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

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

†Hao Helen Zhang is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, Email: hzhang2@stat.ncsu.edu

1. Introduction

1.1. Background

Partial smoothing splines are an important class of semiparametric regression models. Developed in a theoretically elegant framework of reproducing kernel Hilbert spaces (RKHS), these models provide a nice compromise between linear and nonparametric models and have many successful applications. Plenty of work has been done on partial spline models in literature, from the early work of [6, 10, 16, 18, 29, 34, 36], to the relatively newer papers of [24, 25, 30], to the most recent work of [22] and a comprehensive review by [17].

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 called the smooth part or the nonparametric component of the model, since its functional form is not explicitly assumed. The standard approach to compute the partial spline (PS) estimator is minimizing the penalized least squares:

$$(\widetilde{\boldsymbol{\beta}}_{PS}, \tilde{f}_{PS}) = \text{argmin}_{\boldsymbol{\beta}\in R^d, f\in W_m} \frac{1}{n}\sum_{i=1}^{n}[y_i - \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} - f(t_i)]^2 + \lambda_1 \int_0^1 \left[f^{(m)}(t)\right]^2 dt, \tag{1.2}$$

where λ_1 is the smoothing parameter and the operator $J_f^2 = \int_0^1 \left[f^{(m)}(t)\right]^2 dt$ is the roughness penalty on f, see [7, 8, 15, 21] for details. It is known that the solution $\tilde{f}_{PS}$ is a natural spline of order $(2m-1)$ on $[0,1]$ with knots at the points $t_i, i = 1, \cdots, n$. Partial smoothing splines are attractive for its elegant theoretical framework and good performance in real data analysis. Asymptotic theory for partial splines has been developed by several authors including [18, 29, 33, 34]. All of these works studied the convergence rates of $\widetilde{\boldsymbol{\beta}}_{PS}$ under different conditions, while treating f as an infinite dimensional nuisance parameter. In particular, [18] first proved the root-n convergence for parametric components under certain conditions. [25] further studied the large-sample properties of the partial spline in the context of quasi-likelihood estimating.

Partially linear models have also been studied in the context of other linear smoothers such as the kernel smoothing [34] and local polynomial regression [12]. In this paper, we mainly consider partial smoothing splines in the framework of [36].

1.2. Model Selection for Partial Splines

Variable selection is fundamentally important for data analysis and model building, especially for high dimensional data. Effective variable selection could greatly 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 [3], LASSO [35], SCAD [11, 13], and the adaptive LASSO [39, 44]. Contemporary research frequently deals with problems where the input dimension d diverges to infinity as the data sample size increases [13]. There is also active research going on for linear model selection in these situations [13, 14, 19, 20, 45].

Variable selection for semiparametric models has received considerable attention recently. Compared to linear models, model selection for partially linear models is more challenging since it involves several additional issues: estimating the infinite-dimensional nonparametric component, choosing tuning parameters for nonparametric component estimation, and choosing the shrinkage parameter for parametric components estimation and selection. Here we give a brief summary on existing works. [4, 5] proposed some information-type criterion and established a consistency property of the estimate. They made use of the piecewise polynomial to estimate nonparametric part in the Besov space before selecting linear variables. [12] also did some pioneering work by using the SCAD penalty for variable selection in the local polynomial regression setup, and showed that the resulting estimator performs as well as the oracle estimate. [23] further generalized the method by [12] to the generalized varying coefficient partially linear model. Very recently, [41] suggested the use of SCAD penalty for variable selection and polynomial splines to estimate the nonparametric component. In addition, [38, 40] considered the problem of tuning parameter selection and suggested that the BIC assures the consistency of model selection even with a diverging number of parameters.

In this paper, we study an alternative approach to variable selection for partially linear

models. Different from the previously discussed work, our procedure is developed in the framework of smoothing splines. We show that the new procedure leads to a regularization problem in the RKHS, whose unified formulation can greatly facilitate the numerical computation and asymptotic inference of the estimator. To conduct variable selection, different from most of the methods above which use the SCAD penalty, we employ the adaptive LASSO penalty on linear parameters since it can lead to a convex objective function and guarantees the uniqueness of the solution. One big advantage of this procedure is its easy implementation. We show that, by using the representer theory, the optimization problem can be reformulated as a LASSO-type problem so that the entire solution path can be computed by the state-of-the-art LARS algorithm [9]. This also greatly facilitates the tuning procedure. In theory, 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 asymptotic efficiency; (iii) estimate the nonparametric component f at the optimal nonparametric rate. The result (iii) is quite remarkable, since most existing work treat f as a nuisance parameters [18, 34]. This is a new contribution to the theory of partial smoothing spline models. We will also investigate the property of the new procedure with a diverging number of predictors [13].

From now on, we regard $(Y_i, \mathbf{X}_i)$ as i.i.d realizations from some random distribution. Without loss of generality, we assume that 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 $j = 1, \cdots, d$. 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, 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. Technical proofs are presented in Section 6.

2. Method

Without loss of generality, we assume the data points $t_1, \cdots, t_n$ are distinct and sorted, i.e. $0 \le t_1 < t_2 < \cdots < t_n \le 1$. In order to achieve a smooth estimate for the nonparametric component and sparse estimate for the parametric components simultaneously, we consider the following regularization problem:

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

Motivated by the standard partial smoothing spline and the adaptive LASSO for variable selection, 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 and λ_2 are two non-negative tuning parameters which play different roles here: λ_1 controls the smoothness of the estimated nonlinear function while λ_2 controls the degrees of shrinkage on β's. The weight w_j's are pre-specified and their choices will be discussed next. For convenience, we will refer to this procedure as PSA (the partial splines equipped the adaptive penalty). It is easy to see that (2.1) includes the standard partial smoothing spline as a special case, if we set $\lambda_2 = 0$.

Choices of w_j's are essential in (2.1) to assure the optimality of the estimator from (2.1). Similar to the linear model selection, 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:

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

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 the matrix $A(\lambda_1)$ is the smoother or influence matrix [36] and its expression will be given in Section 4. Plugging $\hat{\mathbf{f}}(\boldsymbol{\beta})$

into (2.2), we can obtain an equivalent objective function of a simpler form:

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

The first term $L(\boldsymbol{\beta}) \equiv (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'[I - A(\lambda_1)](\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/n$ in (2.3) can be regarded as a profile least squares, as a function of only $\boldsymbol{\beta}$. We refer to Q as the penalized profile least squares. The form (2.3) provides a convenient form for computing the PSA estimator, since it involves only one type of penalty. In particular, the PSA solution can be computed as:

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

Note that $(\widehat{\boldsymbol{\beta}}_{PSA}, \hat{f}_{PSA})$ does not have an explicit solution form. Special software like Quadratic Programming (QP) or LARS [9] is needed to obtain the solution. Also, the PSA estimator is not linear in $\mathbf{y}$, since the minimizer of $Q(\boldsymbol{\beta})$ is not linear in $\mathbf{y}$. This is quite different from the standard partial smoothing splines, which is a linear smoother with a simple analytic solution form as $\widehat{\boldsymbol{\beta}}_{PS} = [\mathbf{X}'(I - A(\lambda_1))\mathbf{X}]^{-1}\mathbf{X}'[I - A(\lambda_1)]\mathbf{y}$ and $\hat{f}_{PS} = A(\lambda_1)(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_{PS})$.

3. Statistical Theory

Without loss of generality, 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. We also write the estimated vector $\widehat{\boldsymbol{\beta}}_{PSA} = (\hat{\beta}_1, ..., \hat{\beta}_d) = \left(\widehat{\boldsymbol{\beta}}_{PSA,1}', \widehat{\boldsymbol{\beta}}_{PSA,2}'\right)'$ accordingly. In addition, we assume the covariate vector $(\mathbf{X}, T)$ and their observations satisfy the following regularity conditions. $\mathbf{X}$ has zero mean and strictly positive definite covariance matrix $\mathbf{R}$. The observations t_i's are distinct values and 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. These technical conditions are commonly used in the literature, e.g. [18], and are assumed throughout the whole section. Note that the conditions on the smoothing parameters in the below theorems and corollaries are chosen for simplicity of expositions and are not the most general cases.

In the following, we will establish the large-sample properties of the new estimator in two scenarios: d is fixed and d diverges with n. The second dynamic modelling scenario is welcome in analyzing high dimensional data. Both the model estimation and selection properties are established for each scenario.

3.1. Asymptotic Results for Fixed d

In this section, we establish two main theoretical results. Under general regularity conditions, we show that when (λ_1, λ_2) are properly chosen, the estimates $(\widehat{\boldsymbol{\beta}}_{PSA}, \hat{f}_{PSA})$ are both consistent and achieving their optimal rates simultaneously. Secondly, we prove that the proposed procedure has the desired oracle properties for variable selection.

3.1.1. Convergence Rates of $\widehat{\boldsymbol{\beta}}_{PSA}$ and $\hat{f}_{PSA}$

In this section, 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). 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. $\|F\|_n^2 = \sum_{i=1}^n F^2(X_i)/n$.

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

R1. ϵ is assumed to be independent of X, and have the sub-exponential tail, i.e. $E(\exp(|\epsilon|/C_0)) \leq C_0$ for some $0 < C_0 < \infty$ almost surely;

R2. $\sum_k \phi_k \phi_k'/n$ converges to some non-singular matrix with $\phi_k = [1, t_k, \cdots, t_k^{m-1}, x_{k1}, \cdots, x_{kd}]'$.

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*

$$\max_{j=q+1,\ldots,d} \left(\frac{n^{\frac{2m-1}{2(2m+1)}} \lambda_2}{|\tilde{\beta}_j|^\gamma} \right) \to \lambda_{20} > 0 \tag{3.2}$$

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

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

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

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

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

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

In practice, a convenient choice for $\tilde{\beta}_j$'s would be the standard partial spline solution, which can be shown to be $\sqrt{n}$-consistent for $j = 1, \ldots, d$ under general conditions [18]. As a direct consequence of Theorem 3.1, the following corollary states that the double penalized estimators achieve the optimal rates for both parametric and nonparametric estimation, if we use the partial spline solutions to construct the weights in (2.2) and choose the tuning parameters (λ_1, λ_2) properly.

COROLLARY 3.1. *Let* $\gamma = 1$. *Assume* $\tilde{\beta}_j$*'s are the standard partial spline solution, then*

$$\|\widehat{\boldsymbol{\beta}}_{PSA} - \boldsymbol{\beta}_0\| = O_P(n^{-1/2}),$$
$$\|\hat{f}_{PSA} - f_0\|_n = O_P(n^{-m/(2m+1)}),$$

if $n^{2m/(2m+1)}\lambda_1 \to \lambda_{10} > 0$ *and* $n^{2m/(2m+1)}\lambda_2 \to \lambda_{20} > 0$.

REMARK 1. *In Theorem 3.1, the convergence rate for* $\hat{f}_{PSA}$ *is shown to be optimal, i.e.* $O_P(n^{-m/(2m+1)})$. *However, [41] shows that their polynomial spline estimate for* f *achieves the optimal convergence rate* $O_P(n^{-s_f/(2s_f+1)})$, *where* s_f *is some smoothness measurement of* f, *only under some conditions for the smoothness of* f. *Thus, we can conclude that our smoothing spline estimate is more robust in contrast with the polynomial spline estimate in estimating the sparse* $\boldsymbol{\beta}_0$.

3.1.2. Oracle Properties

In this subsection, we study the variable selection consistency of the double penalized estimator $\widehat{\boldsymbol{\beta}}_{PSA}$ and show that it satisfies the oracle properties [11, 44] when the smoothing parameters satisfy the same conditions in Theorem 3.1. In particular, our Theorem 3.2 suggests that the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{PSA,1}$ is exactly the same as that in the linear model (without f) subject to the adaptive Lasso penalty [44]. Hence, we can conclude that the double penalization

procedure can estimate the nonparametric function well enough to ensure the oracle properties of the weighted Lasso estimates in our model framework.

Theorem 3.2. *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*

$$\min_{j=q+1,\ldots,d} \left(\frac{n^{\frac{2m-1}{2(2m+1)}}\lambda_2}{|\tilde{\beta}_j|^\gamma} \right) \to \lambda_{20} > 0 \tag{3.6}$$

as $n \to \infty$, then we can show with probability tending to 1 the local minimizer $\widehat{\boldsymbol{\beta}}_{PSA} = (\widehat{\boldsymbol{\beta}}'_{PSA,1}, \widehat{\boldsymbol{\beta}}'_{PSA,2})'$ must satisfy:

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

(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 X.

REMARK 2. *Note that (3.6) can be relaxed to $\min_{j=q+1,\ldots,d} \sqrt{n}\lambda_2/|\tilde{\beta}_j|^\gamma \to \infty$.*

REMARK 3. *In (2.2), if $f \equiv 0$, then the double penalty estimation reduces to the linear model with an adaptive LASSO penalty. Note that the oracle properties of the adaptive LASSO estimator were established when the related smoothing parameter $\sqrt{n}\lambda_2 \to 0$ and $n\lambda_2 \to \infty$; see Theorem 2 in [44]. In the standard partial smoothing spline without sparsity, i.e. $\lambda_2 = 0$ in (2.2), the asymptotic normality of $\hat{\boldsymbol{\beta}}_{PS}$ and optimal convergence rate for $\hat{f}_{PS}$ are attained when $n^{2m/(2m+1)}\lambda_1 \to \lambda_{10} > 0$, see [18, 25]. By assuming the consistent conditions of the smoothing parameters to the above classical results, our new estimation procedure essentially unifies the oracle properties and optimal convergence results in the partially linear model in one framework.*

The following corollary is a direct consequence of Theorem 3.2 and Corollary 3.1.

COROLLARY 3.2. *Let $\gamma = 1$. Assume $\tilde{\beta}_j$'s are the standard partial spline solution, then*

(a). the optimal convergence rates for $(\widehat{\boldsymbol{\beta}}_{PSA}, \hat{f}_{PSA})$ are achieved;

(b). $\widehat{\boldsymbol{\beta}}$ possesses the oracle properties,

if $n^{2m/(2m+1)}\lambda_1 \to \lambda_{10} > 0$ and $n^{2m/(2m+1)}\lambda_2 \to \lambda_{20} > 0$.

REMARK 4. *Note that in practice it is often unknown which variables are unimportant predictors. Hence, a verifiable version of the conditions (3.2) (or 3.6) is*

$$n^{(2m-1)/(2(2m+1))+\alpha'\gamma}\lambda_2 \to \lambda_{20} > 0,$$

where $n^{-\alpha'}$ is the fastest (or slowest) convergence rate among the consistent $\tilde{\beta}_j$'s for $j = 1, \ldots, d$.

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.

The penalized method for linear models in the context of diverging parameter was first considered by [13]. They showed that the oracle property still holds even when $d_n = o(n^{1/5})$ (relaxed to $o(n^{1/3})$ under some special cases). [41] further extended their results to the partly linear model by taking the polynomial spline estimation approach in which the choice of knots is neither automatic nor data dependent. In contrast, our partial smoothing spline method provides a unified framework to select the parametric components while estimating the non-parametric component. Now, we give the additional regularity conditions required to establish the large-sample theory for the increasing dimensional case except for those assumed for the fixed dimension case, i.e. the sub-exponential tail of ϵ and (3.1):

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

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

R2D. $\lambda_{min}(\sum_k \phi_k \phi_k') \geq c_3 > 0$ for any n.

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

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

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 $\tilde{\boldsymbol{\beta}}$ given the increasing dimension d_n, which is crucial in determining the shrinkage rate of the penalty term in the formulation (2.2). We use the symbol $p_n \asymp q_n$ to indicate some random quantity

$p_n = O_P(q_n)$ and $p_n^{-1} = O_P(q_n^{-1})$ for $q_n \to 0$. Define $x \vee y$ ($x \wedge y$) to be the maximum (minimum) value of x and y.

Lemma 3.1. *Suppose that $\tilde{\boldsymbol{\beta}}$ is a partial smoothing spline estimate given $d_n = n^{1/2} \wedge n\lambda_1^{1/2m}$, then we have*

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

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 interesting results: (i) the convergence rate for $\widehat{\boldsymbol{\beta}}_{PSA}$ coincides with that for the estimator in the linear regression model with increasing dimension [27], thus we can conclude that the presence of nonparametric function and sparsity of $\boldsymbol{\beta}_0$ does not affect the overall convergence rate of $\widehat{\boldsymbol{\beta}}$; (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 component 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 3.3. *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*

$$\|\widehat{\boldsymbol{\beta}}_{PSA} - \boldsymbol{\beta}_0\| = O_P(\sqrt{d_n/n}). \tag{3.8}$$

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

$$\max_{j=q_n+1,\ldots,d_n} \frac{\sqrt{n/d_n}\lambda_2}{|\tilde{\beta}_j|^\gamma} = O_P(n^{1/(2m+1)} d_n^{-1} q_n), \tag{3.9}$$

then we have

$$\|\widehat{f}_{PSA} - f_0\|_n = O_P(\sqrt{d_n/n} \vee n^{-m/(2m+1)} q_n), \tag{3.10}$$

$$J_{\hat{f}_{PSA}} = O_P(1). \tag{3.11}$$

3.2.2. Oracle Properties

In this subsection, we will show that our regularization method gives the efficient sparse estimator even when the dimension of $\boldsymbol{\beta}_0$ diverges to infinity. Specifically, the estimator $\widehat{\boldsymbol{\beta}}_{PSA}$ has achieved the well known oracle properties. In particular, the nonzero estimate $\widehat{\boldsymbol{\beta}}_{PSA,1}$ are

shown to be asymptotically normal with the same means and covariances that they would have if the zero coefficients were known in advance. When showing the asymptotic normality of the important covariate estimator $\widehat{\boldsymbol{\beta}}_{PSA,1}$, we will 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. This is a standard treatment for studying the asymptotic behaviors of the regression parameters with increasing dimension [28, 41].

Theorem 3.4. *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:*

$$\min_{j=q_n+1,\ldots,d_n} \frac{\sqrt{n/d_n}\lambda_2}{|\tilde{\beta}_j|^\gamma} \to \infty, \tag{3.12}$$

we have

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

(b). Asymptotic Normality:

$$\sqrt{n}\mathbf{G}_n R_{11}^{1/2}(\widehat{\boldsymbol{\beta}}_{PSA,1} - \boldsymbol{\beta}_1) \xrightarrow{d} N(\mathbf{0}, \sigma^2\mathbf{G}), \tag{3.13}$$

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

The corollary 3.3, an immediate consequence of the Theorem 3.4, is an important result of our paper. In the Corollary 3.3, we give the fastest possible increasing rates for the dimension of the parametric covariate and its important components in guaranteeing the estimation efficiency and selection consistency. Meanwhile, the range of the smoothing and shrinkage parameters are also given. These conditions are easily verifiable in the applications, see the remark 5.

COROLLARY 3.3. *Let* $\gamma = 1$. *Assume* $\tilde{\boldsymbol{\beta}}$ *is the partial smoothing spline solution, then*

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:

$$d_n = o(n^{1/3}) \text{ and } q_n = o(n^{1/3}), \tag{3.14}$$

$$n\lambda_1^{1/2m} \to \infty, n^{m/(2m+1)}\lambda_1 \to 0 \text{ and } \lambda_1/q_n \asymp n^{-2m/(2m+1)}, \tag{3.15}$$

$$\sqrt{n/d_n}\lambda_2 \to 0, (n/d_n)\lambda_2 \to \infty \text{ and } (n/q_n)\lambda_2 = O(n^{1/(2m+1)}). \tag{3.16}$$

REMARK 5. *The above conditions (3.14)-(3.16) can be verified in many cases. For example, given $m = 2$, we can set the smoothing parameters $\lambda_1, \lambda_2 \asymp n^{-0.55}$ when $d_n \asymp n^{1/4}$, $q_n \asymp n^{1/4}$.*

4. Computation and Tuning

4.1. Algorithm

In the first part of this section, we describe the computational algorithm for any given pair of the tuning parameters λ_1 and λ_2. The tuning issue will be discussed in the second part. For any fixed $\boldsymbol{\beta}$, it is easy to show that the solution $\hat{f}$ is a natural polynomial spline of order $2m-1$. From the reproducing kernel Hilbert space (RKHS) theory [21], $W_m[0,1]$ is an RKHS when equipped with the norm

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

Furthermore, $W_m[0,1]$ can be decomposed as $W_m[0,1] = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 = \{f : f^{(m)} = 0\}$ and $\mathcal{H}_1 = \{f : \int_0^1 f^{(\nu)}(t)dt = 0, \nu = 0, \cdots, m-1; f^{(m)} \in \mathcal{L}_2[0,1].\}$. It is straightforward to show that $\mathcal{H}_0 = \text{span}\{k_\nu(t), \nu = 0, \cdots, m-1\}$, where $k_\nu(t) = B_\nu(t)/\nu!$ and $B_\nu(t)$ are Bernoulli polynomials [1]. And $\mathcal{H}_1$ is an RKHS associated with the reproducing kernel (RK) $K(t,s) = k_m(t)k_m(s) + (-1)^{m-1}k_{2m}([s-t])$, where $[\tau]$ is the fractional part of τ. Let S be an $n \times m$ matrix with the (i,ν)th entry $k_{\nu-1}(t_i)$ and Σ be an $n \times n$ matrix with the (i,j)th entry $K(t_i,t_j)$. The represented theorem suggests that the solution

$$\hat{f}(t) = \sum_{\nu=0}^{m-1} b_\nu k_\nu(t) + \sum_{i=1}^{n} c_i K(t,t_i).$$

In matrix notation, we have $\widehat{\mathbf{f}} = S\mathbf{b} + \Sigma\mathbf{c}$.

Define $M = \Sigma + n\lambda_1 I$. Let the QR decomposition of S be $S = (Q_1, Q_2)\begin{pmatrix} R \\ 0 \end{pmatrix}$, where Q is orthogonal and R is upper triangular with $S'Q_2 = 0$. With $\boldsymbol{\beta}$ fixed, the standard smoothing spline theory suggests that the solution to (2.2) is linear in the residual $(\mathbf{y} - X\boldsymbol{\beta})$, i.e. $\hat{\mathbf{f}}(\boldsymbol{\beta}) = A(\lambda_1)(\mathbf{y} - X\boldsymbol{\beta})$, where the matrix A is the influence matrix and expressed as

$$A = I - n\lambda_1 Q_2(Q_2'MQ_2)^{-1}Q_2'.$$

This leads to the following weighted profile least square

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

Now we can reformulate $Q(\boldsymbol{\beta})$ as the LASSO-type problem by a proper transformation:

$$\frac{1}{n}\|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*\|^2 + \lambda_2 \sum_{j=1}^{d} |\beta_j^*|, \tag{4.1}$$

where $\mathbf{y}^* = [I - A(\lambda_1)]\mathbf{y}$, $\beta_j^* = \beta_j/|\tilde{\beta}_j|^\gamma$, $\mathbf{X}^* = [I - A(\lambda_1)]\mathbf{X}\mathbf{W}$, where $\mathbf{W} = \text{diag}\{|\tilde{\beta}_j|^\gamma\}$. Due to the smart reformulation (4.1) we can compute the solution path of β_j's for the fixed λ_1 via LARS algorithm [9] as shown in the below. This computational advantage is one of the attractive properties of the proposed procedure.

Algorithm:

1. Fit the standard smoothing spline and construct the weights w_j's. Compute $\mathbf{y}^*$ and $\mathbf{X}^*$.
2. Solve (4.1) using the LARS algorithm. Denote the solution as $\widehat{\boldsymbol{\beta}}^* = (\hat{\beta}_1^*, \cdots, \hat{\beta}_d^*)'$.
3. Compute $\widehat{\boldsymbol{\beta}}_{PSA} = (\hat{\beta}_1, \cdots, \hat{\beta}_d)'$ by $\hat{\beta}_j = \hat{\beta}_j^*|\tilde{\beta}_j|^\gamma$ for $j = 1, \cdots, d$.
4. Solve $\widehat{\mathbf{c}} = Q_2(Q_2'MQ_2)^{-1}Q_2'\mathbf{y}$ and $\widehat{\mathbf{b}} = R^{-1}Q_1'(\mathbf{y} - M\widehat{\mathbf{c}})$. Compute $\widehat{\mathbf{f}}_{PSA} = S\widehat{\mathbf{b}} + \Sigma\widehat{\mathbf{c}}$.

4.2. Parameter Tuning

The choice of (λ_1, λ_2) in (2.1) is critical, since λ_1 controls the roughness of f and λ_2 determines the model sparsity. Various criteria have been proposed for parameter tuning; the well-known criteria include Mallow's C_p [26], AIC [2], BIC [31], cross validation (CV) and generalized cross validation (GCV) [37]. Among them, the GCV score was used to choose the smoothing

parameter for the LASSO by [35]. To choose the tuning parameter for the SCAD, [11] used the GCV score and recently [38, 40] suggested the BIC as an alternative criteria. For the adaptive LASSO in the Cox's model, [43] compared the GCV and BIC score and concluded that BIC works better.

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 [35] and [11], 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 then 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 $\text{tr}\,(M(\lambda_1, \lambda_2))$. The GCV score can be constructed as

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

Though the above joint tuning procedure seems to be a natural choice, the two-dimensional search is computationally expensive in practice. In the following, we suggest an alternative approach to tune λ_1 and λ_2 in a more feasible manner. We call this a *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:

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

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 $\lambda_1^* = \arg\min_{\lambda_1} \text{GCV}(\lambda_1)$. 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 [42]. 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. [38, 40, 43] 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. $\hat{\sigma}^2 = (\mathbf{y} - X\widetilde{\boldsymbol{\beta}} - \widetilde{\mathbf{f}})'(\mathbf{y} - X\widetilde{\boldsymbol{\beta}} - \widetilde{\mathbf{f}})/(n - \mathrm{tr}(\tilde{A}) - d)$. Our empirical experiences indicate that the two-stage tuning procedure works well in our simulations.

5. Numerical Studies

We conduct Monte Carlo simulations to evaluate the finite sampling performance of the proposed method, in terms of its model estimation and variable selection. 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, which is given by partial spline models when assuming the true model were known, and use it as a golden rule for easy comparison. We use $\gamma = 1$ for PSA in all the examples.

In each setting, a total of 100 Monte Carlo (MC) simulations are carried out. We report the MC sample mean and standard deviation (given in the parentheses) for the MSEs. As in [12], we use the mean squared error $MSE(\widehat{\boldsymbol{\beta}}) = E\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ and the 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 the data knots in the simulations. To evaluate the variable selection performance of each method, we report the number of correct zero coefficients, the number of coefficients incorrectly set to 0, the model size, and the empirical probability of capturing the true model.

We generate the data from a partially linear model $Y_i = \mathbf{X}_i'\boldsymbol{\beta} + f(T_i) + \varepsilon_i$. Recall that d=dim($\boldsymbol{\beta}$) and q is the number of important covariates in $\boldsymbol{\beta}$. Consider two model settings:

- Model 1: $\boldsymbol{\beta}_1 = (3, 2.5, 2, 1.5, 0, \cdots, 0)'$, $d = 15$ and $q = 4$. And $f_1(t) = 1.5\sin(2\pi t)$.
- Model 2: $\boldsymbol{\beta}_2 = (3, \cdots, 3, 0, \cdots, 0)'$, $d = 20$ 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: Correlated covariates $(X_1, \cdots, X_{20})'$ are from marginally standard normal with AR(1) correlation, i.e. $\text{corr}(X_i, X_j) = \rho^{|i-j|}$. T is generated from Unif$(0,1)$ independently from X_i's. We choose $\rho = 0.3$ and $\rho = 0.6$, respectively.

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) $\epsilon_2 \sim t_{10}$.

We consider two sample sizes in each setting: $n = 100$ and $n = 200$.

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 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 $n = 200$.

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	0.615 (0.028)	0.027 (0.005)	15 (0)	0 (0)	0 (0)
		PSL	0.347 (0.025)	0.026 (0.004)	7.62 (0.23)	7.38 (0.23)	0.00 (0.00)
		PSA	0.243 (0.022)	0.024 (0.004)	4.64 (0.11)	10.36 (0.11)	0.00 (0.00)
		Oracle	0.125 (0.009)	0.016 (0.004)	4 (0)	11 (0)	0 (0)
	200	PS	0.241 (0.009)	0.009 (0.001)	15 (0)	0 (0)	0 (0)
		PSL	0.144 (0.008)	0.009 (0.001)	6.99 (0.21)	8.01 (0.21)	0.00 (0.00)
		PSA	0.118 (0.008)	0.009 (0.003)	4.29 (0.07)	10.71 (0.07)	0.00 (0.00)
		Oracle	0.064 (0.004)	0.008 (0.001)	4 (0)	11 (0)	0 (0)
1	100	PS	2.449 (0.113)	0.100 (0.019)	15 (0)	0 (0)	0 (0)
		PSL	1.401 (0.102)	0.097 (0.018)	7.56 (0.23)	7.44 (0.23)	0.00 (0.00)
		PSA	1.173 (0.103)	0.093 (0.017)	4.78 (0.13)	10.18 (0.12)	0.04 (0.02)
		Oracle	0.499 (0.037)	0.072 (0.011)	4 (0)	11 (0)	0 (0)
	200	PS	0.961 (0.038)	0.030 (0.002)	15 (0)	0 (0)	0 (0)
		PSL	0.578 (0.031)	0.030 (0.002)	7.00 (0.21)	8.00 (0.21)	0.00 (0.00)
		PSA	0.499 (0.034)	0.031 (0.002)	4.38 (0.08)	10.62 (0.08)	0.00 (0.00)
		Oracle	0.256 (0.016)	0.031 (0.003)	4 (0)	11 (0)	0 (0)

TABLE 2
Variable selection frequency over 100 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	100	100	35	34	39	38	28	23	32	29	30	39	35	0.08
		PSA	100	100	7	7	9	6	6	4	6	4	8	5	2	0.64
	200	PSL	100	100	28	25	31	28	29	25	18	23	32	28	32	0.10
		PSA	100	100	3	1	6	1	3	2	4	1	3	4	1	0.81
1	100	PSL	100	100	35	33	37	37	27	24	30	29	30	39	35	0.08
		PSA	100	96	8	9	13	10	9	5	6	5	8	5	4	0.50
	200	PSL	100	100	31	24	31	27	28	28	18	22	32	27	32	0.12
		PSA	100	100	4	3	7	5	2	1	4	1	5	4	2	0.74

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 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.

Table 3 compares the model fitting and variable selection performance of various procedures for the correlated setting in Model 2. Here $\rho = 0.3$ represents a small correlation among X's and $\rho = 0.6$ represents a moderate correlation. 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	0.427 (0.021)	0.503 (0.028)	20 (0)	0 (0)	0 (0)
		PSL	0.308 (0.019)	0.500 (0.027)	12.81 (0.21)	7.19 (0.21)	0.00 (0.00)
		PSA	0.205 (0.015)	0.493 (0.028)	10.22 (0.06)	9.78 (0.06)	0.00 (0.00)
		Oracle	0.173 (0.009)	0.464 (0.015)	10 (0)	10 (0)	0 (0)
	200	PS	0.181 (0.007)	0.406 (0.003)	20 (0)	0 (0)	0 (0)
		PSL	0.133 (0.006)	0.406 (0.003)	13.12 (0.17)	6.88 (0.17)	0.00 (0.00)
		PSA	0.091 (0.005)	0.405 (0.003)	10.10 (0.04)	9.90 (0.04)	0.00 (0.00)
		Oracle	0.085 (0.004)	0.403 (0.003)	10 (0)	10 (0)	0 (0)
0.6	100	PS	0.742 (0.037)	0.503 (0.028)	20 (0)	0 (0)	0 (0)
		PSL	0.411 (0.029)	0.497 (0.027)	12.04 (0.18)	7.96 (0.18)	0.00 (0.00)
		PSA	0.348 (0.027)	0.492 (0.027)	10.16 (0.05)	9.84 (0.05)	0.00 (0.00)
		Oracle	0.292 (0.015)	0.464 (0.015)	10 (0)	10 (0)	0 (0)
	200	PS	0.313 (0.013)	0.407 (0.003)	20 (0)	0 (0)	0 (0)
		PSL	0.178 (0.008)	0.405 (0.003)	12.36 (0.14)	7.64 (0.14)	0.00 (0.00)
		PSA	0.150 (0.008)	0.404 (0.003)	10.09 (0.04)	9.91 (0.04)	0.00 (0.00)
		Oracle	0.142 (0.008)	0.403 (0.003)	10 (0)	10 (0)	0 (0)

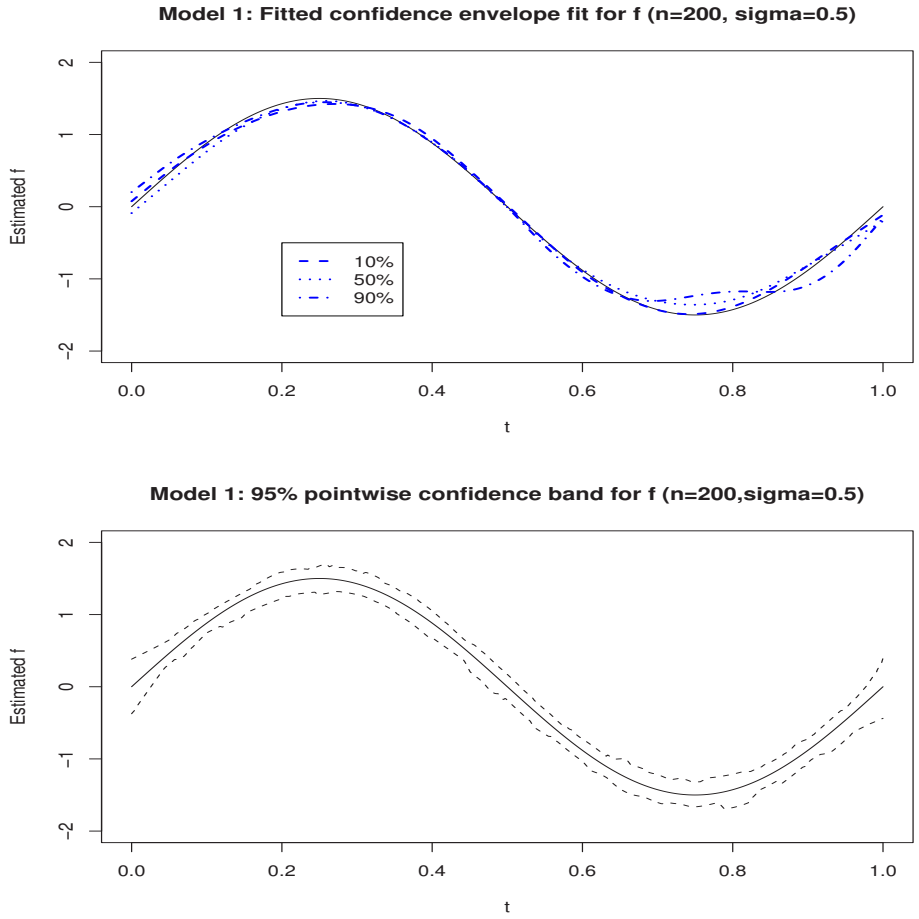
Table 4 compares the variable selection results of PSL and PSA in four scenarios if the covariates are correlated. Since neither of the methods misses any important variable over 100 runs, we only report the selection frequency for unimportant variables. Overall, the PSA results in a more sparse model 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	36	27	29	28	23	29	31	21	33	2	0.13
		PSA	1	3	4	5	1	2	1	1	2	2	0.87
	200	PSL	35	30	33	31	24	38	40	24	27	30	0.02
		PSA	1	1	1	0	0	0	2	3	2	0	0.93
0.6	100	PSL	26	17	23	21	15	20	23	16	20	23	0.17
		PSA	1	2	2	1	1	1	2	2	3	1	0.88
	200	PSL	33	20	26	22	16	19	29	21	22	28	0.06
		PSA	1	0	0	2	0	0	0	2	2	2	0.93

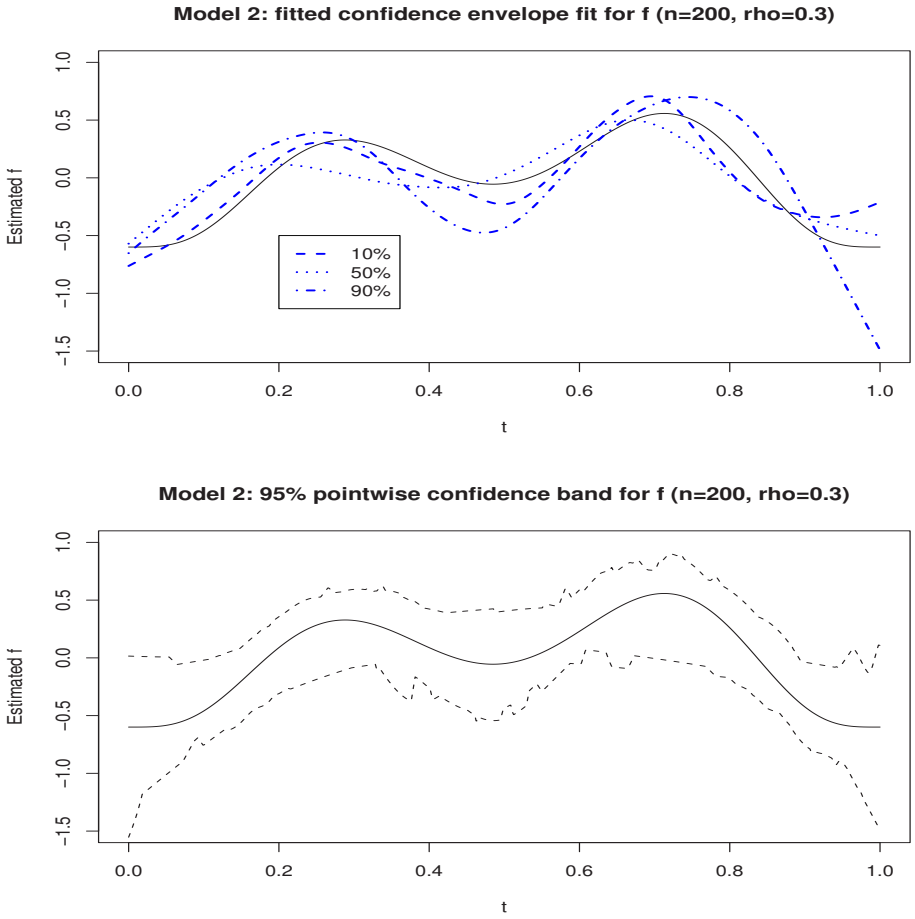
The top row of Figure 2 depicts the typical estimated functions corresponding to the 10th best, the 50th best (median), and the 90th best fits according to MISE among 100 simulations when $n = 200$ and $\rho = 0.3$. It is evident that the estimated curves are able to capture the shape of the true function 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 1. *The estimated nonlinear functions given by the PSA in Model 1 ($n = 200$, $\sigma = 0.5$)*



The estimated nonlinear function, confidence envelop and 95% point-wise confidence interval for Model 1 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 100 simulations. The bottom plot is a 95% pointwise confidence interval.

FIG 2. *The estimated nonlinear functions given by the PSA in Model 2 ($n = 200$, $\rho = 0.3$)*



The estimated nonlinear function, confidence envelop and 95% point-wise confidence interval for Model 2 with $n = 200$ and $\rho = 0.3$. 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 100 simulations. The bottom plot is a 95% pointwise confidence interval.

6. Proofs

For simplicity of notations, we use $\widehat{\boldsymbol{\beta}}$ and $\hat{f}$ to represent $\widehat{\boldsymbol{\beta}}_{PSA}$ and $\hat{f}_{PSA}$, respectively in the proofs. The key technical tool in deriving the nonparametric convergence rate results in Theorem 3.1 is the empirical processes. Hence, we first present the definition for the entropy and related calculations in the below.

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

$$H_B(\delta, \{\eta : \|\eta\|_\infty \leq C, J_\eta \leq C\}, \|\cdot\|_\infty) \leq M \left(\frac{C}{\delta}\right)^{1/k}, \tag{6.1}$$

$$H(\delta, \{\eta : \|\eta\|_\infty \leq C, J_\eta \leq C\}, \|\cdot\|_\infty) \leq M \left(\frac{C}{\delta}\right)^{1/k}, \tag{6.2}$$

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

$$P\left\{\inf_{\|\mathbf{s}\|=M} \Delta(\mathbf{s}) > 0\right\} \geq 1 - \epsilon, \tag{6.3}$$

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.

Note that the quadratic part of $Q(\boldsymbol{\beta})$ is denoted as $L(\boldsymbol{\beta})$. 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

$$\Delta(\mathbf{s}) \geq n^{-1/2}\mathbf{s}'\dot{L}(\boldsymbol{\beta}_0) + \frac{1}{2}\mathbf{s}'[n^{-1}\ddot{L}(\boldsymbol{\beta}_0)]\mathbf{s} + \lambda_2 \sum_{j=1}^{q} \frac{|\beta_{0j} + n^{-1/2}s_j| - |\beta_{0j}|}{|\tilde{\beta}_j|^\gamma}, \tag{6.4}$$

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.

We next figure out the stochastic order of $\dot{L}(\cdot)$ and $\ddot{L}(\cdot)$ at the true value $\boldsymbol{\beta}_0$. Based on the expression of (2.3), it is easy to 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 [18], we can show that

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

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

$$\dot{L}(\boldsymbol{\beta}_0) = O_P(n^{-1/2}), \tag{6.5}$$
$$\ddot{L}(\boldsymbol{\beta}_0) = O_P(1) \tag{6.6}$$

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 $g_0(x,t) = x'\boldsymbol{\beta}_0 + f_0(t)$, and $\hat{g}(x,t) = x'\hat{\boldsymbol{\beta}} + \hat{f}(t)$. Then, by the definition of $(\hat{\boldsymbol{\beta}}, \hat{f})$, we have

$$\|\hat{g} - g_0\|_n^2 + \lambda_1 J_{\hat{f}}^2 + \lambda_2 J_{\hat{\boldsymbol{\beta}}} \le \frac{2}{n}\sum_{i=1}^n \epsilon_i(\hat{g} - g_0)(X_i, t_i) + \lambda_1 J_{f_0}^2 + \lambda_2 J_{\boldsymbol{\beta}_0}, \tag{6.7}$$
$$\|\hat{g} - g_0\|_n^2 \le 2\|\epsilon\|_n\|\hat{g} - g_0\|_n + \lambda_1 J_{f_0}^2 + \lambda_2 J_{\boldsymbol{\beta}_0},$$
$$\|\hat{g} - g_0\|_n^2 \le \|\hat{g} - g_0\|_n O_P(1) + o_P(1),$$

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 $\|g_2(x,t)\|_\infty \le J_{g_2} = J_f$. 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 $\|\hat{g}_2\|_\infty \le J_{\hat{g}}$. We shall now show that $\|\hat{g}\|_\infty/(1 + J_{\hat{g}}) = O_P(1)$ via the above Sobolev decomposition. Then

$$\frac{\|\hat{g}_1\|_n}{1 + J_{\hat{g}}} \le \frac{\|\hat{g}\|_n}{1 + J_{\hat{g}}} + \frac{\|\hat{g}_2\|_n}{1 + J_{\hat{g}}} = O_P(1). \tag{6.8}$$

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, by the entropy calculation in the above, we know the bracketing entropy number for the below constructed class of functions:

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

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 [25] 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:

$$\begin{aligned}\lambda_1 J_{\hat{f}}^2 \le & \left[\|\hat{g}-g_0\|_n^{1-1/2m}(1+J_{\hat{f}})^{1/2m} \vee (1+J_{\hat{f}}) n^{-\frac{2m-1}{2(2m+1)}}\right] O_P(n^{-1/2}) \\ & + \lambda_1 J_{f_0}^2 + \lambda_2 (J_{\boldsymbol{\beta}_0} - J_{\hat{\boldsymbol{\beta}}}),\end{aligned} \tag{6.9}$$

and

$$\begin{aligned}\|\hat{g}-g_0\|_n^2 \le & \left[\|\hat{g}-g_0\|_n^{1-1/2m}(1+J_{\hat{f}})^{1/2m} \vee (1+J_{\hat{f}}) n^{-\frac{2m-1}{2(2m+1)}}\right] O_P(n^{-1/2}) \\ & + \lambda_1 J_{f_0}^2 + \lambda_2 (J_{\boldsymbol{\beta}_0} - J_{\hat{\boldsymbol{\beta}}}).\end{aligned} \tag{6.10}$$

Note that

$$\begin{aligned}\lambda_2 (J_{\boldsymbol{\beta}_0} - J_{\hat{\boldsymbol{\beta}}}) &\le \lambda_2 \sum_{j=1}^q \frac{|\beta_{0j} - \hat{\beta}_j|}{|\tilde{\beta}_j|^\gamma} + \lambda_2 \sum_{j=q+1}^d \frac{|\beta_{0j} - \hat{\beta}_j|}{|\tilde{\beta}_j|^\gamma} \\ &\le O_P(n^{-2m/(2m+1)}).\end{aligned} \tag{6.11}$$

(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

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

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

Proof of Theorem 3.2: For part (a), it suffices to show that

$$Q\{(\bar{\boldsymbol{\beta}}_1, \mathbf{0})\} = \min_{\|\bar{\boldsymbol{\beta}}_2\| \le C n^{-1/2}} Q\{(\bar{\boldsymbol{\beta}}_1, \bar{\boldsymbol{\beta}}_2)\} \text{ with probability approaching to 1} \tag{6.12}$$

for any $\bar{\boldsymbol{\beta}}_1$ satisfying $\|\bar{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\| = O_P(n^{-1/2})$ based on (3.3). In order 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 proof of Theorem 3.1. The assumption (3.6) implies that $\sqrt{n}\lambda_2/|\tilde{\beta}_j|^\gamma \to \infty$ for $j = q+1, \ldots, d$. Thus, we know 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 (a).

We next prove part (b). Similar as the proof in theorem 3.1, it is easy to 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 $L(\boldsymbol{\beta})$ given in (2.3) and the Slutsky's theorem, we have concluded the proof of part (b). □

Important Lemmas. We provide three useful matrix inequalities and two lemmas for preparing the proofs of Theorems 3.3 and 3.4. 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

$$|\mathbf{s}'\mathbf{A}\mathbf{w}| \le \|\mathbf{s}\|\|\mathbf{A}\|\|\mathbf{w}\| \tag{6.13}$$

$$|\mathbf{s}'\mathbf{B}\mathbf{z}| \le |\mathbf{s}'\mathbf{B}\mathbf{s}|^{1/2}|\mathbf{z}'\mathbf{B}\mathbf{z}|^{1/2} \tag{6.14}$$

$$|\mathbf{s}'\mathbf{z}| \le \|\mathbf{s}\|\|\mathbf{z}\| \tag{6.15}$$

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

Lemma 6.1. *Given that $\lambda_1 \to 0$, we have*

$$n^{-k/2}\sum_{l=1}^{n}|[(I-A)\mathbf{f}_0(t)]_l|^k = O(\lambda_1^{k/2}) \qquad \textit{for } k=2,3,\ldots \tag{6.16}$$

Proof: For the case of $k=2$, it has been proved in Lemma 2 of [18]. Next we apply the principle of mathematical induction to prove the cases for arbitrary $k>2$. We first assume that

$$n^{-(k-1)/2}\sum_{l=1}^{n}|[(I-A)\mathbf{f}_0(t)]_l|^{k-1} = O(\lambda_1^{(k-1)/2}) \tag{6.17}$$

for $k=3,4,\ldots$. Then we can write

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

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

Lemma 6.2. *Given that $d_n = n^{1/2}\wedge n\lambda_1^{1/2m}$, we have*

$$[\mathbf{X}'A(\lambda_1)\epsilon]_i = O_P(\lambda_1^{-1/4m}), \tag{6.18}$$

$$[\mathbf{X}'((I-A(\lambda_1))\mathbf{f}_0+\epsilon)]_i = O_P(n^{1/2}), \tag{6.19}$$

$$[\mathbf{X}'(I-A(\lambda_1))\mathbf{X}/n]_{ij} = R_{ij} + O_P(n^{-1/2}\vee n^{-1}\lambda_1^{-1/2m}), \tag{6.20}$$

$$\|\mathbf{X}'(I-A(\lambda_1))\mathbf{X}/n - R\| = o_P(1). \tag{6.21}$$

Proof: We first state the Lemma 4.1 and 4.3 in [7]:

$$n^{-1}\sum_j[(I-A)\mathbf{f}_0]_j^2 \le \lambda_1\int_0^1 (f_0^{(m)}(t))^2dt, \tag{6.22}$$

$$tr(A) = O(\lambda_1^{-1/2m}) \quad \text{and} \quad tr(A^2) = O(\lambda_1^{-1/2m}). \tag{6.23}$$

By the fact that $Var[(\mathbf{X}'A\epsilon)_i] = \sigma^2 R_{ii}tr(A^2)$, we can show that $[\mathbf{X}'A\epsilon]_i = O_P(\lambda_1^{-1/4m})$ 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 $Var(\sum_j W_{ij}) = R_{ii}\sigma^2 + R_{ii}n^{-1}\sum_j[(I-A)\mathbf{f}_0]_j^2$. By (6.22), we have $Var(\sum_j W_{ij}) \to R_{ii}\sigma^2$. We next verify the Liapounov's condition:

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

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 R_{ij}, $[\mathbf{X}'\mathbf{X}/n]_{ij} - R_{ij}$ and $[-\mathbf{X}'A\mathbf{X}/n]_{ij}$. By the central limit theorem, we know the second term in the above decomposition is $O_P(n^{-1/2})$. For the last term, we have $E\{(\mathbf{X}'A\mathbf{X})_{ij}\}^2 =$

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

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

$$\|\mathbf{X}'(I-A)\mathbf{X}/n - R\| = O_P(d_n n^{-1/2} \vee d_n n^{-1}\lambda_1^{-1/2m}). \tag{6.24}$$

Thus (6.21) follows from the dimension condition D1. □

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

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

Let $\delta_n = n^{-1/2}[\mathbf{X}'(I-A)\mathbf{X}]^{1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ and $\omega_n = n^{-1/2}[\mathbf{X}'(I-A)\mathbf{X}]^{-1/2}\mathbf{X}'(I-A)(\mathbf{f}_0 + \epsilon)$. 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 $\|\omega_n\|^2 = K_{1n} + K_{2n} + K_{3n}$, where

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

Applying (6.18), (6.19) and (6.20) to the above three terms, we can conclude that all of them are of the order $O_P(d_n n^{-1})$ by considering the matrix inequalities (6.13)-(6.15). Thus we have proved (3.7) by considering (6.21). □

Proof of Theorem 3.3: We use notations in the proof of Theorem 3.1. Let $\alpha_n = \sqrt{d_n/n}$. Similar as (6.4), we have

$$Q(\boldsymbol{\beta}_0 + \alpha_n \mathbf{s}) - Q(\boldsymbol{\beta}_0) \geq \alpha_n \mathbf{s}' \dot{L}(\boldsymbol{\beta}_0) + \frac{1}{2}\mathbf{s}'[\alpha_n^2 \ddot{L}(\boldsymbol{\beta}_0)]\mathbf{s} + \lambda_2 \sum_{j=1}^{q_n} \frac{|\beta_{0j} + \alpha_n s_j| - |\beta_{0j}|}{|\tilde{\beta}_j|^\gamma}, \quad (6.25)$$

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

$$\alpha_n \mathbf{s}' \dot{L}(\boldsymbol{\beta}_0) = \|\mathbf{s}\| O_P(d_n/n) \quad (6.26)$$

$$\frac{1}{2}\mathbf{s}'[\alpha_n^2 \ddot{L}(\boldsymbol{\beta}_0)]\mathbf{s} = (d_n/n)\mathbf{s}' R \mathbf{s} + O_P(d_n^2 n^{-3/2} \vee d_n^2 n^{-2} \lambda_1^{-1/2m}) \quad (6.27)$$

given any $\|\mathbf{s}\| = C$ independent of n. Thus the first two terms in the right hand side of (6.25) is of the order $O_P(d_n/n)$ by the condition that $d_n = o(n^{1/2} \wedge n\lambda_1^{1/2m})$. The second term, which is positive, dominates the first one by allowing sufficiently large C. Note that 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 of the lower order than $O_P(d_n/n)$. This completes the proof of (3.8).

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 $\lambda_{min}(\sum_k \phi_k \phi_k') \geq c_3 > 0$, 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 [25], we have established the below inequalities:

$$\begin{aligned} \lambda_1 J_{\hat{f}}^2 \leq & \left[\|\hat{g} - g_0\|_n^{1-1/2m}(1 + J_{\hat{f}})^{1/2m} q_n^{1/2m} \vee (1 + J_{\hat{f}}) q_n n^{-\frac{2m-1}{2(2m+1)}}\right] O_P(n^{-1/2}) \\ & + \lambda_1 J_{f_0}^2 + \lambda_2 (J_{\boldsymbol{\beta}_0} - J_{\widehat{\boldsymbol{\beta}}}), \end{aligned} \quad (6.28)$$

and

$$\begin{aligned} \|\hat{g} - g_0\|_n^2 \leq & \left[\|\hat{g} - g_0\|_n^{1-1/2m}(1 + J_{\hat{f}})^{1/2m} q_n^{1/2m} \vee (1 + J_{\hat{f}}) q_n n^{-\frac{2m-1}{2(2m+1)}}\right] O_P(n^{-1/2}) \\ & + \lambda_1 J_{f_0}^2 + \lambda_2 (J_{\boldsymbol{\beta}_0} - J_{\widehat{\boldsymbol{\beta}}}). \end{aligned} \quad (6.29)$$

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

$$\begin{aligned} a_n^2 &\leq O_P(n^{-1/2})a_n^{1-1/2m} \vee O_P(n^{-2m/(2m+1)}) \vee O_P(\lambda_1/q_n) \vee \frac{\lambda_2(J_{\boldsymbol{\beta}_0} - J_{\widehat{\boldsymbol{\beta}}})}{q_n} \\ &\leq O_P(n^{-1/2})a_n^{1-1/2m} \vee O_P(n^{-2m/(2m+1)}) \vee \frac{\lambda_2(J_{\boldsymbol{\beta}_0} - J_{\widehat{\boldsymbol{\beta}}})}{q_n} \\ &\leq O_P(n^{-1/2})a_n^{1-1/2m} \vee O_P(n^{-2m/(2m+1)}) \end{aligned} \tag{6.30}$$

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 (3.9). 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 (3.10). □

Proof of Theorem 3.4: The proof of part (a) is similar as that in the fixed dimensional case, i.e. part (a) in Theorem 3.2. It follows from the regular condition that $\lambda_1/q_n \asymp n^{-2m/(2m+1)}$, Lemma 6.2 and the assumption (3.12).

We next prove the asymptotic normality of $\widehat{\boldsymbol{\beta}}_1$. Similar as the proof for the part (b) in Theorem 3.2, we can establish that

$$\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} = [\mathbf{X}_1'(I - A)\mathbf{X}_1]^{-1} \left[\mathbf{X}_1'(I - A)(\mathbf{f}_0(t) + \epsilon) - \frac{n\lambda_2}{2} Pe(\widehat{\boldsymbol{\beta}}_1)\right], \tag{6.31}$$

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 R, i.e. the condition R3D. Thus, we have

$$\begin{aligned} & \sqrt{n}\mathbf{G}_n R_{11}^{-1/2}(\mathbf{X}_1'(I-A)\mathbf{X}_1/n)(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \\ = \; & \sqrt{n}\mathbf{G}_n R_{11}^{-1/2}\left[\frac{\mathbf{X}_1'(I-A)(\mathbf{f}_0(t)+\epsilon)}{n} - \frac{\lambda_2}{2}Pe(\widehat{\boldsymbol{\beta}}_1)\right] \\ = \; & M_{1n} + M_{2n} + M_{3n}, \end{aligned} \tag{6.32}$$

where

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

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

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

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

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

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

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

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

and apply the Lindeberg's theorem (see Theorem 1.15 in [32]) to show its asymptotic distribution. First, we have

$$Var(\sum_j T_j) = \sum_j Var(T_j) = \mathbf{v}'\mathbf{G}_n\mathbf{G}_n'\mathbf{v}(\sigma^2 + n^{-1}\sum_{l=1}^n((I-A)\mathbf{f}_0)_l^2) \to \sigma^2\mathbf{v}'\mathbf{G}\mathbf{v} \tag{6.33}$$

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

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

In view of (6.33), we obtain

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

and

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

Note that

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

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 $q_n \le d_n = o(n^{1/3})$ 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

$$\begin{aligned}\sqrt{n}\mathbf{G}_n R_{11}^{1/2}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) &= \sqrt{n}\mathbf{G}_n R_{11}^{-1/2}(R_{11} - \mathbf{X}_1'(I-A)\mathbf{X}_1/n)(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) + N(0,\sigma^2\mathbf{G}) + o_P(1) \\ &= N(0,\sigma^2\mathbf{G}) + o(1) + O_P(d_n^{3/2}n^{-1/2} \vee d_n^{3/2}n^{-1}\lambda_1^{-1/2m}) \qquad (6.34)\end{aligned}$$

by the matrix inequality, (6.24) and (3.8). 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 (3.13). □

LITERATURE CITED

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

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

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

[4] Bunea, F. (2004) Consistent Covariate Selection and Post Model Selection Inference in Semiparametric Regression, 2004, *Annals of Statistics* **32**, 898-927.

[5] Bunea, F. and Wegkamp, M. (2004) Two-Stage Model Selection Procedures in Partially Linear Regression, *The Canadian Journal of Statistics* **32**, 105-118.

[6] Chen, H. (1988). Convergence Rates for Parametric Components in a Partly Linear Model. *The Annals of Statistics* **16**, 136-146.

[7] 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.

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

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

[10] Engle, R.F., Granger, C.W.J., Rice, J., Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* **81**, 310-320.

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

[12] 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.

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

[14] 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.

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

[16] Green, P.J., Jennison, C. and Seheult, A. (1985). Analysis of field experiments by least squares. *Journal of Royal Statistical Society B* **47**, 299-315.

[17] Hardle, W., Liang, H. and Gao, J.T. (2000) Partially Linear Models. New York: Springer-Verlag.

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

[19] 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.

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

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

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

[23] Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Annals of Statistics* **36**, 261-286.

[24] Li, Q. (2000). Efficient Estimation of additive partially linear models. *International Economic Review* **41**, 1073-1092.

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

[26] Mallows, C. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

[27] 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.

[28] Portnoy, S. (1985). Asymptotic Behavior of M-Estimator of p Regression Parameters when p^2/n is large. II. Normal Approximation. *Annals of Statistics* **12**, 1298-1309.

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

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

[31] Schwarz, G. (1978) Estimating the Dimension of a Model. *Annals of Statistics* **6**, 461-464.

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

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

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

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

[36] 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.

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

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

[39] 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.

[40] 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* To appear.

[41] Xie, H. and Huang, J. (2009). SCAD-Penalized Regression in High-Dimensional Partially Linear Models. *Annals of Statistics* **37**, 673-696

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

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

[44] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of American Statistical Association* **101**, 1418-1429.

[45] Zou, H. and Zhang, H. H. (2008). On The Adaptive Elastic-Net With A Diverging

Number of Parameters. *Annals of Statistics* To Appear