

B spline variable selection for the single index models

Jianbo Li^{1,2} · Yuan Li¹ · Riquan Zhang³

Received: 15 June 2014 / Revised: 9 October 2015 / Published online: 30 October 2015
© Springer-Verlag Berlin Heidelberg 2015

Abstract Through the nonconcave penalized least squares method, we consider the variable selection in the full nonparametric regression models with the B spline-based single index approximation. Under some regular conditions, we show that the resulting estimates with SCAD and HARD thresholding penalties enjoy \sqrt{n} -consistency and oracle properties. We use some simulation studies and a real example to illustrate the performance of our proposed variable selection procedure.

Keywords Single index model · Lasso · SCAD · Hard Thresholding · Oracle

Mathematics Subject Classification Primary 62G08 · Secondary 62H12

1 Introduction

Single index model and its corresponding statistical inference methods have been extensively investigated over last three decades by many statisticians. For example, [Powell et al. \(1989\)](#) incorporated the average derivative method and kernel smoothing technique to estimate the single index coefficients. [Ichimura \(1993\)](#) gave a least squares estimate by virtue of kernel smoothing technique; [Härdle et al. \(1993\)](#) con-

✉ Jianbo Li
ljianb66@gmail.com

¹ School of Economics and Statistics, Guangzhou University, Guangzhou 510006, Guangdong, People's Republic of China

² School of Mathematics and Statistics, Jiangsu Normal University, Xuzhou 221116, People's Republic of China

³ School of Finance and Statistics, East China Normal University, Shanghai 200241, People's Republic of China

sidered the optimal smoothness of the single index function; Li (1991) discussed the estimation of single index coefficients by sliced regression approach. Xia et al. (2002) studied several index vector estimation methods by a minimum average variance estimation method. Other estimation methods are included in Hall (1989), Klein and Spady (1993), Horowitz and Härdle (1996), Carroll et al. (1997), Xia and Li (1999), Hristache et al. (2001) etc.

Wang and Yang (2009) proposed a B spline-based estimation method for the following fully nonparametric heteroscedastic regression model

$$Y = m(X) + \sigma(X)\varepsilon, \quad m(X) = E(Y|X) \quad (1)$$

through the single index approximation

$$g(v) = E(Y|X'\beta = v), \quad (2)$$

where $Y \in R$ is a response variable, $X \in R^p$ is the corresponding covariate vector, $g(\cdot)$ is a completely unspecified univariate function, termed by single index function, $\beta = (\beta_1, \beta_2, \dots, \beta_p)' \in R^p$ is the regression parameter vector, termed by index parameter vector; X' represents the transpose of X ; ε is the model random error and usually assumed to follow a distribution with $E(\varepsilon|X) = 0$ and $Var(\varepsilon|X) = 1$. For the model identification, it is also assumed that $\|\beta\| = 1$ with last nonzero component positive. Hereafter $\|\cdot\|$ stands for the Euclid norm. There are mainly two advantages of the model (1) with the approximation (2). For one thing, it can avoid the model misspecification; For another one, it can also avoid the so-called ‘‘curse of dimensionality’’ in the nonparametric regression estimation.

In applications, we usually include as many covariates as possible into the working models to improve the modeling accuracy and then establish a high-dimensional statistical model. However, some of the included covariates may be unimportant, which, in turn, may increase estimate variance and so may lead to wrong statistical inference. Therefore variable selection is an important step before application of such models. In the context of linear regression models, many variable selection techniques were proposed and some of them have been extended into the context of semiparametric and nonparametric models. For example, LASSO (Tibshirani 1996, 1997; Knight and Fu 2000; Ciuperca 2014), SCAD (Fan 1997; Fan and Li 2001, Fan and Li 2002; Xu et al. 2014; Neykov et al. 2014), Hard thresholding (Fan 1997; Antoniadis 1997; Fan and Li 2002), Adaptive LASSO (ALASSO) penalty (Zou 2006; Zhang and Lu 2007; Lu and Zhang 2007; Zhang et al. 2010), Dantzig selector (Candes and Tao 2007; Antoniadis et al. 2010) etc. Fan and Lv (2010) overviewed the variable selection methods in details. Variable selection in the single index models also has been studied by some statisticians. Kong and Xia (2007) proposed a separated cross validation-based variable selection method; Zhu et al. (2011) considered the ALASSO approach for a general class of single index models; Zeng et al. (2012) studied the variable selection by a local linear smoothing approximation. Wang (2009) incorporated Bayesian method into the variable selection. Peng and Huang (2011) used a penalized least squares method and local linear approximation to select the important variables in single index model. In

their methods, the bandwidth is a key for the convergence rate of resulting estimates and the implementation may suffer from intensive computation.

In this paper, by combining the B spline-based estimation approach in Wang and Yang (2009) and the nonconcave penalized least squares method in Fan and Li (2001), we consider the variable selection in model (1) with the single index approximation (2), which will be approximated by the B spline polynomial. One advantage of B spline approximation technique is that given the index regression vector β , we can choose a B spline basis and then the unknown single index function can be characterized by the basis expansion coefficients. So we can obtain the approximated estimate of the unknown multivariate function $m(\cdot)$ by estimating the base expansion coefficients of $g(\cdot)$ using the commonly-used least squares method. Thus we can easily implement the proposed variable selection method in the current context. Under some regular conditions, we will show that the resulting estimates with the SCAD and HARD thresholding penalties enjoy the oracle and consistency properties. Using the proposed method, we not only can select the important single index variables but also can estimate the unknown single index functions and the regression parameters simultaneously. Some simulation studies and a real data application will be given to illustrate our proposed variable selection method.

The rest of this paper precedes as follows. In Sect. 2, we will introduce our proposed variable selection method for model (1) with the single index model approximation (2), including B spline approximation technique, penalized least squares method, the main theoretical results and an efficient implementation algorithm. Some numerical studies will be given in Sect. 3. Some conclusions will be made in Sect. 4. We will present the corresponding theoretical proofs in Appendix.

2 Method

2.1 Penalized B spline estimation

It is assumed that $\{(Y_i, X_i)\}_{i=1}^n$ are the realizations of (Y, X) specified by the model (1). Without loss of generality, we assume $\|\beta\| = 1$ with $\beta_p > 0$ and the true value β_0 is in the upper unit hemisphere $S_+^{p-1} = \{\beta : \|\beta\| = 1, \beta_p > 0\}$, otherwise, we can interchange the positions of the last covariate and the one with positive effect. Denote $S_c^{p-1} = \{\beta = (\beta_1, \beta_2, \dots, \beta_p)' : \|\beta\| = 1, \beta_p \geq \sqrt{1 - c^2}\}$ with $c \in (0, 1)$ as a cap shape subset of S_+^{p-1} . With a proper value of c , it is obvious that $\beta_0 \in S_c^{p-1}$. Since S_+^{p-1} is not a compact set, we assume $\beta \in S_c^{p-1}$ in what follows. To define a B spline expansion of $g(v)$, we also suppose that there exists a real positive number M such that $P(\|X\| \leq M) = 1$. Consequently, $X'\beta$ is bounded in some finite interval, say $[a, b]$, with probability 1 for all $\beta \in S_c^{p-1}$.

Following Wang and Yang (2009), we can estimate β by minimizing the following risk function

$$\mathcal{R}(\beta) = E[Y - m_\beta(X_\beta)]^2$$

or the corresponding empirical risk function,

$$R(\beta) = \frac{1}{n} \sum_{i=1}^n [Y_i - m_\beta(X_{\beta,i})]^2,$$

where $X_\beta = X'\beta$, $X_{\beta,i} = X_i'\beta$, $m_\beta(X_\beta) = E(Y|X_\beta) = E[m(X)|X_\beta]$.

To stable the following B spline polynomial expansion of $m_\beta(\cdot)$, Wang and Yang (2009) used the rescaled centered Beta cumulative density function to transform the covariate X_β such as

$$U_\beta = F_p(X_\beta), \quad U_{\beta,i} = F_p(X_{\beta,i}), \quad i = 1, 2, \dots, n,$$

where

$$F_p(v) = \int_{-1}^{v/a} \frac{\Gamma(p+1)}{\Gamma[(p+1)/2]2^{2p}} (1-t^2)^{\frac{p-1}{2}} dt, \quad v \in [-a, a].$$

Then, for fixed β , U_β has a quasi-uniform[0, 1] distribution. So equally-spaced knots can be used to smooth the unknown function in the B spline approximation of $m_\beta(\cdot)$. In terms of U_β , the empirical risk function $R(\beta)$ can be rewritten as

$$R(\beta) = \frac{1}{n} \sum_{i=1}^n [Y_i - \gamma_\beta(U_{\beta,i})]^2, \tag{3}$$

where $\gamma_\beta(\cdot) = m_\beta(F_p^{-1}(\cdot))$ and it is suggested to be approximated by B spline approximation.

Denote r as the order of B spline approximation. Let $\xi_1 = \xi_2 = \dots = \xi_r = a < \xi_{r+1} < \xi_{r+2} < \dots < \xi_{r+N} < b = \xi_{r+N+1} = \xi_{r+N+2} = \dots = \xi_{2r+N}$ be the knot points for the B spline approximation, where $N = n^v$ with $0 < v < 0.5$ such that $\max_{1 \leq k \leq N+1} \{|\xi_{r+k} - \xi_{r+k-1}|\} = O(n^{-v})$. Usually we call $\{\xi_{r+i}\}_{i=1}^N$ as the inner knot points. The number of inner knots, N , can be chosen as a positive integer number between $n^{1/6}$ and $n^{1/5} \log^{-2/5}(n)$. Denote $\{B_j(x)\}_{j=1}^d$ as the B spline basis functions based on the knot set $\{\xi_i\}_{i=1}^{2r+N}$, where $d = r + N$ is the dimension of B spline basis. Following deBoor (1978), the B spline basis functions enjoy the following properties: (i) $B_j(x) = 0$ for $x \notin [\xi_j, \xi_{j+r}]$; (ii) $B_j(x) > 0$ for $x \in [\xi_j, \xi_{j+r}]$; (iii) $\sum_{j=1}^d B_j(x) = 1$ for any $x \in [a, b]$ and 0 otherwise. Consequently, for any $1 \leq j \leq d$ and any $x \in R$, we have $B_j(x) \in [0, 1]$.

Given the B spline basis $\{B_j(x)\}_{j=1}^d$, $\gamma_\beta(\cdot)$ can be approximated by

$$\tilde{\gamma}_\beta(v) = \theta_{\beta,1}B_1(v) + \theta_{\beta,2}B_2(v) + \dots + \theta_{\beta,d}B_d(v) = B(v)'\theta_\beta, \tag{4}$$

where $\theta_\beta = (\theta_{\beta,1}, \theta_{\beta,2}, \dots, \theta_{\beta,d})'$ and $B(x) = (B_1(x), B_2(x), \dots, B_d(x))'$. Hereafter the subscript β indicates that the corresponding quantity depends on the value of β . Plugging (4) into (3), the empirical risk function (3) can be approximated by

$$\tilde{R}(\beta) = \frac{1}{n} \sum_{i=1}^n [Y_i - B(U_{\beta,i})'\theta_\beta]^2. \tag{5}$$

According to the least squares method, θ_β can be estimated by $\hat{\theta}_\beta = [B'_\beta B_\beta]^{-1} B'_\beta Y$ for the fixed β , where $B_\beta = [B(U_{\beta,1}), B(U_{\beta,2}), \dots, B(U_{\beta,n})]'$ and $Y = (Y_1, Y_2, \dots, Y_n)'$. Then the empirical risk function (5) can be estimated by

$$\hat{R}(\beta) = \frac{1}{n} \sum_{i=1}^n [Y_i - B(U_{\beta,i})' \hat{\theta}_\beta]^2. \tag{6}$$

Let $\hat{\beta}$ be the minimizer of $\hat{R}(\beta)$. $g(\cdot)$ can be estimated by

$$\hat{g}(v) = \tilde{\gamma}_{\hat{\beta}}(F_p^{-1}(v)) = B(F_p^{-1}(v))' \hat{\theta}_{\hat{\beta}}. \tag{7}$$

Considering the model restriction $\|\beta\| = 1$ with $\beta_p > 0$, denote $\beta^{(1)} = (\beta_1, \beta_2, \dots, \beta_{p-1})$. Then the index regression parameter vector β can be rewritten as $\beta = \left(\beta^{(1)'}, \sqrt{1 - \|\beta^{(1)}\|^2}\right)'$ with $\|\beta^{(1)}\| < 1$, that is, the free index regression parameters in (2) are just $\beta^{(1)}$. Let $R^*(\beta^{(1)}) = \hat{R}(\beta^{(1)}, \sqrt{1 - \|\beta^{(1)}\|^2})$. Adding the penalty term to the estimated risk function $R^*(\beta^{(1)})$, the penalized risk function is given by

$$Q(\beta^{(1)}) = R^*(\beta^{(1)}) + \sum_{j=1}^{p-1} p_\lambda(|\beta_j|), \tag{8}$$

where $\lambda > 0$ is the tuning parameter and $p_\lambda(\cdot)$ is a penalty function given λ .

Given a proper penalty function $p_\lambda(\cdot)$, we can obtain the penalized least squares estimate of $\beta^{(1)}$ by minimizing the function (8) with respect to $\beta^{(1)}$. To achieve effective variable selection for model (1), the penalty function $p_\lambda(\cdot)$ should be irregular at the origin, that is, $p_\lambda(0+) > 0$ (Fan and Li 2002). Let $\hat{\beta}_n^{(1)}$ be the minimizer of (8) and then $g(\cdot)$ can be estimated by (7) with $\hat{\beta}_n = \left(\hat{\beta}_n^{(1)'}, \sqrt{1 - \|\hat{\beta}_n^{(1)}\|^2}\right)$. With a proper penalty function $p_\lambda(\cdot)$ and a tuning parameter λ , some components of $\hat{\beta}_n^{(1)}$ are shrunk to 0 and so the corresponding covariates will disappear in model (1), which reaches the variable selection. In this paper, we consider three commonly-used penalty functions: SCAD, Hard thresholding and Lasso. Their formula and the corresponding nice properties can be found in Fan and Li (2001).

In what follows, let $\beta_0^{(1)}$ be the true value of $\beta^{(1)}$. Without loss of generality, we partition $\beta_0^{(1)} = \left(\beta_{10}^{(1)'}, \beta_{20}^{(1)'}\right)'$ such that $\beta_{10}^{(1)}$ contains all the nonzero effects in $\beta_0^{(1)}$ and $\beta_{20}^{(1)} = \mathbf{0}$ contains all the zero ones. We also assume that the length of $\beta_{10}^{(1)}$ is s . Correspondingly, $\beta^{(1)}$ and $\hat{\beta}_n^{(1)}$ also have the same partitions, namely, $\beta^{(1)} = \left(\beta_1^{(1)'}, \beta_2^{(1)'}\right)'$, $\hat{\beta}_n^{(1)} = \left(\hat{\beta}_{1n}^{(1)'}, \hat{\beta}_{2n}^{(1)'}\right)'$, where $\beta_1^{(1)'}$ and $\hat{\beta}_{1n}^{(1)'}$ respectively consist of the first s components of $\beta^{(1)}$ and $\hat{\beta}_n^{(1)}$.

Denote

$$a_n = \max\{\dot{p}_{\lambda_n}(|\beta_{j,10}^{(1)}|) : \beta_{j,10}^{(1)} \neq 0\} \quad \text{and} \quad b_n = \max\{|\ddot{p}_{\lambda_n}(|\beta_{j,10}^{(1)}|)| : \beta_{j,10}^{(1)} \neq 0\},$$

where $\beta_{j,10}^{(1)}$ is the j th component of $\beta_{10}^{(1)}$, \dot{p}_{λ_n} and \ddot{p}_{λ_n} respectively are the first- and second-order derivatives of p_{λ_n} . Then we have the following results.

Theorem 1 Under conditions (A1)-(A6) in Wang and Yang (2009), if $b_n \rightarrow 0$, then there exists a minimizer $\hat{\beta}_n^{(1)}$ of $Q(\beta^{(1)})$ such that $\|\hat{\beta}_n^{(1)} - \beta_0^{(1)}\| = O_p(n^{-1/2} + a_n)$.

To present the oracle properties of $\hat{\beta}_n^{(1)}$ denote

$$\tilde{\Sigma}_{\lambda_n} = \text{diag} \left\{ \ddot{p}_{\lambda_n}(|\beta_{1,10}^{(1)}|), \ddot{p}_{\lambda_n}(|\beta_{2,10}^{(1)}|), \dots, \ddot{p}_{\lambda_n}(|\beta_{s,10}^{(1)}|) \right\}$$

and

$$\tilde{\mathbf{b}}_{\lambda_n} = \left[\dot{p}_{\lambda_n}(|\beta_{1,10}^{(1)}|)\text{sign}(\beta_{1,10}^{(1)}), \dot{p}_{\lambda_n}(|\beta_{2,10}^{(1)}|)\text{sign}(\beta_{2,10}^{(1)}), \dots, \dot{p}_{\lambda_n}(|\beta_{s,10}^{(1)}|)\text{sign}(\beta_{s,10}^{(1)}) \right]'$$

Theorem 2 (Oracle properties) Assume that the penalty function $p_{\lambda_n}(\cdot)$ satisfies that

$$\liminf_{n \rightarrow \infty} \liminf_{\theta \rightarrow 0+} \dot{p}_{\lambda_n}(\theta)/\lambda_n = c,$$

where c is a positive constant. If $\lambda_n \rightarrow 0$, $\sqrt{n}\lambda_n \rightarrow \infty$ and $a_n = O(n^{-1/2})$, then under the conditions of Theorem 1, the \sqrt{n} -consistent local minimizer $\hat{\beta}_n^{(1)} = (\hat{\beta}_{1n}^{(1)}, \hat{\beta}_{2n}^{(1)})'$ in Theorem 1 must satisfy:

- (i) (Sparsity) $P(\hat{\beta}_{2n}^{(1)} = \mathbf{0}) \rightarrow 1$;
- (ii) (Asymptotic Normality)

$$\sqrt{n}(V + \tilde{\Sigma}_{\lambda_n}) \left[\hat{\beta}_{1n}^{(1)} - \beta_{10}^{(1)} + (V + \tilde{\Sigma}_{\lambda_n})^{-1} \tilde{\mathbf{b}}_{\lambda_n} \right] \rightarrow N(\mathbf{0}, A) \tag{9}$$

where V and A are the s th leading submatrix of $E[H^*(\beta_0^{(1)})]$ and $\text{Cov}[S^*(\beta_0^{(1)})]$, $H^*(\beta^{(1)})$ and $S^*(\beta^{(1)})$, the Hessian matrix and score function of $R^*(\beta^{(1)})$, are given in Sect. 2.2.

Remark 1 For SCAD and HARD thresholding penalties, when $|\beta_i|$ is large enough, they are constant, which implies that $\tilde{\mathbf{b}}_{\lambda_n} = \mathbf{0}$ and $\tilde{\Sigma}_{\lambda_n} = \mathbf{0}$. So for the variable selection based on SCAD and HARD thresholding penalties, we have

$$\sqrt{n} \left(\hat{\beta}_{1n}^{(1)} - \beta_{10}^{(1)} \right) \rightarrow N \left(\mathbf{0}, V^{-1'} A V^{-1} \right).$$

Remark 2 For LASSO penalty, $a_n = \lambda_n$, which implies that there is a contradict between $a_n = o(n^{-1/2})$ for estimation consistency in Theorem 1 and $\sqrt{n}\lambda_n \rightarrow \infty$ for oracle properties in Theorem 2. Therefore the resulting consistent estimates based on LASSO penalty can not enjoy oracle properties.

2.2 Implementation

For the minimization of $Q(\beta^{(1)})$, we suggest to use the port optimization-based algorithm in Wang and Yang (2009) with the objective function $Q(\beta^{(1)})$ and the gradient vector

$$\begin{aligned}
 S(\beta^{(1)}) &= \frac{\partial Q(\beta^{(1)})}{\partial \beta^{(1)}} = \frac{\partial R^*(\beta^{(1)})}{\partial \beta^{(1)}} + \mathbf{b}_\lambda(\beta^{(1)}) = S^*(\beta^{(1)}) + \mathbf{b}_\lambda(\beta^{(1)}) \\
 &= -\frac{2}{n} \sum_{i=1}^n \left[Y_i - B(U_{\beta,i})' \hat{\theta}_\beta \right] \dot{B}(U_{\beta,i})' \hat{\theta}_\beta \dot{F}_p(X_i' \beta) J_{\beta^{(1)}} X_i + \mathbf{b}_\lambda(\beta^{(1)}),
 \end{aligned}
 \tag{10}$$

where $\mathbf{b}_\lambda(\beta^{(1)}) = [\dot{p}_\lambda(|\beta_1|)\text{sign}(\beta_1), \dot{p}_\lambda(|\beta_2|)\text{sign}(\beta_2), \dots, \dot{p}_\lambda(|\beta_{p-1}|)\text{sign}(\beta_{p-1})]'$, $J_{\beta^{(1)}} = \frac{\partial \beta}{\partial \beta^{(1)}} = (\gamma_1, \gamma_2, \dots, \gamma_p)^T$, $\gamma_i (i \leq p - 1)$ is a $(p - 1)$ -dimensional zero column vector with i th element being 1, $\gamma_p = -\beta^{(1)} / \sqrt{1 - \|\beta^{(1)}\|^2}$, $\dot{B}(v) = (\dot{B}_1(v), \dot{B}_2(v), \dots, \dot{B}_d(v))'$, $\dot{B}_i(v)$, and $\dot{F}_d(\cdot)$ are the first-order derivatives of the B spline basis function $B_i(v)$, and the Beta cumulative density function $F_d(\cdot)$.

In the view of Newton-Raphson iterative algorithm, $\hat{\beta}_n^{(1)}$ can be seen as the values of $\beta^{(1)}$ at the convergence in the following iteration,

$$\beta^{(1)} = \beta_0^{(1)} - \left[H^*(\beta_0^{(1)}) + \Sigma_\lambda(\beta_0^{(1)}) \right]^{-1} \left[S^*(\beta_0^{(1)}) + \mathbf{b}_\lambda(\beta_0^{(1)}) \right],$$

where

$$\begin{aligned}
 H^*(\beta^{(1)}) &= \frac{\partial^2 R^*(\beta^{(1)})}{\partial \beta^{(1)} \partial \beta^{(1)}} = H_1(\beta^{(1)}) + H_2(\beta^{(1)}) + H_3(\beta^{(1)}), \\
 H_1(\beta^{(1)}) &= \frac{2}{n} \sum_{i=1}^n [\dot{B}(U_{\beta,i})' \hat{\theta}_\beta \dot{F}_p(X_i' \beta)]^2 J_{\beta^{(1)}} X_i X_i' J_{\beta^{(1)}}', \\
 H_2(\beta^{(1)}) &= -\frac{2}{n} \sum_{i=1}^n \left[Y_i - B(U_{\beta,i})' \hat{\theta}_\beta \right] \ddot{B}(U_{\beta,i})' \hat{\theta}_\beta [\dot{F}_p(X_i' \beta)]^2 J_{\beta^{(1)}} X_i X_i' J_{\beta^{(1)}}', \\
 H_3(\beta^{(1)}) &= -\frac{2}{n} \sum_{i=1}^n \left[Y_i - B(U_{\beta,i})' \hat{\theta}_\beta \right] \dot{B}(U_{\beta,i})' \hat{\theta}_\beta \ddot{F}_p(X_i' \beta) J_{\beta^{(1)}} X_i X_i' J_{\beta^{(1)}}', \\
 \Sigma_\lambda(\beta^{(1)}) &= \text{diag} \left\{ \dot{p}_\lambda(|\beta_1^{(1)}|) / |\beta_1^{(1)}|, \dot{p}_\lambda(|\beta_2^{(1)}|) / |\beta_2^{(1)}|, \dots, \dot{p}_\lambda(|\beta_{p-1}^{(1)}|) / |\beta_{p-1}^{(1)}| \right\},
 \end{aligned}$$

$\ddot{B}(v) = [\ddot{B}_1(v), \ddot{B}_2(v), \dots, \ddot{B}_d(v)]'$, $\ddot{B}_i(v)$ and $\ddot{F}_d(\cdot)$ are the second-order derivatives of the B spline basis function $B_i(v)$ and the Beta cumulative density function $F_d(\cdot)$. Thus the covariance of the estimate $\hat{\beta}_{1n}^{(1)}$ can be estimated by

$$\text{Cov}(\hat{\beta}_{1n}^{(1)}) = \left[A(\hat{\beta}_{1n}^{(1)}) + C_\lambda(\hat{\beta}_{1n}^{(1)}) \right]^{-1} \text{Cov} \left(D(\hat{\beta}_{1n}^{(1)}) \right) \left[A(\hat{\beta}_{1n}^{(1)}) + C_\lambda(\hat{\beta}_{1n}^{(1)}) \right]^{-1},
 \tag{11}$$

where $A(\beta_1^{(1)})$ and $C_\lambda(\beta_1^{(1)})$ are the s th leading submatrix of $H^*(\beta^{(1)})$ and $\Sigma_\lambda(\beta^{(1)})$ with $\beta_2^{(1)} = \mathbf{0}$; $D(\beta_1^{(1)})$ is a vector consisting of the first s components of $S^*(\beta^{(1)})$ with $\beta_2^{(1)} = \mathbf{0}$. This formula is consistent with the results of Theorem 2. Our numerical studies show that the formula performs very well for the finite sample size.

To sum up, the implementation of the proposed variable selection can be summarized as the following three steps.

- Step 1* Given $\beta \in S_c^{p-1}$, by use of Steps 1-2 in Wang and Yang (2009), we obtain the transformed single index variable U_β and the number of inner knots.
- Step 2* Given the tuning parameter λ , we obtain the penalized estimate of $\beta^{(1)}$ by minimizing $R^*(\beta^{(1)})$ through the port optimization with the linear model least squares estimate as the initial value of $\beta^{(1)}$ and the score function (10).
- Step 3* With the penalized estimate of $\beta^{(1)}$ in Step 2, we can obtain the estimate of single index function through (7) and the covariance matrix of $\hat{\beta}_{1n}^{(1)}$ can be estimated through the formula (11).

Another issue for the variable selection procedure is the selection of tuning parameter λ . Following Wang et al. (2007), we use the following Bayesian information criteria (BIC) to select the tuning parameter λ ,

$$\text{BIC}(\lambda) = R^*(\hat{\beta}_n^{(1)}) + df_n \frac{\log(n)}{n},$$

where df_n is the approximated degree of freedom for model (1) and it can be estimated by the number of nonzero element in $\hat{\beta}_n^{(1)}$. The advantage of BIC is that it tends to identify the true model if the true model is included in the candidate model set.

3 Numerical studies

In this section, we present some simulation examples and an application to illustrate our proposed variable selection method. We use the median of model prediction error (MMPE), $E[(\hat{\beta}_n - \beta_0)' \Sigma_X (\hat{\beta}_n - \beta_0)]$ with $\Sigma_X = E(XX')$, over 500 runs to evaluate the efficiency of our proposed variable selection procedure. The code was compiled using R. It's available for any readers on requirement.

3.1 Simulation examples

We generate a sample $\{(X_i, Y_i)\}_{i=1}^n$ of (X, Y) from the following model

$$Y = m(X) + \sigma(X)\varepsilon,$$

where $X = (X_1, X_2, \dots, X_5)'$ *i.i.d.* $N(0, 1)$, truncated by $[-2.5, 2.5]$, $m(X) = X'\beta + 4 \exp\{-(X'\beta)^2\} + \delta\|X\|$, $\varepsilon \sim N(0, 1)$ and $\beta = (1, 0, -1, 0, 1)'/\sqrt{3}$. When $\delta = 0$, this model is just the single index model. This model was ever used in Wang and Yang (2009) and Xia et al. (2004).

Table 1 Variable selection results (I)

Penalty	Lasso			Hard			Scad			Oracle		
	MMPE	Corr.	ICorr.	MMPE	Corr.	ICorr.	MMPE	Corr.	ICorr.	MMPE	Corr.	ICorr.
<i>n</i>												
100	0.0055	3.0000	0.3520	0.0035	3.0000	0.0580	0.0059	3.0000	0.0368	0.0060	3.0000	0.0000
150	0.0039	3.0000	0.2540	0.0022	3.0000	0.0520	0.0044	3.0000	0.0330	0.0038	3.0000	0.0000
200	0.0026	3.0000	0.2300	0.0013	3.0000	0.0500	0.0029	3.0000	0.0240	0.0029	3.0000	0.0000
300	0.0020	3.0000	0.2280	0.0012	3.0000	0.0480	0.0018	3.0000	0.0168	0.0015	3.0000	0.0000

Table 2 Variable selection results (II)

Model	$\delta = 1, \sigma(x) = 1$						$\delta = 0, \sigma(x) = \frac{1-0.2 \exp\{\ x\ /\sqrt{5}\}}{5+\exp\{\ x\ /\sqrt{5}\}}$					
	Scad			Oracle			Scad			Oracle		
<i>n</i>	MMPE	Corr.	ICorr.	MMPE	Corr.	ICorr.	MMPE	Corr.	ICorr.	MMPE	Corr.	ICorr.
100	0.0066	3.0000	0.0434	0.0066	3.0000	0.0000	0.0031	3.0000	0.0280	0.0035	3.0000	0.0000
150	0.0044	3.0000	0.0352	0.0046	3.0000	0.0000	0.0018	3.0000	0.0080	0.0020	3.0000	0.0000
200	0.0034	3.0000	0.0304	0.0032	3.0000	0.0000	0.0012	3.0000	0.0022	0.0015	3.0000	0.0000
300	0.0013	3.0000	0.0301	0.0019	3.0000	0.0000	0.0009	3.0000	0.0006	0.0010	3.0000	0.0000

In this study, we consider four sample sizes: $n = 100, 150, 200, 300$ and two cases of $\sigma(x)$: $\sigma(x) = 1, \sigma(x) = \frac{1-0.2 \exp\{\|x\|/\sqrt{5}\}}{5+\exp\{\|x\|/\sqrt{5}\}}$. We run each case 500 times. The index function $g(v)$ will be approximated by the cubic B spline technique. We choose tuning parameter λ in all the cases by BIC. We study all the three type of variable selection for the single index models ($\delta = 0, \sigma(x) = 1$). For the models away from single index model and heteroscedastic single index model, we only present the variable selection results based on SCAD penalty.

The summarized results are displayed in Tables 1, 2, 3 and 4. Table 1 shows the variable selection results in the single index model ($\delta = 0, \sigma = 1$). Table 2 displays the variable selection results based on SCAD penalty in the models with ($\delta = 1, \sigma(x) = 1.0$) and ($\delta = 0, \sigma(x) = \frac{1-0.2 \exp\{\|x\|/\sqrt{5}\}}{5+\exp\{\|x\|/\sqrt{5}\}}$). In the two tables, “MMPE”, “Corr.” and “ICorr.” respectively stand for the median of model prediction errors, the average number of nonzero effects correctly detected and the average number of zero effects incorrectly detected by our variable selection procedures. Tables 3 and 4 summarize the estimation results of nonzero effects β_1 and β_3 in the three model cases.

From Table 1, we can see that the variable selection procedure in all the cases can select the same number of important variables. Moreover, the average number of covariates incorrectly detected decreases as the increasing of sample size. In addition, we also find that the medians of model prediction errors decrease as the increasing of sample size and they are reasonably close to the oracle estimation results. Moreover, for all the sample sizes, the median of the model prediction error with Lasso penalty is most far away from the median with oracle estimation while the median with SCAD penalty is closest to the median with oracle estimation. Table 2 reveals that our proposed

Table 3 Summarized estimation results (I)

Penalty	$n = 100$		$n = 150$		$n = 200$		$n = 300$	
	β_1	β_3	β_1	β_3	β_1	β_3	β_1	β_3
Lasso								
Bias	-0.0119	0.0083	-0.0064	0.0044	-0.0076	0.0073	-0.0078	0.0063
SSTD	0.0695	0.0680	0.0368	0.0365	0.0291	0.0295	0.0247	0.0259
MSTD	0.0523	0.0582	0.0424	0.0432	0.0312	0.0312	0.0261	0.0268
Hard								
Bias	0.0001	-0.0028	0.0005	0.0012	0.0013	-0.0007	0.0020	-0.0013
SSTD	0.0395	0.0407	0.0353	0.0324	0.0267	0.0259	0.0222	0.0229
MSTD	0.0435	0.0434	0.0404	0.0409	0.0380	0.0368	0.0339	0.0331
Scad								
Bias	-0.0014	0.0029	-0.0007	0.0011	-0.0026	0.0024	-0.0017	0.0033
SSTD	0.0401	0.0400	0.0345	0.0374	0.0318	0.0323	0.0253	0.0255
MSTD	0.0439	0.0422	0.0407	0.0405	0.0376	0.0386	0.0331	0.0337
Oracle								
Bias	0.0014	-0.0022	0.0003	-0.0014	0.0023	-0.0026	0.0020	-0.0033
SSTD	0.0379	0.0421	0.0317	0.0317	0.0264	0.0289	0.0224	0.0232
MSTD	0.0443	0.0431	0.0402	0.0408	0.0364	0.0377	0.0339	0.0334

variable selection method also performs satisfactory for the true models away from single index model and heteroscedastic single index model.

Tables 3 and 4 summarize the estimation results of nonzero effects β_1 and β_3 . In the table, “Bias”, “SSTD” and “MSTD” respectively represent the estimation bias, sample standard deviation and mean of estimated standard deviation based on the formula (11).

From Table 3, we can see that all the estimation biases are reasonably small and perform very similarly to the oracle estimation especially for the variable selection methods with SCAD and HARD penalties. In most cases, their absolute values decrease as the increasing of sample size. Ignoring the random error, “SSTD” can be seen as the true value of standard deviation. All “MSTD” and “SSTD” are reasonably as small as that based on the oracle estimation especially for large sample size cases. Moreover, their values are vary close in all cases. In addition, “MSTD” and “SSTD” based on the variable selection method with SCAD penalty perform most similarly to the oracle estimation method even for relative small sample size. Table 4 suggests that our proposed method also performs very well in estimation when the models are away from single index models.

In a word, the variable selection procedure with all the three penalties can perform very well and similarly to the oracle estimation according to the variable selection and estimation. Moreover, the variable selection procedure with SCAD outperforms the other two procedures. Therefore we suggest ones to use the variable selection with SCAD penalty in applications.

Table 4 Summarized estimation results (II)

Penalty	$n = 100$		$n = 150$		$n = 200$		$n = 300$	
	β_1	β_3	β_1	β_3	β_1	β_3	β_1	β_3
$\delta = 1, \sigma = 1$								
Scad								
Bias	-0.0017	0.0024	-0.0103	0.0102	-0.0142	0.0064	-0.0085	0.0117
SSTD	0.0713	0.0773	0.0411	0.0421	0.0372	0.0355	0.0279	0.0286
MSTD	0.0750	0.0715	0.0458	0.0443	0.0322	0.0403	0.0246	0.0238
Oracle								
Bias	-0.0020	-0.0026	0.0040	0.0008	0.002	-0.0024	0.0029	-0.0012
SSTD	0.0740	0.0720	0.0367	0.0365	0.0335	0.0329	0.0265	0.0265
MSTD	0.0657	0.0671	0.0477	0.0467	0.0365	0.0365	0.0288	0.0295
$\delta = 0, \sigma = \frac{1-0.2 \exp(\ x\ /\sqrt{5})}{5+\exp(\ x\ /\sqrt{5})}$								
Scad								
Bias	-0.0003	0.0014	-0.0004	-0.0002	-0.0014	0.0020	-0.0024	0.0022
SSTD	0.0306	0.0276	0.0231	0.0248	0.0104	0.0123	0.0085	0.0085
MSTD	0.0316	0.0267	0.0244	0.0250	0.0103	0.0140	0.0084	0.0082
Oracle								
Bias	0.0003	-0.0044	0.0011	-0.0023	0.0020	-0.0028	0.0023	-0.0017
SSTD	0.0274	0.0289	0.0247	0.0243	0.0131	0.0115	0.0082	0.0086
MSTD	0.0349	0.0349	0.0274	0.0283	0.0102	0.0105	0.0086	0.0076

3.2 Application

In this section, we use our proposed variable selection procedures to analyze the body fat data set (Penrose et al. 1985). This data set involves 252 observations with 13 covariates (*age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm and wrist*). This data can be available from R package “*mfp*”. In this application, the response variable is *Percent body fat* (estimated by Brozek’s equation: $457/\text{Density} - 414.2$). In the original data set, we exclude the observations with the percentage body fat estimated as 0 and the density less than 1. Thus the data set used in this application involves 250 observations. Before application, we first standardize all the covariates.

Under the restriction $\|\beta\| = 1$ with $\beta_d > 0$, we use the model (1) with the single index function approximation (2) and our proposed cubic B spline variable selection procedure to select important variables. To compare the performance of our proposed procedures, we also apply the cubic B spline-based estimation method to the analysis of this data set. We summarize the variable selection results and estimates in Table 3, in which “Lasso”, “Scad”, “Hard” and “BE” respectively stand for the estimate results of variable selection procedures with Lasso, SCAD, Hard thresholding and B spline estimation approaches. The plots of estimated single index function $\hat{g}(\cdot)$ are displayed in Fig. 1.

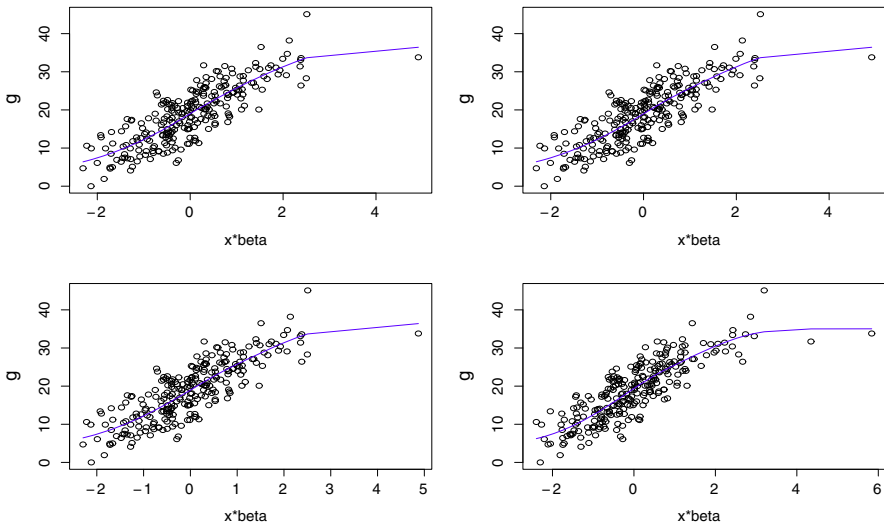


Fig. 1 The plots of estimated single index function $\hat{g}(\cdot)$

Table 5 Summarized estimation results for body fat data

Covaraites	Lasso	Scad	Hard	BE
Age	0.4965	0.4706	0.4695	0.4389
Weight	–	–	–	0.1990
Height	–	–	–	–0.1964
Neck	–	–	–	–0.0663
Chest	–	–	–	0.0534
Abdomen	0.7308	0.7843	0.7806	0.6765
Hip	–	–	–	0.1660
Thigh	0.3241	0.3172	0.3194	0.2662
Knee	–	–	–	–0.0647
Ankle	–	–	–	0.0427
Biceps	–	–	–	0.0562
Forearm	–	–	–	0.0507
Wrist	–0.2564	–0.2507	–0.2611	–0.3925

From the column labeled by “BE” in Table 3, we can see that the predictors *Age*, *Abdomen*, *Thigh* and *Wrist* contains most of information in interpreting the percent body fat while the other factors involves very little information. That is, the fat in a body mainly focuses on Thigh, Abdomen, Forearm and wrist. Moreover, the age also has something important with his percent body fat. From the first three columns in Table 5, all the three variable selection procedures can identify the same important factors. Therefore, we should take age and circumferences of abdomen, forearm, wrist, thigh as the important factors to measure the percent body fat. In addition, Fig. 1 displays the original data points and the B spline fitted lines. From upper-left to below-right, they are from variable selection based on Lasso, SCAD, HARD and

the B spline estimation. This figure significantly indicates that $X'\beta$ as a measurement has a nonlinear effect on the percent body fat, which flexibly shows the relationship between the percent body fat and the four main factors.

4 Conclusion

In this paper, we considered the variable selection for the model (1) with the single index model approximation (4) by incorporating B spline expansion technique. Under some regular conditions, we established the corresponding consistency and oracle properties of resulting penalized estimates. Some numerical studies illustrated that our proposed method performs very well for moderate sample size.

Our experiments shows that our proposed procedure performs very well when the dimension of covariates is less than sample size. For the dimensionality larger than n , some dimension reduction maybe needed such as SIS or ISIS methods. In many applications, some covariates can not be observed exactly and are prone to suffer from measurement error. It is also interesting to consider B spline variable selection with covariate measurement error. All the aspects will be investigated in our sequent research.

Acknowledgements We are grateful to the editor, associate editor, and referees for their helpful comments which led to the revised version of this paper. This work is partially supported by National Natural Science Foundation of China (11201190,11571148, 11271195,11171112), Postdoctoral Science Foundation of China (2014M550432), Humanities and Social Fund of Ministry of Education in China (12YJC910004), the Postdoctoral Initial Foundation in Guangzhou (gzhubsh2013004), Specialized Research Fund for the Doctoral Program of Higher Education(20124410110002), A Project Funded by the Priority Academic Program Development (PAPD) of Jiangsu Higher Education Institutions and “Qinglan” Project in Jiangsu.

Appendix

In this section, we prove Theorems 1, 2 under the Assumptions (A1)–(A6) in Wang and Yang (2009).

Proof of Theorem 1 Let $\alpha_n = n^{-1/2} + a_n$. It is sufficient to show that for any given $\varepsilon \in (0, 1)$, there exists a large constant C such that

$$P \left\{ \inf_{\|\mathbf{u}\|=C} Q(\beta_0^{(1)} + \alpha_n \mathbf{u}) \geq Q(\beta_0^{(1)}) \right\} \geq 1 - \varepsilon. \tag{12}$$

Based on that $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\theta) > 0$, we have

$$\begin{aligned} & Q(\beta_0^{(1)} + \alpha_n \mathbf{u}) - Q(\beta_0^{(1)}) \\ & \geq \left[R^*(\beta_0^{(1)} + \alpha_n \mathbf{u}) - R^*(\beta_0^{(1)}) \right] + \sum_{j=1}^s [p_{\lambda_n}(|\beta_{j0} + \alpha_n u_j|) - p_{\lambda_n}(|\beta_{j0}|)]. \end{aligned} \tag{13}$$

By Theorems 1, 2 in Wang and Yang (2009), for any $\beta^{(1)} \in \{\beta^{(1)} : \beta^{(1)} = \beta_0^{(1)} + \alpha_n \mathbf{u}, \|\mathbf{u}\| = C\}$, we have

$$R^*(\beta^{(1)}) - R^*(\beta_0^{(1)})$$

$$\begin{aligned}
 &= S^*(\beta_0^{(1)})(\beta^{(1)} - \beta_0^{(1)}) + \frac{1}{2}(\beta^{(1)} - \beta_0^{(1)})^T H^*(\beta_0^{(1)})(\beta^{(1)} - \beta_0^{(1)})\{1 + O_P(1)\} \\
 &= \frac{1}{2}(\beta^{(1)} - \beta_0^{(1)})^T [H^*(\beta_0^{(1)}) + O_P(1)](\beta^{(1)} - \beta_0^{(1)}) + O_P(n^{-1/2}) \cdot \|\beta^{(1)} - \beta_0^{(1)}\| \\
 &= \frac{1}{2}\alpha_n^2 \mathbf{u}^T [H^*(\beta_0^{(1)}) + O_P(1)]\mathbf{u} + O_P(n^{-1/2}\alpha_n\|\mathbf{u}\|). \tag{14}
 \end{aligned}$$

Note that $H^*(\beta_0^{(1)})$ is a positive definite matrix. The order for the first term in the last equality of (14) is $C^2\alpha_n^2$ and for second one is $\alpha_n^2 C$. Therefore, for a sufficiently large C , the second term is dominated by the first term in the last equation of (14). On the other hand, by Taylor’s expansion, the second term of (13) is bounded by

$$\sqrt{s}\alpha_n a_n \|\mathbf{u}\| + \alpha_n^2 b_n \|\mathbf{u}\|^2 = C\alpha_n^2(\sqrt{s} + b_n C).$$

If $b_n \rightarrow 0$, the second term of (13) is dominated by the first term of (14). Thus, for a sufficiently large C , (12) holds, which means that there exists a local minimizer in the ball $\{\beta^{(1)} : \beta^{(1)} = \beta_0^{(1)} + \alpha_n \mathbf{u}, \|\mathbf{u}\| \leq C\}$ with probability at least $1 - \varepsilon > 0$. Therefore, there exists a local minimizer $\hat{\beta}_n^{(1)}$ such that $\|\hat{\beta}_n^{(1)} - \beta_0^{(1)}\| = O_P(n^{-1/2} + a_n)$. □

Proof of Theorem 2 (i) It is sufficient to prove that

$$Q((\beta_1^{(1)'}, \boldsymbol{\theta}')') = \min_{\|\beta_2^{(1)}\| \leq Cn^{-1/2}} Q((\beta_1^{(1)'}, \beta_2^{(1)'})') \tag{15}$$

for any given $\beta_1^{(1)}$ satisfying $\|\beta_1^{(1)} - \beta_{10}^{(1)}\| = O_P(n^{-1/2})$ and any constant C .

Denote $S_j^*(\beta^{(1)})$ as the j th element of $S^*(\beta^{(1)})$. By the Taylor expansion of $S_j^*(\beta^{(1)})$ for $\|\beta^{(1)} - \beta_0^{(1)}\| = O_P(n^{-1/2})$ at $\beta_0^{(1)}$, we have

$$S_j^*(\beta^{(1)}) = S_j^*(\beta_0^{(1)}) + \sum_{i=1}^{p-1} \frac{\partial^2 R^*(\beta_0^{(1)})}{\partial \beta_j \partial \beta_i} (\beta_i - \beta_{i0}) + O_P(\|\beta^{(1)} - \beta_0^{(1)}\|^2). \tag{16}$$

From (A.32) and Theorem 2 of Wang and Yang (2009), it can be obtained that

$$\frac{\partial^2 R^*(\beta_0^{(1)})}{\partial \beta_j \partial \beta_i} = l_{ji} + o(1) \quad \text{and} \quad S_j^*(\beta_0^{(1)}) = O_P(n^{-1/2}),$$

where l_{ji} ’s are defined in the Theorem 2 of Wang and Yang (2009). So for $\|\beta^{(1)} - \beta_0^{(1)}\| = O_P(n^{-1/2})$, from (16) we have

$$S_j^*(\beta^{(1)}) = O_P(n^{-1/2}).$$

Therefore, for $\|\beta^{(1)} - \beta_0^{(1)}\| = O_P(n^{-1/2})$ and $j = s + 1, s + 2, \dots, p - 1$, we have that

$$\begin{aligned} \frac{\partial Q(\beta^{(1)})}{\partial \beta_j} &= \frac{1}{n} \left\{ nS_j^*(\beta^{(1)}) + n\dot{p}_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) \right\} \\ &= \frac{1}{n} \left\{ n\dot{p}_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) + O_P(\sqrt{n}) \right\} \\ &= \frac{1}{n} \left\{ n\lambda_n \left[\lambda_n^{-1} \dot{p}_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j) + O_P\left(\frac{1}{\sqrt{n\lambda_n}}\right) \right] \right\}, \end{aligned} \tag{17}$$

Since $\liminf_{n \rightarrow \infty} \liminf_{\theta \rightarrow 0^+} \dot{p}_{\lambda_n}(\theta)/\lambda_n = c > 0$, $\frac{1}{\sqrt{n\lambda_n}} \rightarrow 0$ and $|\text{sign}(\beta_j)| = 1$ for any $\beta_j \neq 0$,

$$\liminf_{n \rightarrow \infty} \liminf_{\beta_j \rightarrow 0} |\lambda_n^{-1} \dot{p}_{\lambda_n}(|\beta_j|)\text{sign}(\beta_j)| = c > 0$$

and so the second term in squared bracket of the last equation in (17) is dominated by the first term when n is large enough. Hence the the derivative and β_j have the same sign. Therefore (15) holds.

(ii) From $a_n = O(n^{-1/2})$ and Theorem 1, there exists a local \sqrt{n} -consistent minimizer, $\hat{\beta}_{1n}^{(1)}$, of $Q((\beta_1^{(1)'}, \mathbf{0})')$ satisfying

$$\frac{\partial Q(\beta^{(1)})}{\partial \beta_j^{(1)}} \Big|_{\beta^{(1)} = (\hat{\beta}_{1n}^{(1)}, \mathbf{0})'} = 0 \quad \text{for } j = 1, 2, \dots, s. \tag{18}$$

Set $\hat{\beta}_n^{(1)} = (\hat{\beta}_{1n}^{(1)'}, \mathbf{0})'$ and $S_1^*(\beta^{(1)})$ as the vector consisting of the first s components of $S^*(\beta^{(1)})$, then

$$\begin{aligned} 0 &= \frac{\partial Q(\beta^{(1)})}{\partial \beta_1^{(1)}} \Big|_{\beta^{(1)} = \hat{\beta}_n^{(1)}} = \frac{\partial Q(\beta^{(1)})}{\partial \beta_1^{(1)}} \Big|_{\beta^{(1)} = \beta_0^{(1)}} + \frac{\partial^2 Q(\beta^{(1)})}{\partial \beta_1^{(1)} \partial \beta_1^{(1)'}} \Big|_{\beta^{(1)} = \beta^{(1)*}} (\hat{\beta}_{1n}^{(1)} - \beta_{10}^{(1)}) \\ &= S_1^*(\beta_0^{(1)}) + \mathbf{b}_{\lambda_n} + \frac{\partial R^*(\beta^{(1)})}{\partial \beta_1^{(1)} \partial \beta_1^{(1)'}} \Big|_{\beta^{(1)} = \beta^{(1)*}} (\hat{\beta}_{1n}^{(1)} - \beta_{10}^{(1)}) + \Sigma_{\lambda_n}(\beta_1^{(1)*}) (\hat{\beta}_{1n}^{(1)} - \beta_{10}^{(1)}) \end{aligned} \tag{19}$$

where $\beta^{(1)*} = (\beta_1^{(1)*'}, \beta_2^{(1)*'})'$ lies on the line segment between $\hat{\beta}_n^{(1)}$ and $\beta_0^{(1)}$. From Theorem 1 above, Theorems 1, 2 in Wang and Yang (2009), (9) holds. This completes the proof. \square

References

Antoniadis A (1997) Wavelets in Statistics: A review (with discussion). *J Ital Stat Soc* 6:97–144
 Antoniadis A, Fryzlewicz P, Frédérique L (2010) The Dantzig selector in Coxs proportional hazards model. *Scand J Stat* 37(4):531–552

- Carroll R, Fan J, Gijbels I, Wand M (1997) Generalized partially linear single-index models. *J Am Stat Assoc* 92:477–489
- Ciuperca G (2014) Model selection by LASSO methods in a change-point model. *Stat Pap* 55:349–374
- Candes E, Tao T (2007) The Dantzig selector: statistical estimation when p is much larger than n . *Ann Stat* 35(6):2313–2351
- deBoor C (1978) *A practical guide to splines*. Springer, New York
- Fan J (1997) Comment on “Wavelet in statistics: a review” by A. Antoniadis. *J Ital Stat Soc* 6(2):131–138
- Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 96(456):1348–1360
- Fan J, Li R (2002) Variable selection for Cox’s proportional hazards model and frailty model. *Ann Stat* 30(1):74–99
- Fan J, Lv J (2010) A selective overview of variable selection in high dimensional feature space. *Stat Sin* 20(1):101–148
- Härdle W, Hall P, Ichimura H (1993) Optimal smoothing in single-index models. *Ann Stat* 21:157–178
- Hall P (1989) On projection pursuit regression. *Ann Stat* 17:573–588
- Horowitz J, Härdle W (1996) Direct semiparametric estimation of single-index models with discrete covariates. *J Am Stat Assoc* 91:1632–1640
- Hristache M, Juditsky A, Spokoiny V (2001) Direct estimation of the index coefficient in a single-index model. *Ann Stat* 29:595–623
- Ichimura H (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J Ecomom* 58:71–120
- Klein R, Spady R (1993) An efficient semiparametric estimator for binary response models. *Econometrica* 61:387–421
- Knight K, Fu W (2000) Asymptotics for lasso-type estimators. *Ann Stat* 28(5):1356–1378
- Kong E, Xia Y (2007) Variable selection for the single-index model. *Biometrika* 94:217–229
- Li K (1991) Sliced inverse regression for dimension reduction. *J Am Stat Assoc* 86:316–342
- Lu W, Zhang H (2007) Variable selection for proportional odds model. *Stat Med* 26(20):3771–3781
- Neykov N, Filzmoser P, Neytchev P (2014) Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. *Stat Pap* 55:187–208
- Peng H, Huang T (2011) Penalized least squares for single index models. *J Stat Plan Inference* 141:1362–1379
- Penrose K, Nelson A, Fisher A (1985) Generalized body composition prediction equation for men using simple measurement techniques. *Med Sci Sports Exerc* 17:189
- Powell J, Stock J, Stoker T (1989) Semiparametric estimation of index coefficients. *Econometrica* 57:1403–1430
- Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B (Methodological)* 58(1):267–288
- Tibshirani R (1997) The lasso method for variable selection in the Cox model. *Stat Med* 16(4):385–395
- Wang L, Yang L (2009) Spline estimation of single index models. *Stat Sin* 19:765–783
- Wang H (2009) Bayesian estimation and variable selection for single index models. *Comput Stat Data Anal* 53:2617–2627
- Wang H, Li R, Tsai C (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94(3):553–568
- Xia Y, Tong H, Li WK, Zhu L (2002) An adaptive estimation of dimension reduction space (with discussion). *J R Stat Soc Ser B* 64:363–410
- Xia Y, Li WK, Tong H, Zhang D (2004) A goodness-of-fit test for single-index models. *Stat Sin* 14:1–39
- Xia Y, Li W (1999) On single-index coefficient regression models. *J Am Stat Assoc* 94:1275–1285
- Xu D, Zhang Z, Wu L (2014) Variable selection in high-dimensional double generalized linear models. *Stat Pap* 55:327–347
- Zeng P, He T, Zhu Y (2012) A lasso-type approach for estimation and variable selection in single index models. *J Comput Graph Stat* 21:92–109
- Zhang H, Lu W (2007) Adaptive lasso for cox’s proportional hazards model. *Biometrika* 94(3):691–703
- Zhang H, Lu W, Wang H (2010) On sparse estimation for semiparametric linear transformation models. *J Multivar Anal* 101(7):1594–1606
- Zhu L, Qian L, Lin J (2011) Variable selection in a class of single-index models. *Ann Inst Stat Math* 63:1277–1293
- Zou H (2006) The adaptive lasso and its oracle properties. *J Am Stat Assoc* 101(476):1418–1429