Semiparametric Modeling for Longitudinal Data: Estimation, Variable Selection, Nonparametric Goodness-of-fit *

JIANQING FAN
Department of Statistics
University of North Carolina
Chapel Hill, NC 27599-3260

RUNZE LI
Department of Statistics
Pennsylvania State University
University Park, PA 16802-2111

Abstract

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, variable selections and nonparametric goodness-of fit that frequently arise from longitudinal data analysis. In this paper, 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 empirically tested. Local polynomial regression techniques are used to estimate the baseline function in the semiparametric model. The generalized likelihood ratio test is introduced to test whether or not the baseline function can be fitted by a family of parametric models.

KEY WORDS: Bootstrap; Goodness-of-fit test; Local polynomial regression; Partial linear model; Penalized least squares; Profile least squares; SCAD.

Fan's research was partially supported by NIH grant R01 HL69720. Li's research was supported by a NSF grant DMS-0102505 and a National Institute on Drug Abuse (NIDA) Grant 1-P50-DA10075.

1 Introduction

Longitudinal data are often highly unbalanced because 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. Diggle, Liang and Zeger (1994) gave a comprehensive account of parametric regression methods in longitudinal data analysis. Verbeke and Molenberghs (2000) systematically summarized the development of linear mixed models for longitudinal data.

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 and functional linear models, have been proposed for longitudinal data analysis. See, for example, Hastie and Tibshirani (1993), Hoover, Rice, Wu and Yang (1998), Wu, Chiang and Hoover (1998), Fan and Zhang (2000), Chiang, Rice and Wu (2001), Huang, Wu and Zhou (2002) and 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. Moyeed and Diggle (1994) and Zeger and Diggle (1994) proposed the following semiparametric model:

$$y(t) = \alpha(t) + \boldsymbol{\beta}^T \mathbf{x}(t) + \varepsilon(t), \tag{1.1}$$

where $y(t)$ and $\mathbf{x}(t)$ are the response variable and the $d \times 1$ covariate vector at time t, respectively, $\alpha(t)$ is an unspecified baseline function of t, $\boldsymbol{\beta}$ is a vector of unknown regression coefficients, and $\varepsilon(t)$ is a zero-mean stochastic process. Model (1.1) 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 (1.1) and its variations have been receiving increasing attention recently. See, for example, Martinussen and Scheike (1999), Lin and Ying (2001) and references therein.

In this paper, we first propose two new estimation procedures for regression coefficients. The difference based estimator (DBE) of $\boldsymbol{\beta}$ provides a simple and good initial estimate of $\boldsymbol{\beta}$. It does not rely on any smoothing techniques. The estimator is then refined by the newly proposed profile least-squares estimator, which is unbiased. It depends, however, on a choice of smoothing parameter. This can be relatively easily selected. With a good initial estimate of $\boldsymbol{\beta}$ such as the DBE, model (1.1) becomes a univariate, nonparametric regression problem. Thus, a wealth of bandwidth selection techniques for univariate nonparametric regression can be employed. The asymptotic normality of the profile least-squares is established and a consistent standard error formula is derived using the sandwich formula. Our study shows that our newly proposed estimators, including the DBE, outperforms that of the Lin and Ying's proposal.

Like parametric regression models, variable selection is important in the semiparametric model (1.1). The number of variables in (1.1) 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 the implementation such as the choice of bandwidth for each submodel. Further, as analyzed by Breiman (1996), they suffer from several drawbacks, including the lack of stability. While they are useful in practice, the stepwise deletion and the best subset method 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 have been proposed to select significant variables for parametric regression models (Fan and Li, 2001). They 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 possess 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. Since the baseline function has not yet been parameterized, a new quadratic loss between the observed data and the theoretical model is introduced, which involves only the unknown parameter $\boldsymbol{\beta}$. This permits us to extend the penalized least-squares technique to the semiparametric model (1.1). 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 will 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 employed to find the solution of the penalized least squares, and a robust standard error formula for estimated coefficients of nonzero components is derived by using a sandwich formula. The standard error formula is empirically tested. It performs very well with moderate sized sample.

Kernel regression has been applied to repeated measurement data (Hart and Wehrly, 1986). In a series of papers by Lin and Carroll (2001b) and references therein, they suggested using a kernel generalized estimating equation (GEE) to estimate the nonparametric regression function. In this paper, the baseline function, $\alpha(t)$, in model (1.1) is proposed using local polynomial regression. A nonparametric goodness-of-fit question which naturally arises is, whether a given parametric family of models adequately fits the baseline function. This kind of nonparametric goodness-of-fit test has been treated in the books by Hart (1997) and Bowman and Azzalini (1997). A generally applicable method, called the generalized likelihood ratio test, is proposed in Fan, Zhang and Zhang (2001), for testing a parametric (or nonparametric) null hypothesis against a semiparametric (or nonparametric) alternative hypothesis. The idea is a generalization of the traditional maximum likelihood ratio test. They unveiled the Wilks phenomenon for a number of testing problems: the asymptotic null distributions are independent of nuisance parameters and follow a χ^2-distribution with diverging degrees of freedom. Further, the procedure has been shown to be asymptotically optimal. The ideas of the generalized likelihood ratio test are extended to the semiparametric model (1.1). A bootstrap procedure is employed to estimate the null distribution of the proposed test. Our simulation studies show that the resulting procedure performs well, in terms of the size and the power of the test.

The rest of this paper is organized as follows. In Section 2, we propose two new estimation procedures for regression coefficients in the parametric component. Asymptotic normality of the proposed estimator is established. We propose in Section 3 a penalized, quadratic loss procedure for selecting significant variables in model (1.1). Statistical inference procedures for the baseline function $\alpha(t)$ are proposed 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 an appendix.

2 Estimation of Parametric Component

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

$$E\{y(t_{ij})|\mathbf{x}_i(t_{ij})\} = \alpha(t_{ij}) + \boldsymbol{\beta}^T\mathbf{x}_i(t_{ij}) \tag{2.1}$$

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

$$\begin{aligned} \mathbf{y}_i &= (y_i(t_{i1}), \cdots, y_i(t_{iJ_i}))^T, \\ \mathbf{X}_i &= (\mathbf{x}_i(t_{i1}), \cdots, \mathbf{x}_i(t_{iJ_i}))^T, \end{aligned}$$

and

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

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

$$\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta})^T W_i(\mathbf{y}_i - \boldsymbol{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta}) \tag{2.2}$$

with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where 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. Misspecification of the working matrix does not affect the consistency of the resulting estimate, but it does affect the efficiency. Following Lin and Ying (2001), we focus on the situation in which the weight matrix W_i is a diagonal matrix (working independence) with diagonal elements possibly data-dependent. The covariance structure on Y_i can also be incorporated into the estimation. We opt for not pursuing this to avoid unduly complicating the problem.

2.1 Lin and Ying's approach

Lin and Ying (2001) introduced the counting process technique to the estimation scheme. The time points where the observations on the i-th 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 regarded 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 paper, 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 above counting process notation, the least-squares problem (2.2) can be written as

$$\frac{1}{2}\sum_{i=1}^{n}\int_0^{+\infty} w(t)\{y_i(t) - \alpha(t) - \boldsymbol{\beta}^T\mathbf{x}_i(t)\}^2\, dN_i(t), \tag{2.3}$$

where the Stieltjes integral $\int_0^\infty h(s)\, dN_i(s) = \sum_{j=1}^{J_i} h(t_{ij})$ for any function $h(\cdot)$.

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

$$\widehat{\alpha}(t; \boldsymbol{\beta}) = \bar{y}(t) - \boldsymbol{\beta}^T\bar{\mathbf{x}}(t), \tag{2.4}$$

where

$$\bar{\mathbf{x}}(t) = \sum_{i=1}^{n}\xi_i(t)\mathbf{x}_i(t) / \sum_{i=1}^{n}\xi_i(t),$$

where $\xi_i(t) = I(c_i \geq t)$, and

$$\bar{y}(t) = \sum_{i=1}^{n}\xi_i(t)y_i(t) / \sum_{i=1}^{n}\xi_i(t).$$

Substituting into (2.3) yields

$$\ell(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\int_0^{+\infty} w(t)[\{y_i(t) - \bar{y}(t)\} - \boldsymbol{\beta}^T\{\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\}]^2\, dN_i(t). \tag{2.5}$$

An estimate of $\boldsymbol{\beta}$ can be found by minimizing (2.5).

Lin and Ying (2001) extended the approach to the situation in which the potential observation times depend on the covariates. Specifically, they assumed that

$$E\{dN_i^*(t)|\mathbf{x}_i(t), y_i(t), c_i \geq t\} = \exp\{\boldsymbol{\gamma}'\mathbf{x}_i(t)\}\, d\Lambda(t), \quad i = 1, \cdots, n, \tag{2.6}$$

where $\boldsymbol{\gamma}$ is a vector of unknown parameter and $\Lambda(\cdot)$ is an arbitrary nondecreasing function. When $\boldsymbol{\gamma} = 0$, it corresponds to the situation where the observation times are independent of the covariates. In this case, $\Lambda(t)$ is the mean cumulative number of observations by time t in the absence of censoring.

Let

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

and

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

Thus, substituting $\bar{\mathbf{x}}(t)$ and $\bar{y}(t)$ by $\bar{\mathbf{x}}(t, \boldsymbol{\gamma})$ and $\bar{y}(t, \boldsymbol{\gamma})$, respectively, (2.5) becomes

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2} \sum_{i=1}^n \int_0^{+\infty} w(t) [\{y_i(t) - \bar{y}(t, \boldsymbol{\gamma})\} - \boldsymbol{\beta}^T \{\mathbf{x}_i(t) - \bar{\mathbf{x}}(t, \boldsymbol{\gamma})\}]^2 \, dN_i(t). \tag{2.7}$$

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

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

(See Pepe and Cai, 1993). Substituting $\widehat{\boldsymbol{\gamma}}$ for $\boldsymbol{\gamma}$ in (2.7), an explicit form for $\widehat{\boldsymbol{\beta}}$ can be derived. It is clear that (2.5) is a specific case of (2.7) with $\boldsymbol{\gamma} = 0$. The standard error of the resulting estimate $\widehat{\boldsymbol{\beta}}$ can be constructed via the corresponding sandwich formula, as proposed in 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 (e.g. partial linear model, Speckman 1988, Severini and Staniswalis, 1994, Carroll *et al.*, 1997).

The weighted least-squares problems (2.5) and (2.7) require that the processes $y_i(t)$ and $\mathbf{x}_i(t)$ are fully observable until the censoring time c_i. This is an unrealistic assumption. Lin and Ying (2001) replaced them by their corresponding values at the nearest time where their values are observed. While this helps practical implementations of the procedure, the method introduces biases due to the nearest neighborhood approximations. Further, since, for each subject, the spaces among observation times $\{t_{ij}, j = 1, \cdots, J_i\}$ do not tend to zero even when the sample size tends to infinity, the approximation biases cannot always be negligible in practice. The approach can also cause some problems in asymptotic theory.

To avoid the above two problems, unbounded loss of efficiency and nonnegligible biases due to approximations, and maintain the simplicity of Lin and Ying's approach, we propose the following simple method from the partial linear model (Fan and Huang, 2001 and Yatchew 1997). Like 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

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

can be expressed in the vector notation as

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

ordered according to the time $\{t_{ij}\}$. By the marginal model (2.1), it follows

$$y_i = \alpha(t_i) + \boldsymbol{\beta}^T \mathbf{x}_i + \varepsilon_i, \quad \text{with} \quad E(\varepsilon_i|\mathbf{x}_i) = 0. \tag{2.8}$$

First of all, observe that

$$y_{i+1} - y_i = \alpha(t_{i+1}) - \alpha(t_i) + \boldsymbol{\beta}^T(\mathbf{x}_{i+1} - \mathbf{x}_i) + e_i, \quad i = 1, \cdots, n^* - 1, \tag{2.9}$$

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 (2.9) is negligible. The least-squares approach can be employed to estimate the parameter $\boldsymbol{\beta}$. The method can be further improved by fitting the following linear model

$$y_{i+1} - y_i = \alpha_0 + \alpha_1(t_{i+1} - t_i) + \boldsymbol{\beta}^T(\mathbf{x}_{i+1} - \mathbf{x}_i) + e_i, \quad i = 1, \cdots, n^* - 1. \tag{2.10}$$

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 $\{t_i, i = 1, \cdots, n^*\}$. Fitting model (2.10) yields an estimate of $\boldsymbol{\beta}$. For simplicity, we will call this method as the Difference Based Estimator (DBE).

Note that the variance of e_i can easily be computed from ε_i. Weighted least-squares can be employed as in (2.5) and (2.7). Further, the variance-covariance of the error vector $\{e_i, i = 1, \cdots, n^* - 1\}$ can be explicitly found from the original covariance structure. We do not proceed in this direction any further, since the purpose for introducing DBE is to get a quick and reliable initial estimate of $\boldsymbol{\beta}$. This estimate will be used for the bandwidth selection of the profile least-squares method, which is more efficient.

We now argue that the loss of efficiency of this approach is limited to the class of estimators using the working independence covariance matrix. To get more insights into the analysis of asymptotic efficiency, pretend that the data are independent. If we use only the data with even indices in (2.10), the resulting data $\{y_{2i+1} - y_{2i}\}$ are still independent. We lose only the data $\{y_{2i+1} + y_{2i}\}$ which contains less information about $\boldsymbol{\beta}$ than $\{y_{2i+1} - y_{2i}\}$, since the former contains nuisance $\alpha(\cdot)$ while the latter does not, when the approximation errors are ignored. Thus, intuitively, there is at most a 50% loss of efficiency in the class of working independence estimators. This will also be observed in our simulation.

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 (1.1) can be written as

$$y^*(t) = \alpha(t) + \varepsilon(t). \tag{2.11}$$

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

$$\alpha(t) \approx \alpha(t_0) + \alpha'(t_0)(t - t_0) \equiv a + b(t - t_0).$$

Let $K(\cdot)$ be a kernel function and h be a bandwidth. The local linear fit is to find local parameters $\widehat{a}$ and $\widehat{b}$ to minimize

$$\sum_{i=1}^{n} \sum_{j=1}^{J_i} \{y_i^*(t_{ij}) - a - b(t_{ij} - t_0)\}^2 w(t_{ij}) K_h(t_{ij} - t_0), \tag{2.12}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$. Here the weight function, $w(t_{ij})$, serves a similar purposes to that in (2.3). The local linear estimate is simply $\widehat{\alpha}(t_0; \boldsymbol{\beta}) = \widehat{a}$.

Before we proceed further, let us introduce some notation. Let $\mathbf{y} = (\mathbf{y}_1^T, \cdots, \mathbf{y}_n^T)$, $\mathbf{X} = (\mathbf{X}_1^T, \cdots, \mathbf{X}_n^T)^T$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_n^T)^T$. Then, model (2.8) can be written as

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.13}$$

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 $\widehat{\boldsymbol{\alpha}} = \mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. The matrix $\mathbf{S}$ is usually called a smoothing matrix of the local linear smoother. It depends only on the observation times $\{t_{ij}, i = 1, \cdots, n, j = 1, \cdots, J_i\}$. Substituting $\widehat{\boldsymbol{\alpha}}$ into (2.13), we obtain

$$(\mathbf{I} - \mathbf{S})\mathbf{y} = (\mathbf{I} - \mathbf{S})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.14}$$

where $\mathbf{I}$ is the identity matrix of order n^*. Applying weighted least-squares to the linear model (2.14), we obtain

$$\widehat{\boldsymbol{\beta}} = \{\mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{y}, \tag{2.15}$$

where $\mathbf{W}$ is a diagonal matrix with the diagonal elements as $w(t_{ij})$s. This estimator is called the profile least-squares estimator. The profile least-squares estimator for the nonparametric component is simply $\alpha(\cdot; \widehat{\boldsymbol{\beta}})$.

It follows from (2.14) and (2.15) that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \{\mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}(\mathbf{I} - \mathbf{S})\mathbf{X}\}^{-1}\mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}\boldsymbol{\varepsilon}.$$

Thus, $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. Furthermore,

$$\operatorname{cov}\{\widehat{\boldsymbol{\beta}}|t_{ij}, \mathbf{x}_i(t_{ij})\} = \mathbf{D}^{-1}\mathbf{V}\mathbf{D}^{-1}, \tag{2.16}$$

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

$$\widehat{\mathbf{V}} = \mathbf{X}^T(\mathbf{I} - \mathbf{S})^T\mathbf{W}\mathbf{C}\mathbf{W}^T(\mathbf{I} - \mathbf{S})\mathbf{X}, \tag{2.17}$$

where $\mathbf{C} = \operatorname{diag}\{\widehat{\boldsymbol{\varepsilon}}_1\widehat{\boldsymbol{\varepsilon}}_1^T, \cdots, \widehat{\boldsymbol{\varepsilon}}_n\widehat{\boldsymbol{\varepsilon}}_n^T\}$, and $\widehat{\boldsymbol{\varepsilon}}_i$ is the residual vector for the i-th subject.

A few questions arise in the practical implementation of the above 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 $\widehat{\alpha}(\cdot; \boldsymbol{\beta})$ does not usually cause a problem, since 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 (2.10), with even indices, can be regarded as the profile least-squares estimate using the local average of two data points as a nonparametric estimator:

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

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

Using (2.15) and noting that $\widehat{\boldsymbol{\alpha}} = \mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, $\widehat{\boldsymbol{\alpha}}$ is linear in $\mathbf{y}$, data driven methods, such as cross-validation (CV), or generalized cross-validation (GCV), can be used to select the bandwidth. But it will be computationally expensive. To avoid expensive computations, our practical choice of bandwidth is as follows. Use the DBE to get an estimate $\widehat{\boldsymbol{\beta}}_{DBE}$. Substituting it into (2.11), we have a univariate nonparametric regression problem. Let $\widehat{h}$ be the bandwidth that is appropriate for this problem. This can be obtained either by a subjective choice via visualization, or a data-driven procedure, such as substitution methods, or cross-validation methods. Use this $\widehat{h}$ for the profile least-squares estimate. From nonparametric theory, this optimal choice of bandwidth is of order $h_n = bn^{-1/5}$. Theorem 2.1 below endorses this choice.

The function $\alpha(\cdot)$ can not be estimated well at some tail of the observation times due to sparsity. Including its estimates at these regions in (2.14) 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.

2.4 Asymptotic result

It is well known that asymptotic theory depends on the formulation on how the data were collected. For longitudinal data, there are many possible formulations. For example, in a series of papers by Wu and his collaborators (see e.g. Hoover *et al.*, 1998; and Wu *et al.*, 1998), it was assumed that time points $\{t_{ij}\}$ are a random sample from a certain population. Diggle *et al.*(1994) used a different formulation. In Lin and Ying (2001), the counting process $N_i(\cdot)$ is assumed to be a random sample from a certain population. To be consistent with the simulation models used in this paper, we adopt the formulation of Lin and Ying (2001). Other formulations can also be accommodated and similar results can be obtained.

From a data analysis point of view, under any reasonable formulation, one should be able to show that the profile least-squares estimate is asymptotically normal with mean $\boldsymbol{\beta}_0$ and variance that can be consistently estimated by (2.16), where $\boldsymbol{\beta}_0$ is the true parameter. Thus, a formulation merely provides a theoretical device to verify the above statement. From this point of view, the theoretical formulations do not affect the practical implementations of our proposed procedure.

In the sequel, we use $\alpha_0(\cdot)$ and $\boldsymbol{\beta}_0$ to denote the true parameters. Since the estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\alpha}(\cdot)$ are linear in the response variable, one can directly demonstrate the asymptotic normality by computing asymptotic counterparts for various terms in (2.15). This will bury many good intuitions in the detailed asymptotic calculations. Instead, in the appendix, we will provide a much simpler idea for establishing the asymptotic normality of the estimators. When the weight function $w(t)$ is data-dependent, we will assume that it tends to a deterministic function in probability. Therefore, for simplicity, assume that $w(t)$ is a deterministic function of t.

Set

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

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

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

Let

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

and

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

Then, we have the following result.

Theorem 2.1 *Suppose that $w(\cdot)$ is continuous 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 $n \to \infty$,*

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

Theorem 2.1 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 $\widehat{\mathbf{V}}$ is a consistent estimator of $\mathbf{B}$. Thus, it can be shown that $\mathbf{D}^{-1}\widehat{\mathbf{V}}\mathbf{D}^{-1}$ is a consistent estimator of $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$. This will also be demonstrated in our empirical studies.

2.5 Numerical Studies

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

Simulation models

Simulation data were generated from the following 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 zero mean and covariance function $E\{\varepsilon(s)\varepsilon(t)\} = \exp(-2|t-s|)$. The covariate vector $\mathbf{x}$ was simulated from a normal distribution with mean zero and $\mathrm{cov}(x_i, x_j) = 0.5^{|i-j|}$. 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. The scheme for generating observation times in Lin and Ying (2001) is adopted here. The counting process $N^*(t)(t > 0)$ for the observation times was generated from a random effects Poisson process with intensity rate η, where η is an independent gamma variable with mean 1 and variance 0.5. 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. The counting process $N^*(t)$ for the observation times was generated from a random effects Poisson process with intensity rate $\eta\exp(0.5x_1)$, 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,.... 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,...,29}, and each scheduled time, except time 0, has a probability of being skipped 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 non-skipped scheduled time to obtain the different observed time point t_{ij} per subject. This scheme for generating observation times is similar to that in Huang, Wu and Zhou (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.

Performance of semiparametric estimators

The performance of an underlying estimator $\widehat{\boldsymbol{\beta}}$ is assessed via its Mean Squares Error (MSE), which equals $E\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$. To evaluate the MSE, we conducted K replicates of Monte Carlo simulations and the MSE is estimated by

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

In our simulations, $K = 400$. The MSEs of DBE and the profile least squares estimator are compared to 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, it can be seen from Table 1 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 (2.10) approximately holds. Therefore, for such three cases, the DBE also outperforms the LY estimator. For Case III, the spacing between sampling time points is wide, and therefore, (2.10) 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 neighborhood approximations used in the LY method.

Now we test the accuracy of the standard error formula (2.16) for the profile least squares estimator. The standard deviation of the 400 estimated coefficients from the 400 simulations can be regarded as the true standard error except for Monte Carlo error (the relative size of Monte Carlo error is approximately of size $\sqrt{1/800}$). The mean and standard deviation of the 400 estimated standard errors gauge the overall performance of the standard error formula. In Table 2, we present only the simulation results of the nonzero coefficients for Case I with the baseline $\tau\sqrt{t/\tau}$. For other cases, the results are similar. From Table 2, the difference between the true standard error and the mean of the estimated standard errors is less than half of a standard deviation of the estimated standard errors. This implies that the proposed standard error formula is accurate and works very well.

3 Variable Selection

Model selection is an indispensable tool for statistical data analysis. However, the problem has rarely been studied in the semiparametric context. In this section, we will 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 one would like to minimize. It can be (2.7) or the weighted quadratic loss induced by model (2.14). It reflects a semiparametric method that one would like to employ.

3.1 Penalized Weighted Least Squares

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

$$\mathcal{L}(\boldsymbol{\beta}) \equiv \ell(\boldsymbol{\beta}) + n \sum_{j=1}^{d} \lambda_j p_j(|\beta_j|), \tag{3.1}$$

where the $p_j(\cdot)$'s are penalty functions, and the λ_j are tuning parameters, which control the model complexity and can be selected by some data-driven methods, such as cross validation, or generalized cross validation. By minimizing (3.1), with a special construction of the penalty function, some coefficients are estimated as zero, which deletes the corresponding variables, while 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 (3.1) can be obtained from the penalized quadratic loss of the semiparametric model (2.1), using the profiling technique. For example, starting from the quadratic loss (2.2) and adding the penalty term $n \sum_{j=1}^{d} \lambda_j p_j(|\beta_j|)$, we obtain the penalized quadratic loss

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

After eliminating the nuisance function $\alpha(\cdot)$ using the profiling technique in §2.3 [see (2.14)], we obtain the following 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) provide various insights into how a penalty function should be chosen. They advocate 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 zero) to reduce model complexity, and *continuity* to avoid unnecessary variation in model prediction. Necessary conditions are given in Antoniadis and Fan (2001). None of the L_q penalties produce estimates that satisfy, simultaneously, the above 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

$$p'_\lambda(\beta) = \lambda \left\{ I(\beta \leq \lambda) + \frac{(a\lambda - \beta)_+}{(a-1)\lambda} I(\beta > \lambda) \right\}, \text{ for some } a > 2 \text{ and } \beta > 0,$$

and $p_\lambda(0) = 0$. For simplicity of presentation, we will use the name "SCAD" for 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 point of view. Hence, this value will be used throughout the rest of the paper.

3.2 Iterated Ridge Regression

It is challenging to find the solution of the penalized least squares of (3.1) because the penalty function $p_{\lambda_j}(|\beta_j|)$, such as the L_q penalty with $(0 < q \leq 1)$ and 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 (3.1), 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

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

In other words,

$$p_{\lambda_j}(|\beta_j|) \approx p_{\lambda_j}(|\beta_{j0}|) + \frac{1}{2}\{p'_{\lambda_j}(|\beta_j^{(0)}|)/|\beta_j^{(0)}|\}(\beta_j^2 - \beta_j^{(0)2}), \quad \text{for} \quad \beta_j \approx \beta_j^{(0)}.$$

With the local quadratic approximation, the Newton-Raphson algorithm can be implemented directly for minimizing $\mathcal{L}(\boldsymbol{\beta})$ defined in (3.1). Furthermore, the Newton-Raphson algorithm is indeed an iterative weighted least squares algorithm. For instance, we update the solution of the penalized profile least squares by

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

where

$$\Sigma_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^0) = \mathrm{diag}\{p'_{\lambda_1}(|\beta_1^{(0)}|)/|\beta_1^{(0)}|, \cdots, p'_{\lambda_d}(|\beta_d^{(0)}|)/|\beta_d^{(0)}|\}.$$

When there is a component $|\beta_j^{(0)}| < \eta$, it is set to zero. 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 above 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

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

When the algorithm converges, the estimator satisfies the condition

$$\partial\ell(\widehat{\boldsymbol{\beta}})/\partial\beta_j + np'_\lambda(|\widehat{\beta}_j|)\mathrm{sgn}(\widehat{\beta}_j) = 0, \tag{3.2}$$

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 standard error formula for the estimated coefficients can be derived from the iterative ridge regression. In other words, the covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be consistently estimated by $n^{-1}\{\widehat{\mathbf{D}} + \Sigma_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}})\}^{-1}\widehat{\mathbf{V}}\{\widehat{\mathbf{D}} + \Sigma_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}})\}^{-1}$ for the non-zero component, where

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

and

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

for the penalized profile least squares estimator. See (2.17) for an explicit form of $\widehat{\mathbf{V}}$. For the penalized least squares estimator derived by using Lin and Ying's approach, $\widehat{\mathbf{D}}$ and $\widehat{\mathbf{V}}$ are as defined in their paper.

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 $\lambda_1, \cdots, \lambda_d$. For linear estimators (in terms of response variable) in nonparametric regression, there is a large amount literature on how to choose 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 cross-validation, or generalized cross-validation can be utilized. Here we estimate $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_d)$ 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

$$e(\boldsymbol{\lambda}) = \text{tr}[\{\tilde{D} + \Sigma_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}})\}^{-1}\tilde{D}],$$

where $\tilde{D}$ is a submatrix of $\widehat{D}$, defined in Section 3.2, corresponding to the nonzero components of $\widehat{\boldsymbol{\beta}}$. Thus, the generalized cross-validation statistic is defined by

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

where RSS $= 2\ell(\widehat{\boldsymbol{\beta}})$ is the residual sum of squares corresponding to $\widehat{\boldsymbol{\beta}}$, given $\boldsymbol{\lambda}$. We select $\widehat{\boldsymbol{\lambda}} = \text{argmin}_{\boldsymbol{\lambda}}\{\text{GCV}(\boldsymbol{\lambda})\}$.

To find an optimal $\boldsymbol{\lambda}$, the GCV needs to be minimized 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 $\boldsymbol{\lambda} = \lambda\,\mathrm{se}(\widehat{\boldsymbol{\beta}}_{LS})$ in practice, where $\mathrm{se}(\widehat{\boldsymbol{\beta}}_{LS})$ stands for the standard error 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. This will be implemented in our simulation.

3.4 Sampling Properties

Now we study the asymptotic properties of the resulting estimate of the penalized least squares (3.1). Express $\mathcal{L}(\boldsymbol{\beta})$ as

$$\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{J_i}\{y_i(t) - \widehat{\alpha}(t_{ij};\boldsymbol{\beta}) - \mathbf{x}^T(t_{ij})\boldsymbol{\beta}\}^2 w(t_{ij}) + n\sum_{j=1}^{d} p_{\lambda_{jn}}(|\beta_j|). \tag{3.3}$$

Expression (3.3) 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, $\widehat{\alpha}(t;\boldsymbol{\beta}) = \bar{y}(t;\widehat{\gamma}) - \boldsymbol{\beta}^T\bar{\mathbf{x}}(t,\widehat{\gamma})$. While for the profile least squares, $\widehat{\boldsymbol{\alpha}}(\cdot;\boldsymbol{\beta}) = \mathbf{S}(\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X})$.

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

$$a_n = \max_j\{|p'_{\lambda_{jn}}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}, \qquad \text{and} \qquad b_n = \max_j\{|p''_{\lambda_{jn}}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}. \tag{3.4}$$

Theorem 3.1 *Under the Conditions of Theorem 2.1, if both a_n and b_n tend to zero as $n \to \infty$, then with probability tending to one, there exists a local minimizer $\widehat{\boldsymbol{\beta}}$ of $\mathcal{L}(\boldsymbol{\beta})$ such that $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_P(n^{-1/2} + a_n)$.*

Theorem 3.1 demonstrates how the rate of convergence of the penalized weighted least squares estimator $\widehat{\boldsymbol{\beta}}$ 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 $\boldsymbol{\beta}_N$ 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 hand side of the above equation is the significant part in the model, while 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 equal 0, i.e., $\boldsymbol{\beta}_{10} = \boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_{20} = \boldsymbol{\beta}_N$. Denote by

$$\Sigma = \mathrm{diag}\{p''_{\lambda_{1n}}(|\beta_{10}|), \cdots, p''_{\lambda_{sn}}(|\beta_{s0}|)\}$$

and

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

Theorem 3.2 *(**Oracle property**) Assume that for $j = 1, \cdots, d$, $\lambda_j \to 0$, $\sqrt{n}\lambda_j \to \infty$ and the penalty function $p_{\lambda_j}(|\beta_j|)$ satisfies that*

$$\liminf_{n\to\infty} \liminf_{\beta_j \to 0+} p_{\lambda_{jn}}(\beta_j)/\lambda_{jn} > 0. \tag{3.5}$$

If $a_n = O_P(n^{-1/2})$, then under the conditions of Theorem 3.1, with probability tending to 1, the root n consistent local minimizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ in Theorem 3.2 must satisfy

*(i) (**Sparsity**) $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$;*

*(ii) (**Asymptotic normality**)*

$$\sqrt{n}\{\mathbf{A}_{11} + \Sigma\}[\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + \{\mathbf{A}_{11} + \Sigma\}^{-1}\mathbf{b}] \to N_s(\mathbf{0}, \mathbf{B}_{11}).$$

in distribution, where $\mathbf{A}_{11}$ and $\mathbf{B}_{11}$, respectively, consist of the first s rows and columns of $\mathbf{A}$ and $\mathbf{B}$ defined in Theorem 2.1.

Parallel to Theorems 3.1 and 3.2, we can prove that the penalized least squares estimate constructed by using Lin and Ying's approach has the same rate of convergence as that in Theorem 3.1. We can further establish its oracle property. From Theorem 3.2, if $\lambda_{n,j} \to 0$, $\sqrt{n}\lambda_{n,j} \to \infty$ for $j = 1, \cdots, d$, $a_n = O_P(n^{-1/2})$, and condition (3.5) is satisfied, then the resulting estimate possesses 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 the penalty functions are SCAD, then $a_n = 0$ when n is sufficiently large, and hence the resulting estimate possesses the oracle property. However, this is not true for the L_1 penalty, since the condition $a_n = \max_j \lambda_{n,j} = O_p(n^{-1/2})$ and the conditions $\sqrt{n}\lambda_{n,j} \to \infty$ can not be satisfied simultaneously.

3.5 Finite Sample Performance of Variable Selection Procedures

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 (1.1) is

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

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

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

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

$$\begin{aligned} &E\int_0^\infty \{\widehat{\alpha}(t) - \alpha_0(t)\}^2\, d\tilde{N}(t) + E\int_0^\infty \{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\tilde{\mathbf{x}}(t)\}^2\, d\tilde{N}(t) \\ &+2E\int_0^\infty \{\widehat{\alpha}(t) - \alpha_0(t)\}\{(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T\tilde{\mathbf{x}}(t)\}\, d\tilde{N}(t). \end{aligned}$$

The first component is the inherent model error due to lack of fit of the nonparametric component $\alpha_0(t)$, the second one is due to lack of fit of the parametric component, and the third one is the covariance between the first two components, which equals

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

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 for assessing the goodness-of-fit of the parametric component. We will call the second term *generalized mean squared error*, denoted by GMSE, and use it to compare the performance of our proposed variable selection procedure with others. The GMSE can be further simplified as

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

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

Comparison

We compare the performance of our procedure with existing ones in terms of reduction of the model complexity and the Relative GMSE (RGMSE), the ratio of GMSE of an underlying procedure to that of the profile least squares estimator without penalization. Table 3 depicts simulation results of some representative cases for the penalized profile least squares. Results for other cases are similar. The means and standard deviations of RGMSEs over 400 simulated data sets are summarized in Table 3. The average number of zero coefficients is also reported in Table 3, in which the column labeled "C" presents the average, restricted only to the true zero coefficients, while the columns label "I" depicts the average of coefficients erroneously set to 0. From Table 3, 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 further, 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 in Table 3. The ratio of the GMSE of the profile penalized least squares estimate to that corresponding to the LY estimator is similar to those in Table 1.

Next, we test the accuracy of the proposed standard error formula for the penalized least squares estimator. Similar to Table 2, Table 4 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 4, the proposed standard error formula works very well.

4 Statistical inference for the baseline function

After we have fitted data to a semiparametric model (1.1), it is often of interest to check whether the data set can be fitted by a less complicated family of parametric models. To address this kind of issue, a nonparametric goodness of fit test is constructed for testing whether $\alpha(t)$ is of a parametric form or not.

4.1 Estimation

As discussed in Section 2.3, the baseline function $\alpha(\cdot)$ can be estimated by the smoothing of the partial residuals $\{(t_i, y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}), i = 1, \cdots, n^*\}$ by using a local linear fit. This results in a nonparametric fit $\widehat{\alpha}(\cdot; \widehat{\boldsymbol{\beta}})$. Since the rate of convergence for $\widehat{\boldsymbol{\beta}}$ is faster than that of the nonparametric estimator, $\widehat{\boldsymbol{\beta}}$ can either be a profile least-square estimator or the difference based estimator. The latter is much easier to obtain, while the former may have better performance in some situations. Since the errors in estimation $\boldsymbol{\beta}$ are negligible in the nonparametric estimation of α, the value of $\boldsymbol{\beta}$ can be regarded as known. Using the fact that $\widehat{\boldsymbol{\alpha}} = \mathbf{S}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$ is linear in $\mathbf{y}$ (ignoring the variability from $\widehat{\boldsymbol{\beta}}$), the standard error for $\widehat{\boldsymbol{\alpha}}$ can be estimated as $\mathbf{SCS}^T$. See (2.16) for a similar expression.

For a given shrinking neighborhood $t \pm h_n$ around a given time t, the chance 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 $N(t)$. Then, it can be shown that

$$\sqrt{nh_n}\{\widehat{\alpha}(t; \widehat{\boldsymbol{\beta}}) - \alpha_0(t) - \frac{1}{2}\alpha_0''(t)\int u^2K(u)duh_n^2\} \xrightarrow{\mathcal{L}} N(0, \sigma^2(t)),$$

where

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

We omit the details of the proof. The bias and variance expressions are similar to those in Fan (1992).

4.2 Nonparametric goodness-of-fit

Consider the following hypothesis testing problem:

$$H_0 : \alpha(t) = \alpha_0(t, \boldsymbol{\theta}) \qquad \text{versus} \qquad H_1 : \alpha(t) \neq \alpha_0(t, \boldsymbol{\theta}), \tag{4.1}$$

where $\alpha_0(t, \boldsymbol{\theta})$ has a parametric form in which we are interested, and $\boldsymbol{\theta}$ is a vector of unknown parameters. For example, taking $\alpha_0(t) = b_0 + b_1t$, the null hypothesis implies that the baseline function is linear in time t.

To gain more insights into the construction of nonparametric likelihood ratio type of tests, assume, tentatively, that the random error $\varepsilon(t)$ is a Gaussian process with zero mean and a covariance function $\mathrm{cov}\{\varepsilon(s), \varepsilon(t)\} = \sigma^2\Gamma(s, t)$, where $\Gamma(\cdot, \cdot)$ is known, but σ^2 is unknown. Furthermore, assume, temporarily, that $\boldsymbol{\beta}$ is known. Under these assumptions, and recalling notation in Section 2, the likelihood function of the data $\{\mathbf{y}_i, \mathbf{X}_i\}_{i=1}^n$ is proportional to

$$(\sigma^2)^{-n^*/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta})^T \Gamma_i^{-1} (\mathbf{y}_i - \boldsymbol{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta})\right\},$$

where Γ_i is the covariance matrix of $(\varepsilon(t_{i1}), \cdots, \varepsilon(t_{iJ_i}))$. Let $\tilde{\alpha}(t)$ and $\widehat{\alpha}(t)$ be estimates of $\alpha(t)$ under H_0 and H_1, respectively. Then, following Fan, Zhang and Zhang (2001), a generalized likelihood ratio (GLR) test statistic is defined as

$$\frac{n^*}{2}\log(\text{RSS}(H_0)/\text{RSS}(H_1))$$

where

$$\text{RSS}(H_0) = \sum_{i=1}^n (\mathbf{y}_i - \tilde{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta})^T \Gamma_i^{-1} (\mathbf{y}_i - \tilde{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta}),$$

and

$$\text{RSS}(H_1) = \sum_{i=1}^n (\mathbf{y}_i - \widehat{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta})^T \Gamma_i^{-1} (\mathbf{y}_i - \widehat{\alpha}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

Under H_0, the GLR test statistic is asymptotically equivalent to

$$T_0 = \frac{n^*}{2}\frac{\text{RSS}(H_0) - \text{RSS}(H_1)}{\text{RSS}(H_1)}.$$

Note that $\widehat{\alpha}_i$ is obtained by using a nonparametric estimator, rather than the maximum likelihood estimator under the alternative model. In fact, as argued in Fan, Zhang and Zhang (2001), the nonparametric MLE usually does not exist, and, even if it does exists, the resulting nonparametric maximum likelihood ratio test is not very powerful. This is the motivation behind the GLR test.

In practice, both $\boldsymbol{\beta}$ and $\Gamma(s,t)$ are unknown. However, under H_1, the unknown regression coefficient $\boldsymbol{\beta}$ can be root n consistently estimated. Furthermore, an ad hoc estimator for Γ_i can be constructed by using the residuals. Specifically, $\widehat{\Gamma}(s,t) = n^{-1}\sum_{i=1}^n \{\xi_i(s)\xi_i(t)\widehat{\varepsilon}_i^*(s)\widehat{\varepsilon}_i^*(t)\}$, where $\widehat{\varepsilon}_i^*(t)$ is the linear interpolation of standardized residuals. From the definition of $\widehat{\Gamma}(s,t)$, the resulting estimate of covariance matrix is non-negative definite. Thus, both $\text{RSS}(H_1)$ and $\text{RSS}(H_0)$ can be evaluated.

Intuitively, under H_0, there will be little difference between $\text{RSS}(H_0)$ and $\text{RSS}(H_1)$. However, under the alternative hypothesis, $\text{RSS}(H_0)$ should become systematically larger than $\text{RSS}(H_1)$, and hence the test statistic T_0 will tend to take a large positive value. Hence, a large value of the test statistic T_0 indicates that the null hypothesis should be rejected.

In the nonparametric regression model and varying coefficient models, Fan, Zhang and Zhang (2001) unveiled the following Wilks phenomenon: The asymptotic null distribution of

$$T \equiv r_K T_0 \tag{4.2}$$

is a chi-square distribution, where r_K is a normalizing constant depending on the kernel function as listed below.

Kernel	Uniform	Epanechnikov	Biweight	Triweight	Gaussian
r_K	1.2000	2.1153	2.3061	2.3797	2.5375

The theoretical justification of the Wilks type of phenomenon is beyond the scope of this study, since the current setup is more complicated. However, as pointed out in Section 3.1, the problem is nearly the same as if the data were sampled independently, since the chance of getting two data points from the same subject in a local neighborhood is small. Thus, we would expect the Wilks type of results to continue to hold here. We will provide empirical justifications in Section 4.3. See Figure 2. Similar to the proposal of Cai, Fan and Li (2000), the null distribution of T can be estimated by using a bootstrap procedure. This usually provides a better estimate than the asymptotic null distribution, since, in the nonparametric situation, the degrees of freedom tends to infinite and the results in Fan, Zhang and Zhang (2001) give only the main order of the degrees of freedom.

4.3 Numerical Studies

Performance of the baseline function

The performance of $\widehat{\alpha}(\cdot)$ is assessed by the square root of average squared errors (RASE),

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

where $\{t_k, k = 1, \cdots, n_{\text{grid}}\}$ are the grid points at which the baseline function $\alpha(\cdot)$ is estimated. In our simulation, we use a the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$, and $n_{\text{grid}} = 200$.

Local linear regression is employed to estimate the baseline function. In our simulation, we take a sample $n = 50$, $\tau = 20$ and set the bandwidth h equal to $h_0\times$(inter-quartile range of observed times t_{ij}). As discussed in Section 3.1, the regression coefficient $\boldsymbol{\beta}$ can be estimated by 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 based on 400 replicates are listed in Table 5. The biases and the standard deviations of the 400 estimated baseline functions at $t = 5$ are also depicted in Table 5. The results presented are very typical. It can be seen that the two approaches have almost the same performance. In addition, from Table 5, it can be seen that the biases are small, and, as the bandwidth increases, the biases increase, but the standard deviations decrease. Figure 1 depicts the typical estimated curves of $\alpha(t)$.

Similar to (2.16), the estimated standard error formula for the resulting estimator $\widehat{\alpha}(t)$ is given by $\mathbf{SCS}^T$. To test the accuracy of the standard error formula, the average of the 400 estimated standard errors and its standard deviation are computed and depicted in Table 5. The results indicate that the standard error formula performs very well.

Size and power calculation

We first verify empirically whether or not the GLR test statistic defined in Section 3 has a chi-square distribution under a null hypothesis. To this end, we take $H_0 : \alpha(t) = a_0 + a_1 t$. The values a_0 and a_1 are taken as follows. Generate a set of data under H_1, compute $\widehat{\boldsymbol{\beta}}$ by using a profile least-squares, obtain the partial residuals, and then fit a line to the partial residuals to obtain $\widehat{a}_0$ and $\widehat{a}_1$. In a sense, the line $\widehat{a}_0 + \widehat{a}_1 t$ is the closest member to $\alpha(t)$ among the linear functions. This poses the challenge of the GLR test in the power calculation below. Then we find the null distribution based on 2000 bootstrap simulations. Figures 2 (a) and (c) depict the estimated density of the GLR test statistic under H_0. We also plot the density of the chi-square distribution in order to examine whether the null distribution is close to a chi-square distribution. The degree of freedom is chosen to be the average of the 2000 bootstrap GLR statistics. It is clear from Figure 2 (a) that the Wilks type results hold.

To examine the power of the proposed nonparametric goodness of fit test, we evaluate the power of the GLR test for the alternative model

$$\alpha(t) = (1 - d_0)(a_0 + a_1 t) + d_0 \alpha_1(t),$$

for each given d_0, where $\alpha_1(t)$ equals $\tau\sqrt{t/\tau}$ with $\tau = 20$. We took $d_0 = 0, 0.05, \cdots, 0.25$. Plots of $\alpha(t)$ are depicted in Figure 2 (b). Figure 2 (c) depicts the three power functions based on 400 Monte Carlo simulations for the sample size, $n = 50$, at three different significance levels: 0.10, 0.05 and 0.01. The powers at $d_0 = 0$ for the foregoing three significance levels are: 0.0950, 0.0475, 0.0150, respectively, when $\alpha = \tau\sqrt{t/\tau}$ and the simulated sample was generated from Case I. This shows that the bootstrap method gives us an approximately right sized test. Further, it demonstrates that the GLR test is powerful, having power near one, even when the alternative model (dotted curve in Figure 2(b)) is so close to the null model (solid line in Figure 2(b)). Results for other cases are similar.

5 An Application

We now illustrate the proposed procedures in Sections 2 - 4 via an analysis of a subset of data from the Multi-Center AIDS Cohort study. The data set contains the HIV status of 283 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991. Details about the design, methods and medical implications of the study can be found in Kaslow *et al.*. (1987). During this study, all participants were scheduled to have their measurements taken during semi-annual visits, but, because many participants missed some of their scheduled visits, and the HIV infections happened randomly during the study, there are unequal numbers of repeated measurements and different measurement times per individual. Wu and Chiang (2000), Fan and Zhang (2000), Huang, Wu and Zhou (2002) analyzed the same data set by using varying-coefficient models. Their analysis aimed to describe the trend 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, denoted by Smoking, Age and PreCD4, respectively. Furthermore, the data were fitted by a varying-coefficient model:

$$y(t) = \beta_0(t) + \beta_1(t)\text{Smoking} + \beta_2(t)\text{Age}(t) + \beta_3(t)\text{PreCD4}(t) + \varepsilon(t). \qquad (5.1)$$

From the results of the hypothesis testing in Huang, Wu and Zhou (2002), only the baseline function varies over time, and PreCD4 has a constant effect over time, namely, $\beta_3(t) = \beta_3$. Neither Smoking nor Age has a significant impact on the mean CD4 percentage. This motivates us to use model (1.1) to fit this data set and to employ 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 $x_3(t)$ 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:

$$\begin{aligned} y(t) &= \alpha(t) + \beta_1 x_1 + \beta_2 x_2(t) + \beta_3 x_3(t) + \beta_4 x_2^2(t) + \beta_5 x_3^2(t) \\ &\quad + \beta_6 x_1 x_2(t) + \beta_7 x_1 x_3(t) + \beta_8 x_2(t) x_3(t) + \varepsilon(t). \end{aligned} \tag{5.2}$$

The DBE estimate for $\boldsymbol{\beta}$ was computed to obtain the partial residuals for $\alpha(\cdot)$, and then the bandwidth $h = 0.5912$ was selected by the Ruppert, Sheather and Wand (1995) plug-in method. After that the profile least squares method with weight $w(t) \equiv 1$ was applied to this model. The resulting estimates and standard errors are depicted in Table 6. We further applied the penalized profile least squares approach to select significant variables. The tuning parameter $\lambda = 0.7213$ 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, compared with the SCAD, the penalized profile least squares with the L_1 penalty shrinks the large coefficients more, and results in a small standard error. From Table 6, the result is in the line with that of Huang, Wu and Zhou (2002), but indicates possible interactions between Smoking status and Age: elder smokers tend to have lower average CD4 counts. The latter is plausible, since the impact of smoking on health tends to be more severe in older men.

Figure 3 depicts the estimated baseline function $\alpha(t)$ using the bandwidth $h = 0.5912$. We also plot 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 now test whether the baseline function is linear in time using the proposed generalized likelihood ratio test. The resulting test statistic is 47.0020 with a p-value 0.0020, which indicates the baseline function varies over time in a nonlinear pattern. This is also evidenced from Figure 3.

Appendix

A.1 Proof of Theorem 2.1

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

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

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

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

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

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

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

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

where

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

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

and

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

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

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

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

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

Using the model

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

and (A.1), we have

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

The minimization of this quadratic function is given by

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

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

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

where

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

and

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

Finally, we apply the convexity Lemma (see, for example, Anderson and Gill, 1982) to show that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \sqrt{n}\widehat{\Sigma}_n^{-1}\xi_n + o_P(1). \tag{A.7}$$

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

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

Similar to (A.4), it can be shown that

$$I'_{n,2}(\widehat{\boldsymbol{\beta}}) = o_P(n^{-1/2}), \quad I'_{n,3}(\widehat{\boldsymbol{\beta}}) = o_P(n^{-1/2}). \tag{A.8}$$

Hence the result follows.

A.2 Proof of Theorem 3.1

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

$$P\left\{\inf_{||\mathbf{u}||=C} \mathcal{L}(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) \geq \mathcal{L}(\boldsymbol{\beta}_0)\right\} \geq 1 - \eta. \tag{A.9}$$

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

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

where $\ell(\boldsymbol{\beta})$ is the first term in the right hand of (3.3). Using (A.5) and (A.8), it can be shown that

$$n^{-1}\{\ell(\boldsymbol{\beta}_0 + \alpha_n \mathbf{u}) - \ell(\boldsymbol{\beta}_0)\} = \frac{\alpha_n^2}{2}\mathbf{u}^T\{\widehat{\Sigma}_n + o_P(1)\}\mathbf{u} - \alpha_n \mathbf{u}^T\{\widehat{\xi}_n + o_P(n^{-1/2})\}, \tag{A.10}$$

as $\ell(\boldsymbol{\beta})$ is a quadratic function of $\boldsymbol{\beta}$. Note that $\widehat{\Sigma}_n \to \mathbf{A}$, a finite positive definite matrix, in probability. The first term in the right-hand side of (A.10) is of the order $O_P(C^2\alpha_n^2)$, and the second term is of the order $O_P(Cn^{-1/2}\alpha_n) = O_P(C\alpha_n^2)$. Furthermore,

$$\sum_{j=1}^{s} \{p_{\lambda_{jn}}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_{jn}}(|\beta_{j0}|)\} \tag{A.11}$$

is bounded by

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

by the Taylor expansion and the Cauchy-Schwarz inequality. As $b_n \to 0$, the first term on the right hand side of (A.10) will dominate (A.11) as well as the second term on the right hand side of (A.10), by taking C sufficiently large. Hence (A.9) holds for sufficiently large C. This completes the proof of the theorem.

A.3 Proof of Theorem 3.2

Lemma A.1 *Under the conditions of Theorem 3.2, 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,*

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

Proof. We are going to 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 $\|\boldsymbol{\beta}_2\| \leq Cn^{-1/2}$, $\partial\ell(\boldsymbol{\beta})/\partial\beta_j$ and β_j have the same signs for $\beta_j \in (-Cn^{-1/2}, Cn^{1/2})$, for $j = s+1, \cdots, d$. Thus, the minimizer attains at $\boldsymbol{\beta}_2 = 0$.

For $\beta_j \neq 0$ and $j = s+1, \cdots, d$,

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_j} = \ell_j'(\boldsymbol{\beta}) + np'_{\lambda_{jn}}(|\beta_j|)\text{sgn}(\beta_j), \tag{A.12}$$

where $\ell_j'(\boldsymbol{\beta}) = \partial\ell(\boldsymbol{\beta})/\partial\beta_j$. By the proof of Theorem 2.1,

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

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

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

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

Proof of Theorem 3.2: Part(i) directly follows by Lemma A.1. Now we prove Part (ii). Using argument similar to the proof of Theorem 3.1, it can be shown that there exists a $\widehat{\boldsymbol{\beta}}_1$ in Theorem 3.1 that is a root n consistent local minimizer of $\mathcal{L}\{(\boldsymbol{\beta}_1^T, 0)^T\}$, satisfying the penalized least squares equations:

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

Following the proof of Theorem 2.1, we have

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

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

Therefore, similar to the proof of Theorem 2.1 and by Slutsky's Theorem, it follows that

$$\sqrt{n}(\mathbf{A}_{11} + \Sigma)\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{A}_{11} + \Sigma)^{-1}\mathbf{b}\} \to N_s(\mathbf{0}, \mathbf{B}_{11}). \qquad \text{(A.13)}$$

This completes the proof of Theorem 3.2.

References

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.*, **10**, 1100–1120.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations (with discussions). *Jour. Amer. Statist. Assoc.*, **96**, 939-967.

Bowman, A.W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis.* Oxford University Press, Oxford.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24**, 2350-2383.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inference for varying-coefficient models. *Jour. Amer. Statist. Assoc.*, **95**, 888-902.

Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P. (1997). Generalized partially linear single-index models. *Jour. Amer. Statist. Assoc.*, **92**, 477-489

Chiang, C.-T., Rice, J.A. and Wu, C.O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Jour. Amer. Statist. Assoc.*, **96**, 605-619.

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

Fan, J. (1992). Design-adaptive nonparametric regression. *Jour. Amer. Statist. Assoc.*, **87**, 998–1004.

Fan, J. and Huang, L. (2001). Goodness-of-fit test for parametric regression models, *Jour. Amer. Statist. Assoc.*, **96**, 640-652.

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

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks Phenomenon, *Ann. Statist.*. **29**, 153-193.

Frank, I.E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-148.

Hart, J.D. (1997). *Nonparametric Smoothing and Lack-of-fit Tests.* Springer, New York.

Hart, J.D. and Wehrly, T.E. (1988). Kernel regression estimation using repeated measurements data. *Jour. Amer. Statist. Assoc.*, **81**, 1080-1088.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models (with discussion). *Jour. Royal Statist. Soc., B*, **55**, 757-796.

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. *Am. J. Epidem.*, **126**, 310-318.

Lin, D.Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data (with discussions). *Jour. Amer. Statist. Assoc.*, **96**, 103-126.

Lin, X. and Carroll, R.J. (2001a). Comment on "Semiparametric and nonparametric regression analysis of longitudinal data". *Jour. Amer. Statist. Assoc.*, **96**, 114-116.

Lin, X. and Carroll, R.J. (2001b). Semiparametric regression for clustered data using generalized estimating equations. *Jour. Amer. Statist. Assoc.*, **96**, 1045-1056.

Martinussen, T. and Scheike, T.H. (1999). A semiparametric additive regression model for longitudinal data. *Biometrika*, **86**, 691-702.

Moyeed, R.A. and Diggle, P.J. (1994). Rates of convergence in semiparametric modeling of longitudinal data. *Austr. Jour. Statist.*, **36**, 75-93.

Müller, H.G. and Stadtmüller U. (1993). On variance function estimation with quadratic forms. *J. Statist. Plann. Inf.* **35**, 213-31.

Pepe, M.S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time-dependent covariates. *Jour. Amer. Statist. Assoc.*, **88**, 811-820.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *Jour. Amer. Statist. Assoc.*, **90**, 1257-1270.

Severini, T.A. and Staniswalis, J.G. (1994). Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.*, **89**, 501–511.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Jour. Royal Statist. Soc. B*, **50**, 413–436.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Jour. Royal Statist. Soc., B*, **58**, 267-288.

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

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. *Jour. Amer. Statist. Assoc.*, **88**, 1388-1402.

Wu, C.O. and Chiang, C.T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, **10**, 433-456.

Yatchew, A. (1997). An elementary estimator of the partial linear model. *Economics Letters*, **57**, 135-143.

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.

Table 1: Ratios of MSEs

Case I		$n = 50$		$n = 75$	
$\alpha(t)$	τ	DBE	Profile LSE	DBE	Profile LSE
$\tau\sqrt{t/\tau}$	4	0.8481	0.6592	0.8407	0.6661
$\tau\sqrt{t/\tau}$	20	0.3450	0.2962	0.3963	0.3246
$\tau\sin(2\pi t/\tau)$	4	0.6632	0.5065	0.6756	0.5377
$\tau\sin(2\pi t/\tau)$	20	0.2798	0.2324	0.3001	0.2359
Case II					
$\tau\sqrt{t/\tau}$	4	0.7868	0.6500	0.7209	0.6004
$\tau\sqrt{t/\tau}$	20	0.2518	0.2138	0.2438	0.2015
$\tau\sin(2\pi t/\tau)$	4	0.5627	0.4641	0.5409	0.4502
$\tau\sin(2\pi t/\tau)$	20	0.1705	0.1395	0.1623	0.1280
Case III					
$\tau\sqrt{t/\tau}$	10	1.0785	0.7040	1.1299	0.7348
$\tau\sqrt{t/\tau}$	20	0.7748	0.5006	0.8316	0.5424
$\tau\sin(2\pi t/\tau)$	10	1.2188	0.6818	1.2560	0.7347
$\tau\sin(2\pi t/\tau)$	20	0.9868	0.5007	1.0721	0.5973
Case IV					
$30\sqrt{t/30}$		0.9666	0.6842	1.0501	0.7086
$30\sin(2\pi t/30)$		0.1434	0.0953	0.1360	0.0869

Table 2: Stds and SEs of Profile LSE for Case I with $\alpha(t) = \tau\sqrt{t/\tau}$

(n, τ)	β_1		β_2		β_5	
	std	se (std(se))	std	se (std(se))	std	se (std(se))
(50, 4)	0.1512	0.1377(0.0327)	0.1683	0.1579(0.0369)	0.1664	0.1543(0.0381)
(75, 4)	0.1200	0.1148(0.0211)	0.1262	0.1273(0.0243)	0.1274	0.1287(0.0240)
(50, 20)	0.0854	0.0820(0.0182)	0.1004	0.0910(0.0211)	0.1012	0.0933(0.0203)
(75, 20)	0.0651	0.0675(0.0130)	0.0718	0.0748(0.0144)	0.0708	0.0749(0.0149)

Table 3: Comparison of Variable Selection Procedures

	$\alpha(t) = \tau\sqrt{t/\tau}$			$\alpha(t) = \tau\sin(2\pi t/\tau)$		
Method	RGMSE	Zero Coefficient		RGMSE	Zero Coefficient	
Case I: $n = 50, \tau = 20$						
Method	mean (std)	C	I	mean (std)	C	I
L_1	0.3936(0.2966)	4.9950	0	0.3923(0.2863)	4.9900	0
SCAD	0.3549(0.2453)	4.9950	0	0.3533(0.2453)	4.9925	0
Oracle	0.3502(0.2412)	5.0000	0	0.3480(0.2425)	5.0000	0
Case II: $n = 75, \tau = 4$						
L_1	0.5772(0.2614)	4.3325	0	0.5733(0.2648)	4.3500	0
SCAD	0.5127(0.2101)	4.4275	0	0.5115(0.2107)	4.4250	0
Oracle	0.3939(0.2326)	5.0000	0	0.3915(0.2318)	5.0000	0
Case III: $n = 50, \tau = 20$						
L_1	0.3975(0.2843)	4.9950	0	0.4002(0.2860)	4.9975	0
SCAD	0.3450(0.2278)	4.9975	0	0.3460(0.2279)	4.9975	0
Oracle	0.3438(0.2269)	5.0000	0	0.3450(0.2271)	5.0000	0
Case IV: $n = 50$, $\tau = 30$						
L_1	0.4091 (0.2716)	4.9975	0	0.4074 (0.2717)	5.0000	0
SCAD	0.3554 (0.2210)	5.0000	0	0.3546 (0.2205)	5.0000	0
Oracle	0.3549 (0.2200)	5.0000	0	0.3542 (0.2199)	5.0000	0

Table 4: Standard Deviations and Standard Errors of $\widehat{\boldsymbol{\beta}}$

	β_1		β_2		β_5	
	std	se (std(se))	std	se (std(se))	std	se (std(se))
L_1	0.0823	0.0798 (0.0176)	0.0826	0.0775 (0.0177)	0.0735	0.0702 (0.0166)
SCAD	0.0810	0.0808 (0.0180)	0.0821	0.0793 (0.0187)	0.0738	0.0708 (0.0169)
Oracle	0.0808	0.0808 (0.0181)	0.0810	0.0794 (0.0188)	0.0737	0.0709 (0.0169)

Table 5: Summary of Simulation Results for $\widehat{\alpha}(t)$ $(\alpha(t) = \tau\sqrt{t/\tau})$

	RASE	$t = 5$		
h_0	mean(std)	bias	std	se (std(se))
DBE $\widehat{\boldsymbol{\beta}}$				
0.25	0.3195 (0.0665))	-0.0284	0.1361	0.1242 (0.0222))
0.35	0.3519 (0.0630))	-0.0661	0.1219	0.1077 (0.0174)
0.45	0.3889 (0.0605))	-0.1207	0.1134	0.1007 (0.0142)
PLS $\widehat{\boldsymbol{\beta}}$				
0.25	0.3195 (0.0656))	-0.0310	0.1349	0.1233 (0.0220)
0.35	0.3525 (0.0618))	-0.0686	0.1206	0.1067 (0.0172)
0.45	0.3900 (0.0592))	-0.1232	0.1118	0.0996 (0.0141)

Table 6: Estimated Coefficients for Model (5.2)

Variable	Profile LS	L_1	SCAD
	$\widehat{\beta}(\text{se}(\widehat{\beta}))$	$\widehat{\beta}(\text{se}(\widehat{\beta}))$	$\widehat{\beta}(\text{se}(\widehat{\beta}))$
Smoking	0.5333(1.0972)	0(0)	0(0)
Age	-0.1010(0.9167)	0(0)	0(0)
PreCD4	2.8252(0.8244)	3.0932(0.5500)	3.1993(0.5699)
Age^2	0.1171(0.4558)	0(0)	0(0)
PreCD4^2	-0.0333(0.3269)	0(0)	0(0)
Smoking*Age	-1.7084(1.1192)	-0.9684(0.4904)	-1.0581(0.5221)
Smoking*PreCD4	1.3277(1.3125)	0(0)	0(0)
Age*PreCD4	-0.1360(0.5413)	0(0)	0(0)

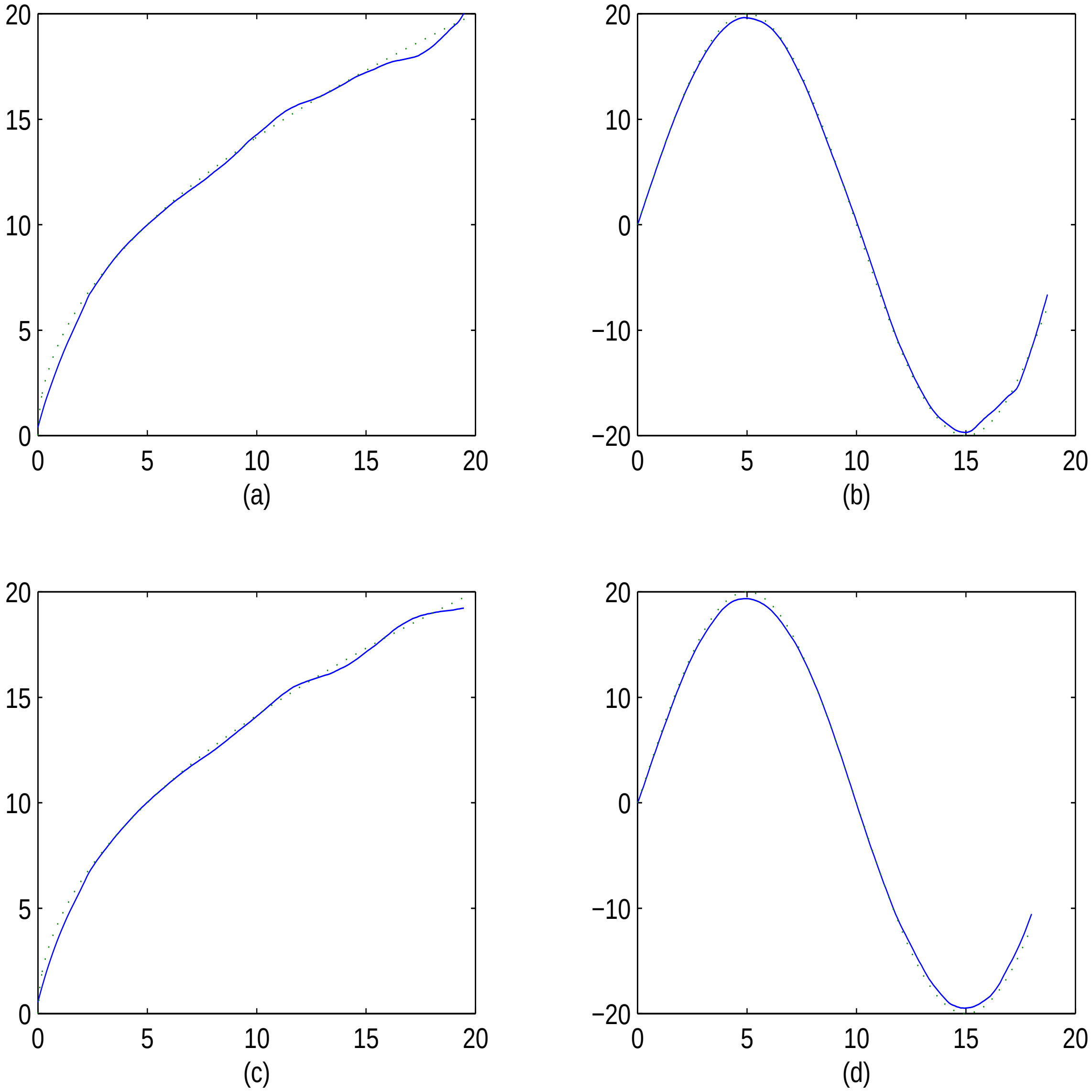

Figure 1: *Typical Estimated Baseline Curves with $n = 50$ and $\tau = 20$. Solid lines stands for estimated curves of $\alpha(t)$, and dotted lines for the true $\alpha(t)$. (a) and (c) are estimated baseline function $\alpha(t) = \tau\sqrt{t/\tau}$ using bandwidth $0.3\times IQR$ for Cases I and II, respectively, when $n = 50$ and $\tau = 20$. (b) and (d) are estimated baseline function $\alpha(t) = \tau\sin(t/\tau)$ using bandwidth $0.2\times IQR$ for Cases I and II, respectively.*

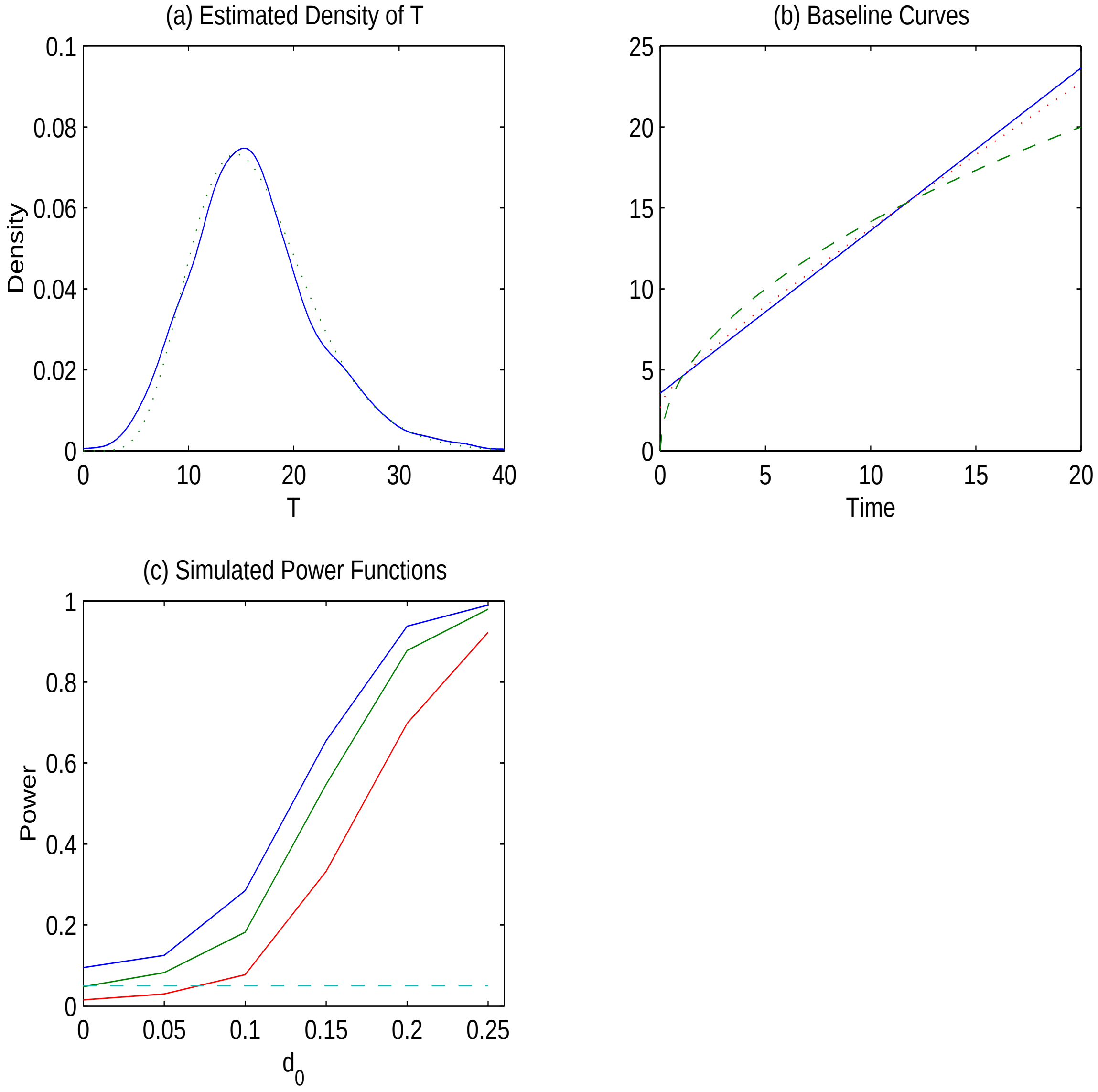

Figure 2: *Plots of estimated density curve, baseline curves and simulated power functions for Case I with* $\alpha(t) = \tau\sqrt{t/\tau}$ *and* $\tau = 20$. *(a) is the estimated density curve of test statistic under* H_0*; the solid line stands for the estimated null density obtained via using bootstrap, and dotted line is the density of a chi-square distributions. (b) is baseline curves* $\alpha(t) = (1 - d_0)(3.5560 + 1.0041t) + d_0\tau\sqrt{t/\tau}$. *The solid line stands for* $d_0 = 0$*, and corresponds to the baseline curve under the null hypothesis, the dotted line stands for* $d_0 = 0.25$ *and the dashed line for* $d_0 = 1$. *(c) is simulated power functions. From the bottom to the top in (c), the simulate power functions corresponds to level at 0.01, 0.05, 0.10. The horizontal dash line stands for the significance level* $\alpha = 0.05$.

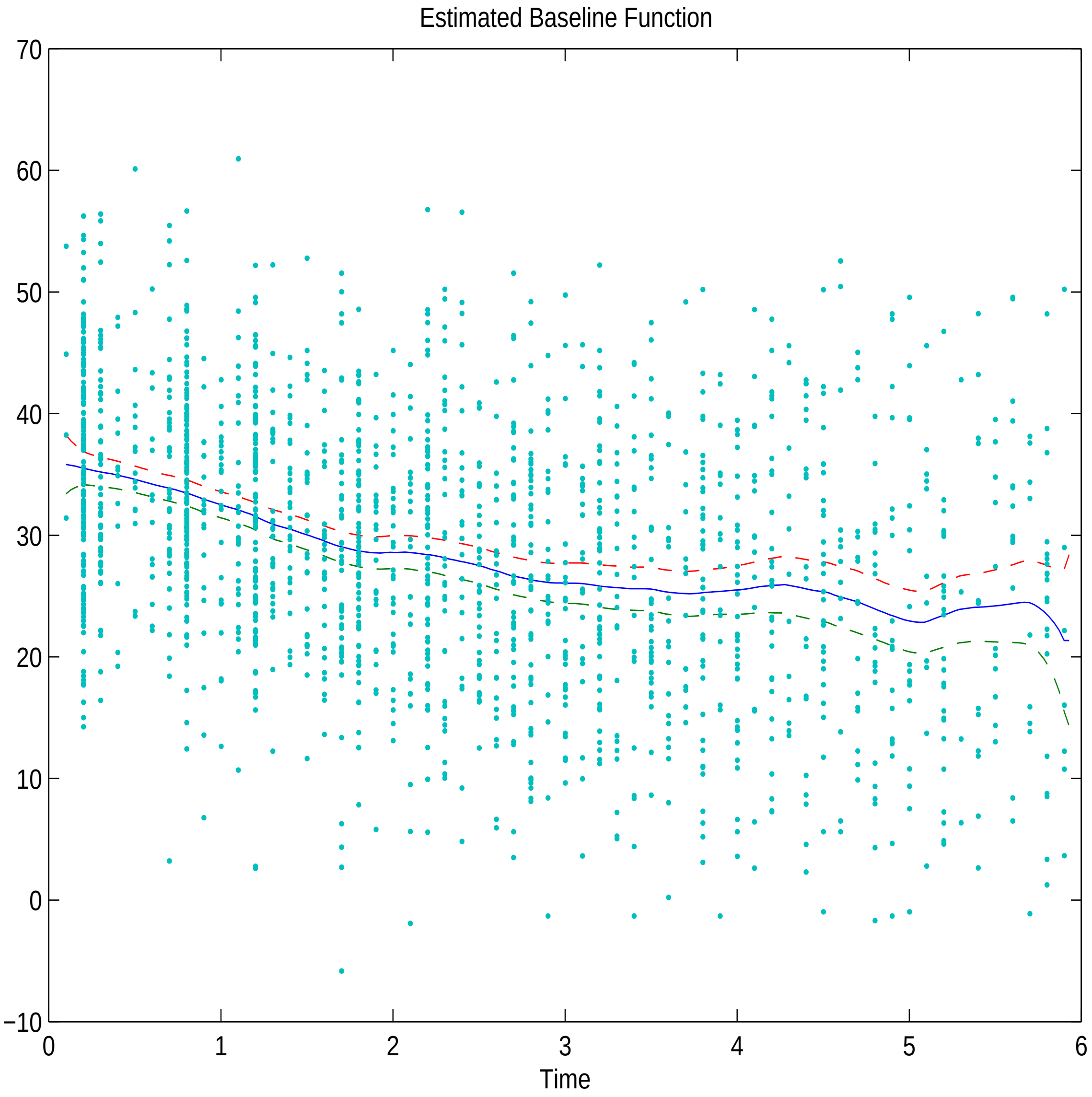

Figure 3: *Estimated Baseline Function. The solid line stands for the estimated baseline function, the dash lines are the estimated baseline function plus/minus twice standard errors. The dots are the residual on parametric part* $r(t) = y(t) - \widehat{\boldsymbol{\beta}}^T \mathbf{x}(t)$.