Published in final edited form as:
J Multivar Anal. 2009 October 1; 100(9): 2100–2111. doi:10.1016/j.jmva.2009.06.009.

Automatic Model Selection for Partially Linear Models

Xiao Ni, **Hao Helen Zhang**, and **Daowen Zhang**
Department of Statistics, North Carolina State University

Xiao Ni: xni@stat.ncsu.edu; Hao Helen Zhang: hzhang2@stat.ncsu.edu; Daowen Zhang: zhang@stat.ncsu.edu

Abstract

We propose and study a unified procedure for variable selection in partially linear models. A new type of double-penalized least squares is formulated, using the smoothing spline to estimate the nonparametric part and applying a shrinkage penalty on parametric components to achieve model parsimony. Theoretically we show that, with proper choices of the smoothing and regularization parameters, the proposed procedure can be as efficient as the oracle estimator (Fan and Li, 2001). We also study the asymptotic properties of the estimator when the number of parametric effects diverges with the sample size. Frequentist and Bayesian estimates of the covariance and confidence intervals are derived for the estimators. One great advantage of this procedure is its linear mixed model (LMM) representation, which greatly facilitates its implementation by using standard statistical software. Furthermore, the LMM framework enables one to treat the smoothing parameter as a variance component and hence conveniently estimate it together with other regression coefficients. Extensive numerical studies are conducted to demonstrate the effective performance of the proposed procedure.

Keywords

Key words and phrases: Semiparametric regression; Smoothing splines; Smoothly clipped absolute deviation; Variable selection

1. Introduction

Partially 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 other additional variables are characterized by nonparametric functions. In particular, we consider a partially linear model $Y = \mathbf{X}^T \boldsymbol{\beta} + f(T) + \varepsilon$, where $\mathbf{X}$ are explanatory variables of primary interest, $\boldsymbol{\beta}$ are regression parameters, $f(\cdot)$ is an unknown smooth function of the auxiliary covariate T, and the errors are uncorrelated. This model is a special case of general additive models (Hastie and Tibshirani, 1990). Estimation of $\boldsymbol{\beta}$ and f has been studied in various contexts including kernel smoothing (Speckman, 1998), smoothing splines (Engle et al., 1986; Heckman, 1986; Wahba, 1990; Green and Silverman, 1994; Gu, 2002), and penalized splines (Ruppert et al., 2003; Liang, 2006).

Often times, the number of potential explanatory variables, d, is large, but only a subset of them are predictive to the response. Variable selection is necessary to improve prediction

Correspondence to: Hao Helen Zhang, hzhang2@stat.ncsu.edu.

accuracy and model interpretability of final models. In this paper, we treat $f(T)$ as a nuisance effect and mainly focus on automatic selection, estimation and inferences for important linear effects in the presence of T. For linear models, numerous variable selection methods have been developed such as stepwise selection, best subset selection, and shrinkage methods like nonnegative garrote (Breiman, 1995), least absolute selection and shrinkage operator (LASSO; Tibshirani, 1996), smoothly clipped absolute deviation (SCAD; Fan and Li, 2001), least angle regression (Efron et al., 2004), adaptive lasso (Zou, 2006; Zhang and Lu, 2006). Information criteria commonly used for model comparison include Mallows C_p (Mallows, 1973), Akaike's Information Criteria (Akaike, 1973) and Bayesian Information Criteria (BIC; Schwarz, 1978). A thorough review on variable selection for linear models is given in Linhart and Zucchini (1986) and Miller (2002).

Though there is a vast amount of work on variable selection for linear models, limited work has been done on model selection for partially linear models as noted in Fan and Li (2004). Model selection for partially linear models is challenging, since it consists of several interrelated estimation and selection problems: nonparametric estimation, smoothing parameter selection, and variable selection and estimation for linear covariates. Fan and Li (2004) has done some pioneering work in this area. In the framework of kernel smoothing, Fan and Li (2004) proposed an effective kernel estimator for nonparametric function estimation while using the SCAD penalty for variable selection; they were among the first to extend the shrinkage selection idea to partially linear models. Bunea (2004) proposed a class of sieve estimators based on penalized least squares for semiparametric model selection, and established the consistency property of their estimator. Bunea and Wegkamp (2004) suggested another two-stage estimation procedure and proved that the estimator is minimax adaptive under some regularity conditions. Recently, variable selection for high dimensional data, either d diverges with n or $d > n$, has been actively studied. Fan and Peng (2004) established asymptotic properties of the nonconcave penalized likelihood estimators for linear model variable selection when d increases with the sample size. Xie and Huang (2007) studied the SCAD-penalized regression for partially linear models for high dimensional data, where polynomial regression splines are employed for model estimation.

In this work, we propose a new regularization approach for model selection in the context of partially smoothing spline models and study its theoretical and computational properties. As we show in the paper, the elegant smoothing spline theory and formulation can be used to develop a simple yet effective procedure for joint function estimation and variable selection. Inspired by Fan and Li (2004), we adopt the SCAD penalty for model parsimony due to its nice theoretical properties. We will show that the new estimator has the oracle property if both smoothing and regularization parameters are chosen properly as $n \to \infty$, when the dimension d is fixed. In the more challenging case when $d_n \to \infty$ as $n \to \infty$, the estimator is shown to be $\sqrt{n/d_n}$-consistent and be able to select important variables correctly with probability tending to one. In addition to these desired asymptotic properties, the new approach also has advantages in computation and parameter estimation. It naturally owns a linear mixed model (LMM) representation, which allows one to take advantage of standard software and implement it without much extra programming effort. This LMM framework further facilitates the process of tuning multiple parameters: the smoothing parameter in the roughness penalty and the regularization parameter associated with the shrinkage penalty. In our work, the smoothing parameter is treated as an additional variance component and estimated jointly with the residual variance using the restricted maximum likelihood (REML) approach, and therefore a two-dimensional grid search can be avoided. We also show that the local quadratic approximation (LQA; Fan and Li, 2001) technique used for computation provides us a convenient and robust sandwich formula for standard errors of the resulting estimates.

The rest of the article is organized as follows. In Section 2 we propose the double penalized least squares method for joint variable selection and model estimation, and establish the asymptotic properties of the resulting estimator $\hat{\boldsymbol{\beta}}$. We further study the large-sample properties of the estimator, such as the estimation consistency and variable selection consistency, in situations when the input dimension increases with the sample size n. In Section 3 we suggest a linear mixed model (LMM) representation for the proposed procedure, which leads to an iterative algorithm with easy implementation. We also discuss how to select the tuning parameters. In Section 4, we derive the covariance estimates for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$, from both Frequentist and Bayesian perspectives. Sections 5 and 6 present simulation results and a real data application. Section 7 concludes the article with a discussion.

2. Double-Penalized Least Squares Estimators and Their Asymptotics

2.1. Double-Penalized Least Squares Estimators

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 partially 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 a 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 $J^2(f)=\int_0^1\{f''(t)\}^2dt$.

To simultaneously achieve the estimation of the nonparametric function $f(t)$ and the selection of important variables, we propose a double-penalized least squares (DPLS) approach by minimizing

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

The first penalty term in (2.2) penalizes the roughness of the nonparametric fit $f(t)$ and the second penalty $p_{\lambda_2}(|\beta_j|)$ is the shrinkage penalty on β_j 's. To our best knowledge, there has been little work on the DPLS in literature. We call the minimizer of (2.2) double-penalized least squares estimators (DPLSEs). There are two tuning parameters in (2.2): $\lambda_1\geq 0$ is a smoothing parameter which balances smoothness of $f(t)$ with fidelity to data, and $\lambda_2\geq 0$ is a regularization parameter controlling the amount of shrinkage 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 is a piecewise quadratic function and satisfies

$$p'_{\lambda_2}(\omega)=\lambda_2\left\{I(\omega\leq\lambda_2)+\frac{(a\lambda_2-\omega)_+}{(a-1)\lambda_2}I(\omega>\lambda_2)\right\}\quad\text{for }\omega>0, \tag{2.3}$$

where $a > 2$ is also a tuning parameter. Fan and Li (2001) showed that the SCAD penalty function results in consistent, sparse and continuous estimators in linear models.

2.2 Asymptotic Theory: *d* fixed

First we lay out regularity conditions on $\boldsymbol{x}_i$, 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 q nonzero components. Assume the uncorrelated random variables ε_i's have uniformly bounded absolute third moments. In addition, we assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independently and identically distributed with mean zero, finite positive definite covariance matrix $\mathbf{R}$, and that the components of $\boldsymbol{x}_i$ have finite third and fourth moments. As in Heckman (1986), we assume that $t_i's$ are distinct values in [0, 1] and satisfy $\int_0^{t_i}u(w)dw=i/n$, where $u(\cdot)$ is a continuous and strictly positive function independent of 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 partially 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

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

where $\mathbf{K}$ is the nonnegative definite smoothing matrix defined by Green and Silverman (1994). Given λ_1, λ_2, and $\boldsymbol{\beta}$, the DPLS minimizer of (2.4) is given by $\hat{\mathbf{f}}(\boldsymbol{\beta})=(\mathbf{I}+n\lambda_1\mathbf{K})^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$, where $\mathbf{A}(\lambda_1)=(\mathbf{I}+n\lambda_1\mathbf{K})^{-1}$ is equivalent to the linear smoother matrix in Craven and Wahba (1979) and Heckman (1986). Plugging $\hat{\mathbf{f}}(\boldsymbol{\beta})$ into (2.4), we obtain a penalized profile least squares only of $\boldsymbol{\beta}$:

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

We call the quadratic term in $Q(\boldsymbol{\beta})$ as the profile least squares and denote it by $L(\boldsymbol{\beta})$.

In the following, we establish the asymptotic theory for our estimator in terms of both estimation and variable selection. Proofs of these results involve the second-order Taylor expansion of $p_{\lambda_2}(|\beta|)$, and we will adapt the derivations of Fan and Li (2001) to our partially linear model context. Compared to the linear models studied in Fan and Li (2001), the major difficulty here is due to the appearance of the nonparametric component f in (2.1), which can affect the linear estimate $\boldsymbol{\beta}$ through the smoother matrix $\mathbf{A}(\lambda_1)$. In Lemma 1, we first establish some theoretical properties of $L(\boldsymbol{\beta})$, which are useful for the proofs of Lemma 2 and Theorems 1 and 2 later in this section.

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

From Lemma 1, we have $n^{-1/2}L'(\boldsymbol{\beta}_0) = O_p(1)$, $n^{-1}L''(\boldsymbol{\beta}_0) = \mathbf{R} + o_p(1)$ and $n^{-1}\frac{\partial L(\boldsymbol{\beta}_0)}{\partial \beta_j} = O_p(n^{-1/2})$, $n^{-1}\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_k} = R_{jk} + o_p(1)$, where R_{jk} is the (j, k)th element of $\mathbf{R}$. Using these results, we can prove the root-n consistency of the DPLSE $\hat{\boldsymbol{\beta}}$ and its oracle properties. Since the derivations of Theorems 1, 2, and Lemma 2 given in the following are similar to those in Fan and Li (2001), they are omitted in the paper.

Theorem 1—As $n \to \infty$, if $\lambda_{1n} \to 0$, $n\lambda_{1n}^{1/4} \to \infty$ and $\lambda_{2n} \to 0$, then there exists a local minimizer $\hat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$.

Theorem 1 says that if we choose proper sequences of λ_{1n} and λ_{2n} as $n \to \infty$, then the DPLSE $\hat{\boldsymbol{\beta}}$ is root-n consistent. In the following, we establish through Lemma 2 and Theorem 2 that $\hat{\boldsymbol{\beta}}$ can perform as well as the oracle procedure in variable selection.

Lemma 2—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 $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O(n^{-1/2})$ and any constant $C > 0$,

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

Theorem 2—As $n \to \infty$, if $\lambda_{1n} \to 0$, $n\lambda_{1n}^{1/4} \to \infty$, $\lambda_{2n} \to 0$, and $n^{1/2} \to \lambda_{2n} \to \infty$, 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: $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$.
- **b.** Asymptotic normality: $n^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \xrightarrow{d} N\{\mathbf{0}, \sigma^2 \mathbf{R}_{11}^{-1}\}$, where $\mathbf{R}_{11}$ is the q × q upper-left sub-matrix of $\mathbf{R}$.

2.3 Asymptotic Theory: $d_n \to \infty$ as $n \to \infty$

In this section, we study the sampling properties of the DPLSEs in the situation where the number of linear predictors tends to ∞ as the sample size n goes to ∞. Similar to Fan and Peng (2004), we show that under certain regularity conditions, the DPLSEs are $\sqrt{n/d_n}$-consistent and also consistent in selecting important variables, where d_n is the dimension of $\boldsymbol{\beta}$ to emphasize its dependence on the sample size n. Similarly, we re-define the number of important parametric effects as q_n. We write the true regression coefficients as $\boldsymbol{\beta}_{n0} = (\boldsymbol{\beta}_{n10}^T, \mathbf{0}^T)^T$ and the DPLSE estimator as $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n1}^T, \widehat{\boldsymbol{\beta}}_{n2}^T)^T$. For any square matrix $\mathbf{G}$, denote its smallest eigenvalue and largest eigenvalue respectively by $\Lambda_{\min}(\mathbf{G})$ and $\Lambda_{\max}(\mathbf{G})$. The following are the regularity conditions assumed to facilitate the technical derivations.

(C1) The elements $\{\beta_{n10,j}\}$'s of $\boldsymbol{\beta}_{n10}$ satisfy

$$\min\{|\beta_{n10,j}|, 1 \le j \le q_n\}/\lambda_{2n} \to \infty.$$

(C2) There exist constants c_1 and c_2 such that

$$0 < c_1 < \Lambda_{\min}(\mathbf{R}) \le \Lambda_{\max}(\mathbf{R}) < c_2 < \infty, \quad \forall n.$$

Both conditions above are adopted from Fan and Peng (2004), which is the first work to study the large-sample properties of the non-concave penalized estimators for linear models when the dimension of data diverges with the sample size *n*. As pointed out by Fan and Peng (2004), the condition (C1) gives the rate at which the penalized estimator can distinguish nonvanishing parameters from 0. Condition (C2) assumes that the **R** is positive definite and its eigenvalues are uniformly bounded.

Theorem 3—Under the conditions (C1) and (C2), as $n \to \infty$, if $\lambda_{1n} \to 0$, $n\lambda_{1n}^{1/4} \to \infty$, $\lambda_{2n} \to 0$, and $d_n = o(n^{1/2} \wedge n\lambda_{1n}^{1/4})$, then there exists a local minimizer $\hat{\boldsymbol{\beta}}_n$ of $Q(\boldsymbol{\beta}_n)$ such that $\| \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0} \| = O_p(\sqrt{d_n/n})$.

Theorem 3 says that if we choose proper sequences of λ_{1n} and λ_{2n} as $n \to \infty$, then the DPLSE $\hat{\boldsymbol{\beta}}_n$ is $\sqrt{n/d_n}$-consistent. This consistency rate is the same as the result of Fan and Peng (2004), where the number of parameters diverges in linear models. It is also the same as the result of the M-estimator studied by Huber (1973) in the diverging dimension situations. In the next, Theorem 4 shows that $\hat{\boldsymbol{\beta}}_n$ is also consistent in variable selection, i.e, unimportant linear predictors will be estimated as exactly zeros with probability tending to one. All the proofs are given in the Appendix.

Theorem 4—Under the regularity conditions (C1) and (C2), as $n \to \infty$, if $\lambda_{1n} \to 0$, $n\lambda_{1n}^{1/4} \to \infty$, $\lambda_{2n} \to 0$, $\sqrt{n/d_n}\lambda_{2n} \to \infty$, and $d_n = o(n^{1/2} \wedge n\lambda_{1n}^{1/4})$, then with probability tending to 1, the local minimizer $\widehat{\boldsymbol{\beta}}_n = (\widehat{\boldsymbol{\beta}}_{n1}^T, \widehat{\boldsymbol{\beta}}_{n2}^T)^T$ in Theorem 3 must satisfy $\hat{\boldsymbol{\beta}}_{n2} = \mathbf{0}$.

3. Computational Algorithm and Parameter Tuning

We reformulate the DPLS into a linear mixed model (LMM) representation for ease of computation. The LMM allows us to treat the smoothing parameter as a variance component and provides a unified estimation and inferential framework. An iterative algorithm is then outlined.

3.1. Linear Mixed Model (LMM) Representation

Let $\boldsymbol{t} = (t_1, \ldots, t_n)^T$ be the vector of distinct t_i's and $\mathbf{f} = (f(t_1), \ldots, f(t_n))^T$. In the case where there are ties in t_i's, an incidence matrix can be used to cast the DPLS into a linear mixed model framework as in Zhang et al. (1998). The partially linear model (2.1) can then be expressed as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{f} + \boldsymbol{\varepsilon}. \tag{3.1}$$

If ε_i's were normally distributed, then minimizing (2.4) with respect to $(\boldsymbol{\beta}, \mathbf{f})$ is equivalent to maximizing the double-penalized likelihood

$$\ell_{dp}(\boldsymbol{\beta}, \mathbf{f}; \boldsymbol{Y}) = \ell(\boldsymbol{\beta}, \mathbf{f}; \boldsymbol{Y}) - \frac{n\lambda_1}{2\sigma^2}\mathbf{f}^T\mathbf{K}\mathbf{f} - \frac{n}{\sigma^2}\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \tag{3.2}$$

where $\ell(\boldsymbol{\beta}, \mathbf{f}; \mathbf{Y}) = -(n/2)\log\sigma^2 - (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{f})/(2\sigma^2)$. Following Green (1987), we may write **f** via a one-to-one linear transformation as $\mathbf{f} = \mathbf{T}\boldsymbol{\delta} + \mathbf{B}\mathbf{a}$, where $\mathbf{T} = [\mathbf{1}, \boldsymbol{t}]$, $\mathbf{1}$ is the vector of 1's with length *n*, $\boldsymbol{\delta}$ and **a** are of length 2 and $n - 2$ respectively, and $\mathbf{B} = \mathbf{L}(\mathbf{L}^T\mathbf{L})^{-1}$

with $\mathbf{L}$ being an $n \times (n-2)$ 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

$$\ell_{dp}(\boldsymbol{\beta}, \delta, \mathbf{a}; \boldsymbol{Y}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{Y} - \mathbf{X}_*\boldsymbol{\beta}_* - \mathbf{B}\mathbf{a})^T(\boldsymbol{Y} - \mathbf{X}_*\boldsymbol{\beta}_* - \mathbf{B}\mathbf{a}) - \frac{n\lambda_1}{2\sigma^2}\mathbf{a}^T\mathbf{a} - \frac{n}{\sigma^2}\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \tag{3.3}$$

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

For fixed $\boldsymbol{\beta}_*$ (and given $\lambda_1, \lambda_2, \sigma^2$), (3.3) can be treated as the joint log-likelihood for the following linear mixed model (LMM) subject to the SCAD penalty on $\boldsymbol{\beta}$

$$\boldsymbol{Y} = \mathbf{X}_*\boldsymbol{\beta}_* + \mathbf{B}\mathbf{a} + \boldsymbol{\varepsilon}, \tag{3.4}$$

where $\boldsymbol{\beta}_*$ represent fixed effects, and $\mathbf{a}$ are random effects with $\mathbf{a} \sim N(0, \tau\boldsymbol{I})$, $\tau = 2\sigma^2/(n\lambda_1)$, and $\boldsymbol{\theta} = (\tau, \sigma^2)$ are variance components. We then conduct variable selection by maximizing the penalized log-likelihood of $\boldsymbol{\beta}_*$ subject to the SCAD penalty

$$\ell_{dp}(\boldsymbol{\beta}_*; \boldsymbol{Y}) = -\frac{1}{2}(\boldsymbol{Y} - \mathbf{X}_*\boldsymbol{\beta}_*)^T\mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{X}_*\boldsymbol{\beta}_*) - \frac{n}{\sigma^2}\sum_{j=1}^{d} p_{\lambda_2}(|\beta_j|), \tag{3.5}$$

where $\mathbf{V} = \sigma^2\mathbf{I}_n + \tau\mathbf{B}\mathbf{B}^T$ is the variance of $\mathbf{Y}$ under mixed model representation (3.4). After selecting important variables and obtaining estimates $\hat{\boldsymbol{\beta}}_*$, we can use $\hat{\boldsymbol{\delta}}$ and the best linear unbiased prediction (BLUP) estimate $\hat{\mathbf{a}}$ to construct the smoothing spline fit $\hat{f}(t)$. This LMM representation suggests 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 the maximum likelihood or restricted maximum likelihood (REML) approach during the variable selection process under the working distributional assumption that $\varepsilon_i's$ were normal. However, it should be noted that the above mixed model representation is merely a framework convenient for computation. The asymptotic results in Section 2 do not depend on the normal error assumption. Simulation results in Section 5 indicate that our procedure is quite robust to the distributional assumption for ε_i's.

The SCAD penalty function defined by (2.3) is not differentiable at the origin, causing difficulty in maximizing (3.5) with gradient based methods such as the Newton-Raphson. Following Fan and Li (2001,2004), we use a local quadratic approximation (LQA) approach. Assuming $\widetilde{\boldsymbol{\beta}}_*^0$ is an initial value close to the maximizer of (3.5), we have the following local approximation:

$$[p_{\lambda_2}(|\beta_{*j}|)]' = p'_{\lambda_2}(|\beta_{*j}|)\mathrm{sign}(\beta_{*j}) \approx \frac{p'_{\lambda_2}(|\widetilde{\beta}^0_{*j}|)}{|\widetilde{\beta}^0_{*j}|}\beta_{*j}, \quad \text{for} |\beta^0_{*j}| \geq \xi, \quad j \geq 3,$$

where ξ is a pre-specified threshold.

Using Taylor expansions, we can approximate (3.5) by

$$\ell_{dp}(\boldsymbol{\beta}_*|\widehat{\boldsymbol{\beta}}_*^0) \approx -\tfrac{1}{2}(\boldsymbol{Y}-\mathbf{X}_*\boldsymbol{\beta}_*)^T\mathbf{V}^{-1}(\boldsymbol{Y}-\mathbf{X}_*\boldsymbol{\beta}_*) - \tfrac{n}{2\sigma^2}\boldsymbol{\beta}_*^T\textstyle\sum_{\lambda_2}(\widehat{\boldsymbol{\beta}}_*^0)\boldsymbol{\beta}_* - \tfrac{n}{\sigma^2}\sum_{j=3}^{d+2}\left\{p_{\lambda_2}(|\widehat{\beta}_{*j}^0|) - \tfrac{1}{2}\tfrac{p'_{\lambda_2}(|\widehat{\beta}_{*j}^0|)}{|\widehat{\beta}_{*j}^0|}(\widehat{\beta}_{*j}^0)^2\right\}, \tag{3.6}$$

where $\sum_{\lambda_2}(\boldsymbol{\beta}_*)=\text{diag}\{0, 0, p'_{\lambda_2}(|\beta_1|)/|\beta_1|, \ldots, p'_{\lambda_2}(|\beta_d|)/|\beta_d|\}$. For fixed $\boldsymbol{\theta} = (\tau, \sigma^2)$we apply the Newton-Raphson method to maximize (3.6) and get the updating formula

$$\widehat{\boldsymbol{\beta}}_*=\left\{\mathbf{X}_*^T\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}_*+n\textstyle\sum_{\lambda_2}(\widehat{\boldsymbol{\beta}}_*^0)/\sigma^2\right\}^{-1}\mathbf{X}_*^T\mathbf{V}^{-1}(\boldsymbol{\theta})\boldsymbol{Y}. \tag{3.7}$$

It is easy to recognize that (3.7) is equivalent to an iterative ridge regression algorithm.

We propose to alternately estimate $(\boldsymbol{\beta}, \mathbf{f})$ and (τ, σ^2) iteratively. The initial values for $\boldsymbol{\beta}_*$, τ and σ^2 are obtained by the `MIXED` procedure in SAS to fit the linear mixed model (3.4) with all the covariates. We then use formula (3.7) to iteratively update $\hat{\boldsymbol{\beta}}_*$. The LMM framework allows us to treat $\tau = \sigma^2/(n\lambda_1)$ as an extra variance component based on selected important linear covariates, so that we can estimate it together with the error variance σ^2 using the restricted maximum likelihood (REML). There is rich literature on the use of REML to estimate smoothing parameters and variance components (e.g. Wahba, 1985;Speed, 1991;Zhang et al., 1998). For example, Zhang et al. (1998) estimated the smoothing parameter via REML for longitudinal data with a nonparametric baseline function and complex variance structures. The partially linear model (3.1) 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.

3.2. Choice of Tuning Parameters

Although the smoothing parameter λ_1 (or equivalently τ) is readily estimated in the LMM framework, we still need to estimate 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.

Many selection criteria, such as cross validation (CV), generalized cross validation (GCV), BIC and AIC selection can be used for parameter tuning. Wang et al. (2007) suggested using the BIC for the SCAD estimator in linear models and partially linear models, and proved its model selection consistency property, i.e. the optimal parameter chosen by the BIC can identify the true model with probability tending to one. We will also use the BIC to select the optimal 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 3. Let $\mathbf{X}_1$ be the sub-matrix of $\mathbf{X}$ for the q important variables and $\boldsymbol{\beta}_1$ be the corresponding $q \times 1$ regression coefficient vector. Then we may use the estimation method of Zhang et al. (1998) to solve the partially linear model (2.1). Consequently $\hat{\mathbf{Y}} = \mathbf{SY}$, where $\mathbf{S}$ is a smoother matrix with $q_1 = trace(\mathbf{S})$. The BIC criterion is then computed as $\text{BIC}(\lambda_2) = -2\ell + q_1 \log n$, where $\ell = -(n/2)\log(2\pi\hat{\sigma}^2) - (\mathbf{Y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \hat{\mathbf{f}})^T(\mathbf{Y}-\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \hat{\mathbf{f}})/(2\hat{\sigma}^2)$. For each grid point of λ_2, the iterative ridge regression results in a model with a set of 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.

4. Frequentist and Bayesian Covariance Estimates

We derive the frequentist and Bayesian covariance formulas for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ parallel to Sections 3.4 and 3.5 in Zhang et al. (1998), except that we also take into account the bias introduced by the imposed penalty for the variable selection. Using these covariance estimates, we are able to construct confidence intervals for the regression coefficients and the nonparametric function. The proposed covariance estimates are evaluated via simulation in Section 5.

4.1. Frequentist Covariance Estimates

From frequentists' point of view, $\text{cov}(\mathbf{Y}|t, \mathbf{x}) = \sigma^2\mathbf{I}$, and we can write $\hat{\boldsymbol{\beta}}_* = (\hat{\boldsymbol{\delta}}^T, \hat{\boldsymbol{\beta}}^T)^T$ as an approximately linear function of $\mathbf{Y}$: $\hat{\boldsymbol{\beta}}_* = \mathbf{QY}$. 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 $\hat{\boldsymbol{\delta}} = \mathbf{Q}_1\mathbf{Y}$, and $\hat{\boldsymbol{\beta}} = \mathbf{Q}_2\mathbf{Y}$. The estimated covariance matrix for $\hat{\boldsymbol{\beta}}$ is given by

$$\widehat{\text{cov}}_F(\widehat{\boldsymbol{\beta}}|t, \mathbf{x})=\mathbf{Q}_2\text{cov}(\boldsymbol{Y})\mathbf{Q}_2^T=\widehat{\sigma}^2\mathbf{Q}_2\mathbf{Q}_2^T, \tag{4.1}$$

where $\hat{\sigma}^2$ is the estimated error variance. It is easy to show that the empirical BLUP estimate of $\mathbf{a}$ is $\hat{\mathbf{a}} = \tilde{\mathbf{A}}(\mathbf{Y} - \mathbf{X}_*\boldsymbol{\beta}_*) = \mathbf{S}_a\mathbf{Y}$, 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 $\hat{\mathbf{f}} = \mathbf{T}\hat{\boldsymbol{\delta}}+ \mathbf{B}\hat{\mathbf{a}} = (\mathbf{TQ}_1 + \mathbf{BS}_a)\mathbf{Y}$ and its covariance

$$\widehat{\text{cov}}_F(\widehat{\mathbf{f}}|t, \mathbf{x})=\widehat{\sigma}^2(\mathbf{TQ}_1+\mathbf{BS}_a)(\mathbf{TQ}_1+\mathbf{BS}_a)^T. \tag{4.2}$$

4.2. Bayesian Covariance Estimates

The LMM representation in Section 3.1 and (3.3) suggests a prior for $f(t)$ of the form $\mathbf{f} = \mathbf{T}\boldsymbol{\delta} + \mathbf{Ba}$, with $\mathbf{a} \sim N(0, \tau\mathbf{I})$ and a flat prior for $\boldsymbol{\delta}$. As a prior for $\boldsymbol{\beta}$, a reasonable choice appears to be the one with kernel $exp\{-\frac{1}{2}\boldsymbol{\beta}^T\sum_{\lambda_2}\boldsymbol{\beta}\}$, where $\boldsymbol{\Sigma}_{\lambda_2}$ is a diagonal matrix defined in Section 3.1. The definition of the SCAD penalty function (2.3) implies that some diagonal elements of the matrix $\boldsymbol{\Sigma}_{\lambda_2}$ can be zero, corresponding to those coefficients with $|\beta_j| > a\lambda_2$. Assume after reordering, $\boldsymbol{\Sigma}_{\lambda_2} = \text{diag}(\mathbf{0}, \boldsymbol{\Sigma}_{22})$, where $\boldsymbol{\Sigma}_{22}$ has positive diagonal elelments. 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 $\boldsymbol{\beta}_2 \sim N(\mathbf{0}, \sum_{22}^{-1})$. The matrix $\mathbf{X}$ is partitioned into $[\mathbf{X}_1, \mathbf{X}_2]$ accordingly. Now we reformulate the mixed model (3.4) as: $\mathbf{Y} = \mathbf{T}\boldsymbol{\delta}+\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{B}_*\mathbf{a} + \boldsymbol{\varepsilon}$, or as $\mathbf{Y} = \boldsymbol{\chi}\boldsymbol{\gamma} + \mathbf{Zb} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\chi} = [\mathbf{T}, \mathbf{X}_1]$, $\boldsymbol{\gamma}=(\boldsymbol{\delta}^T, \boldsymbol{\beta}_1^T)^T$, $\mathbf{Z} = [\mathbf{X}_2, \mathbf{B}_*]$ and $\mathbf{b}=(\boldsymbol{\beta}_2^T, \mathbf{a}^T)^T$ is the new random effect distributed as $\mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with a block diagonal covariance matrix $\sum_b=\text{diag}(\sum_{22}^{-1}, \tau\mathbf{I})$. Under the reformulated linear mixed model, $\boldsymbol{\beta}$ consists of both fixed and random effects. Therefore the Bayesian covariances for $(\hat{\boldsymbol{\beta}},\hat{\mathbf{f}})$ are

$$\text{cov}_B(\widehat{\boldsymbol{\beta}})=\text{cov}\{\widehat{\boldsymbol{\beta}}_1^T, (\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2)^T\}^T, \tag{4.3}$$

$$\mathrm{cov}_B(\widehat{\mathbf{f}}) = [\mathbf{T}, \mathbf{B}]\mathrm{cov}\{\widehat{\boldsymbol{\delta}}^T, (\widehat{\mathbf{a}} - \mathbf{a})^T\}^T [\mathbf{T}, \mathbf{B}]^T. \quad (4.4)$$

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

5. Simulation Studies

We conduct Monte Carlo simulation studies to evaluate the finite sampling performance of the proposed DPLS method in terms of both model estimation and variable selection. Furthermore, we compare our procedure with the SCAD and LASSO methods proposed by Fan and Li (2004). In the following, these three methods are respectively referred to as "DPLSE", "SCAD" and "LASSO". 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). To select the SCAD and LASSO tuning parameters, we tried both BIC and GCV and found that BIC generally gave better performance, so BIC was used for tuning in the SCAD and LASSO.

We simulate the data from a partially linear model $y = \mathbf{x}^T\boldsymbol{\beta} + f(t) + \varepsilon$. Adopting the configuration in Tibshirani (1996) and Fan and Li (2001, 2004), we generate the correlated covariates $\mathbf{x} = (x_1, \ldots, x_8)^T$ from a standard normal distribution with AR(1) $\mathrm{corr}(x_i, x_j) = 0.5^{|i-j|}$, and we set the true coefficients $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. Two types of non-normal errors are used to demonstrate that the proposed normal likelihood based REML estimation is robust to the distributional assumption of errors. We compare three methods in a $2 \times 2 \times 2$ factorial experiment. There are two combinations of (f, ε): 1. $f_1(t) = 4\sin(2\pi t/4)$ with $\varepsilon_1 \sim C_0 t_6$; 2. $f_2(t) = 5\beta(t/20, 11, 5) + 4\beta(t/20, 5, 11)$ where $\beta(t, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1}$, with a mixture normal error $\varepsilon_2 \sim C_0\,(0.5N(1, 1) + 0.5N(-1, 3))$. The scale C_0 is chosen such that the error variance is $\sigma^2 = 1$ or 9. Consider two sample sizes $n = 100$ and $n = 200$. The number of observed unique time points t_i's is chosen to be 50 in all the settings.

As in Fan and Li (2004), we use the mean squares error (MSE) for $\hat{\boldsymbol{\beta}}$ and $\hat{f}$ to respectively evaluate goodness-of-fit for parametric and nonparametric estimation. They are defined as $MSE(\hat{\boldsymbol{\beta}}) = E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2)$, and $MSE(\widehat{f}) = E\left[\int_{T_1}^{T_2}\{\widehat{f}(t) - f(t)\}^2 dt\right]$. In practice, we compute $MSE(\hat{f})$ by averaging over the design knots. Under each setting, we carry out 100 Monte Carlo (MC) simulation runs and report the MC sample mean and standard deviation (given in the parentheses) for the MSEs. To evaluate 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 estimate, bias, and the 95% coverage probability of frequentist and Bayesian confidence intervals for the DPLSE.

5.1. Overall Model Selection and Estimation Results

Table 5.1 compares three variable selection procedures when $\sigma^2 = 1$. The DPLSE outperforms other methods in terms of both estimation and variable selection in all scenarios, and SCAD performs better than LASSO. Overall, the DPLSE achieves a sparser model, with both "Corr." and "Inc." closer to the oracle (5 & 0 respectively). 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.

Table 5.2 presents the results for a high variance case $\sigma^2 = 9$. We notice that, as σ^2 increases from 1 to 9, although there is a substantial amount of increase in the MSEs, the DPLSE still maintains very good performance in model selection. The MSEs of the DPLSE are consistently smaller than those of SCAD and LASSO (not reported here to save space). The incidence of incorrect zero coefficients occurs seldom for $n = 100$ and never occurs for $n = 200$.

5.2. Performance of DPLSE for Parametric Estimation

Table 5.3 presents the point estimate, relative bias, empirical standard error, model-based frequentist and Bayesian standard errors of the estimate. To save space, we only report the point estimation results for the parameters which are truly nonzero. The point estimate is the MC sample average and the empirical standard error is computed by the MC standard deviation. Relative bias is the ratio of the bias and the true value.

We report the results in four scenarios with varying n, σ^2 and $f(t)$, and those in other scenarios are similar and hence omitted. We observe that $\hat{\boldsymbol{\beta}}$ is roughly unbiased in all scenarios. Both Bayesian and frequentist SEs of $\hat{\beta}_j$'s obtained from (4.1) and (4.3) agree well with the empirical SEs; all SEs decrease as n increases or σ^2 decreases. Bayesian SEs are slightly larger than their frequentist counterparts, since they also account for bias in $\hat{\beta}_j$. The confidence intervals based on either Bayesian or frequentist SEs achieve the nominal coverage probability, indicating the accuracy of the SE formulas. Overall, the DPLSE works very well for estimating model parameters.

5.3. Performance of $\hat{f}(t)$ and Pointwise Standard Errors

In Figure 5.1 we plot the pointwise estimates and biases for estimating $f_1(t)$ and $f_2(t)$ when $n = 200$ and $\sigma^2 = 1$ for all three methods.

In plots (a) and (c), the averaged fitted curves are almost indistinguishable from the true nonparametric function, indicating small biases in $\hat{f}(t)$ for all three methods. Pointwise biases are magnified in plots (b) and (d), which show that the DPLSE overall has smaller bias than the other two methods. The SCAD and LASSO fits have slightly larger and rougher pointwise biases, which indicates under-smoothing due to a small bandwidth selected by the plug-in method. Our method is more advantageous in that it automatically estimates the smoothing parameter and controls the amount of smoothing more appropriately by treating $\tau = 1/(n\lambda_1)$ as a variance component.

Figure 5.2 depicts the pointwise standard errors and pointwise coverage probabilities of confidence intervals given by the covariance formulas (4.2) and (4.4). Here $n = 200$ and $\sigma^2 = 1$; (a) and (b) are for f_1 with t_6 errors, and (c) and (d) are for f_2 with mixture normal errors.

We note that the frequentist pointwise SEs interlace with the empirical SEs, whereas the Bayesian pointwise SEs are a little larger than the frequentist counterparts. Accordingly, as shown in plots (b) and (d), the pointwise coverage probability rates for frequentist confidence intervals are around the nominal level, whereas most of the Bayesian coverage probabilities are higher than 95%.

6. Real Data Application

We apply the proposed DPLS method to the *Ragweed Pollen Level* data, which was analyzed in Ruppert et al. (2003). The data was collected in Kalamazoo, Michigan during the 1993 ragweed season, and it consists 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

hour steady or brief but intense rain and $x_1 = 0$ otherwise; x_2 is temperature (°F); x_3 is wind speed (knots). The x-covariates are standardized first. Since the raw response is rather skewed, Ruppert et al. (2003) suggested a square root transformation $y = \sqrt{ragweed}$. Marginal plots suggest 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. Ruppert et al. (2003) fitted a semiparametric model with x_1, x_2 and x_3, whereas we add quadratic and interaction terms and consider a more complex model:

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

The tuning parameter selected by BIC is $\lambda_2 = 0.177$. Table 6.1 gives the DPLSE for the regression coefficients and their corresponding frequentist and Bayesian standard errors.

For comparison, we also fitted the full model via traditional partially splines with only roughness penalty on f. Table 6.1 shows that the final fitted model is $\hat{y} = \hat{f}(day) + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$, indicating that the linear main effect model suffices. All the estimated coefficients are positive, suggesting that the ragweed pollen level increases as each of the covariates increases. The shrinkage estimates have relatively smaller standard errors than those under the full model. Figure 6.1 depicts the estimated nonparametric function $\hat{f}(day)$ and its frequentist and Bayesian 95% pointwise 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.

7. Discussion

We propose a new regularization method for simultaneous variable selection and model estimation in partially linear models via double-penalized least squares. Under certain regularity conditions, the DPLSE $\hat{\boldsymbol{\beta}}$ is root-n consistent and has the oracle property. To facilitate computation, we reformulate the problem into a linear mixed model (LMM) framework, which allows us to estimate the smoothing parameter λ_1 as an additional variance component instead of conducting the conventional two-dimensional grid search together with the other tuning parameter λ_2. Another advantage of the LMM representation is that standard software can be used to implement the DPLS. Simulation studies show that the new method works effectively in terms of both variable selection and model estimation. We have derived both frequentist and Bayesian covariance formulas for the DPLSEs and empirical results favor the frequentist SE formulas for $f(t)$. Furthermore, our empirical results suggest that the DPLSE is robust to the distributional assumption of errors, giving strong support for its application in general situations.

In this paper, we have studied the large sample properties of the new estimators when the dimension d satisfies: (i) d fixed, or (ii) $d_n \to \infty$ as $n \to \infty$ with $d_n < n$. In future research we will investigate the properties and performance of our estimators for the more challenging situation $d \gg n$. Our major challenges will be to study how the convergence rate and asymptotic distributions of the linear components, in the presence of nuisance nonparametric components, will be affected when $d > n$. Very recently, Ravikumar et al. (2008) and Meier et al. (2008) consider the sparse estimation and function smoothing for additive models in high dimensional data settings. We will see how these works can be adapted to tackle our challenges in the future.

The proposed DPLS method assumes that the errors are uncorrelated. In future research, we will generalize it to model selection for correlated data such as longitudinal data. Another interesting problem is model selection for generalized semiparametric models, e.g. $E(Y) = g$

$\{X\beta + f(t)\}$, where g is a link function. In that case we will consider the double-penalized likelihood and investigate asymptotic properties for the resulting estimators.

Acknowledgments

The authors thank the editor, the associate editor, and two reviewers for their constructive comments and suggestions. The research of Hao Zhang was supported by in part by National Science Foundation DMS-0645293 and by National Institute of Health R01 CA085848-08. The research of Daowen Zhang was supported by National Institute of Health R01 CA85848-08.

References

Akaike H. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 1973;60:255–265.

Breiman L. Better subset selection using the nonnegative garrote. Technometrics 1995;37:373–384.

Bunea F. Consistent covariate selection and post model selection inference in semiparametric regression. Annals of Statistics 2004;32:898–927.

Bunea F, Wegkamp M. Two-stage model detection procedures in partially linear regression. The Canadian Journal of Statistics 2004;32:105–118.

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

Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Annals of Statistics 2004;32:407–451.

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

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

Fan J, Li R. New estimation and model selection procedures for semi-parametric modeling in longitudinal data analysis. Journal of American Statistical Association 2004;99:710–723.

Fan J, Peng H. Nonconcave penalized likelihood with a diverging number of parameters. Annals of Statistics 2004;32:928–961.

Green PJ. Penalized likelihood for general semi-parametric regression models. Int Statist Rev 1987;55:245–260.

Green, PJ.; Silverman, BW. Nonparametric Regression and Generalized Linear Models. Chapman and Hall; London: 1994.

Gu, C. Smoothing Spline ANOVA Models. Springer; New York: 2002.

Hastie, T.; Tibshirani, RJ. Generalized Additive Models. Chapman and Hall; London: 1990.

Heckman NE. Spline smoothing in a partly linear model. J R Statist Soc B 1986;48:244–248.

Liang H. Estimation in partially linear models and numerical comparisons. Computational Statistics and Data Analysis 2006;50:675–687. [PubMed: 20174596]

Linhart, H.; Zucchini, W. Model Selection. New York: Wiley; 1986.

Mallows C. Some comments on c_p. Technometrics 1973;15:661–675.

Meier, L.; Van de Geer, S.; Bühlmann, P. High-dimensional additive modeling. 2008. http://arxiv.org/abs/0806.4115v1

Miller, AJ. Subset Selection in Regression. Chapman and Hall; London: 2002.

Ravikumar P, Liu H, Lafferty J, Wasserman L. Spam: Sparse additive models. Advances in Neural Information Processing systems 2008;20:1202–1208.

Ruppert D, Sheather S, Wand M. An effective bandwidth selector for local least squares regression. Journal of American Statistical Association 1995;90:1257–1270.

Ruppert, D.; Wand, MP.; Carroll, RJ. Semiparametric Regression. Cambridge University Press; Cambridge, New York: 2003.

Schwarz G. Estimating the dimension of a model. Annals of Statistics 1978;6:461–464.

Speckman P. Kernel smoothing in partial linear models. J R Statist Soc B 1988;50:413–436.

Speed T. Discussion of 'blup is a good thing: the estimation of random effects' by G. K. Robinson. Statistical Science 1991;6:15–51.

Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc B 1996;58:267–288.

Wahba G. Bayesian 'confidence intervals' for the cross-validated smoothing spline. J R Statist Soc B 1983;45:133–150.

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

Wahba G. Spline Models for Observation Data. Society for Industrial and Applied Mathematics. 1990

Wang H, Li R, Tsai CL. Tuning parameter selector for SCAD. Biometrika 2007;94:553–568. [PubMed: 19343105]

Zhang D, Lin X, Raz J, Sowers M. Semiparametric stochastic mixed models for longitudinal data. Journal of American Statistical Association 1998;93:710–719.

Zhang H, Lu W. Adaptive lasso for cox's proportional hazards model. Biometrika 2007;94:691–703.

Zou H. The adaptive lasso and its oracle properties. Journal of American Statistical Association 2006;101:1418–1429.

Appendix

Proofs

Proof of Lemma 1

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

$$-L'(\boldsymbol{\beta}_0)=\mathbf{X}^T\{\mathbf{I}-\mathbf{A}(\lambda_1)\}(\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}_0), \tag{A1}$$

$$L''(\boldsymbol{\beta}_0)=\mathbf{X}^T\{\mathbf{I}-\mathbf{A}(\lambda_1)\}\mathbf{X}. \tag{A2}$$

For the partially linear model, we have $\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}_0=\mathbf{f}+\boldsymbol{\varepsilon}$. Substitution into (A1) yields

$$\begin{aligned}-n^{-1/2}L'(\boldsymbol{\beta}_0)&=n^{-1/2}\mathbf{X}^T\{\mathbf{I}-\mathbf{A}(\lambda_1)\}(\mathbf{f}+\boldsymbol{\varepsilon})\\ &=n^{-1/2}\mathbf{X}^T[\{\mathbf{I}-\mathbf{A}(\lambda_1)\}\mathbf{f}+\boldsymbol{\varepsilon}]-n^{-1/2}\mathbf{X}^T\mathbf{A}(\lambda_1)\boldsymbol{\varepsilon}.\end{aligned} \tag{A3}$$

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

$$n^{-1/2}\mathbf{X}^T[\{\mathbf{I}-\mathbf{A}(\lambda_1)\}\mathbf{f}+\boldsymbol{\varepsilon}]\xrightarrow{d}N(\mathbf{0},\sigma^2\mathbf{R}), \tag{A4}$$

$$n^{-1/2}\mathbf{X}^T\mathbf{A}(\lambda_1)\boldsymbol{\varepsilon}\xrightarrow{p}\mathbf{0}. \tag{A5}$$

Parts (a) and (b) are obtained by applying Slutsky's theorem to (A1) and (A2).

To prove Theorems 3 and 4, we need the following lemma. Its proof can be derived in the similar fashion as Lemma 1 above and Theorem 1 of Heckman (1986). To save space, we only

state the results below and omit the proof. For any vector $\mathbf{v}$, we use $[\mathbf{v}]_i$ to denote its ith component. For any matrix G, we use $[G]_{ij}$ to denote its (i, j)th element.

Lemma 3

Under the regularity conditions (C1) and (C2), if $\lambda_{1n} \to 0$, then

a. $[L'(\boldsymbol{\beta}_{n0})]_i = O_p(n^{1/2})$,

b. $[L''(\boldsymbol{\beta}_{n0})]_{ij} = nR_{ij} + O_p(n^{1/2} \vee \lambda_{1n}^{-1/4})$.

Proof of Theorem 3

Let $c_n = \sqrt{d_n/n}$. We need to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\{\inf_{\|\boldsymbol{r}\| \geq C} Q(\boldsymbol{\beta}_{n0} + c_n \boldsymbol{r}) > Q(\boldsymbol{\beta}_{n0})\} \geq 1 - \varepsilon. \tag{A6}$$

Let $\Delta_n(\boldsymbol{r}) = Q(\boldsymbol{\beta}_{n0} + c_n \boldsymbol{r}) - Q(\boldsymbol{\beta}_{n0})$. Recall that the first q_n components of $\boldsymbol{\beta}_{n0}$ are nonzero, $p_{\lambda 2n}(0) = 0$ and $p_{\lambda 2n}(\cdot)$ is nonnegative. By Taylor's expansion, we have

$$\begin{aligned}
\Delta_n(\boldsymbol{r}) &\geq L(\boldsymbol{\beta}_{n0} + c_n \boldsymbol{r}) - L(\boldsymbol{\beta}_{n0}) + n \sum_{j=1}^{q_n} \{p_{\lambda_{2n}}(|\beta_{n10,j} + c_n r_j|) - p_{\lambda_{2n}}(|\beta_{n10,j}|)\} \\
&\geq c_n \boldsymbol{r}^T L'(\boldsymbol{\beta}_{n0}) + \tfrac{1}{2} c_n^2 \boldsymbol{r}^T L''(\boldsymbol{\beta}_{n0}) \boldsymbol{r} + \sum_{j=1}^{q_n} [nc_n p'_{\lambda_{2n}}(|\beta_{n10,j}|) \operatorname{sign}(\beta_{n10,j}) r_j] \\
&\quad + \sum_{j=1}^{q_n} [nc_n^2 p''_{\lambda_{2n}}(|\beta_{n10,j}|) r_j^2 \{1 + o(1)\}] \\
&\equiv I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

By Lemma 3(a), we have

$$|I_1| = |c_n \boldsymbol{r}^T L'(\boldsymbol{\beta}_{n0})| \leq c_n \| L'(\boldsymbol{\beta}_{n0}) \| \| \boldsymbol{r} \| = O_p(c_n \sqrt{nd_n}) \| \boldsymbol{r} \| = O_p(nc_n^2) \| \boldsymbol{r} \|.$$

By Lemma 3(b), under the regularity condition (C2), we have

$$\begin{aligned}
I_2 &= \tfrac{1}{2} c_n^2 \boldsymbol{r}^T L''(\boldsymbol{\beta}_{n0}) \boldsymbol{r} = \tfrac{1}{2} nc_n^2 \{\boldsymbol{r}^T \mathbf{R} \boldsymbol{r} + O_p(d_n n^{-1/2} \vee d_n \lambda_{1n}^{-1/4})\} \\
&= \tfrac{1}{2} nc_n^2 \{\boldsymbol{r}^T \mathbf{R} \boldsymbol{r} + o_p(1) \| \boldsymbol{r} \|^2\},
\end{aligned}$$

the last equation above is due to the dimension condition $d_n = o(n^{1/2} \wedge n\lambda_{1n}^{1/4})$. With regard to I_3 and I_4, we have

$$|I_3| \leq \sum_{j=1}^{q_n} |nc_n p'_{\lambda_{2n}}(|\beta_{n10,j}|) \operatorname{sign}(\beta_{n10,j}) r_j| \leq nc_n^2 \| \boldsymbol{r} \|,$$

and

$$|I_4| = \sum_{j=1}^{q_n} nc_n^2 p''_{\lambda_{2n}}(|\beta_{n10,j}|) r_j^2 \{1+o(1)\} \le 2 \cdot \max_{1\le j\le q_n} p''_{\lambda_{2n}}(|\beta_{n10,j}|) \cdot nc_n^2 \| \boldsymbol{r}\|^2.$$

Under the condition (C1), $\max_{1\le j\le q_n} p'_{\lambda_{2n}}(|\beta_{n10,j}|)=0$ and $\max_{1\le j\le q_n} p''_{\lambda_{2n}}(|\beta_{n10,j}|)=0$ when n is large enough and $\lambda_{2n} \to 0$. So, both I_3 and I_4 are dominated by I_2. Therefore, by allowing C to be large enough, all terms I_1, I_3, I_4 are dominated by I_2, which is positive. This proves (A6) and completes the proof.

Proof of Theorem 4

Let $\gamma_n = C\sqrt{d_n/n}$. It suffices to show that as $n \to \infty$ with probability tending to 1, for any $\boldsymbol{\beta}_{n1}$ satisfying $\boldsymbol{\beta}_{n1} - \boldsymbol{\beta}_{n10} = O(\sqrt{d_n/n})$ and $j = q_n + 1, \ldots, d_n$,

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_{nj}} < 0 \text{ for } \beta_{nj} \in (-\gamma_n, 0), \tag{A7}$$

$$> 0 \text{ for } \beta_{nj} \in (0, \gamma_n). \tag{A8}$$

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

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta}_n)}{\partial \beta_{nj}} &= \frac{\partial L(\boldsymbol{\beta}_n)}{\partial \beta_{nj}} + np'_{\lambda_{2n}}(|\beta_{nj}|)\operatorname{sign}(\beta_{nj}) \\ &= \frac{\partial L(\boldsymbol{\beta}_{n0})}{\partial \beta_{nj}} + \sum_{k=1}^{d} \frac{\partial^2 L(\boldsymbol{\beta}_{n0})}{\partial \beta_{nj}\partial \beta_{nk}}(\beta_{nk} - \beta_{nk0}) + np'_{\lambda_{2n}}(|\beta_{nj}|)\operatorname{sign}(\beta_{nj}) \\ &\equiv J_1 + J_2 + J_3. \end{aligned}$$

By Lemma 3 and the regularity conditions (C1) and (C2), we have

$$J_1 = O_p(n^{1/2}) = O_p(\sqrt{nd_n}), \quad J_2 = O_p(\sqrt{nd_n}),$$

so $J_1 + J_2 = O_p(\sqrt{nd_n})$. Since $\sqrt{d_n/n}/\lambda_{2n} \to 0$, from

$$\frac{\partial Q(\boldsymbol{\beta}_n)}{\partial \beta_{nj}} = n\lambda_{2n}\left\{ -\frac{-p'_{\lambda_{2n}}(\beta_{nj})}{\lambda_{2n}} \operatorname{sign}(\beta_{nj}) + O_p\left(\sqrt{d_n/n}/\lambda_{2n}\right)\right\},$$

we can see that the sign of $\frac{\partial Q(\boldsymbol{\beta}_n)}{\partial \beta_{nj}}$ is totally determined by the sign of β_{nj}. Therefore, (A7) and (A8) hold for $j > q_n$, which leads to $\hat{\boldsymbol{\beta}}_{n2} = \mathbf{0}$. Combining with the result of Theorem 2, there is a $\sqrt{n/d_n}$-consistent local minimizer $\hat{\boldsymbol{\beta}}_n$ of $Q(\boldsymbol{\beta}_n)$, and $\hat{\boldsymbol{\beta}}_n$ has the form $(\hat{\boldsymbol{\beta}}_{n1}^T, \mathbf{0}^T)^T$. This completes the proof.

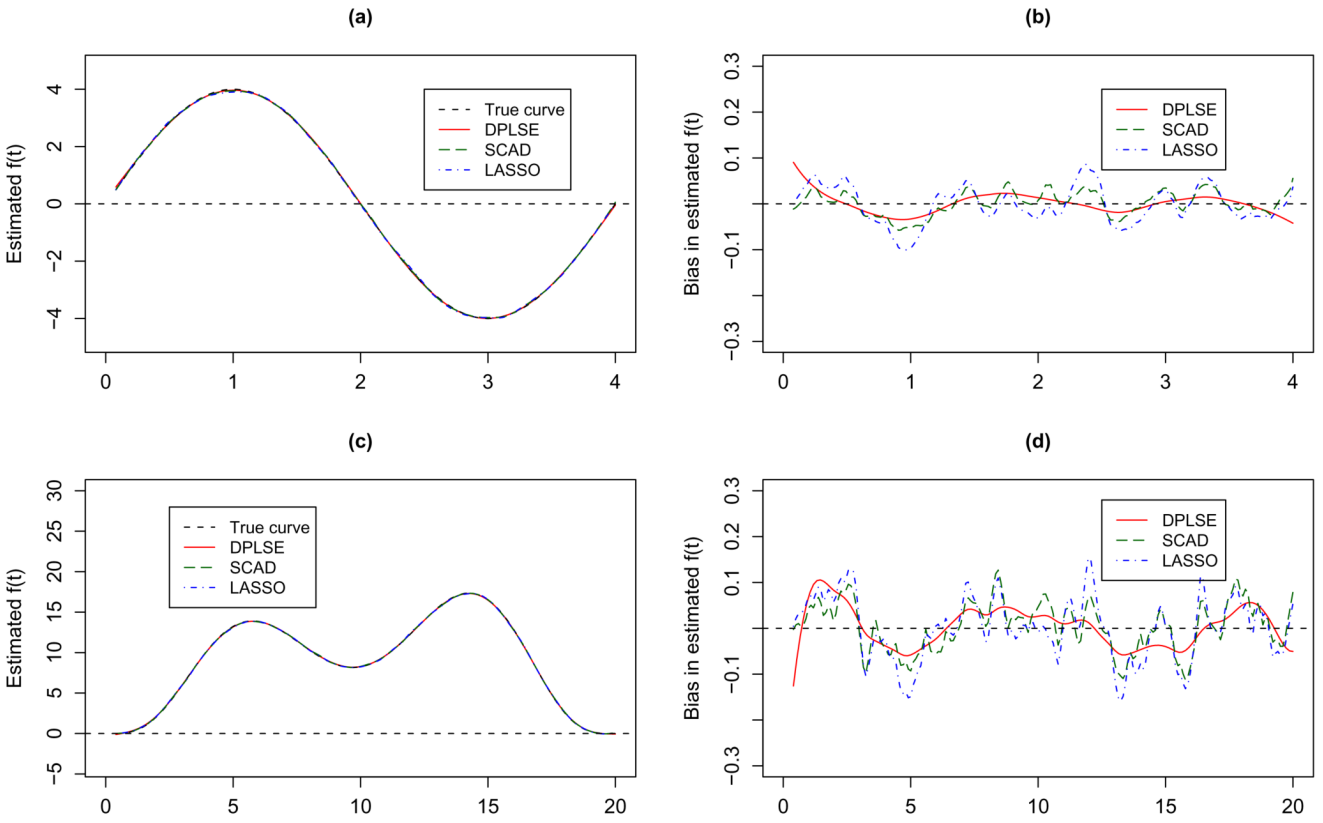

Figure 5.1.
Plots of $\hat{f}(t)$ and pointwise biases (SCAD and LASSO are based on converged MC samples). Plots (a) and (b) are for f_1; plots (c) and (d) are for f_2. Here $n = 200$ and $\sigma^2 = 1$. The horizontal axis is t in all the four plots.

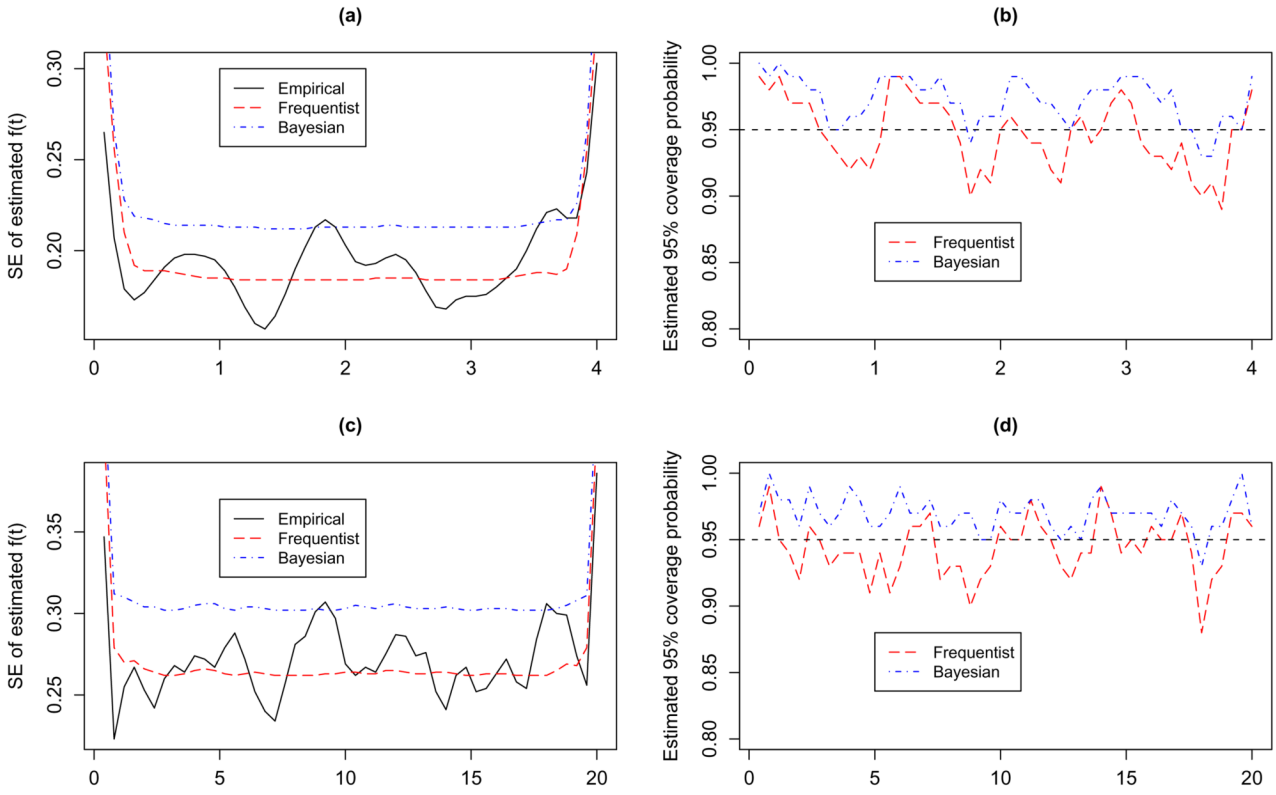

Figure 5.2.
Plots of pointwise frequentist and Bayesian standard errors and coverage probability rates. Plots (a) and (b) are for $f_1(t)$; plots (c) and (d) are for $f_2(t)$. The horizontal axis is t in all plots.

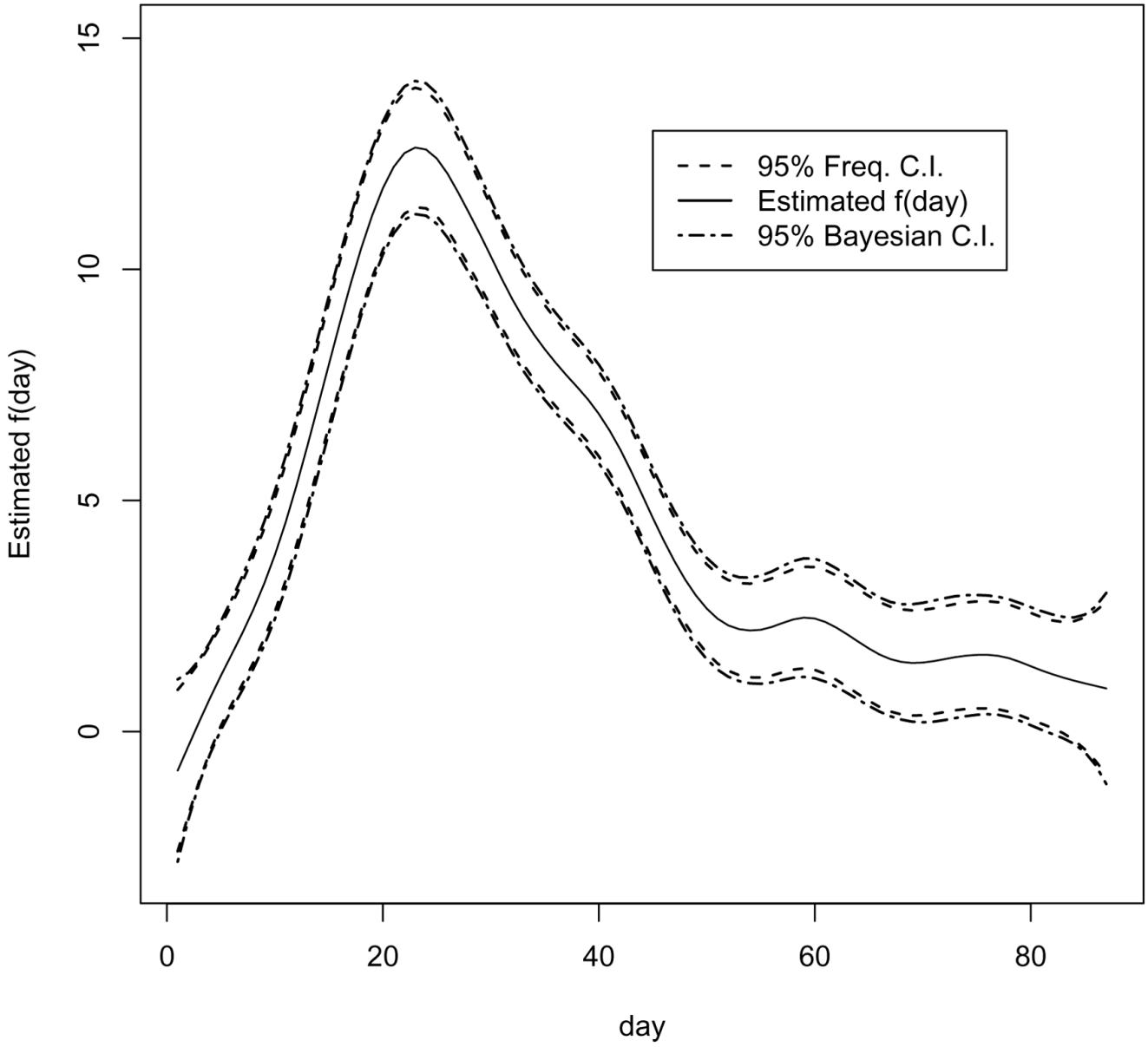

Figure 6.1.
Plot of estimated *f*(*day*) and its frequentist and Bayesian 95% pointwise confidence intervals for the Ragweed Pollen Level data.

Table 5.1

Comparison of variable selection procedures ($\sigma^2 = 1$)[a]

				Model	Zero coef.	
(*n*, *f*)	**Method**	**MSE($\hat{\beta}$)**	**MSE($\hat{f}$)**	**Size (3)**	**Corr.(5)**	**Inc.(0)**
(100, 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, 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, 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, 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

[a]SCAD and LASSO estimates are based on M converged MC samples, where $M \geq 90$ except $M = 72$ for (200, f_1).

Table 5.2

DPLSE model selection and estimation results ($\sigma^2 = 9$)

				Zero coef.	
(*n*, *f*)	**MSE($\hat{\beta}$)**	**MSE($\hat{f}$)**	**Model Size (3)**	**Corr. (5)**	**Inc. (0)**
(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 5.3

DPLSE point estimation results for four selected scenarios.

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	1.497	−0.002	0.088	0.088	0.088	0.96	0.97
	β_5	1.996	−0.002	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

Table 6.1

Estimated coefficients and frequentist and Bayesian SE for ragweed pollen level data

	Full model			Selected model		
Variable	**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	-	-