Sparse and Efficient Estimation for Partial Spline Models with Increasing Dimension

Guang Cheng* **and Hao Helen Zhang**†

Purdue University and North Carolina State University

Abstract: We consider the problem of model selection and estimation for partial spline models and propose a new regularization method in the context of smoothing splines. The regularization has a simple yet elegant form, consisting of a combination of roughness penalty on the nonparametric component and shrinkage penalty on parametric components, which can achieve function smoothing and sparse estimation simultaneously. We establish the convergence rate and oracle properties of the estimator under weak regularity conditions. One remarkable asymptotic result we discover is that, when the model is properly tuned, not only are the estimated parametric components sparse and efficient, but even more interestingly, the nonparametric component can be estimated with the optimal rate. The procedure also has attractive computational properties. Using the representer theory of smoothing splines, we can reformulate the objective function as a LASSO-type problem, which enables us to take advantage of the LARS algorithm to compute the solution path. We then extend the procedure to situations when the number of predictors increases with the sample size and investigate its asymptotic properties in that context. Finite-sampling performance of the procedure is illustrated by simulations.

Keywords and phrases: Smoothing splines, Semiparametric models, RKHS, High dimensionality, Solution path, Oracle property, Shrinkage methods.

Short title: Sparse Partial Spline

1. Introduction

1.1. Background

Partial smoothing splines are an important class of semiparametric regression models. Developed in a framework of reproducing kernel Hilbert spaces (RKHS), these models provide a compromise between linear and nonparametric models.

*Guang Cheng is Assistant Professor, Department of Statistics, Purdue University, West Lafayette, IN 47907-2066, Email: chengg@purdue.edu. Supported by NSF Grant DMS-0906497.

†Hao Helen Zhang is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, Email: hzhang2@stat.ncsu.edu. Zhang's research is partially supported by NSF grant DMS-0645293, NIH grants P01 CA142538 and R01 CA085848. The authors would like to thank the editor, the associate editor, and two reviewers for their helpful comments and suggestions which led to a much improved presentation.

In general, a partial smoothing spline model assumes the data $(\mathbf{X}_i, T_i, Y_i)$ follow

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + f(T_i) + \epsilon_i, \quad i = 1, \cdots, n, \quad f \in W_m[0,1], \tag{1.1}$$

where $\mathbf{X}_i \in R^d$ are linear covariates, $T_i \in [0,1]$ is the nonlinear covariate, and ϵ_i's are independent errors with mean zero and variance σ^2. The space $W_m[0,1]$ is the m^{th} order Sobolev Hilbert space $W_m[0,1] = \{f : f, f^{(1)}, ..., f^{(m-1)}$ are absolutely continuous, $f^{(m)} \in \mathcal{L}_2[0,1]\}$ for $m \geq 1$. Here $f^{(j)}$ denotes the jth derivative of f. The function $f(t)$ is the nonparametric component of the model. Denote the observations of $(\mathbf{X}_i, T_i, Y_i)$ as $(\mathbf{x}_i, t_i, y_i)$ for $i = 1, 2, \ldots, n$. The standard approach to compute the partial spline (PS) estimator is minimizing the penalized least squares:

$$(\widetilde{\boldsymbol{\beta}}_{PS}, \tilde{f}_{PS}) = \underset{\boldsymbol{\beta} \in R^d, f \in W_m}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - f(t_i)\right]^2 + \lambda_1 J_f^2, \tag{1.2}$$

where λ_1 is a smoothing parameter and $J_f^2 = \int_0^1 \left[f^{(m)}(t)\right]^2 dt$ is the roughness penalty on f; see [3, 4, 10, 15] for details. It is known that the solution $\tilde{f}_{PS}$ is a natural spline ([27]) of order $2m-1$ on $[0,1]$ with knots at $t_i, i = 1, \cdots, n$. Asymptotic theory for partial splines has been developed by several authors [12, 18, 21, 22]. In this paper, we mainly consider partial smoothing splines in the framework of [26].

1.2. Model Selection for Partial Splines

Variable selection is important for data analysis and model building, especially for high dimensional data, as it helps to improve the model's prediction accuracy and interpretability. For linear models, various penalization procedures have been proposed to obtain a sparse model, including the non-negative garrote [2], LASSO [24], SCAD [6, 8], and the adaptive LASSO [29, 33]. Contemporary research frequently deals with problems where the input dimension d diverges to infinity as the data sample size increases [8]. There is also active research going on for linear model selection in these situations [8, 9, 13, 14, 34].

In this paper, we propose and study a new approach to variable selection for partially linear models in the framework of smoothing splines. The procedure leads to a regularization problem in the RKHS, whose unified formulation can facilitate numerical computation and asymptotic inferences of the estimator. To conduct variable selection, we employ the adaptive LASSO penalty on linear parameters. One advantage of this procedure is its easy implementation. We show that, by using the representer theory ([27]), the optimization problem can be reformulated as a LASSO-type problem so that the entire solution path can be computed by the LARS algorithm [5]. We show that the new procedure can asymptotically (i) correctly identify the sparse model structure; (ii) estimate the nonzero β_j's consistently and achieve the semiparameric efficiency; (iii) estimate the nonparametric component f at the optimal nonparametric rate. We also investigate the property of the new procedure with a diverging number of predictors [8].

From now on, we regard $(Y_i, \mathbf{X}_i)$ as i.i.d realizations from some probability distribution. We assume that the $\mathbf{x}_i$'s belong to some compact subset in R^d, and they are standardized such that $\sum_{i=1}^n x_{ij}/n = 0$

and $\sum_{i=1}^n x_{ij}^2/n = 1$ for $j = 1, \cdots, d$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})'$. Also assume $t_i \in [0,1]$ for all i. Throughout the paper, we use the convention that $0/0 = 0$. The rest of the article is organized as follows. Section 2 introduces our new double-penalty estimation procedure for partial spline models. Section 3 is devoted to two main theoretical results. We first establish the convergence rates and oracle properties of the estimators in the standard situation with a fixed d, and then extend these results to the situations when d diverges with the sample size n. Section 4 gives the computational algorithm. In particular, we show how to compute the solution path using the LARS algorithm. The issue of parameter tuning is also discussed. Section 5 illustrates the performance of the procedure via simulations and real examples. Discussions and technical proofs are presented in Section 6 and 7.

2. Method

We assume that $0 \le t_1 < t_2 < \cdots < t_n \le 1$. In order to achieve a smooth estimate for the nonparametric component and sparse estimates for the parametric components simultaneously, we consider the following regularization problem:

$$\min_{\boldsymbol{\beta} \in R^d, f \in W_m} \frac{1}{n} \sum_{i=1}^n [y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta} - f(t_i)]^2 + \lambda_1 \int_0^1 \left[f^{(m)}(t) \right]^2 dt + \lambda_2 \sum_{j=1}^d w_j |\beta_j|. \tag{2.1}$$

The penalty term in (2.1) is naturally formed as a combination of roughness penalty on f and the weighted LASSO penalty on $\boldsymbol{\beta}$. Here, λ_1 controls the smoothness of the estimated nonlinear function while λ_2 controls the degree of shrinkage on β's. The weight w_j's are pre-specified. For convenience, we will refer to this procedure as PSA (the Partial Splines with Adaptive penalty).

Note that w_j's should be adaptively chosen such that they take large values for unimportant covariates and small values for important covariates. In particular, we propose using $w_j = 1/|\tilde{\beta}_j|^\gamma$, where $\widetilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_d)'$ is some consistent estimate for $\boldsymbol{\beta}$ in the model (1.1), and γ is a fixed positive constant. For example, the standard partial smoothing spline $\widetilde{\boldsymbol{\beta}}_{PS}$ can be used to construct the weights. Therefore, we get the following optimization problem:

$$(\widehat{\boldsymbol{\beta}}_{PSA}, \hat{f}_{PSA}) = \underset{\boldsymbol{\beta} \in R^d, f \in W_m}{\arg\min} \frac{1}{n} \sum_{i=1}^n [y_i - \mathbf{x}_i' \boldsymbol{\beta} - f(t_i)]^2 + \lambda_1 \int_0^1 \left[f^{(m)}(t) \right]^2 dt + \lambda_2 \sum_{j=1}^d \frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma}. \tag{2.2}$$

When $\boldsymbol{\beta}$ is fixed, the standard smoothing spline theory suggests that the solution to (2.2) is linear in the residual $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, i.e. $\hat{\mathbf{f}}(\boldsymbol{\beta}) = A(\lambda_1)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, where $\mathbf{y} = (y_1, \ldots, y_n)'$, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ and the matrix $A(\lambda_1)$ is the smoother or influence matrix [26]. The expression of $A(\lambda_1)$ will be given in Section 4. Plugging $\hat{\mathbf{f}}(\boldsymbol{\beta})$ into (2.2), we can obtain an equivalent objective function for $\boldsymbol{\beta}$:

$$Q(\boldsymbol{\beta}) = \frac{1}{n} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' [I - A(\lambda_1)] (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda_2 \sum_{j=1}^d \frac{|\beta_j|}{|\tilde{\beta}_j|^\gamma}, \tag{2.3}$$

where I is the identity matrix of size n. The PSA solution can be computed as:

$$\widehat{\boldsymbol{\beta}}_{PSA} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}),$$
$$\hat{f}_{PSA} = A(\lambda_1)(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{PSA}).$$

Special software like Quadratic Programming (QP) or LARS [5] is needed to obtain the solution.

3. Statistical Theory

We can write the true coefficient vector as $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0d})' = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where $\boldsymbol{\beta}_1$ consists of all q nonzero components and $\boldsymbol{\beta}_2$ consists of the rest $(d-q)$ zero elements, and write the true function of f as f_0. We also write the estimated vector $\widehat{\boldsymbol{\beta}}_{PSA} = (\hat{\beta}_1, ..., \hat{\beta}_d) = \left(\widehat{\boldsymbol{\beta}}_{PSA,1}', \widehat{\boldsymbol{\beta}}_{PSA,2}'\right)'$. In addition, assume that $\mathbf{X}_i$ has zero mean and strictly positive definite covariance matrix $\mathbf{R}$. The observations t_i's satisfy

$$\int_0^{t_i} u(w)dw = i/n \quad \text{for i=1,...,n}, \tag{3.1}$$

where $u(\cdot)$ is a continuous and strictly positive function independent of n.

3.1. Asymptotic Results for Fixed d

We show that, for any fixed $\gamma > 0$, if λ_1 and λ_2 converge to zero at proper rates, then both the parametric and nonparametric components can be estimated at their optimal rates. Moreover, our estimation procedure produces the nonparametric estimate $\hat{f}_{PSA}$ with desired smoothness, i.e. (3.5). Meanwhile, we conclude that our double penalization procedure can estimate the nonparametric function well enough to achieve the oracle properties of the weighted Lasso estimates.

In the below we use $\|\cdot\|$, $\|\cdot\|_2$ to represent the Euclidean norm, L_2- norm, and use $\|\cdot\|_n$ to denote the empirical L_2-norm, i.e. $\|F\|_n^2 = \sum_{i=1}^n F^2(s_i)/n$.

We derive our convergence rate results under the following regularity conditions:

- R1. ϵ is assumed to be independent of X, and has a sub-exponential tail, i.e. $E(\exp(|\epsilon|/C_0)) \le C_0$ for some $0 < C_0 < \infty$, see [16];
- R2. $\sum_k \phi_k\phi_k'/n$ converges to some non-singular matrix with $\phi_k = [1, t_k, \cdots, t_k^{m-1}, x_{k1}, \cdots, x_{kd}]'$ in probability.

THEOREM 3.1. *Consider the minimization problem (2.2), where $\gamma > 0$ is a fixed constant. Assume the initial estimate $\tilde{\boldsymbol{\beta}}$ is consistent. If $n^{2m/(2m+1)}\lambda_1 \to \lambda_{10} > 0$, $\sqrt{n}\lambda_2 \to 0$ and*

$$\frac{n^{\frac{2m-1}{2(2m+1)}}\lambda_2}{|\tilde{\beta}_j|^\gamma} \xrightarrow{P} \lambda_{20} > 0 \quad \textit{for } j = q+1, \ldots, d \tag{3.2}$$

as $n \to \infty$, then we have

1. *there exists a local minimizer $\widehat{\boldsymbol{\beta}}_{PSA}$ of (2.2) such that*

$$\|\widehat{\boldsymbol{\beta}}_{PSA} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2}). \tag{3.3}$$

2. *the nonparametric estimate* $\hat{f}_{PSA}$ *satisfies*

$$\|\hat{f}_{PSA} - f_0\|_n = O_P(\lambda_1^{1/2}), \tag{3.4}$$

$$J_{\hat{f}_{PSA}} = O_P(1). \tag{3.5}$$

3. *the local minimizer* $\widehat{\boldsymbol{\beta}}_{PSA} = (\widehat{\boldsymbol{\beta}}'_{PSA,1}, \widehat{\boldsymbol{\beta}}'_{PSA,2})'$ *satisfies*
   (a) *Sparsity:* $P(\widehat{\boldsymbol{\beta}}_{PSA,2} = \mathbf{0}) \to 1$.
   (b) *Asymptotic Normality:*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{PSA,1} - \boldsymbol{\beta}_1) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{R}_{11}^{-1}),$$

   where $\mathbf{R}_{11}$ *is the* $q \times q$ *upper-left sub matrix of covariance matrix of* $\mathbf{X}_i$.

REMARK 1. *Note that t is assumed to be nonrandom and satisfy the condition (3.1), and that $E\mathbf{X} = \mathbf{0}$. In this case, the semiparametric efficiency bound for $\widehat{\boldsymbol{\beta}}_{PSA,1}$ in the partly linear model under sparsity is just $\sigma^2\mathbf{R}_{11}^{-1}$, see [25]. Thus, we can claim that $\widehat{\boldsymbol{\beta}}_{PSA,1}$ is semiparametric efficient.*

If we use the partial spline solutions to construct the weights in (2.2), and choose $\gamma = 1$ and $n^{2m/(2m+1)}\lambda_i \to \lambda_{i0} > 0$ for $i = 1, 2$, the above Theorem 3.1 implies that the double penalized estimators achieve the optimal rates for both parametric and nonparametric estimation, i.e., (3.3)-(3.4), and that $\widehat{\boldsymbol{\beta}}_{PSA}$ possesses the oracle properties, i.e., the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{PSA,1}$ and sparsity of $\widehat{\boldsymbol{\beta}}_{PSA,2}$.

3.2. Asymptotic Results for Diverging d_n

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)' \in R^{q_n} \times R^{m_n} = R^{d_n}$. Let $\mathbf{x}_i = (\mathbf{w}'_i, \mathbf{z}'_i)'$ where $\mathbf{w}_i$ consists of the first q_n covariates, and $\mathbf{z}_i$ consists of the remaining m_n covariates. Thus we can define the matrix $\mathbf{X}_1 = (\mathbf{w}_1, \ldots, \mathbf{w}_n)'$ and $\mathbf{X}_2 = (\mathbf{z}_1, \ldots, \mathbf{z}_n)'$. For any matrix $\mathbf{K}$ we denote its smallest and largest eigenvalue as $\lambda_{min}(\mathbf{K})$ and $\lambda_{max}(\mathbf{K})$, respectively.

Now, we give the additional regularity conditions required to establish the large-sample theory for the increasing dimensional case:

R1D. There exist constants $0 < b_0 < b_1 < \infty$ such that

$$b_0 \le \min\{|\beta_j|, 1 \le j \le q_n\} \le \max\{|\beta_j|, 1 \le j \le q_n\} \le b_1.$$

R2D. $\lambda_{min}(\sum_k \phi_k \phi'_k/n) \ge c_3 > 0$ for any n.

R3D. Let $\mathbf{R}$ be the covariance matrix for the vector $\mathbf{X}_i$. We assume that

$$0 < c_1 \le \lambda_{min}(\mathbf{R}) \le \lambda_{max}(\mathbf{R}) \le c_2 < \infty \ \text{ for any } \ n.$$

3.2.1. Convergence Rate of $\widehat{\boldsymbol{\beta}}_{PSA}$ and $\hat{f}_{PSA}$

We first present a Lemma concerning about the convergence rate of the initial estimate $\tilde{\boldsymbol{\beta}}_{PS}$ given the increasing dimension d_n. For two deterministic sequences $p_n, q_n = o(1)$, we use the symbol $p_n \asymp q_n$ to indicate that $p_n = O(q_n)$ and $p_n^{-1} = O(q_n^{-1})$. Define $x \vee y$ ($x \wedge y$) to be the maximum (minimum) value of x and y.

Lemma 3.1. *Suppose that $\tilde{\boldsymbol{\beta}}_{PS}$ is a partial smoothing spline estimate, then we have*

$$\|\tilde{\boldsymbol{\beta}}_{PS} - \boldsymbol{\beta}_0\| = O_P(\sqrt{d_n/n}) \text{ given } d_n = n^{1/2} \wedge n\lambda_1^{1/2m}. \tag{3.6}$$

Our next theorem gives the convergence rates for $\widehat{\boldsymbol{\beta}}_{PSA}$ and $\hat{f}_{PSA}$ when dimension of $\boldsymbol{\beta}_0$ diverges to infinity. In this increasing dimension set-up, we find three results: (i) the convergence rate for $\widehat{\boldsymbol{\beta}}_{PSA}$ coincides with that for the estimator in the linear regression model with increasing dimension [17], thus we can conclude that the presence of nonparametric function and sparsity of $\boldsymbol{\beta}_0$ does not affect the overall convergence rate of $\widehat{\boldsymbol{\beta}}_{PSA}$; (ii) the convergence rate for $\hat{f}_{PSA}$ is slower than the regular partial smoothing spline, i.e. $O_P(n^{-m/(2m+1)})$, and is controlled by the dimension of important components of $\boldsymbol{\beta}$, i.e. q_n. (iii) the nonparametric estimator $\hat{f}_{PSA}$ always satisfies the desired smoothness condition, i.e. $J_{\hat{f}_{PSA}} = O_P(1)$, even under increasing dimension of $\boldsymbol{\beta}$.

Theorem 3.2. *Suppose that $d_n = o(n^{1/2} \wedge n\lambda_1^{1/2m})$, $n\lambda_1^{1/2m} \to \infty$ and $\sqrt{n/d_n}\lambda_2 \to 0$, we have*

$$\|\widehat{\boldsymbol{\beta}}_{PSA} - \boldsymbol{\beta}_0\| = O_P(\sqrt{d_n/n}). \tag{3.7}$$

If we further assume that $\lambda_1/q_n \asymp n^{-2m/(2m+1)}$ and

$$\max_{j=q_n+1,\ldots,d_n} \frac{\sqrt{n/d_n}\lambda_2}{|\tilde{\beta}_j|^\gamma} = O_P(n^{1/(2m+1)}d_n^{-1}q_n), \tag{3.8}$$

then we have

$$\|\hat{f}_{PSA} - f_0\|_n = O_P(\sqrt{d_n/n} \vee (n^{-m/(2m+1)}q_n)), \tag{3.9}$$

$$J_{\hat{f}_{PSA}} = O_P(1). \tag{3.10}$$

3.2.2. Oracle Properties

In this subsection, we show that the desired oracle properties can also be achieved even in the increasing dimension case. In particular, when showing the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{PSA,1}$, we consider an arbitrary linear combination of $\boldsymbol{\beta}_1$, say $\mathbf{G}_n\boldsymbol{\beta}_1$, where $\mathbf{G}_n$ is an arbitrary $l \times q_n$ matrix with a finite l.

Theorem 3.3. *Given the following conditions:*

D1. $d_n = o(n^{1/3} \wedge (n^{2/3}\lambda_1^{1/3m}))$ *and* $q_n = o(n^{-1}\lambda_2^{-2})$;

S1. λ_1 *satisfies:* $\lambda_1/q_n \asymp n^{-2m/(2m+1)}$ *and* $n^{m/(2m+1)}\lambda_1 \to 0$;

S2. λ_2 *satisfies:*

$$\min_{j=q_n+1,\ldots,d_n} \frac{\sqrt{n/d_n}\lambda_2}{|\tilde{\beta}_j|^\gamma} \xrightarrow{P} \infty, \tag{3.11}$$

we have

(a) Sparsity: $P(\widehat{\boldsymbol{\beta}}_{PSA,2} = \mathbf{0}) \to 1$

(b) Asymptotic Normality:

$$\sqrt{n}\mathbf{G}_n\mathbf{R}_{11}^{1/2}(\widehat{\boldsymbol{\beta}}_{PSA,1} - \boldsymbol{\beta}_1) \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{G}), \tag{3.12}$$

where $\mathbf{G}_n$ *be a non-random* $l \times q_n$ *matrix with full row rank such that* $\mathbf{G}_n\mathbf{G}_n' \to \mathbf{G}$.

In Corollary 3.1, we give the fastest possible increasing rates for the dimensions of $\boldsymbol{\beta}_0$ and its important components to guarantee the estimation efficiency and selection consistency. The range of the smoothing and shrinkage parameters are also given.

Corollary 3.1. *Let* $\gamma = 1$. *Suppose that* $\tilde{\boldsymbol{\beta}}$ *is the partial smoothing spline solution. Then, we have*

1. $\|\widehat{\boldsymbol{\beta}}_{PSA} - \boldsymbol{\beta}_0\| = O_P(\sqrt{d_n/n})$ *and* $\|\hat{f}_{PSA} - f_0\|_n = O_P(\sqrt{d_n/n} \vee (n^{-m/(2m+1)}q_n))$;
2. $\widehat{\boldsymbol{\beta}}_{PSA}$ *possesses the oracle properties.*

if the following dimension and smoothing parameter conditions hold:

$$d_n = o(n^{1/3}) \text{ and } q_n = o(n^{1/3}), \tag{3.13}$$

$$n\lambda_1^{1/2m} \to \infty, n^{m/(2m+1)}\lambda_1 \to 0 \ \text{ and } \ \lambda_1/q_n \asymp n^{-2m/(2m+1)}, \tag{3.14}$$

$$\sqrt{n/d_n}\lambda_2 \to 0, (n/d_n)\lambda_2 \to \infty \ \text{ and } \ (n/q_n)\lambda_2 = O(n^{1/(2m+1)}). \tag{3.15}$$

Define $d_n \asymp n^{\widetilde{d}}$ and $q_n \asymp n^{\widetilde{q}}$, where $0 \le \widetilde{q} \le \widetilde{d} < 1/3$ according to (3.13). For the usual case that $m \ge 2$, we can give a set of sufficient conditions for (3.14)-(3.15) as: $\lambda_1 \asymp n^{-r_1}$ and $\lambda_2 \asymp n^{-r_2}$ for $r_1 = 2m/(2m+1) - \widetilde{q}$, $r_2 = 2m/(2m+1) - \widetilde{q}$ and $r_2 < 1 - \widetilde{d}$. The above conditions are very easy to check. For example, if $m = 2$, we can set $\lambda_1, \lambda_2 \asymp n^{-0.55}$ when $d_n \asymp n^{1/4}$, $q_n \asymp n^{1/4}$.

4. Computation and Tuning

4.1. Algorithm

We propose a two-step procedure to obtain the PSA estimator: first compute $\widehat{\boldsymbol{\beta}}_{PSA}$, then compute $\hat{f}_{PSA}$. As shown in Section 2, we need to minimize (2.3) to estimate β. Define the square root matrix of $I - A(\lambda_1)$ as T, i.e. $I - A(\lambda_1) = T'T$. Then (2.3) can be reformulated into a LASSO-type problem

$$\min \frac{1}{n}(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*)'(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*) + \lambda_2 \sum_{j=1}^{d} |\beta_j^*|, \tag{4.1}$$

where the transformed variables are $\mathbf{y}^* = T\mathbf{y}$, $\mathbf{X}^* = T\mathbf{X}\mathbf{W}$, and $\beta_j^* = \beta_j/|\tilde{\beta}_j|^\gamma, j = 1, \cdots, d$, with $\mathbf{W} = \text{diag}\{|\tilde{\beta}_j|^\gamma\}$. Therefore, (4.1) can be conveniently solved with the LARS algorithm [5].

Now assume $\widehat{\boldsymbol{\beta}}_{PSA}$ has been obtained. Using the standard smoothing spline theory, it is easy to show that $\hat{\mathbf{f}}_{PSA} = A(\lambda_1)(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{PSA})$, where A is the influence matrix. By the reproducing kernel Hilbert space theory [15], $W_m[0,1]$ is an RKHS when equipped with the inner product

$$(f, g) = \sum_{\nu=0}^{m-1} \left[\int_0^1 f^{(\nu)}(t)dt\right] \left[\int_0^1 g^{(\nu)}(t)dt\right] + \int_0^1 f^{(m)} g^{(m)} dt.$$

We can decompose $W_m[0,1] = \mathcal{H}_0 \oplus \mathcal{H}_1$ as a direct sum of two RKHS subspaces. In particular, $\mathcal{H}_0 = \{f : f^{(m)} = 0\} = \text{span}\{k_\nu(t), \nu = 0, \cdots, m-1\}$, where $k_\nu(t) = B_\nu(t)/\nu!$ and $B_\nu(t)$ are Bernoulli polynomials [1]. $\mathcal{H}_1 = \{f : \int_0^1 f^{(\nu)}(t)dt = 0, \nu = 0, \cdots, m-1; f^{(m)} \in \mathcal{L}_2[0,1]\}$, associated with the reproducing kernel $K(t,s) = k_m(t)k_m(s) + (-1)^{m-1}k_{2m}([s-t])$, where $[\tau]$ is the fractional part of τ. Let S be a $n \times n$ square matrix with $s_{i,\nu} = k_{\nu-1}(t_i)$ and Σ be a square matrix with the (i,j)-th entry $K(t_i, t_j)$. Let the QR decomposition of S be $S = (F_1, F_2)\begin{pmatrix} U \\ 0 \end{pmatrix}$, where $F = [F_1, F_2]$ is orthogonal and U is upper triangular with $S'F_2 = 0$. As shown in [26] and [11], the influence matrix A can be expressed as

$$A(\lambda_1) = I - n\lambda_1 F_2(F_2' V F_2)^{-1} F_2',$$

where $V = \Sigma + n\lambda_1 I$. Using the representer theorem ([27]), we can compute the nonparametric estimator as

$$\hat{f}_{PSA}(t) = \sum_{\nu=0}^{m-1} \hat{b}_\nu k_\nu(t) + \sum_{i=1}^{n} \hat{c}_i K(t, t_i),$$

where $\widehat{\mathbf{c}} = F_2(F_2'VF_2)^{-1}F_2'\mathbf{y}$ and $\widehat{\mathbf{b}} = U^{-1}F_1'(\mathbf{y} - \Sigma\widehat{\mathbf{c}})$. We summarize the algorithm in the following:

Step 1. Fit the standard smoothing spline and construct the weights w_j's. Compute $\mathbf{y}^*$ and $\mathbf{X}^*$.

Step 2. Solve (4.1) using the LARS algorithm. Denote the solution as $\widehat{\boldsymbol{\beta}}^* = (\hat{\beta}_1^*, \cdots, \hat{\beta}_d^*)'$.

Step 3. Calculate $\widehat{\boldsymbol{\beta}}_{PSA} = (\hat{\beta}_1, \cdots, \hat{\beta}_d)'$ by $\hat{\beta}_j = \hat{\beta}_j^* |\tilde{\beta}_j|^\gamma$ for $j = 1, \cdots, d$.

Step 4. Obtain the nonparametric fit by $\widehat{\mathbf{f}} = S\widehat{\mathbf{b}} + \Sigma\widehat{\mathbf{c}}$, where the coefficients are computed as $\widehat{\mathbf{c}} = F_2(F_2'VF_2)^{-1}F_2'\mathbf{y}$ and $\widehat{\boldsymbol{\beta}} = U^{-1}F_1'(\mathbf{y} - \Sigma\widehat{\mathbf{c}})$.

4.2. Parameter Tuning

One possible tuning approach for the double penalized estimator is to choose (λ_1, λ_2) *jointly* by minimizing some scores. Following the local quadratic approximation (LQA) technique used in [24] and [6], we can derive the GCV score as a function of (λ_1, λ_2). Define the diagonal matrix $D(\boldsymbol{\beta}) = \text{diag}\{1/|\tilde{\beta}_1\beta_1|, \cdots, 1/|\tilde{\beta}_d\beta_d|\}$. The solution $\widehat{\boldsymbol{\beta}}_{PSA}$ can be approximated by

$$\left[\mathbf{X}'\{I - A(\lambda_1)\}\mathbf{X} + n\lambda_2 D(\widehat{\boldsymbol{\beta}}_{PSA})\right]^{-1} \mathbf{X}'\{I - A(\lambda_1)\}\mathbf{y} \equiv H\mathbf{y}.$$

Correspondingly, $\widehat{\mathbf{f}}_{PSA} = A(\lambda_1)(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{PSA}) = A(\lambda_1)[I - \mathbf{X}H]\mathbf{y}$. Therefore, the predicted response can be approximated as $\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}}_{PSA} + \hat{\mathbf{f}}_{PSA} = M(\lambda_1, \lambda_2)\mathbf{y}$, where

$$M(\lambda_1, \lambda_2) = \mathbf{X}H + A(\lambda_1)[I - \mathbf{X}H].$$

Therefore, the number of effective parameters in the double penalized fit $(\widehat{\boldsymbol{\beta}}_{PSA}, \hat{\mathbf{f}}_{PSA})$ may be approximated by $\mathrm{tr}\,(M(\lambda_1, \lambda_2))$. The GCV score can be constructed as

$$GCV(\lambda_1, \lambda_2) = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{[1 - n^{-1}\mathrm{tr}\,(M(\lambda_1, \lambda_2))]^2}.$$

The two-dimensional search is computationally expensive in practice. In the following, we suggest an alternative *two-stage* tuning procedure. Since λ_1 controls the partial spline fit $(\widetilde{\boldsymbol{\beta}}, \widetilde{\mathbf{b}}, \widetilde{\mathbf{c}})$, we first select λ_1 using the GCV at Step 1 of the computation algorithm:

$$GCV(\lambda_1) = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{[1 - n^{-1}\mathrm{tr}\{\tilde{A}(\lambda_1)\}]^2},$$

where $\widetilde{\mathbf{y}} = (\tilde{y}_1, \cdots, \tilde{y}_n)'$ is the partial spline prediction and $\tilde{A}(\lambda_1)$ is the influence matrix for the partial spline solution. Let $\lambda_1^* = \arg\min_{\lambda_1} \mathrm{GCV}(\lambda_1)$. We can also select λ_1^* using GCV in the smoothing spline problem: $Y_i - \mathbf{X}_i'\tilde{\boldsymbol{\beta}} = f(t_i) + \epsilon_i$, where $\tilde{\boldsymbol{\beta}}$ is the $\sqrt{n}$-consistent difference-based estimator [31]. This substitution approach is theoretically valid for selection λ_1 since the convergence rate of $\tilde{\boldsymbol{\beta}}$ is faster than the nonparametric rate for estimating f, and thus $\tilde{\boldsymbol{\beta}}$ can be treated as the true value. At the successive steps, we fix λ_1 at λ_1^* and only select λ_2 for the optimal variable selection. [28, 30, 32] suggested that the BIC works better than the GCV when tuning λ_2 for the adaptive LASSO in the context of linear models even with diverging dimension. Therefore, we propose to choose λ_2 by minimizing

$$\mathrm{BIC}(\lambda_2) = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{PSA} - \hat{\mathbf{f}}_{PSA})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{PSA} - \hat{\mathbf{f}}_{PSA})/\hat{\sigma}^2 + \log(n)\cdot r,$$

where r is the number of nonzero coefficients in $\widehat{\boldsymbol{\beta}}$, and the estimated residual variance $\hat{\sigma}^2$ can obtained from the standard partial spline model, i.e. $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_{PS} - \widetilde{\mathbf{f}}_{PS})'(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_{PS} - \widetilde{\mathbf{f}}_{PS})/(n - \mathrm{tr}(\tilde{A}(\lambda_1)) - d)$.

5. Numerical Studies

5.1. Simulation 1

We compare the standard partial smoothing spline model with the new procedure under the LASSO and adaptive (ALASSO) penalty. In the following, these three methods are respectively referred to as "PS", "PSL" and "PSA". We also include the "Oracle model" fit assuming the true model were known. In all the examples, we use $\gamma = 1$ for PSA and consider two sample sizes $n = 100$ and $n = 200$.

In each setting, a total of 500 Monte Carlo (MC) simulations are carried out. We report the MC sample mean and standard deviation (given in the parentheses) for the MSEs. Following [7], we use mean squared error $MSE(\widehat{\boldsymbol{\beta}}) = E\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ and mean integrated squared error $MISE(\hat{f}) = E\left[\int_0^1 \{\widehat{f}(t) - f(t)\}^2 dt\right]$ to evaluate goodness-of-fit for parametric and nonparametric estimation, respectively, and compute them by averaging over data knots in the simulations. To evaluate the variable selection performance of each method, we report the number of correct zero ("correct 0") coefficients, the number of coefficients incorrectly set to 0 ("incorrect 0"), model size, and the empirical probability of capturing the true model.

We generate data from a model $Y_i = \mathbf{X}_i'\boldsymbol{\beta} + f(T_i) + \varepsilon_i$, and consider two following model settings:

- Model 1: $\boldsymbol{\beta} = (3, 2.5, 2, 1.5, 0, \cdots, 0)'$, $d = 15$ and $q = 4$. And $f_1(t) = 1.5\sin(2\pi t)$.
- Model 2: $\boldsymbol{\beta} = (3, \cdots, 3, 0, \cdots, 0)'$, $d = 20$ and $q = 10$. The nonparametric function $f_2(t) = t^{10}(1-t)^4/(3B(11,5)) + 4t^4(1-t)^{10}/(15B(5,11))$, where the beta function $B(u,v) = \int_0^1 t^{u-1}(1-t)^{v-1}dt$.

Two possible distributions for the covariates X and T:

- Model 1: $X_1, \cdots, X_{15}, T$ are i.i.d. generated from Unif$(0,1)$.
- Model 2: $\mathbf{X} = (X_1, \cdots, X_{20})'$ are standard normal with AR(1) correlation, i.e. corr$(X_i, X_j) = \rho^{|i-j|}$. T follows Unif$(0,1)$ and is independent with X_i's. We consider $\rho = 0.3$ and $\rho = 0.6$.

Two possible error distributions are used in these two settings:

- Model 1: (normal error) $\epsilon_1 \sim N(0, \sigma^2)$, with $\sigma = 0.5$ and $\sigma = 1$, respectively.
- Model 2: (non-normal error) $\epsilon_2 \sim t_{10}$.

Table 1 compares the model fitting and variable selection performance of various procedures in different settings for Model 1. It is evident that the PSA procedure outperforms both the PS and PSL in terms of both the MSE and variable selection. The three procedures give similar performance in estimating the nonparametric function.

TABLE 1
Variable selection and fitting results for Model 1

σ	n	Method	MSE($\widehat{\boldsymbol{\beta}}_{PSA}$)	MISE($\hat{f}_{PSA}$)	Size	Number of Zeros	
						correct 0	incorrect 0
0.5	100	PS	0.578 (0.010)	0.015 (0.000)	15 (0)	0 (0)	0 (0)
		PSL	0.316 (0.008)	0.015 (0.000)	7.34 (0.09)	7.66 (0.09)	0.00 (0.00)
		PSA	0.234 (0.008)	0.014 (0.000)	4.53 (0.04)	10.47 (0.04)	0.00 (0.00)
		Oracle	0.129 (0.004)	0.014 (0.001)	4 (0)	11 (0)	0 (0)
	200	PS	0.249 (0.004)	0.008 (0.000)	15 (0)	0 (0)	0 (0)
		PSL	0.147 (0.004)	0.008 (0.000)	7.16 (0.09)	7.84 (0.09)	0.00 (0.00)
		PSA	0.111 (0.004)	0.008 (0.000)	4.36 (0.04)	10.64 (0.04)	0.00 (0.00)
		Oracle	0.063 (0.000)	0.007 (0.000)	4 (0)	11 (0)	0 (0)
1	100	PS	2.293 (0.040)	0.055 (0.002)	15 (0)	0 (0)	0 (0)
		PSL	1.256 (0.032)	0.051 (0.002)	7.36 (0.09)	7.64 (0.09)	0.00 (0.00)
		PSA	1.110 (0.036)	0.051 (0.002)	4.72 (0.05)	10.25 (0.05)	0.02 (0.00)
		Oracle	0.511 (0.017)	0.048 (0.002)	4 (0)	11 (0)	0 (0)
	200	PS	0.989 (0.017)	0.028 (0.001)	15 (0)	0 (0)	0 (0)
		PSL	0.587 (0.016)	0.027 (0.001)	7.20 (0.09)	7.80 (0.09)	0.00 (0.00)
		PSA	0.479 (0.014)	0.026 (0.001)	4.42 (0.04)	10.58 (0.04)	0.00 (0.00)
		Oracle	0.252 (0.008)	0.026 (0.001)	4 (0)	11 (0)	0

TABLE 2
Variable selection frequency in percentage over 500 runs for Model 1

σ	n		important index		unimportant variable index											P(correct)
			$1-3$	4	5	6	7	8	9	10	11	12	13	14	15	
0.5	100	PSL	500	500	150	147	159	164	151	124	159	147	158	165	148	0.09
		PSA	500	500	25	23	34	28	30	15	22	16	23	29	21	0.70
	200	PSL	500	500	144	145	131	156	156	142	124	152	141	143	148	0.10
		PSA	500	500	13	13	21	20	14	13	16	19	17	16	18	0.78
1	100	PSL	500	500	147	149	157	164	154	128	162	143	159	165	153	0.08
		PSA	500	489	41	35	45	38	41	19	27	27	34	40	26	0.55
	200	PSL	500	500	146	146	137	158	154	146	122	154	143	146	146	0.10
		PSA	500	500	14	16	27	25	18	15	16	20	21	17	20	0.75

Table 2 shows that the PSA works much better in distinguishing important variables from unimportant variables than PSL. For example, when $\sigma = 0.5$, the PSA identifies the correct model 64 times out of 100 times when $n = 100$ and 81 times when $n = 200$, while the PSL identifies the correct model only 8 times when $n = 100$ and 10 times when $n = 200$.

To present the performance of our nonparametric estimation procedure, we plot the estimated functions for Model 1 in the below Figure 1. The top row of Figure 1 depicts the typical estimated curves corresponding to the 10th best, the 50th best (median), and the 90th best according to MISE among 100 simulations when $n = 200$ and $\sigma = 0.5$. It can be seen that the fitted curves are overall able to capture the shape of the true function very well. In order to describe the sampling variability of the estimated nonparametric function at each point, we also depict a 95% pointwise confidence interval for f in the bottom row of Figure 1. The upper and lower bound of the confidence interval are respectively given by the 2.5th and 97.5th percentiles of the estimated function at each grid point among 100 simulations. The results show that the function f is estimated with very good accuracy.

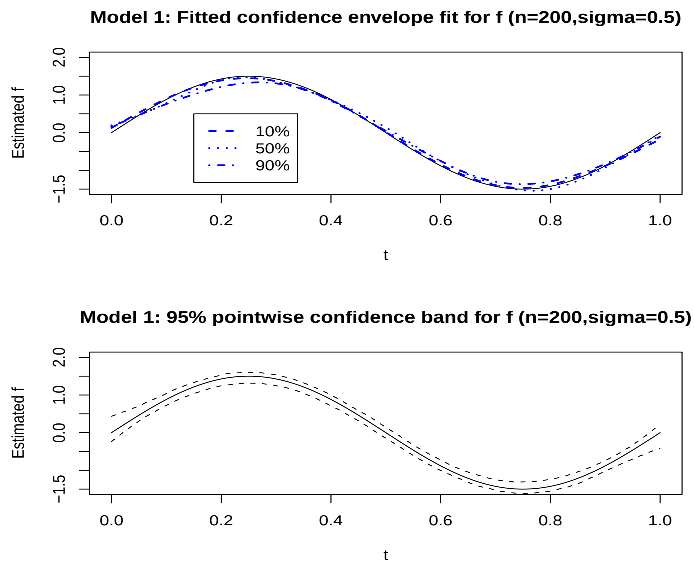

FIG 1. *The estimated nonlinear functions given by the PSA in Model 1.*

The estimated nonlinear function, confidence envelop and 95% point-wise confidence interval for Model 2 with $n = 200$ and $\sigma = 0.5$. In the top plot, the dashed line is for the 10th best fit, the dotted line is for the 50th best fit, and the dashed-dotted line is for the 90th best among 500 simulations. The bottom plot is a 95% pointwise confidence interval.

Table 3 compares the model fitting and variable selection performance in the correlated setting Model 2. The case $\rho = 0.3$ represents a weakl correlation among X's and $\rho = 0.6$ represents a moderate situation. Again, we observe that the PSA performs best in terms of both MSE and variable selection in all settings. In particular, when $n = 200$, the PSA is very close to the "Oracle" results in this example.

TABLE 3
Model selection and fitting results for Model 2

ρ	n	Method	MSE($\widehat{\boldsymbol{\beta}}_{PSA}$))	MISE($\hat{f}_{PSA}$)	Size	Number of Zeros	
						correct 0	incorrect 0
0.3	100	PS	0.416 (0.008)	0.451 (0.002)	20 (0)	0 (0)	0 (0)
		PSL	0.299 (0.006)	0.447 (0.002)	12.99 (0.09)	7.01 (0.09)	0.00 (0.00)
		PSA	0.204 (0.005)	0.443 (0.002)	10.29 (0.03)	9.71 (0.03)	0.00 (0.00)
		Oracle	0.181 (0.004)	0.444 (0.002)	10 (0)	10 (0)	0 (0)
	200	PS	0.179 (0.003)	0.408 (0.001)	20 (0)	0 (0)	0 (0)
		PSL	0.125 (0.003)	0.406 (0.001)	13.22 (0.08)	6.78 (0.8)	0.00 (0.00)
		PSA	0.087 (0.002)	0.404 (0.001)	10.12 (0.02)	9.88 (0.02)	0.00 (0.00)
		Oracle	0.082 (0.002)	0.405 (0.001)	10 (0)	10 (0)	0 (0)
0.6	100	PS	0.721 (0.013)	0.448 (0.002)	20 (0)	0 (0)	0 (0)
		PSL	0.401 (0.009)	0.440 (0.002)	11.91 (0.07)	8.09 (0.07)	0.00 (0.00)
		PSA	0.349 (0.008)	0.438 (0.002)	10.22 (0.03)	9.78 (0.03)	0.00 (0.00)
		Oracle	0.310 (0.004)	0.439 (0.002)	10 (0)	10 (0)	0 (0)
	200	PS	0.311 (0.005)	0.408 (0.001)	20 (0)	0 (0)	0 (0)
		PSL	0.170 (0.004)	0.405 (0.001)	12.47 (0.07)	7.53 (0.07)	0.00 (0.00)
		PSA	0.147 (0.004)	0.404 (0.001)	10.12 (0.02)	9.88 (0.02)	0.00 (0.00)
		Oracle	0.139 (0.004)	0.405 (0.001)	10 (0)	10 (0)	0 (0)

Table 4 compares the variable selection results of PSL and PSA in four scenarios if the covariates are correlated. Since neither of the methods misses any important variable over 100 runs, we only report the selection frequency for unimportant variables. Overall, the PSA results in a more sparse model and identifies the true model with a much higher frequency. For example, when $n = 100$ and the correlation is moderate with $\rho = 0.6$, the PSA identify the correct model 88 times out of 100 runs while the PSL identifies the correct model only 17 times.

TABLE 4
Frequency of variables selected in 100 runs for Model 2.

ρ	n	Method	unimportant variable index										P(correct)
			11	12	13	14	15	16	17	18	19	20	
0.3	100	PSL	161	152	148	145	146	155	149	154	144	141	0.09
		PSA	14	12	18	13	17	17	14	16	9	16	0.80
	200	PSL	175	143	152	169	149	171	174	159	158	161	0.03
		PSA	11	4	5	7	3	4	11	5	6	5	0.91
0.6	100	PSL	142	85	91	90	82	931	82	91	98	101	0.20
		PSA	14	15	11	9	11	9	7	14	8	10	0.87
	200	PSL	163	110	119	128	104	110	131	114	114	142	0.06
		PSA	11	4	2	6	5	8	8	4	7	5	0.91

The top row of Figure 2 depicts the typical estimated functions corresponding to the 10th best, the 50th best (median), and the 90th best fits according to MISE among 100 simulations when $n = 200$ and $\rho = 0.3$. It is evident that the estimated curves are able to capture the shape of the true function very well. The bottom row of Figure 2 depicts a 95% pointwise confidence interval for f. The results show that the function f is estimated with reasonably good accuracy.

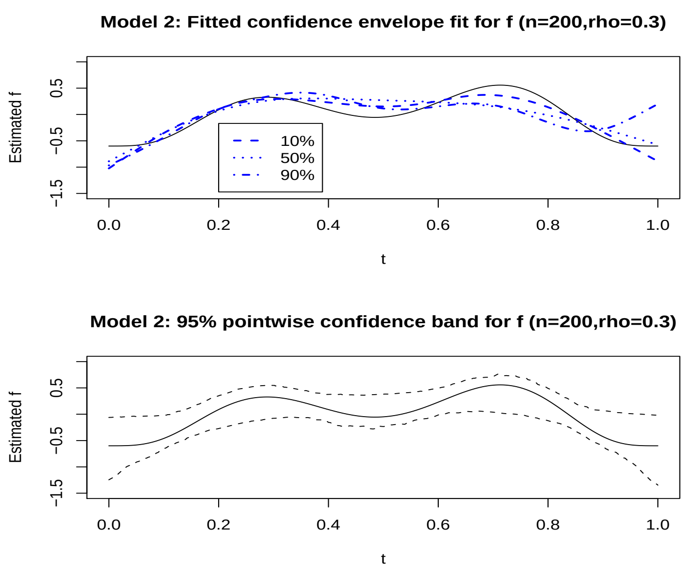

FIG 2. *The estimated nonlinear functions given by the PSA in Model 2.*
The estimated nonlinear function, confidence envelop and 95% point-wise confidence interval for Model 2, $n = 200$ and $\rho = 0.3$. In the top plot, the dashed line is 10th best fit, the dotted line is 50th best fit, and the dashed-dotted line is 90th best of 500 simulations. The bottom plot is a 95% pointwise confidence interval.

5.2. Simulation 2: Large dimensional setting

We consider an example involves with a larger number of linear variables:

- Model 3: The true coefficient $\boldsymbol{\beta} = (4,4,4,4,4,3,3,3,3,3,2,2,2,2,2,0,\cdots,0)'$, $d = 60, q = 15$, and $f_1(t) = 1.5\sin(2\pi t)$. The correlated covariates $(X_1,\cdots,X_{60})'$ are generated from marginally standard normal with AR(1) correlation with $\rho = 0.5$. Consider two settings for the normal error $\epsilon_1 \sim N(0,\sigma^2)$, with $\sigma = 0.5$ and $\sigma = 1$, respectively.

TABLE 5
Variable selection and fitting results for Model 3

σ	n	Method	MSE($\widehat{\boldsymbol{\beta}}_{PSA}$)	MISE($\hat{f}_{PSA}$)	Size	Number of Zeros	
						correct 0	incorrect 0
0.5	100	PS	1.046 (0.027)	0.130 (0.008)	60 (0)	0 (0)	0 (0)
		PSL	0.210 (0.008)	0.040 (0.002)	22.94 (0.28)	37.06 (0.28)	0.00 (0.00)
		PSA	0.180 (0.010)	0.036 (0.002)	17.50 (0.25)	42.50 (0.25)	0.00 (0.00)
		Oracle	0.078 (0.001)	0.016 (0.000)	15 (0)	45 (0)	0 (0)
	200	PS	0.186 (0.002)	0.010 (0.000)	60 (0)	0 (0)	0 (0)
		PSL	0.054 (0.001)	0.008 (0.000)	18.24 (0.15)	41.76 (0.15)	0.00 (0.00)
		PSA	0.035 (0.000)	0.007 (0.000)	15.15 (0.03)	44.85 (0.03)	0.00 (0.00)
		Oracle	0.034 (0.000)	0.007 (0.000)	15 (0)	45 (0)	0 (0)
1	100	PS	3.838 (0.058)	0.438 (0.005)	60 (0)	0 (0)	0 (0)
		PSL	0.695 (0.013)	0.104 (0.003)	22.23 (0.19)	37.77 (0.19)	0.00 (0.00)
		PSA	0.549 (0.012)	0.086 (0.002)	16.35 (0.10)	43.65 (0.10)	0.00 (0.00)
		Oracle	0.306 (0.006)	0.056 (0.002)	15 (0)	45 (0)	0 (0)
	200	PS	0.739 (0.008)	0.035 (0.001)	60 (0)	0 (0)	0 (0)
		PSL	0.188 (0.003)	0.028 (0.001)	23.93 (0.11)	36.07 (0.11)	0.00 (0.00)
		PSA	0.145 (0.003)	0.026 (0.001)	15.18 (0.03)	44.82 (0.03)	0.00 (0.00)
		Oracle	0.134 (0.003)	0.026 (0.001)	15 (0)	45 (0)	0

Table 5 compares the model fitting and variable selection performance of various procedures in different settings for Model 3. It is evident that the PSA procedure outperforms both the PS and PSL in terms of both the MSE and variable selection. The three procedures give similar performance in estimating the nonparametric function.

5.3. Real Example 1: Ragwweed Pollen Data

We apply the proposed method to the *Ragweed Pollen* data analyzed in [19]. The data consists of 87 daily observations of ragweed pollen level and relevant information collected in Kalamazoo, Michigan during the 1993 ragweed season. The main purpose of this analysis is to develop an accurate model for forecasting daily ragweed pollen level based on some climate factors. The raw response *ragweed* is the daily ragweed pollen level (grains/m^3). There are four explanatory variables:

X_1 = rain: the indicator of significant rain for the following day (1 = at least 3 hours of steady or brief but intense rain, 0 = otherwise);

X_2 = temperature: temperature of the following day (^oF);

X_3 = wind: wind speed forecast for the following day (knots);

X_4 = day: the number of days in the current ragweed pollen season.

We first standardize X-covariates. Since the raw response is rather skewed, [19] suggested a square root transformation $Y = \sqrt{ragweed}$. Marginal plots suggest a strong nonlinear relationship between Y and the *day* number. Consequently, a partial linear model with a nonparametric baseline $f(day)$ is reasonable. [19] fitted a semiparametric model with three linear effects X_1, X_2 and X_3 and a nonlinear effect of X_4. For the variable selection purpose, we add the quadratic terms in the model and fit an enlarged model:

$$y = f(\text{day}) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_2^2 + \beta_4 x_3^2 + \varepsilon.$$

Table 5 gives the estimated regression coefficients. We observe that PSL and PSA end up with the same model, and all the estimated coefficients are positive, suggesting that the ragweed pollen level increases as any covariate increases. The shrinkage in parametric terms from the partial spline models resulted from the PSA procedure is overall smaller than that resulted from the PSL procedure.

TABLE 6
Estimated Coefficients for Ragweed Pollen Data

Covariate	PS	PSL	PSA
rain	1.3834	1.3620	1.3816
temperature	0.1053	0.1045	0.1053
wind	0.2407	0.2384	0.2409
temp×temp	0.0042	0.0041	0.0041
wind×wind	- 0.0004	0	0

Figure 3 depicts the estimated nonparametric function $\widehat{f}(\text{day})$ and its 95% pointwise confidence intervals given by the PSA. The plot indicates that $\widehat{f}(day)$ increases rapidly to the peak on around day 25, plunges until day 60, and decreases steadily thereafter. The nonparametric fits given by the other two procedures are similar and hence omitted in the paper.

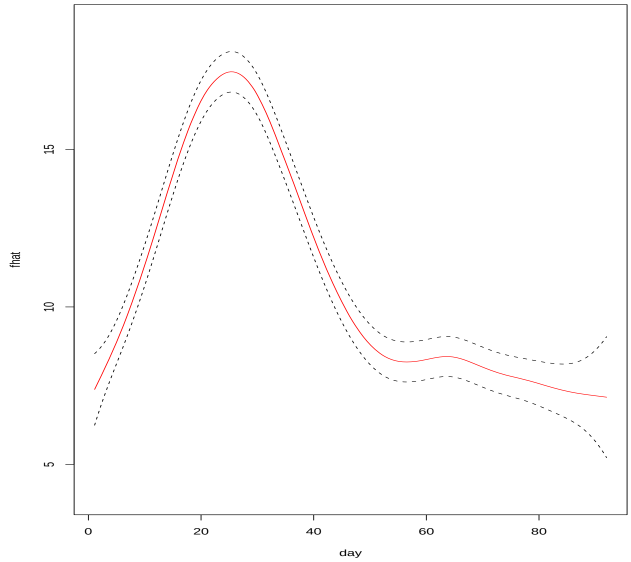

FIG 3. *The estimated nonlinear function $\hat{f}(day)$ for the ragweed pollen data.*

The estimated nonlinear function $\widehat{f}(\text{day})$ with its 95% pointwise confidence interval (dotted lines) given by the PSA for the ragweed pollen data.

5.4. Real Example 2: Prostate Cancer Data

We analyze the *Prostate Cancer* data ([23]). The goal is to predict the log level of prostate specific antigen using a number of clinical measures. The data consists of 97 men who were about to receive a radical prostatectomy. There are eight predictors: X_1 = log cancer volume (lcavol), X_2 = log prostate weight (lweight), X_3 = age, X_4 = log of benign prostatic hyperplasia amount (lbph), X_5 = seminal vesicle invasion (svi), X_6 = log of capsular penetration (lcp), X_7 = Gleason score (gleason), and X_8 = percent of Gleason scores of 4 or 5 (pgg45).

TABLE 7
Estimated Coefficients for Prostate Cancer Data

Covariate	PS	PSL	PSA
lweight	0.587	0.443	0.562
age	-0.020	0	0
lbph	0.107	0	0
svi	0.766	0.346	0.498
lcp	-0.105	0	0
gleason	0.045	0	0
pgg45	0.005	0	0

A variable selection analysis was conducted in [24] using a linear regression model with LASSO, and it selected three important variables *lcavol, lweight, svi* as important variables to predict the prostate specific antigen. We fitted partially linear models by treating *lweight* as a nonlinear term. Table 7 gives estimated coefficients for different methods. Both PSL and PSA select *lcavol* and *svi* as important linear variables, which is consistent to the analysis by [24].

6. Discussion

We propose a new regularization method for simultaneous variable selection for linear terms and component estimation for the nonlinear term in partial spline models. The oracle properties of the new procedure for variable selection are established. Moreover, we have shown that the new estimator can achieve the optimal convergence rates for both the parametric and nonparametric components. All the above conclusions are also proven to hold in the increasing dimensional situation.

The proposed method sets up a basic framework to implement variable selection for partial spline models, and it can be generalized to other types of data analysis. In our future research, we will generalize the results in this paper to the generalized semiparametric models, robust linear regression, or survival data analysis. In this paper, we assume the errors are i.i.d. with constant variance. In practice, the problem of heteroscedastic error, i.e. the variance of ϵ is some non-constant function of (X, T), is often encountered. We will also examine the properties and performance of our approach in that situation.

7. Proofs

For simplicity, we use $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\beta}}_1$ ($\widehat{\boldsymbol{\beta}}_2$) and $\hat{f}$ to represent $\widehat{\boldsymbol{\beta}}_{PSA}$, $\widehat{\boldsymbol{\beta}}_{PSA,1}$ ($\widehat{\boldsymbol{\beta}}_{PSA,2}$) and $\hat{f}_{PSA}$, in the proofs.

Definition: Let $\mathcal{A}$ be a subset of a (pseudo-) metric space $(\mathcal{L}, d)$ of real-valued functions. The δ-covering number $N(\delta, \mathcal{A}, d)$ of $\mathcal{A}$ is the smallest N for which there exist functions $a_1, \ldots, a_N$ in $\mathcal{L}$, such that for each $a \in \mathcal{A}$, $d(a, a_j) \leq \delta$ for some $j \in \{1, \ldots, N\}$. The δ-bracketing number $N_B(\delta, \mathcal{A}, d)$ is the smallest N for which there exist pairs of functions $\{[a_j^L, a_j^U]\}_{j=1}^N \subset \mathcal{L}$, with $d(a_j^L, a_j^U) \leq \delta$, $j = 1, \ldots, N$, such that for each $a \in \mathcal{A}$ there is a $j \in \{1, \ldots, N\}$ such that $a_j^L \leq a \leq a_j^U$. The δ-entropy number (δ-bracketing entropy number) is defined as $H(\delta, \mathcal{A}, d) = \log N(\delta, \mathcal{A}, d)$ ($H_B(\delta, \mathcal{A}, d) = \log N_B(\delta, \mathcal{A}, d)$).

Entropy Calculations: For each $0 < C < \infty$ and $\delta > 0$, we have

$$H_B(\delta, \{\eta : \|\eta\|_\infty \leq C, J_\eta \leq C\}, \|\cdot\|_\infty) \leq M \left(\frac{C}{\delta}\right)^{1/k}, \tag{7.1}$$

$$H(\delta, \{\eta : \|\eta\|_\infty \leq C, J_\eta \leq C\}, \|\cdot\|_\infty) \leq M \left(\frac{C}{\delta}\right)^{1/k}, \tag{7.2}$$

where $\|\cdot\|_\infty$ represents the uniform norm and M is some positive number.

Proof of Theorem 3.1: In the proof of (3.3), we will first show for any given $\epsilon > 0$, there exists a large constant M such that

$$P\left\{\inf_{\|\mathbf{s}\|=M} \Delta(\mathbf{s}) > 0\right\} \geq 1 - \epsilon, \tag{7.3}$$

where $\Delta(\mathbf{s}) \equiv Q(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{s}) - Q(\boldsymbol{\beta}_0)$. This implies with probability at least $(1 - \epsilon)$ that there exists a local minimum in the ball $\{\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{s} : \|\mathbf{s}\| \leq M\}$. Thus, we can conclude that there exists a local minimizer such that $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$ if (6.3) holds. Denote the quadratic part of $Q(\boldsymbol{\beta})$ as $L(\boldsymbol{\beta})$, i.e.,

$$L(\boldsymbol{\beta}) = \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'[I - A(\lambda_1)](\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Then we can obtain the below inequality:

$$\Delta(\mathbf{s}) \geq L(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{s}) - L(\boldsymbol{\beta}_0) + \lambda_2 \sum_{j=1}^{q} \frac{|\beta_{0j} + n^{-1/2}s_j| - |\beta_{0j}|}{|\tilde{\beta}_j|^\gamma},$$

where s_j is the j-th element of vector $\mathbf{s}$. Note that $L(\boldsymbol{\beta})$ is a quadratic function of $\boldsymbol{\beta}$. Hence, by the Taylor expansion of $L(\boldsymbol{\beta})$, we can show that

$$\Delta(\mathbf{s}) \geq n^{-1/2}\mathbf{s}'\dot{L}(\boldsymbol{\beta}_0) + \frac{1}{2}\mathbf{s}'[n^{-1}\ddot{L}(\boldsymbol{\beta}_0)]\mathbf{s} + \lambda_2 \sum_{j=1}^{q} \frac{|\beta_{0j} + n^{-1/2}s_j| - |\beta_{0j}|}{|\tilde{\beta}_j|^\gamma}, \tag{7.4}$$

where $\dot{L}(\boldsymbol{\beta}_0)$ and $\ddot{L}(\boldsymbol{\beta}_0)$ are the first and second derivative of $L(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_0$, respectively. Based on (2.3), we know that $-\dot{L}(\boldsymbol{\beta}_0) = (2/n)\mathbf{X}'[I - A(\lambda_1)](\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)$ and $\ddot{L}(\boldsymbol{\beta}_0) = (2/n)\mathbf{X}'[I - A(\lambda_1)]\mathbf{X}$. Combing the proof of Theorem 1 and its four propositions in [12], we can show that

$$\begin{aligned} n^{-1/2}\mathbf{X}'[I - A(\lambda_1)](f_0 + \epsilon) &\xrightarrow{d} N(0, \sigma^2\mathbf{R}), \\ n^{-1/2}\mathbf{X}'A(\lambda_1)\epsilon &\xrightarrow{P} 0. \end{aligned}$$

provided that $\lambda_1 \to 0$ and $n\lambda_1^{1/2m} \to \infty$. Therefore, the Slutsky's theorem implies that

$$\dot{L}(\boldsymbol{\beta}_0) = O_P(n^{-1/2}), \tag{7.5}$$

$$\ddot{L}(\boldsymbol{\beta}_0) = O_P(1) \tag{7.6}$$

given the above conditions on λ_1. Based on (6.5) and (6.6), we know the first two terms in the right hand side of (6.4) are of the same order, i.e. $O_P(n^{-1})$. And the second term, which converges to some positive constant, dominates the first one by choosing sufficiently large M. The third term is bounded by $n^{-1/2}\lambda_2 M_0$ for some positive constant M_0 since $\tilde{\beta}_j$ is the consistent estimate for the nonzero coefficient for $j = 1, \ldots, q$. Considering that $\sqrt{n}\lambda_2 \to 0$, we have completed the proof of (3.3).

We next show the convergence rate for $\hat{f}$ in terms of $\|\cdot\|_n$-norm, i.e. (3.4). Let $g_0(x,t) = \mathbf{x}'\boldsymbol{\beta}_0 + f_0(t)$, and $\hat{g}(x,t) = \mathbf{x}'\hat{\boldsymbol{\beta}} + \hat{f}(t)$. Then, by the definition of $(\hat{\boldsymbol{\beta}}, \hat{f})$, we have

$$\|\hat{g} - g_0\|_n^2 + \lambda_1 J_{\hat{f}}^2 + \lambda_2 J_{\hat{\boldsymbol{\beta}}} \leq \frac{2}{n}\sum_{i=1}^{n} \epsilon_i(\hat{g} - g_0)(X_i, t_i) + \lambda_1 J_{f_0}^2 + \lambda_2 J_{\boldsymbol{\beta}_0}, \tag{7.7}$$

$$\|\hat{g} - g_0\|_n^2 \leq 2\|\epsilon\|_n\|\hat{g} - g_0\|_n + \lambda_1 J_{f_0}^2 + \lambda_2 J_{\boldsymbol{\beta}_0},$$

$$\|\hat{g} - g_0\|_n^2 \leq \|\hat{g} - g_0\|_n O_P(1) + o_P(1),$$

where $J_\beta \equiv \sum_{j=1}^{d} |\beta_j|/|\tilde{\beta}_j|^\gamma$. The second inequality follows from the Cauchy-Schwartz inequality. The last inequality holds since ϵ has sub-exponential tail, and $\lambda_1, \lambda_2 \to 0$. Then the above inequality implies that $\|\hat{g} - g_0\|_n = O_P(1)$, so that $\|\hat{g}\|_n = O_P(1)$. By Sobolev embedding theorem, we can decompose $g(x,t)$ as

$g_1(x,t)+g_2(x,t)$, where $g_1(x,t)=x'\boldsymbol{\beta}+\sum_{j=1}^m \alpha_j t^{j-1}$ and $g_2(x,t)=f_2(t)$ with $\|g_2(x,t)\|_\infty\leq J_{g_2}=J_f$. Similarly, we can write $\hat g=\hat g_1+\hat g_2$, where $\hat g_1=x'\hat{\boldsymbol{\beta}}+\sum_{j=1}^m\hat\alpha_j t^{j-1}=\hat\delta'\phi$ and $\|\hat g_2\|_\infty\leq J_{\hat g}$. We shall now show that $\|\hat g\|_\infty/(1+J_{\hat g})=O_P(1)$ via the above Sobolev decomposition. Then

$$\frac{\|\hat g_1\|_n}{1+J_{\hat g}}\leq\frac{\|\hat g\|_n}{1+J_{\hat g}}+\frac{\|\hat g_2\|_n}{1+J_{\hat g}}=O_P(1). \tag{7.8}$$

Based on the assumption about $\sum_k\phi_k\phi_k'/n$, (6.8) implies that $\|\hat\delta\|/(1+J_{\hat g})=O_P(1)$. Since (X,t) is in a bounded set, $\|\hat g_1\|_\infty/(1+J_{\hat g})=O_P(1)$. So we have proved that $\|\hat g\|_\infty/(1+J_{\hat g})=O_P(1)$. Thus, the entropy calculation (6.1) implies that

$$H_B\left(\delta,\left\{\frac{g-g_0}{1+J_g}:g\in\mathcal{G},\frac{\|g\|_\infty}{1+J_g}\leq C\right\},\|\cdot\|_\infty\right)\leq M_1\delta^{-1/m},$$

where M_1 is some positive constant, and $\mathcal{G}=\{g(x,t)=x'\boldsymbol{\beta}+f(t):\boldsymbol{\beta}\in R^d,J_f<\infty\}$. Based on Theorem 2.2 in [16] about the continuity modulus of the empirical processes $\{\sum_{i=1}^n\epsilon_i(g-g_0)(z_i)\}$ indexed by g and (6.7), we can establish the following set of inequalities:

$$\begin{aligned}\lambda_1J_{\hat f}^2\leq&\left[\|\hat g-g_0\|_n^{1-1/2m}(1+J_{\hat f})^{1/2m}\vee(1+J_{\hat f})n^{-\frac{2m-1}{2(2m+1)}}\right]O_P(n^{-1/2})\\&+\lambda_1J_{f_0}^2+\lambda_2(J_{\boldsymbol{\beta}_0}-J_{\hat{\boldsymbol{\beta}}}),\end{aligned} \tag{7.9}$$

and

$$\begin{aligned}\|\hat g-g_0\|_n^2\leq&\left[\|\hat g-g_0\|_n^{1-1/2m}(1+J_{\hat f})^{1/2m}\vee(1+J_{\hat f})n^{-\frac{2m-1}{2(2m+1)}}\right]O_P(n^{-1/2})\\&+\lambda_1J_{f_0}^2+\lambda_2(J_{\boldsymbol{\beta}_0}-J_{\hat{\boldsymbol{\beta}}}).\end{aligned} \tag{7.10}$$

Note that

$$\begin{aligned}\lambda_2(J_{\boldsymbol{\beta}_0}-J_{\hat{\boldsymbol{\beta}}})&\leq\lambda_2\sum_{j=1}^q\frac{|\beta_{0j}-\hat\beta_j|}{|\tilde\beta_j|^\gamma}+\lambda_2\sum_{j=q+1}^d\frac{|\beta_{0j}-\hat\beta_j|}{|\tilde\beta_j|^\gamma}\\&\leq O_P(n^{-2m/(2m+1)}).\end{aligned} \tag{7.11}$$

(6.11) in the above follows from $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_P(n^{-1/2})$ and (3.2). Thus, solving the above two inequalities gives $\|\hat g-g_0\|_n=O_P(\lambda_1^{1/2})$ and $J_{\hat f}=O_P(1)$ when $n^{2m/(2m+1)}\lambda_1\to\lambda_{10}>0$. Note that

$$\|X'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_n=\sqrt{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'(\sum_{i=1}^nX_iX_i'/n)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)}\lesssim\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_P(n^{-1/2})$$

by (3.3). Applying the triangle inequality to $\|\hat g-g_0\|_n=O_P(\lambda_1^{1/2})$, we have proved that $\|\hat f-f_0\|_n=O_P(\lambda_1^{1/2})$.

We next prove 3(a). It suffices to show that

$$Q\{(\bar{\boldsymbol{\beta}}_1,\mathbf{0})\}=\min_{\|\bar{\boldsymbol{\beta}}_2\|\leq Cn^{-1/2}}Q\{(\bar{\boldsymbol{\beta}}_1,\bar{\boldsymbol{\beta}}_2)\}\ \text{ with probability approaching to 1} \tag{7.12}$$

for any $\bar{\boldsymbol{\beta}}_1$ satisfying $\|\bar{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\| = O_P(n^{-1/2})$ based on (3.3). To show (6.12), we need to show that $\partial Q(\boldsymbol{\beta})/\partial\beta_j < 0$ for $\beta_j \in (-Cn^{-1/2}, 0)$, and $\partial Q(\boldsymbol{\beta})/\partial\beta_j > 0$ for $\beta_j \in (0, Cn^{-1/2})$, for $j = q+1, \ldots, d$, holds with probability tending to 1. By two term Taylor expansion of $L(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_0$, $\partial Q(\boldsymbol{\beta})/\partial\beta_j$ can be expressed in the following form for $j = q+1, \ldots, d$:

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j} = \frac{\partial L(\boldsymbol{\beta}_0)}{\partial\beta_j} + \sum_{k=1}^{d} \frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_k}(\beta_k - \beta_{0k}) + \lambda_2 \frac{1 \times sgn(\beta_j)}{|\tilde{\beta}_j|^\gamma},$$

where β_k is the k^{th} element of vector $\boldsymbol{\beta}$. Note that $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2})$ by the above constructions. Hence , we have

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j} = O_P(n^{-1/2}) + sgn(\beta_j)\frac{\lambda_2}{|\tilde{\beta}_j|^\gamma}$$

by (6.5) and (6.6) in the above. The assumption (3.2) implies that $\sqrt{n}\lambda_2/|\tilde{\beta}_j|^\gamma \to \infty$ for $j = q+1, \ldots, d$. Thus, the sign of β_j determines that of $\partial Q(\boldsymbol{\beta})/\partial\beta_j$ for $j = q+1, \ldots, d$. This completes the proof of 3(a).

Now we prove 3(b). Following similar proof of (3.3), we can show that there exists a $\sqrt{n}$ consistent local minimizer of $Q(\boldsymbol{\beta}_1, 0)$, i.e. $\hat{\boldsymbol{\beta}}_1$, and satisfies:

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j}|_{\boldsymbol{\beta}=(\hat{\boldsymbol{\beta}}_1, \mathbf{0})} = 0$$

for $j = 1, \ldots, q$. By similar analysis in the above, we can establish the equation:

$$0 = \frac{\partial L(\boldsymbol{\beta}_0)}{\partial\beta_j} + \sum_{k=1}^{q} \left\{\frac{\partial^2 L(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_k}\right\}(\hat{\beta}_k - \beta_{0k}) + \lambda_2 \frac{1 \times sgn(\hat{\beta}_j)}{|\tilde{\beta}_j|^\gamma},$$

for $j = 1, \ldots, q$. Note that the assumption $\sqrt{n}\lambda_2 \to 0$ implies that the third term in the right hand side of the above equation is $o_P(n^{-1/2})$. By the form of $L(\boldsymbol{\beta})$ and the Slutsky's theorem, we conclude the proof of 3(b). □

Important Lemmas. We provide three useful matrix inequalities and two lemmas for preparing the proofs of Theorems 3.2 and 3.3. Given any $n \times m$ matrix $\mathbf{A}$ and symmetric strictly positive definite matrix $\mathbf{B}$, $n \times 1$ vector $\mathbf{s}$ and $\mathbf{z}$, and $m \times 1$ vector $\mathbf{w}$, we have

$$|\mathbf{s}'\mathbf{A}\mathbf{w}| \leq \|\mathbf{s}\|\|\mathbf{A}\|\|\mathbf{w}\| \tag{7.13}$$

$$|\mathbf{s}'\mathbf{B}\mathbf{z}| \leq |\mathbf{s}'\mathbf{B}\mathbf{s}|^{1/2}|\mathbf{z}'\mathbf{B}\mathbf{z}|^{1/2} \tag{7.14}$$

$$|\mathbf{s}'\mathbf{z}| \leq \|\mathbf{s}\|\|\mathbf{z}\| \tag{7.15}$$

where $\|\mathbf{A}\|^2 = \sum_j \sum_i a_{ij}^2$. (6.14) follows from the Cauchy-Schwartz inequality.

LEMMA 7.1. *Given that $\lambda_1 \to 0$, we have*

$$n^{-k/2}\sum_{l=1}^{n} |[(I - A)\mathbf{f}_0(t)]_l|^k = O(\lambda_1^{k/2}) \qquad \textit{for } k = 2, 3, \ldots \tag{7.16}$$

Proof: For the case of $k=2$, it has been proved in Lemma 2 of [12]. Next we apply the principle of mathematical induction to prove the cases for arbitrary $k>2$. We first assume that

$$n^{-(k-1)/2}\sum_{l=1}^{n}|[(I-A)\mathbf{f}_0(t)]_l|^{k-1}=O(\lambda_1^{(k-1)/2}) \tag{7.17}$$

for $k=3$. Then we can write

$$\begin{aligned}
& n^{-k/2}\sum_{l=1}^{n}|[(I-A)\mathbf{f}_0(t)]_l|^{k} \\
\leq\ & n^{-1/2}\max_{l=1,\ldots,n}|[(I-A)\mathbf{f}_0(t)]_l|\times n^{-(k-1)/2}\sum_{l=1}^{n}|[(I-A)\mathbf{f}_0(t)]_l|^{k-1} \\
\leq\ & n^{-1/2}\left[\sum_{l=1}^{n}[(I-A)\mathbf{f}_0(t)]_l^2\right]^{1/2}\times O(\lambda_1^{(k-1)/2})=O(\lambda_1^{k/2})
\end{aligned}$$

The last step follows from (6.17) and the case for $k=2$. □

LEMMA 7.2. *Given that $d_n=n^{1/2}\wedge n\lambda_1^{1/2m}$, we have*

$$[\mathbf{X}'A(\lambda_1)\boldsymbol{\epsilon}]_i=O_P(\lambda_1^{-1/4m}), \tag{7.18}$$

$$[\mathbf{X}'((I-A(\lambda_1))\mathbf{f}_0+\boldsymbol{\epsilon})]_i=O_P(n^{1/2}), \tag{7.19}$$

$$[\mathbf{X}'(I-A(\lambda_1))\mathbf{X}/n]_{ij}=\mathbf{R}_{ij}+O_P(n^{-1/2}\vee n^{-1}\lambda_1^{-1/2m}), \tag{7.20}$$

$$\|\mathbf{X}'(I-A(\lambda_1))\mathbf{X}/n-R\|=o_P(1). \tag{7.21}$$

Proof: We first state the Lemma 4.1 and 4.3 in [3]:

$$n^{-1}\sum_j[(I-A)\mathbf{f}_0]_j^2\leq\lambda_1\int_0^1(f_0^{(m)}(t))^2dt, \tag{7.22}$$

$$tr(A)=O(\lambda_1^{-1/2m})\quad\text{and}\quad tr(A^2)=O(\lambda_1^{-1/2m}). \tag{7.23}$$

By the fact that $Var[(\mathbf{X}'A\boldsymbol{\epsilon})_i]=\sigma^2R_{ii}tr(A^2)$, we can show that $[\mathbf{X}'A\boldsymbol{\epsilon}]_i=O_P(\lambda_1^{-1/4m})$ based on (6.23), thus proved (6.18). We first write the left hand side of (6.19) as $\sqrt{n}\sum_{j=1}^{n}W_{ij}$, where

$$W_{ij}=n^{-1/2}X_{ij}(\epsilon_j+((I-A)\mathbf{f}_0)_j)\ \text{ and }\ X_{ij}\text{ is the }(j,i)-th\text{ element of }\mathbf{X}$$

for $i=1,\ldots,d_n$. We next apply the Lindeberg's theorem to $\sum_j W_{ij}$. It is easy to show that $Var(\sum_j W_{ij})=\mathbf{R}_{ii}\sigma^2+\mathbf{R}_{ii}n^{-1}\sum_j[(I-A)\mathbf{f}_0]_j^2$. By (6.22), we have $Var(\sum_j W_{ij})\to\mathbf{R}_{ii}\sigma^2$. We next verify the Liapounov's condition:

$$\begin{aligned}
\sum_j E|W_{ij}|^3 &= n^{-3/2}E|X_{ij}|^3\sum_j E|\epsilon_j+[(I-A)\mathbf{f}_0]_j|^3 \\
&\leq 3n^{-3/2}\left[nE|\epsilon|^3+\sum_j|[(I-A)\mathbf{f}_0]_j|^3\right]\to 0
\end{aligned}$$

by the sub-exponential tail of ϵ and (6.16). Then the Lindeberg's theorem implies (6.19). As for (6.20), we first write (6.20) as the sum of $\mathbf{R}_{ij}$, $[\mathbf{X}'\mathbf{X}/n]_{ij}-\mathbf{R}_{ij}$ and $[-\mathbf{X}'A\mathbf{X}/n]_{ij}$. By the central limit theorem, the second term in the above decomposition is $O_P(n^{-1/2})$. For the last term, we have $E\{(\mathbf{X}'A\mathbf{X})_{ij}\}^2=$

$$(\mathbf{R}_{ij})^2(tr(A))^2+(\mathbf{R}_{ii}R_{jj}+(\mathbf{R}_{ij})^2)tr(A^2)+(E(X_{1i}X_{1j})^2-2(\mathbf{R}_{ij})^2-\mathbf{R}_{ii}\mathbf{R}_{jj})\sum_r A_{rr}^2$$

for $i\neq j$. When $i=j$, we have $E|(\mathbf{X}'A\mathbf{X})_{ii}|=\mathbf{R}_{ii}tr(A)$. By considering (6.23) we have proved (6.20). (6.20) implies that

$$\|\mathbf{X}'(I-A)\mathbf{X}/n-\mathbf{R}\|=O_P(d_nn^{-1/2}\vee d_nn^{-1}\lambda_1^{-1/2m}). \tag{7.24}$$

Thus (6.21) follows from the dimension condition D1. $\square$

Proof of Lemma 3.1: Based on the definition on $\tilde{\boldsymbol\beta}_{PS}$, we have the below inequality:

$$\frac{1}{n}(\tilde{\boldsymbol\beta}_{PS}-\boldsymbol\beta_0)'\mathbf{X}'(I-A)\mathbf{X}(\tilde{\boldsymbol\beta}_{PS}-\boldsymbol\beta_0)-\frac{2}{n}(\tilde{\boldsymbol\beta}_{PS}-\boldsymbol\beta_0)'\mathbf{X}'(I-A)(\mathbf{f}_0+\boldsymbol\epsilon)\leq 0.$$

Let $\delta_n=n^{-1/2}[\mathbf{X}'(I-A)\mathbf{X}]^{1/2}(\tilde{\boldsymbol\beta}_{PS}-\boldsymbol\beta_0)$ and $\omega_n=n^{-1/2}[\mathbf{X}'(I-A)\mathbf{X}]^{-1/2}\mathbf{X}'(I-A)(\mathbf{f}_0+\boldsymbol\epsilon)$. Then the above inequality can be rewritten as $\|\delta_n\|^2-2\omega_n'\delta_n\leq 0$, i.e. $\|\delta_n-\omega_n\|^2\leq\|\omega_n\|^2$. By Cauchy-Schwartz inequality, we have $\|\delta_n\|^2\leq 2(\|\delta_n-\omega_n\|^2+\|\omega_n\|^2)\leq 4\|\omega_n\|^2$. Examine $\|\omega_n\|^2=K_{1n}+K_{2n}+K_{3n}$, with

$$K_{1n}=n^{-1}\boldsymbol\epsilon'(I-A)\mathbf{X}[\mathbf{X}'(I-A)\mathbf{X}]^{-1}\mathbf{X}'(I-A)\boldsymbol\epsilon$$
$$K_{2n}=2n^{-1}\boldsymbol\epsilon'(I-A)\mathbf{X}[\mathbf{X}'(I-A)\mathbf{X}]^{-1}\mathbf{X}'(I-A)\mathbf{f}_0(t)$$
$$K_{3n}=n^{-1}\mathbf{f}_0(T)'(I-A)\mathbf{X}[\mathbf{X}'(I-A)\mathbf{X}]^{-1}\mathbf{X}'(I-A)\mathbf{f}_0(t)$$

Applying (6.18), (6.19) and (6.20) to the above three terms, we can conclude that all of them are of the order $O_P(d_nn^{-1})$ by considering the matrix inequalities (6.13)-(6.15). Thus we have proved (3.6) by considering (6.21). $\square$

Proof of Theorem 3.2: Let $\alpha_n=\sqrt{d_n/n}$. Similar as (6.4), we have

$$Q(\boldsymbol\beta_0+\alpha_n\mathbf{s})-Q(\boldsymbol\beta_0)\geq\alpha_n\mathbf{s}'\dot{L}(\boldsymbol\beta_0)+\frac{1}{2}\mathbf{s}'[\alpha_n^2\ddot{L}(\boldsymbol\beta_0)]\mathbf{s}+\lambda_2\sum_{j=1}^{q_n}\frac{|\beta_{0j}+\alpha_ns_j|-|\beta_{0j}|}{|\tilde{\beta}_j|^\gamma}, \tag{7.25}$$

where the forms of $\dot{L}(\boldsymbol\beta_0)$ and $\ddot{L}(\boldsymbol\beta_0)$ are specified in the proof of Theorem 3.1. By considering the lemma 6.2, (6.13) and (6.15) in the appendix, we have

$$\alpha_n\mathbf{s}'\dot{L}(\boldsymbol\beta_0)=\|\mathbf{s}\|O_P(d_n/n) \tag{7.26}$$
$$\frac{1}{2}\mathbf{s}'[\alpha_n^2\ddot{L}(\boldsymbol\beta_0)]\mathbf{s}=(d_n/n)\mathbf{s}'\mathbf{R}\mathbf{s}+O_P(d_n^2n^{-3/2}\vee d_n^2n^{-2}\lambda_1^{-1/2m}) \tag{7.27}$$

given any $\|\mathbf{s}\|=C$ independent of n. Thus the first two terms in the right hand side of (6.25) are of the same order $O_P(d_n/n)$ due to $d_n=o(n^{1/2}\wedge n\lambda_1^{1/2m})$. The second term, which is positive, dominates the first one by allowing sufficiently large C. The last term is bounded by $\lambda_2\alpha_n\|\mathbf{s}\|$. Thus, we assume $\sqrt{n}\lambda_2/\sqrt{d_n}\to 0$ so that the last term of (6.25) is $o_P(d_n/n)$. This completes the proof of (3.7).

We next show the nonparametric rate for $\widehat{f}$ by using similar analysis for the fixed dimensional case. Recall that $g(x,t)=x'\boldsymbol{\beta}+f(t)$. Similarly, we can show $\|\widehat{g}-g_0\|_n=O_P(1)$. Combining the fact that $\|g_0\|_\infty=O_P(q_n)$, we have $\|\widehat{g}\|_n=O_P(q_n)$. By assuming that $\lambda_{min}(\sum_k\phi_k\phi_k'/n)\geq c_3>0$, we can obtain

$$\frac{\|\hat{g}\|_\infty}{1+J_{\hat{g}}}=O_P\left(\frac{q_n}{1+J_{\hat{g}}}\right)$$

by similar analysis. Thus, by applying Theorem 2.2 in [16], we have established the below inequalities:

$$\begin{aligned}\lambda_1 J_{\hat{f}}^2 \leq &\left[\|\hat{g}-g_0\|_n^{1-1/2m}(1+J_{\hat{f}})^{1/2m}q_n^{1/2m}\vee(1+J_{\hat{f}})q_n n^{-\frac{2m-1}{2(2m+1)}}\right]O_P(n^{-1/2})\\ &+\lambda_1 J_{f_0}^2+\lambda_2(J_{\boldsymbol{\beta}_0}-J_{\widehat{\boldsymbol{\beta}}}),\end{aligned}\tag{7.28}$$

$$\begin{aligned}\|\hat{g}-g_0\|_n^2 \leq &\left[\|\hat{g}-g_0\|_n^{1-1/2m}(1+J_{\hat{f}})^{1/2m}q_n^{1/2m}\vee(1+J_{\hat{f}})q_n n^{-\frac{2m-1}{2(2m+1)}}\right]O_P(n^{-1/2})\\ &+\lambda_1 J_{f_0}^2+\lambda_2(J_{\boldsymbol{\beta}_0}-J_{\widehat{\boldsymbol{\beta}}}).\end{aligned}\tag{7.29}$$

Let $a_n=\|\widehat{g}-g_0\|_n/[(1+J_{\hat{f}})q_n]$, then (6.29) becomes

$$\begin{aligned}a_n^2 &\leq O_P(n^{-1/2})a_n^{1-1/2m}\vee O_P(n^{-2m/(2m+1)})\vee O_P(\lambda_1/q_n)\vee\frac{\lambda_2(J_{\boldsymbol{\beta}_0}-J_{\widehat{\boldsymbol{\beta}}})}{q_n}\\ &\leq O_P(n^{-1/2})a_n^{1-1/2m}\vee O_P(n^{-2m/(2m+1)})\vee\frac{\lambda_2(J_{\boldsymbol{\beta}_0}-J_{\widehat{\boldsymbol{\beta}}})}{q_n}\\ &\leq O_P(n^{-1/2})a_n^{1-1/2m}\vee O_P(n^{-2m/(2m+1)})\end{aligned}\tag{7.30}$$

In view of the condition $\lambda_1/q_n\asymp n^{-2m/(2m+1)}$, the second inequality in the above follows. The last inequality follows from the below analysis. Note that

$$\begin{aligned}\frac{\lambda_2(J_{\boldsymbol{\beta}_0}-J_{\widehat{\boldsymbol{\beta}}})}{q_n} &\leq \left(\lambda_2\sum_{j=1}^{q_n}\frac{|\beta_{0j}-\hat{\beta}_j|}{|\tilde{\beta}_j|^\gamma}+\lambda_2\sum_{j=q_n+1}^{d_n}\frac{|\beta_{0j}-\hat{\beta}_j|}{|\tilde{\beta}_j|^\gamma}\right)q_n^{-1}\\ &\lesssim \left(\lambda_2\sum_{j=1}^{q_n}|\beta_{0j}-\widehat{\beta}_j|+\max_{j=q_n+1,\ldots,d_n}\frac{\lambda_2}{|\tilde{\beta}_j|^\gamma}\sum_{j=q_n+1}^{d_n}|\beta_{0j}-\widehat{\beta}_j|\right)q_n^{-1}\\ &\lesssim \left[\max_{j=q_n+1,\ldots,d_n}\frac{\lambda_2/q_n}{|\tilde{\beta}_j|^\gamma}\right]O_P(\sqrt{d_n/n})\\ &\leq O_P(n^{-2m/(2m+1)})\end{aligned}$$

since $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_P(\sqrt{d_n/n})$ and (3.8). Therefore (6.30) implies that $a_n=O_P(n^{-m/(2m+1)})$. We next analyze (6.28) which can be rewritten as

$$\begin{aligned}\frac{\lambda_1}{q_n}(J_{\hat{f}}-1) &\leq O_P(n^{-1/2})a_n^{1-1/2m}\vee O_P(n^{-2m/(2m+1)})\\ (J_{\hat{f}}-1) &\leq \frac{q_n}{\lambda_1}O_P(n^{-2m/(2m+1)})\\ J_{\hat{f}} &\leq O_P(1).\end{aligned}$$

in view of the condition that $\lambda_1/q_n \asymp n^{2m/(2m+1)}$. Finally, we have proved that $\|\hat{g}-g_0\|_n = O_P(n^{-m/(2m+1)}q_n)$. Combining the triangle inequality and $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| = O_P(\sqrt{d_n/n})$, we complete the whole proof of (3.9). $\square$

Proof of Theorem 3.3: Proof of part (a) is similar as that in the fixed dimension case, i.e. 3(a) in Theorem 3.1. It follows from the regular condition $\lambda_1/q_n \asymp n^{-2m/(2m+1)}$, Lemma 6.2 and assumption (3.11).

We next prove the asymptotic normality of $\widehat{\boldsymbol{\beta}}_1$. Similar as the proof for 3(b) in Theorem 3.1, we can establish that

$$\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} = [\mathbf{X}_1'(I-A)\mathbf{X}_1]^{-1}\left[\mathbf{X}_1'(I-A)(\mathbf{f}_0(t)+\boldsymbol{\epsilon}) - \frac{n\lambda_2}{2}Pe(\widehat{\boldsymbol{\beta}}_1)\right], \tag{7.31}$$

where $Pe(\widehat{\boldsymbol{\beta}}_1) = (sign(\widehat{\beta}_1)/|\tilde{\beta}_1|^\gamma, \ldots, sign(\widehat{\beta}_{q_n})/|\tilde{\beta}_{q_n}|^\gamma)'$. Note that the invertibility of $\mathbf{X}_1(I-A)\mathbf{X}_1$ follows from (6.21) and the asymptotic invertibility of $\mathbf{R}$, i.e. the condition R3D. Thus, we have

$$\begin{aligned}
& \sqrt{n}\mathbf{G}_n\mathbf{R}_{11}^{-1/2}(\mathbf{X}_1'(I-A)\mathbf{X}_1/n)(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \\
= \;& \sqrt{n}\mathbf{G}_n\mathbf{R}_{11}^{-1/2}\left[\frac{\mathbf{X}_1'(I-A)(\mathbf{f}_0(t)+\boldsymbol{\epsilon})}{n} - \frac{\lambda_2}{2}Pe(\widehat{\boldsymbol{\beta}}_1)\right] \\
= \;& M_{1n} + M_{2n} + M_{3n},
\end{aligned} \tag{7.32}$$

where

$$\begin{aligned}
M_{1n} &= n^{-1/2}\mathbf{G}_n\mathbf{R}_{11}^{-1/2}\mathbf{X}_1'[(I-A)\mathbf{f}_0(t)+\boldsymbol{\epsilon}], \\
M_{2n} &= -n^{-1/2}\mathbf{G}_n\mathbf{R}_{11}^{-1/2}\mathbf{X}_1'A\boldsymbol{\epsilon}, \\
M_{3n} &= -(\sqrt{n}\lambda_2/2)\mathbf{G}_n\mathbf{R}_{11}^{-1/2}Pe(\widehat{\boldsymbol{\beta}}_1).
\end{aligned}$$

In order to derive the asymptotic distribution of $M_{1n}+M_{2n}+M_{3n}$, we apply the Cramer-Wold device. Let $\mathbf{v}$ be a l-vector. We first show that $\mathbf{v}'M_{2n} = o_P(1)$ and $\mathbf{v}'M_{3n} = o_P(1)$. It is easy to show

$$\begin{aligned}
|\mathbf{v}'M_{2n}| \;&\leq n^{-1/2}\|\mathbf{v}\|\|\mathbf{G}_n\mathbf{R}_{11}^{-1/2}\mathbf{X}_1'A\boldsymbol{\epsilon}\| \leq (n\lambda_{min}(\mathbf{R}_{11}))^{-1/2}\|\mathbf{v}\|\|\mathbf{G}_n\mathbf{X}_1'A\boldsymbol{\epsilon}\| \\
&\leq O_P(n^{-1/2}\sqrt{q_n}\lambda_1^{-1/4m}) = o_P(1)
\end{aligned}$$

The last inequality follows from $\mathbf{G}_n\mathbf{G}_n' \to \mathbf{G}$ and (6.18). The conditions that $\lambda_1/q_n \asymp n^{-2m/(2m+1)}$ and $n^{m/(2m+1)}\lambda_1 \to 0$ imply its convergence to zero. As for $\mathbf{v}'M_{3n}$, we have

$$|\mathbf{v}'M_{3n}| \;\leq \frac{\sqrt{n}\lambda_2}{2}\|\mathbf{v}\|\|\mathbf{G}_n\mathbf{R}_{11}^{-1/2}Pe(\widehat{\boldsymbol{\beta}}_1)\| \leq O_P(\sqrt{n}\lambda_2)\|\mathbf{G}_nPe(\widehat{\boldsymbol{\beta}}_1)\| \leq O_P(\sqrt{n}\lambda_2\sqrt{q_n}) = o_P(1)$$

by the stated condition $q_n = o(n^{-1}\lambda_2^{-2})$.

As for $\mathbf{v}'M_{1n}$, we can rewrite it as

$$\mathbf{v}'M_{1n} = \sum_{j=1}^n n^{-1/2}\mathbf{v}'\mathbf{G}_n\mathbf{R}_{11}^{-1/2}\mathbf{w}_j[(I-A)\mathbf{f}_0(t)+\boldsymbol{\epsilon}]_j \equiv \sum_{j=1}^n T_j.$$

and apply Lindeberg's theorem (Theorem 1.15 in [20]) to show its asymptotic distribution. First,

$$Var(\sum_j T_j) = \sum_j Var(T_j) = \mathbf{v}'\mathbf{G}_n\mathbf{G}_n'\mathbf{v}(\sigma^2 + n^{-1}\sum_{l=1}^n((I-A)\mathbf{f}_0)_l^2) \to \sigma^2\mathbf{v}'\mathbf{G}\mathbf{v} \tag{7.33}$$

by $\mathbf{G}_n\mathbf{G}_n' \to \mathbf{G}$ and (6.16). We next verify the condition that

$$\sum_{j=1}^n E(T_j^2 I\{|T_j| > \delta\sigma\sqrt{\mathbf{v}'\mathbf{G}\mathbf{v}}\}) = o(\sigma^2\mathbf{v}'\mathbf{G}\mathbf{v})$$

for any $\delta > 0$. Note that

$$\begin{aligned}\sum_{j=1}^n E(T_j^2 I\{|T_j| > \delta\sigma\sqrt{\mathbf{v}'\mathbf{G}\mathbf{v}}\}) &\leq \sum_{j=1}^n (ET_j^4)^{1/2}(P(|T_j| > \delta\sigma\sqrt{\mathbf{v}'\mathbf{G}\mathbf{v}}))^{1/2} \\ &\leq \left(\sum_{j=1}^n ET_j^4\right)^{1/2}\left(\sum_{j=1}^n P(|T_j| > \delta\sigma\sqrt{\mathbf{v}'\mathbf{G}\mathbf{v}})\right)^{1/2}.\end{aligned}$$

In view of (6.33), we obtain

$$\sum_{j=1}^n P(|T_j| > \delta\sigma\sqrt{\mathbf{v}'\mathbf{G}\mathbf{v}}) \leq \frac{\sum_{j=1}^n ET_j^2}{\delta^2\sigma^2\mathbf{v}'G\mathbf{v}} \to \frac{1}{\delta^2}$$

and

$$\begin{aligned}\sum_{j=1}^n ET_j^4 &\leq \frac{\|\mathbf{v}\|^4\sum_{j=1}^n E\|\mathbf{G}_n\mathbf{R}_{11}^{-1/2}\mathbf{w}_j\|^4 E[(I-A)\mathbf{f}_0 + \boldsymbol{\epsilon}]_j^4}{n^2} \\ &\leq \frac{8\|\mathbf{v}\|^4\sum_{j=1}^n E\|\mathbf{G}_n\mathbf{R}_{11}^{-1/2}\mathbf{w}_j\|^4([(I-A)\mathbf{f}_0]_j^4 + E\epsilon^4)}{n^2}.\end{aligned}$$

Note that

$$E\|\mathbf{G}_n\mathbf{R}_{11}^{-1/2}\mathbf{w}_j\|^4 \leq lq_n^2\lambda_{min}^{-2}(\mathbf{R}_{11})\sum_{i=1}^l \|g_i\|^4 = O(q_n^2),$$

where $\mathbf{G}_n' = (g_1, \ldots, g_l)$, due to $\mathbf{G}_n\mathbf{G}_n' \to \mathbf{G}$. Combined with the above analysis we have $\sum_j ET_j^4 = O(q_n^2\lambda_1^2 \vee q_n^2 n^{-1})$ given the sub-exponential tail of ϵ and (6.16). By the conditions that $q_n \leq d_n = o(n^{1/3})$ and $\lambda_1/q_n \asymp n^{-2m/(2m+1)}$, we have verified the condition that $\sum_{j=1}^n E(T_j^2 I\{|T_j| > \delta\sigma\sqrt{\mathbf{v}'\mathbf{G}\mathbf{v}}\}) = o(\sigma^2\mathbf{v}'\mathbf{G}\mathbf{v})$. Therefore, we have proved that (6.32) $= N(0, \sigma^2\mathbf{G}) + o_P(1)$.

Then we have

$$\begin{aligned}\sqrt{n}\mathbf{G}_n\mathbf{R}_{11}^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) &= \sqrt{n}\mathbf{G}_n\mathbf{R}_{11}^{-1/2}(\mathbf{R}_{11} - \mathbf{X}_1'(I-A)\mathbf{X}_1/n)(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) + N(0, \sigma^2\mathbf{G}) + o_P(1) \\ &= N(0, \sigma^2\mathbf{G}) + o(1) + O_P(d_n^{3/2}n^{-1/2} \vee d_n^{3/2}n^{-1}\lambda_1^{-1/2m})\end{aligned} \tag{7.34}$$

by the matrix inequality, (6.24) and (3.7). The stated condition $d_n = o(n^{1/3} \wedge n^{2/3}\lambda_1^{1/3m})$ implies that the rest term in (6.34) is $o_P(1)$. This completes the proof of (3.12). ☐

LITERATURE CITED

[1] Abramowitz, M. and Stegun, I. (1964) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York.

[2] Breiman, L. (1995). Better subset selection using the nonnegative garrote. *Technometrics* **37**, 373-384.

[3] 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.

[4] Denby, L. (1984). Smooth regression functions. Ph.D. Thesis. Department of Statistics. University of Michigan.

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

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

[7] Fan, J. and Li, R. (2004). New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *Journal of American Statistical Association* **99**, 710-723.

[8] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* **32**, 928-961.

[9] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. (with discussion) *Journal of Royal Statistical Society B* **70**, 849-911.

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

[11] Gu, C. (2002). Smoothing spline ANOVA models. New York: Springer-Verlag.

[12] Heckman, N. (1986). Spline smoothing in a partly linear models. *Journal of Royal Statistical Society, Series B* **48**, 244-248.

[13] Huang, J., Horowitz, J. and Ma, S. (2008), Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *Annals of Statistics* **36**, 587-613.

[14] Huang, J., Ma, S., and Zhang, C. H. (2008), Adaptive LASSO for sparse high dimensional regression, *Statistica Sinica* **18**, 1603-1618.

[15] Kimeldorf, G and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Math. Anal. Applic.* **33**, 82-95.

[16] Mammen, E. and van de Geer, Sara. (1997). Penalized quasi-likelihood estimation in partially linear models. *Annals of Statistics* **25**, 1014-1035.

[17] Portnoy, S. (1984). Asymptotic Behavior of M-Estimator of p Regression Parameters when p^2/n

is large. I. Consistency. *Annals of Statistics* **12**, 1298-1309.

[18] Rice, J. (1986). Convergence Rates for Partially Spline Model. *Statistics and Probability Letters* **4**, 203-208.

[19] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003) Semiparametric Regression, Cambridge University Press.

[20] Shao, J. (2007) Mathematical Statistics. 2nd Ed, Springer.

[21] Shiau, J. and Wahba, G. (1988). Rates of convergence for some estimates of a semi-parametric model. *Commun. Statist.-Simula.* **17**, 111-113.

[22] Speckman, P. (1988). Kernel smoothing in partially linear models. *Journal of Royal Statistical Society-B* **50**, 413-436.

[23] Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freida, F., Redwine, E., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients. *Journal of Urology* **16**, 1076-1083.

[24] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

[25] van der Vaart, A. W., and Wellner, J. A. (1996) Weak Convergence and Empirical Processes: With Applications to Statistics. Springer, New York

[26] Wahba, G. (1984) Partial spline models for the semiparametric estimation functions of several variables. In *Statistics: an Appraisal, Proc. 50th Anniversary Conf.*, eds H. A. David and H. T. David. Ames: Iowa State University Press.

[27] Wahba, G. (1990) *Spline Models for Observational Data*. SIAM. CBMS-NSF Regional Conference Series in Applied Mathematics, volume 59.

[28] Wang, H., Li, R., and Tsai, C.L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

[29] Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection via the LAD-LASSO. *Journal of Business & Economics Statistics* **20**, 347-355.

[30] Wang, H., Li, B., and Leng, C. (2009) Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society, Series B* **71** 671-683.

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

[32] Zhang, H. H. and Lu, W. (2007). Adaptive-LASSO for Cox's proportional hazards model. *Biometrika* **94**, 691-703.

[33] Zou, H. (2006). The adaptive lasso and its oracle properties. *JASA* **101**, 1418-1429.

[34] Zou, H. and Zhang, H. H. (2009). On The Adaptive Elastic-Net With A Diverging Number of Parameters. *Annals of Statistics* **37**, 1733-1751.