Abstract

NI, XIAO. Variable Selection in Partial Linear Models and Semiparametric Mixed Models. (Under the direction of Professors Daowen Zhang and Hao Helen Zhang).

We consider the variable selection problem in two popular semiparametric regression models. For independent data, partial linear models provide good compromises between linear and nonparametric models and have proved very useful for data analysis. Semiparametric mixed models (Zhang et al., 1998) are useful techniques for longitudinal data which model a nonparametric baseline function and complex variance structures. Due to the model complexity, model selection is challenging in these semiparametric regression models. We propose a simple and unified approach for selecting variables in these models. A new type of penalized least squares/likelihood function with double penalty is formulated, using the smoothing spline to estimate the nonparametric part and a shrinkage penalty to achieve parsimony in linear effects at the same time. Distinguished from other methods, the proposed procedure has a linear mixed model (LMM) representation, which greatly facilitates its implementation by using any standard software or statistical packages. Another advantage of the LMM representation is that it allows us to treat the smoothing parameter as a variance component and hence conveniently estimate it together with other regression coefficients. In theory, we show that with proper choices of smoothing and regularization parameters, the proposed variable selection procedure asymptotically performs as well as an oracle estimator. Both frequentist and Bayesian estimates of the covariance and confidence intervals for the estimates are derived. Simulated and real examples show that the new procedure compares favorably with existing methods.

VARIABLE SELECTION IN PARTIAL LINEAR MODELS AND SEMIPARAMETRIC MIXED MODELS

by

XIAO NI

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

STATISTICS

Raleigh, NC

2007

APPROVED BY:

Dr. Daowen Zhang Chair of Advisory Committee	Dr. Hao Helen Zhang Co-Chair of Advisory Committee
Dr. Marie Davidian	Dr. Jason Osborne

Dedication

To My Family

Biography

Xiao Ni was born on July 7, 1978 in Taian, China. He graduated from Xi'an Jiaotong University with a bachelor's degree in Electrical Engineering in 2000. He came to University of Georgia to study Artificial Intelligence and obtained a master's degree in 2002. In 2003 he entered the Department of Statistics at North Carolina State University to study Statistics. Upon completion of his doctorial degree, he will start to work at GlaxoSmithKline as a senior statistician.

Acknowledgements

I would like to express sincere gratitude to my advisors Dr. Daowen Zhang and Dr. Hao Helen Zhang for their constant guidance, encouragement and support in my study. I also thank the other committee members Dr. Marie Davidian and Dr. Jason Osborne for their insights and comments on my dissertation.

It has been a pleasant experience to study in this department, and in this university. I would like to convey my gratitude to all the people who have made the resources available to me, without which I would not be able to complete my study. Special thanks go to Dr. Bill Swallow, for his kind support and encouragement. As a graduate industrial trainee at GlaxoSmithKline, I also thank my manager David Cooper for his guidance and support.

Finally, I would like to thank my parents in China, and my brother in South Carolina, for their unconditional love, encouragement and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Semiparametric regression refers to regression models in which the predictors contain both parametric and nonparametric components. Compared with the ordinary parametric regression, semiparametric regression offers the flexibility to incorporate nonlinear functional relationships. This dissertation concerns model selection and estimation in two common semiparametric regression models: partial linear models for independent data and semiparametric stochastic mixed models for longitudinal data.

1.1 Semiparametric Regression Models

1.1.1 Partial Linear Models

The partial linear model is a popular semiparametric modeling technique which assumes that the relationship between the response variable and the covariates can be represented as

$$Y = \mathbf{X}^T\boldsymbol{\beta} + f(T) + \varepsilon, \tag{1.1}$$

where $\mathbf{X}$ is a $d \times 1$ vector of explanatory variables, T is a scalar covariate, $f(\cdot)$ is an unknown smooth function of T, and ε's are independent random errors with mean zero. Partial linear models (1.1) are special cases of generalized additive models, which were introduced by Hastie and Tibshirani (1990) to solve the curse of dimensionality in multivariate nonparametric regression models. With both parametric and nonparametric components, the model (1.1) is more flexible than the traditional linear model. It is particularly useful when the response variable Y is linearly dependent on $\mathbf{X}$ yet nonlinearly related to covariate T.

Engle et al. (1986) first proposed partial linear models for analyzing the relationship between temperature and electricity usage. Since then partial linear models have gained increasing popularity in a variety of applications, including economics, biometrics and environmental science. For example, Schmalensee and Stoker (1999) used partial lineal models to analyze the United States gasoline consumption data. In environmetric studies, partial linear models have been used to predict Atmospheric SO_2 concentrations (Prada-Sanchez et al. 2000).

Estimation of $\boldsymbol{\beta}$ and $f(t)$ in (1.1) and their theoretical properties have been studied much in literature. Speckman (1988) introduced the idea of partial residuals and and a profile least squares estimator. Robinson (1988) used a Nadaraya-Waston kernel estimator for estimating $f(t)$ and proposed a least squares estimator for $\boldsymbol{\beta}$. Engle et al. (1986), Heckman (1986), Green (1987), Rice (1986) and Green and Silverman (1994) used smoothing splines and proposed a penalized least squares approach. Ruppert et al. (2003) and Liang (2006) estimated $\boldsymbol{\beta}$ and $f(t)$ in a penalized spline context. Heckman (1986) showed that the penalized least squares estimator $\widehat{\boldsymbol{\beta}}$ can be root-n

consistent for independent $\mathbf{X}$ and T under certain regularity conditions. Rice (1986) examined the asymptotic bias of $\widehat{\boldsymbol{\beta}}$ when $\mathbf{X}$ and T are dependent.

1.1.2 Semiparametric Models for Longitudinal Data

The partial linear model (1.1) assumes independence among data. It is natural to extend it to handle correlated data - most commonly longitudinal/clustered data. The defining characteristic of longitudinal data is that experimental units (or subjects) are measured repeatedly over time. Repeated observations on the same subjects tend to be intercorrelated. Therefore special statistical methods are necessary to incorporate the correlation structure in order to draw valid statistical inferences.

Analysis of longitudinal data usually falls into two categories: *marginal models* and *random effects models*. Diggle et al. (2002) provides a detailed survey on these two approaches. In a marginal model, we model the marginal expectation of the response variable as a function of the explanatory variables, and the within-subject correlation is modeled separately. The marginal model approach is appropriate when the center of inference is *population average*. The generalized estimation equations (GEE) method (Liang and Zeger, 1986) is a classic marginal model approach for longitudinal data analysis, which yields consistent estimates even with mis-specified correlation structures. In a random effect model, the effects associated with different subjects are viewed as random samples from a population. Classic linear mixed models (LMM) provide a flexible theoretical and computational framework for longitudinal data analysis (Lard and Ware, 1982; Verbeke and Molenberghs, 2000). In this dissertation, we adopt the random effect model approach and compare it to marginal

model approaches.

Parametric models are not always appropriate when the model assumption is not realistic and is likely to introduce modeling biases. To relax the parametric assumptions, various semiparametric models have been developed for longitudinal data. Different smoothing techniques have been used in these models, including kernel smoothing (see, e.g., Diggle et al., 2002; Chen and Jin, 2006), smoothing splines (see, e.g., Zhang et al., 1998; Wang, 1998), B-splines (He et al., 2002), penalized splines (Ruppert et al., 2003) and local polynomial regression (Fan and Li, 2004). Semiparametric stochastic mixed models (SPMM; Diggle et al., 2002; Zhang et al., 1998) are useful extensions to linear mixed models, which use parametric fixed effects to represent the covariate effects, an arbitrary smooth function to model the time effect, and accounts for the within-subject correlation using random effects and a stationary or nonstationary stochastic process. Unlike the kernel based estimation in Diggle et al. (2002), Zhang et al. (1998) estimated the nonparametric baseline function using smoothing splines by maximizing the penalized likelihood. The most attractive feature is that the smoothing parameter can be treated as an extra variance component in a modified linear mixed model framework and can hence be jointly estimated by restricted maximum likelihood (REML). Because of the generality of the model and great computational advantage, the SPMM is a preferred method for longitudinal data analysis.

1.2 Variable Selection in Semiparametric Regression Models

Variable selection is an important issue in regression analysis. In practice, there are often a large number of covariates or explanatory variables whereas not all of these variables are predicative to the response. Variable selection is therefore necessary to improve model interpretability and prediction accuracy. In this section we briefly survey variable selection methods in linear regression methods, which sheds light on selection of important linear effects in a more general framework - semiparametric regression models.

1.2.1 Variable Selection in Linear Regression Models

For linear regression models, variable selection method can be roughly classified into best subset selection, stepwise regression and shrinkage methods. For a classic linear model with d regressors, there are a total of 2^d possible sub-models. The best subset selection method exhaustively searches through all 2^d models and rank them by some numerical criterion. Such criteria include adjusted R^2, Mallows' C_p (Mallows, 1973), Akaike's Information Criteria (AIC; Akaike, 1973, 1977) and Bayesian Information Criteria (BIC; Schwartz, 1978), etc. Mallows' C_p for a sub-model with q variables defined by

$$C_p = \frac{RSS_q}{\widehat{\sigma}_d^2} + 2q - n,$$

where RSS_q is the residual sum of squares from this sub model fit and $\widehat{\sigma}_d^2$ is the unbiased estimate of σ^2 based on the full model. Mallows' C_p was based on the

attempt to minimize the mean squared error for prediction. The AIC and BIC are information criteria based on likelihood structures. For the log-likelihood of the sub-model with q predictors denoted by ℓ_q, the AIC selects the model which minimizes $-2\ell_q + 2q$; whereas BIC selects the model which minimizes $-2\ell_q + q\ log(n)$.

Even for a moderate d, computation for all 2^d subsets is prohibitively expensive. In this case, a more computationally efficient approach is stepwise selection, which follows a systematic selection path to the final model based on adding/removing one variable at a time. Stepwise regression methods include forward selection and backward elimination. Forward selection starts with no predictors in the model and adds regressors one at a time to the model according to the partial F-statistic until the F-statistic does not exceed a pre-specified threshold. On the other hand, backward selection begins with the full model with all variables and eliminate regressors one-by-one according to the partial F-statistic.

Although the subset selection and stepwise selection methods are useful for variable selection, they have a few intrinsic drawbacks. First, these methods are often very time-consuming and inefficient. Secondly, as pointed out by Breiman (1995), these methods suffer from instability and are sensitive to noise. Thirdly, they tend to ignore the stochastic errors due to their discrete selection process. To achieve better prediction and reduce variances of estimators, many shrinkage estimation approaches have been proposed. The shrinkage methods add a penalty term to the residual sum of squares, for example,

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\beta)^2 + \sum_{j=1}^{d} p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ, which penalizes regression coefficients and shrinks small coefficients to exactly zero. Several forms of $p_\lambda(\cdot)$ have been proposed, including

- L_2 penalty: $p_\lambda(|\theta|) = \lambda|\theta|^2$, also known as ridge regression;
- L_1 penalty: $p_\lambda(|\theta|) = \lambda|\theta|$, also known as (LASSO; Tibshirani, 1996);
- L_q penalty: $p_\lambda(|\theta|) = \lambda|\theta|^q$, $q \geq 0$, also known as bridge regression (Frank and Friedman, 1993).

These shrinkage methods have gained a lot of popularity due to their capability of simultaneously selecting variables and estimating coefficients in a continuous fashion. However, Fan and Li (2001) argued that an ideal penalty function should yield (1) asymptotically unbiased, (2) continuous, (3) sparse estimates, which cannot be achieved by these aforementioned penalty functions. They proposed the smoothly clipped absolute deviation (SCAD) penalty that possesses all three good properties. The SCAD is essentially a symmetric quadratic spline function defined by

$$p'_\lambda(|\theta|) = \begin{cases} \lambda & \text{if } 0 \leq |\theta| \leq \lambda \\ \frac{a\lambda - |\theta|}{a-1} & \text{if } \lambda_2 < |\theta| \leq a\lambda \\ 0 & \text{if } |\theta| > a\lambda \end{cases},$$

where $a > 2$ and λ are two tuning parameters. A plot for SCAD function with $a = 3.7$ and $\lambda = 0.8$ is presented in Figure 1.1. In this thesis, we choose the SCAD penalty for variable selection due to its good properties.

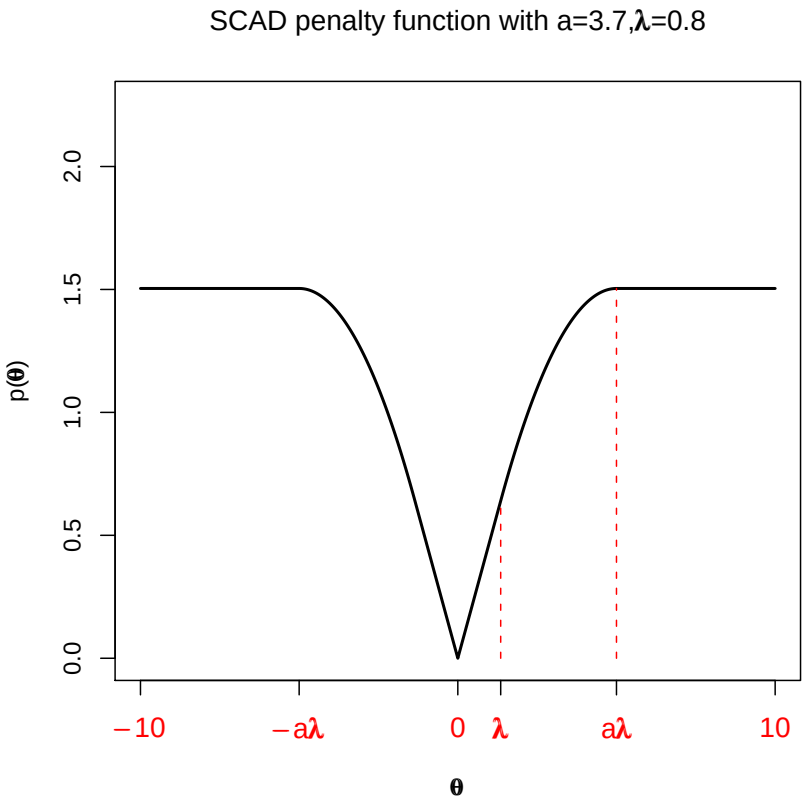

Figure 1.1: Plot of the SCAD penalty function, with $a = 3.7$ and $\lambda = 0.8$.

1.2.2 Variable Selection in Semiparametric Regression Models

Although there is rich literature on variable selection for parametric linear models, very limited work has been done for semiparametric regression models, such as the partial linear models (1.1) and the semiparametric mixed model described in the previous section. Fan and Li (2004) first investigated this problem in longitudinal data context. They developed a penalized least squares approach using local polynomial regressions for estimating the nonparametric baseline function. Their method

was based on a marginal model and essentially ignored the correlation structure by using working independence in implementation. In this thesis, we develop a new approach for variable selection in semiparametric regression models. For partial linear models with independent data, we first propose to minimize a unified double penalized least squares function equipped with two penalty terms, a roughness penalty for the nonparametric component and a shrinkage penalty for the regression coefficients for variable selection. We further extend our method to simultaneously select important linear effects and estimate parameters in semiparametric mixed models. Different from existing methods, our approach has a mixed model representation for the original semiparametric regression models and treats the smoothing parameter as an additional variance component. This allows us to take advantage of existing software for computation. Simulation results show that our methods have better finite sample performance in terms of both model selection and parameter estimation in finite samples than existing methods. Furthermore, our method is based on likelihood structure and therefore is still valid under missing at random (MAR) assumption if there are missing data.

The rest of this thesis is organized as follows. In Chapter 2 we develop automatic model selection methods for partial linear models. In Chapter 3 we extend the methodology to longitudinal (correlation) data settings and work on model selection in semiparametric stochastic mixed models. Separate simulations results and discussions are given in both chapters. Chapter 4 summarizes the thesis and concludes it with a discussion.

Chapter 2

Variable Selection in Partial Linear Models

2.1 Introduction

Partial linear models are popular semiparametric modeling techniques which assume the mean response of interest to be linearly dependent on some covariates, whereas its relation to one or more additional variables is inadequately characterized by parametric functions. They are special cases of the general additive models (Hastie and Tibshirani, 1990). We consider a partial linear model with the form (1.1). Estimation of $\boldsymbol{\beta}$ and f has been studied in various contexts including kernel smoothing (Speckman, 1988), smoothing splines (Engle et al., 1986; Heckman, 1986; Green and Silverman, 1994), and penalized splines (Ruppert et al., 2003; Liang, 2006).

In this paper, we focus on the model selection problem for partial linear models. Due to the complex nature of partial linear models, model selection involves two issues: smoothing parameter selection for the nonparametric part and variable selection for linear covariates. Often times, the number of potential explanatory variables,

d, is large, a variable selection process is necessary in order to improve prediction accuracy and model interpretability of fitted models. Numerous classical and modern procedures have been developed in literature such as stepwise selection, best subset selection, and a rich class of shrinkage methods including bridge regression (Frank and Friedman, 1993), nonnegative garrote (Breiman, 1995), least absolute selection and shrinkage operator (LASSO; Tibshirani, 1996), smoothly clipped absolute deviation (SCAD; Fan and Li, 2001, 2004), least angle regression (LARS; Efron et al., 2004), elastic net (Zou and Hastie, 2005), adaptive lasso (Zou, 2006) and etc. Various information criteria commonly used for model comparison include Mallows' C_p (Mallows, 1973, 1975), Akaike's Information Criteria (Akaike, 1973, 1977) and Bayesian Information Criteria (Schwarz, 1978). A thorough review on variable selection for linear models is given in Linhart and Zucchini (1986), Rao and Wu (2001), and Miller (2002).

Though there is a vast amount of work on variable selection for linear models, very limited work has been done on model selection for partial linear models. This was first noted in Fan and Li (Fan and Li, 2004), where a new model selection procedure was developed in the local polynomial regression setup and shown to give effective performance for longitudinal data analysis. Motivated by Fan and Li's work, we propose a double penalized least squares procedure for model selection in the context of partial smoothing splines. We will employ the SCAD penalty on linear coefficients in (1.1) to select important covariates. There are two tuning parameters in the proposed procedure due to its double penalty form: the smoothing parameter associated with the roughness penalty and the regularization parameter associated with the scad

penalty. We present an idea to avoid two-dimensional grid search, and show it works effectively in practice. Compared to few existing variable selection methods for partial linear models, our procedure is unique in the following aspects.

• It enhances standard partial spline methods by allowing automatic selection of important linear effects.

• Unlike the two-step procedures used in Fan and Li (2004), the new procedure has a unified objective function, which allows the estimation of the nonparametric part and the selection of important linear effects implemented simultaneously and hence improve finite sampling estimation efficiency.

• The new procedure has a linear mixed model (LMM) representation, allowing us to take advantage of standard software for implementation and hence dramatically reducing computational cost. Furthermore, this LMM framework provides a convenient way to make inferences on the nonparametric function as part of the overall inference procedure.

• Instead of using data-driven methods to select the smoothing parameter via grid search, the new procedure treats the smoothing parameter as an additional variance component which can be readily estimated by the restricted maximum likelihood (REML) approach.

The proposed procedure is shown to have oracle properties asymptotically when tuning parameters are chosen properly. This result is similar to Fan and Li (2001,2004). The rest of the article is organized as follows. In Section 2.2 we introduce the double penalized least squares method for simultaneous variable selection and model estimation. Asymptotic properties of the resulting estimator $\widehat{\boldsymbol{\beta}}$ are established. Section 2.3

derives the linear mixed model (LMM) representation for the proposed procedure, leading to an iterative algorithm which is easy to implement. Section 2.4 suggests a feasible way to choose the smoothing parameter for roughness penalty and the regulation parameter for the scad penalty. Furthermore, a REML-based estimate for the error variance is provided. Both frequentist and Bayesian covariance estimates for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{f}}$ are derived in Section 2.5. Section 2.6 presents numerous simulation results, and Section 2.7 illustrates the application of our method to the *Ragweed Pollen Level* data. Section 2.8 concludes the article with a discussion.

2.2 Double-Penalized Least Squares Estimators and Their Asymptotic Properties

In this section, we introduce the double-penalized least squares (DPLS) for partial linear models, which estimates the nonparametric function using a smoothing spline and at the same time selects important covariates for the parametric component. Under regularity conditions, we will establish the asymptotic normality and oracle properties of the double-penalized least squares estimator (DPLSE) $\widehat{\boldsymbol{\beta}}$.

2.2.1 Double-penalized Least Squares

Suppose that the sample consists of n observations. For the ith observation, denote by y_i the response, by $\boldsymbol{x}_i$ the covariate vector from which important covariates are to be selected, and by t_i the covariate whose effect cannot be adequately characterized

by a parametric function. We consider the following partial linear model:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{\beta}$ is a $d \times 1$ vector of regression coefficients, $f(t)$ is an arbitrary twice-differentiable smooth function, and ε_i's are assumed to be uncorrelated random variables with mean zero and common unknown variance σ^2. Define $\mathbf{Y} = (y_1, \cdots, y_n)^T$. Without loss of generality, we further assume that $t_i \in [0, 1]$ and $f(t)$ is in the Sobolev space $\{f(t) : f, f' \text{ are absolutely continuous}, and J^2(f) < \infty\}$, where $J^2(f) = \int_0^1 [f''(t)]^2 dt$.

We intend to simultaneously estimate the nonparametric function $f(t)$ and select important variables from $\boldsymbol{x}$ using the observed data. To this end, we propose a double-penalized least squares (DPLS) approach by minimizing

$$L_{dp}(\boldsymbol{\beta}, f(\cdot); \mathbf{Y}) = \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left\{ y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} - f(t_i) \right\}^2 + \frac{n\lambda_1}{2} \int_0^1 \{f''(t)\}^2 dt + n \sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|). \tag{2.2}$$

The first penalty term in the DPLS penalizes the roughness of the nonparametric component $f(t)$ and the second penalty term $p_{\lambda_2}(|\beta_j|)$ is the shrinkage penalty imposed on β_j's. To our best knowledge, there has been little discussion about DPLS in literature. Lin and Zhang (1999) considered using double-penalized quasi-likelihood in generalized additive mixed models, whereas not in the variable selection context. We call the minimizer of (2.2) double-penalized least squares estimators (DPLSEs) of $\boldsymbol{\beta}$ and $f(t)$. There are two tuning parameters in the objective function (2.2): $\lambda_1 \geq 0$ is, as usual, a smoothing parameter which balances smoothness of $f(t)$ with fidelity to the data, and $\lambda_2 \geq 0$ is a regularization parameter controlling the amount of shrink-

age used in the variable selection. Choices of tuning parameters are very important to assure effective model selection and estimation, which will be discussed later.

In the DPLS (2.2), we adopt the nonconcave SCAD penalty proposed by Fan and Li (2001), which satisfies

$$p'_{\lambda_2}(|\omega|) = \begin{cases} \lambda_2 & \text{if } 0 \le |\omega| \le \lambda_2 \\ \frac{a\lambda_2 - |\omega|}{a-1} & \text{if } \lambda_2 < |\omega| \le a\lambda_2 \\ 0 & \text{if } |\omega| > a\lambda_2 \end{cases} \qquad \text{for} \quad \omega > 0, \tag{2.3}$$

where $a > 2$ is also a tuning parameter. They showed that the SCAD penalty function results in consistent, sparse and continuous estimators in linear models, and have successfully extended it to proportional hazard models for survival data with partial likelihood (Fan and Li, 2002) and semiparametric regression models for longitudinal data (Fan and Li, 2004). The SCAD has been shown to deliver effective performance for variable selection in various models, which motivates us to consider it in this work.

2.2.2 Asymptotic Theory

First we lay out regularity conditions on x_i, t_i and ε which are necessary for the theoretical results. Denote the true coefficients as $\boldsymbol{\beta}_0 = (\beta_{10}, \ldots, \beta_{d0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T$, where $\boldsymbol{\beta}_{20} = \mathbf{0}$ and $\boldsymbol{\beta}_{10}$ consists of all the q nonzero components. Assume ε_i are uncorrelated mean zero random variables with common variance σ^2 and have uniformly bounded absolute third moments. In addition, we assume that the covariate vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are independently and identically distributed with mean zero, $d \times d$ finite positive definite covariance matrix $\mathbf{R}$, and that the components of $\boldsymbol{x_i}$ have finite

fourth moments. For convenience, assume that $t_i's$ are all distinct values in $[0,1]$. As in Heckman (1986), suppose that t_i's satisfy $\int_0^{t_i} u(w)dw = i/n$, where $i = 1,\ldots,n$ and $u(\cdot)$ is a continuous and strictly positive function independent n.

Define $\mathbf{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^T$ and $\mathbf{f} = (f(t_1),\ldots,f(t_n))^T$. The partial linear model (2.1) can then be expressed as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon}$. It can be shown that for given λ_1 and λ_2, minimizing the DPLS (2.2) leads to a smoothing spline estimate for $f(\cdot)$. Hence by theorem (2.1) in Green and Silverman (1994), we can rewrite the DPLS (2.2) as

$$L_{dp}(\boldsymbol{\beta},\mathbf{f};\mathbf{Y}) = \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{f})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{f}) + \frac{n\lambda_1}{2}\mathbf{f}^T\mathbf{K}\mathbf{f} + n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \quad (2.4)$$

where $\mathbf{K}$ is the nonnegative definite smoothing matrix defined by Green and Silverman (1994). Given λ_1,λ_2, σ^2 and $\boldsymbol{\beta}$, the DPLS minimizer of (2.4) is given by

$$\widehat{\mathbf{f}}(\boldsymbol{\beta}) = (\mathbf{I} + n\lambda_1\sigma^2\mathbf{K})^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}). \quad (2.5)$$

Let $\mathbf{A}(\lambda_1) = (\mathbf{I} + n\lambda_1\sigma^2\mathbf{K})^{-1}$, which has an equivalent form as the linear smoother matrix $\mathbf{A}(\lambda)$ in Craven and Wahba (1979) and Heckman (1986). Plugging (2.5) into (2.4), we can derive a profile penalized least squares function only of $\boldsymbol{\beta}$:

$$Q(\boldsymbol{\beta}) = \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\left[\mathbf{I}-\mathbf{A}(\lambda_1)\right](\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) + n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|). \quad (2.6)$$

Denote the quadratic term in (2.6) by $L(\boldsymbol{\beta})$. In the following, we first give two asymptotic results for $L(\boldsymbol{\beta})$ in Lemma 1, which are useful for the proofs of Lemma 2

and Theorems 1 and 2 later in this section. All the proofs are given in the Appendix.

Lemma 1 *Let $L'(\boldsymbol{\beta}_0)$ and $L''(\boldsymbol{\beta}_0)$ be the gradient vector and Hessian matrix of L respectively, evaluated at $\boldsymbol{\beta}_0$. Assume that $\boldsymbol{x}_i$ are independent and identically distributed with finite fourth moments. If $\lambda_{1n} \to 0$ and $n\lambda_{1n}^{1/4} \to \infty$ as $n \to \infty$, then*

(a) $n^{-1/2}L'(\boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \sigma^{-2}\mathbf{R})$,

(b) $n^{-1}L''(\boldsymbol{\beta}_0) \xrightarrow{p} \sigma^{-2}\mathbf{R}$.

Proof of Lemma 1

Differentiating $L(\boldsymbol{\beta})$ and evaluating at $\boldsymbol{\beta}_0$, we get:

$$\begin{aligned} -L'(\boldsymbol{\beta}_0) &= \sigma^{-2}\mathbf{X}^T\left[\mathbf{I} - \mathbf{A}(\lambda_1)\right](\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0), \qquad (2.7) \\ L''(\boldsymbol{\beta}_0) &= \sigma^{-2}\mathbf{X}^T\left[\mathbf{I} - \mathbf{A}(\lambda_1)\right]\mathbf{X}. \qquad (2.8) \end{aligned}$$

For the partial linear model, we have $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{f} + \boldsymbol{\varepsilon}$. Substitution into (2.7) yields

$$\begin{aligned} -n^{-1/2}L'(\boldsymbol{\beta}_0) &= \sigma^{-2}n^{-1/2}\mathbf{X}^T\left[\mathbf{I} - \mathbf{A}(\lambda_1)\right](\mathbf{f} + \boldsymbol{\varepsilon}) \\ &= \sigma^{-2}n^{-1/2}\mathbf{X}^T\left\{\left[\mathbf{I} - \mathbf{A}(\lambda_1)\right]\mathbf{f} + \boldsymbol{\varepsilon}\right\} - \sigma^{-2}n^{-1/2}\mathbf{X}^T\mathbf{A}(\lambda_1)\boldsymbol{\varepsilon}. \qquad (2.9) \end{aligned}$$

Now, the proof of Theorem 1 and its four propositions for $m = 2$ in Heckman (1986) can be used. Under regularity conditions, Heckman showed that, if $\lambda_{1n} \to 0$ and $n\lambda_{1n}^{1/4} \to \infty$, then

$$\begin{aligned} n^{-1/2}\mathbf{X}^T\{(\mathbf{I} - \mathbf{A}(\lambda_1))\mathbf{f} + \boldsymbol{\varepsilon}\} &\xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{R}), \qquad (2.10) \\ n^{-1/2}\mathbf{X}^T\mathbf{A}(\lambda_1)\boldsymbol{\varepsilon} &\xrightarrow{p} \mathbf{0}. \qquad (2.11) \end{aligned}$$

Parts (a) and (b) are obtained by applying Slutsky's theorem to (2.7) and (2.8).

Root-n consistency of the DPLSE $\widehat{\beta}$ is established through the following theorem.

Theorem 1 *As $n \to \infty$, if $\lambda_{1n} \to 0$, $n\lambda_{1n}^{1/4} \to \infty$, and $\max\{|p''_{\lambda_{2n}}(|\beta_{j0}|)| : \beta_{j0} \neq 0\} \to 0$, then there exists a local minimizer $\widehat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2} + b_n)$, where $b_n = \max\{|p'_{\lambda_{2n}}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}$.*

Proof of Theorem 1

We will follow the similar steps in the proofs of Fan and Li (2001). Let $c_n = n^{-1/2} + b_n$. First we show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\left\{\inf_{\|\boldsymbol{r}\|=C} Q(\boldsymbol{\beta}_0 + c_n\boldsymbol{r}) > Q(\boldsymbol{\beta}_0)\right\} \geq 1 - \varepsilon. \tag{2.12}$$

Then with probability at least $1 - \varepsilon$ there exists a local minimum in the ball $\{\boldsymbol{\beta}_0 + c_n\boldsymbol{r} : \|\boldsymbol{r}\| \leq C\}$. This implies that there exists a local minimizer of $Q(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2} + b_n)$.

Let $\Delta_n(\boldsymbol{r}) = Q(\boldsymbol{\beta}_0 + c_n\boldsymbol{r}) - Q(\boldsymbol{\beta}_0)$. Since $p_{\lambda_{2n}}(0) = 0$ and $p_{\lambda_{2n}}(\cdot)$ is nonnegative, we have

$$\Delta_n(\boldsymbol{r}) \geq L(\boldsymbol{\beta}_0 + c_n\boldsymbol{r}) - L(\boldsymbol{\beta}_0) + n\sum_{j=1}^{q}\{p_{\lambda_{2n}}(|\beta_{j0} + c_n r_j|) - p_{\lambda_{2n}}(|\beta_{j0}|)\}, \tag{2.13}$$

where q is the length of $\boldsymbol{\beta}_{10}$. Using Taylor expansion and the quadratic form of $L(\boldsymbol{\beta})$,

we get

$$\Delta_n(\boldsymbol{r}) \geq -c_n \boldsymbol{r}^T\{-L'(\boldsymbol{\beta}_0)\} + \frac{1}{2}\boldsymbol{r}^T\{n^{-1}L''(\boldsymbol{\beta}_0)\}\boldsymbol{r}nc_n^2 + \sum_{j=1}^{q}[nc_n p'_{\lambda_{2n}}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0})r_j + nc_n^2 p''_{\lambda_{2n}}(|\beta_{j0}|)r_j^2\{1+o_p(1)\}]. \quad (2.14)$$

It follows from part (a) of Lemma 1 that $-n^{-1/2}L'(\boldsymbol{\beta}_0) = O_p(1)$. Therefore the first term on the right-hand side of (2.14) has order $O_p(n^{1/2}c_n) = O_p(nc_n^2)$. By part(b) of Lemma 1, $n^{-1}L''(\boldsymbol{\beta}_0) = \sigma^{-2}\mathbf{R} + o_p(1)$. Since $\mathbf{R}$ is a finite positive definite matrix, the second term dominates the first term uniformly in $\|\boldsymbol{r}\| = C$ for sufficiently large C. Using Cauchy-Schwartz inequality, the third term in (2.14) is bounded by

$$q^{1/2}c_n b_n\|\boldsymbol{r}\| + nc_n^2 \max\{|p''_{\lambda_{2n}}(|\beta_{j0}|)| : \beta_{j0} \neq 0\}\|\boldsymbol{r}\|^2. \quad (2.15)$$

By the assumption $\max\{|p''_{\lambda_{2n}}(|\beta_{j0}|)| : \beta_{j0} \neq 0\} \to 0$, (2.15) is also dominated by the second term in (2.14). Therefore by choosing a sufficiently large C, (2.12) holds. This completes the proof of the theorem.

Theorem 1 suggests if we can choose a proper sequence of λ_{2n} such that $b_n \to 0$, as $n \to \infty$, the minimizer of DPLS $\widehat{\boldsymbol{\beta}}$ is root-n consistent. Furthermore, we establish through Lemma 2 and Theorem 2 that $\widehat{\boldsymbol{\beta}}$ is also an oracle estimator, having the sparsity property and the asymptotic normality as well.

Lemma 2 *Suppose that*

$$\liminf_{n\to\infty} \liminf_{\omega\to 0^+} p'_{\lambda_{2n}}(\omega)/\lambda_{2n} > 0. \quad (2.16)$$

As $n \to \infty$, if $\lambda_{1n} \to 0$, $n\lambda_{1n}^{1/4} \to \infty$, $\lambda_{2n} \to 0$, and $n^{1/2}\lambda_{2n} \to \infty$, then with probability tending to 1, for any $\boldsymbol{\beta_1}$ which satisfies $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_p(n^{-1/2})$ and any constant $C > 0$,

$$Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \mathbf{0}\end{pmatrix}\right\} = \min_{\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}} Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}\right\}$$

Proof of Lemma 2

It suffices to show that as $n \to \infty$ with probability tending to 1, for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} = O_p(n^{-1/2})$ and $j = q+1, \ldots, d$,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} < 0 \quad \text{for } \beta_j \in (-Cn^{-1/2},\ 0) \tag{2.17}$$

$$> 0 \quad \text{for } \beta_j \in (0,\ Cn^{-1/2}) \tag{2.18}$$

By Taylor expansion and noting that $L(\boldsymbol{\beta})$ is indeed quadratic in $\boldsymbol{\beta}$, we get

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} + np'_{\lambda_{2n}}(|\beta_j|)\mathrm{sgn}(\beta_j)\\ &= \frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{k=1}^{d}\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_k}(\beta_k - \beta_{k0}) + np'_{\lambda_{2n}}(|\beta_j|)\mathrm{sgn}(\beta_j).\end{aligned}$$

Now by Lemma 1, we have

$$-n^{-1}\frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} = O_p(n^{-1/2}), \quad \frac{1}{n}\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_k} = \sigma^{-2}\mathbf{R}_{jk} + o_p(1).$$

Since the components of $\boldsymbol{\beta}_2$ are confined within $(-Cn^{-1/2}, Cn^{-1/2})$ by construction, it is clear that $\boldsymbol{\beta}_2 - \mathbf{0} = O_p(n^{-1/2})$. Thus jointly with the assumption $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10} =$

$O_p(n^{-1/2})$, we have $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$. Hence it follows that

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = n\lambda_{2n} \left\{ \lambda_{2n}^{-1} p'_{\lambda_{2n}}(|\beta_j|)\mathrm{sgn}(\beta_j) + O_p(n^{-1/2}/\lambda_{2n}) \right\}.$$

By assumptions that $\liminf_{n\to\infty} \liminf_{\omega\to 0^+} p'_{\lambda_{2n}}(\omega)/\lambda_{2n} > 0$ and $n^{-1/2}/\lambda_{2n} \to 0$, the sign of β_j determines that of $\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}$. Therefore, (2.17) and (2.18) hold.

Theorem 2 *Assume that the penalty function $p_{\lambda_{2n}}(|\omega|)$ satisfies* (2.16). *As $n \to \infty$, if $\lambda_{1n} \to 0$, $n\lambda_{1n}^{1/4} \to \infty$, $\lambda_{2n} \to 0$, and $n^{1/2}\lambda_{2n} \to \infty$, then with probability tending to 1, the local minimizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T)^T$ in Theorem 1 must satisfy:*

(a) *Sparsity:* $\widehat{\boldsymbol{\beta}}_2 = 0$.

(b) *Asymptotic normality:*

$$n^{1/2}(\mathbf{R}_{11} + \boldsymbol{\Sigma}_{\lambda_{2n}})\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{R}_{11} + \boldsymbol{\Sigma}_{\lambda_{2n}})^{-1}\mathbf{v}\} \xrightarrow{d} N\{\mathbf{0}, \sigma^2\mathbf{R}_{11}\},$$

where $\mathbf{R}_{11}$ is the $q \times q$ upper-left sub matrix of $\mathbf{R}$,
$\boldsymbol{\Sigma}_{\lambda_{2n}} = \mathrm{diag}\{p''_{\lambda_{2n}}(|\beta_{10}|), \ldots, p''_{\lambda_{2n}}(|\beta_{q0}|)\}$,
and $\mathbf{v} = \left\{p'_{\lambda_{2n}}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}), \ldots, p'_{\lambda_{2n}}(|\beta_{10}|)\mathrm{sgn}(\beta_{q0})\right\}^T$.

Proof of Theorem 2

Part (a) directly follows Lemma 2. Assume $\boldsymbol{\beta}_2 = 0$ is known in advance. Similar as in Theorem 1, it is easy to show that there exists a local minimizer $\widehat{\boldsymbol{\beta}}_1$ of $Q\left\{\begin{pmatrix}\boldsymbol{\beta}_1 \\ 0\end{pmatrix}\right\}$,

which is root-n consistent and also satisfies

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j}\Big|_{\boldsymbol{\beta}=\begin{pmatrix}\widehat{\boldsymbol{\beta}}_1\\ 0\end{pmatrix}} = 0, \quad j = 1, \ldots, q. \tag{2.19}$$

Note $\widehat{\boldsymbol{\beta}}_1$ is a consistent estimator, and by Taylor expansion around $\boldsymbol{\beta}_0$ for (2.19) we have

$$\begin{aligned} 0 = \frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{k=1}^{q} \left\{ \frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_k} + o_p(1) \right\} (\widehat{\beta}_k - \beta_{k0}) \\ + n \left[p'_{\lambda_{2n}}(|\beta_{j0}|)\mathrm{sgn}(\beta_{j0}) + \{p''_{\lambda_{2n}}(|\beta_{j0}|) + o_p(1)\}(\widehat{\beta}_j - \beta_{j0}) \right] \end{aligned} \tag{2.20}$$

Thus by Slutsky's theorem and the central limit theorem, we get the asymptotic normality:

$$n^{1/2}(\mathbf{R}_{11} + \boldsymbol{\Sigma}_{\lambda_{2n}})\{\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{R}_{11} + \boldsymbol{\Sigma}_{\lambda_{2n}})^{-1}\mathbf{v}\} \xrightarrow{d} N\{\mathbf{0}, \sigma^2 \mathbf{R}_{11}\},$$

This concludes the proof for Theorem 2.

2.3 Modified Linear Mixed Model Representation

In this section, we propose a linear mixed model representation for the partial linear model, which allows the smoothing parameter to be estimated as an additional variance component and hence provides a unified estimation and inferential framework. Fan and Li (2001) suggested the local quadratic approximation (LQA)

algorithm to fit the SCAD in linear models. We propose an iterative algorithm which combines the mixed model framework and the LQA to update $(\boldsymbol{\beta}, \mathbf{f})$ successively.

2.3.1 Linear Mixed Model Representation

Consider the general case where there may be ties in observed $\{t_i\}$. Let $\boldsymbol{t^0} = (t_1^0, \ldots, t_r^0)^T$ be an $r \times 1$ vector of ordered distinct values of $\{t_i\}$ $(i = 1, \ldots, n)$, and let $\mathbf{N}$ be the incidence matrix connecting $\{t_i\}$ and $\boldsymbol{t^0}$, such that the (i, j)th element of $\mathbf{N}$ is 1 if $t_i = t_j^0$ and 0 otherwise $(j = 1, \ldots, r)$. The partial linear model (2.1) can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{N}\mathbf{f} + \boldsymbol{\varepsilon}, \tag{2.21}$$

where $\mathbf{f} = \{f(t_1^0), \ldots, f(t_r^0)\}^T$. If ϵ_i's were normally distributed, then (2.4) has an equivalent form to the double-penalized likelihood (DPL)

$$\ell_{dp}(\boldsymbol{\beta}, \mathbf{f}; \mathbf{Y}) = \ell(\boldsymbol{\beta}, \mathbf{f}; \mathbf{Y}) - \frac{n\lambda_1}{2}\mathbf{f}^T\mathbf{K}\mathbf{f} - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \tag{2.22}$$

where $\ell(\boldsymbol{\beta}, \mathbf{f}; \mathbf{Y}) = -(n/2)\log\sigma^2 - (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f})/(2\sigma^2)$. We call the maximizer of (2.22) maximum double-penalized likelihood estimators (MDPLEs). Following Green (1987), write $\mathbf{f}$ via a one-to-one linear transformation as

$$\mathbf{f} = \mathbf{T}\boldsymbol{\delta} + \mathbf{B}\mathbf{a}, \tag{2.23}$$

where $\mathbf{T} = [\mathbf{1}, \boldsymbol{t}^0]$ and $\mathbf{1}$ is an $r \times 1$ vector of 1's, $\boldsymbol{\delta}$ and $\mathbf{a}$ are 2×1 and $(r-2) \times 1$ vectors respectively, and $\mathbf{B} = \mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}$ with $\mathbf{L}$ being an $r \times (r-2)$ full rank matrix

satisfying $\mathbf{K} = \mathbf{L}\mathbf{L}^T$ and $\mathbf{L}^T\mathbf{T} = 0$. It follows that $\mathbf{f}^T\mathbf{K}\mathbf{f} = \mathbf{a}^T\mathbf{a}$ and yields an equivalent double-penalized log-likelihood

$$\begin{aligned}\ell_{dp}(\boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{a}; \mathbf{Y}) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_* - \mathbf{B}_*\mathbf{a})^T(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_* - \mathbf{B}_*\mathbf{a}) \\ &\quad -\frac{n\lambda_1}{2}\mathbf{a}^T\mathbf{a} - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \end{aligned} \tag{2.24}$$

where $\mathbf{X}_* = [\mathbf{N}\mathbf{T}, \mathbf{X}]$, $\mathbf{B}_* = \mathbf{N}\mathbf{B}$, $\boldsymbol{\beta}_* = (\boldsymbol{\delta}^T, \boldsymbol{\beta}^T)^T$.

For fixed $\boldsymbol{\beta}_*$ (and given λ_1,λ_2,σ^2), the DPL (2.24) can be treated as the joint log-likelihood for the following linear mixed model subject to the SCAD penalty for $\boldsymbol{\beta}$

$$\mathbf{Y} = \mathbf{X}_*\boldsymbol{\beta}_* + \mathbf{B}_*\mathbf{a} + \boldsymbol{\varepsilon}, \tag{2.25}$$

where $\boldsymbol{\beta}_* = (\boldsymbol{\delta}^T, \boldsymbol{\beta}^T)^T$ are fixed effects, and $\mathbf{a}$ are treated as random effects distributed as $\mathbf{a} \sim N(0, \tau\boldsymbol{I})$ with $\tau = 1/(n\lambda_1)$, and $\boldsymbol{\theta} = (\tau, \sigma^2)$ are the variance components. In this mixed model framework, we can conduct variable selection for $\boldsymbol{x}$ by maximizing the penalized log-likelihood of $\boldsymbol{\beta}_*$ with the same penalty for $\boldsymbol{\beta}$

$$\ell_{dp}(\boldsymbol{\beta}_*; \mathbf{Y}) = -\frac{1}{2}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*)^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*) - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \tag{2.26}$$

where $\mathbf{V} = \sigma^2\mathbf{I}_n + \tau\mathbf{B}_*\mathbf{B}_*^T$ is the variance of $\mathbf{Y}$ under mixed model representation (2.25). After important variables are selected and $\boldsymbol{\beta}_*$ are estimated, $\widehat{\delta}$ and the best linear unbiased prediction (BLUP) estimate of $\mathbf{a}$ can be used to construct the smoothing spline estimate of $f(t)$.

This mixed model representation also indicates that the inverse of the smoothing

parameter τ can be treated as a variance component and hence can be jointly estimated with σ^2 using maximum likelihood or restricted maximum likelihood (REML) approach during the variable selection process under the working distributional assumption that $\varepsilon_i' s$ were normal errors. However, it should be noted that the above mixed model representation is merely a computational framework. The asymptotic results given in the previous section does not depend on the normal distributional assumption for ϵ_i and our simulation results in Section 2.6 indicate that our overall estimation procedure is robust to the distributional assumption for ϵ_i.

2.3.2 Local Quadratic Approximation (LQA)

The SCAD penalty function defined by (2.3) is not differentiable at the origin, causing difficulty for the maximization of (2.26) using gradient based methods, such as the Newton-Raphson method. Following Fan and Li (2001, 2004), we use a local quadratic approximation (LQA) approach to maximize (2.26). Assuming $\widehat{\boldsymbol{\beta}}_*^0$ is an initial value close to the maximizer of (2.26), we have the following local approximation:

$$\left\{p_{\lambda_2}(|\widehat{\beta}_{*j}|)\right\}' = p'_{\lambda_2}(|\widehat{\beta}_{*j}|)\mathrm{sgn}(\widehat{\beta}_{*j}) \approx \frac{p'_{\lambda_2}(|\widehat{\beta}_{*j}^0|)}{|\widehat{\beta}_{*j}^0|}\widehat{\beta}_{*j}, \quad \text{for } |\widehat{\beta}_{*j}^0| \geq \xi, \quad \text{for } \mathrm{j} \geq 3,$$

where ξ is a pre-specified threshold. Using Taylor expansions, we can locally approximate the DPL (2.26) by

$$\ell_{dp}(\widehat{\boldsymbol{\beta}}_*|\widehat{\boldsymbol{\beta}}_*^0) \approx -\frac{1}{2}(\mathbf{Y}-\mathbf{X}_*\widehat{\boldsymbol{\beta}}_*)^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}_*\widehat{\boldsymbol{\beta}}_*) - \frac{1}{2}n\widehat{\boldsymbol{\beta}}_*^T\boldsymbol{\Sigma}_{\lambda_2}(\widehat{\boldsymbol{\beta}}_*^0)\widehat{\boldsymbol{\beta}}_* - n\sum_{j=3}^{d+2}\left\{p_{\lambda_2}(|\widehat{\beta}_{*j}^0|) - \frac{1}{2}\frac{p'_{\lambda_2}(|\widehat{\beta}_{*j}^0|)}{|\widehat{\beta}_{*j}^0|}(\widehat{\beta}_{*j}^0)^2\right\}, \quad (2.27)$$

where $\boldsymbol{\Sigma}_{\lambda_2}(\boldsymbol{\beta}_*) = \text{diag}\{0, 0, p'_{\lambda_2}(|\beta_1|)/|\beta_1|, \ldots, p'_{\lambda_2}(|\beta_d|)/|\beta_d|\}$. For fixed $\boldsymbol{\theta} = (\tau, \sigma^2)$, we apply the Newton-Raphson method to maximize (2.27) and get the updating formula

$$\widehat{\boldsymbol{\beta}}_* = \left\{\mathbf{X}_*^T\mathbf{W}(\boldsymbol{\theta})\mathbf{X}_* + n\boldsymbol{\Sigma}_{\lambda_2}(\widehat{\boldsymbol{\beta}}_*^0)\right\}^{-1}\mathbf{X}_*^T\mathbf{W}(\boldsymbol{\theta})\mathbf{Y}, \quad (2.28)$$

where $\mathbf{W}(\boldsymbol{\theta}) = \mathbf{V}^{-1} = \sigma^{-2}\{\mathbf{I}_n - \mathbf{B}_*(\lambda_1\sigma^2\mathbf{I}_r + \mathbf{B}_*^T\mathbf{B}_*)^{-1}\mathbf{B}_*^T\}$, which is a computationally more efficient formula when r (number of distinct t_i's) is much smaller than n (sample size). It is easy to recognize from (2.28) that it is in fact an iterative ridge regression algorithm.

2.4 Estimation of Tuning Parameters and Error Variance

The linear mixed effect model framework allows us to treat $\tau = 1/(n\lambda_1)$ as an extra variance component, so that we can estimate it together with the error variance σ^2. We first focus on the estimation of (τ,σ^2), and then discuss how to estimate λ_2.

2.4.1 Estimation of λ_1 and σ^2

The iterative algorithm discussed in Section 2.3 is based on fixed or known smoothing parameter λ_1 (or equivalently τ) and σ^2. However, they are usually unknown and need to be estimated. Using the mixed model representation, one can estimate (τ, σ^2) with the restricted maximum likelihood (REML) estimators (Zhang et al., 1998). We propose to alternately estimate $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{f}})$ and (τ, σ^2) iteratively until the algorithm converges. The initial values for $\boldsymbol{\beta}_*$, τ and σ^2 are obtained by the `MIXED` procedure in SAS to fit the linear mixed model (2.25) with all the covariates.

There is rich literature on the use of REML to estimate the smoothing parameter and variance components (e.g. Wahba, 1985; Kohn et al., 1991; Speed, 1991; Zhang et al., 1998; Lin and Zhang, 1999). In particular, Zhang et al. (1998) estimated the smoothing parameter and variance components simultaneously using REML for longitudinal data with a nonparametric baseline function and complex variance structures. The partial linear model (2.21) has a similar form as (2) of Zhang et al. (1998), with only two variance components (τ, σ^2), and hence the estimation proceeds similarly.

Specifically, denote by $\mathbf{X}_{[s]}$ the subset of important variables selected from $\mathbf{X}$ at the sth iteration. The REML log-likelihood of (τ, σ^2) at this iteration is

$$\ell_R^{[s]}(\tau, \sigma^2; \mathbf{Y}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log\left|\mathbf{X}_{*[s]}^T\mathbf{V}^{-1}\mathbf{X}_{*[s]}\right| - \frac{1}{2}(\mathbf{Y}-\mathbf{X}_{*[s]}\widehat{\boldsymbol{\beta}}_{*[s]})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}_{*[s]}\widehat{\boldsymbol{\beta}}_{*[s]}),$$

where $\mathbf{X}_{*[s]} = \left[\mathbf{NT}, \mathbf{X}_{[s]}\right]$, $\widehat{\boldsymbol{\beta}}_{*[s]} = (\mathbf{X}_{*[s]}^T\mathbf{V}^{-1}\mathbf{X}_{*[s]})^{-1}\mathbf{X}_{*[s]}^T\mathbf{V}^{-1}\mathbf{Y}$ is the MLE of $\boldsymbol{\beta}_*$ based on the selected important variables $\mathbf{X}_{[s]}$ and $\mathbf{V}$ is defined in Section 2.3.1. Differentiating $\ell_R^{[s]}(\tau, \sigma^2; \mathbf{Y})$ with respect to τ and σ^2, we obtain the REML estimating

equations for τ and σ^2 (Harville, 1977):

$$-\frac{1}{2}\text{tr}(\mathbf{P}\mathbf{B}_*\mathbf{B}_*^T) + \frac{1}{2}(\mathbf{Y} - \mathbf{X}_{*[s]}\widehat{\boldsymbol{\beta}}_{*[s]})^T\mathbf{V}^{-1}\mathbf{B}_*\mathbf{B}_*^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}_{*[s]}\widehat{\boldsymbol{\beta}}_{*[s]}) = 0 \quad (2.29)$$

$$-\frac{1}{2}\text{tr}(\mathbf{P}) + \frac{1}{2}(\mathbf{Y} - \mathbf{X}_{*[s]}\widehat{\boldsymbol{\beta}}_{*[s]})^T\mathbf{V}^{-2}(\mathbf{Y} - \mathbf{X}_{*[s]}\widehat{\boldsymbol{\beta}}_{*[s]}) = 0, \quad (2.30)$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}_{*[s]}(\mathbf{X}_{*[s]}^T\mathbf{V}^{-1}\mathbf{X}_*^T)^{-1}\mathbf{X}_{*[s]}^T\mathbf{V}^{-1}$. The Fisher scoring algorithm can then be used to solve (2.29) and (2.30) for τ and σ^2. We alternate the estimation of $\boldsymbol{\beta}_*$ and the estimation of (τ, σ^2) until the subset $\mathbf{X}_{[s]}$ converges to a stable set. Note that if $\mathbf{X}_{[s]}$ is the correct subset, equations (2.29) and (2.30) are unbiased estimating equations for τ and σ^2 even without normality assumption for $\mathbf{a}$ and ϵ_i. This is the reason that above algorithm is robust to the distributional assumption of ϵ_i.

2.4.2 Choice of the SCAD Tuning Parameters

The aforementioned methods are for fixed values of the SCAD tuning parameters (λ_2, a). To find their optimal values, one common approach could be a two-dimensional grid search using some data-driven criteria, such as CV and GCV (Craven and Wahba, 1979), which can be rather computationally prohibitive. Fan and Li (2001) showed numerically that $a = 3.7$ minimizes the Bayesian risk and recommended its use in practice. Thus we set $a = 3.7$ and only tune λ_2 in our implementation.

We propose using the Bayesian Information Criterion (BIC) (Schwarz, 1978) to select the regularization parameter λ_2 from a gridded range under working normal distributional assumption for ε_i. Given λ_2, suppose q variables are selected by the

algorithm in Section 2.3. Let $\mathbf{X}_1$ be the sub matrix of $\mathbf{X}$ for the q important variables and $\boldsymbol{\beta}_1$ the $q\times 1$ corresponding regression coefficient vector. Then we may use methods of Zhang et al. (1998) to solve the partial linear model (2.1). Consequently $\widehat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$, where $\mathbf{S}$ is a smoother matrix with $q_1 = trace(\mathbf{S})$. The BIC criterion is computed as

$$\mathrm{BIC}(\lambda_2) = -2\ell + q_1 \log n, \tag{2.31}$$

where $\ell = -(n/2)\log(2\pi\widehat{\sigma}^2) - (\mathbf{Y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{N}\widehat{\mathbf{f}})^T(\mathbf{Y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{N}\widehat{\mathbf{f}})/(2\widehat{\sigma}^2)$.

Notice that (2.31) is an implicit function of λ_2, which plays a similar role as CV or GCV for selecting tuning parameters. For each grid point of λ_2, the iterative ridge regression results in a model with q important covariates, and we compute the BIC for this selected model. Based on our empirical evidence and the fact that BIC is consistent in selecting correct models under certain conditions (Schwarz, 1978), we chose BIC over GCV for tuning λ_2 in our numerical analysis.

2.5 Frequentist and Bayesian Covariance Estimates

In this section, we derive the frequentist and Bayesian covariance formulas for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{f}}$ parallel to Sections 3.4 and 3.5 in Zhang et al. (1998), except that we also take into account the extra variability introduced by the variable selection process. Using these covariance estimates, we are able to calculate standard errors and construct confidence intervals for the regression coefficients and the nonparametric function. The proposed covariance estimates are evaluated via simulations.

2.5.1 Frequentist Covariance Estimates

From frequentists' point of view, $\text{cov}(\mathbf{Y}|t,\mathbf{x}) = \sigma^2\mathbf{I}$, and we can write $\widehat{\boldsymbol{\beta}}_* = (\widehat{\boldsymbol{\delta}}^T, \widehat{\boldsymbol{\beta}}^T)^T$ as an approximately linear function of $\mathbf{Y}$:

$$\widehat{\boldsymbol{\beta}}_* = \left\{\mathbf{X}_*^T\mathbf{W}\mathbf{X}_* + n\boldsymbol{\Sigma}_{\lambda_2}(\widehat{\boldsymbol{\beta}})\right\}^{-1}\mathbf{X}_*^T\mathbf{W}\mathbf{Y} \equiv \mathbf{Q}\mathbf{Y}. \tag{2.32}$$

Let $\mathbf{Q} = (\mathbf{Q}_1^T, \mathbf{Q}_2^T)^T$, where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are partitions of $\mathbf{Q}$ with dimensions corresponding to $(\boldsymbol{\delta}^T, \boldsymbol{\beta}^T)^T$, so that $\widehat{\boldsymbol{\delta}} = \mathbf{Q}_1\mathbf{Y}$, and $\widehat{\boldsymbol{\beta}} = \mathbf{Q}_2\mathbf{Y}$. The estimated covariance matrix for $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{\text{cov}}_F(\widehat{\boldsymbol{\beta}}|t,\mathbf{x}) = \mathbf{Q}_2\text{cov}(\mathbf{Y})\mathbf{Q}_2^T = \widehat{\sigma}^2\mathbf{Q}_2\mathbf{Q}_2^T, \tag{2.33}$$

where $\widehat{\sigma}^2$ is the estimated error variance. It is easy to show that the empirical BLUP estimate of $\mathbf{a}$ is $\widehat{\mathbf{a}}(\boldsymbol{\beta}_*) = \tilde{\mathbf{A}}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*) = \mathbf{S}_a\mathbf{Y}$, where $\mathbf{S}_a = \tilde{\mathbf{A}}(\mathbf{I} - \mathbf{X}_*\mathbf{Q})$ and $\tilde{\mathbf{A}} = (n\lambda_1\sigma^2\mathbf{I} + \mathbf{B}_*^T\mathbf{B}_*)^{-1}\mathbf{B}_*^T$. Therefore $\widehat{\mathbf{f}} = \mathbf{T}\widehat{\boldsymbol{\delta}} + \mathbf{B}\widehat{\mathbf{a}} = (\mathbf{T}\mathbf{Q}_1 + \mathbf{B}\mathbf{S}_a)\mathbf{Y}$ and its covariance can be computed as

$$\widehat{\text{cov}}_F(\widehat{\mathbf{f}}|t,\mathbf{x}) = \widehat{\sigma}^2(\mathbf{T}\mathbf{Q}_1 + \mathbf{B}\mathbf{S}_a)(\mathbf{T}\mathbf{Q}_1 + \mathbf{B}\mathbf{S}_a)^T. \tag{2.34}$$

2.5.2 Bayesian Covariance Estimates

The mixed model representation in Section 3.1 and the DPL (2.24) indicate that $f(t)$ has a prior in the form of $\mathbf{f} = \mathbf{T}\boldsymbol{\delta} + \mathbf{B}\mathbf{a}$, with $\mathbf{a} \sim N(0, \tau\mathbf{I})$ and a flat prior for $\boldsymbol{\delta}$, and the selected important coefficient $\boldsymbol{\beta}$ has a prior whose log-density is approximately equal to $-\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{\lambda_2}\boldsymbol{\beta}/2$ up to a constant, where $\boldsymbol{\Sigma}_{\lambda_2}$ is a diagonal matrix

defined in section 2.3.2. The definition of the SCAD penalty function (2.3) implies that some diagonal elements of the matrix $\mathbf{\Sigma}_{\lambda_2}$ can be zero, corresponding to those coefficients with $|\beta_j| > a\lambda_2$. Assume after reordering, $\mathbf{\Sigma}_{\lambda_2} = \text{diag}(\mathbf{0}, \mathbf{\Sigma}_{22})$, where $\mathbf{\Sigma}_{22}$ is nonzero. It follows that $\boldsymbol{\beta}$ can be partitioned into $(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1$ can be regarded as"fixed" effects and $\boldsymbol{\beta}_2$ as "random" effects with $\boldsymbol{\beta}_2 \sim N(\mathbf{0}, \mathbf{\Sigma}_{22}^{-1})$. The matrix $\mathbf{X}$ is partitioned into $[\mathbf{X}_1, \mathbf{X}_2]$ accordingly. Now we reformulate the mixed model (2.25) as: $\mathbf{Y} = \mathbf{NT}\boldsymbol{\delta} + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{B}_*\mathbf{a} + \boldsymbol{\varepsilon}$, or in the form of a linear mixed model

$$\mathbf{Y} = \mathcal{X}\boldsymbol{\gamma} + \mathbf{Zb} + \boldsymbol{\varepsilon}, \tag{2.35}$$

where $\mathcal{X} = [\mathbf{NT}, \mathbf{X}_1]$, $\boldsymbol{\gamma} = (\boldsymbol{\delta}^T, \boldsymbol{\beta}_1^T)^T$, $\mathbf{Z} = [\mathbf{X}_2, \mathbf{B}_*]$ and $\mathbf{b} = (\boldsymbol{\beta}_2^T, \mathbf{a}^T)^T$ is the new random effect distributed as $\mathbf{b} \sim N(\mathbf{0}, \mathbf{\Sigma}_b)$ with a block diagonal covariance matrix $\mathbf{\Sigma}_b = \text{diag}(\mathbf{\Sigma}_{22}^{-1}, \tau\mathbf{I})$. Under the linear mixed model (2.35), $\boldsymbol{\beta}$ consists of both fixed and random effects. Therefore the Bayesian covariances for $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{f}})$ are

$$\begin{aligned} \text{cov}_B(\widehat{\boldsymbol{\beta}}) &= \text{cov}\{\widehat{\boldsymbol{\beta}}_1^T, (\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)^T\}^T, & (2.36) \\ \text{cov}_B(\widehat{\mathbf{f}}) &= [\mathbf{T}, \mathbf{B}]\, \text{cov}\{\widehat{\boldsymbol{\delta}}^T, (\widehat{\mathbf{a}} - \mathbf{a})^T\}^T [\mathbf{T}, \mathbf{B}]^T. & (2.37) \end{aligned}$$

These Bayesian variance estimates can be viewed to account for the bias in $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{f}}$ due to imposed penalties (Wahba, 1983).

2.6 Simulation Studies

We conduct Monte Carlo simulations to evaluate the finite-sample performance of our proposed methods, examining the estimation of both the regression coefficients

and the nonparametric function, as well as the variable selection results. Simulations are implemented using SAS. We compare our estimators with LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2004) and demonstrate that our method outperforms the other two methods in various situations. For the kernel-based estimators with L_1 and SCAD penalties (Fan and Li, 2004), we adopt Fan and Li's approach to choose the bandwidth parameter: first compute the difference based estimator (DBE) for $\boldsymbol{\beta}$ and then select the bandwidth using the plug-in method of Ruppert et al. (1995). The SCAD and LASSO tuning parameters are selected by BIC instead of GCV to achieve a better performance. Two types of non-normal errors are used to demonstrate that the proposed normal likelihood based REML estimation procedure is robust to the distributional assumption of errors.

The mean squares error (MSE) for both fitted $\widehat{\boldsymbol{\beta}}$ and $\widehat{f}$ is computed as an indicator of goodness-of-fit. As in Fan and Li (2004), the MSE for $\widehat{\boldsymbol{\beta}}$ is defined by $MSE_\beta = E(\|\widehat{\boldsymbol{\beta}}_l - \boldsymbol{\beta}\|^2)$. For $\widehat{f}(t)$, we have $MSE_f = E\left[\int_{T_1}^{T_2}\{\widehat{f}(t) - f(t)\}^2 dt\right]$, and we estimate it by averaging over 50 grid points (knots) $\frac{1}{50}\sum_{l=1}^{50}\{\widehat{f}(t_l) - f(t_l)\}^2$. We report the MC sample mean and standard deviation for the MSEs. In our simulations, the bandwidth selected by the plug-in method sometimes caused numerical problem and thus the results of SCAD and LASSO are based on M MC samples ($M \geq 90$, except $M = 72$ for one scenario). Our method always converges and hence $M = 100$. "Corr." gives the average number of zero coefficients, restricted only to the true zero coefficients (5 in our design), whereas the average number of coefficients incorrectly set to 0 (ideally should be 0) is referred to as "Inc.". Model size is the MC average for the total number of nonzero coefficients in the fitted model (true size is 3 by design).

2.6.1 Simulation Design

We simulate the data from a partial linear model $y = \mathbf{x}^T\boldsymbol{\beta} + f(t) + \varepsilon$. We adopt the configuration in Tibshirani (1996) and Fan and Li (2001) and generate correlated covariates $\mathbf{x} = (x_1, \ldots, x_8)^T$ from a standard normal distribution but with AR(1) correlation $\text{corr}(x_i, x_j) = 0.5^{|i-j|}$. Following Fan and Li (2004), we set the true coefficients $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. To assess the performance of our methods in various situations, we consider a factorial experiment with total 8 scenarios. Each experiment is repeated using 100 Monte Carlo (MC) samples:

- Two combinations of f and random errors. We choose $f_1(t) = 4\sin(2\pi t/T)$ with ε generated from t_6. The second form is $f_2(t) = 5\beta(t/T, 11, 5) + 4\beta(t/T, 5, 11)$ where $\beta(t, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1}$, with mixture normal random error $\varepsilon \sim 0.5N(1, 1) + 0.5N(-1, 3)$. The periods T are 4 and 20 for f_1 and f_2 respectively.
- Two sample sizes $n = 100$ and $n = 200$ with 50 equally spaced knots in $[0, T]$.
- Two noise levels (error variances): $\sigma^2 = 1$ and $\sigma^2 = 9$.

2.6.2 Model Fitting and Selection Performance

Table 2.1 presents the model selection and fitting results of three methods where $\sigma = 1$. Our method is labeled as "DPLSE"; "SCAD" and "LASSO" correspond to the SCAD and L_1 penalized least squares estimators of Fan and Li (2004). From Table 2.1, it can be seen that the DPLSE outperforms the other two methods in terms of model estimation and selection in all situations, and SCAD yields better results than LASSO. Our selected models are not only sparser (with correct number

Table 2.1: Model fitting and selection result comparison

(n, σ^2, f)	$Method$	MSE_β $Mean(SD)$	MSE_f $Mean(SD)$	$Model$ $Size$	*Zero coef.*	
					$Corr.$	$Inc.$
(100,1, f_1)	DPLSE	0.05 (0.06)	0.07 (0.04)	3.22	4.78	0
	SCAD	0.09 (0.09)	0.17 (0.07)	3.39	4.61	0
	LASSO	0.10 (0.09)	0.17 (0.07)	3.82	4.18	0
(100,1, f_2)	DPLSE	0.06 (0.06)	0.14 (0.05)	3.21	4.79	0
	SCAD	0.08 (0.08)	0.28 (0.10)	3.31	4.69	0
	LASSO	0.13 (0.10)	0.29 (0.11)	3.69	4.31	0
(200,1, f_1)	DPLSE	0.02 (0.02)	0.04 (0.02)	3.08	4.92	0
	SCAD	0.02 (0.02)	0.09 (0.03)	3.26	4.74	0
	LASSO	0.03 (0.02)	0.09 (0.03)	3.45	4.55	0
(200,1, f_2)	DPLSE	0.02 (0.02)	0.08 (0.03)	3.07	4.93	0
	SCAD	0.03 (0.03)	0.19 (0.05)	3.24	4.76	0
	LASSO	0.04 (0.03)	0.19 (0.05)	3.53	4.47	0

of zero coefficients closer to 5), but also closer to the true model (closer to true model size 3). The "Inc" column are all zeros, which means that the important variable are always selected. The MSEs produced by DPLSE are always smaller than those by SCAD and LASSO. Although Table 2.2 reports only the DPLSE results for higher noise cases $\sigma^2 = 9$ to save space, DPLSE still achieves better performance than the other two methods. In Table 2.2, incorrect zero coefficients occur only once when $n = 100$ and they reduce to 0 when n doubles. Comparing Table 2.1 and Table 2.2, we see that as σ^2 increases from 1 to 9, although there is a substantial amount of increase in the MSEs, our method still maintains very good performance in terms of model size and number of correct zero coefficients.

Table 2.2: Model selection and estimation summary where $\sigma^2 = 9$

	MSE_β	MSE_f	*Model*	*Zero coef.*	
$(n,\ f)$	$Mean(SD)$	$Mean(SD)$	*Size*	*Corr.*	*Inc.*
$(100,\ f_1)$	0.58 (0.67)	0.55 (0.39)	3.23	4.75	0.02
$(200,\ f_1)$	0.22 (0.24)	0.27 (0.15)	3.12	4.88	0
$(100,\ f_2)$	0.71 (0.75)	0.92 (0.49)	3.21	4.77	0.02
$(200,\ f_2)$	0.22 (0.22)	0.48 (0.19)	3.97	4.93	0

Table 2.3: Variable selection results on estimates of model parameters based on 100 replications for four scenarios, where $\varepsilon \sim t_6$ for $f_1(t)$ and mixture normal for f_2

Scenario	*Model*	*Point*	*Relative*	*Empirical*	*Model-based SE*		95% *CP*	
(n, σ^2, f)	*parameter*	*estimate*	*bias*	*SE*	*Freq.*	*Bayesian*	*Freq.*	*Bayesian*
$(100, 1, f_1)$	β_1	3.011	0.004	0.129	0.128	0.129	0.95	0.95
	β_2	1.500	0.000	0.113	0.106	0.107	0.94	0.94
	β_5	2.024	0.012	0.134	0.105	0.107	0.89	0.90
$(200, 1, f_1)$	β_1	3.006	0.002	0.086	0.087	0.087	0.94	0.95
	β_2	1.502	0.002	0.086	0.087	0.088	0.95	0.96
	β_5	1.994	−0.002	0.075	0.076	0.077	0.96	0.96
$(200, 1, f_2)$	β_1	3.009	0.003	0.088	0.087	0.088	0.94	0.94
	β_2	1.502	0.001	0.088	0.088	0.088	0.96	0.97
	β_5	2.018	0.009	0.074	0.077	0.078	0.98	0.99
$(200, 9, f_2)$	β_1	3.037	0.012	0.242	0.261	0.263	0.94	0.94
	β_2	1.487	−0.009	0.302	0.264	0.265	0.96	0.96
	β_5	1.983	−0.012	0.246	0.230	0.232	0.96	0.96

2.6.3 Performance of Estimators for Parametric Model Parameters

We now evaluate our DPLSEs for model parameters $\boldsymbol{\beta}$. We check the accuracy of the frequentist and Bayesian standard error (SE) formulas for $\widehat{\boldsymbol{\beta}}$ by comparing them to the empirical SE computed from the 100 estimated coefficients. We calculate coverage probability rates for the 95% confidence intervals constructed using the covariance formulas in Section 2.5. We report the results for four representative scenarios by

varying the sample size n, noise level σ^2 and the form of $f(t)$. The results for other cases are similar and are omitted.

Table 2.3 presents the point estimates, relative biases, empirical standard errors, model-based frequentist and Bayesian standard errors of $\widehat{\beta}_j$ in four selected scenarios. The relative bias is defined as the bias in the parameter estimate divided by the true value. The point estimates are the Monte Carlo sample averages and the empirical standard errors are computed by the MC standard deviations of each $\widehat{\beta}_j$. The average coverage probabilities (CP) for both Bayesian and frequentist 95% confidence intervals of nonzero components of $\boldsymbol{\beta}$ are presented as well. From Table 2.3, it can be seen that the MDPLEs $\widehat{\beta}_j's$ and the estimated error variance $\widehat{\sigma}^2$ are almost unbiased in all scenarios. Both Bayesian and frequentist SEs for $\widehat{\boldsymbol{\beta}}$ computed respectively from (2.33) and (2.36) agree well with the empirical SE; all SEs decrease as the sample size n increases or as the noise level σ^2 decreases. The model-based Bayesian SEs are slightly larger than the frequentist counterparts, since the Bayesian SEs account for the biases in $\widehat{\beta}_j$. The confidence intervals for β_j's constructed with both Bayesian and frequentist SEs achieve or approach the nominal level, indicating the accuracy of the SE formulas for $\widehat{\beta}_j$'s. Overall, our method works very well for estimating model parameters.

2.6.4 Performance of $\widehat{f}(t)$ and its Point-wise Standard Errors

In this section we assess the performance of the DPL estimate of $\widehat{f}(t)$ and its point-wise standard errors. We visualize and compare results using MC samples in which SCAD and LASSO did not have numerical problems (our method always converges).

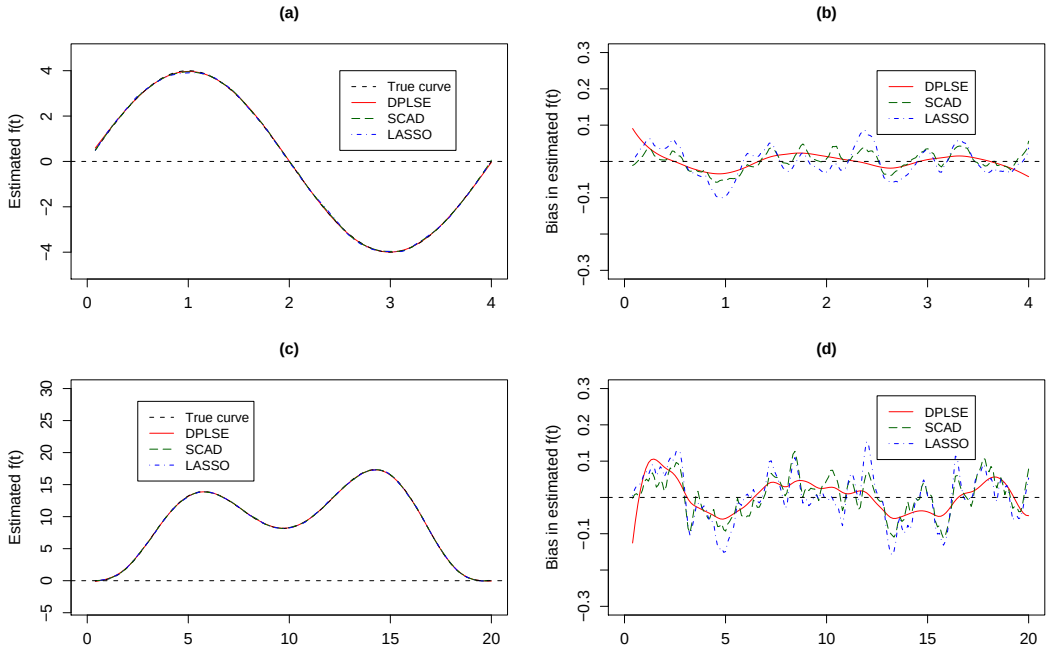

Figure 2.1: Plots of $\widehat{f}(t)$ and point-wise biases based on valid MC samples. The top two plots (a) and (b) are $\widehat{f}$ and point-wise bias for $\widehat{f}_1(t)$; (c) and (d) are for $\widehat{f}_2(t)$. Here $f_1(t) = 4\sin(2\pi/4)$, $n = 200$, and $\varepsilon_i \sim t_6$ with $\sigma^2 = 1$; $f_2(t) = 5\beta(t/20, 5, 11) + 4\beta(t/20, 11, 5)$, $n = 200$, and ε_i's are mixture normal random errors with variance $\sigma^2 = 1$. The horizontal axis is t in all four plots.

Figure 2.1 shows the point-wise estimates and biases for both $f_1(t)$ and $f_2(t)$ when $n = 200$ and $\sigma^2 = 1$ for all three methods (DPLSE, SCAD and LASSO). In plot (a) and (c), the averaged fitted curves are almost indistinguishable from the true nonparametric function, indicating small biases in $\widehat{f}(t)$ for all three methods. Point-wise biases are magnified in plot (b) and (d). It can be seen that the DPLSE fit has overall ssmaller bias than the other two methods of Fan and Li (2004). The SCAD and LASSO fits produce not only larger but also rougher point-wise biases, which indicates under-smoothing due to small bandwidth selected by the plug-in method. Our method is more advantageous in that it automatically estimates the smoothing

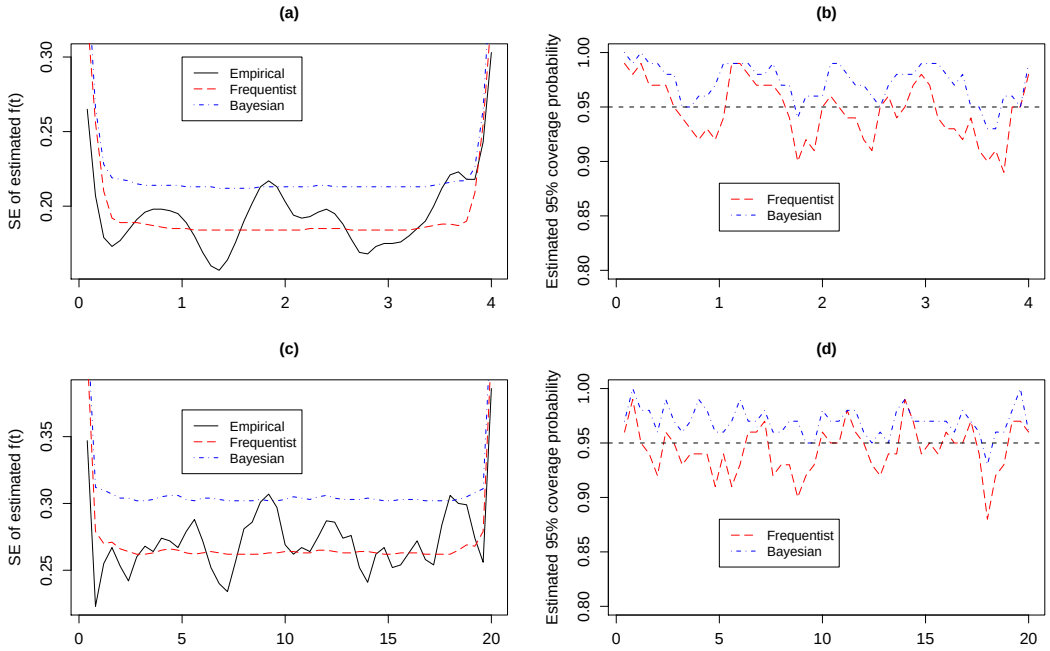

Figure 2.2: Plots of point-wise standard errors and coverage probability rates based on valid replications. The top two plots (a) and (b) are $\widehat{f}(t)$ and point-wise bias for $\widehat{f}_1(t)$; (c) and (d) are for $f_2(t)$, where $f_1(t)$ and $f_2(t)$ are described as before. The horizontal axis is t in all four plots.

parameter and control the amount the smoothing more appropriately by treating $\tau = 1/\lambda_1$ as a variance component. Plots (c) and (d) show similar patterns for the bimodal f_2 with mixture normal errors. Based on comparisons shown in Table 2.1 and Figure 2.1, we conclude that our method has better performance than those in Fan and Li (2004).

Figure 2.2 depicts the DPLSE point-wise standard errors and point-wise coverage probabilities for confidence intervals constructed by using the covariance formulas (2.34) and (2.37). Same as in Figure 2.1, here $n = 200$ and $\sigma^2 = 1$; (a) and (b) are for f_1 and t_6 errors, and (c) and (d) are for f_2 and mixture normal errors. Plots (a) and (c) shows three point-wise SEs: the empirical SE, the model-based frequentist and

Bayesian SEs average over the 100 samples. It can be seen that the frequentist point-wise SEs interlace with the empirical SEs, whereas the Bayesian point-wise SEs are a little larger than the frequentist counterparts. Accordingly, plot (b) and (d) shows that the point-wise coverage probability rates for the frequentist confidence intervals are around the nominal level, whereas the Bayesian ones are uniformly higher than 95%. Contrasting plots (a) and (b) reveals that larger standard errors are associated with poorer coverage probabilities.

2.7 Application of the Proposed Method to a Real Data Set

We apply the proposed method to the *Ragweed Pollen Level* data, which was analyzed in Ruppert et al. (2003). Collected in Kalamazoo, Michigan during the 1993 ragweed season, the data consist of 87 daily observations of ragweed pollen level and relevant information. The main interest is to develop accurate models to forecast daily ragweed pollen level. The raw response *ragweed* is the daily ragweed pollen level (grains/m^3). Among the explanatory variables, x_1 is an indicator of significant rain, where $x_1 = 1$ if there is at least 3 hours of steady or brief but intense rain and $x_1 = 0$ otherwise; x_2 is temperature (^oF); x_3 is wind speed (knots). Since the raw response is rather skewed, Ruppert et al. (2003) suggested a square root transformed response $y = \sqrt{ragweed}$. Marginal plots indicate a strong nonlinear relationship between y and the day number in the current ragweed pollen season. Consequently, a semiparametric regression model with a nonparametric baseline $f(day)$ is reasonable.

Table 2.4: Estimated coefficients and frequentist and bayesian SEs for ragweed pollen level data

Variable	Full model			Selected model		
	Parameter estimate	Frequentist SE	Bayesian SE	Parameter estimate	Frequentist SE	Bayesian SE
x_1	0.64	0.22	0.23	0.70	0.18	0.18
x_2	1.31	0.37	0.39	1.16	0.36	0.37
x_3	0.87	0.19	0.20	0.76	0.19	0.20
x_2^2	0.53	0.23	0.24	0	-	-
x_3^2	0.04	0.19	0.19	0	-	-
x_1x_2	0.26	0.19	0.19	0	-	-
x_1x_3	0.02	0.22	0.23	0	-	-
x_2x_3	0.34	0.20	0.20	0	-	-

Ruppert et al. (2003) fitted a semiparametric model with x_1, x_2 and x_3, whereas here we add quadratic and interaction terms and consider a more complex model:

$$y = f(day) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon.$$

We fit the proposed semiparametric model to select important covariates, as well as estimate both the parametric and nonparametric components. The x-covariates are standardized beforehand. The SCAD tuning parameter selected by BIC is $\lambda_2 = 0.177$. Table 2.4 gives the MDPLEs of the regression coefficients and their corresponding frequentist and Bayesian standard errors. For comparison, we also include in Table 2.4 the estimates via the penalized likelihood with only roughness penalty, namely, the traditional partial spline estimates. The model selected by our procedure is $\widehat{y} = \widehat{f}(day) + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3$, indicating that the linear main effect model suffices. All the estimated coefficients are positive, suggesting that the ragweed pollen level

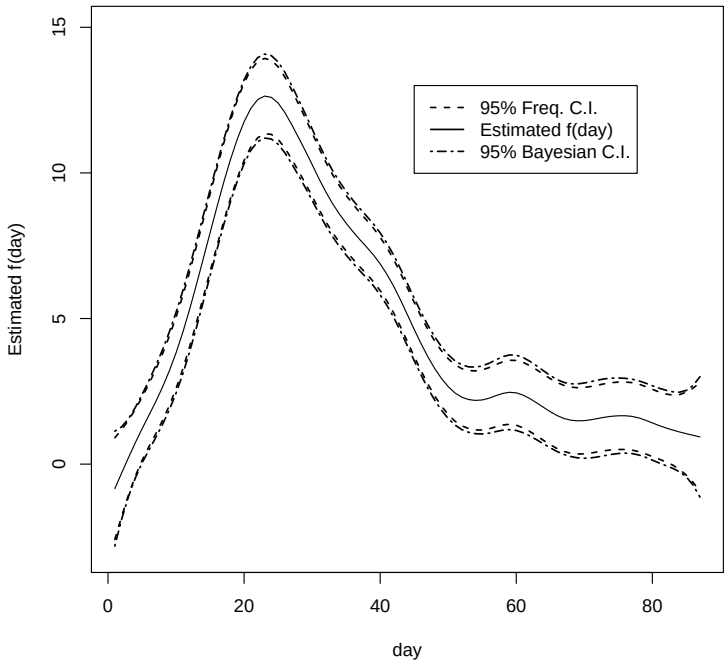

Figure 2.3: Plot of Estimated $f(day)$ and Its Frequentist and Bayesian 95% Point-wise Confidence Intervals for the Ragweed Pollen Level Data.

increases as either of the covariates increases. It can be seen that the shrinkage estimates have relatively smaller standard errors than those under the full model. Figure 2.3 shows the estimated nonparametric function $\widehat{f}(day)$ and its frequentist and Bayesian 95% point-wise confidence intervals. The plot indicates that the baseline $f(day)$ climbs rapidly to the peak on around day 25 and plunges until day 60, and decreases steadily thereafter.

2.8 Discussion

In this article we have proposed a new approach for simultaneous model selection and estimation via double-penalized least squares for partial linear models. The DPLS is equipped with the roughness penalty for estimating the nonparametric function and the SCAD penalty for selecting important covariates. By minimizing the DPLS, we obtain a natural smoothing spline estimate for $f(t)$ and shrinkage estimates for the parametric coefficients β. We showed that under regularity conditions, the resulting DPLSE $\widehat{\beta}$ is root-n consistent and has oracle property and asymptotic normality.

To facilitate computation, we cast the partial linear model into a linear mixed model framework and estimated the smoothing parameter as an additional variance component using REML. This is a very attractive feature since the estimation of the smoothing parameter λ_1 is separated from the estimation of the other tuning parameter λ_2, which is usually done through a grid search. Another advantage of using mixed model representation is that existing software for mixed models can be readily used to implement the overall variable selection and estimation procedure. We proposed a successive local quadratic approximation algorithm to iteratively compute the MDPLEs. The BIC is used to select the SCAD tuning parameters. Simulation studies indicate that the proposed method performed very well and outperformed existing methods in terms of model selection and estimation of the regression coefficients and the nonparametric function. The Bayesian and frequentist SEs for $\widehat{\beta}$ differ little and both are close to the empirical estimates. However, the Bayesian point-wise SEs for $\widehat{f}(t)$ are visually larger than the frequentist SEs, resulting in wider confidence intervals and inflated coverage probability rates. Therefore we recommend the frequentist

SE formulas for $f(t)$. Results show that the DPL approach is robust to the distributional assumption of errors, giving empirical support for using our approach in the general DPLS problems.

The proposed DPLS method is for partial linear models, where the errors are assumed to be uncorrelated. The next step is to generalize our method for correlated data. In particular, we plan to extend our approach into the longitudinal data setting and accommodate complex variance structures and random stochastic process. Another extension is model selection and estimation in semiparametric generalized linear models, e.g. $E(Y) = g\{X\beta + f(t)\}$, where g is a smooth link function. In that case we directly work with the double-penalized likelihood, and the asymptotic properties for the resulting estimators need to be investigated.

Chapter 3

Variable Selection in Semiparametric Mixed Models

3.1 Introduction

Longitudinal data are often encountered in many biomedical and clinical studies where subjects are repeatedly measured over time. From a statistical point of view, the data consist of individual clusters and are correlated. For longitudinal data, traditional multivariate regression methods are hardly useful because of the unbalanced data and their lack of flexibility for modeling covariance structures. In contrast, linear mixed models provide a flexible parametric framework for longitudinal data analysis (Lard and Ware, 1982; Verbeke and Molenberghs, 2000; Diggle et al., 2002). Parametric linear mixed models are sometimes inappropriate and likely to introduce modeling biases when underlying models are complicated. To relax the parametric assumptions, various semiparametric models have been developed for longitudinal data. Different smoothing techniques have been used in these models, including kernel smoothing (Diggle et al., 2002; Chen and Jin, 2006), smoothing splines (Zhang

et al., 1998; Wang, 1998), B-splines (He et al., 2002), penalized splines (Ruppert et al., 2003) and local polynomial regressions (Fan and Li, 2004). Semiparametric mixed models (SPMM; Diggle et al., 2002; Zhang et al., 1998) are useful extensions to linear mixed models, which use parametric fixed effects to represent the covariate effects and an arbitrary smooth function to model the time effect, and accounts for the within-subject correlation using random effects and a stationary or nonstationary stochastic process. Unlike the kernel based estimation in Diggle et al. (2002), Zhang et al. (1998) estimated the nonparametric baseline function using smoothing splines by maximizing the penalized likelihood. The most attractive feature of this approach is that the smoothing parameter can be treated as an extra variance component in a modified linear mixed model framework and can be jointly estimated by restricted maximum likelihood (REML).

Variable selection is an important issue in regression analysis. In longitudinal studies, there can be a large number of covariates for each subject, whereas only a part of them are predictive to the response. Inclusion of redundant or unimportant predictors causes inflated variances and lack of stability. Therefore variable selection is necessary for improving prediction accuracy and model interpretability of fitted models. For linear models, traditional approaches for variable selection include stepwise selection and best subset selection based on information criteria such as Mallows' C_p (Mallows, 1973, 1975), Akaike's Information Criteria (Akaike, 1973, 1977) and Bayesian Information Criteria (Schwarz, 1978). As pointed out by Breiman (1995), these methods suffer from instability and relative lack of accuracy. To achieve better prediction and reduce variances of estimators, many shrinkage estimation approaches

have been proposed such as bridge regression (Frank and Friedman, 1993), nonnegative garrote (Breiman, 1995), and least absolute selection and shrinkage operator (LASSO; Tibshirani, 1996). They have gained a lot of popularity due to their capability of simultaneously selecting variables and estimating coefficients in a continuous fashion. More recently, Fan and Li (2001, 2004) proposed a nonconcave penalized likelihood approach with the smoothly clipped absolute deviation (SCAD) penalty function, which led to a consistent, sparse and continuous solution having oracle properties provided suitable choices of regularization parameters.

Although there is rich literature on variable selection in linear regression models for independent data, little work has been done in semiparametric mixed models for correlated data, especially for longitudinal data. In this chapter we propose a double penalized likelihood approach to simultaneously perform model selection and estimation in the semiparametric mixed model for longitudinal data. We incorporate the SCAD penalty into the penalized likelihood in Zhang et al. (1998) for selecting important variables. Fan and Li (2004) considered a similar problem, whereas their approach differs from ours in that: (i) they used local polynomial regressions for estimating the nonparametric baseline function; (ii) they used marginal models for longitudinal data in the framework of the generalized estimation equation (GEE; Liang and Zeger, 1986); (iii) they in fact ignored the correlation among data by using working independence covariance structure. As a contrast, our approach explicitly models the within-subject correlation by using random effects and various Gaussian stochastic processes.

Compared with the marginal approach in Fan and Li (2004), our approach is

more advantageous in the following aspects. First, if the model is correctly specified, our estimators are more efficient. Secondly, missing data problems often occur in longitudinal studies, and the marginal model approach is only valid under missing completely at random (MCAR) assumption, which may be too stringent in practice; whereas our approach still yields valid inference when data are missing at random (MAR). Finally our method is based on linear mixed models, and hence can be easily implemented using existing software packages, such as the popular SAS PROC MIXED. However, like any model-based inference procedures, the gain in efficiency is conditional on the correct specification of the model. Mis-specified models may lead to biased inference. Performance and comparisons of our methods with others in various scenarios will be illustrated through simulations later in this chapter.

The rest of this chapter is organized as follows. In Section 3.2 we introduce the double-penalized likelihood approach. We describe the algorithm for maximizing the double-penalized log-likelihood in Section 3.3. The standard error formula are derived in Section 3.4. Section 3.5 reports the results from simulation studies. We illustrate our method using a real data analysis example in Section 3.6. Section 3.7 concludes the chapter with a discussion.

3.2 Double Penalized Likelihood Methodology

In this section, we first specify the semiparametric mixed model for longitudinal data. Then we introduce the double penalized likelihood methodology for automatic model selection for semiparametric mixed models.

3.2.1 Semiparametric Mixed Models

Suppose there are m subjects in the longitudinal study, with the ith subject having n_i repeated observations over time. Denote by y_{ij} $(i = 1, \ldots, m,\ j = 1, \ldots, n_i)$ the response at time point t_{ij}. As in Zhang et al. (1998), we consider the following semiparametric mixed model (SPMM):

$$y_{ij} = f(t_{ij}) + \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{b}_i + U_i(t_{ij}) + \varepsilon_{ij}, \tag{3.1}$$

where $f(t)$ is an arbitrary smooth and twice-differentiable baseline function, $\mathbf{x}_{ij}$ represent the covariate vector, β is a $d \times 1$ vector of regression coefficients, $\mathbf{z}_{ij}$ correspond to covariate vector for random coefficients, $\mathbf{b}_i$ is an $s \times 1$ vector of random coefficients, the $U_i(t)$ is independent mean zero Gaussian process modeling serial correlation, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ are independent measurement errors. Assume $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{G}(\phi))$, where $\mathbf{G}$ is a positive definite matrix parameterized by vector ϕ. The Gaussian process $U_i(t)$ has mean zero and covariance function $\mathrm{cov}(U_i(t), U_i(s)) = \gamma(\boldsymbol{\xi}, \alpha; t, s)$ for a specific parametric function $\gamma(\cdot)$ that depends on a parameter vector $\boldsymbol{\xi}$ and α, which are used to characterize the variance and correlation of the process $U_i(t)$. We further assume that $\mathbf{b}_i$, $U_i(t)$ and ε_{ij} are mutually independent.

We propose to model the within-subject serial correlation using a variety of stationary or non-stationary Gaussian stochastic processes $U_i(t)$. For example, a stationary Ornsterin-Uhlenbeck (OU) process can be used to model homogeneous within-subject covariance structure, which assumes constant variance and the correlation function to decay exponentially over time with the rate α, i.e. $\mathrm{corr}(U_i(t), U_i(s)) =$

$\exp(-\alpha|t-s|)$ (Diggle et al., 2002). If we assume that the variance function changes over time, for instance, in the form of $\log(\xi(t)) = \xi_0 + \xi_1(t)$, then the OU process generalizes to a nonhomogeneous Ornsterin-Uhlenbeck (NOU) process (Zhang et al., 1998). There are many other choices of Gaussian stochastic processes, such as Wiener process (Taylor et al., 1994) and integrated Wiener process (Wahba, 1978). Their detailed specifications and further references can be found in Zhang et al. (1998).

For convenience, we write the models in a matrix format. Define $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})^T$, and $\mathbf{X}_i$, $\mathbf{Z}_i$, $\mathbf{U}_i$, $\boldsymbol{\varepsilon}_i$ similarly for $i = 1, \ldots, m$. Let the total number of observations $n = \sum_{i=1}^{m} n_i$. Let $\boldsymbol{t^0} = (t_1^0, \ldots, t_r^0)^T$ be an $r \times 1$ vector of ordered distinct values of $\{t_i\}$ $(i = 1, \ldots, n)$, and let $\mathbf{N}$ be the incidence matrix connecting $\{t_i\}$ and $\boldsymbol{t^0}$, such that the (i,j)th element of $\mathbf{N}$ is 1 if $t_i = t_j^0$ and 0 otherwise $(j = 1, \ldots, r)$. Let $\mathbf{t}_i$ $(i = 1, \ldots, m)$ denote the vector of observed times for the i-th subject. Further denote $\mathbf{Y} = (Y_1^T, \ldots, Y_m^T)^T$ and $\mathbf{X}$, $\mathbf{N}$, $\boldsymbol{\varepsilon}$ similarly, and $\mathbf{Z} = \mathrm{diag}(\mathbf{Z}_1, \ldots, \mathbf{Z}_m)$. Then model (3.1) can be written in a matrix format as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{N}\mathbf{f} + \mathbf{Z}\mathbf{b} + \mathbf{U} + \boldsymbol{\varepsilon}, \tag{3.2}$$

where $\mathbf{f} = (f(t_1^0), \ldots, f(t_r^0))^T$, $\mathbf{b} = (\mathbf{b}_1^T, \ldots, \mathbf{b}_m^T)^T$ is normal $(\mathbf{0}, \mathcal{G}(\phi))$ with covariance $\mathcal{G}(\phi) = \mathrm{diag}(\mathbf{G}, \ldots, \mathbf{G})$; $\mathbf{U} = (\mathbf{U}_1^T, \ldots, \mathbf{U}_m^T)^T$ is normal$(0, \boldsymbol{\Gamma}(\boldsymbol{\xi}, \alpha))$ with covariance $\boldsymbol{\Gamma}(\boldsymbol{\xi}, \alpha) = \mathrm{diag}(\boldsymbol{\Gamma}_\mathbf{1}(\mathbf{t_1}, \mathbf{t_1}), \ldots, \boldsymbol{\Gamma}_\mathbf{m}(\mathbf{t_m}, \mathbf{t_m}))$ and the (j, j')th element $(j, j' = 1, \ldots, n_i)$ of $\boldsymbol{\Gamma}(\mathbf{t_i}, \mathbf{t_i})$ being $\gamma(\boldsymbol{\xi}, \alpha; t_{ij}, t_{ij'})$; and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}_n)$ with $\mathbf{I}_n$ being a $n \times 1$ identity matrix.

As we assume a Gaussian distribution for the data, we can easily write out the marginal likelihood for the SPMM model (3.1). For fixed variance components $\boldsymbol{\theta} =$

$(\boldsymbol{\phi}^T, \boldsymbol{\xi}^T, \alpha, \sigma^2)^T$, the log-likelihood function of $(\beta, \mathbf{f})$ is (up to a constant):

$$\ell(\beta, \mathbf{f}; \mathbf{Y}) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{N}\mathbf{f}), \tag{3.3}$$

where $\mathbf{V} = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_m)$ and $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \boldsymbol{\Gamma}_i + \sigma^2\mathbf{I}_{n_i}$.

3.2.2 Double Penalized Likelihood

One main purpose of this chapter is to select important linear effects in the semiparametric mixed model (3.1) for longitudinal data. To this end, we propose to maximize the double penalized likelihood (DPL) function:

$$\ell_{dp}(\boldsymbol{\beta}, f(\cdot); \mathbf{Y}) = \ell(\beta, \mathbf{f}; \mathbf{Y}) - \frac{\lambda_1}{2}\int_{T_1}^{T_2}\{f''(t)\}^2dt - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \tag{3.4}$$

where $\lambda_1 \geq 0$ is a smoothing parameter controlling the balance between the goodness of fit and the roughness of the estimated $f(t)$, and T_1 and T_2 specify the range of t; $p_{\lambda_2}(\cdot)$ is a shrinkage penalty function with $\lambda_2 \geq 0$ controlling the amount of shrinkage.

It can be shown that for given λ_1 and λ_2, maximizing the DPL (3.4) leads to a natural cubic spline estimate for $f(t)$. We call the maximizers $(\widehat{\beta}, \widehat{f})$ maximum double-penalized likelihood estimators (MDPLEs). Note that when $\lambda_2 = 0$, the MDPLEs reduce to the maximum penalized estimators (MPLEs) of Zhang et al. (1998). Choices of tuning parameters are important to assure effective model selection and estimation, and we will discuss it later. In the DPL (3.4), we adopt the same SCAD penalty defined by (2.3) in Chapter 2.

3.3 A Unified Algorithm for Obtaining MDPLEs

In this section, we formulate a linear mixed model (LMM) representation for the SPMM similar to Zhang et al. (1998), based on which we propose an iterative algorithm for solving for the MDPLEs and estimating all other model parameters.

3.3.1 Linear Mixed Model Representation

By Theorem 2.1 in Green and Silverman (1994), we can further rewrite the DPL (3.4) as

$$\ell_{dp}(\boldsymbol{\beta}, f(\cdot); \mathbf{Y}) = \ell(\beta, \mathbf{f}; \mathbf{Y}) - \frac{\lambda_1}{2}\mathbf{f}^T\mathbf{K}\mathbf{f} - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \tag{3.5}$$

where $\mathbf{K}$ is the nonnegative definite smoothing matrix defined by Green and Silverman (1994). Following Green (1987), we write $\mathbf{f}$ via a one-to-one linear transformation as

$$\mathbf{f} = \mathbf{T}\boldsymbol{\delta} + \mathbf{B}\mathbf{a}, \tag{3.6}$$

where $\mathbf{T} = [\mathbf{1}, \boldsymbol{t}^0]$ and $\mathbf{1}$ is an $r \times 1$ vector of 1's, $\boldsymbol{\delta}$ and $\mathbf{a}$ are respectively 2×1 and $(r-2) \times 1$ vectors, and $\mathbf{B} = \mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}$ with $\mathbf{L}$ being an $r \times (r-2)$ full rank matrix satisfying $\mathbf{K} = \mathbf{L}\mathbf{L}^T$ and $\mathbf{L}^T\mathbf{T} = 0$. It follows that $\mathbf{f}^T\mathbf{K}\mathbf{f} = \mathbf{a}^T\mathbf{a}$.

For fixed λ_2 and a in the second penalty in (3.5), we now propose to cast the semiparametric stochastic mixed model (3.1) into a unified linear mixed model framework (Zhang et al., 1998). As we show later, this framework will greatly facilitate the estimation of MDPLEs $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{f}})$, the smoothing parameter λ_1, and variance components

$\boldsymbol{\theta}$. Plugging (3.6) into (3.5) yields an equivalent double-penalized log-likelihood

$$\begin{aligned}
\ell_{dp}(\boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{a}; \mathbf{Y}) &= -\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{NT}\boldsymbol{\delta} - \mathbf{NBa})^T \mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{NT}\boldsymbol{\delta} - \mathbf{NBa}) \\
&\quad -\frac{\lambda_1}{2}\mathbf{a}^T\mathbf{a} - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|) - \frac{1}{2}\log|\mathbf{V}|, \\
&= -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_* - \mathbf{B}_*\mathbf{a})^T \mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_* - \mathbf{B}_*\mathbf{a}) \\
&\quad -\frac{\lambda_1}{2}\mathbf{a}^T\mathbf{a} - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \qquad (3.7)
\end{aligned}$$

where $\mathbf{X}_* = [\mathbf{NT}, \mathbf{X}]$, $\mathbf{B}_* = \mathbf{NB}$, $\boldsymbol{\beta}_* = (\boldsymbol{\delta}^T, \boldsymbol{\beta}^T)^T$. For fixed $\boldsymbol{\beta}_*$ (and given λ_1,λ_2,σ^2,a), maximizing (3.7) with respect to $\mathbf{a}$ leads to

$$\widehat{\mathbf{a}}(\boldsymbol{\beta}_*) = (\lambda_1\mathbf{I} + \mathbf{B}_*^T\mathbf{V}^{-1}\mathbf{B}_*)^{-1}\mathbf{B}_*^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*) \equiv \tilde{\mathbf{A}}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*), \qquad (3.8)$$

where $\tilde{\mathbf{A}} = (\lambda_1\mathbf{I} + \mathbf{B}_*^T\mathbf{V}^{-1}\mathbf{B}_*)^{-1}\mathbf{B}_*^T\mathbf{V}^{-1}$. Plugging $\widehat{\mathbf{a}} = \tilde{\mathbf{A}}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*)$ back into (3.7) and ignoring $-(1/2)\log|\mathbf{V}|$, we get a profile penalized likelihood:

$$\ell_{dp}(\boldsymbol{\beta}_*; \mathbf{Y}) = -\frac{1}{2}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*)^T\mathbf{W}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*) - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \qquad (3.9)$$

where $\mathbf{W} = (\mathbf{I} - \mathbf{B}_*\tilde{\mathbf{A}})^T\mathbf{V}^{-1}(\mathbf{I} - \mathbf{B}_*\tilde{\mathbf{A}}) + \lambda_1\tilde{\mathbf{A}}^T\tilde{\mathbf{A}}$. Note that the first two components of $\boldsymbol{\beta}_*$, namely $\boldsymbol{\delta}$, are not penalized in (3.9). We can then conduct variable selection by maximizing (3.9). After important variables are selected and their coefficients are estimated, the estimates $\widehat{\delta}$ and $\widehat{\boldsymbol{a}}$ in (3.8) can be used to construct the smoothing spline estimate of $f(t)$.

From an alternative point of view, the new DPL (3.7) can be regarded as the joint

log-likelihood for the following linear mixed model subject to the SCAD penalty for $\boldsymbol{\beta}$

$$\mathbf{Y} = \mathbf{X}_*\boldsymbol{\beta}_* + \mathbf{B}_*\mathbf{a} + \boldsymbol{\varepsilon}_*, \tag{3.10}$$

where $\boldsymbol{\beta}_* = (\boldsymbol{\delta}^T, \boldsymbol{\beta}^T)^T$ are fixed effects, and $\mathbf{a}$ is treated as random effects distributed as $\mathbf{a} \sim N(0, \tau\boldsymbol{I})$ with $\tau = 1/\lambda_1$, $\boldsymbol{\varepsilon}_* = \mathbf{Z}\mathbf{b} + \mathbf{U} + \boldsymbol{\varepsilon}$ distributed as $N(\mathbf{0}, \mathbf{V})$, and $\boldsymbol{\theta}_* = (\tau, \boldsymbol{\theta}^T)^T$ are the variance components. The variance matrix of $\mathbf{Y}$ under mixed model representation (3.10) is $\mathbf{V}_* = \mathbf{V} + \tau\mathbf{B}_*\mathbf{B}_*^T$. Simple matrix algebra then shows $\mathbf{V}_*^{-1} = \mathbf{W}$. If we conduct variable selection for $\mathbf{X}$ in the linear mixed (3.10) by maximizing the penalized log-likelihood of $\boldsymbol{\beta}_*$ with the same penalty for $\boldsymbol{\beta}$, i.e.

$$\ell_{dp}(\boldsymbol{\beta}_*; \mathbf{Y}) = -\frac{1}{2}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*)^T\mathbf{V}_*^{-1}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*) - n\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \tag{3.11}$$

we will get exactly the same solution as before due to the equivalence between this penalized likelihood (3.11) and the profile penalized likelihood (3.9). It is also easy to show that $\widehat{\mathbf{a}}(\boldsymbol{\beta}_*)$ in (3.8) is the same as the best linear unbiased prediction (BLUP) estimate of $\mathbf{a}$ from the mixed model (3.10) after $\boldsymbol{\beta}_*$ is estimated. The above argument therefore implies that the estimation of the nonparametric function $f(t)$ and the variable selection for $\mathbf{X}$ for the SPMM (3.1) can be realized through this unified mixed model framework.

As in Chapter 2, the above mixed model representation also indicates that the inverse of the smoothing parameter τ can be treated as a variance component and hence can be jointly estimated with $\boldsymbol{\theta}$ using maximum likelihood or restricted maximum likelihood (REML) approach during the variable selection process. This treatment gives

us great computational advantage by avoiding expensive two-dimensional search for two tuning parameters, and also enabling the convenient implementation in standard software packages.

3.3.2 Local Quadratic Approximation

Similar to the steps in Section 2.3.2, we locally approximate the DPL (3.11) by

$$\ell_{dp}(\widehat{\boldsymbol{\beta}}_*|\widehat{\boldsymbol{\beta}}_*^0) \approx -\frac{1}{2}(\mathbf{Y}-\mathbf{X}_*\widehat{\boldsymbol{\beta}}_*)^T\mathbf{V}_*^{-1}(\mathbf{Y}-\mathbf{X}_*\widehat{\boldsymbol{\beta}}_*) - \frac{1}{2}\widehat{\boldsymbol{\beta}}_*^T\boldsymbol{\Sigma}_{\lambda_2}(\widehat{\boldsymbol{\beta}}_*^0)\widehat{\boldsymbol{\beta}}_* - n\sum_{j=3}^{d+2}\left\{p_{\lambda_2}(|\widehat{\beta}_{*j}^0|) - \frac{1}{2}\frac{p'_{\lambda_2}(|\widehat{\beta}_{*j}^0|)}{|\widehat{\beta}_{*j}^0|}(\widehat{\beta}_{*j}^0)^2\right\},$$

where $\boldsymbol{\Sigma}_{\lambda_2}(\boldsymbol{\beta}_*) = \text{diag}\{0, 0, p'_{\lambda_2}(|\beta_1|)/|\beta_1|, \ldots, p'_{\lambda_2}(|\beta_d|)/|\beta_d|\}$. For fixed $\boldsymbol{\theta}_* = (\tau, \boldsymbol{\theta}^T)^T$, we apply the Newton-Raphson method to maximize $\ell_{dp}(\widehat{\boldsymbol{\beta}}_*|\widehat{\boldsymbol{\beta}}_*^0)$ and get the updating formula

$$\widehat{\boldsymbol{\beta}}_* = \left\{\mathbf{X}_*^T\mathbf{W}\mathbf{X}_* + n\boldsymbol{\Sigma}_{\lambda_2}(\widehat{\boldsymbol{\beta}}_*^0)\right\}^{-1}\mathbf{X}_*^T\mathbf{W}\mathbf{Y}, \tag{3.12}$$

where $\mathbf{W} = \mathbf{V}_*^{-1} = (\mathbf{V} + \tau\mathbf{B}_*\mathbf{B}_*^T)^{-1}$. Note that $\mathbf{V}_*$ is $n \times n$ and no longer has block-diagonal structure, and therefore direct inverse will be difficult. Instead we use $\mathbf{V}_*^{-1} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{B}_*(\tau^{-1}\mathbf{I}_r + \mathbf{B}_*^T\mathbf{V}^{-1}\mathbf{B}_*)^{-1}\mathbf{B}_*^T\mathbf{V}^{-1}$, which is computationally a more efficient formula when r (number of distinct t_i's) is much smaller than n (total sample size). It is easy to recognize that (3.12) is in fact an iterative ridge regression algorithm.

3.3.3 Estimation of λ_1 and Variance Components

The iterative algorithm discussed in Section 3.3.2 is based on fixed or known smoothing parameter λ_1 (or equivalently τ) and $\boldsymbol{\theta}$. However, they are usually unknown and need to be estimated. Parallel to Section 2.4.1 in Chapter 2, let us denote by $\mathbf{X}_{[s]}$ the subset of important variables selected from $\mathbf{X}$ at the sth iteration. The REML log-likelihood of $(\tau, \boldsymbol{\theta})$ at this iteration is

$$\ell_R^{[s]}(\tau, \boldsymbol{\theta}; \mathbf{Y}) = -\frac{1}{2}\log|\mathbf{V}_*| - \frac{1}{2}\log\left|\mathbf{X}_{*[s]}^T\mathbf{V}_*^{-1}\mathbf{X}_{*[s]}\right| - \frac{1}{2}(\mathbf{Y}-\mathbf{X}_{*[s]}\widehat{\boldsymbol{\beta}}_{*[s]})^T\mathbf{V}_*^{-1}(\mathbf{Y}-\mathbf{X}_{*[s]}\widehat{\boldsymbol{\beta}}_{*[s]}),$$

where $\mathbf{X}_{*[s]} = \left[\mathbf{NT}, \mathbf{X}_{[s]}\right]$, $\widehat{\boldsymbol{\beta}}_{*[s]} = (\mathbf{X}_{*[s]}^T\mathbf{V}_*^{-1}\mathbf{X}_{*[s]})^{-1}\mathbf{X}_{*[s]}^T\mathbf{V}_*^{-1}\mathbf{Y}$ is the MLE of $\boldsymbol{\beta}_*$ based on the selected important variables $\mathbf{X}_{[s]}$ and $\mathbf{V}_*$ is defined in Section 3.1. Differentiating $\ell_R^{[s]}(\tau, \sigma^2; \mathbf{Y})$ with respect to τ and $\boldsymbol{\theta}$ and use the identity $\mathbf{V}_*^{-1}(\mathbf{Y} - \mathbf{X}_{*[s]}\widehat{\boldsymbol{\beta}}_{[s]}) = \mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}_{[s]}\widehat{\boldsymbol{\beta}}_{[s]} - \mathbf{N}\widehat{\mathbf{f}})$, we obtain the REML estimating equations for τ and $\boldsymbol{\theta}$ (Harville, 1977):

$$-\frac{1}{2}\mathrm{tr}(\mathbf{P}\mathbf{B}_*\mathbf{B}_*^T) + \frac{1}{2}(\mathbf{Y} - \mathbf{X}_{[s]}\widehat{\boldsymbol{\beta}}_{[s]} - \mathbf{N}\widehat{\mathbf{f}})^T\mathbf{V}^{-1}\mathbf{B}_*\mathbf{B}_*^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}_{[s]}\widehat{\boldsymbol{\beta}}_{[s]} - \mathbf{N}\widehat{\mathbf{f}}) = 0$$
$$-\frac{1}{2}\mathrm{tr}(\mathbf{P}) + \frac{1}{2}(\mathbf{Y} - \mathbf{X}_{[s]}\widehat{\boldsymbol{\beta}}_{[s]} - \mathbf{N}\widehat{\mathbf{f}})^T\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\theta_i}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}_{[s]}\widehat{\boldsymbol{\beta}}_{[s]} - \mathbf{N}\widehat{\mathbf{f}}) = 0,$$

where $\mathbf{P} = \mathbf{V}_*^{-1} - \mathbf{V}_*^{-1}\mathbf{X}_{*[s]}(\mathbf{X}_{*[s]}^T\mathbf{V}_*^{-1}\mathbf{X}_*^T)^{-1}\mathbf{X}_{*[s]}^T\mathbf{V}_*^{-1}$. Zhang et al. (1998) showed that the computation of matrix $\mathbf{P}$ can be dramatically simplified by using the identity: $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathcal{H}\mathbf{D}^{-1}\mathcal{H}^T\mathbf{V}^{-1}$, where $\mathcal{H}$ is the coefficient matrix of the following

systems of equations

$$\mathcal{H}\begin{bmatrix}\boldsymbol{\beta}_* \\ \mathbf{a}\end{bmatrix} \equiv \begin{bmatrix}\mathbf{X}_*^T\mathbf{W}\mathbf{X}_* & \mathbf{X}_*^T\mathbf{W}\mathbf{B}_* \\ \mathbf{B}_*^T\mathbf{W}\mathbf{X}_* & \mathbf{B}_*^T\mathbf{W}\mathbf{B}_* + \lambda_1\mathbf{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_* \\ \mathbf{a}\end{bmatrix} = \begin{bmatrix}\mathbf{X}_*^T\mathbf{W}\mathbf{Y} \\ \mathbf{B}_*^T\mathbf{W}\mathbf{Y}\end{bmatrix}.$$

The Fisher scoring algorithm can then be used to solve the above REML equations for τ and $\boldsymbol{\theta}$. The standard errors for $\widehat{\boldsymbol{\theta}}$ can be obtained from the inverse of the expected fisher information matrix at convergence.

3.3.4 Choice of the SCAD Tuning Parameters

The aforementioned methods are based on fixed SCAD tuning parameters (λ_2, a), which need to be tuned for the optimal values. By the same argument in Chapter 2, we set $a = 3.7$. In addition, we propose using the Bayesian Information Criterion (BIC) (Schwarz, 1978) to select the other regularization parameter λ_2 from a gridded range. Given λ_2, suppose q variables are selected. Let $\mathbf{X}_1$ be the sub-matrix of $\mathbf{X}$ for q important variables and $\boldsymbol{\beta}_1$ the $q \times 1$ corresponding regression coefficient vector. Then we may use methods of Zhang et al. (1998) to solve SPMM (3.1). Consequently $\widehat{\mathbf{Y}} = \mathbf{S}\mathbf{Y}$, where $\mathbf{S}$ is a smoother matrix with $q_1 = \text{trace}(\mathbf{S})$. The BIC criterion is computed as

$$\text{BIC}(\lambda_2) = -2\ell + q_1 \log n, \tag{3.13}$$

where $\ell = -(n/2)\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - (1/2)(\mathbf{Y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{N}\widehat{\mathbf{f}})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 - \mathbf{N}\widehat{\mathbf{f}})$. It can be shown that $\widehat{\mathbf{Y}} = \mathcal{H}\mathbf{D}^{-1}\mathcal{H}^T\mathbf{W}\mathbf{Y}$. Then it is easy to compute

$$\text{trace}(\mathbf{S}) = \text{trace}(\mathcal{H}\mathbf{D}^{-1}\mathcal{H}^T\mathbf{W}) = \text{trace}(\mathbf{D}^{-1}\mathcal{H}^T\mathbf{W}\mathcal{H}).$$

3.3.5 A Unified Algorithm

We propose to alternately estimate $(\boldsymbol{\beta}, \mathbf{f})$ and $\boldsymbol{\theta}_* = (\tau, \boldsymbol{\theta})$ iteratively until the algorithm converges.

1. Use the method of Zhang et al. (1998) to obtain the MPLEs of $\boldsymbol{\beta}_*$ and θ_* with the *full* model. They will be used as the initial values.

2. Choose a gridded range for λ_2 (the SCAD tuning parameter). For each grid point λ_{2i}, iteratively update $\widehat{\boldsymbol{\beta}}_*$ and $\widehat{\theta}_*$ until convergence. Specifically, at the k-th iteration:

   (i) Threshold $\widehat{\boldsymbol{\beta}}_*$ by a small threshold η, say, $\eta = 10^{-5}$. That is, if $|\widehat{\beta}_j| < \eta$, then set $\widehat{\beta}_j = 0$.

   (ii) Compute $\widehat{\boldsymbol{\beta}}_*^{(k)}$ from $\widehat{\boldsymbol{\beta}}_*^{(k-1)}$ by the updating formula (3.12).

   (iii) Use the REML to re-estimate $\boldsymbol{\theta}_*$ with selected variables.

   (iv) Check convergence: if $\max_j \left\{ \frac{|\widehat{\beta}_{*j}^{(k)} - \widehat{\beta}_{*j}^{(k-1)}|}{|\widehat{\beta}_{*j}^{(k-1)}|} \right\} < tol$, where tol is a small tolerance, and compute the BIC score BIC($\lambda_{2\mathrm{i}}$); otherwise go to Step 2(i).

3. Identify the value of λ_2 which minimizes the BIC score. Obtain the final estimates for $(\boldsymbol{\beta}, \boldsymbol{\theta})$.

4. Compute the estimate $\widehat{\mathbf{f}}$ (through $\widehat{\mathbf{f}} = \mathbf{T}\widehat{\boldsymbol{\delta}} + \mathbf{B}\widehat{\mathbf{a}}$).

3.4 Frequentist and Bayesian Standard Errors

In this section, we derive the frequentist and Bayesian covariance formula for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{f}}$ parallel to Section 2.5 in Chapter 2. Using these covariance estimates, we are able to calculate standard errors and construct confidence intervals for the regression coefficients and the nonparametric function. The proposed covariance estimates are evaluated via simulations.

3.4.1 Frequentist Covariance Estimates

From frequentist point of view, $\text{cov}(\mathbf{Y}|t, \mathbf{x}) = \mathbf{V}$, where $\mathbf{V}$ is defined in Section 3.2.1. By the LQA, at convergence we can write $\widehat{\boldsymbol{\beta}}_* = (\widehat{\boldsymbol{\delta}}^T, \widehat{\boldsymbol{\beta}}^T)^T$ as an approximately linear function of $\mathbf{Y}$:

$$\widehat{\boldsymbol{\beta}}_* = \left\{\mathbf{X}_*^T\mathbf{W}\mathbf{X}_* + n\boldsymbol{\Sigma}_{\lambda_2}(\widehat{\boldsymbol{\beta}})\right\}^{-1}\mathbf{X}_*^T\mathbf{W}\mathbf{Y} \equiv \mathbf{Q}\mathbf{Y}.$$

Let $\mathbf{Q} = (\mathbf{Q}_1^T, \mathbf{Q}_2^T)^T$, where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are partitions of $\mathbf{Q}$ with dimensions corresponding to $(\boldsymbol{\delta}^T, \boldsymbol{\beta}^T)^T$, so that $\widehat{\boldsymbol{\delta}} = \mathbf{Q}_1\mathbf{Y}$, and $\widehat{\boldsymbol{\beta}} = \mathbf{Q}_2\mathbf{Y}$. The estimated covariance matrix for $\widehat{\boldsymbol{\beta}}$ is given by

$$\widehat{\text{cov}}(\widehat{\boldsymbol{\beta}}) = \mathbf{Q}_2\text{cov}(\mathbf{Y})\mathbf{Q}_2^T = \mathbf{Q}_2\widehat{\mathbf{V}}\mathbf{Q}_2^T, \tag{3.14}$$

where $\widehat{\mathbf{V}}$ is the estimated variance matrix. By the profile formula (3.8), we have $\widehat{\mathbf{a}}(\boldsymbol{\beta}_*) = \mathbf{S}_a\mathbf{Y}$, where $\mathbf{S}_a = \tilde{\mathbf{A}}(\mathbf{I} - \mathbf{X}_*\mathbf{Q})$. Therefore $\widehat{\mathbf{f}} = \mathbf{T}\widehat{\boldsymbol{\delta}} + \mathbf{B}\widehat{\mathbf{a}} = (\mathbf{T}\mathbf{Q}_1 + \mathbf{B}\mathbf{S}_a)\mathbf{Y}$

and its covariance can be computed as

$$\widehat{\text{cov}}_F(\widehat{\mathbf{f}}|t,\mathbf{x}) = (\mathbf{TQ}_1 + \mathbf{BS}_a)\widehat{\mathbf{V}}(\mathbf{TQ}_1 + \mathbf{BS}_a)^T. \tag{3.15}$$

3.4.2 Bayesian Covariance Estimates

Following the same steps in Section 2.5.2 in Chapter 2, we can reformulate the mixed model (3.10) and arrive at a modified mixed model similar to (2.35) as

$$\mathbf{Y} = \mathcal{X}_*\boldsymbol{\gamma} + \mathbf{Z}_*\mathbf{b}_* + \boldsymbol{\varepsilon}_*, \tag{3.16}$$

where $\mathcal{X}_* = [\mathbf{NT}, \mathbf{X}_1]$, $\boldsymbol{\gamma} = (\boldsymbol{\delta}^T, \boldsymbol{\beta}_1^T)^T$, $\mathbf{Z}_* = [\mathbf{X}_2, \mathbf{B}_*]$, and $\mathbf{b}_* = (\boldsymbol{\beta}_2^T, \mathbf{a}^T)^T$ are the new random effects distributed as $N(\mathbf{0}, \mathbf{G}_b)$ with a block diagonal covariance matrix $\mathbf{G}_b = \text{diag}(\boldsymbol{\Sigma}_{22}^{-1}, \tau\mathbf{I}_{r-2})$, and $\boldsymbol{\varepsilon}_* = \mathbf{Zb} + \mathbf{U} + \boldsymbol{\varepsilon}$ distributed as $N(\mathbf{0}, \mathbf{V})$. Let $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\mathbf{b}}_*$ be the estimators by solving model (3.16) using conventional mixed model theory. Under model (3.16), the coefficient matrix of Henderson's mixed model equations is given by

$$\mathbf{C}_* = \begin{pmatrix} \mathcal{X}_*^T\mathbf{V}^{-1}\mathcal{X}_* & \mathcal{X}_*^T\mathbf{V}^{-1}\mathbf{Z}_* \\ \mathbf{Z}_*^T\mathbf{V}^{-1}\mathcal{X}_* & \mathbf{Z}_*^T\mathbf{V}^{-1}\mathbf{Z}_* + \mathbf{G}_b^{-1} \end{pmatrix}.$$

Henderson (1975) showed that provided $\mathcal{X}_*$ is a full rank matrix, the covariance matrix for $(\widehat{\boldsymbol{\gamma}}, \widehat{\mathbf{b}})$ is

$$\text{cov}\begin{pmatrix} \widehat{\boldsymbol{\gamma}} \\ \widehat{\mathbf{b}}_* - \mathbf{b}_* \end{pmatrix} = \mathbf{C}_*^{-1}.$$

Therefore the Bayesian covariances for $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{f}})$ are

$$\begin{aligned} \mathrm{cov}_B(\widehat{\boldsymbol{\beta}}) &= \mathrm{cov}\{\widehat{\boldsymbol{\beta}}_1^T, (\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)^T\}^T, \quad (3.17) \\ \mathrm{cov}_B(\widehat{\mathbf{f}}) &= [\mathbf{T}, \mathbf{B}]\, \mathrm{cov}\{\widehat{\boldsymbol{\delta}}^T, (\widehat{\mathbf{a}} - \mathbf{a})^T\}^T [\mathbf{T}, \mathbf{B}]^T. \quad (3.18) \end{aligned}$$

The covariance terms in the right hand side of (3.17) and (3.18) can be extracted from $\mathbf{C}_*^{-1}$. These Bayesian variance estimates can be viewed to account for the bias in $\widehat{\beta}$ and $\widehat{\mathbf{f}}$ due to imposed penalties (Wahba, 1983).

3.5 Simulation Studies

We conduct Monte Carlo simulations to evaluate the performance of the proposed method. We also compare our procedure with the SCAD and LASSO methods in Fan and Li (2004). When implementing Fan and Li (2004), we adopt their approach to choose the kernel bandwidth: first compute the difference based estimator (DBE) for $\boldsymbol{\beta}$ and then select the bandwidth using the plug-in method of Ruppert et al. (1995). Same as in Fan and Li (2004), the SCAD and LASSO tuning parameters are selected by GCV. As in Fan and Li (2004), we use the mean squares error (MSE) for $\widehat{\boldsymbol{\beta}}$ and $\widehat{f}$ to respectively evaluate goodness-of-fit for parametric and nonparametric estimation. They are defined as

$$MSE(\widehat{\beta}) = E(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2), \quad (3.19)$$

and

$$MSE(\widehat{f}) = E\left[\int_{T_1}^{T_2} \{\widehat{f}(t) - f(t)\}^2 dt\right]. \quad (3.20)$$

For each Monte Carlo sample, we estimate the integration $\int_{T_1}^{T_2}\{\widehat{f}(t) - f(t)\}^2 dt$ empirically by averaging over the design knots. Then we take the Monte Carlo sample means to estimate $MSE(\widehat{\beta})$ and $MSE(\widehat{f})$.

To evaluate the variable selection performance of each method, we report the number of correct zero coefficients (denoted as "Corr."), the number of coefficients incorrectly set to 0 (denoted as "Inc."), and the model size. In addition, we report the point estimates, biases, and the 95% coverage probabilities of frequentist and Bayesian confidence intervals for the DPLSE. In our implementation for the SCAD and LASSO, the bandwidth selected using the plug-in method occasionally caused numerical problems and failed to converge. Therefore, the results of SCAD and LASSO are only based on converged cases. In contrast, our procedure always converges.

3.5.1 Simulation Design

We simulate data from a longitudinal clinical trial with staggered entry time. Suppose there are totally $m = 40$ patients, divided into 5 groups, each one with 8 patients. Consider a staggered entry design, where the k-th group entered on the k-th time points. Each patient has 10 equally spaced scheduled visits, with 5 time units between adjacent visits. For example, subjects numbered by 1 to 8 enter at $t = 1$ and the second scheduled visit will be at $t = 6$. Hence there are 50 distinct time points, and on each time point there are 8 patients. The total number of observations for the full data is 400. We let the total length of study be 4. Table 3.1 shows the scheduled visit times for subject 1 to subject 40.

Table 3.1: Staggered entry design for simulated longitudinal data.

Subject	Visit times									
1-8	1	6	11	16	21	26	31	36	41	46
9-16	2	7	12	17	22	27	32	37	42	47
17-24	3	8	13	18	23	28	33	38	43	48
25-32	4	9	14	19	24	29	34	39	44	49
33-40	5	10	15	20	25	30	35	40	45	50

We simulate longitudinal data from the following semiparametric mixed model:

$$y_{ij} = f(t_{ij}) + \boldsymbol{x}_i^T \boldsymbol{\beta} + b_{0i} + U_i(t_{ij}) + \varepsilon_{ij}, \tag{3.21}$$

where $i = 1, \ldots, 40$ and $j = 1, \ldots, 10$. Now we describe the detailed specifications of model parameters for (3.21). We choose $f(t) = 4\sin(2\pi t/4)$ for the nonparametric baseline function. By the staggered entry design, there are 50 knots for the smoothing spline fit. There are eight mutually independent x-covariates $(x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)^T$, which are simulated from a standard normal distribution. For simplicity we assume that the covariates are constant over time. The true regression coefficients are $\boldsymbol{\beta} = (.8, .8, 0, 0, .8, 0, 0, 0)^T$. The random intercept b_{0i} in model (3.21) represents the among-subjects variation, and b_{0i} is simulated from $N(0, v_1)$, with $v_1 = 0.36$.

We consider a stationary Ornstein-Uhlenbeck (OU) process for the within-subject covariance structure. The OU process assumes a constant variance σ_u^2 and exponential serial correlation:

$$corr[U_i(t), U_i(s)] = \exp(-\alpha|t-s|) = \rho^{|t-s|}.$$

Parameters (ρ, σ_u^2) are two variance components. The OU process is an ante-dependence (Markovian) model of order 1 (Diggle et al. 2002, sec 5.2, pg 89), which means that the conditional distribution of $U_i(t_j)$ given all its prior states depends only on the value of $U_i(t_{j-1})$. This allows us to simulate the vector of $\{U_i(t_1), \ldots, U_i(t_{10})\}^T$ sequentially after $U_i(t_1) \sim N(0, \sigma_u^2)$ is generated:

$$U_i(t_j)|U_i(t_{j-1}) \sim N\left(\varphi_j U_i(t_{j-1}), \sigma_u^2(1-\varphi_j^2)\right),\ j = 2, \ldots, 10,$$

where $\varphi_j = \rho^{|t_j - t_{j-1}|}$. In our simulations we set $\rho = 0.4$ and $\sigma_u^2 = 0.5$. We also simulate the measurement error in model (3.21): $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, where $\sigma_\varepsilon^2 = 0.25$. Therefore the true variance components are $\boldsymbol{\theta} \equiv (v_1, \rho, \sigma_u^2, \sigma_\varepsilon^2) = (0.36, 0.4, 0.5, 0.25)$.

We compare our DPL approach to the methods in Fan and Li (2004) in three scenarios: full data, missing observations and mis-specified models. Specifically, these three scenarios are described as follows.

- Full data. Each subject has 10 repeated observations. There are 400 observations in total.

- Missing data. Imagine that some patients may drop out of the trial at some time point during the study. Dropout corresponds to monotone missingness, which means that if a subject misses visit t_j, all subsequent observations at $t_k > t_j$ are missing. We consider the missing at random case, where the missing mechanism depends only on observed data, not on unobserved data. Specifically, each patient has scheduled measurements up to the 6-th visit; from this point on, the probability of missing the current observation depends on the last observation

through a logit model: $logit(p_m) = -1 - 2y_{j-1}$, for $j \geq 7$. This is the missing at random model used in Section 13.3 of Diggle et al. (2002). In our simulation, this missing data model yielded on average 278.5 non-missing observations and approximately 30% missing data.

- Model mis-specification. When simulating data, we add a random slope to the model (3.21). Now the simulated data are generated from the following model:

$$y_{ij} = f(t_{ij}) + \boldsymbol{x}_i^T \boldsymbol{\beta} + b_{0i} + b_{1i} t_{ij} + U_i(t_{ij}) + \varepsilon_{ij},$$

where $b_{1i} \sim N(0, 0.16)$ is the random slope. If we still fit the random intercept model (3.21), we expect to get biased or invalid results from our method since it is a likelihood-based approach.

3.5.2 Simulation Results

Table 3.2 summarizes and compares the variable selection and model fitting results of the DPL, SCAD and LASSO methods in the three scenarios given in Section 3.5.1. Table 3.3 gives the point estimates, relative biases, empirical and model-based standard errors and 95% coverage probabilities. Table 3.4 presents the variance component estimation results. Figure 3.1 contains four sub-plots for $\widehat{f}(t)$: the estimated $f(t)$ provided by three procedures, point-wise biases by three procedures, point-wise empirical and model-based frequentist and Bayesian SEs using the DPL method, and 95% coverage probabilities using DPL. Similar plots are given in Figure 3.2 for the missing data scenario and in Figure 3.3 for the model mis-specification case.

Table 3.2: Comparison of variable selection procedures for simulated longitudinal data based on 100 replications [a].

Scenario	Method	MSE($\widehat{\beta}$)	MSE($\widehat{f}$)	Model Size (3)	Zero coef. Corr.(5)	Zero coef. Inc.(0)
Full data	DPL	0.09 (0.17)	0.03 (0.03)	3.09	4.87	0.04
	SCAD	0.08 (0.09)	0.05 (0.03)	3.26	4.79	0
	LASSO	0.08 (0.09)	0.03 (0.02)	3.30	4.70	0
Missing data [b]	DPL	0.10 (0.23)	0.14 (0.05)	3.11	4.83	0.06
	SCAD	0.12 (0.14)	0.28 (0.10)	3.38	4.59	0.02
	LASSO	0.11 (0.11)	0.29 (0.11)	3.47	4.52	0.01
Mis-specified	DPL	0.62 (0.66)	0.55 (1.46)	2.24	4.95	0.81
	SCAD	0.23 (0.31)	0.09 (0.06)	3.35	4.57	0.08
	LASSO	0.22 (0.29)	0.05 (0.05)	3.41	4.53	0.07

[a] SCAD and LASSO estimates are based on 87, 86 and 91 converged Monte Carlo samples for the full data, missing data and mis-specification scenarios respectively.

[b] The amount of missing data is approximately 30% on average.

Full data case. It can be seen from Table 3.2 that the DPL yields better model sizes (closer to the true model size 3) with better number of correct zeros (true model has 5) than the other two methods. It did better overall in terms of model selection, although the DPL on average incorrectly produced 0.04 incorrect zeroes. The MSEs for $\widehat{\boldsymbol{\beta}}$ and $\widehat{f}$ from DPL are comparable with those of the SCAD and LASSO estimates. In terms of the estimation for $\boldsymbol{\beta}$, Table 3.3 shows that the relative biases of $\widehat{\beta}_j$ by using our DPL method are the smallest among all the three methods. All three methods produced confidence intervals with good coverage probabilities. From Table 3.4, we see that the variance components using our DPL are estimated with small bias. Figure 3.1 shows that the DPL has smaller overall bias in $\widehat{f}(t)$ than the SCAD and LASSO.

Missing data. When data are missing at random, the DPL is still valid and maintains good performance in model selection and parameter estimation, whereas

the SCAD and LASSO no longer produce consistent estimates. For the SCAD and LASSO, although the model selection and relative biases for $\widehat{\beta}_j$ only get slightly worse compared with the full model case, significant biases can be seen in nonparametric function estimate $\widehat{f}(t)$ presented in plot (b) of Figure 3.2. In the same plot, the $\widehat{f}(t)$ using DPL maintains remarkably small biases, even though about 30% of the data were missing. As shown in plot (d) of Figure 3.2, the coverage probabilities using the DPL method still balance around the nominal level and seem to be unaffected by the missing data.

Model mis-specification. As seen from Table 3.2, 3.3 and 3.4 and Figure 3.3. the DPL method can be severely affected by model mis-specification. The SCAD and LASSO methods in Fan and Li (2004) are robust in this case because they are GEE type approaches that do not require the full marginal likelihood and possess consistent property under mis-specified models.

3.6 Real Data Analysis

We illustrate the DPL method using a subset of data from the Multi-Center AIDS Cohort study. The data contain the repeated measurements of CD4 cell counts and percentage and other related covariates of 283 homosexual men who were infected with the human immunodeficiency virus (HIV) between 1984 and 1991. All patients in this study had semi-annual visits for physical examinations. However, because many subjects missed some of their scheduled visits and the HIV infections occurred randomly in the study, there are unequal number of observations and different measurement times for each subject. Details for the study such as design, methods and

Table 3.3: Point estimation results for $\boldsymbol{\beta}$ in three scenarios, based on 100 replications.

	Model	Point	Relative	Empirical	Model-based SE		95% CP	
Method	Parameter	Estimate	Bias	SE	Freq.	Bayes.	Freq.	Bayes.
Scenario 1: Full data								
DPL	β_1	0.789	−0.014	0.167	0.129	0.129	0.93	0.93
	β_2	0.807	0.008	0.156	0.135	0.135	0.92	0.92
	β_5	0.787	−0.016	0.150	0.130	0.130	0.91	0.91
SCAD	β_1	0.774	−0.032	0.138	0.116	-	0.90	-
	β_2	0.788	−0.016	0.145	0.117	-	0.91	-
	β_5	0.760	−0.050	0.148	0.114	-	0.84	-
LASSO	β_1	0.761	−0.049	0.132	0.115	-	0.94	-
	β_2	0.772	−0.035	0.140	0.115	-	0.96	-
	β_5	0.746	−0.067	0.145	0.114	-	0.96	-
Scenario 2: Missing data								
DPL	β_1	0.780	−0.026	0.177	0.144	0.144	0.93	0.93
	β_2	0.795	−0.006	0.151	0.151	0.151	0.96	0.96
	β_5	0.782	−0.022	0.144	0.144	0.144	0.93	0.93
SCAD	β_1	0.765	−0.044	0.170	0.120	-	0.81	-
	β_2	0.778	−0.027	0.170	0.120	-	0.85	-
	β_5	0.760	−0.050	0.168	0.117	-	0.87	-
LASSO	β_1	0.752	−0.059	0.159	0.118	-	0.83	-
	β_2	0.768	−0.040	0.150	0.120	-	0.87	-
	β_5	0.746	−0.068	0.161	0.115	-	0.87	-
Scenario 3: Model mis-specification								
DPL	β_1	0.679	−0.151	0.409	0.160	0.160	0.73	0.73
	β_2	0.613	−0.234	0.431	0.151	0.151	0.69	0.69
	β_5	0.642	−0.198	0.418	0.156	0.156	0.71	0.71
SCAD	β_1	0.789	−0.014	0.215	0.149	-	0.88	-
	β_2	0.763	−0.046	0.233	0.157	-	0.86	-
	β_5	0.772	−0.036	0.229	0.148	-	0.84	-
LASSO	β_1	0.773	−0.033	0.207	0.147	-	0.88	-
	β_2	0.756	−0.055	0.215	0.156	-	0.87	-
	β_5	0.758	−0.053	0.221	0.145	-	0.87	-

Table 3.4: Point estimation results for $\boldsymbol{\theta}$ in three scenarios from proposed DPL, based on 100 replications.

Scenario	Model Parameter	True Value	Point Estimate(SD)	Relative Bias
Full Data	v_1	0.36	0.36 (0.22)	0.00
	ρ	0.40	0.38 (0.23)	−0.05
	σ_u^2	0.50	0.58 (0.19)	0.16
	σ_u^2	0.25	0.21 (0.11)	−0.16
Missing Data	v_1	0.36	0.42 (0.34)	0.00
	ρ	0.40	0.36 (0.30)	−0.10
	σ_u^2	0.50	0.67 (0.19)	0.34
	σ_u^2	0.25	0.33 (0.21)	0.32
Mis-specification	v_1	0.36	0.03 (0.12)	−0.92
	ρ	0.40	0.89 (0.07)	1.23
	σ_u^2	0.50	2.01 (0.36)	3.02
	σ_u^2	0.25	0.66 (0.23)	1.64

medical implications can be found in Kaslow et al. (1987). Huang et al. (2002) used a varying-coefficient model to describe the trend of the mean CD4 percentage over time and to assess the effects of smoking, age and pre-HIV infection CD4 percentage on the mean CD4 percentage after the infection. Therefore they considered CD4 cell percentage to be the response with three covariates x_1, x_2 and x_3: x_1 is the smoking status after HIV infection, where $x_1 = 1$ for smokers and $x_1 = 0$ for non-smokers; x_2 is the centered age for individual's age at HIV infection; and x_3 is the centered pre-infection CD4 percentage. They fit the data using the following model:

$$y_{ij} = \beta_0(t_{ij}) + x_1\beta_1(t_{ij}) + x_2\beta_2(t_{ij}) + x_3\beta_3(t_{ij}) + \varepsilon_{ij}, \tag{3.22}$$

where ε_{ij}'s are the marginal errors. It is evident from the analysis of Huang et al. (2002) that only the baseline function varies over time and PreCD4 has a constant

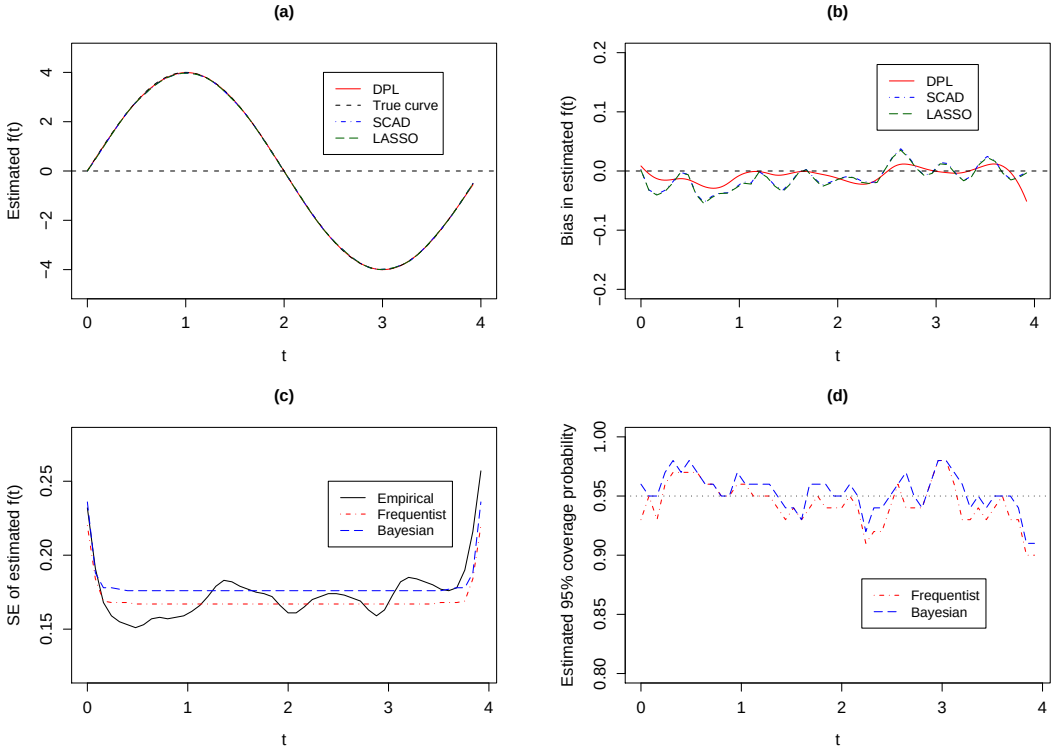

Figure 3.1: Plots for $\widehat{f}(t)$ in the full data scenario based on 100 replications.

effect over time, and neither smoking status or age has a significant effect on the mean CD4 percentage.

Fan and Li (2004) further analyzed the same data set with the purpose of variable selection in semiparametric regression models. After standardizing x_2 and x_3, they added the interactions of the three covariates and quadratic terms of x_2 and x_3 and considered the following semiparametric model:

$$y_{ij} = f(t_{ij}) + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_2^2\beta_4 + x_3^2\beta_5 + x_1x_2\beta_6 + x_1x_3\beta_7 + x_2x_3\beta_8 + \varepsilon_{ij}, \quad (3.23)$$

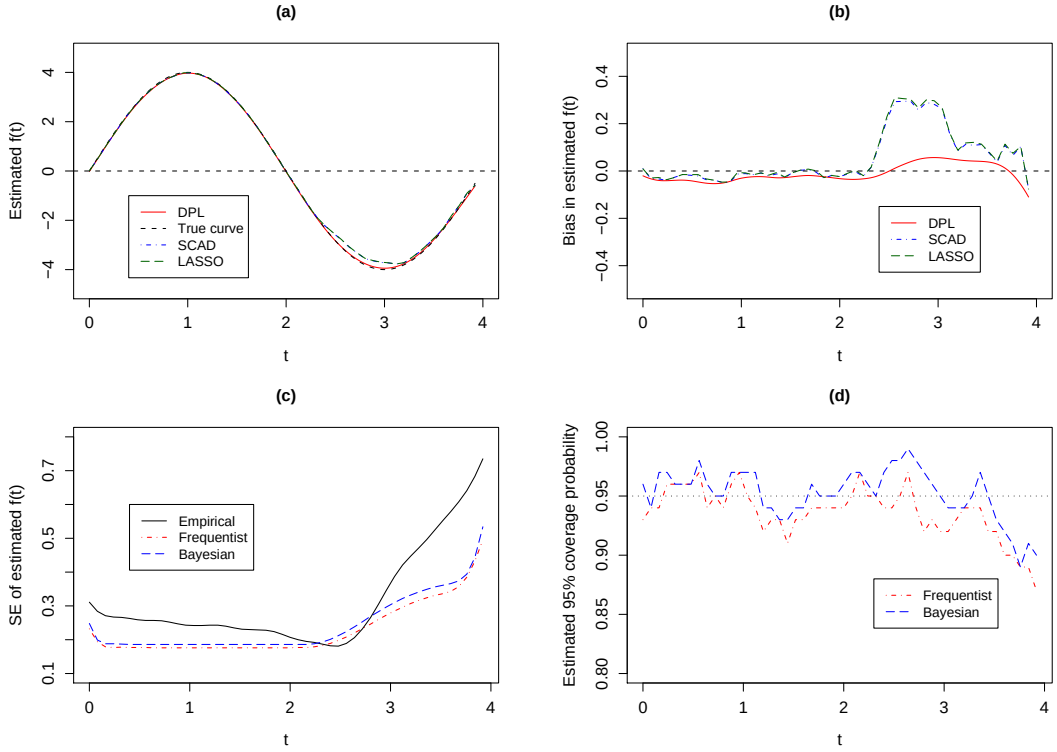

Figure 3.2: Plots for $\widehat{f}(t)$ in the missing at random data scenario based on 100 replications. Approximately 30% of the data are missing.

where ε_{ij}'s are the marginal errors. They used the local polynomial for estimating $f(t)$ and the SCAD/LASSO penalty for selecting important variables. They adopted a two-stage approach for selecting the bandwidth and SCAD/LASSO tuning parameters. They first used a difference-based method to get a coarse estimate $\widehat{\boldsymbol{\beta}}$, and then selected the bandwidth $h = 0.5912$ using the plug-in method (Ruppert et al., 1995); with this bandwidth they tuned the SCAD/LASSO parameters using GCV and obtained the final coefficient estimates. In their analysis, they assumed a working independence correlation matrix for ε_{ij}. As a result, the SCAD and LASSO methods both selected a main effect and an interaction: PreCD4 and Smoking $\times$ Age.

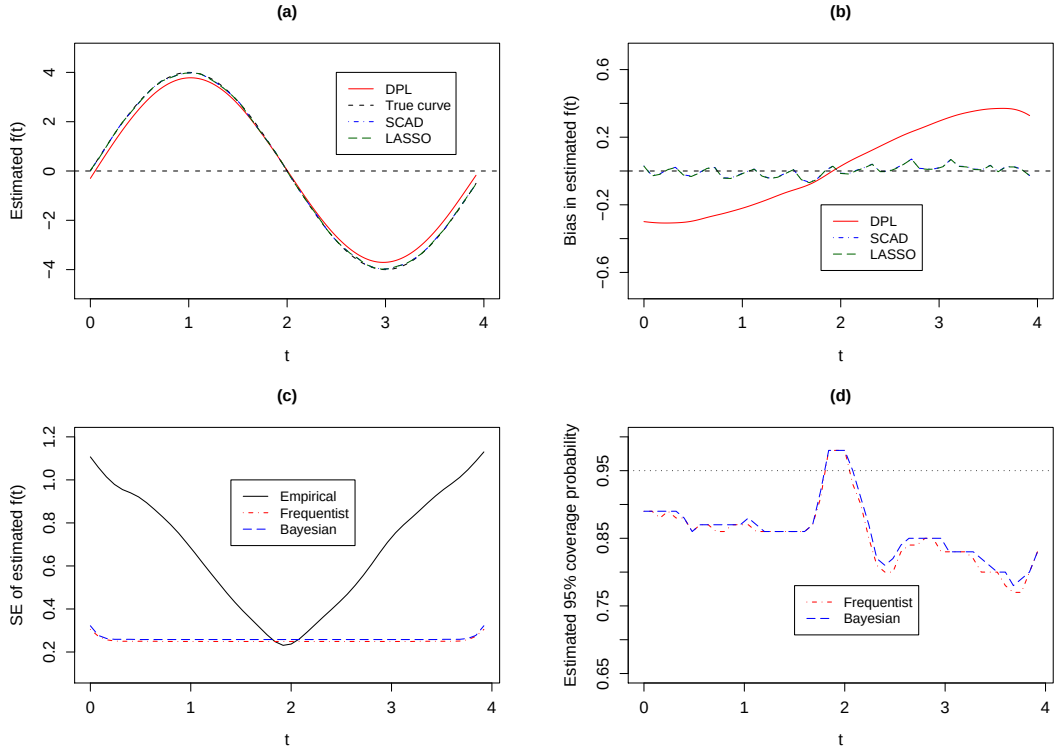

Figure 3.3: Plots for $\widehat{f}(t)$ in the mis-specified model scenario based on 100 replications. The random slope in the true model was left out in model fitting.

We applied our DPL method to this data set for model selection and estimation. We fit a semiparametric mixed model for the data which has the same fixed effects as in model (3.23), although we explicitly model the correlation structure using random coefficients and Gaussian stochastic processes. Specifically, we considered the following model:

$$y_{ij} = f(t_{ij}) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + b_i + U_i(t_{ij}) + \varepsilon_{ij}, \quad (3.24)$$

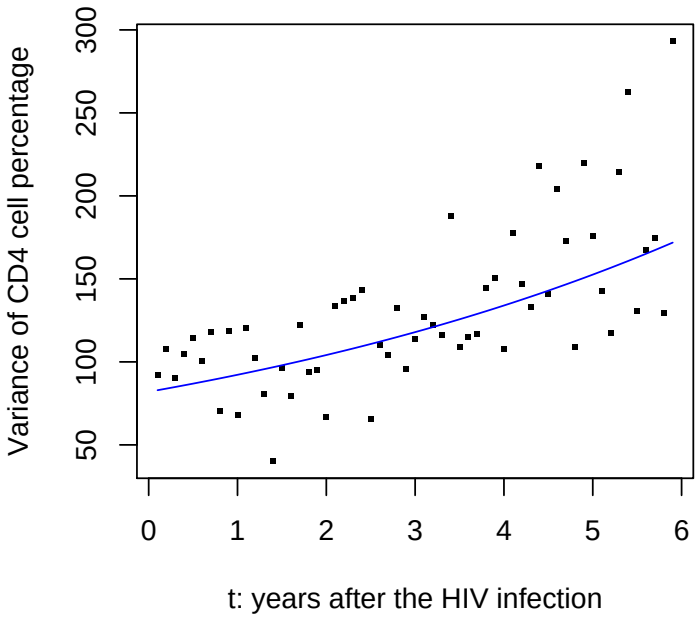

Figure 3.4: Sample variance of the CD4 cell percentage at 59 distinct knots, with the solid line being the estimated variance function from model (3.25).

where b_i's are independent random intercepts following $N(0, v_1)$, the $U_i(t)$ is mean zero Gaussian process modeling serial correlation and the ε_{ij}'s are independent measurement errors following $N(0, \sigma^2)$. The estimated baseline and residual plot in Fan and Li (2004) indicates that the variance varies (increases) over time, and this led us to use a Nonhomogeneous Ornstein-Uhlenbeck (NOU) process with $\log[var(U_i(t))] = a_0 + a_1 t$ and $corr(U_i(t), U_i(s)) = \rho^{|t-s|}$. In Figure 3.4 we also plotted the empirical marginal variances for the observed CD4 cell percentage values on the distinct time points (knots). It suggests that the variance of the CD4 cell percentage increases over time. Based on these considerations, we fit this data set using model (3.24). Table 3.5

Table 3.5: Estimated coefficients and frequentist and Bayesian SE under model (3.24) (with random intercept) for CD4 count data from the multicenter AIDS cohort study.

	Full Model	Selected Model
Variable	$\widehat{\beta}(SE_{freq}, SE_{Bayes})$	$\widehat{\beta}(SE_{freq}, SE_{Bayes})$
Smoking	0.4349 (1.1208, 1.1209)	0(0,0)
Age	0.3074 (0.7087, 0.7087)	0(0,0)
PreCD4	4.1174 (0.7112, 0.7112)	4.2052 (0.5229, 0.5229)
Age^2	0.0006 (0.3984, 0.3984)	0(0,0)
PreCD4^2	0.7465 (0.3721, 0.3721)	0(0,0)
Smoking*Age	−0.6431 (1.2176, 1.2176)	0(0,0)
Smoking*PreCD4	−0.5754 (1.2561, 1.2561)	0(0,0)
Age*PreCD4	0.4681 (0.5291, 0.5291)	0(0,0)

Table 3.6: Model variance component and tuning parameter estimation results for CD4 count data from the multicenter AIDS cohort study.

	Model variance parameters					Tuning Parameters	
Model	v_1	ρ	a_0	a_1	σ^2	τ	λ_2
(3.24)	0.00001	0.8941	3.9894	0.3289	20.8126	0.7960	0.125
(3.25)	-	0.8790	4.1560	0.1490	18.1543	1.0156	0.250

presents the estimated coefficients along with their frequentist and Bayesian standard errors for both the full and selected models. As shown in Table 3.5, PreCD4 is the only important variable selected by our DPL method. The selected model is more parsimonious and interpretable than the full model with smaller SEs for the estimated coefficients. The Bayesian and frequentist SE's are almost identical.

The parameter estimates for variance components are shown in Table 3.6. It can be seen that the estimated variance of the random intercept is 0.00001, which means that there is virtually no among-subject variation and hence the random intercept b_i in model (3.24) may be unnecessary. Therefore we refit the data without random

Table 3.7: Estimated coefficients and frequentist and Bayesian SE under model (3.25) (without random intercept) for CD4 count data from the multicenter AIDS cohort study.

Variable	Full Model $\widehat{\beta}(SE_{freq}, SE_{Bayes})$	Selected Model $\widehat{\beta}(SE_{freq}, SE_{Bayes})$
Smoking	0.5092 (1.1216, 1.1216)	0(0,0)
Age	0.2326 (0.7096, 0.7096)	0(0,0)
PreCD4	3.7058 (0.7112, 0.7112)	3.7682 (0.5188, 0.5188)
Age^2	−0.0279 (0.4002, 0.4002)	0(0,0)
PreCD4^2	0.4097 (0.3724, 0.3724)	0(0,0)
Smoking*Age	−0.7427 (1.2152, 1.2152)	0(0,0)
Smoking*PreCD4	−0.2262 (1.2579, 1.2579)	0(0,0)
Age*PreCD4	0.4172 (0.5291, 0.5291)	0(0,0)

intercept:

$$y_{ij} = f(t_{ij}) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{22} x_2^2 + \beta_{33} x_3^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + U_i(t_{ij}) + \varepsilon_{ij}. \quad (3.25)$$

The estimates for variance parameters given in Table 3.6 are very close under two models. The variable selection result remains the same, as shown in Table 3.7, except that the point estimate is slightly different. Our analysis result is consistent with Huang et al. (2002): only the PreCD4 has significant effect on the mean CD4 cell percentage. Our selected model is similar to Fan and Li (2004), although their final model also has the Smoking×Age interaction in addition to PreCD4. Their estimate (standard error) for the PreCD4 effect using SCAD penalty is 3.1993 (0.5699).

Figure 3.5 depicts the estimated baseline function $f(t)$ by using our DPL method, which is comparable to the fit in Fan and Li (2004) and indicates that missing completely at random (MCAR) missing data mechanism may be appropriate for the

missing data mechanism in this data set. It shows that the CD4 cell percentage baseline has a declining trend over time. Again it can be seen that the variance for the residuals exhibits an increase trend over time, and our NOU process properly captures this pattern. We may also consider other types of random processes for modeling serial correlation, and this will be discussed in Section 3.7.

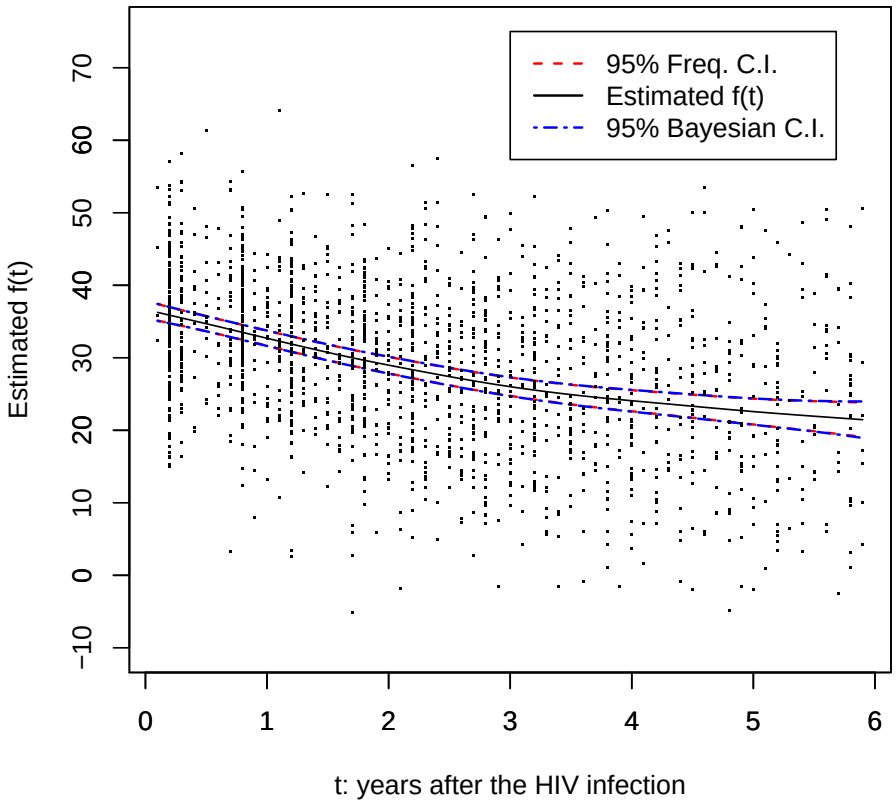

Figure 3.5: Plot of estimated baseline function with 95% frequentist and Bayesian confidence intervals. The dots are the residual, on parametric part $r = y - \widehat{\boldsymbol{\beta}}^T \boldsymbol{x}$.

3.7 Discussion

In this chapter we have proposed a new double penalized likelihood (DPL) approach for selecting important linear effects in semiparametric mixed models (SPMM; Zhang et al., 1998) for longitudinal data. The SPMM is a useful modeling technique for longitudinal data which extends the usual linear mixed models. It relaxes linear assumptions and uses a nonparametric smooth function to model the nonlinear relationships between the response and covariates. The DPL includes two penalty terms: the roughness penalty for the nonparametric part $f(t)$ and a shrinkage penalty for the fixed covariate effects $\boldsymbol{\beta}$. Maximizing the DPL leads to a parsimonious model and a smoothing spline fit for $f(t)$. We have cast the SPMM into a modified linear mixed model framework in which the smoothing parameter is treated as an extra variance component and can be conveniently estimated using REML.

Our DPL method is distinguished from that in Fan and Li (2004), which is a two-stage procedure using local polynomial regression for estimating $f(t)$ in a marginal model context. We compared the DPL approach with their SCAD and LASSO methods through simulations. Simulations show that our method has overall better performance in terms of variable selection and parameter estimation. Under correct model specification, our method is more efficient than Fan and Li's method, which ignores the correlation using working independence correlation matrix. When the data set contains missing data subject to missing at random (MAR), their approach suffered severely and led to biased estimation; whereas our DPL still performed well. However, we emphasize that the DPL has to be based on correct model specifications and the inference can be biased otherwise. It may be useful for us to conduct model di-

agnostics through exploratory analysis and compare different models via information criteria to ensure a proper model for the variance structure.

In this work we assumed Gaussian distributions and continuous responses. In the future, we also plan to extend the DPL methodology to generalized semiparametric mixed models to incorporate other types of endpoints and more likelihood structures. We hope to derive theoretical properties for the estimators. Another interesting direction is to also work on the selection of variance components automatically.

Chapter 4

Conclusion and Discussion

In this thesis we have proposed a new double-penalized least squares (likelihood) approach for variable selection in two semiparametric regression models: partial linear models for independent data and semiparametric mixed models (SPMM) for longitudinal data. Our main interest is to select important linear effects from these semiparametric models, and hence we treat the nonparametric component $f(t)$ as a nuisance effect. We use a smoothing spline to estimate the nonparametric part and impose a shrinkage penalty on $\boldsymbol{\beta}$ to achieve parsimony in linear effects. We cast these semiparametric models into a linear mixed model (LMM) framework which can be easily implemented using existing software packages. In this LMM framework, we avoid expensive two-dimensional search for tuning parameters by treating the smoothing parameter as a variance component and directly estimating it via REML. We have derived standard error formulas for the DPL estimator from both frequentist and Bayesian perspectives. Under regularity conditions, we showed in Chapter 2 that the DPLSEs are root-n consistent and have the oracle property for partial linear models. We also found in Chapter 2 that our method is robust to error distributional

assumptions in partial linear models.

We compared our method with those in Fan and Li (2004) through extensive simulations in both independent and longitudinal data settings. Simulation results show that our method has overall better performance in model selection and estimation. Since we adopt a SPMM approach for longitudinal data, a correct model specification is very important for the DPL to yield valid inference. Furthermore, our method is still consistent when data are missing at random (MAR). The marginal model approach in Fan and Li (2004) requires that the data are missing completely at random (MCAR), which is a much stronger assumption. We illustrated these points in Chapter 3.

In the future, we plan to extend the DPL methodology to generalized semiparametric mixed models to incorporate other types of endpoints and more likelihood structures. In the semiparametric models, the response variable is assumed to be linearly related to some covariates and nonlinearly related to other covariates. Identification of the nonlinear effects is based on empirical evidence such as a basic marginal plot. Therefore we want to find a systematic way to identify linear and nonlinear effects in the semiparametric models. Another future direction is to simultaneously select important variance components in addition to the linear effects in the semiparametric mixed models.

Bibliography

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60**, 255–265.

Akaike, H. (1977). On entropy maximization principle. *Application of Statistics*. P. R. Krishnaiah (eds.), Amsterdam: NorthHolland, 27–41.

Breiman, L. (1995). Better subset selection using the nonnegative garrotte. *Technometrics* **37**, 373–384.

Chen, K. and Jin, Z. Z. (2006). Partial linear regression models for clustered data. *Journal of the American Statistical Association* **101**, 195–204.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerishe Mathematik* **31**, 377–403.

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

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–451.

Engle, R., Granger, C., Rice, J. and Weiss, A. (1986). Nonparametric estimates of the raltion between weather and electricity sales. *Journal of the American Statistical Association* **81**, 310–386.

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

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics* **30**, 74–99.

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

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

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review* **55**, 245–260.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models*. London: Chapman and Hall.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association* **72**, 320–340.

Hastie, T. and Tibshirani, R. J. (1990). *Generalized additive models*. London: Chapman and Hall.

He, X., Zhu, Z. and Fung, W. K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89**, 579–590.

Heckman, N. E. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society, Ser. B* **48**, 244–248.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447.

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: rational, organization and selected characteristics of the participants. *American Journal of Epidemiology* **126**, 310–318.

Kohn, R., Ansley, C. and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameter. *Journal of the American Statistical Association* **86**, 1042–1050.

Lard, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Liang, H. (2006). Estimation in partially linear models and numerical comparisons. *Computational Statistics and Data Analysis* **50**, 675–687.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society, Ser. B* **61**, 381–400.

Linhart, H. and Zucchini, W. (1986). *Model selection*. New York: Wiley.

Mallows, C. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

Mallows, C. (1975). More comments on C_p. *Technometrics* **37**, 362–372.

Miller, A. J. (2002). *Subset selection in regression*. London: Chapman and Hall.

Rao, C. and Wu, Y. (2001). On model selection (with discussion). *Institute of Mathematical Statistical Lecture Notes* **38**, 1–64.

Rice, J. (1986). Convergence rates for partially splined models. *Statistics and Probability Letters* **4**, 203–208.

Ruppert, D., Sheather, S. and Wand, M. (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* **90**, 1257–1270.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge, New York: Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society, Ser. B* **50**, 413–436.

Speed, T. (1991). Discussion of 'BLUP is a good thing: the estimation of random effects' by G. K. Robinson. *Statistical Sciences* **6**, 15–51.

Taylor, M. G., Cumberland, W. G. and Sy, J. P. (1994). A stochastic-model for analysis of longitudinal AIDS data. *Journal of the American Statistical Association* **89**, 727–736.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B* **58**, 267–288.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society, Ser. B* **40**, 364–372.

Wahba, G. (1983). Bayesian 'confidence intervals' for the cross-validated smoothing spline. *Journal of the Royal Statistical Society, Ser. B* **45**, 133–150.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics* **13**, 1378–1402.

Wang, Y. D. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society, Ser. B* **60**, 159–174.

Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93**, 710–719.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Ser. B* **67**, 301–320.