theoretical result has also been empirically tested. In addition, with the aid of local quadratic approximations to the penalty functions, an iterative ridge regression algorithm is used to find the solution of the penalized least squares, and a robust standard error formula for estimated coefficients of nonzero components is derived using a sandwich formula. The SE formula is empirically tested; it performs very well with moderate-sized samples.

The rest of the article is organized as follows. In Section 2 we propose two new estimation procedures for regression coefficients in the parametric component and establish asymptotic normality of one of the proposed estimators. In Section 3 we propose a penalized quadratic loss procedure for selecting significant variables in model (1). We investigate finite-sample performance of the proposed procedures in Section 4. We further illustrate the proposed methodology through an analysis of a subset of data from the Multi-Center AIDS cohort study. Technical proofs are relegated to the Appendix.

2. NEW ESTIMATION PROCEDURES

Suppose that we have a sample of n subjects. For the ith subject, the response variable $y_i(t)$ and the covariate vector $\mathbf{x}_i(t)$ are collected at time points $t = t_{i1}, \ldots, t_{iJ_i}$, where J_i is the total number of observations on the ith subject. Consider the marginal model

$$E\{y(t_{ij})|\mathbf{x}_i(t_{ij})\} = \alpha(t_{ij}) + \boldsymbol{\beta}^T\mathbf{x}_i(t_{ij}) \tag{2}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, J_i$. Denote

$$\mathbf{y}_i = \big(y_i(t_{i1}), \ldots, y_i(t_{iJ_i})\big)^T,$$
$$\mathbf{X}_i = \big(\mathbf{x}_i(t_{i1}), \ldots, \mathbf{x}_i(t_{iJ_i})\big)^T,$$

and

$$\boldsymbol{\alpha}_i = \big(\alpha(t_{i1}), \ldots, \alpha(t_{iJ_i})\big)^T.$$

Thus a weighted least squares fit is obtained by minimizing the weighted least squares function

$$\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta})^T\mathbf{W}_i(\mathbf{y}_i - \boldsymbol{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta}) \tag{3}$$

with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where $\mathbf{W}_i$ is a $J_i \times J_i$ weight matrix, called a working covariance matrix. The most efficient choice of the working covariance matrix is the inverse of the true covariance matrix. Following Lin and Ying (2001), we focus on the situation in which the weight matrix $\mathbf{W}_i$ is a diagonal matrix (working independence) with diagonal elements possibly data-dependent. Although this $\mathbf{W}_i$ may not be correct, misspecification of the working matrix will not affect the consistency of the resulting estimate, only the efficiency.

In what follows, we first introduce notation for counting process and give a brief review of Lin and Ying's method, and then propose two new estimation procedures for regression coefficients: DBE and profile least squares estimator. We further discuss the issue of bandwidth selection. The DBE serves as a good initial estimator for bandwidth selection. The sampling property of the profile least squares estimator will be established. Finally, we present estimation of the baseline function and its asymptotic properties in Section 2.6.

2.1 Counting Process Approaches

Martinussen and Scheike (1999) and Lin and Ying (2001) introduced the counting process technique to the estimation scheme. The time points where the observations on the ith subject are made are characterized by the counting process

$$N_i(t) \equiv \sum_{j=1}^{J_i} I(t_{ij} \le t),$$

where $I(\cdot)$ is the indicator function. Both $y(t)$ and time-varying covariates $\mathbf{x}(t)$ were observed at the jump points of $N_i(t)$. The observation times are considered as realizations from an arbitrary counting process that is censored at the end of follow-up. Specifically, $N_i(t) = N_i^*(t \wedge c_i)$, where $N_i^*(t)$ is a counting process in discrete or continuous time, c_i is the follow-up or censoring time, and $a \wedge b = \min(a, b)$. The censoring time c_i is

allowed to depend on the vector of covariates $\mathbf{x}_i(\cdot)$ in an arbitrary manner. In this article we assume that the censoring mechanism is noninformative in the sense that

$$E\{y_i(t)|\mathbf{x}_i(t), c_i \geq t\} = E\{y_i(t)|\mathbf{x}_i(t)\}.$$

Using the foregoing counting process notation, the least squares problem (3) can be written as

$$\frac{1}{2}\sum_{i=1}^{n}\int_0^{+\infty} w(t)\big\{y_i(t) - \alpha(t) - \boldsymbol{\beta}^T\mathbf{x}_i(t)\big\}^2\, dN_i(t) \quad (4)$$

for a working independence $\mathbf{W}_i$ with diagonal elements $w(t_{ij})$.

The essence of Lin and Ying's approach is to first estimate the function $\alpha(t)$ and then apply a substitution technique. Lin and Ying (2001) considered two situations, depending on whether the potential observation times are independent of the covariates $\mathbf{x}(t)$. When the potential observation times depend on the covariates, they assumed that

$$E\big\{dN_i^*(t)|\mathbf{x}_i(t), y_i(t), c_i \geq t\big\} = \exp\{\boldsymbol{\gamma}'\mathbf{x}_i(t)\}\, d\Lambda(t), \quad i = 1, \ldots, n, \quad (5)$$

where $\boldsymbol{\gamma}$ is a vector of unknown parameter and $\Lambda(\cdot)$ is an arbitrary nondecreasing function. When $\boldsymbol{\gamma} = 0$, the observation times are independent of the covariates. Denote

$$\bar{\mathbf{x}}(t, \boldsymbol{\gamma}) = \frac{\sum_{i=1}^n \xi_i(t)\exp\{\boldsymbol{\gamma}^T\mathbf{x}_i(t)\}\mathbf{x}_i(t)}{\sum_{i=1}^n \xi_i(t)\exp\{\boldsymbol{\gamma}^T\mathbf{x}_i(t)\}}$$

and

$$\bar{y}(t, \boldsymbol{\gamma}) = \frac{\sum_{i=1}^n \xi_i(t)\exp\{\boldsymbol{\gamma}^T\mathbf{x}_i(t)\}y_i(t)}{\sum_{i=1}^n \xi_i(t)\exp\{\boldsymbol{\gamma}^T\mathbf{x}_i(t)\}},$$

where $\xi_i(t) = I(c_i \geq t)$. Lin and Ying (2001) estimated the baseline function by

$$\hat{\alpha}(t; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \bar{y}(t, \boldsymbol{\gamma}) - \boldsymbol{\beta}^T\bar{\mathbf{x}}(t, \boldsymbol{\gamma}). \quad (6)$$

In practice, $\bar{\mathbf{x}}(t, \boldsymbol{\gamma})$ and $\bar{y}(t, \boldsymbol{\gamma})$ may not be evaluable because $\mathbf{x}(t)$ and $y(t)$ are not fully observed until the censoring time c_i. Lin and Ying (2001) replaced $\mathbf{x}(t)$ and $y(t)$ by their corresponding values at the nearest time when their values are observed. Thus, substituting $\alpha(t)$ by $\hat{\alpha}(t; \boldsymbol{\beta}, \boldsymbol{\gamma})$ in (4) yields

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2}\sum_{i=1}^{n}\int_0^{+\infty} w(t)\big[\{y_i(t) - \bar{y}(t, \boldsymbol{\gamma})\} - \boldsymbol{\beta}^T\{\mathbf{x}_i(t) - \bar{\mathbf{x}}(t, \boldsymbol{\gamma})\}\big]^2\, dN_i(t). \quad (7)$$

The parameter $\boldsymbol{\gamma}$ can be consistently estimated by its moment estimator $\hat{\boldsymbol{\gamma}}$, the solution to

$$\sum_{i=1}^{n}\int_0^{+\infty}\{\mathbf{x}_i(t) - \bar{\mathbf{x}}(t, \boldsymbol{\gamma})\}\, dN_i(t) = 0.$$

Substituting $\hat{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma}$ in (7), we can derive an explicit form for $\hat{\boldsymbol{\beta}}$. The standard error of the resulting estimate $\hat{\boldsymbol{\beta}}$ can be constructed via the corresponding sandwich formula, as proposed by Lin and Ying (2001).

2.2 A Difference-Based Method

An advantage of Lin and Ying's approach is its simplicity. It does not involve any smoothing parameter. Lin and Ying (2001) realized that efficiency can be gained by incorporating smoothing techniques into the baseline estimation, and Lin and Carroll (2001a) argued further that the efficiency gain can be infinite for certain specific cases, such as the partial linear model (Speckman 1988; Severini and Staniswalis 1994; Carroll, Fan, Gijbels, and Wand 1997).

The weighted least squares problem (7) requires that the processes $y_i(t)$ and $\mathbf{x}_i(t)$ be fully observable until the censoring time c_i. This is an unrealistic assumption. Lin and Ying (2001) replaced the processes by their corresponding values at the nearest time where their values are observed. Although this helps practical implementations of the procedure, the method introduces biases due to the nearest-neighbor approximations. Further, because for each subject, the spaces among observation times $\{t_{ij}, j = 1, \ldots, J_i\}$ do not tend to 0 even when the sample size tends to infinity, the approximation biases cannot always be negligible in practice. The approach may cause some problems in asymptotic theory.

To avoid the forementioned two problems, unbounded loss of efficiency and nonnegligible biases due to approximations, and to maintain the simplicity of Lin and Ying's approach, we propose the following simple method from the partial linear model (Fan and Huang 2001). As in Lin and Ying's approach, we ignore the within-subject correlation and use the working independence covariance matrix, for simplicity of presentation and implementation. Dropping the subscript j, the observed data

$$\big\{\big(t_{ij}, \mathbf{x}(t_{ij})^T, \mathbf{y}(t_{ij})\big), j = 1, \ldots, J_i, i = 1, \ldots, n\big\}$$

can be expressed in the vector notation as

$$\big\{(t_i, \mathbf{x}_i^T, \mathbf{y}_i), i = 1, \ldots, n^*\big\} \quad \text{with} \quad n^* = \sum_{i=1}^{n} J_i,$$

ordered according to time $\{t_{ij}\}$. By the marginal model (2), it follows that

$$y_i = \alpha(t_i) + \boldsymbol{\beta}^T\mathbf{x}_i + \varepsilon_i \quad \text{with} \quad E(\varepsilon_i|\mathbf{x}_i) = 0. \quad (8)$$

First, observe that

$$y_{i+1} - y_i = \alpha(t_{i+1}) - \alpha(t_i) + \boldsymbol{\beta}^T(\mathbf{x}_{i+1} - \mathbf{x}_i) + e_i, \quad i = 1, \ldots, n^* - 1, \quad (9)$$

where stochastic error $e_i = \varepsilon_{i+1} - \varepsilon_i$. Under some mild conditions, the spacing between t_i and t_{i+1} is of order $O(1/n)$. Hence, the term $\alpha(t_{i+1}) - \alpha(t_i)$ in (9) is negligible. The least squares approach can be used to estimate the parameter $\boldsymbol{\beta}$. The method can be further improved by fitting the linear model

$$y_{i+1} - y_i = \alpha_0 + \alpha_1(t_{i+1} - t_i) + \boldsymbol{\beta}^T(\mathbf{x}_{i+1} - \mathbf{x}_i) + e_i, \quad i = 1, \ldots, n^* - 1. \quad (10)$$

In the presence of ties among observation times t_{ij}, this linear approximation to $\alpha(t_{i+1}) - \alpha(t_i)$ still holds. The linear term $\alpha_0 + \alpha_1(t_{i+1} - t_i)$ is introduced to correct for the finite-sample bias when the gap of the spacing is wide. This can occur at the tails of the distribution of the time $\{t_i, i = 1, \ldots, n^*\}$. Fitting model (10) yields an estimate of $\boldsymbol{\beta}$. For simplicity, we call this

method the DBE. Our limited experience shows that the DBE performs quite well; see the numerical comparison in Section 4. The purpose of the DBE is to obtain a quick and reliable initial estimate of $\boldsymbol{\beta}$. The DBE is also used in the bandwidth selection of the profile least squares method. We do not pursue the sampling property of the DBE any further herein.

2.3 Profile Least Squares Approach

For a given $\boldsymbol{\beta}$, let $y^*(t) \equiv y(t) - \boldsymbol{\beta}^T \mathbf{x}(t)$. Then the model (1) can be written as

$$y^*(t) = \alpha(t) + \varepsilon(t). \tag{11}$$

This is a nonparametric regression problem. Thus we can use a nonparametric regression technique to estimate $\alpha(t)$. We focus only on the local linear regression technique (Fan 1992). For t in a neighborhood of t_0, it follows by the Taylor expansion that

$$\alpha(t) \approx \alpha(t_0) + \alpha'(t_0)(t - t_0) \equiv a_0 + a_1(t - t_0).$$

Let $K(\cdot)$ be a kernel function and let h be a bandwidth. The local linear fit is to find $(\hat{a}_0, \hat{a}_1)$ minimizing

$$\sum_{i=1}^{n} \sum_{j=1}^{J_i} \{y_i^*(t_{ij}) - a_0 - a_1(t_{ij} - t_0)\}^2 w(t_{ij}) K_h(t_{ij} - t_0), \tag{12}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$. Here the weight function, $w(t_{ij})$, serves a similar purpose to that in (4). The local linear estimate is simply $\hat{\alpha}(t_0; \boldsymbol{\beta}) = \hat{a}_0$. To improve the efficiency, Wang, Carroll, and Lin (2004) recently proposed an innovative iterated profile likelihood method for estimating $\boldsymbol{\beta}$ when the true covariance function is given. The procedure is very computationally intensive. In contrast, implementing our procedure is very simple.

Before we proceed further, we introduce some notation. Let $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)$, $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$, and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_n^T)^T$. Then model (8) can be written as

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{13}$$

where $\boldsymbol{\varepsilon}$ is the vector of stochastic errors. It is well known that the local linear fit is linear in $y_i^*(t_{ij})$ (Fan 1992). Thus the estimate of $\alpha(t)$ is linear in $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. Hence, the estimate for the vector $\boldsymbol{\alpha}$ can be expressed as $\hat{\boldsymbol{\alpha}} = \mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. The matrix $\mathbf{S}$, usually called a smoothing matrix of the local linear smoother, depends only on the observation times $\{t_{ij}, i = 1, \ldots, n, j = 1, \ldots, J_i\}$ and the amount of smoothing. Substituting $\hat{\boldsymbol{\alpha}}$ into (13), we obtain

$$(\mathbf{I} - \mathbf{S})\mathbf{y} = (\mathbf{I} - \mathbf{S})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{14}$$

where $\mathbf{I}$ is the identity matrix of order n^*. Applying weighted least squares to the linear model (14), we obtain

$$\hat{\boldsymbol{\beta}} = \{\mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{y}, \tag{15}$$

where $\mathbf{W}$ is a diagonal matrix with the diagonal elements as $w(t_{ij})$s. The form of $\hat{\boldsymbol{\beta}}$ is similar to that of Speckman (1988) for partial linear models with independent observations. Speckman's work is motivated by partial residuals; Moyeed and Diggle (1994) extended the idea of partial residuals to longitudinal data. Our approach is motivated from the principle of profile likelihood, which is also applicable to the penalized least squares in Section 3.1. However, the asymptotic normality of Moyeed and Diggle's estimator has not yet been shown. Furthermore, our current statistical setting is different from theirs. The estimator in (15) is called the *profile least squares estimator*. The profile least squares estimator for the nonparametric component is simply $\alpha(\cdot; \hat{\boldsymbol{\beta}})$.

From (15), we can derive an estimate for covariance matrix of $\hat{\boldsymbol{\beta}}$,

$$\operatorname{cov}\{\hat{\boldsymbol{\beta}} | t_{ij}, \mathbf{x}_i(t_{ij})\} = \mathbf{D}^{-1}\mathbf{V}\mathbf{D}^{-1}, \tag{16}$$

where $\mathbf{D} = \mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{X}$ and $\mathbf{V} = \operatorname{cov}\{\mathbf{X}^T \times (\mathbf{I} - \mathbf{S})^T\mathbf{W}\boldsymbol{\varepsilon}\}$, which is linear in $\boldsymbol{\varepsilon}$ and can be easily estimated by

$$\hat{\mathbf{V}} = \mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}\mathbf{C}\mathbf{W}^T(\mathbf{I} - \mathbf{S})\mathbf{X}, \tag{17}$$

where $\mathbf{C} = \operatorname{diag}\{\hat{\boldsymbol{\varepsilon}}_1\hat{\boldsymbol{\varepsilon}}_1^T, \ldots, \hat{\boldsymbol{\varepsilon}}_n\hat{\boldsymbol{\varepsilon}}_n^T\}$ and $\hat{\boldsymbol{\varepsilon}}_i$ is the residual vector for the ith subject.

2.4 Bandwidth Selection

A few questions arise in the practical implementation of the foregoing procedure. The first question is how to select the bandwidth so that $\boldsymbol{\beta}$ can be estimated well. The variance inherited in the nonparametric estimate $\hat{\alpha}(\cdot; \boldsymbol{\beta})$ does not usually cause a problem, because it will be averaged out in the parametric least squares fitting. Thus, a general strategy is to select a small bandwidth so that the bias is negligible. In fact, the procedure in (10), with even indices, can be considered as the profile least squares estimate using the local average of two data points as a nonparametric estimator,

$$\hat{\alpha}(t_{2i+1}) = 2^{-1}\{(y_{2i+1} - \boldsymbol{\beta}^T\mathbf{x}_{2i+1}) + (y_{2i} - \boldsymbol{\beta}^T\mathbf{x}_{2i})\},$$

where the notation is the same as that in Section 2.2. This provides stark evidence that for a large range of the smoothing parameter, as long as it is small enough, the resulting profile least squares estimate is root-n consistent. However, the efficiency for estimating $\boldsymbol{\beta}$ can be affected by the choice of bandwidth.

Using (15) and noting that $\hat{\boldsymbol{\alpha}} = \mathbf{S}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, we see that $\hat{\boldsymbol{\alpha}}$ is linear in $\mathbf{y}$. Data-driven methods, such as cross-validation (CV) and generalized cross-validation (GCV), can be used to select the bandwidth. But this will be computationally expensive. To avoid expensive computations and possibly unstable numerical implementation, our practical choice of bandwidth is to use the DBE to get an estimate $\hat{\boldsymbol{\beta}}_{\mathrm{DBE}}$. Substituting it into (11), we have a univariate nonparametric regression problem. Let $\hat{h}$ be the bandwidth that is appropriate for this problem. This can be obtained by a subjective choice via visualization, or by a data-driven procedure such as a substitution method, or a CV method. Use this $\hat{h}$ for the profile least squares estimate. From nonparametric theory, this optimal choice of bandwidth is of order $h_n = bn^{-1/5}$. Theorem 1 endorses this choice.

The function $\alpha(\cdot)$ cannot be estimated well at some tail of the observation times because of sparsity. Including its estimates at these regions in (14) can have an adverse effect on the estimation of $\boldsymbol{\beta}$. To avoid this, we can simply exclude 5%, say, of the data at the tail in the analysis.

2.5 Asymptotic Result

It is well known that asymptotic theory depends on the formulation of how the data were collected. For longitudinal data, there are many possible formulations. For example, in a series of works by Wu and collaborators (see, e.g., Hoover et al. 1998; Wu et al. 1998), it was assumed that time points $\{t_{ij}\}$ are a random sample from a certain population. Diggle et al. (2002) used a different formulation. Lin and Ying (2001) assumed the counting process $N_i(\cdot)$ to be a random sample from a certain population. To be consistent with the simulation models used in this article, we adopt the formulation of Lin and Ying (2001). Other formulations can also be accommodated, with similar results obtained.

In the sequel, we use $\alpha_0(\cdot)$ and $\boldsymbol{\beta}_0$ to denote the true parameters. Because the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}(\cdot)$ are linear in the response variable, we can directly demonstrate the asymptotic normality by computing asymptotic counterparts for various terms in (15). This will bury much good intuition in the detailed asymptotic calculations. Instead, in the Appendix we provide a much simpler idea for establishing the asymptotic normality of the estimators. When the weight function $w(t)$ is data-dependent, we assume that it tends to a deterministic function in probability. Therefore, for simplicity, assume that $w(t)$ is a deterministic function of t.

Set

$$\hat{\boldsymbol{\Sigma}}_n = n^{-1}\sum_{i=1}^{n}\int_0^\infty \{\mathbf{x}_i(t) - E\mathbf{x}_i(t)\}^{\otimes 2} w(t)\, dN_i(t),$$

where $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$, and

$$\hat{\boldsymbol{\xi}}_n = n^{-1}\sum_{i=1}^{n}\int_0^\infty \{\mathbf{x}_i(t) - E\mathbf{x}_i(t)\}\varepsilon_i(t) w(t)\, dN_i(t).$$

Let

$$\mathbf{A} = E\int_0^\infty \{\mathbf{x}(t) - E\mathbf{x}(t)\}^{\otimes 2} w(t)\, dN(t)$$

and

$$\mathbf{B} = E\left\{\int_0^\infty \{\mathbf{x}(t) - E\mathbf{x}(t)\}\varepsilon(t) w(t)\, dN(t)\right\}^{\otimes 2}.$$

We then have the following result.

Theorem 1. Suppose that $w(\cdot)$ is continuous, the J_i's are bounded, and the matrices $\mathbf{A}$ and $\mathbf{B}$ exist. If $\mathbf{A}$ is finite positive definite and $h_n = bn^{-a}$ for $1/8 < a < 1/2$, then, as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \sqrt{n}\hat{\boldsymbol{\Sigma}}_n^{-1}\hat{\boldsymbol{\xi}}_n + o_P(1) \xrightarrow{\mathcal{L}} \mathrm{N}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}),$$

where n is the number of subjects.

Theorem 1 gives the asymptotic representation for the profile least squares estimator. This allows one to establish asymptotic normality under a different formulation. The asymptotic normality easily follows from the asymptotic representation.

It is intuitively clear that the matrix $\mathbf{D}$ is a consistent estimator of $\mathbf{A}$ and the matrix $\hat{\mathbf{V}}$ is a consistent estimator of $\mathbf{B}$. Thus it can be shown that $\mathbf{D}^{-1}\hat{\mathbf{V}}\mathbf{D}^{-1}$ is a consistent estimator of $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$. This will also be demonstrated in our empirical studies.

2.6 Estimation of Baseline Function

As discussed in Section 2.3, the baseline function $\alpha(\cdot)$ can be estimated by smoothing the partial residuals $\{(t_i, y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}), i = 1, \ldots, n^*\}$ using a local linear fit. This results in a nonparametric fit, $\hat{\alpha}(\cdot; \hat{\boldsymbol{\beta}})$. Because the rate of convergence for $\hat{\boldsymbol{\beta}}$ is faster than that for the nonparametric estimator, $\hat{\boldsymbol{\beta}}$ can either be a profile least squares estimator or a DBE. The latter is much easier to obtain, whereas the former may have better performance in some situations. Because the errors in estimation $\boldsymbol{\beta}$ are negligible in the nonparametric estimation of α, the value of $\boldsymbol{\beta}$ can be considered as known. Using the fact that $\hat{\boldsymbol{\alpha}} = \mathbf{S}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ is linear in $\mathbf{y}$ (ignoring the variability from $\hat{\boldsymbol{\beta}}$), the standard error for $\hat{\boldsymbol{\alpha}}$ can be estimated as $\mathbf{S}\mathbf{C}\mathbf{S}^T$; see (16) for a similar expression.

For a given shrinking neighborhood $t_0 \pm h_n$ around a given time t_0, the likelihood of getting two or more data points from the same subject is negligible as $h_n \to 0$. Hence, the problem is the same as the nonparametric regression for independent data. Let $\lambda(t)$ be the intensity function of the process $N(t)$, where $\lambda(t)$ may depend on covariates $\mathbf{x}$. For ease of presentation, we suppress $\mathbf{x}$ in the notation of λ. Then it can be shown that, under some mild regularity conditions,

$$\sqrt{nh_n}\left\{\hat{\alpha}(t_0; \hat{\boldsymbol{\beta}}) - \alpha_0(t_0) - \frac{1}{2}\alpha_0''(t_0)\int u^2 K(u)\, du\, h_n^2\right\} \xrightarrow{\mathcal{L}} \mathrm{N}\big(0, \sigma^2(t_0)\big), \quad (18)$$

where

$$\sigma^2(t_0) = \frac{\mathrm{var}\{\varepsilon(t_0)\}}{\lambda(t_0)}\int K(u)^2\, du.$$

We omit the details of the proof. The bias and variance expressions are similar to those of Fan (1992). We outline the key steps for deriving the bias and variance in Appendix A.2.

3. VARIABLE SELECTION

Model selection is an indispensable tool for statistical data analysis. However, it has rarely been studied in the semiparametric context. In this section we introduce the penalized least squares approach. The first step is to eliminate the nuisance parameters, the nonparametric function $\alpha(\cdot)$. Let $\ell(\boldsymbol{\beta})$ be the weighted least squares that we would like to minimize. It can be (7) or the weighted quadratic loss induced by model (14), and reflects a semiparametric method that we would like to use.

3.1 Penalized Weighted Least Squares

Suppose that $\mathbf{x}_i$ consists of d variables, some of which are not statistically significant. A penalized least squares takes the form

$$\mathcal{L}(\boldsymbol{\beta}) \equiv \ell(\boldsymbol{\beta}) + n\sum_{j=1}^{d}\lambda_j p_j(|\beta_j|), \quad (19)$$

where the $p_j(\cdot)$'s are penalty functions and the λ_j's are tuning parameters, which control the model complexity and can be selected by some data-driven methods, such as CV and GCV. By minimizing (19), with special construction of the penalty function, some coefficients are estimated as 0, which deletes the

corresponding variables, whereas others are not. Thus the procedure selects variables and estimates coefficients simultaneously. The resulting estimate is called a penalized least squares estimate.

The penalized least squares (19) can be obtained from the penalized quadratic loss of the semiparametric model (2), using the profiling technique. For example, starting from the quadratic loss (3) and adding the penalty term $n\sum_{j=1}^{d}\lambda_j \times p_j(|\beta_j|)$, we obtain the penalized quadratic loss

$$\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\alpha}_i-X_i\boldsymbol{\beta})^T\mathbf{W}_i(\mathbf{y}_i-\boldsymbol{\alpha}_i-X_i\boldsymbol{\beta})+n\sum_{j=1}^{d}\lambda_j p_j(|\beta_j|).$$

After eliminating the nuisance function $\alpha(\cdot)$ using the profiling technique in Section 2.3 [see (14)], we obtain the penalized least squares

$$\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+n\sum_{j=1}^{d}\lambda_j p_j(|\beta_j|).$$

The penalty functions $p_j(\cdot)$ and the regularization parameters λ_j are not necessarily the same for all j. This allows us to incorporate prior information for the unknown coefficients by using different penalty functions or taking different values of λ_j. For instance, we may wish to keep important predictors in linear regression models, and hence do not want to penalize their coefficients. For ease of presentation, we denote $\lambda_j p_j(\cdot)$ by $p_{\lambda_j}(\cdot)$.

Many penalty functions, such as the family of L_q penalty ($q \geq 0$), have been used for penalized least squares and penalized likelihood in various parametric models. For instance, $q = 0$ corresponds to the entropy penalty, L_1 penalty results in the LASSO proposed by Tibshirani (1996), and bridge regression (Frank and Friedman 1993) corresponds to $0 < q < 1$. Antoniadis and Fan (2001) and Fan and Li (2001) have provided various insights into how a penalty function should be chosen. They advocated that a good penalty function should yield an estimator with the following three properties: *unbiasedness* for a large true coefficient, to avoid unnecessary estimation bias; *sparsity* (estimating a small coefficient as 0), to reduce model complexity; and *continuity*, to avoid unnecessary variation in model prediction. Necessary conditions were given by Antoniadis and Fan (2001). None of the L_q penalties produces estimates that simultaneously satisfy the foregoing three properties. A simple penalty function, which results in an estimator with the three desired properties, is the smoothly clipped absolute deviation (SCAD) penalty. Its first derivative is defined by

$$p'_\lambda(\beta)=\lambda\left\{I(\beta\leq\lambda)+\frac{(a\lambda-\beta)_+}{(a-1)\lambda}I(\beta>\lambda)\right\}$$

for some $a > 2$ and $\beta > 0$,

and $p_\lambda(0) = 0$. For simplicity of presentation, we use the term SCAD to denote all procedures using the SCAD penalty. The SCAD involves two unknown parameters, λ and a. Fan and Li (2001) suggested using $a = 3.7$ from a Bayesian standpoint; hence we use this value throughout the rest of the article.

3.2 Iterated Ridge Regression

Finding the solution of the penalized least squares of (19) is challenging, because the penalty function $p_{\lambda_j}(|\beta_j|)$, such as the L_q penalty ($0 < q \leq 1$) or the SCAD penalty, is irregular at the origin and may not have a second derivative at some points. Following Fan and Li (2001), we locally approximate the penalty functions by quadratic functions as follows. Given an initial value $\boldsymbol{\beta}^{(0)}$ that is close to the minimizer of (19), when $|\beta_j^{(0)}| \geq \eta$ (a prescribed value), the penalty $p_{\lambda_j}(|\beta_j|)$ can be locally approximated by the quadratic function as

$$\left[p_{\lambda_j}(|\beta_j|)\right]' = p'_{\lambda_j}(|\beta_j|)\operatorname{sgn}(\beta_j) \approx \left\{p'_{\lambda_j}\left(|\beta_j^{(0)}|\right)\Big/|\beta_j^{(0)}|\right\}\beta_j.$$

With the local quadratic approximation, the Newton–Raphson algorithm can be implemented directly for minimizing $\mathcal{L}(\boldsymbol{\beta})$, (19). Furthermore, the Newton–Raphson algorithm is indeed an iterative ridge regression algorithm. For instance, we update the solution of the penalized profile least squares by

$$\boldsymbol{\beta}^{(1)} = \left[\mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{X}+n\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta}^{(0)}\right)\right]^{-1} \times \mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{y},$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^0) = \operatorname{diag}\left\{p'_{\lambda_1}\left(|\beta_1^{(0)}|\right)\Big/|\beta_1^{(0)}|, \ldots, p'_{\lambda_d}\left(|\beta_d^{(0)}|\right)\Big/|\beta_d^{(0)}|\right\}.$$

When there is a component $|\beta_j^{(0)}| < \eta$, it is set to 0. In the implementation, we take the unpenalized profile least squares estimator as an initial value and iteratively update $\boldsymbol{\beta}^{(1)}$.

Similar to the penalized profile least squares method, the foregoing thresholding-shrinkage idea can also be applied to Lin and Ying's estimator. The penalized least squares estimator derived by using Lin and Ying's approach can be obtained by iteratively updating

$$\boldsymbol{\beta}^{(1)} = \left[\frac{1}{n}\sum_{i=1}^{n}\int_0^{+\infty} w(t)\{\mathbf{x}_i(t)-\bar{\mathbf{x}}(t,\hat{\boldsymbol{\gamma}})\}^{\otimes 2}\,dN_i(t) + \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta}^{(0)}\right)\right]^{-1} \times \left[\frac{1}{n}\sum_{i=1}^{n}\int_0^{+\infty} w(t)\{\mathbf{x}_i(t)-\bar{\mathbf{x}}(t;\hat{\boldsymbol{\gamma}})\} \times \{y_i(t)-\bar{y}(t;\hat{\boldsymbol{\gamma}})\}\,dN_i(t)\right].$$

When the algorithm converges, the estimator satisfies the condition

$$\frac{\partial \ell(\hat{\boldsymbol{\beta}})}{\partial \beta_j} + np'_{\lambda_j}\left(|\hat{\beta}_j|\right)\operatorname{sgn}(\hat{\beta}_j) = 0, \tag{20}$$

the penalized weighted least squares equation for nonzero components.

With the local quadratic approximation, the iterative ridge regression is similar to the Newton–Raphson algorithm. Thus a robust empirical SE formula for the estimated coefficients can be derived from the iterative ridge regression. In other words, the covariance matrix of $\hat{\boldsymbol{\beta}}$ can be consistently estimated

by $n^{-1}\{\hat{\mathbf{D}} + \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}})\}^{-1}\hat{\mathbf{V}}\{\hat{\mathbf{D}} + \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}})\}^{-1}$ for the nonzero component, where

$$\hat{\mathbf{D}} = \frac{1}{n}\mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{X}$$

and

$$\hat{\mathbf{V}} = \frac{1}{n}\widehat{\operatorname{cov}}\{\mathbf{X}^T(\mathbf{I}-\mathbf{S})^T\mathbf{W}(\mathbf{I}-\mathbf{S})\mathbf{y}\}$$

for the penalized profile least squares estimator; see (17) for an explicit form of $\hat{\mathbf{V}}$. For the penalized least squares estimator derived by using Lin and Ying's approach, $\hat{\mathbf{D}}$ and $\hat{\mathbf{V}}$ are as defined in their article.

3.3 Choice of Regularization Parameters

To implement the methods described in the previous sections, it is desirable to have an automatic data-driven method for estimating the tuning parameters $\lambda_1, \ldots, \lambda_d$. For linear estimators (in terms of response variable) in nonparametric regression, there is much literature on choosing a smoothing parameter. The resulting estimators of the penalized weighted least squares are not linear, but, with the aid of local quadratic approximation, they are approximately linear. Therefore various smoothing parameter selectors, such as CV or GCV, can be used. Here we estimate $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ by minimizing an approximate GCV score. Recall that the iterative ridge regression algorithm is used to obtain the penalized weighted least squares estimator. By some straightforward calculation, the effective number of parameters in the last step of the iterative ridge regression algorithm is

$$e(\boldsymbol{\lambda}) = \operatorname{tr}\big[\{\tilde{\mathbf{D}} + \boldsymbol{\Sigma}_{\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}})\}^{-1}\tilde{\mathbf{D}}\big],$$

where $\tilde{\mathbf{D}}$ is a submatrix of $\hat{\mathbf{D}}$, defined in Section 3.2, corresponding to the nonzero components of $\hat{\boldsymbol{\beta}}$. Thus the GCV statistic is defined by

$$\mathrm{GCV}(\boldsymbol{\lambda}) = \frac{\mathrm{RSS}}{n\{1 - e(\boldsymbol{\lambda})/n\}^2},$$

where $\mathrm{RSS} = 2\ell(\hat{\boldsymbol{\beta}})$ is the residual sum of squares corresponding to $\hat{\boldsymbol{\beta}}$, given $\boldsymbol{\lambda}$. We select $\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}}\{\mathrm{GCV}(\boldsymbol{\lambda})\}$.

To find an optimal $\boldsymbol{\lambda}$, we need to minimize the GCV over a d-dimensional space. This is an unduly onerous task. Intuitively, it is expected that the magnitude of λ_j should be proportional to the standard error of the weighted least squares estimate of β_j. Therefore, we may set $\boldsymbol{\lambda} = \lambda\,\mathrm{se}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}})$ in practice, where $\mathrm{se}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}})$ denotes the SE of the unpenalized weighted least squares estimate. Thus we minimize the GCV score over the one-dimensional space, which will save a great deal of cost. We implement this in our simulation.

3.4 Sampling Properties

We now study the asymptotic properties of the resulting estimate of the penalized least squares (19). Express $\mathcal{L}(\boldsymbol{\beta})$ as

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\big\{y_i(t) - \hat{\alpha}(t_{ij};\boldsymbol{\beta}) - \mathbf{x}^T(t_{ij})\boldsymbol{\beta}\big\}^2 w(t_{ij}) + n\sum_{j=1}^{d} p_{\lambda_{jn}}(|\beta_j|). \quad (21)$$

Expression (21) provides a unified form of penalized least squares for Lin and Ying's approach and the profile least squares approach. Specifically, for Lin and Ying's approach, $\hat{\alpha}(t;\boldsymbol{\beta}) = \bar{y}(t;\hat{\boldsymbol{\gamma}}) - \boldsymbol{\beta}^T\bar{\mathbf{x}}(t,\hat{\boldsymbol{\gamma}})$, whereas for the profile least squares, $\hat{\boldsymbol{\alpha}}(\cdot;\boldsymbol{\beta}) = \mathbf{S}(\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X})$.

First we establish the convergence rate of the penalized profile least squares estimator. Assume that all penalty functions $p_{\lambda_{jn}}(\cdot)$ are negative and nondecreasing with $p_{\lambda_{jn}}(0) = 0$. Let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$, and let

$$a_n = \max_j\big\{\big|p'_{\lambda_{jn}}(|\beta_{j0}|)\big| : \beta_{j0} \neq 0\big\}$$

and

$$b_n = \max_j\big\{\big|p''_{\lambda_{jn}}(|\beta_{j0}|)\big| : \beta_{j0} \neq 0\big\}. \quad (22)$$

Theorem 2. Under the conditions of Theorem 1, if both a_n and b_n tend to 0 as $n \to \infty$, then with probability tending to 1, there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $\mathcal{L}(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2} + a_n)$.

Theorem 2 demonstrates how the rate of convergence of the penalized weighted least squares estimator $\hat{\boldsymbol{\beta}}$ depends on λ_j. To achieve the root n convergence rate, we have to take λ_j small enough so that $a_n = O_P(n^{-1/2})$. Next we establish the oracle property for the penalized profile least squares estimator. Let $\boldsymbol{\beta}_S$ consist of all nonzero components of $\boldsymbol{\beta}_0$, and let $\boldsymbol{\beta}_N$ consist of all zero components of $\boldsymbol{\beta}_0$. Let

$$\mathbf{x}^T(t)\boldsymbol{\beta}_0 = \mathbf{x}_S^T(t)\boldsymbol{\beta}_S + \mathbf{x}_N^T(t)\boldsymbol{\beta}_N = \mathbf{x}_S^T(t)\boldsymbol{\beta}_S,$$

where $\mathbf{x}_S(t)$ and $\mathbf{x}_N(t)$ are two subsets of covariates. The first part in the right side of the foregoing equation is the significant part in the model, whereas the second part is not significant. Thus, for ease of presentation, we assume, without loss of generality, that all of the first s components of $\boldsymbol{\beta}_0$ are not equal to 0, and all other components do equal 0, that is, $\boldsymbol{\beta}_{10} = \boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_{20} = \boldsymbol{\beta}_N$. Denote

$$\boldsymbol{\Sigma} = \operatorname{diag}\big\{p''_{\lambda_{1n}}(|\beta_{10}|), \ldots, p''_{\lambda_{sn}}(|\beta_{s0}|)\big\},$$

and

$$\mathbf{b} = \big(p'_{\lambda_{1n}}(|\beta_{10}|)\operatorname{sgn}(\beta_{10}), \ldots, p'_{\lambda_{sn}}(|\beta_{s0}|)\operatorname{sgn}(\beta_{s0})\big)^T.$$

Further, let $\hat{\boldsymbol{\beta}}_1$ consist of the first s components of $\hat{\boldsymbol{\beta}}$ and let $\hat{\boldsymbol{\beta}}_2$ consist of the last $d - s$ components of $\hat{\boldsymbol{\beta}}$.

Theorem 3 (Oracle property). Assume that for $j = 1, \ldots, d$, $\lambda_{jn} \to 0$, $\sqrt{n}\lambda_{jn} \to \infty$, and the penalty function $p_{\lambda_{jn}}(|\beta_j|)$ satisfies that

$$\liminf_{n\to\infty}\liminf_{\beta_j\to 0+}\frac{p'_{\lambda_{jn}}(\beta_j)}{\lambda_{jn}} > 0. \quad (23)$$

If $a_n = O_P(n^{-1/2})$, then under the conditions of Theorem 2, with probability tending to 1, the root n consistent local minimizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ in Theorem 2 must satisfy the following:

a. (sparsity) $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$
b. (asymptotic normality)

$$\sqrt{n}\{\mathbf{A}_{11} + \boldsymbol{\Sigma}\}\big[\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + \{\mathbf{A}_{11} + \boldsymbol{\Sigma}\}^{-1}\mathbf{b}\big] \to \mathrm{N}_s(\mathbf{0}, \mathbf{B}_{11})$$

in distribution, where $\mathbf{A}_{11}$ and $\mathbf{B}_{11}$ consist of the first s rows and columns of $\mathbf{A}$ and $\mathbf{B}$ defined in Theorem 1.

From Theorem 3, if $\lambda_{jn} \to 0$, $\sqrt{n}\lambda_{jn} \to \infty$ for $j = 1, \ldots, d$, $a_n = O_P(n^{-1/2})$, and condition (23) is satisfied, then the resulting estimate has an oracle property. This implies that the resulting procedure correctly specifies the true model and estimates the unknown regression coefficients as efficiently as if we knew the submodel. If all of the penalty functions are SCAD, then $a_n = 0$ when n is sufficiently large, and hence the resulting estimate has the oracle property. However, this is not true for the L_1 penalty, because the condition $a_n = \max_j \lambda_{jn} = O_p(n^{-1/2})$ and the conditions $\sqrt{n}\lambda_{jn} \to \infty$ cannot be satisfied simultaneously.

4. NUMERICAL STUDIES

We now assess the finite-sample performance of the newly proposed procedures via Monte Carlo simulations.

4.1 Simulation Models

We generated simulation data from the semiparametric model

$$y(t) = \alpha(t) + \boldsymbol{\beta}^T \mathbf{x}(t) + \varepsilon(t),$$

where $\alpha(t) = \tau\sqrt{t/\tau}$ or $\tau \sin(2\pi t/\tau)$, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\varepsilon(t)$ is a Gaussian process with mean 0 and covariance function $E\{\varepsilon(s)\varepsilon(t)\} = \exp(-2|t-s|)$. We simulated the covariate vector $\mathbf{x}$ from a normal distribution with mean 0 and $\operatorname{cov}(x_i, x_j) = .5^{|i-j|}$. Following Lin and Ying (2001), we set $w(t) \equiv 1$ for simplicity. The mechanism for generating simulation data is as follows:

- Case I. Observation times are independent of covariates. We adopt the scheme for generating observation times of Lin and Ying (2001). We generated the counting process $N^*(t)(t > 0)$ for the observation times from a random-effects Poisson process with intensity rate η, where η is an independent gamma variable with mean 1 and variance .5. Thus the observation times within the same subject are positively correlated. We set $N^*(0) \equiv 1$ so that each subject has at least one observation. The censoring time was an independent uniform $(0, \tau)$ variable, where τ equals either 4 or 20, which yields, on average, 3 and 11 observations per subject.
- Case II. Observation times depend on covariates. We generated the counting process $N^*(t)$ for the observation times from a random-effects Poisson process with intensity rate $\eta \exp(.5x_1)$, where η is the same as that in case I. Further, the censoring time and the covariate vector are also the same as those in case I.
- Case III. Observation times are fixed. The censoring time is the same as that in case I, but the observation times are set to be integers, $0, 1, 2, \ldots$. The censoring time was an independent uniform $(0, \tau)$ variable, where τ equals either 10 or 20, which yields, on average, 6 and 11 observations per subject.
- Case IV. Observation times are scheduled but can be randomly missed. Each individual has a set of "scheduled" time points, $\{0, 1, 3, \ldots, 29\}$, and each scheduled time (except time 0) has a probability of being skipped of 60%. The actual observation time is a random perturbation of the scheduled time; a uniform distribution over a $[-1, 1]$ random deviate is added to the nonskipped scheduled time to obtain the different observed time point t_{ij} per subject. This scheme for generating observation times is similar to that of Huang et al. (2002). The baseline function $\alpha(t)$ is taken to be either $30\sqrt{t/30}$ or $30 \sin(2\pi t/30)$, and the covariate vector is the same as that in case I.

4.2 Performance of Semiparametric Estimators

4.2.1 Performance of $\hat{\boldsymbol{\beta}}$. We assess the performance of an underlying estimator $\hat{\boldsymbol{\beta}}$ via its mean squares error (MSE), $E\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$. To evaluate the MSE, we conducted K replicates of Monte Carlo simulations; the MSE is estimated by

$$\frac{1}{K}\sum_{k=1}^{K} \|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}\|^2.$$

In our simulations, $K = 400$. The MSEs of the DBE and the profile least squares estimator are compared with Lin and Ying's (LY for short) estimator. The relative MSE (RMSE), the ratio of the MSE of an underlying estimator to that of the LY estimator, is depicted in Table 1. The profile least squares estimator improves the LY estimator by reducing interpolation bias and efficiently estimating the baseline function. In addition, the profile least squares estimator improves the DBE by reducing variance. Thus Table 1 shows that the profile least squares estimator performs best in all four cases. For cases I, II, and IV, there are many distinct sampling time points, and (10) approximately holds. Therefore, for such cases, the DBE also outperforms the LY estimator. For case III, the spacing between sampling time points is wide, and thus (10) does not hold. Thus the DBE does not improve the performance of the LY estimator. From Table 1, the relative performance of the LY method gets poorer as τ increases. This is due to the bias of the nearest-neighbor approximations used in the LY method.

Now we test the accuracy of the SE formula (16) for the profile least squares estimator. The standard deviation (SD) of the 400 estimated coefficients from the 400 simulations can be considered the true SE except for Monte Carlo error. (The relative size of Monte Carlo error is approximately of size $\sqrt{1/800}$.)

Table 1. Ratios of MSEs

		$n = 50$		$n = 75$	
$\alpha(t)$	τ	DBE	Profile LSE	DBE	Profile LSE
Case I					
$\tau\sqrt{t/\tau}$	4	.8481	.6592	.8407	.6661
$\tau\sqrt{t/\tau}$	20	.3450	.2962	.3963	.3246
$\tau\sin(2\pi t/\tau)$	4	.6632	.5065	.6756	.5377
$\tau\sin(2\pi t/\tau)$	20	.2798	.2324	.3001	.2359
Case II					
$\tau\sqrt{t/\tau}$	4	.7868	.6500	.7209	.6004
$\tau\sqrt{t/\tau}$	20	.2518	.2138	.2438	.2015
$\tau\sin(2\pi t/\tau)$	4	.5627	.4641	.5409	.4502
$\tau\sin(2\pi t/\tau)$	20	.1705	.1395	.1623	.1280
Case III					
$\tau\sqrt{t/\tau}$	10	1.0785	.7040	1.1299	.7348
$\tau\sqrt{t/\tau}$	20	.7748	.5006	.8316	.5424
$\tau\sin(2\pi t/\tau)$	10	1.2188	.6818	1.2560	.7347
$\tau\sin(2\pi t/\tau)$	20	.9868	.5007	1.0721	.5973
Case IV					
$30\sqrt{t/30}$		.9666	.6842	1.0501	.7086
$30\sin(2\pi t/30)$		.1434	.0953	.1360	.0869

Table 2. SDs, SEs, and 95% CP of Profile LSE for Case I With $\alpha(t) = \tau\sqrt{t/\tau}$

	β_1			β_2		
(n, τ)	SD	$SE_{(SD(SE))}$	95% CP	SD	$SE_{(SD(SE))}$	95% CP
(50, 4)	.151	$.138_{(.033)}$	.923	.168	$.158_{(.037)}$	.918
(75, 4)	.120	$.115_{(.021)}$	.930	.126	$.127_{(.024)}$	.955
(50, 20)	.085	$.082_{(.018)}$	.943	.100	$.091_{(.021)}$	.945
(75, 20)	.065	$.068_{(.013)}$	.953	.072	$.075_{(.014)}$	.960

The mean and SD of the 400 estimated SEs gauge the overall performance of the SE formula. The coverage probability (CP) indicates the how accuracy of the confidence interval. Table 2 presents only the simulation results of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ for case I with the baseline $\tau\sqrt{t/\tau}$. The results are similar for other cases. From Table 2, the difference between the true SE and the mean of the estimated SEs is less than half of a standard deviation of the estimated SEs. This implies that the proposed SE formula is accurate. The CP of 95% confidence interval of $\hat{\beta}_j$ is also presented.

4.2.2 Performance of $\hat{\alpha}(t)$. The performance of $\hat{\alpha}(\cdot)$ is assessed by the square root of average squared errors (RASE),

$$\mathrm{RASE}^2 = n_{\mathrm{grid}}^{-1} \sum_{k=1}^{n_{\mathrm{grid}}} \{\hat{\alpha}(t_k) - \alpha(t_k)\}^2,$$

where $\{t_k, k = 1, \ldots, n_{\mathrm{grid}}\}$ are the grid points at which the baseline function $\alpha(\cdot)$ is estimated. In our simulation, we use the Epanechnikov kernel $K(t) = .75(1 - t^2)_+$ and $n_{\mathrm{grid}} = 200$.

Local linear regression is used to estimate the baseline function. In our simulation, we take a sample $n = 50$, $\tau = 20$, and set the bandwidth h equal to $h_0 \times$ (interquartile range of observed times t_{ij}). As discussed in Section 2.6, the regression coefficient $\boldsymbol{\beta}$ can be estimated using either the DBE method or the profile least squares approach. So we also compare the performance of these two approaches. The RASEs for 3 different h_0's based on 400 replicates are listed in Table 3, as are the biases and the SDs of the 400 estimated baseline functions at $t = 5$. The results are very typical; the two approaches have almost the same performance. In addition, Table 3 also shows that the biases are small and that as the bandwidth increases, the biases increase but the SDs decrease. Figure 1 depicts the typical estimated curves of $\alpha(t)$.

Similar to (16), the estimated SE formula for the resulting estimator $\hat{\alpha}(t)$ is given by $\mathbf{SCS}^T$. To test the accuracy of the SE formula, we computed the average of the 400 estimated SEs and its standard deviation, as listed in Table 3. The results indicate that the SE formula performs very well.

Table 3. Summary of Simulation Results for $\hat{\alpha}(t)$ [$\alpha(t) = \tau\sqrt{t/\tau}$]

	RASE	$t = 5$		
h_0	$\mathrm{Mean}_{(SD)}$	Bias	SD	$SE_{(SD(SE))}$
DBE $\hat{\boldsymbol{\beta}}$				
.25	$.3195_{(.0665)}$	−.0284	.1361	$.1242_{(.0222)}$
.35	$.3519_{(.0630)}$	−.0661	.1219	$.1077_{(.0174)}$
.45	$.3889_{(.0605)}$	−.1207	.1134	$.1007_{(.0142)}$
PLS $\hat{\boldsymbol{\beta}}$				
.25	$.3195_{(.0656)}$	−.0310	.1349	$.1233_{(.0220)}$
.35	$.3525_{(.0618)}$	−.0686	.1206	$.1067_{(.0172)}$
.45	$.3900_{(.0592)}$	−.1232	.1118	$.0996_{(.0141)}$

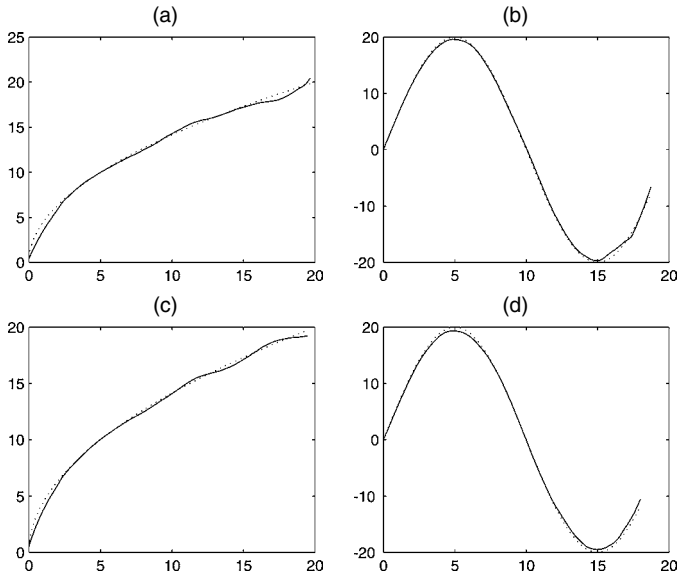

Figure 1. Typical Estimated Baseline Curves With $n = 50$ and $\tau = 20$. The solid lines represent the estimated curves of $\alpha(t)$; dotted lines, the true $\alpha(t)$. (a) and (c) Estimated baseline function $\alpha(t) = \tau\sqrt{t/\tau}$ using bandwidth .35 × IQR for cases I and II, when $n = 50$ and $\tau = 20$. (b) and (d) Estimated baseline function $\alpha(t) = \tau \sin(t/\tau)$ using bandwidth .2 × IQR for cases I and II.

4.3 Finite-Sample Performance of Variable Selection Procedures

4.3.1 Prediction Error. The prediction error is defined as the average error in the prediction of the dependent variable given the independent variables for future cases that are not used in the construction of a prediction equation. Let $\{\tilde{\mathbf{x}}(t), \tilde{y}(t), \tilde{N}(t)\}$ be a new observation from the underlying model; then the prediction error for model (1) is

$$\mathrm{PE}(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = E \int_0^\infty \left\{\tilde{y}(t) - \hat{\alpha}(t) - \hat{\boldsymbol{\beta}}^T \tilde{\mathbf{x}}(t)\right\}^2 d\tilde{N}(t),$$

where the expectation is a conditional expectation given the data used in constructing the prediction procedure. The prediction error can be decomposed as

$$\mathrm{PE}(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = E \int_0^\infty \sigma_\varepsilon^2(t) \exp\left\{\boldsymbol{\gamma}^T \tilde{\mathbf{x}}_i(t)\right\} \xi(t)\, d\Lambda(t) + E \int_0^\infty \left\{\hat{\alpha}(t) - \alpha_0(t) - (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \tilde{\mathbf{x}}(t)\right\}^2 d\tilde{N}(t),$$

where $\sigma_\varepsilon^2(t) = \mathrm{var}\{\varepsilon(t)\}$. The first component is the inherent prediction error due to noise; the second is due to lack of fit with an underlying model. This component, termed *model error*, can be further decomposed as

$$E \int_0^\infty \{\hat{\alpha}(t) - \alpha_0(t)\}^2 d\tilde{N}(t) + E \int_0^\infty \left\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \tilde{\mathbf{x}}(t)\right\}^2 d\tilde{N}(t) + 2E \int_0^\infty \{\hat{\alpha}(t) - \alpha_0(t)\}\left\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \tilde{\mathbf{x}}(t)\right\} d\tilde{N}(t).$$

The first component is the inherent model error due to lack of fit of the nonparametric component $\alpha_0(t)$, the second component is due to lack of fit of the parametric component, and the

third component is the covariance between the first two components, which equals

$$2\int_0^\infty \{\hat{\alpha}(t)-\alpha_0(t)\}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T \times E\big[\tilde{\mathbf{x}}(t)\exp\{\boldsymbol{\gamma}^T\tilde{\mathbf{x}}(t)\}\big]\xi(t)\,d\Lambda(t).$$

When $\gamma = 0$ (the observation times are independent of covariates) and $E\{\tilde{\mathbf{x}}(t)\} = 0$, the cross-product term is equal to 0. Therefore, the second term in the decomposition of model error plays a role in assessing the goodness of fit of the parametric component. We call the second term the *generalized mean squared error* (GMSE) and use it to compare the performance of our proposed variable selection procedure with others. The GMSE can be further simplified as

$$\text{GMSE} = (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\left\{\int_0^\infty E\tilde{\mathbf{x}}(t)^{\otimes 2}\exp\{\boldsymbol{\gamma}^T\tilde{\mathbf{x}}(t)\}\xi(t)\,d\Lambda(t)\right\} \times (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0).$$

When $\mathbf{x}(t)$ is randomly generated from a normal distribution, the GMSE has an analytic form, as shown by some straightforward calculations.

4.3.2 Comparison. We compare the performance of our procedure with existing procedures in terms of reduction of model complexity and relative GMSE (RGMSE), the ratio of GMSE of an underlying procedure to that of the profile least squares estimator without penalization. Table 4 depicts simulation results of some representative cases for the penalized profile least squares; results for other cases are similar. The means and table summarizes the SDs of RGMSEs over 400 simulated datasets and also reports the average number of zero coefficients. The column labeled "C" presents the average, restricted only to the true zero coefficients, whereas the column labeled "I" depicts the average of coefficients erroneously set to 0. From Table 4, it can be seen that for both kinds of penalized least squares, the penalized least squares with the SCAD and L_1 penalties effectively reduce model complexity, and that the SCAD performs as well as the oracle estimator and outperforms the penalized quadratic loss with the L_1 penalty. We have also conducted simulations to assess the performance of the penalized least squares constructed, based on the LY method. From our simulations, the relative performance of the penalized least squares estimate with the L_1 penalty and the SCAD penalty is similar to those given in Table 4. The ratio of the GMSE of the profile penalized least squares estimate to that corresponding to the LY estimator is similar to those shown in Table 1.

Table 5. Standard Deviations, Standard Errors, and 95% Coverage Probability of $\hat{\boldsymbol{\beta}}$

	β_1			β_2		
	SD	*SE*$_{(SD(SE))}$	*95% CP*	*SD*	*SE*$_{(SD(SE))}$	*95% CP*
L_1	.082	$.080_{(.018)}$	.935	.083	$.078_{(.018)}$	.918
SCAD	.081	$.081_{(.018)}$	.940	.082	$.079_{(.019)}$	.948
Oracle	.081	$.081_{(.018)}$	.942	.081	$.079_{(.019)}$	.950

Next we test the accuracy of the proposed standard error formula for the penalized least squares estimator. Similar to Table 2, Table 5 summarizes the simulation results for case I with $n = 50$, $\alpha(t) = \tau\sqrt{t/\tau}$, and $\tau = 20$. Results for other cases are similar. From Table 5, we see that the proposed standard error formula works very well.

4.4 An Application

We now illustrate the proposed procedures in Sections 2 and 3 via an analysis of a subset of data from the Multi-Center AIDS Cohort study. The dataset contains the human immunodeficiency virus (HIV) status of 283 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991. Details of the study design, methods, and medical implications have been given by Kaslow et al. (1987). During this study, all participants were scheduled to have their measurements taken during semiannual visits, but, because many participants missed some of their scheduled visits and the HIV infections occurred randomly during the study, there are unequal numbers of repeated measurements and different measurement times per individual. Fan and Zhang (2000) and Huang et al. (2002) analyzed the same dataset using varying-coefficient models. Their analysis aimed to describe the trend

Table 4. Comparison of Variable Selection Procedures

	$\alpha(t)=\tau\sqrt{t/\tau}$			$\alpha(t)=\tau\sin(2\pi t/\tau)$		
	RGMSE	*Zero coefficient*		*RGMSE*	*Zero coefficient*	
Method	*Mean*$_{(SD)}$	*C*	*I*	*Mean*$_{(SD)}$	*C*	*I*
Case I: $n=50$, $\tau=20$						
L_1	$.3936_{(.2966)}$	4.9950	0	$.3923_{(.2863)}$	4.9900	0
SCAD	$.3549_{(.2453)}$	4.9950	0	$.3533_{(.2453)}$	4.9925	0
Oracle	$.3502_{(.2412)}$	5.0000	0	$.3480_{(.2425)}$	5.0000	0
Case II: $n=75$, $\tau=4$						
L_1	$.5772_{(.2614)}$	4.3325	0	$.5733_{(.2648)}$	4.3500	0
SCAD	$.5127_{(.2101)}$	4.4275	0	$.5115_{(.2107)}$	4.4250	0
Oracle	$.3939_{(.2326)}$	5.0000	0	$.3915_{(.2318)}$	5.0000	0
Case III: $n=50$, $\tau=20$						
L_1	$.3975_{(.2843)}$	4.9950	0	$.4002_{(.2860)}$	4.9975	0
SCAD	$.3450_{(.2278)}$	4.9975	0	$.3460_{(.2279)}$	4.9975	0
Oracle	$.3438_{(.2269)}$	5.0000	0	$.3450_{(.2271)}$	5.0000	0
Case IV: $n=50$, $\tau=30$						
L_1	$.4091_{(.2716)}$	4.9975	0	$.4074_{(.2717)}$	5.0000	0
SCAD	$.3554_{(.2210)}$	5.0000	0	$.3546_{(.2205)}$	5.0000	0
Oracle	$.3549_{(.2200)}$	5.0000	0	$.3542_{(.2199)}$	5.0000	0

of the mean CD4 percentage depletion over time and to evaluate the effects of cigarette smoking, pre-HIV infection CD4 percentage, and age at infection on the mean CD4 percentage after the infection. Therefore, they took the CD4 cell percentage of a subject at distinct time points after HIV infection and considered the three covariates: Smoking, Age, and PreCD4. They fit the data by a varying-coefficient model,

$$y(t) = \beta_0(t) + \beta_1(t)\text{Smoking} + \beta_2(t)\text{Age}(t) + \beta_3(t)\text{PreCD4} + \varepsilon(t). \quad (24)$$

The results of the hypothesis testing of Huang et al. (2002) indicate that at significance level .05, only the baseline function varies over time, and PreCD4 has a constant effect over time, $\beta_3(t) = \beta_3$. Neither Smoking nor Age has a significant impact on mean CD4 percentage. This motivates us to use model (1) to fit this dataset and to use variable selection techniques to select a parsimonious model.

In our analysis, we took x_1 to be the smoking status: (1 for a smoker and 0 for a nonsmoker), $x_2(t)$ to be the standardized variable for age, and x_3 to be the standardized variable for PreCD4. It is of interest to examine whether there are any interaction effects and quadratic effects from these covariates. So we introduce the interactions of the three covariates and quadratic terms of x_2 and x_3 to the initial full model, and consider the following semiparametric model:

$$y(t) = \alpha(t) + \beta_1 x_1 + \beta_2 x_2(t) + \beta_3 x_3 + \beta_4 x_2^2(t) + \beta_5 x_3^2 + \beta_6 x_1 x_2(t) + \beta_7 x_1 x_3 + \beta_8 x_2(t) x_3 + \varepsilon(t). \quad (25)$$

We computed the DBE estimate for $\boldsymbol{\beta}$ to obtain the partial residuals for $\alpha(\cdot)$, and then selected the bandwidth $h = .5912$ by the Ruppert, Sheather, and Wand (1995) plug-in method. After that, we applied the profile least squares method with weight $w(t) \equiv 1$ to this model. The resulting estimates and standard errors are given in Table 6. Figure 2 depicts the estimated baseline function $\alpha(t)$ using the bandwidth $h = .5912$. It also plots the estimated baseline function plus/minus two standard errors, which can serve as a pointwise confidence interval ignoring the bias of the nonparametric fit.

We further applied the penalized profile least squares approach to select significant variables. The tuning parameter $\lambda = .7213$ for both the SCAD and the L_1 penalties. The results are also shown in Table 6. The penalized profile least squares with the SCAD penalty and the L_1 penalty yield almost the same results, except that the penalized profile least squares with the L_1 penalty shrinks the large coefficients more and results in a small SE. The results in Table 6 are in line with those of Huang et al. (2002), but indicate possible interactions between Smoking and Age; elder smokers tend to have lower average CD4 counts.

Table 6. Estimated Coefficients for Model (25)

Variable	Profile LS $\hat{\beta}_{(SE(\hat{\beta}))}$	L_1 $\hat{\beta}_{(SE(\hat{\beta}))}$	SCAD $\hat{\beta}_{(SE(\hat{\beta}))}$
Smoking	$.5333_{(1.0972)}$	$0_{(0)}$	$0_{(0)}$
Age	$-.1010_{(.9167)}$	$0_{(0)}$	$0_{(0)}$
PreCD4	$2.8252_{(.8244)}$	$3.0932_{(.5500)}$	$3.1993_{(.5699)}$
Age^2	$.1171_{(.4558)}$	$0_{(0)}$	$0_{(0)}$
PreCD4^2	$-.0333_{(.3269)}$	$0_{(0)}$	$0_{(0)}$
Smoking*Age	$-1.7084_{(1.1192)}$	$-.9684_{(.4904)}$	$-1.0581_{(.5221)}$
Smoking*PreCD4	$1.3277_{(1.3125)}$	$0_{(0)}$	$0_{(0)}$
Age*PreCD4	$-.1360_{(.5413)}$	$0_{(0)}$	$0_{(0)}$

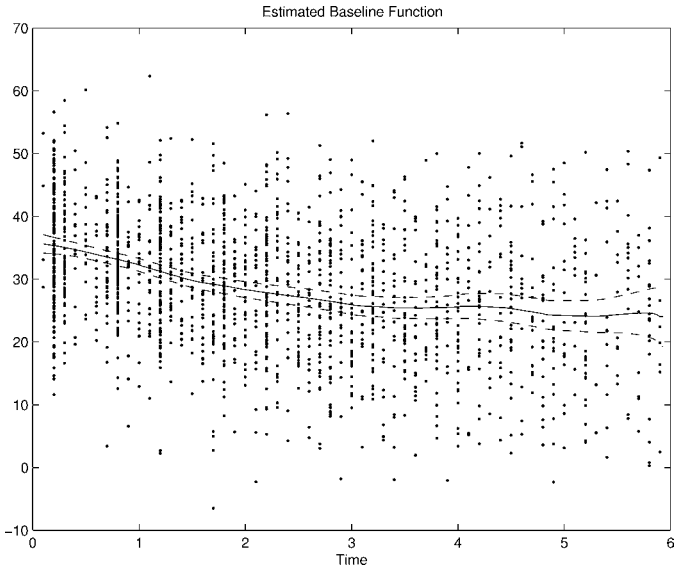

Figure 2. Estimated Baseline Function. The solid line represents the estimated baseline function; the dashed lines, the estimated baseline function plus/minus twice standard errors. The dots are the residual, on parametric part $r(t) = y(t) - \hat{\boldsymbol{\beta}}^T \mathbf{x}(t)$.

5. CONCLUDING REMARKS

In this article we have proposed two new approaches for estimating the regression coefficients in a semiparametric model. The asymptotic normality is established. We have further proposed an innovative class of variable selection procedures for the semiparametric model. With a proper choice of regularization parameters and penalty functions, we have demonstrated that the proposed variable selection procedures perform as well as an oracle estimator. Our method for the nonparametric component is distinguished from those of Martinussen and Scheike (1999, 2001) and Lin and Ying (2001), which focus on cumulatives of the nonparametric terms of the model.

As a referee pointed out, a very important model is the semiparametric varying coefficient model

$$y(t) = \mathbf{x}_1(t)\boldsymbol{\beta}(t) + \mathbf{x}_2(t)\boldsymbol{\gamma} + \varepsilon(t), \quad (26)$$

which has been studied by Martinussen and Scheike (2001) using a point process approach. Our profile least squares approach can be applied to model (26). Furthermore, our variable selection procedure can readily be extended to model (26). It is of interest to test whether or not the effect $\boldsymbol{\beta}(t)$ is really time-dependent. One may deal with this issue using the strategy proposed by Martinussen and Scheike (1999, 2001), Fan, Zhang, and Zhang (2001), and Lin and Ying (2001).

Under certain regularity conditions, the profile likelihood approach provides a semiparametric efficient estimator for independent observations. This is also true for our estimator; under the setting of independent and identically distributed observations, the covariance matrix given in Theorem 1 is the same as the semiparametric information bound given by, for example, Carroll et al. (1997). For longitudinal data analysis, the efficient bound is very complicated, as demonstrated by Lin and Carroll

(2001b). Finding a simple and effective nonparametric efficient method for the problem is an interesting subject of future research. Our simple profile likelihood method provides a useful tool for this investigation.

APPENDIX: PROOFS

A.1 Proof of Theorem 1

For each given $\boldsymbol{\beta}$, the estimator $\hat{\alpha}(t;\boldsymbol{\beta})$ is a local linear estimator of the bivariate data

$$\big\{\big(t_{ij}, y_i^*(t_{ij})\big), j = 1,\ldots,J_i, i = 1,\ldots,n\big\}.$$

Thus from the theory of local linear fitting (Fan 1992), it is a consistent estimate of the function

$$\alpha(t;\boldsymbol{\beta}) = E\big\{y(t) - \boldsymbol{\beta}^T\mathbf{x}(t)\big\} = \alpha_0(t) - (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T E\mathbf{x}(t). \quad \text{(A.1)}$$

Let $\ell_n(\boldsymbol{\beta})$ denote the weighted quadratic loss,

$$\ell_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\big\{y_i(t_{ij}) - \hat{\alpha}(t_{ij};\boldsymbol{\beta}) - \mathbf{x}_i(t_{ij})^T\boldsymbol{\beta}\big\}^2 w(t_{ij}). \quad \text{(A.2)}$$

Then $\hat{\boldsymbol{\beta}}$ minimizes the convex function $\ell_n(\boldsymbol{\beta})$. (In fact, it is a quadratic function of $\boldsymbol{\beta}$.) Decompose

$$\ell_n(\boldsymbol{\beta}) = I_{n,1}(\boldsymbol{\beta}) + I_{n,2}(\boldsymbol{\beta}) + I_{n,3}(\boldsymbol{\beta}), \quad \text{(A.3)}$$

where

$$I_{n,1}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\big\{y_i(t_{ij}) - \alpha(t_{ij};\boldsymbol{\beta}) - \mathbf{x}_i(t_{ij})^T\boldsymbol{\beta}\big\}^2 w(t_{ij}),$$

$$I_{n,2}(\boldsymbol{\beta}) = 2n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\big\{y_i(t_{ij}) - \alpha(t_{ij};\boldsymbol{\beta}) - \mathbf{x}_i(t_{ij})^T\boldsymbol{\beta}\big\} \times \big\{\alpha(t_{ij};\boldsymbol{\beta}) - \hat{\alpha}(t_{ij};\boldsymbol{\beta})\big\}w(t_{ij}),$$

and

$$I_{n,3}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\big\{\alpha(t_{ij};\boldsymbol{\beta}) - \hat{\alpha}(t_{ij};\boldsymbol{\beta})\big\}^2 w(t_{ij}).$$

Note that $\ell_n(\boldsymbol{\beta})$ is really the weighted residuals sum of squares of the local linear estimator $\hat{\alpha}(\cdot;\boldsymbol{\beta})$. Following some tedious calculations, similar to those of Müller and Stadtmüller (1993),

$$I_{n,2}(\boldsymbol{\beta}) = O_P\{I_{n,3}(\boldsymbol{\beta})\} = O\bigg(h^4 + \frac{1}{nh}\bigg). \quad \text{(A.4)}$$

We now deal with the main term $I_{n,1}$ in (A.2). This can be written as

$$I_{n,1}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\int_0^\infty \big\{y_i(t) - \alpha(t;\boldsymbol{\beta}) - \mathbf{x}_i(t)^T\boldsymbol{\beta}\big\}^2 w(t)\,dN_i(t).$$

Using the model

$$y(t) = \alpha_0(t) + \mathbf{x}(t)^T\boldsymbol{\beta}_0 + \varepsilon(t)$$

and (A.1), we have

$$\begin{aligned} I_{n,1}(\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^{n}\int_0^\infty \big[\varepsilon_i(t) - \{\mathbf{x}_i(t) - E\mathbf{x}_i(t)\}^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\big]^2 \\ &\qquad \times w(t)\,dN_i(t) \\ &= n^{-1}\sum_{i=1}^{n}\int_0^\infty \varepsilon_i^2(t)w(t)\,dN_i(t) \\ &\quad - 2(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\hat{\boldsymbol{\xi}}_n + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\beta}-\boldsymbol{\beta}_0). \end{aligned} \quad \text{(A.5)}$$

The minimization of this quadratic function is given by

$$\hat{\boldsymbol{\beta}}_0 = \boldsymbol{\beta}_0 + \hat{\boldsymbol{\Sigma}}_n^{-1}\hat{\boldsymbol{\xi}}_n.$$

By the law of large numbers and the central limit theorem,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0) \xrightarrow{\mathcal{L}} \mathrm{N}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}), \quad \text{(A.6)}$$

where

$$\mathbf{A} = E\int_0^\infty \{\mathbf{x}(t) - E\mathbf{x}(t)\}^{\otimes 2} w(t)\,dN(t)$$

and

$$\mathbf{B} = E\bigg\{\int_0^\infty \{\mathbf{x}(t) - E\mathbf{x}(t)\}\varepsilon(t)w(t)\,dN(t)\bigg\}^2.$$

Finally, we apply the convexity lemma (see, e.g., Andersen and Gill 1982) to show that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \sqrt{n}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{\xi}_n + o_P(1). \quad \text{(A.7)}$$

This together with (A.6) proves the results. To show that, first of all, by the convexity lemma, $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$, from (A.3), we have

$$\begin{aligned} &\mathbf{I}'_{n,1}(\hat{\boldsymbol{\beta}}) + \mathbf{I}'_{n,2}(\hat{\boldsymbol{\beta}}) + \mathbf{I}'_{n,3}(\hat{\boldsymbol{\beta}}) \\ &\quad = 2\hat{\boldsymbol{\Sigma}}_n(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - 2\hat{\boldsymbol{\xi}}_n + \mathbf{I}'_{n,2}(\hat{\boldsymbol{\beta}}) + \mathbf{I}'_{n,3}(\hat{\boldsymbol{\beta}}) \\ &\quad = \mathbf{0}. \end{aligned}$$

Similar to (A.4), we can show that

$$\mathbf{I}'_{n,2}(\hat{\boldsymbol{\beta}}) = o_P(n^{-1/2}) \qquad \text{and} \qquad \mathbf{I}'_{n,3}(\hat{\boldsymbol{\beta}}) = o_P(n^{-1/2}). \quad \text{(A.8)}$$

Hence, the result follows.

A.2 Derivation of Bias and Variance in (18)

In this section we outline the key steps to derive the bias and variance in (18). Detailed proofs are similar to those given by Fan (1992); also see chapter 3 of Fan and Gijbels (1996).

Denote $\hat{y}_i(t_{ij}) = y_i(t_{ij}) - \mathbf{x}_i(t_{ij})^T\hat{\boldsymbol{\beta}}$. From (12), the local linear fit involves finding $\hat{\mathbf{a}} = (\hat{a}_0, \hat{a}_1)^T$ minimizing

$$\sum_{i=1}^{n}\sum_{j=1}^{J_i}\big\{\hat{y}_i(t_{ij}) - a_0 - a_1(t_{ij} - t_0)\big\}^2 w(t_{ij})K_h(t_{ij} - t_0).$$

Using matrix notation, we have a closed form for $\hat{\mathbf{a}}$. Denote $\hat{\mathbf{y}} = (y_1(t_{11}),\ldots,y_n(t_{nJ_n}))^T$; $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2)$ is $n^* \times 2$ matrix, where $\mathbf{z}_1 = \mathbf{1}_{n^*}$ and $\mathbf{z}_2 = (t_{11} - t_0,\ldots,t_{nJ_n} - t_0)^T$. Furthermore, let $\mathbf{U} = \mathrm{diag}\{w(t_{11})K_h(t_{11} - t_0),\ldots,w(t_{nJ_n})K_h(t_{nJ_n} - t_0)\}$, an $n^* \times n^*$ diagonal matrix. Thus

$$\hat{\mathbf{a}} = (\mathbf{Z}^T\mathbf{U}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{U}\hat{\mathbf{y}}.$$

We further have

$$E\{\hat{\mathbf{a}}|\mathbf{X},\mathbf{t}\} = (\mathbf{Z}^T\mathbf{U}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{U}\{\boldsymbol{\alpha} + \mathbf{X}E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\} \quad \text{(A.9)}$$

and

$$\mathrm{var}\{\hat{\mathbf{a}}|\mathbf{X},\mathbf{t}\} = (\mathbf{Z}^T\mathbf{U}\mathbf{Z})^{-1}\big(\mathbf{Z}^T\mathbf{U}\,\mathrm{var}\{\hat{\mathbf{y}}|\mathbf{X},\mathbf{t}\}\mathbf{U}^T\mathbf{Z}\big)(\mathbf{Z}^T\mathbf{U}\mathbf{Z})^{-1}. \quad \text{(A.10)}$$

We first calculate the order of $S_n \equiv \mathbf{Z}^T\mathbf{U}\mathbf{Z}$. Denote

$$\begin{aligned} S_{n,l} &= \sum_{i=1}^{n}\sum_{j=1}^{J_i} w(t_{ij})K_h(t_{ij} - t_0)(t_{ij} - t_0)^l \\ &= \sum_{i=1}^{n}\int_0^\infty w(t)K_h(t - t_0)(t - t_0)^l\,dN_i(t), \qquad l = 0, 1, 2. \end{aligned}$$

Thus by some calculations,

$$E(S_{n,l}) = nE\int_0^\infty w(t)K_h(t-t_0)(t-t_0)^l\,dN(t)$$
$$= nh^l\Big\{\lambda(t_0)w(t_0)\int u^l K(u)\,du + o(1)\Big\}.$$

Using $S_{n,l} = ES_{n,l} + O_P\{\sqrt{\operatorname{var}(S_{n,l})}\}$, we can further show that

$$S_{n,l} = nh^l\lambda(t_0)w(t_0)\int u^l K(u)\,du\,\{1+o_P(1)\}, \tag{A.11}$$

provided that $h\to 0$ and $nh\to\infty$. Next we calculate the order of $S_n^* \equiv \mathbf{Z}^T\mathbf{U}\operatorname{var}\{\mathbf{y}|\mathbf{X},\mathbf{t}\}\mathbf{U}^T\mathbf{Z}$. Note that the off-diagonal elements of $\mathbf{U}\operatorname{var}\{\mathbf{y}|\mathbf{X},\mathbf{t}\}\mathbf{U}^T$ either are equal to 0 or have the form

$$w(t_{ij})K_h(t_{ij}-t_0)\operatorname{cov}\{\hat{y}_i(t_{ij}),\hat{y}_i(t_{ij'})\}w(t_{ij'})K_h(t_{ij'}-t_0),$$

which is negligible as $h\to 0$, because $K_h(t_{ij}-t_0)K_h(t_{ij'}-t_0)$ is negligible under mild conditions on the kernel function. Hence, the leading term of S_n^* involves its diagonal elements. Denote

$$S_{n,l}^* = \sum_{i=1}^n\sum_{j=1}^{J_i}(t_{ij}-t_0)^l w^2(t_{ij})K_h^2(t_{ij}-t_0)\operatorname{var}\{\varepsilon_i(t_{ij})\}$$
$$= \sum_{i=1}^n\int_0^\infty w^2(t)K_h^2(t-t_0)\operatorname{var}\{\varepsilon(t)\}\,dN_i(t).$$

Similar to $S_{n,l}$, we can show that

$$S_{n,l}^* = nh^{l-1}w^2(t_0)\lambda(t_0)\operatorname{var}\{\varepsilon(t_0)\}\int u^l K^2(u)\,du\,\big(1+O_P(1)\big). \tag{A.12}$$

Note that $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\| = O_P(n^{-1/2})$, the leading term in the right side of (A.9) is $(\mathbf{Z}^T\mathbf{U}\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{U}\boldsymbol{\alpha}$. Similar to Fan (1992), we can further derive the asymptotic bias given in (18) using (A.11). Again because $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\| = O_P(n^{-1/2})$, it can be shown that the difference between $\operatorname{var}\{\hat{y}_i(t_{ij})|\mathbf{X},\mathbf{t}\}$ and $\operatorname{var}\{\varepsilon_i(t_{ij})\}$ is negligible. So we can obtain the asymptotic variance by calculating the order of (A.10) using (A.11) and (A.12).

A.3 Proof of Theorem 2

Denote $\alpha_n = n^{-1/2}+a_n$. It is sufficient to show that, for any given $\eta>0$, there exists a large constant C such that

$$P\Big\{\inf_{\|\mathbf{u}\|=C}\mathcal{L}(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})\ge\mathcal{L}(\boldsymbol{\beta}_0)\Big\}\ge 1-\eta. \tag{A.13}$$

This implies, with probability at least $1-\eta$, that there exists a local minimizer in the ball $\{\boldsymbol{\beta}_0+\alpha_n\mathbf{u}:\|\mathbf{u}\|\le C\}$. Define

$$D_n(\mathbf{u}) = \mathcal{L}(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-\mathcal{L}(\boldsymbol{\beta}_0).$$

Note that $p_{\lambda_{jn}}(0)=0$ and $p_{\lambda_{jn}}(|\beta_j|)$ is nonnegative,

$$n^{-1}D_n(\mathbf{u}) \ge n^{-1}\big\{\ell(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-\ell(\boldsymbol{\beta}_0)\big\}$$
$$+\sum_{j=1}^s\big\{p_{\lambda_{jn}}(|\beta_{j0}+\alpha_n u_j|)-p_{\lambda_{jn}}(|\beta_{j0}|)\big\},$$

where $\ell(\boldsymbol{\beta})$ is the first term on the right side of (21). Using equation (A.5) and the rates in (A.8), it can be shown that

$$n^{-1}\{\ell(\boldsymbol{\beta}_0+\alpha_n\mathbf{u})-\ell(\boldsymbol{\beta}_0)\}$$
$$= \frac{\alpha_n^2}{2}\mathbf{u}^T\big\{\hat{\boldsymbol{\Sigma}}_n+o_P(1)\big\}\mathbf{u}-\alpha_n\mathbf{u}^T\big\{\hat{\boldsymbol{\xi}}_n+o_P(n^{-1/2})\big\}, \tag{A.14}$$

as $\ell(\boldsymbol{\beta})$ is a quadratic function of $\boldsymbol{\beta}$. Note that $\hat{\boldsymbol{\Sigma}}_n\to\mathbf{A}$, a finite positive definite matrix in probability. The first term in the right side of (A.14) is of order $O_P(C^2\alpha_n^2)$, and the second term is of order $O_P(Cn^{-1/2}\alpha_n) = O_P(C\alpha_n^2)$. Furthermore,

$$\sum_{j=1}^s\big\{p_{\lambda_{jn}}(|\beta_{j0}+\alpha_n u_j|)-p_{\lambda_{jn}}(|\beta_{j0}|)\big\} \tag{A.15}$$

is bounded by

$$\sqrt{s}\alpha_n a_n\|\mathbf{u}\|+\alpha_n^2 b_n\|\mathbf{u}\|^2 = C\alpha_n^2(\sqrt{s}+b_n C)$$

by the Taylor expansion and the Cauchy–Schwarz inequality. As $b_n\to 0$, the first term on the right side of (A.14) will dominate (A.15) as well as the second term on the right side of (A.14), by taking C sufficiently large. Hence (A.13) holds for sufficiently large C. This completes the proof of the theorem.

A.4 Proof of Theorem 3

Lemma A.1. Under the conditions of Theorem 3, with probability tending to 1, for any given $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and any constant C,

$$\mathcal{L}\big\{(\boldsymbol{\beta}_1^T,\mathbf{0})^T\big\} = \min_{\|\boldsymbol{\beta}_2\|\le Cn^{-1/2}}\mathcal{L}\big\{(\boldsymbol{\beta}_1^T,\boldsymbol{\beta}_2^T)^T\big\}.$$

Proof. We show that with probability tending to 1, as $n\to\infty$, for any $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}\| = O_P(n^{-1/2})$ and $\|\boldsymbol{\beta}_2\|\le Cn^{-1/2}$, $\partial\ell(\boldsymbol{\beta})/\partial\beta_j$ and β_j have the same signs for $\beta_j\in(-Cn^{-1/2},Cn^{1/2})$, for $j=s+1,\ldots,d$. Thus the minimizer attains at $\boldsymbol{\beta}_2=\mathbf{0}$.

For $\beta_j\ne 0$ and $j=s+1,\ldots,d$,

$$\frac{\partial\mathcal{L}(\boldsymbol{\beta})}{\partial\beta_j} = \ell_j'(\boldsymbol{\beta})+np_{\lambda_{jn}}'(|\beta_j|)\operatorname{sgn}(\beta_j), \tag{A.16}$$

where $\ell_j'(\boldsymbol{\beta}) = \partial\ell(\boldsymbol{\beta})/\partial\beta_j$. By the proof of Theorem 1,

$$\ell_j'(\boldsymbol{\beta}) = -n\big\{\hat{\xi}_j-(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\hat{\boldsymbol{\Sigma}}_j+o_P(n^{-1/2})\big\},$$

where $\hat{\xi}_j$ is the jth component of $\hat{\boldsymbol{\xi}}_n$ and $\hat{\boldsymbol{\Sigma}}_j$ is the jth column of $\hat{\boldsymbol{\Sigma}}_n$. Note that $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| = O_P(n^{-1/2})$ by the assumption and that $\hat{\boldsymbol{\Sigma}}_n\to\mathbf{A}$ in probability. Thus $n^{-1}\ell_j(\boldsymbol{\beta})$ is of order $O_P(n^{-1/2})$. Therefore,

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_j} = n\lambda_{jn}\big\{\lambda_{jn}^{-1}p_{\lambda_{jn}}'(|\beta_j|)\operatorname{sgn}(\beta_j)+O_P(n^{-1/2}/\lambda_n)\big\}.$$

Because $\liminf_{n\to\infty}\liminf_{\beta_j\to 0^+}\lambda_{jn}^{-1}p_{\lambda_{jn}}'(|\beta_j|)>0$ and $n^{-1/2}\times\lambda_{jn}\to 0$, the sign of the derivative is completely determined by that of β_j. This completes the proof.

Proof of the Theorem 3. Part (a) directly follows by Lemma A.1. Now we prove part (b). Using an argument similar to the proof of Theorem 2, it can be shown that there exists a $\hat{\boldsymbol{\beta}}_1$ in Theorem 2 that is a root-n consistent local minimizer of $\mathcal{L}\{(\boldsymbol{\beta}_1^T,0)^T\}$, satisfying the penalized least squares equations

$$\frac{\partial\mathcal{L}\{(\hat{\boldsymbol{\beta}}_1^T,\mathbf{0})^T\}}{\partial\boldsymbol{\beta}_1} = \mathbf{0}.$$

Following the proof of Theorem 1, we have

$$\frac{\partial\mathcal{L}\{(\hat{\boldsymbol{\beta}}_1^T,\mathbf{0})^T\}}{\partial\boldsymbol{\beta}_1}$$
$$= n\big[-\hat{\boldsymbol{\xi}}_{(1)}+o_P(n^{-1/2})+\big\{\hat{\boldsymbol{\Sigma}}_{(1)}+o_P(1)\big\}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})\big]$$
$$+n\big[\mathbf{b}_n+\boldsymbol{\Sigma}\{1+o_P(1)\}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10})\big],$$

where $\hat{\xi}_{(1)}$ consists of the first s components of $\hat{\boldsymbol{\xi}}_n$ and $\hat{\boldsymbol{\Sigma}}_{(1)}$ consists of the first s rows and columns of $\hat{\boldsymbol{\Sigma}}_n$.

Therefore, similar to the proof of Theorem 1 and by Slutsky's theorem, it follows that

$$\sqrt{n}(\mathbf{A}_{11}+\boldsymbol{\Sigma})\big\{\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{10}+(\mathbf{A}_{11}+\boldsymbol{\Sigma})^{-1}\mathbf{b}\big\}\to\mathrm{N}_s(\mathbf{0},\mathbf{B}_{11}). \tag{A.17}$$

This completes the proof of Theorem 3.

[Received November 2002. Revised March 2004.]

REFERENCES

Andersen, P. K., and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100–1120.

Antoniadis, A., and Fan, J. (2001), "Regularization of Wavelet Approximations" (with discussion), *Journal of the American Statistical Association*, 96, 939–967.

Breiman, L. (1996), "Heuristics of Instability and Stabilization in Model Selection," *The Annals of Statistics*, 24, 2350–2383.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489.

Cheng, S. C., and Wei, L. J. (2000), "Inferences for a Semiparametric Model With Panel Data," *Biometrika*, 87, 89–97.

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

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002), *Analysis of Longitudinal Data* (2nd ed.), Oxford, U. K.: Oxford University Press.

Fan, J. (1992), "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall.

Fan, J., and Huang, L. (2001), "Goodness-of-Fit Test for Parametric Regression Models," *Journal of the American Statistical Association*, 96, 640–652.

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

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29, 153–193.

Fan, J., and Zhang, J. (2000), "Two-Step Estimation of Functional Linear Models With Applications to Longitudinal Data," *Journal of the Royal Statistical Society*, Ser. B, 62, 303–322.

Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–148.

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

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

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.

Lin, D. Y., and Ying, Z. (2001), "Semiparametric and Nonparametric Regression Analysis of Longitudinal Data" (with discussion), *Journal of the American Statistical Association*, 96, 103–126.

Lin, X., and Carroll, R. J. (2001a), Comment on "Semiparametric and Nonparametric Regression Analysis of Longitudinal Data," by D. Y. Lin and Z. Ying, *Journal of the American Statistical Association*, 96, 114–116.

——— (2001b), "Semiparametric Regression for Clustered Data Using Generalized Estimating Equations," *Journal of the American Statistical Association*, 96, 1045–1056.

Martinussen, T., and Scheike, T. H. (1999), "A Semiparametric Additive Regression Model for Longitudinal Data," *Biometrika*, 86, 691–702.

——— (2001), "Sampling-Adjusted Analysis of Dynamic Additive Regression Models for Longitudinal Data," *Scandinavian Journal of Statistics*, 28, 303–323.

Moyeed, R. A., and Diggle, P. J. (1994), "Rates of Convergence in Semiparametric Modeling of Longitudinal Data," *Australian Journal of Statistics*, 36, 75–93.

Müller, H. G., and Stadtmüller, U. (1993), "On Variance Function Estimation With Quadratic Forms," *Journal Statistical Planning and Inferences* 35, 213–231.

Ruppert, D., Sheather, S. J., and Wand, M. P. (1995), "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association*, 90, 1257–1270.

Severini, T. A., and Staniswalis, J. G. (1994), "Quasi-Likelihood Estimation in Semiparametric Models," *Journal of the American Statistical Association*, 89, 501–511.

Speckman, P. (1988), "Kernel Smoothing in Partial Linear Models," *Journal of the Royal Statistical Society*, Ser. B, 50, 413–436.

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

Verbeke, G., and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, New York: Springer-Verlag.

Wang, N., Carroll, R. J., and Lin, X. (2004), "Efficient Semiparametric Marginal Estimation for Longitudinal/Clustered Data," *Journal of the American Statistical Association*, to appear.

Wu, C. O., Chiang, T., and Hoover, D. R. (1998), "Asymptotic Confidence Regions for Kernel Smoothing of a Time-Varying Coefficient Model With Longitudinal Data," *Journal of the American Statistical Association*, 88, 1388–1402.

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.