Acta Mathematicae Applicatae Sinica, English Series
Vol. 33, No. 2 (2017) 373–388
DOI: 10.1007/s10255-017-0666-1
http://www.ApplMath.com.cn & www.SpringerLink.com

Variable Selection for the Partial Linear Single-Index Model

Wu WANG, Zhong-yi ZHU[†]

Department of Statistics, School of Management, Fudan University, Shanghai 200433, China
(E-mail: zhuzy@fudan.edu.cn)

Abstract In this paper, we consider the issue of variable selection in partial linear single-index models under the assumption that the vector of regression coefficients is sparse. We apply penalized spline to estimate the nonparametric function and SCAD penalty to achieve sparse estimates of regression parameters in both the linear and single-index parts of the model. Under some mild conditions, it is shown that the penalized estimators have oracle property, in the sense that it is asymptotically normal with the same mean and covariance that they would have if zero coefficients are known in advance. Our model owns a least square representation, therefore standard least square programming algorithms can be implemented without extra programming efforts. In the meantime, parametric estimation, variable selection and nonparametric estimation can be realized in one step, which incredibly increases computational stability. The finite sample performance of the penalized estimators is evaluated through Monte Carlo studies and illustrated with a real data set.

Keywords nonparametric link function; SCAD penalty; semiparametric model; spline estimation; variable selection

2000 MR Subject Classification 62F10; 62G10

1 Introduction

We consider a partial linear single-index model

$$E(Y|X,Z) = X^\top \beta_{\mathbf{0}} + \eta(\mathbf{Z}^\top \gamma_{\mathbf{0}}), \tag{1}$$

where β_0 and γ_0 are regression coefficients associated with covariates X and Z, respectively. We assume that the first element of γ_0 is positive and $\|\gamma_0\| = 1$ to ensure identifiability. In Model (1), X is linearly related to the mean response, while $Z^\top\gamma_0$ is nonlinearly related to the mean response with an unknown link function $\eta(\cdot)$. This model is quite flexible to include many known models. When the dimension of Z is one, it includes partial linear model, $E(Y|X,Z) = X^\top\beta + \eta(Z)$ (see [4,12,27] and the comprehensive book by [8] for details. As we all know, fully nonparametric model suffers from curse of dimensionality and cannot accommodate high dimensional covariate Z. By combining the multivariate predictors into a univariate index $Z^\top\gamma_0$, Model (1) avoids this problem, and still captures important features of high-dimensional data. When there is no linear term $X^\top\beta$, Model (1) also includes the single-index model, $E(Y|X,Z) = \eta(Z^\top\gamma_0)$, as its special case (see [2,16,20,23] for details).

Statistical inferences on the partial linear single-index model have been studied by some authors. [3] considered a generalized partial linear single-index model where they proposed a method to iteratively estimate the nonparametric function $\eta(.)$ and parameters (β_0, γ_0) based on local linear quasi-likelihood functions. [28] found that above approach was computationally

Manuscript received July 1, 2013. Revised October 31, 2014.
Supported by the National Natural Science Foundation of China (No. 11671096).
[†] Corresponding author.

unstable and proposed a penalized spline-based estimation of $\eta(.)$. [25] proposed a two-stage estimation procedure to estimate the link function and the parameters in the partial linear single-index model. [30] proposed a estimation and variable selection procedure based on minimum average variance estimation with measurement errors in the response and covariates. [6] considered semiparametric estimation in a partial linear single-index panel data model with fixed effects. [15] developed an efficient estimating equations procedure for performing variable selection and semiparametric efficient estimates for the heteroscedastic partial linear single-index model. [18] proposed a class of consistent estimators by using a proper weighting strategy with an unspecified error variance function.

In biomedical, environmental or econometric studies, there are many variables involved in the model, but generally the number of important variables is relatively small which means that the true model should be sparse. In these situations, variable selection, or more generally model selection, is often the most important objective, because the final model would be easy to interpret and accurate in prediction. Traditional variable selection methods include best subset, stepwise regression and so on. Stepwise regression methods are often trapped into a local minimum solution rather than global optimal solution as indicated out by [31]. Moreover, as pointed out by [7] and [31], these procedures also ignore the stochastic errors or uncertainty in the variable selection stage.

Penalty-based variable selection methods, such as [7,24,31], have gained a lot of attention for well-studied theoretical properties and computational advantages. So far, there are many works on the issue of variable selection under the framework of linear models, but similar works on Model (1) are not too much. [13] did some pioneer works. Their method was a discrete variable selection method which suffered from high computational burden and inherent high variability as mentioned above. [17] studied the SCAD penalty-based regression for Model (1), where kernel smoother-based profile local linear regression was employed for estimating nonparametric link function, and penalized least square estimator was employed for variable selection and parameter estimation. [29] studied the estimation and variable selection for a partial linear single-index model when some linear covariates are not observed but their ancillary variables are available. [14] extended studies on variable selection for partial linear single-index model to binary and count responses using kernel smoothers.

Besides the kernel smoother, spline smoother is another important nonparametric smoothing method. It is necessary and meaningful to understand the theoretic and practical properties of the spline smoother-based estimates for the partial linear single-index model. As indicated by [10,11], splines provide good approximations of a smoothing function with a small number of basis functions, and algorithms designed for parametric models can be used for spline estimators. [26] showed that penalized splines can outperform kernel methods in nonparametric models with clustered data. In this paper, we study issues of the variable selection and parameter estimation in Model (1) by combining the penalized spline-based nonparametric estimation and shrinkage variable selection methods. As we shown in the following context, our model owns a least square representation, and so standard least square programming algorithms can be directly implemented without extra programming effort. In the meantime, the estimator for the nonparametric link function, active variables as well as the estimators of their coefficients can be obtained in one step which increases computational stability. However, these good futures were not presented in kernel smoother-based methods[17]. Liang's kernel smoother-based profile local linear regression bears great computational burden, in each evaluation of Liang's objective function, the whole dataset has to be used O(n) times, while the dataset only has to be used O(1) times in each evaluation of our objective function. When the dataset is large, Liang's estimator could be infeasible.

We organize the rest of the paper as follows. In Section 2, we propose our estimation and variable selection procedure. The issues of computational details and some other practical problems are also discussed in this section. Large sample properties, e.g. the oracle property of proposed estimators, are investigated in Section 3. Simulation studies are conducted in Section 4 to illustrate the finite sample performance of the proposed estimation and variable selection procedure. In Section 5, we apply our proposed method to analyze the Boston housing data. Finally, a conclusion and future work are summarized in Section 6.

2 Model and Method

2.1 Estimating and Variable Selection Procedure

Without loss of generality, we consider a sample of n observations. For the ith observation, denote y_i as the response variable, by x_i the covariate vector of the linear effect and by z_i the covariate vector of the single-index. We consider the following model

$$y_i = x_i^\top \beta_0 + \eta(z_i^\top \gamma_0) + \varepsilon_i, \qquad 1 \le i \le n, \tag{2}$$

where β_0 is $p \times 1$ regression coefficients associated with covariate x, and γ_0 is $d \times 1$ regression coefficients associated with covariate z. The data $\{(y_i, x_i, z_i),\ i = 1, \cdots, n\}$ are independent and identically distributed. The noise ε_i is independent of (x_i, z_i) with mean 0 and variance σ^2, and $\eta_0(\cdot)$ is an unknown link function. Following [2] and [28], the unknown link function $\eta(\cdot)$ can be approximated by a m-degree spline function with K fixed knots $\Psi = \{\xi_1, \cdots, \xi_K\}$, that is $\eta(t) \approx \alpha^\top B(t)$, where

$$B(t) = (1, t, \cdots, t^m, (t-\xi_1)_+^m, \cdots, (t-\xi_K)_+^m)^\top \tag{3}$$

is a truncated power basis with knots $\xi_1, \cdots, \xi_K$, and α is the spline coefficients. To simplify notations, denote $\theta_\gamma = (\beta^\top, \gamma^\top, \alpha^\top)^\top$, and $v_i = (x_i^\top, z_i^\top)^\top$. Define the mean function

$$\mu(v_i, \theta_\gamma) = \alpha^\top B(\gamma^\top z_i) + \beta^\top x_i, \qquad 1 \le i \le n. \tag{4}$$

To selection important variables, [7] proposed the smoothly clipped absolute deviation(SCAD) penalty function:

$$p'_\lambda(\theta; \lambda, a) = \lambda I_{\{\theta \le \lambda\}} + \frac{(a\lambda - \theta)_+}{(a-1)} I_{\{\theta > \lambda\}}, \qquad \text{for} \quad \theta > 0, \quad a > 2, \tag{5}$$

where λ and a are tunning parameters. This penalty function can produce sparse estimation which automatically set small estimated coefficients to zero, while large coefficients are unbiased. In this article, we use the SCAD penalty to effectively select important variables. We set $a = 3.7$ as in [7].

The penalized least squares objective function for estimating β, γ and α is

$$Q(\theta_\gamma) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mu(v_i, \theta_\gamma))^2 + \tau\alpha^\top D\alpha + \sum_{j=1}^{p} p_{\lambda_j}(|\beta_j|) + \sum_{j=p+1}^{p+d} p_{\lambda_j}(|\gamma_j|), \tag{6}$$

where D is an positive semidefinite symmetric matrix. Following [22] and [28], we set D to be a diagonal matrix with its last K diagonal elements equal to 1 and the rest elements equal to 0. The first penalty is added to avoid the overfitting problem caused by using penalized splines to approximate the nonparametric function and we call it the smoothing penalty. The second and third penalty terms are shrinkage penalties on the linear parameters and single-index parameters, respectively. The smoothing parameter τ controls the smoothness of the nonparametric function fitting while tunning parameters $\lambda_j,\ j = 1, \cdots, p+d$ control the amount of shrinkage in the variable selection. Here we allow different penalties for the parameters in linear and single-index parts. For each evaluation of $Q(\theta_\gamma)$, we need to calculate the pseudo design points $B(z_i^\top \gamma), i = 1, \cdots, n$, in which we use the data m+K times, the dataset is used O(1) times in total. To evaluate [17]'s objective function, the profile estimator $\eta(.)$ must be evaluated at all the n index values $z_1\gamma^\top, \cdots, z_n\gamma^\top$, one evaluation of $\eta(z_i\gamma^\top)$ correspond to use the dataset once, the dataset is used O(n) times accordingly. Liang's algorithm could be infeasible when the dataset is very large.

2.2 Computation Algorithm

Recall that γ satisfies $\|\gamma\| = 1$ and $\gamma_1 > 0$, let ω be a $d-1$ dimension vector, we reparameterize γ as

$$\gamma(\omega) = \frac{(1\ \omega^\top)^\top}{\sqrt{1+\omega^\top\omega}}. \tag{7}$$

[28] also adopted this reparametrization. With this definition, $\gamma(\omega)$ is a bijection between unit vector γ with positive first element and $d-1$-dimensional vector ω without any restriction. After reparameterizing γ with ω, the objective function is now a function of ω and can be written as follows:

$$Q(\beta,\omega,\alpha) = \frac{1}{2n}\left\| \begin{pmatrix} Y \\ 0 \\ 0 \\ 0 \end{pmatrix} - \begin{pmatrix} \alpha^\top B(Z\gamma(\omega)) + X\beta \\ \sqrt{2np_\lambda(|\beta|)} \\ \sqrt{2np_\lambda(|\omega|)} \\ \sqrt{2n\tau} D\alpha \end{pmatrix} \right\|^2, \tag{8}$$

where $Y = (y_1, \cdots, y_n)^\top$, $\alpha^\top B(Z\gamma(\omega)) + X\beta = (\alpha^\top B(z_1^\top \gamma(\omega)) + x_1^\top \beta, \cdots, \alpha^\top B(z_n^\top \gamma(\omega)) + x_n^\top \beta,)^\top$, $\sqrt{2np_\lambda(|\beta|)} = (\sqrt{2np_{\lambda_1}(|\beta_1|)}, \cdots, \sqrt{2np_{\lambda_p}(|\beta_p|)})^\top$ and $\sqrt{np_\lambda(|\omega|)} = (\sqrt{np_{\lambda_{p+1}}(|\omega_1|)}, \cdots, \sqrt{np_{\lambda_{p+d-1}}(|\omega_{d-1}|)})^\top$. Minimization of $Q(\beta,\omega,\alpha)$ is a nonlinear least square problem, it can be solved by Gauss-Newton or Levenberg-Marquardt algorithms. In our experience, these algorithms may cause convergent problem in practice. We recommend using standard nonlinear least square functions such as lsqnonlin(.) or lsqcurvefit(.) in MATLAB to get better estimates.

2.3 Practical Issue

In practice, the effective estimation of the penalized partial linear single-index model relies on careful selection of the number and location of the knots, the smoothing parameter τ as well as tunning parameters λ_j. A detailed study of the choice of the number of knots K has been given by [21]. According to Ruppert's suggestion and our simulation investigation, for most cases, especially for monotonic or unimodal link functions, 5–10 knots are adequate. When the number of knots K is determined, we recommend that the knots should be placed at equally spaced sample quantiles of the single-index $z_i'\gamma$. During the estimation process, the knots change with the updating of γ.

The smoothing parameter τ controls the smoothness of nonparametric fitting, the tunning parameter λ_j control the amount of shrinkage in variable selection. There are many criteria can be used to select the tunning parameters, such as minimizing cross validation(CV) score, generalized cross validation(GCV) score, BIC and AIC. To choose τ and λ_j, a high dimension grid search can be applied, but it is rather time prohibitive. [19] proposed a double penalized variable selection procedure in partial linear model using smoothing spline to estimate the nonparametric effects. To select the smoothing parameter and tunning parameter, they first derived a linear mixed model representation under normal error assumptions and estimated the smoothing parameter as a variance component; by fixing the smoothing parameter, they proposed a BIC criteria to select the tunning parameter. We propose a two-step selection procedure as follows: first, we consider the penalized spline partial linear single-index model without shrinkage penalties, a good choice of the smoothing parameter τ is the minimizer of the following GCV score proposed by [28],

$$\mathrm{GCV}(\tau) = \frac{\frac{1}{n}\sum(y_i - \tilde{\alpha}^{\top}B(z_i^{\top}\gamma(\tilde{\omega})) - x_i^{\top}\tilde{\beta})^2}{(1 - \frac{1}{n}\mathrm{tr}A(\tau))^2}, \tag{9}$$

where $\tilde{\alpha}$, $\tilde{\omega}$, $\tilde{\beta}$ are the estimators with given τ and there is no shrinkage penalty in the objective function.The hat matrix $A(\tau)$ is

$$A(\tau) = A_s(\tau) + I - A_s(\tau)BB^{\top}(I - A_s(\tau))B^{-1}B^{\top}I - A_s(\tau), \tag{10}$$

where $A_s(\tau)$ is

$$A_s(\tau) = B(B^{\top}B + n\tau D)^{-1}B^{\top}. \tag{11}$$

Second, with the fixed $\widehat{\tau}$, we propose BIC score to choose λ_j. To reduce the dimension of tunning parameters, we set $\lambda_j = \lambda\mathrm{SE}(\tilde{\beta}_j), j = 1, \cdots, p$; $\lambda_j = \lambda\mathrm{SE}(\tilde{\omega}_j), j = p+1, \cdots, p+d-1$ as in [7], $\mathrm{SE}(\tilde{\beta}_j)$, $\mathrm{SE}(\tilde{\omega}_j)$ are the standard errors of estimators without shrinkage penalties. Define BIC score as

$$\mathrm{BIC}(\lambda) = \log\{Q_{\lambda}/n\} + \mathit{df}\log(n)/n, \tag{12}$$

where

$$Q_{\lambda} = (Y - X\widehat{\beta} - B(Z\gamma(\widehat{\omega}))\widehat{\alpha})^{\top}(Y - X\widehat{\beta} - B(Z\gamma(\widehat{\omega}))\widehat{\alpha}). \tag{13}$$

and *df* is the number of nonzero coefficients in $\widehat{\beta}$, $\widehat{\omega}$. The GCV and BIC score can select satisfactory smoothing parameter and tunning parameter in practice.

3 Asymptotic Theory and Oracle Properties

In this section, we will establish the asymptotic theory and oracle properties of the proposed estimator. Before presenting the results, we have to handle the constraints $\|\gamma\| = 1$ and $\gamma_{10} > 0$ on the single-index parameter γ. As in [2] and [28], define $\phi = (\phi_1, \cdots, \phi_{d-1})^{\top}$ be a d-1 dimensional vector and

$$\gamma_{\phi} = \begin{pmatrix} \sqrt{1 - \phi^{\top}\phi} \\ \phi_1 \\ \vdots \\ \phi_{d-1} \end{pmatrix}. \tag{14}$$

Note that γ_{ϕ} is equivalent to $\gamma(\omega)$ in the sense that there exists a one to one map between γ_{ϕ} and $\gamma(\omega)$, but γ_{ϕ} is more convenient in theoretical justification. The true parameter must satisfy the constraint $\|\phi_0\| < 1$, with this reparameterization, γ_{ϕ_0} satisfies all the constraints and is infinitely differentiable in a neighborhood of ϕ_0. Let $\theta_{\phi} = (\beta^{\top}, \phi^{\top}, \alpha^{\top})^{\top}$, and $\pi = (\beta^{\top}, \phi^{\top})^{\top}$. The dimension of π is $s = p + d - 1$, The mean function is

$$\mu(v_i; \theta_{\phi}) = \alpha^{\top}B(\gamma_{\phi}^{\top}z_i) + \beta^{\top}x_i. \tag{15}$$

The first derivative of the mean function with respect to θ_{ϕ} is

$$\frac{\partial\mu(V; \theta_{\phi})}{\partial\theta_{\phi}} = \begin{pmatrix} X \\ \alpha^{\top}B'(\gamma_{\phi}^{\top}Z)[-(1 - \phi^{\top}\phi)^{-\frac{1}{2}}\phi : \mathrm{I}_{d-1}]Z \\ B(\gamma_{\phi}^{\top}Z) \end{pmatrix}.$$

The Jacobian matrix from θ_γ to θ_ϕ is

$$\mathrm{J}(\phi) = \begin{pmatrix} \mathrm{I}_p & 0 & 0 \\ 0 & -(1-\phi^\top\phi)^{-\frac{1}{2}}\phi^\top & 0 \\ 0 & \mathrm{I}_{d-1} & 0 \\ 0 & 0 & \mathrm{I} \end{pmatrix}.$$

The penalized least squares objective function is

$$Q(\theta_\phi) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mu(v_i, \theta_\phi))^2 + \tau\alpha^\top D\alpha + \sum_{j=1}^{s} p_{\lambda_j}(|\pi_j|). \tag{16}$$

Let

$$L(\theta_\phi) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mu(v_i, \theta_\phi))^2 + \tau\alpha^\top D\alpha.$$

In the following, we denote θ_ϕ by θ. All of the assumptions are imposed on θ_ϕ and its corresponding parameter space Θ.

We require the following assumptions to derive asymptotic properties of the proposed estimator:

Assumption 1. The parameter space Θ is compact, and the true parameter θ_0 is an interior point of Θ.

Assumption 2. $\frac{1}{n}\sum_{i=1}^{n}(\mu(v_i;\theta_1) - \mu(v_i;\theta_2))^2$ converges to some limit function uniformly in θ_1, $\theta_2 \in \Theta$, and

$$\mathrm{R}(\theta) = \lim_{n\to\infty}\sum_{i=1}^{n}(\mu(v_i;\theta) - \mu(v_i;\theta_0))^2$$

has a unique minimum at $\theta = \theta_0$.

Assumption 3. The mean function is twice continuously differentiable in a neighborhood of θ_0, and

$$\Omega(\theta_0) = \lim_{n\to\infty}\frac{1}{n}\sum\frac{\partial\mu(v_i;\theta_0)}{\partial\theta}\frac{\partial\mu(v_i;\theta_0)^\top}{\partial\theta}$$

exists and is nonsingular. Furthermore,

$$\frac{1}{n}\sum\frac{\partial\mu(v_i;\theta)}{\partial\theta}\frac{\partial\mu(v_i;\theta)^\top}{\partial\theta} \qquad \text{and} \qquad \frac{1}{n}\sum\frac{\partial^2\mu(v_i;\theta)}{\partial\theta_j\partial\theta_k}$$

converge uniformly in θ in a neighborhood of θ_0.

Under above assumptions, [28] proved consistency and asymptotic normality of partial linear single-index model fitted by penalized splines. We cite their results as the following lemma:

Lemma 1. *Under assumptions A1–A3 and $\sqrt{n}\tau \to 0$, we have*

$$\sqrt{n}\frac{\partial L(\theta)}{\partial\theta}\Big|_{\theta_0} \longrightarrow_d N(0, \sigma^2\Omega(\theta_0))),$$
$$\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta^\top}\Big|_{\theta_0} \longrightarrow_p \Omega(\theta_0),$$

where $\to_d$ denotes 'convergence in distribution', $\to_p$ denotes 'convergence in probability'.

Divide $\pi_0 = (\beta_0^\top, \phi_0^\top)^\top$ into two parts, $\pi_0 = (\pi_{10}^\top, \pi_{20}^\top)^\top$, without loss of generality, assume that dimension of π_{10} is s_1, dimension of π_{20} is s_2, $s_1 + s_2 = s$ and $\pi_{20} = 0$. Parameter θ

can be divided into 3 parts, $\theta = (\pi_1^\top,\ \alpha^\top,\ \pi_2^\top)^\top = (\theta_1^\top, \theta_2^\top)^\top$, $\theta_1 = (\pi_1^\top,\ \alpha^\top)^\top$, $\theta_2 = \pi_2$, and $\theta_0 = (\theta_{10}^\top, 0)^\top$. Let Ω_{11} be submatrix of $\Omega(\theta)$ correspond to θ_1. Lemma 1 implies that $L'(\theta_0) = O_p(n^{-\frac{1}{2}})$, $L''(\theta_0) = \Omega(\theta_0) + o_p(1)$. Using these results, we can prove root-n consistency of $\widehat{\theta}$.

Theorem 1. *If $\sqrt{n}\tau \to 0$ and $\lambda_j \to 0$, then there exists a local minimizer $\widehat{\theta}$ of $Q(\theta)$ such that $\|\widehat{\theta} - \theta_0\| = O_p(n^{-\frac{1}{2}} + a_n)$, where $a_n = \max\{p'_{\lambda_j}(|\theta_{j0}|),\ \theta_{j0} \neq 0\}$.*

Proof. Denote $\varsigma_n = n^{-\frac{1}{2}} + a_n$, we only have to show that for any given $\varepsilon > 0$, there exists a large constant C such that

$$P\Big\{\sup_{\|u\|=C} Q(\theta_0 + \varsigma_n u) > Q(\theta_0)\Big\} \geq 1 - \varepsilon.$$

This implies that with probability at least $1-\varepsilon$ that exists a local minimum in the ball $\{\theta_0 + \varsigma_n u : \|u\| \leq C\}$.

Using $p_{\lambda_j}(0) = 0$, we have

$$\begin{aligned} D_n(u) =& Q(\theta_0 + \varsigma_n u) - Q(\theta_0) \\ \geq& L(\theta_0 + \varsigma_n u) - L(\theta_0) + \sum_{j=1}^{s_1}\{p_{\lambda_j}(|\pi_{10j} + \varsigma_n u_j|) - p_{\lambda_j}(|\pi_{10j}|)\}. \end{aligned}$$

By Taylor expansion we have

$$\begin{aligned} D_n(u) \geq& -\varsigma_n L'(\theta_0)^\top u + \frac{1}{2} u^\top \Omega(\theta_0) u \varsigma_n^2 \{1 + o_p(1)\} \\ &+ \sum_{j=1}^{s_1}\{\varsigma_n p'_{\lambda_j}(|\pi_{10j}|) u_j + \varsigma_n^2 p''_{\lambda_j}(|\pi_{10j}|) u_j^2 (1 + o(1))\}. \end{aligned}$$

Note that $L'(\theta_0) = O_p(n^{-1/2})$, the first term on the right hand side is of order $O_p(n^{-1/2}\varsigma_n)$, by choosing a sufficiently large C, the second term dominates the first term uniformly in $\|u\| = C$.

The third term is bounded by

$$\sqrt{s_1}\varsigma_n a_n \|u\| + \varsigma_n^2 \max\{p''_{\lambda_j}(|\pi_{10j}|) : \pi_{10j} \neq 0\}\|u\|^2,$$

because $\lambda_j \to 0$, $\max\{p''_{\lambda_j}(|\pi_{10j}|) : \pi_{10j} \neq 0\} \to 0$. Hence the third term is also dominated by the second term, this completes the proof. □

For SCAD penalty, $a_n = 0$ for large enough n, Theorem 1 implies that if we choose proper τ and λ_j, the penalized estimator $\widehat{\theta}$ is root-n consistent. Theorem 2 shows that $\widehat{\theta}$ can perform as well as the oracle estimator.

Theorem 2. *If $\sqrt{n}\tau \to 0$, $\lambda_j \to 0$ and $\sqrt{n}\lambda_j \to \infty$, then with probability tending to 1, the root-n consistent local minimizers $\widehat{\theta} = (\widehat{\theta}_1^\top, \widehat{\theta}_2^\top)^\top$ must satisfy*

$$\begin{aligned} &\widehat{\theta}_2 = 0, \\ &\sqrt{n}(\widehat{\theta}_1 - \theta_{10}) \to_d N(0, \sigma^2 \Omega_{11}). \end{aligned}$$

Proof. To prove part (a), we only have to show that for any θ_1 satisfying $\theta_1 - \theta_{10} = O_p(n^{-1/2})$ and some small $\varepsilon_n = Cn^{-1/2}$ for π_{20j} which is a zero component of π_{20},

$$\begin{aligned} &\frac{\partial Q(\theta)}{\partial \pi_{2j}} > 0, \qquad 0 < \pi_{2j} < \varepsilon_n, \\ &\frac{\partial Q(\theta)}{\partial \pi_{2j}} < 0, \qquad -\varepsilon_n < \pi_{2j} < 0. \end{aligned}$$

By Taylor's expansion, we have

$$\begin{aligned}\frac{\partial Q(\theta)}{\partial \pi_{2j}} =&\frac{\partial L(\theta)}{\partial \pi_{2j}} + p'_{\lambda_j}(|\pi_{2j}|)\mathrm{sgn}(\pi_{2j})\\ =&\frac{\partial L(\theta_0)}{\partial \pi_{2j}} + \sum_l \frac{\partial^2 L(\bar\theta)}{\partial \pi_{2j}\partial\theta_l}(\theta_l-\theta_{l0}) + p'_{\lambda_j}(|\pi_{2j}|)\mathrm{sgn}(\pi_{2j})\\ =&I_1+I_2+I_3,\end{aligned}$$

where $\bar\theta$ lies between θ and θ_0. By Lemma 1, we have $I_1 = O_p(n^{-1/2})$ and $\frac{\partial^2 L(\bar\theta)}{\partial\pi_{2j}\partial\theta_l} = O_p(1)$. So we have $I_2 = O_p(n^{-1/2})$. Therefore

$$\frac{\partial Q(\theta)}{\partial \pi_{2j}} = \lambda_j\Big\{\frac{p'_{\lambda_j}(|\pi_{2j}|)}{\lambda_j}\mathrm{sgn}(\pi_{2j}) + O_p(n^{-1/2}/\lambda_j)\Big\}.$$

Since for SCAD penalty, $\liminf\limits_{n\to\infty}\liminf\limits_{\theta\to 0^+} p'_{\lambda_j}(\theta) > 0$, and $\sqrt{n}\lambda_j \to \infty$, the sign of the derivative is completely determined by π_{2j}. This completes the proof of part (a).

Now we prove part (b) of Theorem 2. From Theorem 1 and the sparsity property of the proposed estimator, there exists a $\widehat\theta_1$ that is a root-n consistent local minimizer of $Q\{(\theta_1^\top, 0^\top)^\top\}$ that satisfies

$$\frac{\partial Q(\widehat\theta)}{\partial\theta_j} = 0, \qquad \text{for } \ j=1,\cdots,s_1+m+K+1,$$

where $\widehat\theta = (\widehat\theta_1^\top, 0^\top)^\top$. Note that $\widehat\theta_1$ is a consistent estimator, by Taylor expansion,

$$\begin{aligned}\frac{\partial Q(\widehat\theta)}{\partial\theta_j} =&\frac{\partial L(\widehat\theta)}{\partial\theta_j} - p'_{\lambda_j}(|\widehat\theta_j|)\mathrm{sgn}(\widehat\theta_j)\\ =&\frac{\partial L(\theta_0)}{\partial\theta_j} + \sum_{l=1}^{s}\Big\{\frac{\partial^2 L(\theta_0)}{\partial\theta_j\partial\theta_l} + o_p(1)\Big\}(\widehat\theta_l - \theta_{l0})\\ &- p'_{\lambda_j}(|\theta_{j0}|)\mathrm{sgn}(\theta_{j0}) + \{p''_{\lambda_j}(|\theta_{j0}|) + o_p(1)\}(\widehat\theta_j - \theta_{j0}).\end{aligned}$$

For SCAD penalty, when n is large enough, $p'_{\lambda_j}(|\theta_{j0}|)$ and $p''_{\lambda_j}(|\theta_{j0}|)$ equals 0 exactly. By Lemma 1 and Slutsky's theorem, we have

$$\sqrt{n}(\widehat\theta_1 - \theta_{10}) \to_d N(0,\, \sigma^2\Omega_{11}).$$

This completes the proof. □

Theorem 1 shows that under mild conditions the proposed estimator is root-n consistent if we choosing proper τ and λ_j. Furthermore, theorem 2 shows that the root-n consistent estimator must satisfy $\widehat\theta_2 = 0$ in probability tending to 1 and $\widehat\theta_1$ is asymptotic normal with covariance matrix Ω_{11}. This means that the proposed estimator performs as well as if $\theta_2 = 0$ is known in advance, this is called the oracle property in [7].

4 Monte Carlo Study

In this section, Monte Carlo studies are presented to illustrate the finite sample performance of the proposed estimation and variable selection procedure. We use cubic penalized spline with 10 knots to approximate the nonparametric link function, smoothing parameter and tunning parameter are selected by the GCV and BIC score proposed in Section 2.

Table 1. Simulation Results for Scenario 1

		γ			β		
		MRME	C	I	MRME	C	I
$n = 100$	OP	0.21	4.00	0.00	0.46	6.00	0.00
$\sigma = 0.1$	OL	0.26	4.00	0.00	0.28	6.00	0.00
	PP	0.28	3.47	0.00	0.61	5.33	0.10
	PL	0.37	3.60	0.08	0.91	5.32	0.29
$n = 100$	OP	0.11	4.00	0.00	0.38	6.00	0.00
$\sigma = 0.25$	OL	0.20	4.00	0.00	0.27	6.00	0.00
	PP	0.63	3.34	0.00	0.86	5.14	0.76
	PL	0.73	3.29	0.30	0.86	4.91	1.02
$n = 100$	OP	0.29	4.00	0.00	0.64	6.00	0.00
$\sigma = 0.5$	OL	0.15	4.00	0.00	0.28	6.00	0.00
	PP	0.77	1.54	0.37	0.86	3.26	1.11
	PL	0.90	1.90	0.40	0.94	3.69	1.45
$n = 100$	OP	0.28	4.00	0.00	0.44	6.00	0.00
$\sigma = 1$	OL	0.20	4.00	0.00	0.27	6.00	0.00
	PP	0.95	2.03	0.68	0.92	4.36	2.03
	PL	0.97	1.02	0.61	0.93	3.43	1.59
$n = 200$	OP	0.14	4.00	0.00	0.34	6.00	0.00
$\sigma = 0.1$	OL	0.27	4.00	0.00	0.34	6.00	0.00
	PP	0.34	3.79	0.00	0.54	5.61	0.02
	PL	0.33	3.89	0.02	0.85	5.55	0.02
$n = 200$	OP	0.10	4.00	0.00	0.32	6.00	0.00
$\sigma = 0.25$	OL	0.31	4.00	0.00	0.39	6.00	0.00
	PP	0.29	3.67	0.00	0.54	5.58	0.83
	PL	0.36	3.86	0.03	0.94	5.50	0.57
$n = 200$	OP	0.48	4.00	0.00	0.50	6.00	0.00
$\sigma = 0.5$	OL	0.26	4.00	0.00	0.34	6.00	0.00
	PP	0.72	1.94	0.20	0.96	3.56	0.78
	PL	0.89	2.83	0.23	0.96	4.23	1.06
$n = 200$	OP	0.31	4.00	0.00	0.65	6.00	0.00
$\sigma = 1$	OL	0.08	4.00	0.00	0.30	6.00	0.00
	PP	0.85	2.06	0.58	0.97	4.57	1.92
	PL	0.99	1.21	0.42	0.88	3.79	1.64

Table 2. Simulation Results for Scenario 2

		γ			β		
		MRME	C	I	MRME	C	I
$n = 100$	OP	0.29	4.00	0.00	0.47	6.00	0.00
$\sigma = 0.1$	OL	0.29	4.00	0.00	0.35	6.00	0.00
	PP	0.30	3.77	0.00	0.54	5.44	0.00
	PL	0.36	3.75	0.05	0.88	5.44	0.19
$n = 100$	OP	0.25	4.00	0.00	0.46	6.00	0.00
$\sigma = 0.25$	OL	0.24	4.00	0.00	0.24	6.00	0.00
	PP	0.37	3.44	0.00	0.86	4.97	0.03
	PL	0.66	3.47	0.27	0.94	5.11	1.07
$n = 100$	OP	0.37	4.00	0.00	0.28	6.00	0.00
$\sigma = 0.5$	OL	0.15	4.00	0.00	0.33	6.00	0.00
	PP	0.75	0.94	0.18	0.62	4.25	1.32
	PL	0.99	1.85	0.42	0.95	3.92	1.40
$n = 100$	OP	0.39	4.00	0.00	0.30	6.00	0.00
$\sigma = 1$	OL	0.14	4.00	0.00	0.35	6.00	0.00
	PP	0.97	1.78	0.42	0.62	4.92	2.26
	PL	0.95	0.89	0.46	0.93	3.56	1.60
$n = 200$	OP	0.18	4.00	0.00	0.44	6.00	0.00
$\sigma = 0.1$	OL	0.31	4.00	0.00	0.36	6.00	0.00
	PP	0.19	3.87	0.00	0.50	5.72	0.00
	PL	0.36	3.91	0.01	0.79	5.64	0.01
$n = 200$	OP	0.21	4.00	0.00	0.49	6.00	0.00
$\sigma = 0.25$	OL	0.32	4.00	0.00	0.30	6.00	0.00
	PP	0.27	3.59	0.00	0.74	5.21	0.00
	PL	0.40	3.87	0.03	0.85	5.53	0.50
$n = 200$	OP	0.36	4.00	0.00	0.27	6.00	0.00
$\sigma = 0.5$	OL	0.26	4.00	0.00	0.37	6.00	0.00
	PP	0.56	0.61	0.14	0.83	4.30	0.52
	PL	0.90	2.80	0.21	0.95	4.26	1.05
$n = 200$	OP	0.36	4.00	0.00	0.33	6.00	0.00
$\sigma = 1$	OL	0.10	4.00	0.00	0.38	6.00	0.00
	PP	0.80	1.41	0.28	0.60	5.12	1.71
	PL	0.93	1.26	0.33	0.90	4.02	1.58

We repeat 500 times from the following model:

$$y_i = \sin\left\{\frac{\pi(z_i^T\gamma - A)}{B - A}\right\} + x_i^\top\beta + \sigma\varepsilon, \qquad i = 1, \cdots, n, \tag{17}$$

where ε_i is standard normally distributed, constants $A = 0.3912$, $B = 1.3409$. The mean function has the coefficients $\beta = (3, 2, 0, 0, 0, 1.5, 0, 0.2, 0.3, 0.15, 0, 0)^\top$, $\gamma = (1, 3, 1.5, 0.5, 0, 0, 0, 0)^\top / \sqrt{12.5}$. The sample size n is set to be 100 and 200 respectively with $\sigma = 0.1$, 0.25, 0.5 and 1. As in [17], we generate the linear and single-index covariates from the following 3 scenarios to assess the robustness of the estimates: (1) The covariate X and Z are independent and uniformly distributed on [0,1]; (2) The first five and last five elements of X are independent and standard normally distributed, the 6th and 7th elements are independently Bernoulli distributed with success probability 0.5; (3) The covariates Z are independent and uniformly distributed on [0,1], covariates $X = W + \{1.5\exp 1.5z_1, 5z_1, 5\sqrt{z_2}, 3z_1 + z_2^2, 0, 0, 0, 0, 0, 0, 0, 0\}$, where covariates W are generated from a 12-dimensional normal distribution with mean 0 and variance 0.25, the correlation between w_i and w_j is $0.4^{|i-j|}$.

To assess the performance of the estimators, we consider the median of relative model error in [7] and [17], and define relative model error as RME=ME/ME$_{full}$, ME$_\beta = E[X\beta - X\widehat{\beta}]^2$, ME$_\gamma = E[Z\gamma - Z\widehat{\gamma}]^2$. A full model estimator is calculated by fitting the data with all variables in the model, while an oracle estimator is calculated by fitting the model without all the irrelevant variables. MRME values smaller than one indicates that the estimate performs better than the unpenalized estimator. Let C be the average number of true zero coefficients that were correctly set to zero, I be the average number of true nonzero coefficients incorrectly set to zero. In Tables 1–3, OP stands for the oracle estimator of proposed method, OL stands for the oracle estimator of Liang's method, PP stands for the penalized estimator of the proposed method, PL stands for the penalized estimator of Liang's method.

Table 3. Simulation Results for Scenario 3

		γ			β		
		MRME	C	I	MRME	C	I
$n = 100$	OP	0.33	4.00	0.00	0.28	6.00	0.00
$\sigma = 0.1$	OL	0.28	4.00	0.00	0.15	6.00	0.00
	PP	0.43	3.60	0.01	0.43	5.08	0.00
	PL	0.48	3.67	0.03	0.82	5.24	0.05
$n = 100$	OP	0.51	4.00	0.00	0.28	6.00	0.00
$\sigma = 0.25$	OL	0.19	4.00	0.00	0.18	6.00	0.00
	PP	0.65	3.70	0.01	0.80	5.06	0.60
	PL	0.50	3.35	0.21	0.85	4.99	0.56
$n = 100$	OP	0.35	4.00	0.00	0.11	6.00	0.00
$\sigma = 0.5$	OL	0.11	4.00	0.00	0.31	6.00	0.00
	PP	0.82	1.50	0.42	0.95	3.62	1.22
	PL	0.86	1.61	0.34	0.92	3.82	1.02
$n = 100$	OP	0.42	4.00	0.00	0.21	6.00	0.00
$\sigma = 1$	OL	0.16	4.00	0.00	0.37	6.00	0.00
	PP	0.99	1.94	0.69	0.87	4.31	2.04
	PL	0.99	0.86	0.47	0.84	3.85	1.66
$n = 200$	OP	0.18	4.00	0.00	0.34	6.00	0.00
$\sigma = 0.1$	OL	0.32	4.00	0.00	0.18	6.00	0.00
	PP	0.33	3.97	0.00	0.38	5.52	0.00
	PL	0.39	3.89	0.00	0.73	5.52	0.01
$n = 200$	OP	0.29	4.00	0.00	0.38	6.00	0.00
$\sigma = 0.25$	OL	0.29	4.00	0.00	0.17	6.00	0.00
	PP	0.47	3.90	0.00	0.69	5.44	0.17
	PL	0.39	3.80	0.04	0.83	5.29	0.12
$n = 200$	OP	0.42	4.00	0.00	0.11	6.00	0.00
$\sigma = 0.5$	OL	0.27	4.00	0.00	0.41	6.00	0.00
	PP	0.70	1.66	0.42	0.93	3.76	0.76
	PL	0.92	2.66	0.14	0.90	4.43	0.56
$n = 200$	OP	0.47	4.00	0.00	0.15	6.00	0.00
$\sigma = 1$	OL	0.12	4.00	0.00	0.40	6.00	0.00
	PP	0.86	2.08	0.67	0.89	4.50	1.90
	PL	0.98	1.16	0.32	0.93	4.09	1.42

In summary, we can see from Tables 1–3 that the MRME values of all of the proposed estimators are less than one, which indicates that the proposed penalized estimators perform better than the unpenalized estimators, regardless of the sample size or noise level. Comparing to [17], the proposed estimator has better performance in reducing model errors in almost all the three scenarios. As for the estimation of the single- index parameter γ, our estimator outperforms Liang's estimator in terms of reducing the number 'I', which means our estimator has less true nonzero coefficients that is incorrectly set to zero. Note that set a nonzero coefficients into zero is more problematic than include a irrelevant variable in a regression model, it introduces biases into the estimators and the estimators may not be consistent. The number 'I' act as the counterpart of power in hypothesis testing. When the standard error of noise σ increases, it generally becomes harder to estimate the single index coefficients. The correctly selected zero coefficients are less than 2 obtained by both our method and Liang's method in most cases. The incorrectly selected nonzero coefficients of β increase as σ increases, the small coefficients of β are harder to identify. In scenario 1, the proposed method performed better than Liang's method in terms of reducing model error; In scenario 2, the proposed method performs better in terms of reducing the number 'I'. For example, when n is 100 and σ^2 is 0.25, the number 'I' of β is 1.07 and the number 'I' of γ is 0.27 for Liang's estimator. While for our estimator, the number 'I' of β is 0.03 and the number 'I' of γ is 0. When σ increases to 0.5, our estimator perform better than Liang's estimator in estimating β, our estimator has higher 'C' and lower 'I'. When σ increases to 1, our estimator has better performance in estimating α,while the incorrectly selected nonzero elements of β also increase and larger than Liang's. In scenario 3, our method performs better in specifying true zero coefficients and true nonzero coefficients of γ. In all the three scenarios, the proposed estimator improves as the sample size increases and the error variance σ^2 decreases as the our theory predicted.

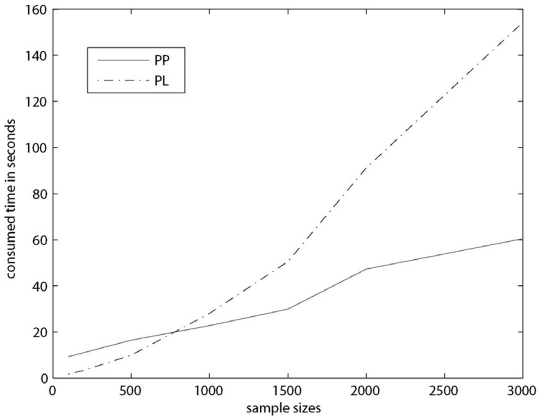

Fig 1. Computational Time in Different Sample Sizes. PP Stands For The Proposed Method, PL Stands For Liang'S Method

We compare the computational cost in penalized spline estimator of our proposed method and kernel estimator of [17]'s method in scenario 1 with $\sigma = 0.5$ of one simulation run, since the computational cost is very similar in other cases, we will not show them here. We run both procedures in Matlab version 8.1.0.604 on a Linux platform with AMD CPU Opteron 6212(2.7GHz), the median CPU times in 200 simulation runs are recorded in Fig 1. When the sample sizes are small, Liang's procedure actually has faster speed, but the computational cost of Liang's procedure increases rapidly as sample size increases, when the sample size are 3000, Liang's procedure takes 154 seconds in one simulation run while our procedure takes 60 seconds.

5 Data Example

In this section we apply the penalized partial linear single-index estimator to the Boston Housing data analyzed in [9]. The dependent variable is the logarithm of the median value of owner occupied homes in each of the 506 census tracts in Boston Standard Metropolitan Statistical Areas, 13 independent variables are: RM, average number of rooms in owner units; AGE, proportion of owner units built prior to 1940; B, black proportion of the population; LSTAT, proportion of population that is in the lower status; CRIM, crime rate by town; ZN, proportion of town' residential land zoned for lots greater than 25,000 square feet; INDUS, proportion of nonretail business acres per town; TAX, full property tax rate; PTRATIO, pupil-teacher ratio by town school district; CHAS, Charles River dummy: =1 if tract bound the Charles River, =0 if otherwise; DIS, weighted distances to five employment centers in the Boston region; RAD, index of accessibility to radial highways; NOX, nitrogen oxide concentration in pphm.

[9] measured the willingness to pay for clear air, the final model was

$$\begin{aligned}\log(\mathrm{MV}) =& a_1 + a_2\mathrm{RM}^2 + a_3\mathrm{AGE} + a_4\log(\mathrm{DIS}) + a_5\log(\mathrm{RAD}) + a_6\mathrm{TAX}\\ &+ a_7\mathrm{PTRATIO} + a_8(\mathrm{B} - 0.63)^2 + a_9\log(\mathrm{LSTAT}) + a_{10}\mathrm{CRIM}\\ &+ a_{11}\mathrm{ZN} + a_{12}\mathrm{INDUS} + a_{13}\mathrm{CHAS} + a_{14}\mathrm{NOX}^p + e.\end{aligned}$$

[5] analyzed Boston Housing data set, where sliced inverse regression and forward stepwise regression were used, they found that the main contributors to the housing values were CRIM, LSTAT, RM. [25] found that some of the variables were discrete and the assumptions of sliced inverse regression may not met. They constructed a partial linear single-index model with only the CHAS variable as linear covariates and all other variables as single-index, they found that the nonparametric function $\eta(.)$ was a nonincreasing function but with a upward trend on the right tail.

As in [9], we also make a logarithm transformation to the dependent variable. The final model of [9] includes 7 linear effects: AGE, TAX, PTRATIO, CRIM, ZN, INDUS, CHAS; and 6 nonlinear effects: RM, DIS, RAD, B, LSTAT, NOX. We construct a partial linear single-index model where we put all linear effects of [9] in the linear part, and put all nonlinear effects of [9] in the single-index part. We use cubic penalized spline with 10 knots to approximate the nonparametric link function $\eta(.)$, use GCV score to select the smoothing parameter, and use BIC score to select the tunning parameters. The selected value for the smoothing parameter is 2.5×10^{-4}, the tunning parameter is 0.41. We also estimate with [17]'s kernel method, the bandwidth is selected to be 0.3.

Fig.2 depicts the estimated nonparametric function $\eta(.)$, and shows that $\eta(.)$ is a nonincreasing function, Our estimate confirms [25]'s claim that the upward curvature of the function at high values of the single-index value may not be true. While Liang's estimate has this upward trend. Since $\eta(.)$ is monotone, all the single-index have the simple explanation as effects[16], means that covariates with positive parameters have negative effect on housing values, while covariates with negative parameters have positive effect on housing values. The estimation and variable selection result are recorded in Table 4. Comparing to Liang's result, we select one more variable CHAS, which is a indicator variable records whether the tract bounds the river. Its coefficient is estimated positive which means that it has a positive effect on housing prices. This result is consistent with [9] which pointed out that including of this variable captures the amenities of a riverside location. Three variables are ruled out, which are AGE, ZN and INDUS. The air pollution covariate NOX has a negative effect on housing values, which confirms the important result of [9] that we are paying prices for clear air.

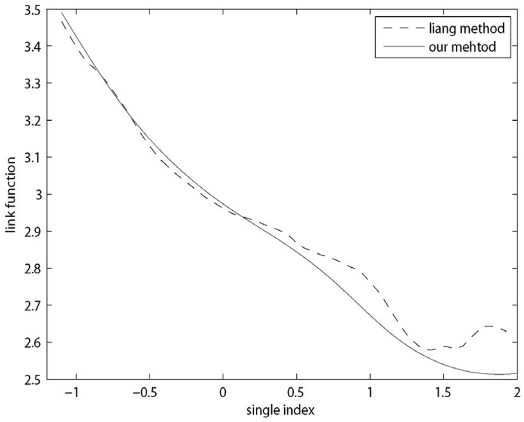

Fig 2. The Estimation for the Nonparametric Link Function $\eta(.)$

Table 4. Parameter Estimation for Boston Housing Data, s.e. Stands for Standard Error

	Liang			New		
var	no penalty	s.e.	penalized	no penalty	s.e.	penalized
NOX	0.118	0.031	0.103	0.174	0.031	0.174
RM	−0.227	0.053	−0.213	−0.181	0.029	−0.203
DIS	0.163	0.039	0.162	0.290	0.028	0.285
RAD	−0.525	0.072	−0.504	−0.357	0.047	−0.333
B	−0.152	0.031	−0.117	−0.115	0.060	−0.080
LSTAT	0.780	0.115	0.806	0.844	0.180	0.854
AGE	−0.001	0.012	-	−0.003	0.132	-
TAX	−0.137	0.015	−0.130	−0.086	0.016	−0.045
PTRATIO	−0.059	0.008	−0.062	−0.055	0.009	−0.016
CRIM	−0.088	0.014	−0.090	−0.089	0.009	−0.049
ZN	0.011	0.007	-	0.021	0.016	-
INDUS	0.008	0.010	-	0.001	0.009	-
CHAS	0.060	0.026	-	0.081	0.010	0.030

Remark. [9] studied the Boston housing data and found that several variables were essentially nonlinearly related to the housing price especially the air pollutant proxy NOX. To add more evidence that a linear model is not enough to model the Boston housing data, we simply use a test statistic to test a linear parametric model against a nonparametric alternative. The test statistic was proposed by [1]. It was designed to test parameter constancy in linear models, but it also had nontrivial local power in detecting nonlinearity in regression functions. Let $X = (x_1, \dots, x_n)$ be the design matrix and $X_k = (x_1, \dots, x_k)$ be the first k rows of the design matrix. First we fit a linear regression model $y = X\beta + \varepsilon$, let $\widehat{\beta}$ be an estimator of β and $\widehat{\varepsilon}_t = y_t - x_t'\widehat{\beta}$. Define the process T_n^* by

$$T_n^*(k/n, z) = (X'X)^{-1/2}\Big(\sum_{t=1}^{k} x_t I(\widehat{\varepsilon}_t \le z) - (X_k'X_k)(X'X)^{-1}\sum_{t=1}^{n} I(\widehat{\varepsilon}_t \le z)\Big)$$

and the test statistic by

$$M_n^* = \max_k \sup_z ||T_n^*(k/n, z)||_\infty,$$

where $\|\cdot\|_\infty$ denotes the maximum norm. The test statistic M_n^* has nonstandard asymptotic distributions, [1] provided a table of critical values. For the Boston housing data, the test statistic $M_n^* = 2.242$ and the critical value is 1.091 at the significance level 0.01. Thus the null hypothesis of linear model is rejected at the significance level 0.01, we'd better use a nonlinear model for the data.

6 Conclusion

We propose a simultaneous estimation and variable selection procedure for partial linear single-index model by combing penalized spline nonparametric estimation and shrinkage variable selection methods. Under certain conditions, we show consistency and oracle property of the estimators. The penalized spline approach owns a least squares representation, therefore standard software can be implemented and estimation procedure is computationally expedient and stable in practice. The simulation study and data example presented illustrate the effectiveness of our method.

The proposed procedure assume that the errors are independent, it can be generalized to a longitudinal data model, in which observations are correlated within each subjects,

$$g(\mu_{ij}) = x_{ij}^\top + \eta(z_{ij}^\top \gamma),$$

in which g is a known link function, and $\eta(.)$ is a unknown nonparametric function. [3] studied generalized partial linear single-index model but the observations are uncorrelated. Variable selection in generalized longitudinal data model deserves further study.

References

[1] Bai, J. Testing for parameter constancy in linear regressions: an empirical distribution function approach. *Econometrica*, 64: 597–622 (1996)

[2] Bai, Y., Fung, W.K., Zhu, Z.Y. Penalized quadratic inference functions for single-index models with longitudinal data. *Journal of Multivariate Analysis*, 100: 152–161 (2009)

[3] Carroll, R.J., Fan, J., Gijbels, I., Wand, M.P. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92: 477–489 (1997)

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

[5] Chen, C.H., Li, K.C. Can SIR be as popular as multiple linear regression. *Statistica Sinica*, 8: 289–316 (1998)

[6] Chen, J., Gao, J., Li, D. Estimation in partially linear single-index panel data models with fixed effects. *Journal of Business and Economic Statistics*, 31: 315–330 (2013)

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

[8] Hardle, W., Gao, J., Liang, H. Partially Linear Models. Springer-Verlag, New York, 2000

[9] Harrison, D., Rubinfeld, D.L. Hedonic housing prices and the demand for clean air. *Environmental Economics Management*, 5: 81–102 (1978)

[10] He, X., Fung, W.K., Zhu, Z.Y. Robust Estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association*, 100: 1176–1184 (2005)

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

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

[13] Kong, E., Xia, Y. Variable selection for the single-index model. *Biometrika*, 94: 217–229 (2006)

[14] Lai P., Tian, Y., Lian, H. Estimation and variable selection for generalised partially linear single-index models. *Journal of Nonparametric Statistics*, 26: 171–185 (2014)

[15] Lai, P., Wang Q., Zhou, X. Variable selection and semiparametric efficient estimation for the heteroscedastic partially linear single-index model. *Computational Statistics and Data Analysis*, 70: 241–256 (2014)

[16] Li, K.C. Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86: 316–342 (1991)

[17] Liang, H., Liu, X., Li, R., Tsai, C. Estimation and testing for partially linear single-index models. *The Annals of Statistics*, 38: 3811–3836 (2010)

[18] Ma, Y., Zhu, L. Doubly robust and efficient estimators for heteroscedastic partially linear single-index model allowing high-dimensional covariates. *Journal of the Royal Statistical Society, Series B*, 75: 305–322 (2013)

[19] Ni, X., Zhang, H., Zhang, D. Automatic model selection for partially linear models. *Journal of Multivariate Analysis*, 100: 2100–2111 (2009)

[20] Powell, J.L., Stock, J.H., Stoker, T.M. Semiparametric estimation of index coefficient. *Econometrica*, 51: 1403–1430 (1989)

[21] Ruppert, D. Selecting the Number of Knots for Penalized Splines. *Journal of Computational and Graphical Statistics*, 4: 735–757 (2002)

[22] Ruppert, D., Carroll, R. Penalized Regression Splines. Working paper, Cornell University, School of Operations Research and Industrial Engineering, available at www.orie.cornell.edu/1davidr/papers, 1997

[23] Stoker, T.M. Consistent estimation of scaled coefficients. *Econometrica*, 54: 1461–1481 (1986)

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

[25] Wang, J.L., Xue, L.G., Zhu, L.X., Chong, Y.S. Estimation for a partial-linear single-index model. *The Annals of Statistics*, 38: 246–274 (2010)

[26] Welsh, A.H., Lin, X., Carroll, R.J. Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel methods. *Journal of the American Statistical Association*, 97: 482–493 (2002)

[27] Xie, H., Huang, J. SCAD-penalized Regression in High-dimensional Partially Linear Models. *The Annals of Statistics*, 37: 673–696 (2009)

[28] Yu, Y., Ruppert, D. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97: 1042–1054 (2002)

[29] Zhang, J., Wang X., Yu, Y., Gai, Y. Estimation and variable selection in partial linear single index models with error-prone linear covariates. *Statistics*, 48: 1048–1070 (2014)

[30] Zhang, J., Yu, Y., Zhu, L., Liang, H. Partial linear single index models with distortion measurement errors, *Annals of the Institute of Statistical Mathematics*, 65: 237–267 (2013)

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