Journal of Statistical Planning and Inference 141 (2011) 1362–1379

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Penalized least squares for single index models

Heng Peng [a,*,1], Tao Huang [b,2]

[a] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong
[b] Department of Statistics, University of Virginia, Charlottesville, VA 22904, United States

ARTICLE INFO

Article history:
Received 8 December 2009
Received in revised form
20 September 2010
Accepted 12 October 2010
Available online 21 October 2010

Keywords:
Local polynomial regression
Nonconcave penalized least squares
SCAD penalty
Single index model
Variable selection

ABSTRACT

The single index model is a useful regression model. In this paper, we propose a nonconcave penalized least squares method to estimate both the parameters and the link function of the single index model. Compared to other variable selection and estimation methods, the proposed method can estimate parameters and select variables simultaneously. When the dimension of parameters in the single index model is a fixed constant, under some regularity conditions, we demonstrate that the proposed estimators for parameters have the so-called oracle property, and furthermore we establish the asymptotic normality and develop a sandwich formula to estimate the standard deviations of the proposed estimators. Simulation studies and a real data analysis are presented to illustrate the proposed methods.

1. Introduction

In practice, one often uses a linear regression model $E(y|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$ to study the relationship between a response variable y and a vector of covariates $\mathbf{x}$. However, the linear relationship between the response variable and covariates is too restricted and limits the application of the linear regression model. To make the model more flexible, a single index model $E(y|\mathbf{x}) = g(\mathbf{x}^T\boldsymbol{\beta})$ can be used where the link function $g(\cdot)$ is a smooth unknown function. The single index model not only can mitigate the risk of misspecifying the link function, but also can overcome the curse of dimensionality. The single index model is a classical semiparametric model. The link function $g(\cdot)$ or the nonparametric part of the model and parameters $\boldsymbol{\beta}$ or the parametric part of the model need to be estimated simultaneously. The single index model has been widely studied. See, for example, Powell et al. (1989), Duan and Li (1991), Härdle et al. (1993), Ichimura (1993), Horowitz and Härdle (1996), Carroll et al. (1997), and the references therein.

There are two estimation problems for the single index model. One is the estimation of parameters, and the other is the estimation of the link function. Since the optimal convergence rate of the parametric estimator is faster than the optimal convergence rate of the nonparametric estimator, intuitively if the parameters $\boldsymbol{\beta}$ can be estimated accurately and efficiently, one can plug the estimate into the single index model, and subsequently obtain a good estimate for the link function. Hence the estimation of the parameters $\boldsymbol{\beta}$ seems more important than the estimation of the link function $g(\cdot)$. However, if the variable set includes some irrelevant covariates or includes hundreds of covariates, the accuracy of the estimation of parameters will deteriorate by the curse of dimensionality. Hence similar to the linear regression model, excluding irrelevant

* Corresponding author.
E-mail address: hpeng@math.hkbu.edu.hk (H. Peng).

[1] Heng Peng's research is supported by the CERG Grant of Hong Research Grant Council, HKBU201707, HKBU201809 and HKBU201610, the FRG Grant of Hong Kong Baptist University, FRG/08-09/II-33 and National Nature Science Foundation of China, NNSF 10871054.
[2] Tao Huang's research is partially supported by NSF Grant 0906661.

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.10.003

variables and selecting the most important variables from a set of covariates is an important issue for the single index model. Furthermore, it is important to establish the large sample properties of the estimators after the variable selection. Traditional variable selection methods, such as AIC, BIC, C_p or cross validation methods (Kong and Xia, 2007; Naik and Tsai, 2001), have to estimate parameters and calculate a criterion value for every candidate model first, and then select the best model according to the criterion values. The estimation and the selection are separated into two steps. The computation burden for selecting the best model using these methods can be unbearable even the dimension of the parameters $\boldsymbol{\beta}$ is moderate high due to their combinatorial nature. In most situations, these variable selection methods have to adopt stepwise deletion, subset selection or other procedures to alleviate the computation load. Moreover, these procedures ignore stochastic errors inherited in the process of variable selection, and cannot estimate and select variable simultaneously. Therefore it can be difficult to derive the large sample properties for these stepwise selection procedures. Fan and Li (2001) studied the penalized least squares method for parametric models and showed that the nonconcave penalized least squares method can estimate parameters and select variables simultaneously and has the oracle property, namely, the nonsignificant parameters can be automatically set to zero with probability tending to one and the significant parameters can be estimated efficiently as if the true model were known.

There are direct and indirect methods to estimate parameters for the single index model. Direct methods include the slice inverse regression (SIR) (Duan and Li, 1991) and the average derivative estimation (ADE) (Stoker, 1986; Powell et al., 1989; Härdle and Tsybakov, 1993; Horowitz and Härdle, 1996; Hristache et al., 2001). Once an accurate estimate of parameters is obtained, one can plug it into the model and use univariate nonparametric regression methods to estimate the link function. However, the single index model is more complicated than the linear regression model, and these direct estimation methods have their drawbacks. SIR requires the distribution of covariates to be elliptically symmetric, and is sensitive to the link function since it cannot diagnose symmetric (Cook and Weisberg, 1991). Zhu and Zhu (2009) proposed a method based on SIR to select variables for high-dimensional generalized single index model, however, their variable selection method lacks efficiency. ADE requires high-dimensional nonparametric regression and suffers the curse of dimensionality. Indirect methods, such as Carroll et al. (1997), use an iterative algorithm to estimate parameters and the link function simultaneously, and the resulting estimators have nice theoretical properties though the computation may lead to an optimization problem in a high-dimensional space.

Nonconcave penalized least squares and nonconcave penalized likelihood methods were proposed by Fan and Li (2001), and subsequently were applied to Cox's proportional hazards models and frailty models (Fan and Li, 2002). Fan and Peng (2004) studied the nonconcave penalized likelihood estimation and testing problems when the number of parameters is diverging with the sample size. Li and Liang (2008) investigated the nonconcave penalized method for generalized semiparametric partial linear models. Hunter and Li (2005) studied the convergence of the nonconcave penalized least squares method. Zou and Li (2008) proposed a fast algorithm for the nonconcave penalized likelihood method. Fan and Li (2001) proposed to use the generalized cross validation method (GCV) to select the tuning parameter for the nonconcave penalized likelihood method. Wang et al. (2007) showed that GCV may select an inconsistent tuning parameter, and proposed a BIC type of criterion for selecting the tuning parameter which can consistently identify the true model.

In this paper, when p, the dimension of $\boldsymbol{\beta}$, is a fixed constant, we extend the idea of nonconcave penalized least squares to the single index model, and propose an iterative estimation algorithm for the single index model. The basic idea of our proposed method is to replace the ordinary least squares estimate by the nonconcave penalized least square estimate, and iteratively estimate the parameters and the link function. We show that the proposed method can estimate parameters and select variables simultaneously and has nice theoretical properties. In practice, the iterative algorithm can be computationally intensive because an optimal tuning parameter is needed for each iteration. We propose a simple plug-in tuning parameter to alleviate the computation load. In fact, with the proposed plug-in tuning parameter, the computation load of the proposed nonconcave penalized least squares method is in the same magnitude as the computation load of the ordinary least squares method.

The rest of the paper is organized as follows. In Section 2 we introduce the proposed nonconcave penalized least squares method and the iterative estimation algorithm for the single index model. In addition, a simple plug-in method for selecting the tuning parameter is described. In Section 3 we derive the large sample and oracle properties of the proposed estimators. A sandwich formula for estimating the standard deviations of the proposed estimators is also described in this section. In Section 4, simulation studies and a real data analysis are presented to illustrate the proposed methods. Some discussions are given in Section 5. Technical proofs are relegated to the Appendix.

2. Nonconcave penalized least squares for single index models

2.1. Nonconcave penalized least squares method

Fan and Li (2001) studied the penalized least squares problem

$$\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+n\sum_{j=1}^{p}p_\lambda(|\beta_j|), \tag{2.1}$$

where $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T$, $\mathbf{X}$ is a $n\times p$ matrix and $\mathbf{y}$ is a $n\times 1$ vector. They showed that a good penalty function should result in an estimator with three properties:

(1) Unbiasedness: to avoid unnecessary bias, the penalty function should not over-penalize large parameters;
(2) Sparsity: to reduce model complexity, the penalty function should act as a thresholding rule and set nonsignificant parameters to zero automatically;
(3) Continuity: to avoid instability in model prediction, the penalty function should be chosen such that the resulting penalized least squares estimator is continuous in data.

By the discussion of Fan and Li (2001), the penalty functions satisfying the sparsity and the continuity must be singular at the origin. The condition $p'_\lambda(|\beta|)=0$ for large $|\beta|$ is a sufficient condition for the unbiasedness. In particular, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty function which is defined as follows:

$$p'_\lambda(\theta)=\lambda\left\{I(\theta\le\lambda)+\frac{(a\lambda-\theta)_+}{(a-1)\lambda}I(\theta>\lambda)\right\}, \tag{2.2}$$

for some $a>2$ and $\theta>0$. Fan and Li showed that the Bayes risk is not sensitive to the choice of a, and $a=3.7$ is a good choice for various problems. By Fan (1997), the simple penalized least squares problem

$$(z-\theta)^2/2+p_\lambda(|\theta|) \tag{2.3}$$

with SCAD penalty yields the solution

$$\hat{\theta}=\begin{cases}\operatorname{sgn}(z)(|z|-\lambda)_+, & \text{when } |z|\le 2\lambda,\\ \{(a-1)z-\operatorname{sgn}(z)a\lambda\}/(a-2), & \text{when } 2\lambda<|z|\le a\lambda,\\ z, & \text{when } |z|>a\lambda.\end{cases} \tag{2.4}$$

To minimize (2.1), Fan and Li (2001) proposed a modified Newton–Raphson algorithm, and showed that SCAD penalty function can be locally approximated by a quadratic function as follows:

$$p_{\lambda_j}(|\beta_j|)\approx p_\lambda(|\beta_{j0}|)+\tfrac{1}{2}\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\}(\beta_j^2-\beta_{j0}^2),$$

where β_j is in the neighborhood of β_{j0}, and $\beta_j\neq 0$. Therefore the minimizer for (2.1) can be obtained by iteratively solving

$$\boldsymbol{\beta}_1=\{\mathbf{X}^T\mathbf{X}+n\Sigma_\lambda(\boldsymbol{\beta}_0)\}^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.5}$$

where $\Sigma_\lambda(\boldsymbol{\beta}_0)=\operatorname{diag}\{p_\lambda'(|\beta_{10}|)/|\beta_{10}|,\ldots,p_\lambda'(|\beta_{d0}|)/|\beta_{d0}|\}$, and d is the dimension of $\boldsymbol{\beta}_0$.

2.2. Estimation method for single index models

Consider the single index model

$$y_i=g(\mathbf{X}_i^T\boldsymbol{\beta})+\varepsilon_i,\quad i=1,2,\ldots,n,$$

where $g(\cdot)$ is a smooth unknown link function, $\mathbf{X}_i$ is a vector of covariates of length p, and ε_i is a white noise with unknown variance σ^2. For identifiability we assume that $\|\boldsymbol{\beta}\|=1$ and $\operatorname{sign}(\beta_1)=1$. We follow the idea of Carroll et al. (1997) and use an iterative algorithm to estimate parameters $\boldsymbol{\beta}$ and the link function $g(\cdot)$ simultaneously.

Given an estimate of $\boldsymbol{\beta}$, the link function $g(\cdot)$ can be locally approximated by a linear function

$$g(v)\approx g(u)+g'(u)(v-u)\equiv a+b(v-u),$$

for v in the neighborhood of u, where $a=g(u)$ and $b=g'(u)$ are local constants. In practice, local linear approximation/estimation is done at evaluation time. With a symmetric kernel function $K_h(t)=K(t/h)/h$ where h is the bandwidth, one can estimate a and b by local linear regression

$$\sum_{i=1}^n[y_i-a-b(\mathbf{X}_i^T\boldsymbol{\beta}-u)]^2K_h(\mathbf{X}_i^T\boldsymbol{\beta}-u). \tag{2.6}$$

Given an estimate of the link function $g(\cdot)$, say $\hat{g}(\cdot)$, one can update the estimate of $\boldsymbol{\beta}$ by minimizing the following nonconcave penalized least squares function

$$\sum_{i=1}^n[y_i-\hat{g}(\mathbf{X}_i^T\boldsymbol{\beta})]^2+n\sum_{j=1}^p p_\lambda(|\beta_j|). \tag{2.7}$$

However, since $\hat{g}(\cdot)$ may not be a linear function, solving (2.7) is a nonlinear optimization problem, and the computation can be challenging. Instead we propose to use the local approximation idea and update the estimate of $\boldsymbol{\beta}$ by minimizing the

following nonconcave penalized least squares function given the current estimates $\boldsymbol{\beta}_0$ and $\hat{g}(\cdot)$,

$$\sum_{i=1}^{n}[y_i-\hat{g}(\mathbf{X}_i^T\boldsymbol{\beta}_0)-\hat{g}'(\mathbf{X}_i^T\boldsymbol{\beta}_0)(\mathbf{X}_i^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}_0)]^2+n\sum_{j=1}^{p}p_\lambda(|\beta_j|). \tag{2.8}$$

Minimizing (2.8) is a quadratic optimization problem. Similar to the problem of minimizing (2.1), we can use (2.5) to iteratively update the estimate of $\boldsymbol{\beta}$ until convergence.

Our estimation procedure for $\boldsymbol{\beta}$ and $g(\cdot)$ is described in details as follows:

Step 0: Initialization step. Obtain an initial estimate of $\boldsymbol{\beta}$. For example, let $\hat{\boldsymbol{\beta}}_1$ be the least squares estimate for

$$y_i=\mathbf{X}_i^T\boldsymbol{\beta}+\varepsilon_i,\quad i=1,2,\cdots n,$$

and set $\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}_1/\|\hat{\boldsymbol{\beta}}_1\|$, and $\text{sign}(\hat{\beta}_{11})=1$.

Step 1: Find $\hat{g}(u,\hat{\boldsymbol{\beta}})=\hat{a}$ and $\hat{g}'(u,\hat{\boldsymbol{\beta}})=\hat{b}$ by minimizing (2.6).

Step 2: Let $\boldsymbol{\beta}_0=\hat{\boldsymbol{\beta}}$, update the estimate of $\boldsymbol{\beta}$ by minimizing (2.8).

Step 3: Continue Step 1 and Step 2 until convergence.

Step 4: Given the final estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ from Step 3, refine the estimate $\hat{g}(u,\hat{\boldsymbol{\beta}})$ of $g(\cdot)$ again by minimizing (2.6).

In fact, our proposed iterative algorithm can be thought as an EM algorithm. In Step 1, estimating the link function $g(\cdot)$ is a classical univariate nonparametric regression problem, and the estimator is equivalent to a condition expectation. Therefore Step 1 can be thought as the E-step of EM algorithm. Step 2 is to minimize an objective function, and can be regarded as the M-step of EM algorithm. The convergence of the algorithm for minimizing (2.8) has been shown by Hunter and Li (2005). Together our proposed iterative algorithm should converge intuitively because of the convergence of EM algorithm.

Same as the discussion of Carroll et al. (1997), our proposed estimation procedure needs to select an optimal smoothing parameter or the bandwidth h on two different levels. In Step 1 and Step 2, it is of primary interest to assure the accuracy of the estimation of $\boldsymbol{\beta}$, hence the bandwidth h should be optimal for estimating $\boldsymbol{\beta}$, and it means that the link function $g(\cdot)$ should be undersmoothed. In Step 4, it is of primary interest to obtain an accurate estimate of the link function, hence the bandwidth h should be optimal for estimating $g(\cdot)$ as if $\boldsymbol{\beta}$ were known. Selecting different smoothing parameters is standard for semiparametric models using the kernel regression. There are other techniques to estimate $\boldsymbol{\beta}$ and $g(\cdot)$ efficiently with a single smoothing parameter. For details, see Härdle et al. (1993).

In addition to the selection of optimal smoothing parameters h, our proposed estimation procedure also needs to select the tuning parameter λ. Fan and Li (2001) used GCV to select λ, and Wang et al. (2007) proposed a BIC type criterion to select λ. However, both methods need to search the optimal λ over a set of fine grids, and the computation can be intensive for the proposed iterative algorithm because an optimal tuning parameter is needed for each iteration. We propose a simple plug-in method for selecting the tuning parameter in the following section.

2.3. Selection of bandwidth and tuning parameters

When $\boldsymbol{\beta}$ is known, estimating $g(\cdot)$ is a classical univariate nonparametric regression problem. There are many bandwidth selection methods. However, since our proposed estimation procedure is iterative in nature, GCV type of bandwidth selection methods can be computationally intensive. Hence we use the plug-in global bandwidth $\hat{h}_{opt}$ proposed by Ruppert et al. (1995) to estimate $g(\cdot)$ in Step 4. On the other hand, in Steps 1 and 2, one needs to undersmooth (smaller bandwidth) the estimation of $g(\cdot)$ so that $\hat{\beta}$ has the optimal $\sqrt{n}$ convergence rate. As discussed by Carroll et al. (1997), a sensible rule for selecting h for Step 1 is much more difficult. By their suggestion, a relatively *ad hoc* possibility is

$$\hat{h}_{opt}\times n^{1/5}\times n^{-1/3}=\hat{h}_{opt}\times n^{-2/15},$$

since this bandwidth gives the correct order of magnitude for the asymptotic results shown in the next section.

The selection of tuning parameters λ and a is another practical issue for the proposed nonconcave penalized least squares method. Because of the iterative nature of the proposed estimation method, standard methods, such as CV and GCV, are computationally intensive and a simple plug-in method is desired. By the discussion of Fan and Li (2001), $a=3.7$ is a good choice for various problems. Empirical results also showed that $a=3.7$ is close to the optimal tuning parameter selected by CV or GCV. Furthermore, by the asymptotic results of Fan and Li (2001), when $\lambda=\sqrt{C_1\log n/n}\sigma$, the nonconcave penalized least squares method has the so-called oracle property, namely, the zero coefficients are set to zero automatically and the nonzero coefficients can be estimated efficiently as if the true model were known. Denote the true parameter $\boldsymbol{\beta}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$, where $\boldsymbol{\beta}_1\neq 0$ and $\boldsymbol{\beta}_2=0$. Since the nonzero coefficients are not supposed to be penalized and should be larger than $a\lambda$, we can plug the final estimate of $\boldsymbol{\beta}_1$ into the nonconcave penalized least squares function and rewrite it as the following:

$$\sum_{i=1}^{n}[y_i-g(\mathbf{X}_{1i}^T\hat{\boldsymbol{\beta}}_1)]^2+C_1C_2\log n\cdot d\cdot\sigma^2, \tag{2.9}$$

where d is the dimension of $\boldsymbol{\beta}_1$, and C_2 is a constant determined by the nonconcave penalty function $p_\lambda(\cdot)$. For SCAD penalty function, we know $p_\lambda(|\beta|)=(a+1)\lambda^2/2$ when $|\beta|>a\lambda$, so $C_2=(a+1)/2$ when all nonzero coefficients in the model are larger than $a\lambda$. Note that (2.9) is the same as BIC if $C_1C_2=1$. As shown by Bai et al. (1999), variable selection using BIC is asymptotical consistent. Hence BIC selection and the nonconcave penalized least squares method should have the same estimation in the

asymptotical sense under some general conditions, though the computation for BIC selection is much more intensive than the computation for the nonconcave penalized least squares method. By the above analogy between the nonconcave penalized least squares method and BIC selection, we propose

$$\lambda = \sqrt{\frac{2\log n}{n(a+1)}}\hat{\sigma}$$

as the plug-in tuning parameter for SCAD penalized least squares method, where $\hat{\sigma}$ is an estimate of the standard deviation of white noises. A simple estimate for σ is

$$\sqrt{\frac{1}{n-d}\sum_{i=1}^{n}[y_i - \hat{g}(\mathbf{X}_i^T\hat{\boldsymbol{\beta}})]^2},$$

where $\hat{g}(\cdot)$ and $\hat{\boldsymbol{\beta}}$ are the current estimates of $g(\cdot)$ and $\boldsymbol{\beta}$, and d is the dimension of $\hat{\boldsymbol{\beta}}$.

3. Asymptotical properties

We now study the large sample properties of the proposed estimators. First, the asymptotic property and the oracle property for the estimator of $\boldsymbol{\beta}$ are shown in the following section. Since the estimation of $\boldsymbol{\beta}$ is as efficient as if the true model were known, the sampling property of the link function $g(\cdot)$ can be derived similarly as that of Carroll et al. (1997). Furthermore, by the asymptotical distribution of the estimator of $\boldsymbol{\beta}$, we are able to construct a sandwich formula to estimate the covariance matrix of the estimator of $\boldsymbol{\beta}$.

3.1. Sampling property and oracle property

We first establish the asymptotic property of the nonconcave penalized least squares estimator. Let

$$Q(g,\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i - g(\mathbf{X}_i^T\boldsymbol{\beta})]^2 + n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|). \tag{3.1}$$

Denote $\boldsymbol{\beta}_0$ as the true value of $\boldsymbol{\beta}$ in the model with a fixed constant dimension p, and

$$\boldsymbol{\beta}_0 = (\beta_{10},\ldots,\beta_{p0})^T = (\boldsymbol{\beta}_{10}^T,\boldsymbol{\beta}_{20}^T)^T,$$

where $\boldsymbol{\beta}_{20} = 0$, $\boldsymbol{\beta}_{10} = (\beta_{10},\ldots,\beta_{s0})^T, \beta_{j0} \neq 0, j = 1,\ldots,s$, $\beta_{10} > 0$ and $\|\boldsymbol{\beta}_0\| = 1$.

We impose the following regularity conditions:

(A) the marginal density of $\mathbf{X}^T\boldsymbol{\beta}$ is positive and uniformly continuous in a neighborhood of $\boldsymbol{\beta}_0$. Furthermore, $\mathbf{X}^T\boldsymbol{\beta}$ has a positive density on its support D.
(B) the function $g''(\cdot)$ is continuous and bounded in D.
(C) $\mathbf{X}$ is bounded and its density function has a continuous second derivative, Cov($\mathbf{X}$) is nonsingular, and for any vector $\mathbf{v}$, if $\mathbf{v}^T\mathrm{Cov}(\mathbf{X})\boldsymbol{\beta}_0 = 0$, then $\mathbf{v}^T\Sigma\mathbf{v} > c\|\mathbf{v}\|^2$, where $\Sigma = \mathbf{E}g'^2(U)\{\mathbf{X}-\mathbf{E}(\mathbf{X}|U=\mathbf{X}^T\boldsymbol{\beta}_0)\}\{\mathbf{X}-\mathbf{E}(\mathbf{X}|U=\mathbf{X}^T\boldsymbol{\beta}_0)\}^T$.
(D) the kernel function K is a symmetric density function with bounded support and bounded first derivative.
(E) $\mathrm{E}(\varepsilon_i|\mathbf{X}_i) = 0$, $\mathrm{E}(\varepsilon_i^2|\mathbf{X}_i) = \sigma^2 > 0$, and $\mathrm{E}(\varepsilon_i^4|\mathbf{X}_i)$ exists.
(F) $nh^3 \to \infty$ and $nh^4 \to 0$.

It is obvious that regularity condition (A) and condition (B) are similar to the regularity conditions used by Carroll et al. (1997) and Härdle et al. (1993) for the single index model and the generalized single index model. Condition (C) is imposed to facilitate the technical arguments though it is somewhat complex. For instance, the condition that Cov($\mathbf{X}$) is nonsingular is a basic condition that Fan and Li (2001) used to study the sampling properties of the nonconcave penalized likelihood estimators. If $\mathbf{X}$ follows a multivariate normal distribution, then $\mathbf{X}^T\mathbf{v}$ and $\mathbf{X}^T\boldsymbol{\beta}_0$ are independent. Moreover it is easy to show that condition (C) is satisfied and can be simplified to Cov($\mathbf{X}$) being nonsingular. Condition (D) and condition (E) are standard regularity conditions for nonparametric regressions. Condition (F) is used for the bandwidth selection. In our theoretical derivation, the bandwidth h is selected to satisfy the condition, and assumed to be constant in all iterative steps. But to improve the efficiency of the estimation of the link function, whenever $\boldsymbol{\beta}$ is updated, the bandwidth should be also updated according to our proposed plug-in bandwidth selection method. Though there is a small gap between the bandwidths we use for the theoretical study and for practical implementation, such a small gap can be neglected because it is bounded by a constant according to our proposed plug-in bandwidth selection method.

By the following theorem, we show that a penalized least squares estimator converges at the rate

$$\mathrm{O_p}(n^{-1/2} + a_n), \tag{3.2}$$

where $a_n = \max\{p'_{\lambda_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\}$. This implies that for SCAD penalty function, the penalized least squares estimator is root-n consistent if $\lambda_n \to 0$ and all nonzero coefficients are larger than $a\lambda_n$.

Theorem 3.1. *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *be independent and identically distributed with a density* $f(\mathbf{x})$ *that satisfies conditions* (A)–(F). *If* $\max\{|p_{\lambda_n}''(|\beta_{j0}|) : \beta_{j0} \neq 0\} \to 0$ *and* $a_n = O(n^{-1/2})$, *then there exists a local minimizer* $\hat{\boldsymbol{\beta}}$ *of* $Q(g_\beta, \boldsymbol{\beta})$ *such that* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2} + a_n)$, *where* $\|\boldsymbol{\beta}\| = \|\boldsymbol{\beta}_0\| = 1$, *and* g_β *is the local linear estimate of the link function* $g(\cdot)$ *given* $\boldsymbol{\beta}$.

By Theorem 3.1, there exists a root-n consistent penalized least squares estimate for $\boldsymbol{\beta}$ if we can properly select the tuning parameter λ_n. Next we show that the estimator possesses the sparsity property, i.e. $\hat{\boldsymbol{\beta}}_2 = 0$, in the following lemma.

Lemma 3.1. *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *be independent and identically distributed with a density* $f(\mathbf{x})$ *that satisfies conditions* (A)–(F). *Assume that*

$$\liminf_{n\to\infty} \liminf_{\theta \to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0. \tag{3.3}$$

If $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$, *then with probability tending to* 1, *for any given* $\boldsymbol{\beta}_1$ *that satisfies* $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{10}\| = O_p(n^{-1/2})$ *and any constant* C,

$$Q\left\{g_{(\beta_1^T, 0)^T}, \begin{pmatrix} \boldsymbol{\beta}_1 \\ 0 \end{pmatrix}\right\} = \max_{\|\boldsymbol{\beta}_2\| \le Cn^{-1/2}} Q\left\{g_{(\beta_1^T, \beta_2^T)^T}, \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}\right\}.$$

Theorem 3.2. *Let* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *be independent and identically distributed with a density* $f(\mathbf{x})$ *that satisfies conditions* (A)–(F). *Assume that the penalty function* $p_{\lambda_n}(|\theta|)$ *satisfies condition* (3.3). *If* $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$, *then with probability tending to* 1, *the root-n consistent local maximizer* $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ *in Theorem* 3.1 *must satisfy*:

(a) *Sparsity*: $\hat{\boldsymbol{\beta}}_2 = 0$.
(b) *Asymptotic normality*:

$$\sqrt{n}(W_0 + \Sigma)\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (W_0 + \Sigma)^- \mathbf{b}\} \to \mathcal{N}(0, \sigma^2 W_0),$$

where

$$\Sigma = \mathrm{diag}\{p_{\lambda_n}''(|\beta_{10}|), \ldots, p_{\lambda_n}''(|\beta_{s0}|)\},$$

$$\mathbf{b} = (p_{\lambda_n}'(|\beta_{10}|)\mathrm{sgn}(\beta_{10}), \ldots, p_{\lambda_n}'(|\beta_{10}|)\mathrm{sgn}(\beta_{s0}))^T,$$

where s *is the number of components of* $\boldsymbol{\beta}_{10}$, *and*

$$W_0 = \mathrm{E}[\{\mathbf{X}_1 g'(\mathbf{X}_1\boldsymbol{\beta}_{10})\}\{\mathbf{X}_1 g'(\mathbf{X}_1\boldsymbol{\beta}_{10})\}^T] - \mathrm{E}\{\mathrm{E}[\{\mathbf{X}_1 g'(\mathbf{X}_1\boldsymbol{\beta}_{10})\}|\mathbf{X}_1\boldsymbol{\beta}_{10} = U]\mathrm{E}[\{\mathbf{X}_1 g'(\mathbf{X}_1\boldsymbol{\beta}_{10})\}|\mathbf{X}_1\boldsymbol{\beta}_{10} = U]^T\}.$$

Remark 1. By Lemma 3.1 and Theorem 3.2, it is easy to see that the nonconcave SCAD penalized least squares estimator has the so-called oracle property when $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$.

3.2. Sandwich formula

Similar to Fan and Li (2001) and Carroll et al. (1997), we can derive a sandwich formula to estimate the covariance matrix of the estimator of $\boldsymbol{\beta}$.

Let $\tilde{Q}$ be a $n \times p$ matrix with the ith row being $g'(U_i)\mathbf{X}_i^T$, where $U_i = \mathbf{X}_i^T\boldsymbol{\beta}$. Let $P_\beta^* = \mathbf{I} - \boldsymbol{\beta}\boldsymbol{\beta}^T$, $\tilde{\boldsymbol{\varepsilon}}$ be the vector with the ith element being $g(U_i) + g'(U_i)\mathbf{X}_i^T\boldsymbol{\beta} + \varepsilon_i$, $\tilde{\mathbf{g}} = \{g(U_1), \ldots, g(U_n)\}^T$, and $\Sigma_\lambda(\boldsymbol{\beta}) = \mathrm{diag}\{p_\lambda'(|\beta_1|)/|\beta_1|, \ldots, p_\lambda'(|\beta_d|)/|\beta_d|\}$.

Note that the constraint $\|\boldsymbol{\beta}\| = 1$ is necessary for identifiability. Estimate $\boldsymbol{\beta}$ by solving

$$0 = \tilde{Q}^T(\tilde{\boldsymbol{\varepsilon}} - \tilde{\mathbf{g}}) - (\tilde{Q}^T\tilde{Q} + n\Sigma_{\lambda_n})\boldsymbol{\beta} + \theta\boldsymbol{\beta},$$

where θ is a Lagrange multiplier. Multiplying both sides by P_β^*, then solving it with respect to $\boldsymbol{\beta}$ yields

$$\hat{\boldsymbol{\beta}} = (P_\beta^*(\tilde{Q}^T\tilde{Q} + n\Sigma_{\lambda_n}))^- P_\beta^*\tilde{Q}^T(\tilde{\boldsymbol{\varepsilon}} - \tilde{\mathbf{g}}).$$

Note that $\tilde{\mathbf{g}} = \mathbf{S}_\beta(\tilde{\boldsymbol{\varepsilon}} - \tilde{Q}\boldsymbol{\beta})$. It is easy to show that $\hat{\boldsymbol{\beta}} = \tilde{H}\tilde{\boldsymbol{\varepsilon}}$, where

$$\tilde{H} = \{P_\beta^*(\tilde{Q}^T(I - S_\beta)\tilde{Q} + n\Sigma_{\lambda_n})\}^- P_\beta^*\tilde{Q}^T(I - S_\beta).$$

Hence the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is $\hat{\sigma}^2\tilde{H}\tilde{H}^T$, where $\hat{\sigma}^2$ can be estimated by

$$\frac{1}{n}\sum_{i=1}^n [y_i - \hat{g}_\beta(\mathbf{X}_i\hat{\boldsymbol{\beta}})]^2.$$

4. Numerical study

In this section we illustrate the performance of the proposed nonconcave penalized least squares method and the proposed sandwich formula by two simulation studies and a real data analysis.

4.1. Application

We apply the proposed nonconcave penalized least squares method to a body fat dataset (Penrose et al., 1985). The dataset consists of 252 observations. The response variable is the percentage of body fat which is determined by the underwater weighting technique (Siri, 1956; Katch and McArdle, 1977). This underwater weighting technique can be inconvenient in practice since it requires to estimate the body density which in turn requires to measure the difference of body weight measured in air and during water submersion. Hence it is desirable to build a simple model to estimate the percentage of body fat by using only a few measurements. The predictors in this dataset include age, weight, height, and ten body circumference measurements (neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm and wrist). We exclude the samples that shows inconsistency estimate between the percentage of body fat and the body density and samples with the percentage of body fat less than one. After the exclusion, there are 246 observations for the analysis.

We consider all thirteen predictors to model the logarithm of the percentage of body fat. In particular, we consider three models, a linear model with variable selection via the nonconcave penalized least squares method (Fan and Li, 2001), a single index model without variable selection (Duan and Li, 1991; Carroll et al., 1997), and a single index model with variable selection via the proposed nonconcave penalized least squares method. For the nonconcave penalized least squares method, we use the SCAD penalty, and the constant a in the SCAD is taken as 3.7 and the unknown parameter λ is chosen by the proposed plug-in method. For the single index model, the constraints $\|\boldsymbol{\beta}\| = 1$ and $\mathrm{sgn}(\beta_1) = 1$ are needed for identifiability. The estimated coefficients are reported in Table 1.

From Table 1, the proposed nonconcave least squares method for the single index model chooses four out of 13 predictors, where the nonconcave least squares method for the linear model chooses six additional predictors. Among all the predictors, all three models shows that abdomen is the most important measurement for the prediction of the percentage of body fat. In addition, age, neck circumference and wrist circumference are selected by the proposed method. It might seem counter-intuitive that wrist circumference is a more important measurement than hip and thigh circumferences for predicting the percentage of body fat. One explanation might be that the percentage of body fat is highly related to the measurements at three parts of the body, the limps, upper and middle parts of the body, and the wrist, neck and abdomen circumferences are the best measurements for that three parts of the body with abdomen circumference being the most informative one. The estimated link function $g(\cdot)$ of the single index model is depicted in Fig. 1. It shows that the link function is a truly nonlinear function in which case a linear model will be inadequate to describe the relationship between the predictors and the response. In addition, it shows that the estimated link functions are essentially the same when one chooses only four predictors instead of all 13 predictors.

4.2. Simulation I

In this example we simulate 200 datasets consisting of 200 observations from the following model

$$y = g(\mathbf{X}^T\boldsymbol{\beta}) + \sigma\varepsilon,$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|$ with $\boldsymbol{\beta}_0 = (3,1.5,0,0,2,0,0,0)^T$ and $\sigma^2 = 0.1$. The components of $\mathbf{X}$ and ε are standard normal and the correlation between X_i and X_j is $\rho^{|i-j|}$ with $\rho = 0.5$, and the unknown link function is $g(\cdot) = \sin(\cdot)$.

Similar to the real data analysis, for the simulated data we consider three models, a linear model with variable selection via the nonconcave penalized least squares method, a single index model without variable selection, and a single index model with variable selection via the penalized least squares methods. For the proposed nonconcave penalized least squares

Table 1
The body fat data.

Method	SIM-SCAD	SIM[a]	SIM[b]	LM-SCAD
Age	0.0149	0.0863	0.0438	0.0489
Weight	0	0.0447	0.1000	0.1457
Height	0	−0.0612	0.0514	−0.0395
Neck	−0.1691	−0.1578	0.1974	−0.1408
Chest	0	−0.0574	0.2638	−0.0943
Abdomen	0.9606	0.9373	−0.6818	0.7663
Hip	0	−0.1741	−0.4339	−0.3638
Thigh	0	0.1096	0.3769	0.1461
Knee	0	−0.0115	−0.0114	0
Ankle	0	0.0076	−0.0069	0
Biceps	0	0.0466	−0.0056	0
Forearm	0	0.0510	−0.2601	0.0413
Wrist	−0.2202	−0.1809	0.1179	−0.1186

SIM-SCAD: single index model with SCAD penalty; LM-SCAD: linear model with SCAD penalty.
[a] Single index model estimated by the iterative method.
[b] Single index model estimated by SIR.

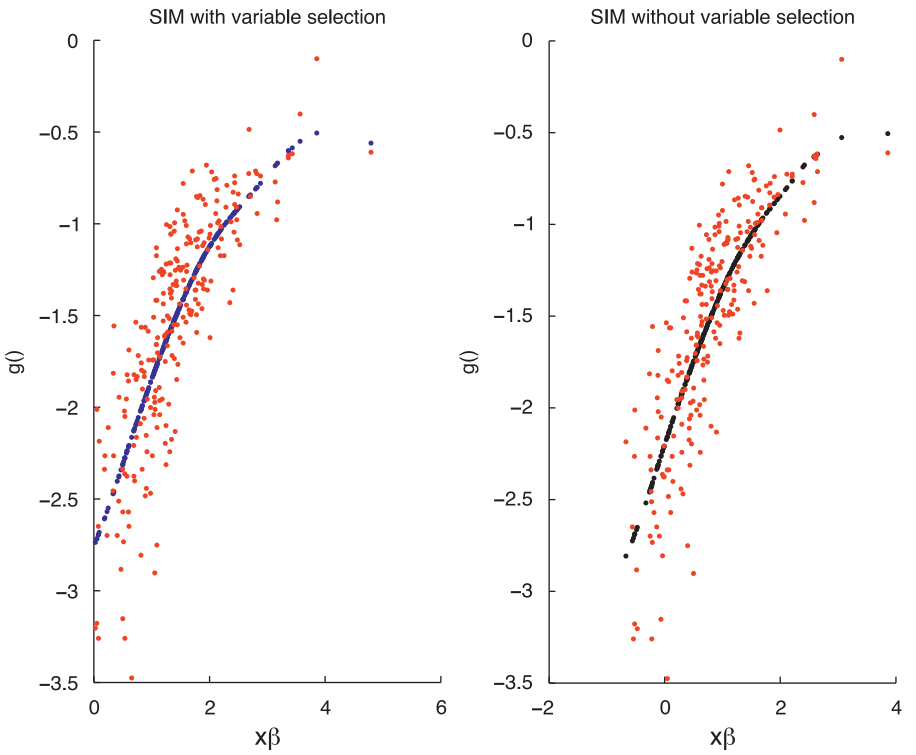

Fig. 1. Estimated link function $g(\cdot)$. Left: Single index model with variable selection; Right: Single index model without variable selection.

Table 2
Estimation result.

Method	β_1		β_2		β_5		Average no. of nonzero coefficient	
	Bias	SD	Bias	SD	Bias	SD	Correct	Incorrect
SIM-SCAD[1]	−0.0050	0.0304	0.0015	0.0463	0.0032	0.0254	3.0000	0.1700
SIM-SCAD[2]	−0.0059	0.0310	0.0015	0.0457	0.0029	0.0258	3.0000	0.4400
SIM-LASSO[1]	0.0144	0.0329	−0.0091	0.0541	−0.0207	0.0323	3.0000	0.5400
SIM-LASSO[2]	0.0095	0.0306	−0.0069	0.0467	−0.0118	0.0323	3.0000	0.7000
AICc	−0.0041	0.0300	−0.0012	0.0455	−0.0007	0.0319	3.0000	0.8600
LM-SCAD	−0.3838	0.0419	−0.1888	0.0378	−0.2537	0.0411	2.9950	1.3350
SIM[1]	−0.0134	0.0365	−0.0016	0.0507	0.0038	0.0393	*	*
SIM[2]	−0.2369	0.2271	−0.1346	0.3069	−0.1957	0.2920	*	*
Oracle-SIM	−0.0060	0.0268	−0.0021	0.0378	0.0041	0.0312	3.0000	0

method, the initial value can be estimated either by the iterative procedure proposed by Carroll et al. (1997) least squares method or by the SIR method. For this example, both methods give similar results and therefore we choose the estimate by the iterative procedure as the initial value for the following numerical study. However, based on our experience, the initial value sometimes can be an important factor in the process of estimation, specially when the function $g(\cdot)$ is quite complicated and/or the dimension of $\boldsymbol{\beta}$ is relatively high, because an inappropriate initial value might result in a local minimum instead of a global minimum. Under these situations, the estimate using the iterative procedure is preferred for our proposed least squares method comparing to that using SIR.

The estimated coefficients are reported in Table 2 in which "Bias" is the estimation bias, and "SD" is the median absolute deviation (MAD) of estimated coefficients divided by 0.6745. The average number of nonzero coefficients are also reported in Table 2, in which the column labeled "Correct" presents the average restricted only to the true nonzero coefficients, and the column labeled "Incorrect" depicts the average number of coefficients erroneously set to nonzero, "*" means "not available".

Throughout the paper, we use SCAD penalty function to illustrate our proposed penalized least square method. In practice, SCAD penalty function can be replaced by L_1 penalty function. The modified Newton–Raphson method can also be applied to minimize (2.7) and (2.8) with L_1 penalty function. We use SIM-LASSO to denote such methods, and then numerically compare the performance of different penalty functions for the proposed penalized least squares method. However, the theoretical

properties obtained in Section 3 for the proposed nonconcave penalized least squares method, namely for SIM-SCAD, cannot be directly extended to SIM-LASSO, and further theoretical investigations are needed for SIM-LASSO.

In Table 2, the tuning parameter λ is selected by the plug-in method for SIM-SCAD^1 and SIM-LASSO^1 and by the cross validation method for SIM-SCAD^2 and SIM-LASSO^2. AICc denotes the estimate by AICc criterion proposed by Naik and Tsai (2001). From Table 2, it can be seen that the proposed SIM-SCAD performs quite well. It not only estimates parameters with great accuracy compared to the estimate of the oracle single index model (Oracle-SIM) using the iterative algorithm (Carroll et al., 1997) but also selects the significant variables with high probability. Table 2 also shows that the proposed plug-in selector is very effective. Fig. 2 shows one estimation of the link function $g(\cdot)$. SIM-SCAD^{1,2} perform a bit better than SIM-LASSO^{1,2} in both estimation and selection, and specifically, SIM-SCAD^{1,2} have similar standard deviation but smaller bias than SIM-LASSO^{1,2}. Because SIM-LASSO^{1,2} are estimated by using the iterative algorithm, we conjecture that the bias of LASSO estimate would be accumulated to affect the final estimation and selection results. In comparison to SIM-SCAD and SIM-LASSO methods, AICc method yields similar estimation and no better selection results. But AICc is computationally intensive, and we have to estimate and compare 255 candidate single index models in order to select the best model for this simulation example.

We also test the accuracy of the proposed sandwich formula. The median absolute deviation (MAD) of estimated coefficients divided by 0.6745, denoted by SD in Table 3, can be regarded as the true standard error. The median of estimated SD's, denoted by SD_m, and their MAD divided by 0.6745, denoted by SD_{mad}, gauge the overall performance of the proposed sandwich formula. The results are presented in Table 3. Since our proposed sandwich formula does not count the variation resulted from the iteration process, the estimated standard deviations (SD_m) tend to be smaller than the true standard deviations (SD). However, when the sample size increases, the accuracy of the proposed sandwich formula increases since the variation resulted from the iteration process is reduced.

4.3. Simulation II

In this example, we explore the effect of high dimensionality on the proposed penalized nonconcave least squares method. Similar to Example 1, we simulate 200 datasets consisting of n observations from the following model

$$y = g(\mathbf{X}^T\boldsymbol{\beta}) + \sigma\varepsilon,$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_0/\|\boldsymbol{\beta}_0\|$ with $\boldsymbol{\beta}_0 = (3,1.5,0,0,2,0,\ldots,0)_{p\times 1}$, a p-dimensional parameter vector with only three nonzero significant parameters. The components of $\mathbf{X}$ and ε follow standard normal distribution and the correlation between X_i and X_j is $\rho^{|i-j|}$ with $\rho = 0.5$, $\sigma^2 = 0.1$, and the unknown link function is $g(\cdot) = \sin(\cdot)$.

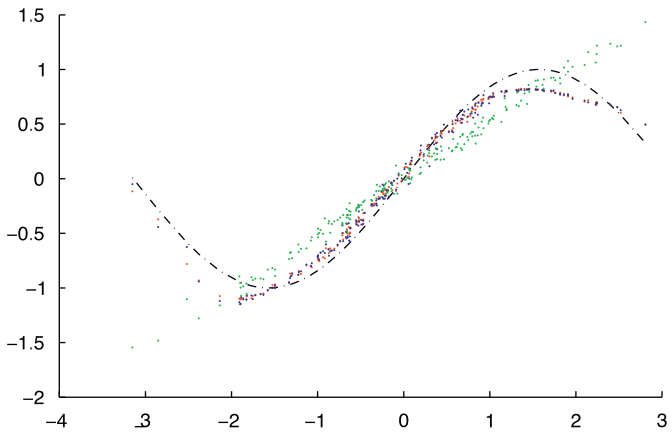

Fig. 2. Estimated unknown function $g(\cdot)$. Black: true function; Red: estimated function by SIM-SCAD; Blue: estimated function by SIM^1; Green: estimated linear function by LM-SCAD. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Standard deviations of estimates.

Sample size	β_1		β_2		β_5	
	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})	SD	SD_m (SD_{mad})
100	0.0418	0.0368 (0.0073)	0.0737	0.0576 (0.0075)	0.0369	0.0406 (0.0060)
200	0.0304	0.0250 (0.0032)	0.0463	0.0406 (0.0052)	0.0254	0.0282 (0.0034)
400	0.0181	0.0173 (0.0017)	0.0296	0.0282 (0.0023)	0.0213	0.0198 (0.0015)
800	0.0153	0.0123 (0.0009)	0.0212	0.0199 (0.0012)	0.0143	0.0140 (0.0011)

Table 4
Probability of identifying.

$p=n^{\alpha}$	n=100		n=200		n=400	
	TS	*TN*	*TS*	*TN*	*TS*	*TN*
$\alpha=1/3$	1	0.9175	1	0.9410	1	0.9182
$\alpha=2/5$	1	0.9287	1	0.9171	1	0.9147
$\alpha=1/2$	1	0.9192	1	0.9218	1	0.9248

Table 5
Estimation result.

Method	β_1		β_2		β_5		Avg. no. of nonzero coefficient	
	Bias	SD	Bias	SD	Bias	SD	Correct	Incorrect
n=200, p=8								
SIM-SCAD^1	−0.0050	0.0304	0.0015	0.0463	0.0032	0.0254	3.0000	0.1700
SIM-SCAD^2	−0.0059	0.0310	0.0015	0.0457	0.0029	0.0258	3.0000	0.4400
SIM-LASSO^1	0.0144	0.0329	−0.0091	0.0541	−0.0207	0.0323	3.0000	0.5400
SIM-LASSO^2	0.0095	0.0306	−0.0069	0.0467	−0.0118	0.0323	3.0000	0.7000
n=200, p=16								
SIM-SCAD^1	−0.0017	0.0290	−0.0019	0.0462	−0.0017	0.0319	3.0000	0.6800
SIM-SCAD^2	−0.0050	0.0287	−0.0024	0.0475	−0.0011	0.0360	3.0000	1.3600
SIM-LASSO^1	0.0188	0.0289	−0.0118	0.0537	−0.0246	0.0418	3.0000	1.1250
SIM-LASSO^2	0.0142	0.0294	−0.0091	0.0489	−0.0212	0.0385	3.0000	1.7150
n=200, p=32								
SIM-SCAD^1	−0.0097	0.0300	0.0089	0.0504	0.0025	0.0331	2.9800	2.3000
SIM-SCAD^2	−0.0149	0.0321	0.0075	0.0505	0.0037	0.0380	2.9850	4.1800
SIM-LASSO^1	0.0124	0.0309	−0.0017	0.0527	−0.0216	0.0376	2.9850	1.7300
SIM-LASSO^2	0.0072	0.0294	−0.0018	0.0538	−0.0177	0.0322	2.9900	3.2950

First, let n=(100, 200, 400) and $p=n^{\alpha}, \alpha=(1/3, 2/5, 1/2)$. For all nine possible combinations of (n, p), we apply our proposed penalized least squares method SIM-SCAD^1 for the above model. Table 4 reports the result, where *TS* is the probability of selecting all true significant parameters, and *TN* is the probability of identifying all true nonsignificant parameters. From Table 4, it shows that our proposed penalized least squares method can still identify both significant and nonsignificant parameters with high probability when p is diverging with the sample size.

Next, let n=200 and p=(8,16,32). We compare the performance of SIM-SCAD and SIM-LASSO. AICc is not included because of its computation burden. From Table 5, it is easy to see that even when the dimension of parameters is large, our propose penalized least squares methods still work effectively. When p=8 and 16, SIM-SCAD^{1,2} perform better than SIM-LASSO^{1,2}, but when p=32, SIM-LASSO^{1,2} perform better than SIM-SCAD^{1,2}. When the dimension of parameters is moderate large, with a reasonable initial estimate, SIM-SCAD yields more accurate estimate than SIM-LASSO because SCAD penalty does not over penalized large parameters. When the dimension of parameters is relatively high, a good initial estimate may be hard to obtain, and SIM-SCAD may converge to a local minimum and yield inconsistent variable selection results. On the other hand, SIM-LASSO^{1,2} always converge to the global minimum because of the convexity of L_1 penalty, and yield more stable variable selection results than SIM-SCAD^{1,2}.

5. Discussion

In this paper, we propose a nonconcave penalized least squares method for the single index model. Compared to other methods, the proposed method can select variables and estimate parameters simultaneously. The oracle property of the proposed nonconcave penalized estimator is established when the dimension of parameters is a fixed constant. In addition, a simple plug-in method for selecting the tuning parameter λ is proposed and is shown to be effective by simulations and a real data analysis. In fact, with this simple plug-in tuning parameter, the computation cost of the proposed nonconcave penalized least squares method is similar to that of the ordinary least squares method.

For our proposed method, we simply use a modified Newton–Raphson algorithm to find the minimizer of the nonconcave penalized least squares function. Other new algorithms for minimizing the nonconcave penalized least squares function, such as Hunter and Li (2005) and Zou and Li (2008), can be also applied to reduce computation burden without any difficulty. Similar to our numerical study, other penalized methods, such as LASSO (Tibshirani, 1996) or Adaptive LASSO (Zou, 2006), can be also applied to the single index model to select variables and estimate parameters. Here we only provide a general applicable estimation and selection procedure for the single index model.

We preliminarily explore the situation when the number of parameters $\boldsymbol{\beta}$ is diverging with the number of observations for the single index model through simulations. According to our numerical results, the proposed nonconcave penalized least squares method can be applied to some high-dimensional single index models. However, if the dimension of significant parameters is larger than $n^{1/5}$, the proposed method may not be applicable because of the curse of dimensionality. In turn, the link function may not be estimated efficiently, and the efficiency of the proposed iterative algorithm may be reduced. Hence, in this situation, further theoretical investigations are needed to understand the effectiveness of our proposed nonconcave penalized least squares method.

Appendix A. Proof

Lemma A.1. *Let $U_0=\mathbf{X}^T\boldsymbol{\beta}_0, U=\mathbf{X}^T\boldsymbol{\beta}$ and $f_\beta(\cdot)$ be the density function of U. Let $g(u_0,h,\boldsymbol{\beta})$ be the local linear estimate of the link function $g(\cdot)$. Under the conditions in Section* 3.1, *we have*

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le Cn^{-1/2}}\sup_{u_0\in D}\left|g(u_0,h,\boldsymbol{\beta})-g(u_0)-\frac{1}{nf_{\beta_0}(u_0)}\sum_{i=1}^{n}K_h(U_{0i}-u_0)\varepsilon_i-(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^Tg'_0(u_0)\mathbf{E}(\mathbf{X}|U=u_0)\right|=\mathrm{O_p}(h^2)+\mathrm{O_p}(c_n^2),$$

where $c_n=\mathrm{O_p}(h^2+\sqrt{\log n/(nh)})$, and C is a constant.

Proof. To prove the lemma, we first show

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le Cn^{-1/2}}\sup_{u_0\in D}\left|\frac{1}{n}\sum_{i=1}^{n}\{K_h(U_i-u_0)Z_i-\mathrm{E}[K_h(U_i-u_0)Z_i]\}\right|=\mathrm{O_p}\left(\sqrt{\frac{\log n}{nh}}\right), \tag{A.1}$$

where Z_i, $i=1,2,\ldots,n$ is a smoothing function of $(\mathbf{X}_i, Y_i)$, and it is easy to show that $\mathrm{E}Z_i^4<\infty$, $i=1,2,\ldots,n$.

Similar to the proof of Lemma 6.1 in Fan and Yao (2003), for a given $\boldsymbol{\beta}$ and $\delta_n=\sqrt{a\log n/(nh)}$ with a sufficiently large a, we have

$$\mathrm{P}\left(\sup_{u_0\in D}\left|\frac{1}{n}\sum_{i=1}^{n}\{K_h(U_i-u_0)Z_i-\mathrm{E}[K_h(U_i-u_0)Z_i]\}\right|>\delta_n\right)=4n^{1-a}.$$

Next, by Lemma 2.5 in van de Geer (2000), we can partition the area $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le Cn^{-1/2}$ into $N\le(Ch^{-3/2}+1)^p$ sub-area I_j, $j=1,2,\ldots,N$, with center at $\boldsymbol{\beta}_j$ such that $\|\boldsymbol{\beta}-\boldsymbol{\beta}_j\|\le Cn^{-1/2}h^{3/2}$ when $\boldsymbol{\beta}\in I_j$.

Again similar to the proof of Lemma 6.1 in Fan and Yao (2003), it is easy to show that, for $\boldsymbol{\beta}\in I_j$, and $U_{ij}=\mathbf{X}_i\boldsymbol{\beta}_j$,

$$\sup_{\boldsymbol{\beta}\in I_j}\left|\frac{1}{n}\sum_{i=1}^{n}\{K_h(U_i-u_0)Z_i-\mathrm{E}[K_h(U_i-u_0)Z_i]\}-\frac{1}{n}\sum_{i=1}^{n}\{K_h(U_{ij}-u_0)Z_i-\mathrm{E}[K_h(U_{ij}-u_0)Z_i]\}\right|=\mathrm{O}(\sqrt{\log n/(nh)}).$$

Define

$$f(\boldsymbol{\beta})=\sup_{u_0\in D}\left|\frac{1}{n}\sum_{i=1}^{n}\{K_h(U_i-u_0)Z_i-\mathrm{E}[K_h(U_i-u_0)Z_i]\}\right|,$$

then

$$\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le Cn^{-1/2}}|f(\boldsymbol{\beta})|\le\min_j\sup_{\boldsymbol{\beta}\in I_j}|f(\boldsymbol{\beta})-f(\boldsymbol{\beta}_j)|+\sup_j|f(\boldsymbol{\beta}_j)|\le\mathrm{O}(\sqrt{\log n/(nh)})+\sup_j|f(\boldsymbol{\beta}_j)|.$$

Hence given a sufficiently large a, we have

$$\begin{aligned}&\mathrm{P}\left(\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le Cn^{-1/2}}\left|\frac{1}{n}\sum_{i=1}^{n}\{K_h(U_i-u_0)Z_i-\mathrm{E}[K_h(U_i-u_0)Z_i]\}\right|>\delta_n+\mathrm{O}(\sqrt{\log n/(nh)})\right)\\&\quad\le\sum_{j=1}^{N}\mathrm{P}\left(\sup_{u_0\in D}\left|\frac{1}{n}\sum_{i=1}^{n}\{K_h(U_{ij}-u_0)Z_i-\mathrm{E}[K_h(U_{ij}-u_0)Z_i]\}\right|>\delta_n\right)\le N\cdot4n^{1-a}\longrightarrow0,\end{aligned}$$

and this proves (A.1).

Since $g(u_0,h,\boldsymbol{\beta})$ is the local linear estimate of the link function $g(\cdot)$, we have

$$g(u_0,h,\boldsymbol{\beta})=\sum_{i=1}^{n}W_0^n\left(\frac{U_i-u_0}{h}\right)Y_i=\sum_{i=1}^{n}W_0^n\left(\frac{U_i-u_0}{h}\right)g(U_{0i})+\sum_{i=1}^{n}W_0^n\left(\frac{U_i-u_0}{h}\right)\varepsilon_i\hat{=}I+II, \tag{A.2}$$

where $W_0^n(\cdot)$ is the equivalent kernel for the local linear regression (Fan and Gijbels, 1996).

For (I), by Taylor expansion, we have

$$\sum_{i=1}^{n} W_0^n\left(\frac{U_i-u_0}{h}\right)g(U_{0i}) = \sum_{i=1}^{n} W_0^n\left(\frac{U_i-u_0}{h}\right)\{g(U_i)+g'(U_i)(U_{0i}-U_i)+g''(\kappa_i)(U_{i0}-U_i)^2\} \hat{=} T_1+T_2+T_3, \tag{A.3}$$

where

$$T_1 = \sum_{i=1}^{n} W_0^n\left(\frac{U_i-u_0}{h}\right)\{g(u_0)+g'(u_0)(U_i-u_0)+g''(\tau_i)(U_i-u_0)^2\},$$

$$T_2 = \sum_{i=1}^{n} W_0^n\left(\frac{U_i-u_0}{h}\right)\{g'(u_0)(U_{0i}-U_i)+g''(\eta_i)(U_i-u_0)(U_{0i}-U_i)\},$$

$$T_3 = \sum_{i=1}^{n} W_0^n\left(\frac{U_i-u_0}{h}\right)\{g''(\kappa_i)(U_{i0}-U_i)^2\},$$

and κ_i, τ_i and η_i are values between U_i and u_0.

Then by properties of the equivalent kernel property of $W_0^n(\cdot)$ and (A.1), it is easy to show that

$$\sup_{\|\beta-\beta_0\|\le Cn^{-1/2}}\sup_{u_0\in D}|T_1-g(u_0)| = O(h^2(1+c_n)),$$

$$\sup_{\|\beta-\beta_0\|\le Cn^{-1/2}}\sup_{u_0\in D}|T_2-(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T g'(u_0)\mathbf{E}(\mathbf{X}|U=u_0)| = O_p(\|\beta-\beta_0\|\cdot(c_n+h)),$$

and

$$\sup_{\|\beta-\beta_0\|\le Cn^{-1/2}}\sup_{u_0\in D}|T_3| = O_p(\|\beta-\beta_0\|^2).$$

Together we have

$$\sup_{\|\beta-\beta_0\|\le Cn^{-1/2}}\sup_{u_0\in D}\left|\sum_{i=1}^{n} W_0^n\left(\frac{U_i-u_0}{h}\right)g(U_{0i})-g(u_0)-(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T g'(u_0)\mathbf{E}\{\mathbf{X}|U=u_0\}\right| = O(h^2). \tag{A.4}$$

Similarly, for (II), it is easy to show that

$$\sup_{\|\beta-\beta_0\|\le Cn^{-1/2}}\sup_{u_0\in D}\left|\sum_{i=1}^{n} W_0^n\left(\frac{U_i-u_0}{h}\right)\varepsilon_i-\frac{1}{nf_{\beta_0}(u_0)}\sum_{i=1}^{n} K_h(U_{0i}-u_0)\varepsilon_i\right| = O_p(c_n^2). \tag{A.5}$$

By (A.2)–(A.5), it concludes the proof of Lemma A.1. □

A.1. Proof of Theorem 3.1

Let

$$Q(\hat{g}_\beta,\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i-\hat{g}_\beta(\mathbf{X}_i^T\boldsymbol{\beta}))^2 + n\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|),$$

where $\hat{g}_\beta$ is the local linear regression of the link function $g(\cdot)$ with $\boldsymbol{\beta}$.

We need only to show that for any given ε there exists a large C such that

$$P\left\{\sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|=C(n^{-1/2}+a_n),\|\boldsymbol{\beta}\|=1} Q(\hat{g}_\beta,\boldsymbol{\beta}) > Q(\hat{g}_{\beta_0},\boldsymbol{\beta}_0)\right\} \ge 1-\varepsilon. \tag{A.6}$$

This implies with probability at least $1-\varepsilon$ that there exists a local minimum in the ball $\{\boldsymbol{\beta}: \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| \le C(n^{-1/2}+a_n), \|\boldsymbol{\beta}\|=1\}$. Hence there exists a local minimizer such that $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| = O_p(n^{-1/2}+a_n)$.

Since $p_{\lambda_n}(0)=0$, we have

$$Q(\hat{g}_\beta,\boldsymbol{\beta})-Q(\hat{g}_{\beta_0},\boldsymbol{\beta}_0) \ge \sum_{i=1}^{n}\{(y_i-\hat{g}_\beta(\mathbf{X}_i^T\boldsymbol{\beta}))^2-(y_i-\hat{g}_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0))^2\} + \sum_{j=1}^{s}\{np_{\lambda_n}(|\beta_j|)-np_{\lambda_n}(|\beta_{j0}|)\}. \tag{A.7}$$

As shown by Fan and Li (2001), the second term in (A.7) is bounded by

$$\sqrt{s}Cn(n^{-1/2}+a_n)a_n + C^2n(n^{-1/2}+a_n)^2\max\{|p''_{\lambda_n}(|\beta_{j0}|: \beta_{j0}\ne 0)\}. \tag{A.8}$$

Next we show that (A.8) is bounded by the first term in (A.7). Denote $\hat{U}_i = \mathbf{X}_i^T\hat{\boldsymbol{\beta}}$ and $U_i = \mathbf{X}_i^T\boldsymbol{\beta}_0$. We then can write the first term in (A.7) as

$$\begin{aligned}\sum_{i=1}^n\{(y_i-\hat{g}_{\beta}(\mathbf{X}_i^T\boldsymbol{\beta}))^2-(y_i-\hat{g}_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0))^2\} &= \sum_{i=1}^n\left(y_i-\sum_{j=1}^n W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)y_i\right)^2-\sum_{i=1}^n\left(y_i-\sum_{j=1}^n W_0^n\left(\frac{U_j-U_i}{h}\right)y_i\right)^2\\ &= 2\sum_{i=1}^n\left(y_i-\sum_{j=1}^n W_0^n\left(\frac{U_j-U_i}{h}\right)y_i\right)\left(\sum_{j=1}^n W_0^n\left(\frac{U_j-U_i}{h}\right)y_i-\sum_{j=1}^n W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)y_i\right)\\ &\quad+\sum_{i=1}^n\left(\sum_{j=1}^n W_0^n\left(\frac{U_j-U_i}{h}\right)y_i-\sum_{j=1}^n W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)y_i\right)^2 \hat{=} I_1+I_2. \end{aligned}\tag{A.9}$$

First consider I_2 in (A.9). It is easy to know that

$$\begin{aligned}\sum_{j=1}^n W_0^n\left(\frac{U_j-U_i}{h}\right)y_j-\sum_{j=1}^n W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)y_j &= \sum_{j=1}^n W_0^n\left(\frac{U_j-U_i}{h}\right)g(U_j)-\sum_{j=1}^n W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)g(U_j)\\ &\quad+\sum_{j=1}^n W_0^n\left(\frac{U_j-U_i}{h}\right)\varepsilon_j-\sum_{j=1}^n W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)\varepsilon_j.\end{aligned}\tag{A.10}$$

For the first two terms in (A.10), we have

$$\begin{aligned}\sum_{j=1}^n W_0^n\left(\frac{U_j-U_i}{h}\right)g(U_j)-\sum_{j=1}^n W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)g(U_j) &= g(U_i)-g(\hat{U}_i)-g'(\hat{U}_i)\mathbf{E}(\mathbf{X}_i|\hat{U}=\hat{U}_i)(\boldsymbol{\beta}_0-\boldsymbol{\beta})+\mathrm{O_p}(h^2)\\ &= g'(\hat{U}_i)(\mathbf{X}_i-\mathbf{E}(\mathbf{X}_i|\hat{U}=\hat{U}_i))(\boldsymbol{\beta}_0-\boldsymbol{\beta})+\mathrm{O_p}(h^2).\end{aligned}\tag{A.11}$$

Hence

$$\sum_{i=1}^n\left(\sum_{j=1}^n W_0^n\left(\frac{U_j-U_i}{h}\right)g(U_j)-\sum_{j=1}^n W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)g(U_j)\right)^2 = n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\Sigma^*(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+n\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\cdot\mathrm{O_p}(h^2)+\mathrm{O_p}(nh^4)\tag{A.12}$$

where $\Sigma^* = \mathbf{E}g'^2(\hat{U}_i)\{\mathbf{X}_i-\mathbf{E}(\mathbf{X}_i|\hat{U}=\hat{U}_i)\}\{\mathbf{X}_i-\mathbf{E}(\mathbf{X}_i|\hat{U}=\hat{U}_i)\}^T$. Since $\|\boldsymbol{\beta}\| = \|\boldsymbol{\beta}_0\| = 1$, $\boldsymbol{\beta}$ can be written as $\boldsymbol{\beta} = \tau\mathbf{v}+(1-\tau^2)^{1/2}\boldsymbol{\beta}_0$ where $\mathbf{v}^T\mathrm{Cov}(\mathbf{X})\boldsymbol{\beta}_0 = 0$. Then by Condition (C) it is easy to show that $(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^T\Sigma^*(\boldsymbol{\beta}-\boldsymbol{\beta}_0) > c\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2$, where c is a positive constant. This shows that the order of (A.12) is no smaller than $nc\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2$.

On the other hand, define

$$\sum_{i=1}^n\left\{\sum_{j=1}^n W_0^n\left(\frac{U_j-U_i}{h}\right)\varepsilon_j-\sum_{j=1}^n W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)\varepsilon_j\right\}^2 = \boldsymbol{\varepsilon}^T(S_\beta-S_{\beta_0})^T(S_\beta-S_{\beta_0})\boldsymbol{\varepsilon} \hat{=} V_1.$$

Note that

$$\mathbf{E}V_1 = \sigma^2\mathbf{E}[\mathrm{Tr}((S_\beta-S_{\beta_0})^T(S_\beta-S_{\beta_0}))] = \sum_{i=1}^n\mathbf{E}S_{ii},$$

where $\mathrm{Tr}(\cdot)$ is trace operator, and

$$S_{ii} = [(S_\beta-S_{\beta_0})^T(S_\beta-S_{\beta_0})]_{ii} = \sum_{j=1}^n\left\{W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)-W_0^n\left(\frac{U_j-U_i}{h}\right)\right\}^2.$$

By Taylor expansion, it is easy to show that

$$W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right) = \frac{f_\beta^{-1}(U_i)+\hat{c}_{ni}}{n}\left\{K_h(U_j-U_i)+K'_h(U_j^*-U_i^*)\cdot\frac{(X_j-X_i)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{h}\right\},$$

$$W_0^n\left(\frac{U_j-U_i}{h}\right) = \frac{f_\beta^{-1}(U_i)+c_{ni}}{n}K_h(U_j-U_i).$$

Hence

$$W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)-W_0^n\left(\frac{U_j-U_i}{h}\right) = \frac{\hat{c}_{ni}+c_{ni}}{n}K_h(U_j-U_i)+\left\{\frac{\hat{c}_{ni}}{n}+\frac{1}{nf(U_i)}\right\}K'_h(U_j^*-U_i^*)\cdot\frac{(X_j-X_i)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{h},$$

and

$$S_{ii} \le 2\sum_{j=1}^{n}\left\{(\hat{c}_{ni}+c_{ni})\cdot\frac{1}{n}K_h(U_j-U_i)\right\}^2+2\sum_{j=1}^{n}\left\{\frac{\hat{c}_{ni}}{n}+\frac{1}{nf(U_i)}\right\}^2\left\{K'_h(U_j^*-U_i^*)\cdot\frac{(X_j-X_i)(\boldsymbol{\beta}-\boldsymbol{\beta}_0)}{h}\right\}^2 \hat{=} S_{i1}+S_{i2}.$$

where by Lemma A.1, c_{ni} and $\hat{c}_{ni}$ should have the same order as $c_n\mathrm{O_p}(h^2+\sqrt{\log n/(nh)})$. It is obvious that

$$S_{i1}=\mathrm{O_p}(c_n^2/(nh)) \quad \text{and} \quad S_{i2}=\mathrm{O_p}(1/n\cdot h^3\cdot\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2).$$

Then by conditions $nh^3\to\infty$ and $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|=C\cdot n^{-1/2}$, we have

$$\mathbf{E}V_1=\sigma^2\mathbf{E}[\mathrm{Tr}((S_\beta-S_{\beta_0})^T(S_\beta-S_{\beta_0}))]=\sum_{i=1}^{n}\mathbf{E}S_{ii}=n\cdot\mathrm{O_p}(c_n^2/(nh))+n\cdot\mathrm{O_p}(1/n\cdot h^3\cdot\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2)=\mathrm{o_p}(1). \tag{A.13}$$

Let $\eta=\mathbf{E}\varepsilon_i^4$, then

$$\mathbf{E}V_1^2\le\eta\cdot 2\cdot\mathbf{E}[\mathrm{Tr}((S_\beta-S_{\beta_0})^T(S_\beta-S_{\beta_0})(S_\beta-S_{\beta_0})^T(S_\beta-S_{\beta_0}))].$$

By the definition of S_{ij},

$$\sum_{j=1}^{n}|S_{ij}|\le\frac{1}{2}\sum_{j=1}^{n}(S_{ii}+S_{jj})=\frac{n}{2}S_{ii}+\frac{1}{2}\mathrm{Tr}((S_\beta-S_{\beta_0})^T(S_\beta-S_{\beta_0}))=\mathrm{o}(1),$$

and by Gersgörin theorem (Quarteroni et al., 2000),

$$|\lambda_i((S_\beta-S_{\beta_0})(S_\beta-S_{\beta_0})^T)|=\mathrm{o}(1)<1.$$

where $\lambda_i, i=1,\ldots,p$ are eigenvalues of $(S_\beta-S_{\beta_0})(S_\beta-S_{\beta_0})^T$. Then we have

$$\mathbf{E}V_1^2\le 2\eta\cdot\mathbf{E}\,\mathrm{Tr}\{(S_\beta-S_{\beta_0})^T(S_\beta-S_{\beta_0})(S_\beta-S_{\beta_0})^T(S_\beta-S_{\beta_0})\}\le 2\eta\cdot\mathbf{E}\,\mathrm{Tr}\{(S_\beta-S_{\beta_0})^T(S_\beta-S_{\beta_0})\}=\mathrm{o}(1). \tag{A.14}$$

By (A.13) and (A.14), it is easy to show that

$$V_1=\mathrm{o_p}(1). \tag{A.15}$$

Consider

$$V_2=\sum_{i=1}^{n}\left\{\sum_{j=1}^{n}W_0^n\left(\frac{U_j-U_i}{h}\right)g(U_j)-\sum_{j=1}^{n}W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)g(U_j)\right\}\times\left\{\sum_{j=1}^{n}W_0^n\left(\frac{U_j-U_i}{h}\right)\varepsilon_j-\sum_{j=1}^{n}W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)\varepsilon_j\right\}.$$

By (A.11),

$$\sum_{j=1}^{n}W_0^n\left(\frac{U_j-U_i}{h}\right)g(U_j)-\sum_{j=1}^{n}W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)g(U_j)=\mathrm{O_p}(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+h^2),$$

and together with

$$\sum_{i=1}^{n}W_0^n\left(\frac{U_j-U_i}{h}\right)-W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)=\mathrm{O_p}(c_n),$$

we have

$$\begin{aligned}V_2&=\sum_{i=1}^{n}\mathrm{O_p}(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+h^2)\left\{\sum_{j=1}^{n}W_0^n\left(\frac{U_j-U_i}{h}\right)\varepsilon_j-\sum_{j=1}^{n}W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)\varepsilon_j\right\}\\&=\sum_{j=1}^{n}\sum_{i=1}^{n}\mathrm{O_p}(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+h^2)\left\{W_0^n\left(\frac{U_j-U_i}{h}\right)-W_0^n\left(\frac{\hat{U}_j-\hat{U}_i}{h}\right)\right\}\varepsilon_j=\mathrm{o_p}(\sqrt{n}c_n(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+h^2))=\mathrm{o_p}(\sqrt{n}(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+h^2)).\end{aligned} \tag{A.16}$$

By (A.12), (A.15) and (A.16), we have

$$I_2\ge nc\|\boldsymbol{\beta}_0-\boldsymbol{\beta}\|^2(1+\mathrm{o_p}(1)). \tag{A.17}$$

Now we consider I_1 in (A.9),

$$\begin{aligned}\tfrac{1}{2}I_1&=\mathbf{y}^T(I-S_{\beta_0})^T(S_{\beta_0}-S_\beta)\mathbf{y}=\boldsymbol{\varepsilon}^T(I-S_{\beta_0})^T(S_{\beta_0}-S_\beta)\boldsymbol{\varepsilon}+g^T(U)(I-S_{\beta_0})^T(S_{\beta_0}-S_\beta)g(U)\\&\quad+\boldsymbol{\varepsilon}^T(I-S_{\beta_0})^T(S_{\beta_0}-S_\beta)g(U)+g^T(U)(I-S_{\beta_0})^T(S_{\beta_0}-S_\beta)\boldsymbol{\varepsilon}\hat{=}L_1+L_2+L_3.\end{aligned} \tag{A.18}$$

For L_2 in (A.18), by local linear regression,

$$g^T(U)(I-S_{\beta_0})=\mathrm{O_p}(h^2),$$

and

$$(S_{\beta_0}-S_\beta)g(U)=\mathrm{O_p}(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|+h^2),$$

so we have

$$L_2=(nh^4+n\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|h^2)=\mathrm{o_p}(n\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2). \tag{A.19}$$

Similar to the computation of V_2, for the second term of L_3 we have

$$g^T(U)(I-S_{\beta_0})^T(S_{\beta_0}-S_\beta)\boldsymbol{\varepsilon}=\mathrm{O_p}(h^2\cdot\sqrt{\mathrm{Tr}\{(S_{\beta_0}-S_\beta)(S_{\beta_0}-S_\beta)^T\}})=\mathrm{O}(h^2). \tag{A.20}$$

For the first term of L_3, we have

$$\boldsymbol{\varepsilon}^T(I-S_{\beta_0})^T(S_{\beta_0}-S_\beta)g(U)=\boldsymbol{\varepsilon}^T(I-S_{\beta_0})\mathrm{O}(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|)=\mathrm{O_p}(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\sqrt{\mathrm{Tr}\{(I-S_{\beta_0})^T(I-S_{\beta_0})\}})=\mathrm{O_p}(\sqrt{n}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|). \tag{A.21}$$

Finally for L_1, we have

$$L_1=\boldsymbol{\varepsilon}^T(S_{\beta_0}-S_\beta)\boldsymbol{\varepsilon}+\boldsymbol{\varepsilon}^TS_{\beta_0}^T(S_{\beta_0}-S_\beta)\boldsymbol{\varepsilon}\hat{=}F_1+F_2.$$

As shown before, for F_1,

$$\mathbf{E}F_1^2\le 2\eta\cdot\mathrm{Tr}((S_{\beta_0}-S_\beta)^T(S_{\beta_0}-S_\beta))=\mathrm{o}(1), \tag{A.22}$$

and for F_2,

$$\mathbf{E}F_2^2\le 2\eta\cdot\mathrm{Tr}((S_{\beta_0}-S_\beta)^TS_{\beta_0}^TS_{\beta_0}(S_{\beta_0}-S_\beta))\le 2\eta\cdot\mathrm{Tr}((S_{\beta_0}-S_\beta)^T(S_{\beta_0}-S_\beta))=\mathrm{o}(1). \tag{A.23}$$

Then by (A.22) and (A.23), we have

$$L_1=\mathrm{o_p}(1). \tag{A.24}$$

Then by (A.19)–(A.21) and (A.24), we have

$$I_1=\mathrm{O_p}(\sqrt{n}\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|). \tag{A.25}$$

By (A.17) and (A.25), we know that I_1 and I_2 are dominated by $n\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|^2\mathbf{E}g'^2(\hat{U}_i)\{\mathbf{X}_i-\mathbf{E}(\mathbf{X}_i|\hat{U}=\hat{U}_i)\}^2$, and it dominates the penalty term, see (A.8). Hence it is easy to show (A.6) when C is large enough. □

A.2. Proof of Lemma 3.1

It is sufficient to show that, with probability tending to 1 as $n\to\infty$, for any $\boldsymbol{\beta}_1$ satisfying $\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10}=\mathrm{O_p}(n^{-1/2})$, and for some small $\varepsilon_n=Cn^{-1/2}$, we have

$$\frac{\partial Q(\hat{g}_\beta,\boldsymbol{\beta})}{\partial\beta_j}<0 \text{ for } 0<\beta_j<\varepsilon_n \tag{A.26}$$

$$>0 \text{ for } -\varepsilon_n<\beta_j<0, \tag{A.27}$$

where $j=s+1,\ldots,p$.

To show (A.26), by (2.6) and (2.8) we have

$$\begin{aligned}\frac{\partial Q(\hat{g}_\beta,\boldsymbol{\beta})}{\partial\beta_j}&=\sum_{i=1}^n(Y_i-\hat{g}_\beta(\mathbf{X}_i^T\boldsymbol{\beta}))\hat{g}'_\beta(\mathbf{X}_i^T\boldsymbol{\beta})X_{ij}+np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j)\\&=\sum_{i=1}^n\varepsilon_i\hat{g}'_\beta(\mathbf{X}_i^T\boldsymbol{\beta})X_{ij}+\sum_{i=1}^n(g(\mathbf{X}_i^T\boldsymbol{\beta}_0)-\hat{g}_\beta(\mathbf{X}_i^T\boldsymbol{\beta}))\hat{g}'_\beta(\mathbf{X}_i^T\boldsymbol{\beta})X_{ij}+np'_{\lambda_n}(|\beta_j|)\mathrm{sgn}(\beta_j)\hat{=}L_1+L_2+L_3.\end{aligned} \tag{A.28}$$

By the definition of $\hat{g}_\beta$, L_1 can be written as

$$L_1=\sum_{i=1}^n\varepsilon_iX_{ij}\sum_{l=1}^nW_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right)Y_l=\sum_{i=1}^n\varepsilon_iX_{ij}\sum_{l=1}^nW_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right)g(\mathbf{X}_l^T\boldsymbol{\beta}_0)+\sum_{i=1}^n\varepsilon_iX_{ij}\sum_{l=1}^nW_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right)\varepsilon_l\hat{=}N_1+N_2.$$

By the conditions (A)–(C) in Section 3.1, it is easy to show that

$$X_{ij}\sum_{l=1}^nW_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right)g(\mathbf{X}_l^T\boldsymbol{\beta}_0),\quad j=1,2,\ldots,n,$$

is bounded, and independent of $\boldsymbol{\varepsilon}$. Hence

$$N_1=\sum_{i=1}^n\varepsilon_iX_{ij}\sum_{l=1}^nW_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right)g(\mathbf{X}_l^T\boldsymbol{\beta}_0)=\mathrm{O_p}(\sqrt{n}). \tag{A.29}$$

For N_2, by the definition of $W_1^n(\cdot)$ and the results of local polynomial regression, we have

$$\sum_{i=1}^{n}\sum_{l=1}^{n} X_{ij}^2\left\{W_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right)\right\}^2 = \mathrm{O_p}(h^{-3}).$$

Hence

$$\mathrm{Var}(N_2) = \mathrm{O_p}(h^{-3}). \tag{A.30}$$

In addition, it is easy to show that $W_1^n(0) = \mathrm{O_p}((h+1/\sqrt{nh})/nh^2)$, and then

$$\mathrm{E}(N_2) = \sum_{i=1}^{n} \mathrm{E}X_{ij}\varepsilon_i^2 W_1^n(0) = \mathrm{O}(h^{-3/2}). \tag{A.31}$$

Therefore by (A.30) and (A.31), we have

$$N_2 = \mathrm{O_p}(h^{-3/2}) = \mathrm{O_p}(\sqrt{n}).$$

Together with (A.29),

$$L_1 = \mathrm{O_p}(\sqrt{n}). \tag{A.32}$$

For L_2, it can be expressed as

$$L_2 = \sum_{i=1}^{n}(g(\mathbf{X}_i^T\boldsymbol{\beta}_0)-g(\mathbf{X}_i^T\boldsymbol{\beta})+g(\mathbf{X}_i^T\boldsymbol{\beta})-\hat{g}_{\beta}(\mathbf{X}_i^T\boldsymbol{\beta}))\hat{g}'_{\beta}(\mathbf{X}_i^T\boldsymbol{\beta})X_{ij}.$$

By the Taylor expansion, it is easy to see that

$$\sum_{i=1}^{n}(g(\mathbf{X}_i^T\boldsymbol{\beta}_0)-g(\mathbf{X}_i^T\boldsymbol{\beta}))\hat{g}'_{\beta}(\mathbf{X}_i^T\boldsymbol{\beta})X_{ij} = \mathrm{O_p}(n\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|) = \mathrm{O_p}(\sqrt{n}). \tag{A.33}$$

By Lemma 5.1 and similar to the equation of (A.13) in Carroll et al. (1997)

$$\begin{aligned}\sum_{i=1}^{n}(g(\mathbf{X}_i^T\boldsymbol{\beta})-\hat{g}_{\beta}(\mathbf{X}_i^T\boldsymbol{\beta}))\hat{g}'_{\beta}(\mathbf{X}_i^T\boldsymbol{\beta})X_{ij} = \sum_{i=1}^{n}\Bigg(&\frac{1}{nf_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0)}\sum_{k=1}^{n}K_h(\mathbf{X}_k^T\boldsymbol{\beta}_0-\mathbf{X}_i^T\boldsymbol{\beta}_0)\varepsilon_k-(\boldsymbol{\beta}^T-\boldsymbol{\beta}_0^T)g'_0(u_0)\mathbf{E}\{\mathbf{X}|\hat{U}=u_0\}\\ &+g'(\mathbf{X}_i^T\boldsymbol{\beta}_0)\mathbf{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\mathrm{o_p}(n^{-1/2})\Bigg)\hat{g}'_{\beta}(\mathbf{X}_i^T\boldsymbol{\beta})X_{ij}.\end{aligned}$$

Since $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\| = \mathrm{O}(n^{-1/2})$, it is easy to show that

$$\sum_{i=1}^{n}(-(\boldsymbol{\beta}^T-\boldsymbol{\beta}_0^T)g'_0(u_0)\mathbf{E}\{\mathbf{X}|\hat{U}=u_0\}+g'(\mathbf{X}_i^T\boldsymbol{\beta}_0)\mathbf{X}_i^T(\boldsymbol{\beta}-\boldsymbol{\beta}_0)+\mathrm{o_p}(n^{-1/2}))\hat{g}'_{\beta}(\mathbf{X}_i^T\boldsymbol{\beta})X_{ij} = \mathrm{O_p}(\sqrt{n}). \tag{A.34}$$

In addition,

$$\hat{g}'_{\beta}(\mathbf{X}_i^T\boldsymbol{\beta})X_{ij} = (\mathrm{E}(Y|\mathbf{X}_i^T\boldsymbol{\beta}))^T X_{ij}+\mathrm{O_p}(h)+X_{ij}\sum_{l=1}^{n}W_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right)\varepsilon_l$$

and

$$\sum_{i=1}^{n}\left(\frac{1}{nf_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0)}\sum_{k=1}^{n}K_h(\mathbf{X}_k^T\boldsymbol{\beta}_0-\mathbf{X}_i^T\boldsymbol{\beta}_0)\varepsilon_k\right)\hat{g}'_{\beta}(\mathbf{X}_i^T\boldsymbol{\beta})X_{ij} \doteq M_1+M_2+\mathrm{O_p}(\sqrt{n}).$$

where

$$M_1 = \sum_{i=1}^{n}\left(\frac{1}{nf_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0)}\sum_{k=1}^{n}K_h(\mathbf{X}_k^T\boldsymbol{\beta}_0-\mathbf{x}_i\boldsymbol{\beta}_0)\varepsilon_k\right)(\mathrm{E}(Y|\mathbf{X}_i^T\boldsymbol{\beta}))^T X_{ij},$$

$$M_2 = \sum_{i=1}^{n}\left(\frac{1}{nf_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0)}\sum_{k=1}^{n}X_{ij}K_h(\mathbf{X}_k^T\boldsymbol{\beta}_0-\mathbf{X}_i^T\boldsymbol{\beta}_0)\varepsilon_j\right)\sum_{l=1}^{n}W_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right)\varepsilon_l.$$

Since

$$\sum_{i=1}^{n}\frac{1}{nf_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0)}K_h(\mathbf{X}_k^T\boldsymbol{\beta}_0-\mathbf{X}_i^T\boldsymbol{\beta}_0)(\mathrm{E}(Y|\mathbf{X}_i^T\boldsymbol{\beta}))'X_{ij} = \mathrm{O_p}(1),$$

we have

$$M_1 = \sum_{i=1}^{n}\left(\frac{1}{nf_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0)}\sum_{k=1}^{n}K_h(\mathbf{X}_k^T\boldsymbol{\beta}_0-\mathbf{X}_i^T\boldsymbol{\beta}_0)\varepsilon_j\right)(\mathrm{E}(Y|\mathbf{X}_i^T\boldsymbol{\beta}))'X_{ij}$$

$$= \sum_{k=1}^{n}\left(\sum_{i=1}^{n}\frac{1}{nf_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0)}K_h(\mathbf{X}_k^T\boldsymbol{\beta}_0-\mathbf{X}_i^T\boldsymbol{\beta}_0)(\mathrm{E}(Y|\mathbf{X}_i^T\boldsymbol{\beta}))'X_{ij}\right)\varepsilon_k = \mathrm{O_p}(\sqrt{n}). \tag{A.35}$$

M_2 can be written as

$$M_2 = \sum_{k=1}^{n}\sum_{l=1}^{n}\varepsilon_j\varepsilon_l\sum_{i=1}^{n}\frac{X_{ij}}{nf_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0)}K_h(\mathbf{X}_k^T\boldsymbol{\beta}_0-\mathbf{X}_i^T\boldsymbol{\beta}_0)W_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right).$$

By the property of equivalent kernel function, it is easy to show that

$$W_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right) = \frac{1}{nh^2f(\mathbf{X}_i^T\boldsymbol{\beta})\mu_2}\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}K\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right)(1+\mathrm{o_p}(1)).$$

Hence

$$\mathrm{E}\left\{\sum_{i=1}^{n}\frac{X_{ij}}{nf_{\beta_0}(\mathbf{X}_i^T\boldsymbol{\beta}_0)}K_h(\mathbf{X}_k^T\boldsymbol{\beta}_0-\mathbf{x}_i\boldsymbol{\beta}_0)W_1^n\left(\frac{\mathbf{X}_l^T\boldsymbol{\beta}-\mathbf{X}_i^T\boldsymbol{\beta}}{h}\right)\right\}^2 = \mathrm{O_p}\left(\frac{1}{n^2h^3}\right),$$

and then we have

$$\mathrm{E}M_2^2 = \mathrm{O_p}\left(n^2\left(\frac{1}{n^2h^3}\right)\right) \quad \text{and} \quad M_2 = \mathrm{O_p}(h^{-3/2}) = \mathrm{O_p}(\sqrt{n}). \tag{A.36}$$

By (A.33)–(A.36), we have

$$L_2 = \mathrm{O_p}(\sqrt{n}). \tag{A.37}$$

By (A.32), (A.37) and the assumption $\boldsymbol{\beta}_1-\boldsymbol{\beta}_{10} = \mathrm{O_p}(n^{-1/2})$, we have

$$\frac{\partial Q(\hat{g}_\beta,\boldsymbol{\beta})}{\partial\beta_j} = n\lambda_n(\lambda_n^{-1}p_{\lambda_n}'\mathrm{sgn}(\beta_j)+\mathrm{O_p}(n^{-1/2}/\lambda_n)), \quad j=s+1,\ldots,p$$

Whereas $\liminf_{n\to\infty}\liminf_{\theta\to0^+}\lambda_n^{-1}p'_{\lambda_n}(\theta)>0$ and $n^{-1/2}/\lambda\to0$, the sign of the derivative is completely determined by that of β_j. Hence, (A.26) and (A.27) follow. This completes the proof of Lemma 3.1. □

A.3. Proof of Theorem 3.2

It follows by Lemma 3.1 that part (a) holds. Now we prove part (b). It can be easily shown that there exists a $\hat{\boldsymbol{\beta}}_1$ in Theorem 3.1 which is a root-n consistent local maximizer of $Q\{g_{(\beta_1^T,0)^T},(\beta_1^T,0)^T\}$ under the constraint $\|\boldsymbol{\beta}_1\|=1$. Hence with θ being the Lagrange multiplier, we know that $(\hat{\boldsymbol{\beta}}_1^T,0)^T$ is the solution to

$$0 = \theta\hat{\boldsymbol{\beta}}_1+n^{-1/2}\sum_{i=1}^{n}(Y_i-\hat{g}_{\hat{\beta}}(\mathbf{X}_i^T\hat{\boldsymbol{\beta}}))\hat{g}'_{\hat{\beta}}(\mathbf{X}_i^T\hat{\boldsymbol{\beta}})\mathbf{X}_i+n^{1/2}(p_{\lambda_n}'(|\beta_1|)\mathrm{sgn}(\beta_1),\ldots,p_{\lambda_n}'(|\beta_s|)\mathrm{sgn}(\beta_s))^T.$$

By the Taylor expansion and similar to Eq. (37) in Carroll et al. (1997), given $nh^4\to0$, we have

$$0 = \theta\hat{\boldsymbol{\beta}}_1+n^{-1/2}\sum_{i=1}^{n}\mathbf{X}_i\hat{g}_{\beta_0}'(U_i)\varepsilon_i-n^{-1/2}\mathbf{X}_i\hat{g}_{\beta_0}'(U_i)\{\hat{g}_{\hat{\beta}}(\mathbf{X}_i^T\hat{\boldsymbol{\beta}})-g(\mathbf{X}_i^T\boldsymbol{\beta}_0)\}+n^{1/2}\mathbf{b}+n^{1/2}\Sigma(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01})+\mathrm{o_p}(1). \tag{A.38}$$

Next by expanding $\hat{g}(\cdot)$ and by Lemma A.1, we have

$$0 = \theta\hat{\boldsymbol{\beta}}_1+n^{-1/2}\sum_{i=1}^{n}\mathbf{X}_ig'(U_i)\varepsilon_i+Wn^{1/2}(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01})-n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mathbf{X}_ig'(U_i)}{nf(U_i)}K_h(U_j-U_i)\varepsilon_j+n^{1/2}\mathbf{b}+n^{1/2}\Sigma(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01})+\mathrm{o_p}(1). \tag{A.39}$$

Interchanging the summations we get

$$n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\mathbf{X}_ig'(U_i)}{nf(U_i)}K_h(U_j-U_i)\varepsilon_j = n^{-1/2}\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{\mathbf{X}_ig'(U_i)}{nf(U_i)}K_h(U_j-U_i)\varepsilon_j = n^{-1/2}\sum_{j=1}^{n}\varepsilon_j\mathrm{E}\{\mathbf{X}g'(U)|U_j\}+\mathrm{o_p}(1). \tag{A.40}$$

Combining (A.39) and (A.40), and multiplying by $P_\beta = I-\hat{\boldsymbol{\beta}}_1\hat{\boldsymbol{\beta}}_1^T = I-\boldsymbol{\beta}_{01}\boldsymbol{\beta}_{01}^T+\mathrm{o_p}(1)$, we obtain

$$P_\beta n^{1/2}\{(W+\Sigma)(\hat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01})+\mathbf{b}\} = n^{-1/2}\sum_{i=1}^{n}P_\beta\{\mathrm{E}\{\mathbf{X}_ig'(U)|U_i\}-\mathbf{X}_ig'(U_i)\}\varepsilon_i. \tag{A.41}$$

Since $\mathrm{Var}(\varepsilon_i)=\sigma^2$, the covariance matrix of the right-hand side of (A.41) is $P_\beta W_0P_\beta$, and this completes the proof of Theorem 3.2. □

References

Bai, Z.D., Rao, C.R., Wu, Y., 1999. Model selection with data-oriented penalty. Journal of Statistical Planning and Inference 77, 103–117.

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

Cook, D.R., Weisberg, S., 1991. Sliced inverse regression for dimension reduction: comment. Journal of the American Statistical Association 86, 328–332.

Duan, N., Li, K.-C., 1991. Slicing regression: a link-free regression method. Annals of Statistics 19, 505–530.

Fan, J., 1997. Comments on 'Wavelets in Statistics: A Review' by A. Antoniadis. Journal of the Italian Statistical Association 6, 131–138.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, London.

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

Fan, J., Li, R., 2002. Variable selection for Cox's proportional hazards model and frailty model. Annals of Statistics 30, 74–99.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Annals of Statistics 32, 928–961.

Fan, J., Yao, Q., 2003. Nonlinear Time Series. Springer.

Härdle, W., Hall, P., Ichimura, H., 1993. Optimal smoothing in single-index models. Annals of Statistics 21, 157–178.

Härdle, W., Tsybakov, A.B., 1993. How sensitive are average derivatives. Journal Econometrics 58, 31–48.

Hristache, M., Juditsky, A., Spokoiny, V., 2001. Direct estimation of the index coefficient in a single-index model. Annals of Statistics 29, 595–623.

Horowitz, J.L., Härdle, W., 1996. Direct semiparametric estimation of single-index models with discrete covariates. Journal of the American Statistical Association 91, 1632–1640.

Hunter, D., Li, R., 2005. Variable selection using MM algorithms. Annals of Statistics 33, 1617–1642.

Ichimura, H., 1993. Semiparametric least square (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics 58, 71–120.

Katch, F., McArdle, W., 1977. Nutrition, Weight Control, and Exercise. Houghton Mifflin Co., Boston.

Kong, E., Xia, Y., 2007. Variable selection for the single-index model. Biometrika 94, 217–229.

Li, R., Liang, H., 2008. Variable selection in semiparametric regression modeling. Annals of Statistics 36, 261–286.

Naik, P.A., Tsai, C.-L., 2001. Single-index model selections. Biometrika 88, 821–832.

Penrose, K.W., Nelson, A.G., Fisher, A.G., 1985. Generalized body composition prediction equation for men using simple measurement techniques. Medicine and Science in Sports and Exercise 17, 189.

Powell, J.L., Stock, J.M., Stoker, T.M., 1989. Semiparametric estimation of index coefficients. Econometrica 57, 1403–1430.

Quarteroni, A., Sacco, R., Saleri, F., 2000. Numerical Mathematics. Springer-Verlag, New York.

Ruppert, D., Sheather, S.J., Wand, M.P., 1995. An effective bandwidth selector for local least square regression. Journal of the American Statistical Association 90, 1257–1270.

Siri, W.E., 1956. Gross composition of the body. Advances in Biological and Medical Physics, IV. Academic Press, Inc., New York.

Stoker, T.M., 1986. Consistent estimation of scale coefficients. Econometrica 54, 1461–1481.

Tibshirani, R.J., 1996. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, Series B 58, 267–288.

van de Geer, S.A., 2000. Empirical Processes in M-estimation. Cambridge University Press.

Wang, H., Li, R., Tsai, C.-L., 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94, 553–568.

Zhu, L.P., Zhu, L.X., 2009. Nonconcave penalized inverse regression in single-index models with high dimensional predictors. Journal of Multivariate Analysis 100, 862–875.

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101, 1418–1429.

Zou, H., Li, R., 2008. One-step sparse estimates in nonconcave penalized likelihood models. Annals of Statistics 36, 1509–1533.