SCAD-Penalized Regression in High-Dimensional Partially Linear Models

Huiliang Xie and Jian Huang

University of Iowa

July 2006

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 367

Summary. We consider the problem of simultaneous variable selection and estimation in partially linear models with a divergent number of covariates in the linear part, under the assumption that the vector of regression coefficients is sparse. We apply the SCAD penalty to achieve sparsity in the linear part and use polynomial splines to estimate the nonparametric component. Under reasonable conditions it is shown that consistency in terms of variable selection and estimation can be achieved simultaneously for the linear and nonparametric components. Furthermore, the SCAD-penalized estimators of the nonzero coefficients are shown to be asymptotically normal with the same means and covariances that they would have if the zero coefficients were known in advance. Simulation studies are conducted to evaluate the finite sample behavior of the SCAD-penalized estimators.

Key Words and phrases. Asymptotic normality, high-dimensional data, oracle property, penalized estimation, semiparametric models, variable selection, .

Short title. High-dimensional PLM

AMS 2000 subject classification. Primary 62J05, 62G08; secondary 62E20

1. Introduction. Consider a partially linear model (PLM)

$$Y = \mathbf{X}'\boldsymbol{\beta} + g(T) + \varepsilon,$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients associated with $\mathbf{X}$, and g is an unknown function of T. In this model, the mean response is linearly related to $\mathbf{X}$, while its relation with T is not specified up to any finite number of parameters. This model combines the flexibility of nonparametric regression and parsimony of linear regression. When the relation between Y and $\mathbf{X}$ is of main interest and can be approximated by a linear function, it offers more interpretability than a purely nonparametric model.

We consider the problem of simultaneous variable selection and estimation in the PLM when p is large in the sense that $p \to \infty$ as the sample size $n \to \infty$. For finite-dimensional $\boldsymbol{\beta}$, several approaches have been proposed to estimate $\boldsymbol{\beta}$ and g. Examples include the partially spline estimator [Wahba (1984), Engle, Granger, Rice and Weiss (1986) and Heckman (1986)] and the partially residual estimator [Robinson (1988), Speckman (1988) and Chen (1988)]. A detailed treatment of the PLM can be found in Härdel, Liang and Gao (2000). Under appropriate assumptions about the smoothness of g and the structure of $\mathbf{X}$, these estimators of $\boldsymbol{\beta}$ were shown to be $\sqrt{n}$-consistent and asymptotically normal. It was also shown that the estimators of g can converge at the optimal rate in the purely nonparametric regression determined in Stone (1980, 1982). In these studies, the dimension of the covariate vector $\mathbf{X}$ was fixed and the problem of variable selection in $\mathbf{X}$ via penalization was not considered. The PLM is a basic and one of the most studied semiparametric models. In addition to the work on the PLM, there has also been extensive research on efficient estimation in a large class of semiparametric models, see for example, Bickel, Klaassen, Ritov and Wellner (1998) and the references cited therein. However, the results for the PLM with a finite-dimensional β and those for the semiparametric models in general are not applicable to the PLM with a divergent number of covairates. Indeed, it appears that there is no systematic theoretical investigation of estimation in semiparametric models with a high-dimensional parametric component.

We are particularly interested in $\boldsymbol{\beta}$ when it is sparse, in the sense that many of its elements are zero. Our work is motivated from biomedical studies that investigate the relationship between a phenotype of interest and genomic measurements such as microarray data. In many such studies, in addition to genomic measurements, other types of measurements such as clinical or environmental covariates are also available. To obtain unbiased estimates of genomic effects, it is necessary to take into account these covariates. Assuming a sparse model is often reasonable with genomic data. This is because although the total number of measurements can be large, the number of important ones is usually relatively small. In these problems, selection of important covariates is often one of the most important goals in the analysis.

We use the SCAD method to achieve simultaneous consistent variable selection and estimation of $\boldsymbol{\beta}$. The SCAD method is proposed by Fan and Li (2001) in a general parametric framework for variable selection and efficient estimation. This method uses a specially designed penalty function, the smoothly clipped absolute deviation (hence the name SCAD). We estimate the nonparametric component g using the partial residual method with the B-spline bases. The resulting estimator of $\boldsymbol{\beta}$ maintains the oracle property of the SCAD-penalized estimators in parametric settings. Here the oracle property means that the estimator can correctly select the nonzero coefficients with probability converging to one and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariances that they would have if the zero coefficients were known in advance. Therefore, an oracle estimator is asymptotically as efficient as the ideal estimator assisted by an oracle who knows which coefficients are nonzero. For the nonparametric component, convergence of the estimator of g in the SCAD-penalized partially linear regression still reaches the optimal global rate.

There have been several investigations on asymptotic properties of penalized estimation in parametric models. Knight and Fu (2000) studied the asymptotic distributions of bridge estimators when the number of covariates is fixed. Fan and Li (2001) studied asymptotic properties of SCAD penalized likelihood methods when the number of parameters is finite. Fan and Peng (2004) considered the same problem when the number of parameters diverges. Under certain regularity conditions, they showed that there exist local maximizers of the penalized likelihood that have an oracle property. Huang, Horowitz and Ma (2006) studied the bridge estimators with a divergent number of covariates in a linear regression model. They showed that the bridge estimators have an oracle property under appropriate conditions if the bridge index is strictly between 0 and 1. Several earlier studies have investigated the properties of regression estimators with a divergent number of covariates. See, for example, Huber (1981) and Portnoy (1984, 1985). Portnoy proved consistency and asymptotic normality of a class of M-estimators of regression parameters under appropriate conditions. However, he did not consider penalized regression or selection of variables in sparse models.

The rest of this article is organized as follows. In Section 2, we define the SCAD-penalized estimator $(\widehat{\boldsymbol{\beta}}_n, \widehat{g}_n)$ in the PLM, abbreviated as SCAD-PLM estimator hereafter. The main results for the SCAD-PLM estimator are given in Section 3, including the consistency and oracle property of $\widehat{\boldsymbol{\beta}}_n$ as well as the rate of convergence of $\widehat{g}_n$. Section 4 is describes an algorithm for computing the SCAD-PLM estimator and the criterion to choose the penalty parameter. Section 5 offers simulation studies that illustrate the finite sample bevavior of this estimator. Some concluding remarks are given in Section 6. The proofs are relegated to the Appendix.

2. Penalized estimation in PLM with the SCAD penalty. To make it explicit that the covariates and regression coefficients depend on n, we write the PLM

$$Y_i = \mathbf{X}_i^{(n)\prime}\boldsymbol{\beta}^{(n)} + g(T_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $(\mathbf{X}_i^{(n)}, T_i, Y_i)$ are independent and identically distributed as $(\mathbf{X}^{(n)}, T, Y)$ and ε_i is independent of $(\mathbf{X}_i^{(n)}, T_i)$ with mean 0 and variance σ^2. We assume that T takes values in a compact interval, and for simplicity, we assume this interval to be $[0, 1]$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ and let $\mathbb{X}^{(n)} = (X_{ij}, 1 \leq i \leq n, 1 \leq j \leq p_n)$ be the $n \times p_n$ design matrix associated with $\boldsymbol{\beta}^{(n)}$. In sparse models, the p_n covariates can be classified into two categories: the important ones whose corresponding coefficients are nonzero and the trivial ones that actually are not present in the underlying model. For convenience of notation, we write

$$\boldsymbol{\beta}^{(n)} = (\boldsymbol{\beta}_1^{(n)\prime}, \boldsymbol{\beta}_2^{(n)\prime})', \tag{1}$$

where $\boldsymbol{\beta}_1^{(n)\prime} = (\beta_1^{(n)}, \ldots, \beta_{k_n}^{(n)})$ and $\boldsymbol{\beta}_2^{(n)\prime} = (0, \ldots, 0)$. Here $k_n (\leq p_n)$ is the number of nontrivial covariates. Let $m_n = p_n - k_n$ be the number of zero coefficients.

We use the polynomial splines to approximate g. For a positive integer M_n, let $\Delta_n = \{\xi_{n\nu}\}_{\nu=1}^{M_n}$ be a partition of $[0, 1]$ into $M_n + 1$ subintervals $I_{n\nu} = [\xi_{n\nu}, \xi_{n,\nu+1}) : \nu = 0, \ldots, M_n - 1$ and $I_{nM_n} = [\xi_{nM_n}, 1]$. Here $\xi_{n0} = 0$ and $\xi_{n,M_n+1} = 1$. Denote the largest mesh size of Δ_n, $\max_{0 \leq \nu \leq M_n}\{\xi_{n,\nu+1} - \xi_{n\nu}\}$, by $\overline{\Delta}_n$. Throughout the article we assume $\overline{\Delta}_n = O(M_n^{-1})$. Let $\mathcal{S}_m(\Delta_n)$

be the space of polynomial splines of order m with simple knots at the points $\xi_{n1}, \ldots, \xi_{nM_n}$. This space consists of all functions s with these two properties:

(i) Restricted to any interval $I_{n\nu}(0 \leq \nu \leq M_n)$, s is a polynomial of order m;

(ii) If $m \geq 2$, s is $m-2$ times continuously differentiable on $[0,1]$.

According to Corollary 4.10 in Schumaker (1981), there is a local basis $\{B_{nw}, 1 \leq w \leq q_n\}$ for $\mathcal{S}_m(\Delta_n)$, where $q_n = M_n + m$ is the dimension of $\mathcal{S}_m(\Delta_n)$. Let

$$\mathbf{Z}(t; \Delta_n)' = (B_{n1}(t), \ldots, B_{nq_n}(t))$$

and $\mathbb{Z}^{(n)}$ be the $n \times q_n$ matrix whose ith row is $\mathbf{Z}(T_i; \Delta_n)'$. Any $s \in \mathcal{S}_m(\Delta_n)$ can be written $s(t) = \mathbf{Z}(t; \Delta_n)'\mathbf{a}^{(n)}$ for a $q_n \times 1$ vector $\mathbf{a}^{(n)}$. We try to find the s in $\mathcal{S}_m(\Delta_n)$ that is close to g. Under reasonable smoothness conditions, g can be well approximated by elements in $\mathcal{S}$. Thus the problem of estimating g becomes that of estimating $\mathbf{a}^{(n)}$.

Given $a > 2$ and $\lambda > 0$, the SCAD penalty at θ is

$$p_\lambda(\theta; a) = \begin{cases} \lambda|\theta| & , \quad |\theta| \leq \lambda, \\ -(\theta^2 - 2a\lambda|\theta| + \lambda^2)/[2(a-1)], & \lambda < |\theta| \leq a\lambda, \\ (a+1)\lambda^2/2 & , \quad |\theta| > a\lambda. \end{cases}$$

More insight into it can be gain through its first derivative:

$$p'_\lambda(\theta; a) = \begin{cases} \mathrm{sgn}(\theta)\lambda & , \quad |\theta| \leq \lambda, \\ \mathrm{sgn}(\theta)(a\lambda - |\theta|)/(a-1), & \lambda < |\theta| \leq a\lambda, \\ 0 & , \quad |\theta| > a\lambda. \end{cases}$$

The SCAD penalty is continuously differentiable on $(-\infty, 0) \cup (0, \infty)$, but singular at 0. Its derivative vanishes outside $[-a\lambda, a\lambda]$. As a consequence, SCAD penalized regression can produce sparse solutions and unbiased estimates for large coefficients. More details of the penalty can be found in Fan and Li (2001).

The penalized least squares objective function for estimating $\boldsymbol{\beta}^{(n)}$ and $\mathbf{a}^{(n)}$ with the SCAD penalty is

$$Q_n(\mathbf{b}^{(n)}, \mathbf{a}^{(n)}; \lambda_n, a, \Delta_n, m) = \|\mathbf{Y} - \mathbb{X}^{(n)}\mathbf{b}^{(n)} - \mathbb{Z}^{(n)}\mathbf{a}^{(n)}\|^2 + n\sum_{j=1}^{p_n} p_{\lambda_n}(b_j^{(n)}; a). \tag{2}$$

Let

$$(\widehat{\boldsymbol{\beta}}_n^{(n)}, \widehat{\boldsymbol{\alpha}}_n^{(n)}) = \arg\min Q_n(\mathbf{b}^{(n)}, \mathbf{a}^{(n)}; \lambda_n, a, \Delta_n, m).$$

The SCAD-PLM estimators of $\boldsymbol{\beta}$ and g are $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{g}_n(t) \equiv \mathbf{Z}(t; \Delta_n)'\widehat{\boldsymbol{\alpha}}_n^{(n)}$, respectively.

The polynomial splines were also used by Huang (1999) in the partially linear Cox models. Some computational conveniences were also discussed there. We limit our search for the estimate of g to the space of polynomial splines of order m instead of the larger space of piecewise polynomials of order m, with the goal to find a smooth estimator of g. Unlike the basis pursuit in nonparametric regression, no penalty is imposed on the estimator of the nonparametric part as our interest lies in the variable selection with regard to the parametric part.

For any $\mathbf{b}^{(n)}$, the $\mathbf{a}^{(n)}$ that minimizes Q_n necessarily satisfies

$$\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)}\mathbf{a}^{(n)} = \mathbb{Z}^{(n)\prime}(\mathbf{Y} - \mathbb{X}^{(n)\prime}\mathbf{b}^{(n)}).$$

Let $P_{\mathbf{Z}} = \mathbb{Z}^{(n)}(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}$ be the projection matrix of the column space of $\mathbb{Z}^{(n)}$. The profile objective function of the parametric part becomes

$$\widetilde{Q}_n(\mathbf{b}^{(n)}; \lambda_n, a, \Delta_n, m) = \|(I - P_{\mathbf{Z}}^{(n)})(\mathbf{Y} - \mathbb{X}^{(n)}\mathbf{b}^{(n)})\|^2 + n\sum_{j=1}^{p_n} p_{\lambda_n}(b_j^{(n)}; a). \tag{3}$$

Then

$$\widehat{\boldsymbol{\beta}}_n^{(n)} = \arg\min Q_n(\mathbf{b}^{(n)}; \lambda_n, a, \Delta_n, m).$$

Because the profile objective function does not involve $\mathbf{a}^{(n)}$ and has an explicit form, it is useful for both theoretical investigation and computation. We will use it to established the asymptotic properties of $\widehat{\boldsymbol{\beta}}_n^{(n)}$. Computationally, this expression can be used to first obtain $\widehat{\boldsymbol{\beta}}_n^{(n)}$. Then $\widehat{\mathbf{a}}_n^{(n)}$ can be computed using the resulting residuals as the response for the covariate matrix $\mathbb{Z}^{(n)}$.

3. Asymptotic properties of the SCAD-PLM estimator. In this section we state the results of the asymptotic properties of the SCAD-PLM estimator. First, we define some notations. Let $\theta_j^{(n)}(t) = E[X_j^{(n)} | T = t]$ for $j = 1, \ldots, p_n$. Let the $p_n \times p_n$ conditional variance-covariance matrix of $(\mathbf{X}^{(n)} | T = t)$ be $\Sigma^{(n)}(t)$. Let $\mathbf{e}^{(n)} = \mathbf{X}^{(n)} - E\left[\mathbf{X}^{(n)} | T\right]$. We can write $\Sigma^{(n)}(t) = \text{Var}(\mathbf{e}^{(n)} | T = t)$. Denote the unconditional variance-covariance matrix of $\mathbf{e}^{(n)}$ by $\Xi^{(n)}$. We have $\Xi^{(n)} = E[\Sigma^{(n)}(T)]$. We assume the following conditions on the smoothness of g and $\theta_j^{(n)}, 1 \leq j \leq p_n$.

Condition 1. There are absolute constants $\gamma_\theta > 0$ and $M_\theta > 0$ such that

$$\sup_{n \geq 1} \sup_{1 \leq j \leq p_n} |\theta_{nj}^{(r_\theta)}(t_2) - \theta_{nj}^{(r_\theta)}(t_1)| \leq M_\theta |t_2 - t_1|^{\gamma_\theta}, \quad \text{for } 0 \leq t_1, t_2 \leq 1,$$

and the degree of the polynomial spline $m - 1 \geq r_\theta$. Let $s_\theta = r_\theta + \gamma_\theta$.

Condition 2. There exists an absolute constant σ_{4e} such that for all n and $1 \leq j \leq p_n$,

$$E[e_j^{(n)^4} | T] \leq \sigma_{4e}, \text{ almost surely.}$$

Condition 3. There are absolute constants $\gamma_g > 0$ and $M_g > 0$ such that

$$|g^{(r_g)}(t_2) - g^{(r_g)}(t_1)| \leq M_g |t_2 - t_1|^{\gamma_g}, \quad \text{for } 0 \leq t_1, t_2 \leq 1,$$

with $r_g \leq m - 1$. Let $s_g = r_g + \gamma_g$.

As in nonparametric regression, we allow $M_n \to \infty$ but $M_n = o(n)$. In addition, we assume that the tuning parameter $\lambda_n \to 0$ as $n \to \infty$. This is the assumption adopted in nonconcave penalized regression (Fan and Peng 2004). For convenience, all the other conditions required for the conclusions in this section are listed here.

(A1) (a) $\lim_{n\to\infty} p_n^2/n = 0$; (b) $\lim_{n\to\infty} p_n^2 M_n^2/n^2 = 0$; (c) $\lim_{n\to\infty} p_n/M_n^{s_\theta} = 0$.

(A2) The smallest eigenvalue of $\Xi^{(n)}$, denoted by $\lambda_{\min}(\Xi^{(n)})$, satisfies

$$\liminf_{n\to\infty} \lambda_{\min}(\Xi^{(n)}) = c_\lambda > 0.$$

(A3) $\lambda_n = o(k_n^{-1/2})$.

(A4) $\liminf_{n\to\infty} \min_{1\le j\le k_n} |\beta_j^{(n)}| = c_\beta > 0$.

(A5) Let $\lambda_{\max}(\Xi^{(n)})$ be the largest eigenvalue of $\Xi^{(n)}$. (a) $\lim \sqrt{p_n \lambda_{\max}(\Xi^{(n)})}/(\sqrt{n}\lambda_n) = 0$; (b) $\lim \sqrt{p_n \lambda_{\max}(\Xi^{(n)})}/(M_n^{s_g}\lambda_n) = 0$.

(A6) Suppose for all t in $[0,1]$, $\text{tr}(\Sigma_{11}^{(n)}(t)) \le \text{tr}(\Sigma_{u,11}^{(n)})$ and the latter satisfies $\lim \sqrt{\text{tr}(\Sigma_{u,11}^{(n)})} M_n^{-s_g} = 0$ and $\lim \text{tr}(\Sigma_{u,11}^{(n)}) M_n/n = 0$.

(A7) $\lim \sqrt{n} M_n^{-(s_g+s_\theta)} = 0$.

Theorem 1. *(Consistency of $\widehat{\boldsymbol{\beta}}^{(n)}$) Under (A1)–(A2),*

$$\|\widehat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^{(n)}\| = O_P(\sqrt{p_n/n} + M_n^{-s_g} + \sqrt{k_n}\lambda_n).$$

Thus under (A1)–(A3),

$$\|\widehat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^{(n)}\| \xrightarrow{P} 0.$$

This theorem establishes the consistency of the SCAD-PLM estimator of the parametric part. (A1) requires the number of covariates considered not to increase at rates faster than $\sqrt{n}$ and M_n^{1/s_θ}. (A2) is a requirement for model identifiability. It assumes that $\Xi^{(n)}$ is positive definite so that no random variable of the form $\sum_{j=1}^{p_n} c_j X_j^{(n)}$, where c_j's are constants, can be functionally related to T. When p_n increases with n, $\Xi^{(n)}$ needs to be bounded away from any singular matrix. The assumption about λ_n, (A3), says that λ_n should converge to 0 fast enough so that the penalty would not introduce any bias. The rate at which λ_n goes to 0 only depends on k_n. It is interesting to note that the smoothness index s_g of g and the number of spline bases M_n affects the rate of convergence of $\widehat{\boldsymbol{\beta}}^{(n)}$ by contributing a term $M_n^{-s_g}$. When p_n is bounded and no SCAD penalty is imposed ($\lambda_n = 0$), the convergence rate is $O(n^{-1/2} + M_n^{-s_g})$, which is consistent with Theorem 2 of Chen (1988).

Corresponding to the partition in (1), write $\widehat{\boldsymbol{\beta}}^{(n)} = (\widehat{\boldsymbol{\beta}}_1^{(n)\prime}, \widehat{\boldsymbol{\beta}}_2^{(n)\prime})'$, where $\widehat{\boldsymbol{\beta}}_1^{(n)\prime}$ and $\widehat{\boldsymbol{\beta}}_2^{(n)\prime}$ are vectors of length k_n and m_n, respectively. The theorem below shows that all the covariates with zero coefficients can be detected simultaneously with probability tending to 1, provided that λ_n does not converge to 0 too fast.

Theorem 2. *(Variable selection in $\mathbf{X}^{(n)}$) Assume all the $e_j^{(n)}$'s support sets are contained in a compact set in $\mathcal{R}$. Under (A1)–(A5)*

$$\lim_{n\to\infty} P(\widehat{\boldsymbol{\beta}}_2^{(n)} = \mathbf{0}) = 1.$$

(A5) puts restriction on the largest eigenvalue of $\Xi^{(n)}$). In general, $\lambda_{\max}(\Xi^{(n)}) = O(p_n)$, as can be seen from

$$\lambda_{\max}(\Xi^{(n)}) < \text{tr}(\Xi^{(n)}) \leq p_n\sqrt{\sigma_{4e}}.$$

There is the question of whether there exists a λ_n that satisfies both (A3) and (A5). It can be checked that, if $p_n = o(n^{1/3})$ there exists λ_n such that (A3) and (A5) hold. When k_n is bounded, the existence of such λ_n only requires that $p_n = o(n^{1/2})$. This relaxation also holds for the case when $\lambda_{\max}(\Xi^{(n)})$ is bounded from above.

By Theorem 2, $\widehat{\boldsymbol{\beta}}_2^{(n)}$ degenerates at $\mathbf{0}_{m_n}$ with probability converging to 1. We now consider the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_1^{(n)}$. According to the partition of $\boldsymbol{\beta}^{(n)}$ in (1), write $\mathbb{X}^{(n)}$ and $\Xi^{(n)}$ in the block form:

$$\mathbb{X}^{(n)} = (\underbrace{\mathbb{X}_1^{(n)}}_{n\times k_n} \underbrace{\mathbb{X}_2^{(n)}}_{n\times m_n}),$$

$$\Xi^{(n)} = \begin{array}{c c} & \begin{array}{cc} k_n & m_n \end{array} \\ \begin{array}{c} k_n \\ m_n \end{array} & \begin{pmatrix} \Xi_{11}^{(n)} & \Xi_{12}^{(n)} \\ \Xi_{21}^{(n)} & \Xi_{22}^{(n)} \end{pmatrix} \end{array}.$$

Let $\mathbf{A}_n$ be a non-random $\iota \times k_n$ matrix with full row rank and

$$\Sigma_n = n^2 \mathbf{A}_n \left[\mathbb{X}_1^{(n)\prime}(I - P_{\mathbf{Z}})\mathbb{X}_1^{(n)}\right]^{-1} \Xi_{11}^{(n)} \left[\mathbb{X}_1^{(n)\prime}(I - P_{\mathbf{Z}})\mathbb{X}_1^{(n)}\right]^{-1} \mathbf{A}_n'.$$

Theorem 3. *(Asymptotic distribution of $\widehat{\boldsymbol{\beta}}^{(n)}$) Suppose that all the support sets of $e_j^{(n)}$'s are contained in a compact set in $\mathcal{R}$, $j = 1, \ldots, p_n$. Then under (A1)–(A7),*

$$\sqrt{n}\Sigma_n^{-1/2}\mathbf{A}_n(\widehat{\boldsymbol{\beta}}_1^{(n)} - \boldsymbol{\beta}_1^{(n)}) \xrightarrow{d} N(\mathbf{0}_\iota, \sigma^2\mathbf{I}_\iota). \tag{4}$$

The asymptotic distribution result can be used to construct asymptotic confidence intervals for any fixed number of coefficients simultaneously.

In (4), we used the inverse of $\mathbb{X}_1^{(n)\prime}(I - P_{\mathbf{Z}})\mathbb{X}_1^{(n)}$ and that of Σ_n. Under assumption (A2), by Theorem 4.3.1 in Wang and Jia (1993), the smallest eigenvalue of $\Xi_{11}^{(n)}$ is no less than c_λ and bounded away from 0. By Lemma 1 in the Appendix, $\mathbb{X}_1^{(n)\prime}(I - P_{\mathbf{Z}})\mathbb{X}_1^{(n)}$ is invertible with probability tending to 1. The invertibility of Σ_n then follows from the full row rank restriction on $\mathbf{A}_n$.

(A6) may appear a little abrupt. It requires $\sum_{j=1}^{k_n}\text{Var}(e_j^{(n)}|\,T = t)$ to be less than the trace of a $k_n \times k_n$ matrix $\Sigma_{u,11}^{(n)}$ as t ranges over $[0, 1]$, which is considerably weaker than the assumption that $\Sigma_{u,11}^{(n)} - \Sigma_{11}^{(n)}(t)$ is a nonnegative definite matrix for any $t \in [0, 1]$. We can also replace tr$(\Sigma_{u,11}^{(n)})$ by k_n in the assumption, since for all t,

$$\sum_{j=1}^{k_n}\text{Var}(e_j^{(n)}|\,T = t) \le k_n\sqrt{C_e}.$$

(A7) requires that g and $\theta_j^{(n)}$ be smooth enough. Intuitively, a smooth g makes it easier to estimate $\boldsymbol{\beta}$. The smoothness requirement on $\theta_j^{(n)}$ also makes sense, since this helps to remove effect of T on $X_j^{(n)}$ and the estimation of $\boldsymbol{\beta}$ is based on the relationship

$$Y - E[Y|\,T] = (\mathbf{X} - E[\mathbf{X}|\,T])\,\beta + \varepsilon.$$

We now consider the consistency of $\widehat{g}_n$. Suppose that T is an absolutely continuous random variable on $[0, 1]$ with density f_T. We use the L_2 distance

$$\|\widehat{g}_n - g\|_T = \left\{\int_0^1 [\widehat{g}_n(t) - g(t)]^2 f_T(t)\,dt\right\}^{1/2}.$$

This is the measure of distance between two functions that were used in Stone (1982, 1985). If our interest is confined to the estimation of $\boldsymbol{\beta}^{(n)}$, we should choose large M_n unless computing comes into consideration. However, too large an M_n would introduce too much variation and is detrimental to the estimation of g.

Theorem 4. *(Rate of convergence of $\widehat{g}_n$) Suppose $M_n = o(\sqrt{n})$ and $f_T(t)$ is bounded away from 0 and infinity on $[0,1]$. Under (A1)–(A5),*

$$\|\widehat{g}_n - g\|_T = O_P(k_n/\sqrt{n} + \sqrt{M_n/n} + \sqrt{k_n}M_n^{-s_g}).$$

In the special case of bounded k_n, Theorem 4 simplifies to the well-known result in nonparametric regression:

$$\|\widehat{g}_n - g\|_T = O_P(\sqrt{M_n/n} + M_n^{-s_g}).$$

When $M_n \sim n^{-1/(2s_g+1)}$, the convergence rate is optimal. However, the feasibility of such a choice requires $s_g > 1/2$. To have the asymptotic normality of $\widehat{\boldsymbol{\beta}}_1^{(n)}$ hold simultaneously, we also need $s_\theta > 1/2$. In the diverging k_n case, the rate of convergence is determined by k_n, p_n, M_n, s_g and s_θ jointly. With appropriate s_g, s_θ and p_n, the rate of convergence can be $n^{-1/2}k_n + k_n^{1/(4s_g+2)}n^{-s_g/(2s_g+1)}$.

4. Computation.

The computation of the SCAD-PLM estimator involves the choice of λ_n. We first consider the estimation as well the standard error approximation of the estimator with a given λ_n and then describe the generalized cross validation approach to choose appropriate λ_n in the PLM.

4.1. Computation of $\widehat{\boldsymbol{\beta}}^{(n)}$ and $\widehat{g}_n$.

The computation of $(\widehat{\boldsymbol{\beta}}^{(n)}, \widehat{g}_n)$ requires the minimization of (2). The projection approach adopted here converts this problem to the minimization of (3). In particular, given m and a partition Δ_n, a basis of $\mathcal{S}_m(\Delta_n)$ is given by $(B_{n1}, \ldots, B_{nq_n})$. The basis functions are evaluated at $T_i, i = 1, \ldots, n$ and form $\mathbb{Z}_n$. In `Splus` or `R`, this can be realized with the bs function. Regress each columns of $\mathbb{X}^{(n)}$ and $\mathbf{Y}$ on $\mathbb{Z}_n$ separately. Denote the residuals by $\widetilde{\mathbb{X}}^{(n)}$ and $\widetilde{\mathbf{Y}}$. The minimization of (3) is now a nonconcave penalized regression problem, with observations $(\widetilde{\mathbb{X}}^{(n)}, \widetilde{\mathbf{Y}})$. We also standardize the columns of $\widetilde{\mathbb{X}}^{(n)}$ so the covariates with smaller variations will not be discriminated against.

Fan and Li (2001) proposed to approximate the nonconcave penalty with a local quadratic function, to facilitate the use of the Newton-Raphson method. At the kth step, suppose the initial value of the jth element of the kth step estimator of $\boldsymbol{\beta}^{(n)}$ is $b_{(k),j}^{(n)}$. The penalty function $p_\lambda(b_j)$ for b_j around $b_{k,j}^{(n)}$ is approximated by

$$p_\lambda(|b_j|) \approx p_\lambda(|b_{(k),j}^{(n)}|) + \frac{p'_\lambda(|b_{k,j}^{(n)}|)}{2|b_{(k),j}^{(n)}|}(b_j^2 - b_{(k),j}^{(n)}{}^2). \tag{5}$$

An undesirable outcome of this approximation is the estimate of $\beta_j^{(n)}$ has to end up being 0 once it reached 0 in any step.

Hunter and Li (2005) described a minorize-maximize (MM) algorithm to compute the nonconcave penalized estimator. In this algorithm, the approximation in (5) is improved with a small perturbation $\xi > 0$ to handle the non-differentiability at 0. This prevents the estimation from being trapped at 0. Let

$$p_{\lambda,\xi}(|b_j|) = p_\lambda(|b_j|) - \xi \int_0^{|b_j|} \frac{p'_\lambda(u)}{\xi + u}\, du.$$

In the $(k+1)$th step, the penalty function $p_\lambda(b_j)$ for b_j around $b_{k,j}^{(n)}$ is approximated by

$$p_{\lambda,\xi}(|b_j|) \approx p_{\lambda,\xi}(|b_{(k),j}^{(n)}|) + \frac{p'_\lambda(|b_{k,j}^{(n)}|+)}{2(\xi + |b_{(k),j}^{(n)}|)}(b_j^2 - b_{(k),j}^{(n)}{}^2). \tag{6}$$

When ξ is small, the difference between (5) and (6) should be small. We adopt this algorithm for computing $\widehat{\boldsymbol{\beta}}^{(n)}$.

Given λ_n and a, the profile objective function is

$$\widetilde{Q}_{n,\xi}(\mathbf{b}^{(n)}; \lambda_n, a) = \sum_{i=1}^{n} (\widetilde{Y}_i - \widetilde{\mathbf{X}}_i^{(n)\prime}\mathbf{b}^{(n)})^2 + n \sum_{j=1}^{p_n} p_{\lambda,\xi}(b_j^{(n)}; a).$$

Let $\mathbf{b}_k^{(n)}$ be the solution at the kth step in the iteration. In the $(k+1)$th step, around $\mathbf{b}_{(k)}^{(n)}$, we approximate $\widetilde{Q}_{n,\xi}$ by

$$S_{k,\xi}(\mathbf{b}^{(n)};\lambda_n,a)=\sum_{i=1}^{n}(\widetilde{Y}_i-\widetilde{\mathbf{X}}_i^{(n)\prime}\mathbf{b}^{(n)})^2+n\sum_{j=1}^{p_n}p_{\lambda_n,\xi}(|b_{(k),j}^{(n)}|;a)+\frac{np'_\lambda(|b_{(k),j}|+;a)}{2(\xi+|b_{(k),j}^{(n)}|)}(b_j^{(n)2}-b_{(k),j}^{(n)}{}^2).$$

Therefore the one-step estimator starting at $\boldsymbol{\beta}_{(k)}$ is

$$\mathbf{b}_{(k+1)}^{(n)}=(\widetilde{\mathbb{X}}^{(n)\prime}\widetilde{\mathbb{X}}^{(n)}+n\mathbf{D}_\xi(\mathbf{b}_{(k)}^{(n)};\lambda_n,a))^{-1}\widetilde{\mathbb{X}}^{(n)\prime}\widetilde{\mathbf{Y}},$$

where $\mathbf{D}_\xi(\mathbf{b}_{(k)}^{(n)};\lambda_n,a)$ is the diagonal matrix whose diagonal elements are $\frac{1}{2}p'_\lambda(|b_{(k),j}^{(n)}|+;a)/(\xi+|b_{(k),j}^{(n)}|)$, $j=1,\ldots,p_n$.

Given the tolerance τ, convergence is claimed when

$$\left|\frac{\partial\widetilde{Q}_{n,\xi}(\mathbf{b}^{(n)})}{\partial b_j^{(n)}}\right|<\frac{\tau}{2},\quad\forall j=1,\ldots,p.$$

In the final $\mathbf{b}^{(n)}$, the $b_j^{(n)}$'s $(1\le j\le p_n)$ that satisfy

$$\left|\frac{\partial\widetilde{Q}_{n,\xi}(\mathbf{b}^{(n)})}{\partial b_j^{(n)}}-\frac{\partial\widetilde{Q}_{n}(\mathbf{b}^{(n)})}{\partial b_j^{(n)}}\right|=\frac{n\xi p'_\lambda(|\beta_j|;a)}{\xi+|\beta_j|}>\frac{\tau}{2}$$

are set to 0. To start, we may take the least square estimator

$$\mathbf{b}_{\mathrm{LS}}^{(n)}=(\widetilde{\mathbb{X}}^{(n)\prime}\widetilde{\mathbb{X}}^{(n)})^{-1}\widetilde{\mathbb{X}}^{(n)\prime}\widetilde{\mathbf{Y}}$$

as $\mathbf{b}_{(0)}^{(n)}$. The perturbation ξ should be kept small so that the difference between $\widetilde{Q}_{n,\xi}(\widehat{\beta})$ and $\widetilde{Q}_n(\widehat{\beta})$ is negligible. Hunter and Li (2005) suggested using

$$\xi=\frac{\tau}{2n\lambda_n}\min\{|b_{(0),j}^{(n)}|:b_{(0),j}^{(n)}\neq 0\}.$$

Simulation in the next section shows that this algorithm works very well for our problem.

Once we have computed $\widehat{\boldsymbol{\beta}}^{(n)}$, the value of g at some $t\in[0,1]$ is estimated by

$$\widehat{g}_n(t)=\mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}(\mathbf{Y}-\mathbb{X}^{(n)}\widehat{\boldsymbol{\beta}}^{(n)}).$$

4.2. Standard errors. Like all estimators computed with the Newton-Raphson algorithm, the standard errors for the nonzero coefficient estimates can be derived from the Hessian matrix. Specifically, by the local linear approximation,

$$\frac{\partial S_\xi(\mathbf{b}_1^{(n)};\lambda_n,a)}{\partial \mathbf{b}_1^{(n)}} \approx \frac{\partial S_\xi(\boldsymbol{\beta}_1^{(n)};\lambda_n,a)}{\partial \boldsymbol{\beta}_1^{(n)}} + \frac{\partial^2 S_\xi(\boldsymbol{\beta}_1^{(n)};\lambda_n,a)}{\partial \boldsymbol{\beta}_1^{(n)}\partial \boldsymbol{\beta}_1^{(n)\prime}}(\mathbf{b}_1^{(n)} - \boldsymbol{\beta}_1^{(n)}),$$

we have

$$\begin{aligned}\mathbf{b}_1^{(n)} - \boldsymbol{\beta}_1^{(n)} &\approx -\Big(\frac{\partial^2 S_\xi(\boldsymbol{\beta}_1^{(n)};\lambda_n,a)}{\partial \boldsymbol{\beta}_1^{(n)}\partial \boldsymbol{\beta}_1^{(n)\prime}}\Big)^{-1}\frac{\partial S_\xi(\boldsymbol{\beta}_1^{(n)};\lambda_n,a)}{\partial \boldsymbol{\beta}_1^{(n)}} \\ &\approx -\Big(\frac{\partial^2 S_\xi(\mathbf{b}_1^{(n)};\lambda_n,a)}{\partial \mathbf{b}_1^{(n)}\partial \mathbf{b}_1^{(n)\prime}}\Big)^{-1}\frac{\partial S_\xi(\mathbf{b}_1^{(n)};\lambda_n,a)}{\partial \mathbf{b}_1^{(n)}}.\end{aligned}$$

Since

$$\begin{aligned}\frac{\partial S_\xi(\mathbf{b}_1^{(n)};\lambda_n,a)}{\partial b_j^{(n)}} &= -2\widetilde{\mathbb{X}}_{\cdot j}^{(n)\prime}\widetilde{\mathbf{Y}} + 2\widetilde{\mathbb{X}}_{\cdot j}^{(n)\prime}\widetilde{\mathbb{X}}^{(n)}\mathbf{b} + n\frac{b_j^{(n)}p'_{\lambda_n}(|b_j^{(n)}|)}{\xi + |b_j^{(n)}|} \\ &= \sum_{i=1}^{n}\left[-2\widetilde{X}_{ij}^{(n)}\widetilde{Y}_i + 2\widetilde{X}_{ij}^{(n)}\widetilde{\mathbf{X}}_{i1}^{(n)\prime}\mathbf{b}_1^{(n)} + \frac{b_j^{(n)}p'_{\lambda_n}(|b_j^{(n)}|)}{\xi + |b_j^{(n)}|}\right], \\ &\triangleq 2\sum_{i=1}^{n}U_{ij}(\xi;\lambda_n,a),\end{aligned}$$

it follows that, for $j, l = 1, \dots, p_n$,

$$\text{Cov}\Big(n^{-1/2}\frac{\partial S_\xi(\mathbf{b}_1^{(n)};\lambda_n,a)}{\partial b_j^{(n)}}, n^{-1/2}\frac{\partial S_\xi(\mathbf{b}_1^{(n)};\lambda_n,a)}{\partial b_j^{(n)}}\Big) \approx \frac{4}{n}\sum_{i=1}^{n}U_{ij}U_{il} - \frac{4}{n^2}\sum_{i=1}^{n}U_{ij}\sum_{i=1}^{n}U_{il} = 4(\text{Cov}(\mathbb{U}))_{jl}.$$

Therefore the variance-covariance matrix of the nonzero estimators can be approximated by

$$\widehat{\text{Cov}(\widehat{\boldsymbol{\beta}}_1^{(n)})} = n(\widetilde{\mathbb{X}}_1^{(n)\prime}\widetilde{\mathbb{X}}_1^{(n)} + n\mathbf{D}_\xi(\widehat{\boldsymbol{\beta}}_1^{(n)};\lambda_n,a))^{-1}\text{Cov}(\mathbb{U})(\widetilde{\mathbb{X}}_1^{(n)\prime}\widetilde{\mathbb{X}}_1^{(n)} + n\mathbf{D}_\xi(\widehat{\boldsymbol{\beta}}_1^{(n)};\lambda_n,a))^{-1}.$$

4.3. Selection of λ_n. The above computation procedures are for the case when λ_n and a are specified. We choose λ_n by minimizing the generalized cross validation score [Wahba (1990)], which is defined to be

$$\text{GCV}(\lambda,a) = \frac{\|\widetilde{\mathbf{Y}} - \widetilde{\mathbb{X}}_1\widehat{\boldsymbol{\beta}}_1^{(n)}(\lambda,a)\|^2/n}{(1-\text{eff}(\lambda,a)/n)^2},$$

where

$$\mathrm{eff}(\lambda, a) = \mathrm{tr}\left[\widetilde{\mathbb{X}}_1^{(n)}\left(\widetilde{\mathbb{X}}_1^{(n)\prime}\widetilde{\mathbb{X}}_1^{(n)} + n\mathbf{D}_\xi(\widehat{\boldsymbol{\beta}}_1^{(n)}(\lambda, a); \lambda, a)\right)^{-1}\widetilde{\mathbb{X}}_1^{(n)\prime}\right]$$

is the number of effective parameters and $\mathbf{D}_\xi(\widehat{\boldsymbol{\beta}}_1^{(n)}(\lambda, a); \lambda, a)$ is a submatrix of $\mathbf{D}_\xi(\widehat{\boldsymbol{\beta}}^{(n)}(\lambda, a); \lambda, a)$. Likewise, $\widetilde{\mathbb{X}}_1^{(n)}$ only includes the columns of which the corresponding elements of $\widehat{\boldsymbol{\beta}}^{(n)}(\lambda, a)$ are non-vanishing.

The requirement that $a > 2$ is needed in the SCAD penalty function. Simulation studies indicate that the generalized cross validation score would not change much for a range of values of a. So for computational efficiency, we fix $a = 3.7$, as suggested by Fan and Li (2001).

Note that here we use fixed partition Δ_n and m in estimating the nonparametric component g. Data-driven choice of them may be desirable. However, in our simulations, $m = 4$ (cubic splines) and $M_n \leq 3$ with even partition of $[0, 1]$ serves the purpose well.

5. Simulation study and data example. In this section we illustrate the SCAD-PLM estimator's finite sample properties with simulated examples. In both examples, we use $m = 4$, $M_n = 3$ and the sample quantiles of T_i's as the knots.

Example 1. In this study, we simulate $n = 100$ points $T_i, i = 1, \ldots, 100$ from the uniform distribution on $[0, 1]$. For each i, e_{ij}'s are simulated to be normally distributed with autocorrelated variance structure $AR(\rho)$ such that

$$\mathrm{Cov}(e_{ij}, e_{il}) = \rho^{|j-l|}, 1 \leq j, l \leq 10,$$

X_{ij}'s are then formed as follows:

$$X_{i1} = \sin(2T_i) + e_{i1}, X_{i2} = (0.5 + T_i)^{-2} + e_{i2}, X_{i3} = \exp(T_i) + e_{i3}, X_{i5} = (T_i - 0.7)^4 + e_{i5},$$

$$X_{i6} = T_i(1 + T_i^2)^{-1} + e_{i6}, X_{i7} = \sqrt{1 + T_i} + e_{i7}, X_{i8} = \log(3T_i + 8) + e_{i8}, X_{ij} = e_{ij}, j = 4, 9, 10.$$

The response Y_i is computed as

$$Y_i = \sum_{j=1}^{10} X_{ij}\beta_{ij} + \cos(T) + \varepsilon_i, \qquad i = 1, \ldots, 100.$$

where $\beta_j = j, 1 \leq j \leq 4$, $\beta_j = 0, 5 \leq j \leq 10$, and ε_i's are sampled from $N(0,1)$, and For each $\rho = 0, 0.2, 0.5, 0.8$, we generated $N = 100$ data sets. For comparison we apply the SCAD penalized regression method, treating T_i as a linear predictor like X_{ij}'s. The corresponding estimator is abbreviated as LS-SCAD estimator. Also linear regression and partially linear regression without penalty are applied for comparison. The linear regression with model selection based on AIC (abbreviated as LS-AIC) is also included in the comparison.

The results are summarized in Table 1 and 2. Columns 3 through 6 in Table 1 are the averages of the estimates of $\beta_j, j = 1, \ldots, 4$ respectively. Column 7 is the numbers of estimates of $\beta_j, 5 \leq j \leq 10$ that are 0, averaged over 100 simulations, and their medians are given in Column 8. Column 9 gives the sums of the numbers of estimates of $\beta_j, j = 1, \ldots, 4$ that are 0. The last column only makes sense for the LS-SCAD and LS-AIC estimators. It gives the percentage of times in the 100 simulations in which the coefficient estimate of T equals 0.

In this simulation model, the nonparametric part $g(T) = \cos(T)$ can be fairly well approximated by a linear function on $[0, 1]$. As a result, the LS-SCAD estimator is expected to give good estimates. It is shown in Table 1 that the estimates of $\beta_j, 1 \leq j \leq 4$ are all very close to the underlying values. The SCAD-penalized counterparts are comparable to the traditional estimators. Neither the linear regression method nor the partially linear regression does any variable selection. LS-SCAD and PLM-SCAD pick out the covariates with zero coefficients efficiently. LS-AIC has similar performance to the LS-SCAD estimator. On average each time 83% of the covariates with zero coefficients are selected and none of the covariates with nonzero coefficients are incorrectly chosen as trivial in the 100 simulations. However, in each setting, about $2/3$ of the time, the LS-SCAD method attributes no effect to T, which does have a nonlinear effect on Y. This is due to the relatively small variation caused in $g(T)$ (with a range less than 0.5) compared with the random variation.

Table 2 summarizes the performance of the sandwich estimator of the standard error of the PLM-SCAD estimator. Columns 2, 4, 6 and 8 are the standard errors of $\beta_j, 1 \leq j \leq 4$ in the 100 simulations, respectively, while Columns 3, 5, 7 and 9 are the average of the standard deviation estimates of these coefficients, obtains via the Hessian matrices. It is seen that the sandwich estimator of the standard error works well, although it slightly underestimates the sampling variation.

Example 2. This study is similar to Example 1 except that the responses are simulated from

$$Y_i = \sum_{j=1}^{10} X_{ij}\beta_j + \cos(2\pi T_i) + \varepsilon_i.$$

So here $g(T) = \cos(2\pi T)$. In Example 1, $g(T) = \cos(T)$. This change in $g(T)$ makes it difficult to have a linear approximation of $g(T)$ on $[0, 1]$. So the LS-SCAD estimator is expected to fail in this situation. Besides, the variation in $g(\cdot)$ (with a range of 2) is relatively large compared to the variation in the error term. Thus misspecification of $g(T)$ will cause bias in the estimation of the linear part. This is reflected in Table 3. The LS-SCAD estimates of the nonzero coefficients are clearly biased and the biases become larger as the correlation between covariates increases. It can be seen in Column 9, when $\rho = 0.8$, the nonzero covariates are even estimated to be trivial, which is not seen when the model is not misspecified.

6. Discussion. In this paper, we studied the SCAD-penalized method for variable selection and estimation in the PLM with a divergent number of covariates. B-spline basis functions are used for fitting the nonparametric part. Variable selection and coefficient estimation in the parametric part are achieved simultaneously. The oracle property of the SCAD-PLM estimator of the parametric part was established and consistency of the SCAD-PLM estimator of the nonparametric part was shown. Compared to the classical variable selection methods such as subset selection, the SCAD has two advantages. First, the variable selection with SCAD is continuous and hence more stable than the subset selection, which is a discrete and non-continuous. Second, the SCAD is computationally feasible for high-dimensional data. In contrast, computation in subset selection is combinatorial and not feasible when p is large.

We have focused on the case where there is one variable in the nonparametric part. Nonetheless, this may be extended to the case of d covariates $T_1, \ldots, T_d$. Specifically, consider the model

$$Y = \mathbf{X}^{(n)\prime}\boldsymbol{\beta}^{(n)} + g(T_1, \ldots, T_d) + \varepsilon. \tag{7}$$

The SCAD-PLM estimator $(\widehat{\boldsymbol{\beta}}^{(n)}, \widehat{g}_n)$ can be obtained via

$$\min_{(\mathbf{b}^{(n)} \in \mathcal{R}^{p_n}, \phi \in \mathcal{S})} \{\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^{(n)\prime}\mathbf{b}^{(n)} - \phi)^2 + n\sum_{j=1}^{p_n} p_{\lambda_n}(b_j^{(n)}; a)\}.$$

Here $\mathcal{S}$ is the space of all the d-variate functions on $[0,1]^d$ that meet some requirement of smoothness. In particular, we can take $\mathcal{S}$ to be the space of the products of the B-spline basis functions, then project $\mathbb{X}^{(n)}$ and $\mathbf{Y}$ onto this space with this basis and perform the SCAD-penalized regression to $\widetilde{\mathbf{Y}}$ on $\widetilde{\mathbb{X}}^{(n)}$. This has already been discussed in Friedman (1991). However, for large d and moderate sample size, even with very small M_n, this model may suffer from the "curse of dimensionality."

A more parsimonious extension is the partially linear additive model (PLAM)

$$Y = \mu + \mathbf{X}^{(n)\prime}\boldsymbol{\beta}^{(n)} + \sum_{l=1}^{d} g_l(T_l) + \varepsilon, \tag{8}$$

where $E[g_l(T_l)] = 0$ holds for $l = 1, \ldots, d$. To estimate $\boldsymbol{\beta}$ and g_l, for each T_l, we first determine the partition Δ_{nl}. For simplicity, we assume that the numbers of knots are M_n and the mesh sizes are $O(M_n^{-1})$ for all l. Suppose that X and Y are centered. The SCAD-PLAM estimator $(\widehat{\boldsymbol{\beta}}^{(n)}, \widehat{g}_{n1}, \ldots, \widehat{g}_{n1})$ is then defined to be the minimizer of

$$\sum_{i=1}^{n}[Y_i - \mathbf{X}_i^{(n)\prime}\mathbf{b}^{(n)} - \sum_{l=1}^{d}\phi_l(T_{il})]^2 + n\sum_{j=1}^{p_n} p_{\lambda_n}(b_j^{(n)}; a)\},$$

subject to

(i) $\sum_{i=1}^{n} \phi_l(T_{il}) = 0$,

(ii) ϕ_l is an element of $\mathcal{S}_m(\Delta_{nl})$.

Under the assumptions similar to those for the SCAD-PLM estimator, $\widehat{\boldsymbol{\beta}}^{(n)}$ can be shown to possess the oracle property. Furthermore, if the joint distribution of $(T_1, \ldots, T_d)$ is absolutely continuous and its density is bounded away from 0 and infinity on $[0,1]^d$, following the proof of Lemma 7 in Stone (1985) and that of Theorem 4 here, we can obtain the same global consistency rate for each additive component, i.e.

$$\|\widehat{g}_{nl} - g_l\|_{T_l} = O_P(k_n/\sqrt{n} + \sqrt{M_n/n} + \sqrt{k_n}M_n^{-s_g}), \quad l = 1, \ldots, d.$$

One way to compute the SCAD-PLAM estimator is the following. First, form the B-spline basis $\{B_{nw}, 1 \le w \le q_n\}$ as follows: the first $M_n + m - 1$ components are the B-spline basis functions corresponding to T_1 ignoring the intercept, the second $M_n + m - 1$ components corresponding to T_2, and so on. The intercept is the last component. So here $q_n = dM_n + dm - d + 1$. Now computation can proceed in a similar way to that for the SCAD-PLM estimator.

Our results require that $p_n < n$. While this condition is often satisfied in applications, there are important settings in which it is violated. For example, in studies with microarray data as covariate measurements, the number of genes (covariates) is typically greater than the sample size. Without any further assumptions on the structure of covariate matrix, the regression parameter is in general not identifiable if $p_n > n$. It is an interesting topic of future research to identify conditions under which the SCAD-PLM estimator achieves consistent variable selection and asymptotic normality even when $p_n > n$.

References

[1] Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York.

[2] Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, **16**, 136–146.

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

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

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

[6] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1–68.

[7] Härdel, W., Liang, H and Gao, J. (2000). *Partially Linear Models*. Springer Contributions to Statistics Physica-Verlag, New York.

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

[9] Huang, J. (1999). Efficient estimation of the partially linear additive Cox model. *The Annals of Statistics*, **27**, 1536–1563.

[10] Huang, J., Horowitz, J. L. and Ma, S. G. (2006). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Technical report # 360, Department of Statistics and Actuarial Science, University of Iowa.

[11] Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, **1**, 799–821.

[12] Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, **33**, 1617-1642.

[13] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, **28**,No. 6, 1356–1378.

[14] Portnoy, S. (1984). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large: I. Consistency. *Ann. Statist.* **12**, 1298-1309.

[15] Portnoy, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large: II. Normal approximation. *Ann. Statist.* **13**, 1403-1417.

[16] Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, **56**, 931–954.

[17] Schumaker, L. (1981). *Spline Functions: Basic Theory*. Wiley, New York.

[18] Speckman, P. (1988). Kernel smoothing in partial lineal models. *Journal of the Royal Statistical Society. Series B*, **50**, 413–436.

[19] Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, **8**, 1348–1360.

[20] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, **10**, 1040–1053.

[21] Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 689–705.

[22] Wahba, G. (1984). Partial spline models for the semiparametric estimation of functions of several variables. *Analyses for Time Series, Japan-US Joint Seminar*, 319–329. Tokyo: Institute of Statistical Mathematics.

[23] Wahba, G. (1990). *Spline models for observational data.* CBMS-NSF Regional Conference Series in Applied Mathematics.

[24] Wang, S. and Jia, Z. (1993). *Inequalities in Matrix Theory*, Anhui Education Press.

Department of Statistics and Actuarial Science

University of Iowa

Iowa City, Iowa 52242

E-mail: huiliang-xie@uiowa.edu, jian-huang@uiowa.edu

Appendix. We now give the proofs of the results stated in Section 3. Write

$$\mathbb{X}^{(n)} = (X_{ij})_{\substack{i=1,\ldots,n\\ j=1,\ldots,p_n}} = (\theta_j^{(n)}(T_i))_{\substack{i=1,\ldots,n\\ j=1,\ldots,p_n}} + (e_{ij}^{(n)})_{\substack{i=1,\ldots,n\\ j=1,\ldots,p_n}} \triangleq \boldsymbol{\theta}^{(n)}(\mathbf{T}) + \mathbb{E}_n.$$

Lemma 1. *Under (A1),*

$$\|\mathbb{X}^{(n)\prime}(I - P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}/n - \Xi^{(n)}\| \xrightarrow{P} 0.$$

Proof. (Lemma 1) For simplicity, write

$$\mathbf{A}^{(n)} = \mathbb{X}^{(n)\prime}(I - P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}/n, \quad \mathbf{C}^{(n)} = \mathbf{A}^{(n)} - \Xi^{(n)}.$$

Note that $\mathbb{X}_{\cdot j}^{(n)} = \mathbf{e}_{\cdot j}^{(n)} + \theta_{nj}(\mathbf{T})$, where $\mathbf{e}_{\cdot j}^{(n)} = (e_{1j}^{(n)}, \ldots, e_{nj}^{(n)})'$.

$$\begin{aligned}|C_{jl}^{(n)}| &= |(\frac{\mathbf{e}_{\cdot j}^{(n)\prime}\mathbf{e}_{\cdot l}^{(n)}}{n} - \Xi_{jl}^{(n)}) + \frac{\mathbf{e}_{\cdot j}^{(n)\prime}P_{\mathbf{Z}}^{(n)}\mathbf{e}_{\cdot l}^{(n)}}{n} + \frac{\mathbf{e}_{\cdot j}^{(n)\prime}(I - P_{\mathbf{Z}}^{(n)})\boldsymbol{\theta}_l^{(n)}(\mathbf{T})}{n}\\ &\quad + \frac{\mathbf{e}_{\cdot l}^{(n)\prime}(I - P_{\mathbf{Z}}^{(n)})\boldsymbol{\theta}_j^{(n)}(\mathbf{T})}{n} + \frac{\boldsymbol{\theta}_j^{(n)}(\mathbf{T})'(I - P_{\mathbf{Z}}^{(n)})\boldsymbol{\theta}_l^{(n)}(\mathbf{T})}{n}|\end{aligned}$$

By Condition 2,

$$E\left[n^{-1}\mathbf{e}_{\cdot j}^{(n)\prime}\mathbf{e}_{\cdot l}^{(n)} - \Xi_{jl}^{(n)}\right]^2 = n^{-1}\text{Var}(e_j^{(n)}e_l^{(n)}) \le n^{-1}\sigma_{4e}.$$

Since

$$\begin{aligned} E\left[n^{-1}\mathbf{e}_{\cdot j}^{(n)\prime}P_{\mathbf{Z}}^{(n)}\mathbf{e}_{\cdot j}^{(n)}\right]^2 &= n^{-2}E\left\{E[(\mathbf{e}_{\cdot j}^{(n)\prime}P_{\mathbf{Z}}^{(n)}\mathbf{e}_{\cdot j}^{(n)})^2|\mathbb{Z}^{(n)}]\right\} \\ &= n^{-2}E\left\{\sum_{i=1}^n\sum_{i'=1}^n\sum_{\iota=1}^n\sum_{\iota'=1}^n P_{ii'}P_{\iota\iota'}E\left[e_{ij}^{(n)}e_{i'j}^{(n)}e_{\iota j}^{(n)}e_{\iota' j}^{(n)}|\mathbb{Z}^{(n)}\right]\right\}, \end{aligned}$$

and

$$P_{ii'}P_{\iota\iota'}E\left[e_{ij}^{(n)}e_{i'j}^{(n)}e_{\iota j}^{(n)}e_{\iota' j}^{(n)}|\mathbb{Z}^{(n)}\right] = \begin{cases} P_{ii}P_{\iota\iota}\Sigma_{jj}^{(n)}(T_i)\Sigma_{jj}^{(n)}(T_\iota), & i = i' \ne \iota = \iota', \\ P_{ii'}^2\Sigma_{jj}^{(n)}(T_i)\Sigma_{jj}^{(n)}(T_{i'}), & i = \iota \ne i' = \iota', \\ P_{ii'}^2\Sigma_{jj}^{(n)}(T_i)\Sigma_{jj}^{(n)}(T_{i'}), & i = \iota' \ne i' = \iota, \\ P_{ii}^2E[e_{ij}^{(n)4}|T_i], & i = i' = \iota = \iota', \\ 0, & \text{otherwise}, \end{cases}$$

together with $\Sigma_{jj}^{(n)}(T_i) \le \sigma_{4e}^{1/2}$ and $P_{\mathbf{Z},ii}^{(n)} \le 1$, we have

$$\begin{aligned} E\left[n^{-1}\mathbf{e}_{\cdot j}^{(n)\prime}P_{\mathbf{Z}}^{(n)}\mathbf{e}_{\cdot j}^{(n)}\right]^2 &\le n^{-2}\sigma_{4e}\left\{E[\text{tr}^2(P_{\mathbf{Z}}^{(n)})] + 2E[\text{tr}(P_{\mathbf{Z}}^{(n)^2})]\right\} + n^{-2}\sigma_{4e}E[\text{tr}(P_{\mathbf{Z}}^{(n)})] \\ &\le n^{-2}\sigma_{4e}(q_n^2 + 3q_n). \end{aligned}$$

By Corollary 6.21 in Schumaker (1981) and the properties of least square regression,

$$E[n^{-1}\boldsymbol{\theta}_j^{(n)}(\mathbf{T})(I - P_{\mathbf{Z}}^{(n)})\boldsymbol{\theta}_j^{(n)}(\mathbf{T})] \le C_1M_\theta(\overline{\Delta}_n)^{2s_\theta},$$

where C_1 is a constant determined only by r_θ. By the Cauchy-Schwarz inequality and C_r inequality we have

$$\|\mathbf{C}^{(n)}\|^2 = O_P(p_n^2/n + p_n^2M_n^2/n^2 + p_n^2M_n^{-2s_\theta}).$$

The convergence follows from (A1). □

Lemma 2. $E[tr(\mathbb{X}^{(n)\prime}(I - P_{\mathbf{Z}})\mathbb{X}^{(n)})] = O(np_n)$.

Proof.

$$\begin{aligned}
& E[\mathrm{tr}(\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}})\mathbb{X}^{(n)})] \\
= \;& E\left[\mathrm{tr}([\mathbb{E}^{(n)}+\boldsymbol{\theta}^{(n)}(\mathbf{T})]'(I-P_{\mathbf{Z}})[\mathbb{E}^{(n)}+\boldsymbol{\theta}^{(n)}(\mathbf{T})])\right] \\
= \;& E\left[\mathrm{tr}(\mathbb{E}^{(n)\prime}(I-P_{\mathbf{Z}})\mathbb{E}^{(n)}+2\mathbb{E}^{(n)\prime}(I-P_{\mathbf{Z}})\boldsymbol{\theta}^{(n)}(\mathbf{T})+\boldsymbol{\theta}^{(n)}(\mathbf{T})'(I-P_{\mathbf{Z}})\boldsymbol{\theta}^{(n)}(\mathbf{T}))\right] \\
= \;& E\left\{E\left[\mathrm{tr}(\mathbb{E}^{(n)\prime}(I-P_{\mathbf{Z}})\mathbb{E}^{(n)}+2\mathbb{E}^{(n)\prime}(I-P_{\mathbf{Z}})\boldsymbol{\theta}^{(n)}(\mathbf{T})+\boldsymbol{\theta}^{(n)}(\mathbf{T})'(I-P_{\mathbf{Z}})\boldsymbol{\theta}^{(n)}(\mathbf{T}))|\mathbf{T}\right]\right\} \\
= \;& E\left\{E\left[\mathrm{tr}(\mathbb{E}^{(n)\prime}(I-P_{\mathbf{Z}})\mathbb{E}^{(n)})|\mathbf{T}\right]\right\}+E\left[\mathrm{tr}(\boldsymbol{\theta}^{(n)}(\mathbf{T})'(I-P_{\mathbf{Z}})\boldsymbol{\theta}^{(n)}(\mathbf{T}))\right] \\
\leq \;& E\left\{E\left[\mathrm{tr}(\mathbb{E}^{(n)\prime}(I-P_{\mathbf{Z}})\mathbb{E}^{(n)})|\mathbf{T}\right]\right\}+C_1np_nM_\theta M_n^{-2s_\theta} \\
= \;& E\left\{E\left[\sum_{j=1}^{p_n}\mathbf{e}_{\cdot j}^{(n)\prime}(I-P_{\mathbf{Z}})\mathbf{e}_{\cdot j}^{(n)}|\mathbf{T}\right]\right\}+C_1np_nM_\theta M_n^{-2s_\theta} \\
= \;& E[\sum_{j=1}^{p_n}\mathrm{tr}((I-P_{\mathbf{Z}})\Sigma_{jj}^{(n)}(\mathbf{T}))]+C_1np_nM_\theta M_n^{-2s_\theta} \\
\leq \;& np_n\sigma_{4e}^{1/2}+C_1np_nM_\theta M_n^{-2s_\theta}. \qquad (\mathrm{tr}(AB)\leq\lambda_{\max}(B)\mathrm{tr}(A))
\end{aligned}$$

Here $\Sigma_{jj}^{(n)}(\mathbf{T})) = \mathrm{diag}(\Sigma_{jj}^{(n)}(T_1),\ldots,\Sigma_{jj}^{(n)}(T_n))$. □

Proof of Theorem 1

Proof. Let $\boldsymbol{\varepsilon}=(\varepsilon_1,\ldots,\varepsilon_n)'$ and $\mathbf{g}(\mathbf{T})=(g(T_1),\ldots,g(T_n))'$. Since $\widehat{\boldsymbol{\beta}}^{(n)}$ minimizes $Q_n(\mathbf{b}^{(n)})$, it necessarily holds that

$$Q_n(\widehat{\boldsymbol{\beta}}^{(n)})\leq Q_n(\boldsymbol{\beta}^{(n)}).$$

Rewriting this inequality, we have

$$\|(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}(\widehat{\boldsymbol{\beta}}^{(n)}-\boldsymbol{\beta}^{(n)})\|^2-2(\boldsymbol{\varepsilon}+\mathbf{g}(\mathbf{T}))'(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}(\widehat{\boldsymbol{\beta}}^{(n)}-\boldsymbol{\beta}^{(n)})\leq\frac{nk_n}{2}(a+1)\lambda_n^2.$$

Let

$$\boldsymbol{\delta}_n=n^{-1/2}\left[\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}\right]^{1/2}(\widehat{\boldsymbol{\beta}}^{(n)}-\boldsymbol{\beta}^{(n)}),$$

and

$$\boldsymbol{\omega}_n=n^{-1/2}\left[\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}\right]^{-1/2}\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})(\boldsymbol{\varepsilon}+\mathbf{g}(\mathbf{T})).$$

Then

$$\|\boldsymbol{\delta}_n\|^2 - 2\boldsymbol{\omega}_n'\boldsymbol{\delta}_n \leq \frac{k_n}{2}(a+1)\lambda_n^2.$$

i.e.

$$\|\boldsymbol{\delta}_n - \boldsymbol{\omega}_n\|^2 \leq \|\boldsymbol{\omega}_n\|^2 + \frac{k_n}{2}(a+1)\lambda_n^2.$$

By the C_r inequality,

$$\begin{aligned}\|\boldsymbol{\delta}_n\|^2 &\leq 2\left(\|\boldsymbol{\delta}_n - \boldsymbol{\omega}_n\|^2 + \|\boldsymbol{\omega}_n\|^2\right) \\ &\leq 4\|\boldsymbol{\omega}_n\|^2 + k_n(a+1)\lambda_n^2.\end{aligned}$$

Examine

$$\begin{aligned}\|\boldsymbol{\omega}_n\|^2 &= n^{-1}(\boldsymbol{\varepsilon}+\mathbf{g}(\mathbf{T}))'(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}\left[\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}\right]^{-1}\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})(\boldsymbol{\varepsilon}+\mathbf{g}(\mathbf{T})) \\ &\triangleq I_{n1}+I_{n2}+I_{n3},\end{aligned}$$

where

$$\begin{aligned}I_{n1} &= n^{-1}\boldsymbol{\varepsilon}'(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}\left[\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}\right]^{-1}\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})\boldsymbol{\varepsilon}, \\ I_{n2} &= 2n^{-1}\boldsymbol{\varepsilon}'(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}\left[\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}\right]^{-1}\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})\mathbf{g}(\mathbf{T}), \\ I_{n3} &= n^{-1}\mathbf{g}(\mathbf{T})'(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}\left[\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})\mathbb{X}^{(n)}\right]^{-1}\mathbb{X}^{(n)\prime}(I-P_{\mathbf{Z}}^{(n)})\mathbf{g}(\mathbf{T}).\end{aligned}$$

Obviously,

$$I_{n1} = E[E(I_{n1}|\mathbb{X}^{(n)},\mathbf{T})]O_P(1) = p_n n^{-1}O_P(1).$$

By the property of projection matrices,

$$I_{n3} \leq n^{-1}\mathbf{g}(\mathbf{T})'(I-P_{\mathbf{Z}}^{(n)})\mathbf{g}(\mathbf{T}) = M_n^{-2s_g}O(1).$$

Thus $\|\boldsymbol{\omega}_n\|^2 = O_P(p_n/n + M_n^{-2s_g})$. Furthermore,

$$\|\widehat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^{(n)}\|^2 = O_P(p_n/n + M_n^{-2s_g} + k_n\lambda_n^2)$$

follows from Lemma 1 with (A2). Thus (A3) immediately leads to the consistency. □

Lemma 3. *(Rate of convergence) Suppose (A1)–(A4) hold. Then*

$$\|\widehat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^{(n)}\| = O_P(\sqrt{p_n/n} + \sqrt{p_n}/M_n^{s_g}).$$

Proof. Let $u_n = \sqrt{p_n/n} + M_n^{-s_g} + \sqrt{k_n}\,\lambda_n$. When $u_n = o(\min_{1\le j\le k_n}|\beta_j^{(n)}|)$, with probability tending to 1, $\min_{1\le j\le k_n}|\widehat{\beta}_j^{(n)}| > a\lambda_n$.

Given a sequence $\{h_n : h_n > 0\}$ that converges to 0, partition $\mathcal{R}^{p_n}\backslash\{\mathbf{0}_{p_n}\}$ into shells $\{S_{n,l}, l = 0, 1, \ldots\}$ where $S_{n,l} = \{\mathbf{b}^{(n)} : 2^{l-1}h_n \le \|\mathbf{b}^{(n)} - \boldsymbol{\beta}^{(n)}\| < 2^l h_n\}$. Then

$$\begin{aligned}
P\left(\|\widehat{\boldsymbol{\beta}}_n^{(n)} - \boldsymbol{\beta}^{(n)}\| \ge 2^L h_n\right) &\le o(1) + \sum_{\substack{l>L \\ 2^l h_n \le 2^{L_1} u_n}} P\left(\widehat{\boldsymbol{\beta}}_n^{(n)} \in S_{n,l}, \|\mathbf{C}^{(n)}\| \le c/2\right) \\
&\le o(1) + \sum_{\substack{l>L \\ 2^l h_n \le 2^{L_1} u_n}} P\left(\inf_{\mathbf{b}^{(n)}\in S_{n,l}} Q_n(\mathbf{b}^{(n)}) \le Q_n(\boldsymbol{\beta}^{(n)}), \|\mathbf{C}^{(n)}\| \le c_\lambda/2\right) \\
&\le o(1) + \sum_{l>L} P\Bigg(\sup_{\mathbf{b}^{(n)}\in S_{n,l}} 2(\boldsymbol{\varepsilon} + \mathbf{g}(\mathbf{T}))'(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}(\mathbf{b}^{(n)} - \boldsymbol{\beta}^{(n)}) \ge \\
&\quad \inf_{\mathbf{b}^{(n)}\in S_{n,l}} (\mathbf{b}^{(n)} - \boldsymbol{\beta}^{(n)})'\mathbb{X}^{(n)\prime}(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}(\mathbf{b}^{(n)} - \boldsymbol{\beta}^{(n)}), \|\mathbf{C}^{(n)}\| \le c_\lambda/2\Bigg) \\
&\le \sum_{l>L} P\Bigg(\sup_{\mathbf{b}^{(n)}\in S_{n,l}} (\boldsymbol{\varepsilon} + \mathbf{g}(\mathbf{T}))'(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}(\mathbf{b}^{(n)} - \boldsymbol{\beta}^{(n)}) \ge 2^{2l-4} n c_\lambda h_n^2\Bigg) \\
&\quad + o(1).
\end{aligned}$$

Since

$$\begin{aligned}
& E \sup_{\mathbf{b}^{(n)}\in S_{n,l}} |(\boldsymbol{\varepsilon} + \mathbf{g}(\mathbf{T}))'(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}(\mathbf{b}^{(n)} - \boldsymbol{\beta}^{(n)})| \\
\le\ & 2^l h_n \sqrt{E[(\boldsymbol{\varepsilon} + \mathbf{g}(\mathbf{T}))'(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}\mathbb{X}^{(n)\prime}(I - P_{\mathbf{Z}})(\boldsymbol{\varepsilon} + \mathbf{g}(\mathbf{T}))]} \\
\le\ & 2^{l+1/2} h_n \sqrt{E[\boldsymbol{\varepsilon}'(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}\mathbb{X}^{(n)\prime}(I - P_{\mathbf{Z}})\boldsymbol{\varepsilon}] + E[\mathbf{g}(\mathbf{T})'(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}\mathbb{X}^{(n)\prime}(I - P_{\mathbf{Z}})\mathbf{g}(\mathbf{T})]} \\
\le\ & 2^{l+1/2} h_n \sqrt{C_3 n p_n + E[\mathbf{g}(\mathbf{T})'(I - P_{\mathbf{Z}})\mathbf{g}(\mathbf{T})\mathrm{tr}(\mathbb{X}^{(n)}\mathbb{X}^{(n)\prime}(I - P_{\mathbf{Z}}))]} \\
\le\ & 2^l h_n C_4 \left(\sqrt{np_n} + n\sqrt{p_n} M_n^{-s_g}\right).
\end{aligned}$$

Continuing the previous arguments, by the Markov inequality,

$$P\left(\|\widehat{\boldsymbol{\beta}}_n^{(n)} - \boldsymbol{\beta}^{(n)}\| \ge 2^L h_n\right) \le o(1) + \sum_{l>L} \frac{C_5(\sqrt{p_n} + \sqrt{np_n} M_n^{-s_g})}{2^{l-4} h_n \sqrt{n}}.$$

This suggests that $\|\widehat{\boldsymbol{\beta}}^{(n)} - \boldsymbol{\beta}^{(n)}\| = O_P(\sqrt{p_n/n} + \sqrt{p_n}/M_n^{s_g})$. □

Proof of Theorem 2

Proof. Consider the partial derivatives of $Q_n(\boldsymbol{\beta}^{(n)} + \mathbf{v}^{(n)})$. We assume $\|\mathbf{v}^{(n)}\| = O_P(\sqrt{p_n/n} + \sqrt{p_n}M_n^{-s_g})$. Suppose the support sets of $e_j^{(n)}$ are all contained in a compact set $[-C_e, C_e]$. For $j = k_n+1, \ldots, p_n$, if $\|\mathbf{v}^{(n)}\| \leq \lambda_n$,

$$\begin{aligned}\frac{\partial\, Q_n(\boldsymbol{\beta}^{(n)} + \mathbf{v}^{(n)})}{\partial\, v_j^{(n)}} &= 2\mathbb{X}_{\cdot j}^{(n)\prime}(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}\mathbf{v}^{(n)} + 2\mathbb{X}_{\cdot j}^{(n)\prime}(I - P_{\mathbf{Z}})(\boldsymbol{\varepsilon} + \mathbf{g}(\mathbf{T})) + n\lambda_n \mathrm{sgn}(v_j^{(n)}) \\ &\triangleq II_{n1,j} + II_{n2,j} + II_{n3,j}.\end{aligned}$$

$$\begin{aligned}\max_{k_n+1\leq j\leq p_n} |II_{n1,j}| &= 2|\mathbb{X}_{\cdot j}^{(n)\prime}(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}\mathbf{v}^{(n)}| \\ &\leq 2\|\mathbf{v}^{(n)}\| \max_{k_n+1\leq j\leq p_n} \left\|\mathbb{X}_{\cdot j}^{(n)\prime}(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}\right\| \\ &\leq (\sqrt{p_n/n} + \sqrt{p_n}M_n^{-s_g})O_P(1) \max_{k_n+1\leq j\leq p_n} \|(I - P_{\mathbf{Z}})\mathbb{X}_{\cdot j}^{(n)}\|\lambda_{\max}^{1/2}(\mathbb{X}^{(n)\prime}(I - P_{\mathbf{Z}})\mathbb{X}^{(n)}) \\ &= (\sqrt{p_n n} + n\sqrt{p_n}M_n^{-s_g})O_P(1)\sqrt{\lambda_{\max}(\Xi^{(n)}) + o_P(1)} \\ &= \sqrt{p_n(n + n^2M_n^{-2s_g})\lambda_{\max}(\Xi^{(n)})O_P(1)}.\end{aligned}$$

So this term is dominated by $\frac{1}{2}II_{n3,j}$ as long as

$$\lim \frac{\sqrt{n}\lambda_n}{\sqrt{p_n\lambda_{\max}(\Xi^{(n)})}} = \infty \qquad \text{and} \qquad \lim \frac{\lambda_n M_n^{s_g}}{\sqrt{p_n\lambda_{\max}(\Xi^{(n)})}} = \infty,$$

both of which are stated in (A5). To sift out all the trivial components, we need

$$P(\max_{k_n+1\leq j\leq p_n} |II_{n2,j}| > n\lambda_n/2) \to 0.$$

This is also implied by (A5), as can be seen from

$$\begin{aligned}
& P(\max_{k_n+1\leq j\leq p_n} |II_{n2,j}| > n\lambda_n/2) \\
\leq\ & \frac{2E[\max_{k_n+1\leq j\leq p_n} |II_{n2,j}|]}{n\lambda_n} \\
\leq\ & \frac{2\sqrt{\sum_{j=k_n+1}^{p_n} E[II_{n2,j}^2]}}{n\lambda_n} \\
\leq\ & \frac{2\sqrt{2}\sqrt{\sum_{j=k_n+1}^{p_n}\{E[\boldsymbol{\varepsilon}'(I-P_{\mathbf{Z}})\mathbb{X}_j^{(n)}\mathbb{X}_j^{(n)\prime}(I-P_{\mathbf{Z}})\boldsymbol{\varepsilon}] + E[g(\mathbf{T})'(I-P_{\mathbf{Z}})\mathbb{X}_j^{(n)}\mathbb{X}_j^{(n)\prime}(I-P_{\mathbf{Z}})g(\mathbf{T})]\}}}{n\lambda_n} \\
\leq\ & \frac{C_1\sqrt{nm_n + nM_n^{-2s_g}nm_n}}{n\lambda_n}.
\end{aligned}$$

This completes the proof. □

Proof of Theorem 3

Proof. Let $\mathbf{A}_n$ be any $\iota \times k_n$ matrix with full row rank and $\Sigma_n = \mathbf{A}_n\mathbf{A}_n'$. From the variable selection conclusion, with probability tending to 1, we have

$$\widehat{\boldsymbol{\beta}}_1^{(n)} - \boldsymbol{\beta}_1^{(n)} = \left[\mathbb{X}_1^{(n)\prime}(I-P_{\mathbf{Z}})\mathbb{X}_1^{(n)}\right]^{-1}\mathbb{X}_1^{(n)\prime}(I-P_{\mathbf{Z}})(g(\mathbf{T})+\boldsymbol{\varepsilon}).$$

We consider the limit distribution of

$$\begin{aligned}
\mathbf{V}_n &= n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\left[\mathbb{X}_1^{(n)\prime}(I-P_{\mathbf{Z}})\mathbb{X}_1^{(n)}\right](\widehat{\boldsymbol{\beta}}_1^{(n)} - \boldsymbol{\beta}_1^{(n)}) \\
&= n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\mathbb{X}_1^{(n)\prime}(I-P_{\mathbf{Z}})(g(\mathbf{T})+\boldsymbol{\varepsilon}) \\
&\triangleq I_{n1} + I_{n2},
\end{aligned}$$

where

$$\begin{aligned}
I_{n1} &= n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\mathbb{X}_1^{(n)\prime}(I-P_{\mathbf{Z}})g(\mathbf{T}), \\
I_{n2} &= n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\mathbb{X}_1^{(n)\prime}(I-P_{\mathbf{Z}})\boldsymbol{\varepsilon}.
\end{aligned}$$

Note that the conclusion of Theorem 3 is equivalent to

$$\mathbf{V}_n \xrightarrow{d} N(\mathbf{0}_\iota, \sigma^2 I_\iota).$$

The first term is a $o_P(1)$ term under (A6) and (A7), as shown in

$$\begin{aligned} I_{n1} &= n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\mathbb{E}_1^{(n)\prime}(I-P_{\mathbf{Z}})g(\mathbf{T}) + n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\boldsymbol{\theta}_1^{(n)\prime}(\mathbf{T})(I-P_{\mathbf{Z}})g(\mathbf{T}), \\ &= II_{n1} + II_{n2}, \end{aligned}$$

where

$$\begin{aligned} \|II_{n1}\|^2 &= E\|II_{n1}\|^2 O_P(1) \\ &= n^{-1}E\left[g(\mathbf{T})'(I-P_{\mathbf{Z}})\mathbb{E}_1^{(n)}\Xi_{11}^{(n)^{-1/2}}\mathbf{A}_n'\Sigma_n^{-1}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\mathbb{E}_1^{(n)\prime}(I-P_{\mathbf{Z}})g(\mathbf{T})\right]O_P(1) \\ &= n^{-1}E\left\{g(\mathbf{T})'(I-P_{\mathbf{Z}})E\left[\mathbb{E}_1^{(n)}\Xi_{11}^{(n)^{-1/2}}\mathbf{A}_n'\Sigma_n^{-1}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\mathbb{E}_1^{(n)\prime}|\mathbf{T}\right](I-P_{\mathbf{Z}})g(\mathbf{T})\right\}O_P(1) \\ &\leq n^{-1}E\left\{g(\mathbf{T})'(I-P_{\mathbf{Z}})E\left[\mathbb{E}_1^{(n)}\Xi_{11}^{(n)^{-1}}\mathbb{E}_1^{(n)\prime}|\mathbf{T}\right](I-P_{\mathbf{Z}})g(\mathbf{T})\right\}O_P(1) \\ &= n^{-1}E\left\{g(\mathbf{T})'(I-P_{\mathbf{Z}})\text{Diag}\left(\text{tr}(\Xi_{11}^{(n)^{-1}}\Sigma_{11}^{(n)}(T_1)),\ldots,\text{tr}(\Xi_{11}^{(n)^{-1}}\Sigma_{11}^{(n)}(T_n))\right)(I-P_{\mathbf{Z}})g(\mathbf{T})\right\}O_P(1) \\ &\leq n^{-1}\|(I-P_{\mathbf{Z}})g(\mathbf{T})\|^2\text{tr}\left(\Sigma_{u,11}^{(n)}\right) \\ &= \text{tr}\left(\Sigma_{u,11}^{(n)}\right)M_n^{-2s_g}O_P(1) = o_P(1), \end{aligned}$$

and

$$\begin{aligned} \|II_{n2}\|^2 &\leq n^{-1}\|(I-P_{\mathbf{Z}})g(\mathbf{T})\|^2\lambda_{\max}\left((I-P_{\mathbf{Z}})\boldsymbol{\theta}_1^{(n)}(\mathbf{T})\Xi_{11}^{(n)^{-1}}(I-P_{\mathbf{Z}})\boldsymbol{\theta}_1^{(n)\prime}\right) \\ &\leq n^{-1}\|(I-P_{\mathbf{Z}})g(\mathbf{T})\|^2\|(I-P_{\mathbf{Z}})\boldsymbol{\theta}_1^{(n)}\|^2 \\ &= n^{-1}nM_n^{-2s_g}nM_n^{-2s_\theta}O(1) \\ &= nM_n^{-2(s_g+s_\theta)}O(1). \end{aligned}$$

Decompose the second term I_{n2} as

$$\begin{aligned} I_{n2} &= n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\mathbb{E}_1^{(n)\prime}\boldsymbol{\varepsilon} - n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\mathbb{E}_1^{(n)\prime}P_{\mathbf{Z}}\boldsymbol{\varepsilon} \\ &\quad + n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\boldsymbol{\theta}_1^{(n)\prime}(\mathbf{T})(I-P_{\mathbf{Z}})\boldsymbol{\varepsilon}, \\ &= III_{n1} + III_{n2} + III_{n3}. \end{aligned}$$

Actually, the last two terms above are trivial:

$$
\begin{aligned}
\|III_{n2}\|^2 &= n^{-1}O_P(1)E[\text{tr}(P_{\mathbf{Z}}\mathbb{E}_1^{(n)}\Xi_{11}^{(n)^{-1/2}}\mathbf{A}_n'\Sigma_n^{-1}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\mathbb{E}_1^{(n)\prime}P_{\mathbf{Z}})] \\
&\le n^{-1}O_P(1)E[\text{tr}(P_{\mathbf{Z}}\mathbb{E}_1^{(n)}\Xi_{11}^{(n)^{-1}}\mathbb{E}_1^{(n)\prime}P_{\mathbf{Z}})] \\
&= n^{-1}O_P(1)E[\text{tr}(P_{\mathbf{Z}}\mathbb{E}_1^{(n)}\mathbb{E}_1^{(n)\prime})] \\
&\le n^{-1}O_P(1)\text{tr}(\Sigma_{u,11}^{(n)})E[\text{tr}\,(P_{\mathbf{Z}})] \\
&= \text{tr}(\Sigma_{u,11}^{(n)})M_n/nO_P(1) = o_P(1). \\
\|III_{n3}\|^2 &= n^{-1}O_P(1)E\left[\text{tr}\left((I-P_{\mathbf{Z}})\boldsymbol{\theta}_1^{(n)}(\mathbf{T})\Xi_{11}^{(n)^{-1}}\boldsymbol{\theta}_1^{(n)\prime}(\mathbf{T})(I-P_{\mathbf{Z}})\right)\right] \\
&= k_nM_n^{-2s_\theta}O_P(1) = o_P(1).
\end{aligned}
$$

So we focus on $III_{n1} = n^{-1/2}\Sigma_n^{-1/2}\mathbf{A}_n\Xi_{11}^{(n)^{-1/2}}\mathbb{E}_1^{(n)\prime}\boldsymbol{\varepsilon}$. Since

$$\text{Var}(III_{n1}) = E[\text{Var}(III_{n1}|\mathbb{X}^{(n)},\mathbf{T})] = \sigma^2 I_\iota,$$

by the central limit theorem we have

$$III_{n1} \xrightarrow{d} N(\mathbf{0}_\iota, \sigma^2 I_\iota).$$

The conclusion follows from the Slutsky's theorem. □

Lemma 4. *Sequences of random variables A_n and random vectors $\mathbf{B}_n$ satisfy $E[A_n^2|\mathbf{B}_n] = O_P(u_n^2)$, where $\{u_n\}$ is a sequence of positive numbers. Then*

$$A_n = O_P(u_n).$$

Proof. For any $\varepsilon > 0$, there is some M_1 such that

$$P(E[A_n^2|\mathbf{B}_n] > M_1u_n^2) < \varepsilon/2.$$

Let $M_2^2 = 2M_1/\varepsilon$. Then

$$\begin{aligned}
P(|A_n| > M_2 u_n) &\leq P(|A_n| > M_2 u_n, E[A_n^2|\mathbf{B}_n] \leq M_1 u_n^2) + P(E[A_n^2|\mathbf{B}_n] > M_1 u_n^2) \\
&< E[1_{(|A_n|>M_2u_n)} 1_{(E[A_n^2|\mathbf{B}_n]\leq M_1 u_n^2)}] + \varepsilon/2 \\
&= E\{1_{(E[A_n^2|\mathbf{B}_n]\leq M_1 u_n^2)} E[1_{(|A_n|>M_2u_n)}|\mathbf{B}_n]\} + \varepsilon/2 \\
&\leq E\left[1_{(E[A_n^2|\mathbf{B}_n]\leq M_1 u_n^2)} \frac{E[A_n^2|\mathbf{B}_n]}{M_2^2 u_n^2}\right] + \varepsilon/2 \\
&\leq \varepsilon.
\end{aligned}$$

The arbitrariness of ε implies the conclusion. □

Proof of Theorem 4

Proof. The nonparametric component $g()$ at a point $t \in [0,1]$ is estimated with

$$\widehat{g}_n(t) = \mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}(\mathbf{Y} - \mathbb{X}^{(n)}\widehat{\boldsymbol{\beta}}^{(n)}).$$

With probability tending to 1,

$$\begin{aligned}
\widehat{g}_n(t) - g(t) &= \mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}(\mathbf{Y} - \mathbb{X}_1^{(n)}\widehat{\boldsymbol{\beta}}_1^{(n)}) - g(t) \\
&= \mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}(\mathbf{Y} - \mathbb{X}_1^{(n)}\boldsymbol{\beta}_1^{(n)}) - g(t) \\
&\quad -\mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}\mathbb{X}_1^{(n)}(\widehat{\boldsymbol{\beta}}_1^{(n)} - \boldsymbol{\beta}_1^{(n)}) \\
&= \mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}g(\mathbf{T}) - g(t) \\
&\quad + \mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}\boldsymbol{\varepsilon} \\
&\quad - \mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}\boldsymbol{\theta}_1^{(n)}(\mathbf{T})(\widehat{\boldsymbol{\beta}}_1^{(n)} - \boldsymbol{\beta}_1^{(n)}) \\
&\quad - \mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}\mathbb{E}_1^{(n)}(\widehat{\boldsymbol{\beta}}_1^{(n)} - \boldsymbol{\beta}_1^{(n)}) \\
&\triangleq I_{n1} + I_{n2} + I_{n3} + I_{n4}.
\end{aligned}$$

Consider $\|\widehat{g}_n - g\|_T^2 = \int[\widehat{g}_n(t) - g(t)]^2 f_T(t)dt$. Without further assumptions, by Lemma 9 in Stone (1985),

$$\|I_{n1}\|_T^2 = O_P(M_n^{-2s_g}).$$

When $M_n = o(\sqrt{n})$, by Lemma 4 in Stone (1985),

$$E[\|I_{n2}\|_T^2 | \mathbf{T}] = O_P(M_n/n).$$

Therefore

$$\|I_{n2}\|_T^2 = O_P(M_n/n).$$

When $\{\theta_j^{(n)}(\cdot), n \geq 1, 1 \leq j \leq k_n\}$ are uniformly bounded on $[0,1]$,

$$\begin{aligned}
\|I_{n3}\|_T^2 &\leq \|\mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}\boldsymbol{\theta}_1^{(n)}(\mathbf{T})\|_T^2 \, \|\widehat{\boldsymbol{\beta}}_1^{(n)} - \boldsymbol{\beta}_1^{(n)}\|^2 \\
&\leq [O(k_n) + O_P(k_n M_n^{-2s_\theta})] \, [O_P(1) M_n^{-2s_g} + k_n/n O_P(1)] \\
&= O_P(1)\left(k_n M_n^{-2s_g} + k_n^2 n^{-1}\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|I_{n4}\|_T^2 &\leq \|\mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}\mathbb{E}_1^{(n)}\|_T^2 \, \|\widehat{\boldsymbol{\beta}}_1^{(n)} - \boldsymbol{\beta}_1^{(n)}\|^2 \\
&= \|\widehat{\boldsymbol{\beta}}_1^{(n)} - \boldsymbol{\beta}_1^{(n)}\|^2 \, \|\mathbf{Z}(t;\Delta_n)'(\mathbb{Z}^{(n)\prime}\mathbb{Z}^{(n)})^{-1}\mathbb{Z}^{(n)\prime}\mathbb{E}_1^{(n)}\|_T^2 \\
&\leq O_P(k_n M_n/n)[O_P(1) M_n^{-2s_g} + k_n/n O_P(1)] \\
&= O_P(1)\left(M_n^{1-2s_g} k_n/n + M_n k_n^2/n^2\right).
\end{aligned}$$

To sum up, when $k_n = o(\sqrt{n})$, we have

$$\|\widehat{g}_n - g\|_T^2 = O_P(k_n^2/n + M_n/n + k_n M_n^{-2s_g}). \quad \square$$

Table 1: Simulation 1, comparison of estimators

Estimator	ρ	β_1	β_2	β_3	β_4	$\overline{K}$	$\widetilde{K}$	$\sum IK$	$\%(\widehat{g}(T)=0)$
LS	0	0.9969	1.9932	2.9959	4.0084	0	0	0	0
	0.2	1.0212	1.9845	3.0084	3.9824	0	0	0	0
	0.5	1.0186	1.9748	3.0022	4.0098	0	0	0	0
	0.8	1.0252	1.9786	2.9961	4.0120	0	0	0	0
LS-AIC	0	0.9806	1.9948	2.9928	4.0161	4.71	5	0	57%
	0.2	1.0065	2.0099	2.9834	4.0107	4.78	5	0	56%
	0.5	1.0139	2.0016	2.9654	4.0195	4.84	5	0	74%
	0.8	1.0071	2.0181	2.9732	3.9989	4.95	5	0	78%
PLM	0	0.9903	1.9981	3.0038	3.9768	0	0	0	0
	0.2	0.9960	1.9857	3.0056	3.9983	0	0	0	0
	0.5	0.9790	2.0147	2.9858	4.0156	0	0	0	0
	0.8	1.0100	1.9726	3.0363	3.9772	0	0	0	0
LS-SCAD	0	0.9843	2.0135	2.9857	3.9889	4.60	5	0	61%
	0.2	0.9896	2.0250	2.9658	3.9926	4.53	5	0	75%
	0.5	0.9971	2.0175	2.9512	4.0149	4.75	5	0	73%
	0.8	1.0113	1.9907	3.0031	4.0018	4.64	5	0	64%
PLM-SCAD	0	0.9938	2.0000	3.0018	4.0050	4.41	5	0	0
	0.2	0.9723	2.0101	3.0083	4.0052	4.47	5	0	0
	0.5	0.9854	1.9822	3.0228	4.0095	4.72	5	0	0
	0.8	0.9934	1.9961	3.0090	3.9948	4.87	5	0	0

Table 2: Simulation 1, standard error estimate

ρ	$se(\beta_1)$	$\widehat{se}(\beta_1)$	$se(\beta_2)$	$\widehat{se}(\beta_2)$	$se(\beta_3)$	$\widehat{se}(\beta_3)$	$se(\beta_4)$	$\widehat{se}(\beta_4)$
0	0.1059	0.0969	0.1015	0.0983	0.1155	0.0978	0.1123	0.0978
0.2	0.1273	0.1002	0.1080	0.1036	0.0977	0.1014	0.1122	0.1028
0.5	0.1291	0.1131	0.1426	0.1268	0.1271	0.1266	0.1395	0.1153
0.8	0.1855	0.1613	0.2455	0.2104	0.2157	0.2150	0.1999	0.1767

Table 3: Simulation 2, comparison of estimators

Estimator	ρ	β_1	β_2	β_3	β_4	$\overline{K}$	$\widetilde{K}$	$\sum IK$	$\%(\widehat{g}(T)=0)$
LS	0	0.9032	2.1906	3.0205	4.0005	0	0	0	0
	0.2	0.8771	2.2009	2.9758	3.9851	0	0	0	0
	0.5	0.8113	2.2996	2.9471	3.9511	0	0	0	0
	0.8	0.5723	2.6451	2.7889	3.9200	0	0	0	0
LS-AIC	0	0.9230	2.1822	3.0578	4.0286	4.96	5	0	69%
	0.2	0.8908	2.1733	3.0175	3.9969	4.82	5	0	66%
	0.5	0.8130	2.3084	2.9525	3.9734	4.85	5	0	38%
	0.8	0.5742	2.5791	2.9127	3.8749	4.94	5	2	17%
PLM	0	0.9922	1.9965	3.0259	3.9898	0	0	0	0
	0.2	0.9903	1.9927	3.0047	4.0036	0	0	0	0
	0.5	0.9999	1.9917	2.9977	4.0101	0	0	0	0
	0.8	0.9684	2.0355	2.9970	4.0008	0	0	0	0
LS-SCAD	0	0.9403	2.1625	3.0763	4.0157	4.61	5	0	65%
	0.2	0.9003	2.1621	3.0295	3.9772	4.41	5	0	60%
	0.5	0.8112	2.2540	2.9974	3.9741	4.88	5	0	42%
	0.8	0.5480	2.6442	2.8314	3.9459	4.94	5	2	12%
PLM-SCAD	0	0.9780	1.9996	2.9895	4.0117	4.74	5	0	0
	0.2	1.0021	2.0049	3.0129	4.0003	4.44	5	0	0
	0.5	0.9899	2.0130	3.0046	4.0012	4.73	5	0	0
	0.8	1.0023	1.9870	2.9852	4.0297	4.63	5	0	0