Robust variable selection in partially varying coefficient single-index model

Huiming Zhu[a], Zhike Lv[a,*], Keming Yu[b], Chao Deng[a]

[a] College of Business Administration, Hunan University, Changsha, 410082, PR China

[b] Department of Mathematical Sciences, Brunel University, London UB8 3PH, UK

Abstract

By combining basis function approximations and smoothly clipped absolute deviation (SCAD) penalty, this paper proposes a robust variable selection procedure for partially varying coefficient single-index model based on modal regression. The proposed procedure simultaneously selects significant variables in the parametric components and the nonparametric components. With appropriate selection of the tuning parameters, we establish the theoretical properties of our procedure, including consistency in variable selection and the oracle property in estimation. Furthermore, we also discuss the bandwidth selection and propose a modified expectation-maximisation (EM)-type algorithm for the proposed estimation procedure. The finite sample properties of the proposed estimators are illustrated by some simulation examples.

Keywords: Variable selection; Spline approximation; Modal regression; SCAD; Oracle property

1. Introduction

Partially varying coefficient single-index model (PVCSIM) combines naturally the advantages of both the single-index models and the varying coefficient models. Ever since Wong, Ip, and Zhang (2008) proposed the PVCSIM, studies in this class of model have raised the great interest of research in Statistics field. We formulate a PVCSIM as

$$Y = Z^T\theta(U) + g(X^T\beta) + \varepsilon, \tag{1.1}$$

where Y is a response variable, X and Z are of dimensions $p \times 1$ vectors and $q \times 1$ vectors, $\theta(\cdot) = (\theta_1(\cdot), ..., \theta_q(\cdot))^T$ is a vector of unknown function, $\beta = (\beta_1, ..., \beta_p)^T$ is a vector of unknown parameters, $g(\cdot)$ is an unknown link function, and ε is random error with mean zero. Due to the curse of dimensionality, we assume, for simplicity, that U is univariate. And we also assume that $\|\beta\| = 1$ and $\text{sigh}(\beta_1) = 1$ to ensure identifiability, where $\|\cdot\|$ denotes the Euclidean metric.

Model (1.1) is quite flexible enough to cover a variety of existing statistical models. For example, if $g(\cdot) = 0$, it reduces to the standard varying-coefficient model. When $\theta(\cdot)$ is an unknown constant

*Corresponding author. Tel.: +86 731 88684461. *E-mail address*: lzk2011@hnu.edu.cn(Z.Lv).

parameter, then model (1.1) is a partially linear single-index model (Wang, Xue & Chong, 2010; Wang & Wu, 2013). In addition, model (1.1) becomes the standard single-index model when $Z = 0$ or $\theta(\cdot) = 0$ (Wu, Yu & Yu, 2010). Due to its flexibility and generality, model (1.1) has gained much attention in recent years. Wang and Xue (2011) developed a stepwise approach to obtain asymptotic normality estimators of the varying-coefficient vector and the parametric vector. Huang and Zhang (2010) constructed the confidence region for the parameter β in model (1.1) based on the empirical likelihood technique. While Huang (2011) used the empirical likelihood method to study the confidence regions of the varying-coefficient parts. Huang, Lin, Feng and Pang (2013) proposed a class of efficient penalized estimating equations to estimate the index parametric components in the PVCSIM. Feng and Xue (2013) also considered the problem of variable selection in the PVCSIM. However, the aforementioned existing researches were mainly built on either the least-square or empirical likelihood method, which are expected to be very sensitive to the outliers and its efficiency may be significantly decreased for many commonly used non-normal errors.

Recently, Yao, Lindsay and Li (2012) proposed a new estimation approach based on a local modal regression for the nonparametric model. Then, Zhang, Zhao and Liu (2013) and Zhao, Zhang, Liu and Lv (2014) investigated the partially linear varying coefficient model based on modal regression, respectively. And Liu, Zhang, Zhao and Lv (2013) developed a new robust and efficient estimation procedure based on local modal regression for single index models. A distinguishing characteristic of their method is that it introduces an additional tuning parameter which is automatically selected using the observed data to achieve both robustness and efficiency of the resulting estimate. Namely, their method is not only robust when there are outliers or the error distribution is heavy-tail, but as asymptotically efficient as the ordinary least-square-based estimator when the data include no outliers and the error distribution is a Gaussian distribution. Due to its nice property, it has attracted increasing attention. Here, we extend the modal regression approach to the model (1.1).

Variable selection is a crucial issue in regression analysis. In practice, a number of variables are available for inclusion in an initial analysis, but many of them may not be significant and should be excluded from the final model to increase the accuracy of prediction. Traditional variable selection methods such as stepwise regression and best subset selection are computationally infeasible when the number of predictors is large. Therefore, various shrinkage methods such as the LASSO (Tibshirani 1996), the adaptive LASSO (Zou 2006) and the SCAD (Fan & Li 2001) have gained much attention in recent years. However, the LASSO is known to be near mini-max optimal as well as consistent under certain regularity conditions, Zou (2006) showed that it falls short of attaining the oracle property. By this property, an estimator estimates a zero coefficient exactly as zero with probability approaching one, while still being asymptotically normal for the non-zero coefficients in large samples. In this respect, the LASSO is inferior to the SCAD estimator which possesses the oracle property. So in the present paper, we prefer the SCAD of Fan and Li (2001) since it simultaneously satisfies the mathematical conditions for unbiasedness, sparsity, and continuity. More detail can be found in Fan and Li (2001). Since the SCAD was proposed, there has been a large number of literature focused on its applications in many important nonparametric and semiparametric models.

In this paper, we investigated the variable selection for the varying coefficient function $\theta(\cdot)$ and the unknown parametric index β in model (1.1) based on modal regression. By combining the basis function approximate and the SCAD penalty, we develop a variable selection procedure for PVCSIM. More specifically, we first use the B-spline functions to approximate the unknown coefficient functions and link function in model (1.1). And then combine with the restraint $\|\beta\| = 1$ to construct the penalized estimation function for PVCSIM based on modal regression. Under certain regularity conditions, we are able to establish this variable selection procedure is consistent, and the estimators have oracle property. Moreover, a modified version of a modal expectation-maximisation (MEM) algorithm is proposed to obtain the solutions for the object function. Some simulation studies show that, when data is contaminated by outliers, the proposed variable selection procedure can perform well in finite samples.

The layout of the remainder of the paper is as follows. In Section 2, following the idea of the modal regression approach, we propose the regularized estimation produce using basis expansion and the SCAD penalty function. Then, Under some regularity conditions, we establish some theoretical properties of the proposed variable selection procedure. We describe the detail of bandwidth selection and propose a modified MEM algorithm. In addition, we give the method of choosing the tuning parameters in Section 3. In Section 4, we conduct some simulation studies to examine the finite sample performance of the proposed procedures. Finally, in section 5 we conclude the paper. All the regularity conditions and the technical proofs are relegated to Appendix.

2. Estimation and variable selection procedure

As a measure of centre, the median and mode have the common advantage of robustness, when there exist outliers. Furthermore, since the modal regression focuses on the relationship between the majority data points and summaries the "most likely"conditional values, it can provide more meaningful point prediction than the mean regression when the error density is skewed. Suppose that $\{(Y_i, X_i, Z_i, U_i), i = 1, ..., n\}$ is an i.i.d. sample from model (1.1). Then following the method of Yao, Lindsay and Li (2012), the robust modal estimate of the PVCSIM is to maximize

$$\frac{1}{n}\sum_{i=1}^{n}\phi_h\Big(Y_i - Z_i^T\theta(U_i) - g(X_i^T\beta)\Big), \tag{2.1}$$

where $\phi_h(t) = h^{-1}\phi(t/h)$, $\phi(t)$ is a kernel density function, and the choice of $\phi(\cdot)$ is not very important. h is a bandwidth. For ease of computation, we use the standard normal density for $\phi(t)$ throughout the present article.

Remark 1. The choice of kernel is not very important because it is possible to obtain estimators with somewhat improved asymptotic properties by using different kernels (see, e.g., Eddy, 1980; Romano, 1988). For the simplicity of the calculation, we use the Gaussian density for $\phi(t)$.

2.1. Spline-based estimation

Since $\theta(\cdot)$ and $g(\cdot)$ are unknown functions in (2.1), here, we use polynomial splines to approximate it. More specifically, let $B_1(u) = (B_{11}(u), ..., B_{1L_1}(u))^T$ and $B_2(t) = (B_{21}(t), ..., B_{1L_2}(t))^T$ be the B-

spline basis functions with the order of $M_1 + 1$ and $M_2 + 1$, respectively, where $L_1 = K_1 + M_1 + 1$ and $L_2 = K_2 + M_2 + 1$, with K_1 and K_2 are the number of interior knots, Then $\theta_j(u)$ and $g(t)$ can be approximated by

$$\theta_j(u) \approx B_1(u)^T \gamma_j, \ j = 1, ..., q. \quad \text{and} \quad g(t) \approx B_2(t)^T \eta.$$

Then, we can obtain $\hat{\gamma}$, $\hat{\eta}$ and $\hat{\beta}$ by maximizing

$$\frac{1}{n} \sum_{i=1}^{n} \phi_h \Big(Y_i - W_i^T \gamma - B_2(X_i^T \beta)^T \eta \Big), \tag{2.2}$$

where $W_i = I_q \otimes B_1(U_i) \cdot Z_i$ with I_q is a $q \times q$ identity matrix, and $\gamma = (\gamma_1^T, ..., \gamma_q^T)^T$.

2.2. Variable selection for PVCSIM

In this subsection, our main goal is to find zero components (i.e., $\theta_j(\cdot) = 0$ and $\beta_s = 0$) in PVCSIM. Thus, we define the following semiparametric penalized estimation for PVCSIM based on modal regression as

$$\mathcal{L}(\gamma, \eta, \beta) \equiv \frac{1}{n} \sum_{i=1}^{n} \phi_h \Big(Y_i - W_i^T \gamma - B_2(X_i^T \beta)^T \eta \Big) - \sum_{j=1}^{q} p_{\lambda_{1j}}(\|\gamma_j\|_H) - \sum_{s=1}^{p} p_{\lambda_{2s}}(|\beta_s|), \tag{2.3}$$

where $p_{\lambda_{1j}}$ and $p_{\lambda_{2s}}$ are two penalized parameters for the jth varying coefficient function and the sth parameter component, respectively. $\|\gamma_j\|_H = (\gamma_j^T H \gamma_j)^{1/2}$ with $H = \int B_1(u) B_1(u)^T \mathrm{d}u$.

Remark 2. Formulation (2.3) includes many popular variable selection methods, for example, the Lasso (Tibshirani 1996) uses the L_1 penalty with $p_{\lambda_1}(\|\cdot\|) = \lambda_1 \|\cdot\|$. Bridge regression (Frank & Friedman 1993) uses the L_q penalty with $p_{\lambda_1}(\|\cdot\|) = \lambda_1 \|\cdot\|^q$. When $0 < q < 1$ the L_q penalty is concave over $(0, \infty)$ and nondifferentiable at zero. Fan and Li (2001) proposed the use of the SCAD penalty defined by its first derivative as

$$p'_\lambda(x) = \lambda \Big\{ I(x \le \lambda) + \frac{(a\lambda - x)_+}{(a - \lambda)} I(x > \lambda) \Big\},$$

for some $a > 2$. The SCAD penalty is a spline function on an interval near zero and constant outside, so that it can shrink small value of an estimate to zero while having no impact on a large one. As illustrated in Fan and Li (2001), this penalty function satisfies three requirements for variable selection, namely, asymptotic unbiasedness, sparsity and continuity of the estimated parameters. Therefore, we only focus on the SCAD penalty function throughout of this paper.

Recalling that we assume that $\|\beta\| = 1$, this means that true value of β is a boundary point on the unit sphere, which causes some difficulty since $g(X_i^T \beta)$ does not have a derivative at the point β. To solve this problem, we suggest the popularly used "delete-one-component" method proposed by Yu and Rupper (2002). The detail is as follows, without loss of generality, we assume that the true parameter β has a positive component β_1 (otherwise, consider -β_1). For $\beta = (\beta_1, ..., \beta_p)^T$, let $\beta^{(1)} = (\beta_2, ..., \beta_p)^T$ be a $(p-1)$-dimensional parameter vector after deleting the 1th component β_1 in β. Then, we can rewrite

$$\beta = \beta(\beta^{(1)}) = \big((1 - \|\beta^{(1)}\|^2)^{1/2}, \beta_2, ..., \beta_p \big)^T.$$

The true parameter $\beta^{(1)}$ satisfies the constraint $\|\beta^{(1)}\| < 1$. Thus, β is infinitely differential in a neighborhood of $\beta^{(1)}$. Let $\widetilde{\alpha} = (\gamma^T, \eta^T, \beta^T)^T$ and $\alpha = (\gamma^T, \eta^T, \beta^{(1)T})^T$, and the Jacobian matrix is

$$J_{\beta^{(1)}} = \partial\beta/\partial\beta^{(1)} = (\kappa_1, ..., \kappa_p)^T, \tag{2.4}$$

where κ_s $(1 < s \leq p)$ is a $(p-1)$-dimensional unit vector with sth component 1, and $\kappa_1 = -(1 - \|\beta^{(1)}\|^2)^{1/2}\beta^{(1)}$.

By this reparametrization, and note that α is one dimension lower that $\widetilde{\alpha}$, so the objective function (2.3) is transformed to

$$\mathcal{L}(\alpha) \equiv \frac{1}{n}\sum_{i=1}^{n}\phi_h\Big(Y_i - W_i^T\gamma - B_2(X_i^T\beta(\beta^{(1)}))^T\eta\Big) - \sum_{j=1}^{q}p_{\lambda_{1j}}(\|\gamma_j\|_H) - \sum_{s=1}^{p-1}p_{\lambda_{2s}}(|\beta_s^{(1)}|), \tag{2.5}$$

In order to facilitate use, we note that maximizing the objective function (2.5) is equivalent to minimizing

$$\widetilde{\mathcal{L}}(\alpha) \equiv -\frac{1}{n}\sum_{i=1}^{n}\phi_h\Big(Y_i - W_i^T\gamma - B_2(X_i^T\beta(\beta^{(1)}))^T\eta\Big) + \sum_{j=1}^{q}p_{\lambda_{1j}}(\|\gamma_j\|_H) + \sum_{s=1}^{p-1}p_{\lambda_{2s}}(|\beta_s^{(1)}|), \tag{2.6}$$

Let $\hat{\gamma}$, $\hat{\eta}$ and $\hat{\beta}^{(1)}$ be the solution by minimizing (2.6). Thus, the estimators of β, $\theta_j(u)$ and $g(t)$ can be obtained by

$$\hat{\beta} = (\sqrt{1 - \|\hat{\beta}^{(1)}\|^2}, \hat{\beta}^{(1)T})^T, \quad \hat{g}(t) = B_2(t)^T\hat{\eta} \text{ and } \hat{\theta}_j(u) = B_1(u)^T\hat{\gamma}_j, \quad j = 1, ..., q. \tag{2.7}$$

Next, we study the theoretical property of the proposed penalized estimators. We first introduce some notations. Let β_0 and $\theta_0(\cdot)$ be the true values of β and $\theta(\cdot)$, respectively. Without loss of generality, we assume that $\beta_{0s} = 0$ for $s = d_2 + 1, ..., p$, and $\beta_{0s} \neq 0$ for $s = 1, ..., d_2$. Similarly, we assume that $\theta_{0j}(\cdot) = 0$ for $j = d_1 + 1, ..., q$, and $\theta_{0j}(\cdot) \neq 0$ for $j = 1, ..., d_1$. In addition, $F(u, t, z, h) = E[\phi_h''(\varepsilon)|U = u, X^T\beta = t, Z = z]$ and $G(u, t, z, h) = E[\phi_h'(\varepsilon)^2|U = u, X^T\beta = t, Z = z]$. The following theorem gives the consistency of the proposed penalized estimators.

Theorem 1 *Suppose that the regularity conditions C1∼C9 in the Appendix hold and the number of knots $K = O_p(n^{1/(2r+1)})$, Then we have*

(a) $\|\hat{\beta} - \beta_0\| = O_p(K^{-r} + a_n)$,

(b) $\|\hat{\theta}_j(\cdot) - \theta_{0j}(\cdot)\| = O_p(K^{-r} + a_n), \quad j = 1, ..., q,$

where $a_n = \max_{j,s}\{|p'_{\lambda_{1j}}(|\beta_{0j}|)|, |p'_{\lambda_{2s}}(\|\gamma_{0s}\|_H)| : \beta_{0j} \neq 0, \gamma_{0s} \neq 0\}$, and r is defined in the Appendix.

Remark 3. Theorem 1 implies that, the rates of convergence of the proposed penalized estimators depend on λ_{1j}, λ_{2s} and K. So the rates of convergence of the proposed penalized estimators can be further improved to $\|\hat{\beta} - \beta_0\| = O_p(K^{-r})$ and $\|\hat{\theta}_j(\cdot) - \theta_{0j}(\cdot)\| = O_p(K^{-r})$ if $\lambda_{\max} \to 0$, then $a_n = 0$. Furthermore, under some conditions, we show that such consistent estimators must possess the sparsity property, which is stated as follows.

Theorem 2 *Suppose that the regularity conditions C1∼C9 given in the Appendix hold, and that $\lambda_{\max} \to 0$*

and $K^r\lambda_{\min} \to \infty$ *as* $n \to \infty$. *Then with probability tending to 1,* $\hat{\beta}$ *and* $\hat{\theta}_j(\cdot)$ *satisfy*

(a) $\hat{\beta}_s = 0, \quad s = d_2 + 1, ..., p,$

(b) $\hat{\theta}_j(\cdot) = 0, \quad j = d_1 + 1, ..., q,$

where $\lambda_{\max} = \max_{j,s}(\lambda_{1j}, \lambda_{2s})$ *and* $\lambda_{\min} = \min_{j,s}(\lambda_{1j}, \lambda_{2s})$.

Next, we show that the estimators for nonzero coefficients in the parametric components have the same asymptotic distribution as that based on the correct submodel. To demonstrate this, we require further notations in order to present the oracle properties of the resulting estimators. Define $\gamma_{0\mathcal{A}} = (\gamma_{01}^T, ..., \gamma_{0d_1}^T)^T$ and $\beta_{0\mathcal{A}} = (\beta_{01}, ..., \beta_{0d_2})^T$ to be true values of $\gamma_{\mathcal{A}}$ and $\beta_{\mathcal{A}}$, respectively. Corresponding covariates are denoted by $W_{i\mathcal{A}}$ and $G_{i\mathcal{A}} = g'(X_{i\mathcal{A}}^T\beta(\beta_{0\mathcal{A}}^{(1)}))J_{\beta_{0\mathcal{A}}^{(1)}}^T X_{i\mathcal{A}}$. In addition, let $\Sigma = \mathrm{E}(G(U,T,Z,h)\widetilde{G}_{\mathcal{A}}\widetilde{G}_{\mathcal{A}}^T)$ and $\Sigma_1 = \mathrm{E}(F(U,T,Z,h)\widetilde{G}_{\mathcal{A}}\widetilde{G}_{\mathcal{A}}^T)$ with $\widetilde{G}_{\mathcal{A}} = G_{\mathcal{A}} - \Phi^T\Lambda^{-1}W_{\mathcal{A}}$, where $\Phi = \mathrm{E}(F(U,T,Z,h)W_{\mathcal{A}}G_{\mathcal{A}}^T)$ and $\Lambda = \mathrm{E}(F(U,T,Z,h)W_{\mathcal{A}}W_{\mathcal{A}}^T)$.

Theorem 3 *Suppose that the regularity conditions C1∼C9 in the Appendix hold,*

$$\sqrt{n}(\hat{\beta}_{\mathcal{A}} - \beta_{0\mathcal{A}}) \xrightarrow{D} N(0, J_{\beta_{0\mathcal{A}}^{(1)}}\Sigma_1^{-1}\Sigma\Sigma_1^{-1}J_{\beta_{0\mathcal{A}}^{(1)}}^T), \tag{2.8}$$

where "$\xrightarrow{D}$*" represents the convergence in distribution.*

3. Bandwidth selection and estimation algorithm

The purpose of this section is twofold. we will first discuss the selection of bandwidth in theory. Then, we will discuss a modified MEM algorithm for the PVCSIM.

3.1. Optimal bandwidth

For the sake of simplicity, we assume that the error variable independent of U, X and Z, then based on (2.8) and the asymptotic variance of the least-square B-spline estimator(LSBS) given in Feng and Xue (2013), we can show that the ratio of the asymptotic variance of the modal regression (MR) estimator to that of the LSBS estimator is given by

$$R(h) \triangleq \frac{G(h)F^{-2}(h)}{\sigma^2}, \tag{3.1}$$

where $G(h) = E[\phi_h'(\varepsilon)]^2$, $F(h) = E[\phi_h''(\varepsilon)]^2$ and $\sigma^2 = E(\varepsilon^2)$. Obviously, the ratio $R(h)$ only depends on h, and it plays a key role in efficiency and robustness of estimators. So the ideal choice of h is

$$h_{opt} = \arg\min_h R(h) = \arg\min_h G(h)F^{-2}(h). \tag{3.2}$$

From above, we can see that h_{opt} does not depend on n and only depends on the conditional error distribution of ε.

3.2. Algorithm

In this subsection, we extend the local quadratic algorithm(LQA, Fan & Li 2001) and the MEM algorithm (Li, Ray, & Lindsay, 2007) to maximize (2.5) in Section 2.

Since the SCAD penalty is irregular at the origin, maximizing (2.5) directly may be difficult. Here, we use an iterative algorithm based on the local quadratic approximation of the penalty function $p_\lambda(\cdot)$ as in Fan and Li (2001). More specifically, in a neighborhood of a given nonzero u_0, an approximation of the penalty function at value u_0 can be given by

$$p_\lambda(|u|) \approx p_\lambda(|u_0|) + \frac{1}{2}\frac{p'_\lambda(|u_0|)}{|u_0|}(u^2 - u_0^2).$$

Therefore, for the given initial $\phi_s^{(0)}$ ($\phi_s = \beta_s^{(1)}$) with $|\phi_s^{(0)}| > 0$ for $s = 1, ..., d_2 - 1$, and γ_j^0 with $\|\gamma_j^0\|_H > 0$ for $j = 1, ..., d_1 - 1$, we get

$$p_{\lambda_{1j}}(\|\gamma_j\|_H) \approx p_{\lambda_{1j}}(\|\gamma_j^{(0)}\|_H) + \frac{1}{2}\frac{p'_{\lambda_{1j}}(\|\gamma_j^{(0)}\|_H)}{\|\gamma_j^{(0)}\|_H}(\|\gamma_j\|_H^2 - \|\gamma_j^{(0)}\|_H^2),$$

$$p_{\lambda_{2s}}(|\phi_s|) \approx p_{\lambda_{2s}}(|\phi_s^{(0)}|) + \frac{1}{2}\frac{p'_{\lambda_{2s}}(|\phi_s^{(0)}|)}{|\phi_s^{(0)}|}(|\phi_s|^2 - |\phi_s^{(0)}|^2).$$

With the aid of LQA and MEM algorithm, we propose a modified MEM algorithm as follows.

Step 0. Obtain the initial estimator $\hat{\beta}^{(0)}$ by fitting the partially varying coefficient single-index model based on Huang et al.(2013), or the unpenalized estimator obtained by minimizing (2.6) with $p_\lambda(\cdot) = 0$. Normalize $\hat{\beta}^{(0)}$ satisfy $\|\hat{\beta}^{(0)}\| = 1$ and impose the constraint that its first element is positive for identifiability.

Step 1. For the index values $\{t_i = X_i^T\hat{\beta}^{(0)}, i = 1, ..., n\}$, obtain $(\hat{\gamma}, \hat{\eta})$ by maximizing

$$\mathcal{L}(\gamma, \eta, \lambda_{1j}) \equiv \frac{1}{n}\sum_{i=1}^{n}\phi_h\Big(Y_i - W_i^T\gamma - B_2(t_i)^T\eta\Big) - \sum_{j=1}^{q}p_{\lambda_{1j}}(\|\gamma_j\|_H). \tag{3.3}$$

In this step, we can use the modified MEM algorithm. First, let us define $\Gamma_i = (W_i^T, B_2^T(t_i))^T$ and $\Upsilon = (\gamma^T, \eta^T)^T$, and set $\Upsilon^{(0)}$ be the initial value and start with $m = 0$:

E-step: in this step, we update $\pi_1(j|\Upsilon^{(m)})$ by

$$\pi_1(j|\Upsilon^{(m)}) = \frac{\phi_h(Y_i - \Gamma_i^T\Upsilon^{(m)})}{\sum_{i=1}^{n}\phi_h(Y_i - \Gamma_i^T\Upsilon^{(m)})}, \quad j = 1, ..., n.$$

M-step: Then we update Υ obtain $\hat{\Upsilon}^{(m+1)}$

$$\begin{aligned}\hat{\Upsilon}^{(m+1)} &= \arg\max_{\Upsilon}\left(\sum_{i=1}^{n}\big\{\pi_1(j|\Upsilon^{(m)})\log\phi_h\big(Y_i - \Gamma_i^T\Upsilon\big)\big\} - \sum_{j=1}^{q}p_{\lambda_{1j}}(\|\gamma_j\|_H)\right)\\ &= (\Gamma^T W\Gamma + \Sigma_{\lambda_1})^{-1}\Gamma^T WY,\end{aligned}$$

where $\Gamma = (\Gamma_1, ..., \Gamma_n)^T$, $Y = (Y_1, ..., Y_n)^T$, W is an $n \times n$ diagonal matrix with diagonal elements $\pi_1(j|\Upsilon^{(m)})$s, and

$$\Sigma_{\lambda_1} = \text{diag}\left\{ \frac{p'_{\lambda_{11}}(\|\gamma_1^{(m)}\|_H)}{\|\gamma_1^{(m)}\|_H} H, ..., \frac{p'_{\lambda_{1q}}(\|\gamma_q^{(m)}\|_H)}{\|\gamma_q^{(m)}\|_H} H, 0, ..., 0 \right\}.$$

Step 2. Utilize the estimator $\hat{\gamma}$ and $\hat{\eta}$ obtained by Step 1, by maximizing

$$\mathcal{L}(\beta^{(1)}, \lambda_{2s}) \equiv \frac{1}{n} \sum_{i=1}^{n} \phi_h\Big(Y_i - W_i^T\hat{\gamma} - \hat{g}(X_i^T\beta(\hat{\phi})) - \hat{g}'(X_i^T\beta(\hat{\phi}))J_{\hat{\phi}}^T X_i(\phi - \hat{\phi})\Big) - \sum_{s=1}^{p-1} p_{\lambda_{2s}}(|\beta_s^{(1)}|). \quad (3.4)$$

Let $\phi^{(0)} = \hat{\phi}$ ($\phi = \beta^{(1)}$) be the initial value and start with $m = 0$:

E-step: in this step, we update $\pi_2(j|\phi^{(m)})$ by

$$\pi_2(j|\phi^{(m)}) = \frac{\phi_h\Big(Y_i - W_i^T\hat{\gamma} - \hat{g}(X_i^T\beta(\hat{\phi})) - \hat{g}'(X_i^T\beta(\hat{\phi}))J_{\hat{\phi}}^T X_i(\phi^{(m)} - \hat{\phi})\Big)}{\sum_{i=1}^n \phi_h\Big(Y_i - W_i^T\hat{\gamma} - \hat{g}(X_i^T\beta(\hat{\phi})) - \hat{g}'(X_i^T\beta(\hat{\phi}))J_{\hat{\phi}}^T X_i(\phi^{(m)} - \hat{\phi})\Big)}, \quad j = 1, ..., n.$$

M-step: Then we update ϕ obtain $\hat{\phi}^{(m+1)}$

$$\begin{aligned}\hat{\phi}^{(m+1)} &= \arg\max_{\phi}\Bigg(\sum_{i=1}^{n}\{\pi_2(j|\phi^{(m)})\log\phi_h\Big(Y_i - W_i^T\hat{\gamma} - \hat{g}(X_i^T\beta(\hat{\phi})) - \hat{g}'(X_i^T\beta(\hat{\phi}))J_{\hat{\phi}}^T X_i(\phi - \hat{\phi})\}\\ &\quad - \sum_{s=1}^{p-1} p_{\lambda_{2s}}(|\beta_s^{(1)}|)\Bigg)\\ &= (G^{*T}\widetilde{W}G^* + \Sigma_{\lambda_2})^{-1}(G^{*T}\widetilde{W}\widetilde{Y} + G^{*T}\widetilde{W}G^*\hat{\phi}),\end{aligned}$$

where $\widetilde{Y} = (Y_1 - W_1^T\hat{\gamma} - \hat{g}(X_1^T\beta(\hat{\phi})), ..., Y_n - W_n^T\hat{\gamma} - \hat{g}(X_n^T\beta(\hat{\phi})))^T$, $G^* = (G_1^*, ..., G_n^*)^T$ with $G_i^* = \hat{g}'(X_i^T\beta(\hat{\phi}))X_i^T J_{\hat{\phi}}$, $\widetilde{W}$ is an $n \times n$ diagonal matrix with diagonal elements $\pi_2(j|\phi^{(m)})$s, and

$$\Sigma_{\lambda_2} = \text{diag}\left\{ \frac{p'_{\lambda_{21}}(|\phi_1^{(m)}|)}{|\phi_1^{(m)}|}, ..., \frac{p'_{\lambda_{2p}}(|\phi_{p-1}^{(m)}|)}{|\phi_{p-1}^{(m)}|} \right\}.$$

Step 3. Iterate Step 1 and Step 2 until convergence, and denote the final estimators of ϕ, γ and η as $\hat{\phi}$, $\hat{\gamma}$ and $\hat{\eta}$, respectively, and then obtain $\hat{\beta}$ via the transformation.

To implement this method, the number of interior knots K_1 and K_2, and the tuning parameters a, λ_{1j}'s and λ_{2s}'s in the penalty functions should be chosen. Fan and Li (2001) showed that the choice of $a = 3.7$ performs well in a variety of situations, therefore, we choose $a = 3.7$ throughout this paper. Since it is difficult to choose too many tuning parameters and knots simultaneously, therefore, we borrow the idea of Zhao and Xue (2009), and simplify the tuning parameters as $\lambda_{1j} = \lambda_1/\|\hat{\gamma}_j^*\|_H$ and $\lambda_{2s} = \lambda_2/|\hat{\phi}_s^*|$ with $\hat{\gamma}_j^*$ and $\hat{\phi}_s^*$ are the unpenalized estimators of γ_j and ϕ_s, respectively. Moreover, we set the number of interior knots $K_1 = K_2 = K$ in our simulation studies. Then we choose the parameters λ and K by

maximizing the following cross-validation score

$$\mathrm{CV}(\lambda_1, K) = \sum_{i=1}^{n} \phi_h\Big(Y_i - W_i^T\hat{\gamma}_{[-i]} - B_2(X_i^T\beta(\hat{\beta}^{(1)}))^T\hat{\eta}_{[-i]}\Big), \tag{3.5}$$

$$\mathrm{CV}(\lambda_2) = \sum_{i=1}^{n} \phi_h\Big(Y_i - W_i^T\hat{\gamma} - B_2(X_i^T\beta(\hat{\beta}^{(1)}_{[-i]}))^T\hat{\eta}\Big), \tag{3.6}$$

where $\hat{\gamma}_{[-i]}$, $\hat{\eta}_{[-i]}$ and $\hat{\beta}^{(1)}_{[-i]}$ are the solutions based on Equations (3.3) and (3.4) after deleting the ith subject.

4. Simulation study

In this section, we first consider how to select the bandwidth h in practice, and then assess the performance of the proposed procedure by some simulation studies.

4.1. Bandwidth selection in practice

In this subsection, we present the details of bandwidth selection in our simulation studies. We first need to estimate $F(h)$ and $G(h)$ defined in Eq.(3.1) to obtain the optimal bandwidth h_{opt} based on Eq.(3.2). And $F(h)$ and $G(h)$ can be estimated by

$$\hat{F}(h) = \frac{1}{n}\sum_{i=1}^{n}\phi_h''(\hat{\varepsilon}) \quad \text{and} \quad \hat{G}(h) = \frac{1}{n}\sum_{i=1}^{n}[\phi_h'(\hat{\varepsilon})]^2, \tag{4.1}$$

where $\hat{\varepsilon} = Y_i - Z_i^T\hat{\theta}(U_i) - \hat{g}(X_i^T\hat{\beta})$ with $\hat{g}(\cdot)$, $\hat{\theta}(\cdot)$ and $\hat{\beta}$ are estimated based on the pilot estimates. Therefore, we can estimate $R(h)$ by $\hat{R}(h) = \hat{G}(h)/\hat{F}^2(h)\hat{\sigma}^2$, where $\hat{\sigma}$ is also estimated based on the pilot estimate. However, since there is no explicit solution for h, thus, according to Yao et al. (2012), we use grid search method. As pointed out by Yao et al. (2012) and Zhang et al. (2013), the possible grids points for h can be $h = 0.5\hat{\sigma} \times 1.02^j$, $j = 0, ..., k$ for some fixed k. (such as k=120). Therefore, by the grid search method, we can obtain the optimal bandwidth h_{opt} based on Eq.(3.2).

4.2. Simulation study

In this subsection, we conduct simulation studies to assess the finite-sample performance of the proposed procedures. We generate data from the following partially varying coefficient single-index model

$$Y_i = Z_i^T\theta(U_i) + 3sin(\pi X_i^T\beta) + \varepsilon_i, \tag{4.2}$$

where $\beta = (0.5, \sqrt{2}, 0.5, 0, 0, 0, 0, 0, 0)$, $\theta(u) = (\theta_1(u), \theta_2(u), 0, 0, 0, 0, 0)$ with $\theta_1(u) = sin(2\pi u)$ and $\theta_2(u) = 7(u-1)^2$. $U_i \sim$ uniform $(0, 1)$, $X_i = (X_{i1}, ..., X_{i9})^T$ are generated from uniform distribution on (0, 1) and $Z_i = (Z_{i1}, ..., Z_{i7})^T$ are generated from a standard normal distribution. We considered the following four different error distributions: Case (1): $\varepsilon_i \sim N(0, 1)$; Case (2): $\varepsilon_i \sim t(3)$; Case (3): $\varepsilon_i \sim Laplace(0, 2)$; Case (4): $\varepsilon_i \sim 0.95N(0, 1) + 0.05N(0, 10^2)$. In the following simulations, we use the cubic B-splines, and the sample size n is set to be 100, 200 and 400. These simulations are all replicated over 500 times.

In addition, the interior knots are taken equidistantly, the tuning parameter and the number of interior knots are obtained by (3.5) and (3.6) in Section 3, and we choose the optimal bandwidth h based on the introduction in subsection 4.1.

We use the generalized mean square error (GMSE), as defined in Li and Liang (2008)

$$\text{GMSE} = (\hat{\beta} - \beta)^T \text{E}(XX^T)(\hat{\beta} - \beta),$$

to assess the performance of variable selection procedures for the parametric component. And the performance of estimator $\hat{\theta}(\cdot)$ will be assessed by using the square root of average square errors (RASE)

$$\text{RASE} = \left\{ N_{\text{grid}}^{-1} \sum_{j=1}^{N_{\text{grid}}} \|\hat{\theta}(u_j) - \theta(u_j)\|^2 \right\}^{1/2},$$

where $u_j, j = 1, ..., N_{\text{grid}}$ are the regular grid points at which the function $\hat{\theta}(u)$ is evaluated. In our simulation, $N_{\text{grid}} = 201$ is used.

To examine the robustness of the proposed variable selection procedure, we compare the performance of the variable selection procedure based on modal regression (MR) proposed in this paper with that based on the least square B-spline (LSBS) estimator used in Feng and Xue (2013). The simulated results are reported in Table 1. The column labeled "C_β" in Table 1 gives the average number of zero coefficients correctly estimated to be zero for parametric β. Column "IC_β" presents the average number of nonzero coefficients incorrectly estimated to be zero. While "$C_{\theta(\cdot)}$" and "$IC_{\theta(\cdot)}$" present the average number of zero coefficients correctly estimated to be zero and the average number of nonzero coefficients incorrectly estimated to be zero for varying coefficient functions $\theta(u)$, respectively. Furthermore, Table 1 also presents the median of GMSE for the parametric components and the median of RASE for the nonparametric components.

From Table 1, we can make the following observations: (a) For given n, the penalised MR estimate performs better than the penalised LSBS estimator method especially for the non-normal error distribution; (b) For given error distribution, we can see that the variable selection method based on MR and LSBS both become better and better as n increases. In addition, the performance of the variable selection method based on MR becomes more and more closer to that based on the Oracle procedure as n increases; (c) For the error distribution is $0.95N(0, 1) + 0.05N(0, 10^2)$, we can see that the superiority of MR becomes more and more obvious as sample size n increases.

5. Conclusions

In this paper, we have proposed a robust variable selection procedure for PVCSIM based on modal regression. The main contributions of the present article can be summarized as follows: (a) our procedures are computationally efficient and theoretically reliable; (b) the variable selection procedure has the oracle property; (c) the estimators of the single-index parametric components, which are of primary interest, are still asymptotically normal.

There are several possible extensions that deserve further study. In our work, we are just concerned with the situation that the covariates are errors free, while it might be interesting to investigate the case where the covariates are subject to measurement errors. Furthermore, Our approach described in this paper can be easily extended to other models, such as partially linear single index model and partially linear additive model. On another direction, it would be interesting to consider the dimensions q go to infinity as $n \to \infty$, the variable selection procedure proposed by this paper will not work any more, for such high-dimensional problems, it is the subject of ongoing research.

Acknowledgements

The authors are deeply grateful to Editor-in-Chief Byeong U. Park, the Associate Editor and two anonymous reviewers for their constructive comments that greatly improved the article. Special thanks go to professor Yao W. (Kansas State University) for his help. The research of Zhu is partially supported by National Natural Science Foundation of China (NNSFC) under Grants 71171075, 71221001 and 71031004. The research of Yu is supported by NNSFC under Grant 11261048.

Appendix: Proofs

For convenience and simplicity, let C denote a positive constant that may be different at each appearance throughout this paper. For any two sequences $\{a_n, b_n, n = 1, 2, ...\}$, we write $a_n \asymp b_n$ if there are constants $0 < c_1 < c_2 < \infty$ such that $c_1 \leq a_n/b_n \leq c_2$ for all n sufficiently large. Before we prove our main theorems, we list some regularity conditions that are used in this paper.

C1. The variable U has a bounded support $\mathscr{U}([a_1, b_2])$ and its density function $f_U(\cdot)$ is positive and has a continuous second derivative.

C2. The function $\theta_j(u)$ is the $r(r > 2)$th continuously differentiable on $[0, 1]$, and the function $g(t)$ has bounded and continuous derivatives up to order r on $[a_2, b_2]$.

C3. Let the matrices $\mathrm{E}(ZZ^T|U = u)$ and $\mathrm{E}(ZX^T|U = u)$ be continuous with respect to u. Furthermore, for given u, $\mathrm{E}(ZZ^T|U = u)$ and $\mathrm{E}(ZX^T|U = u)$ are all positive definite matrix and their eigenvalues are bounded. In addition, we also assume that $\max_i \|X_i\|/\sqrt{n} = o_p(1)$ and $\max_i \|Z_i\|/\sqrt{n} = o_p(1)$.

C4. $K_1 \asymp K_2 \asymp K$.

C5. $F(u, t, z)$ and $G(u, t, z)$ are continuous with respect to (u, t, z). Furthermore, $F(u, t, z) < 0$ for any $h > 0$.

C6. $\mathrm{E}(\phi_h'(\varepsilon)|U = u, X^T\beta = t, Z = z) = 0$ and $\mathrm{E}(\phi_h''(\varepsilon)^2|U = u, X^T\beta = t, Z = z)$, $\mathrm{E}(\phi_h'(\varepsilon)^3|U = u, X^T\beta = t, Z = z)$ and $\mathrm{E}(\phi_h'''(\varepsilon)|U = u, X^T\beta = t, Z = z)$ are continuous with respect to t at the point t_0.

C7. Let $t_{j1}, ..., t_{jK_j}$ be the interior knots of $[a_j, b_j]$ for $j = 1, 2$. Moreover, let $t_{j0} = a_j$, $t_{j,K_j+1} = b_j$, $h_{ji} = c_{ji} - c_{j,i-1}$ and $h_j = \max_{1\leq i\leq K_{j+1}}\{h_{ji}\}$, Then, there exist a constant C_{0j} satisfy

$$\frac{h_j}{\min_{1\leq i\leq K_{j+1}}\{h_{ji}\}} < C_{0j}, \quad \max\{\|h_{j,i+1} - h_{ji}\|\} = o(K^{-1}).$$

C8. Let $b_n = \max_{j,s}\{|p''_{\lambda_{1j}}(\|\gamma_{0j}\|_H)|, |p''_{\lambda_{2s}}(|\beta_{0s}^{(1)}|)| : \gamma_{0j} \neq 0, \beta_{0s}^{(1)} \neq 0\}$, then $b_n \to 0$ as $n \to \infty$.
C9. $\liminf_{n\to\infty} \liminf_{\beta_s^{(1)}\to 0^+} \lambda_{2s}^{-1}|p'_{\lambda_{2s}}(|\beta_s^{(1)}|)| > 0$ and $\liminf_{n\to\infty} \liminf_{\|\gamma_j\|_H \to 0} \lambda_{1j}^{-1} p'_{\lambda_{1j}}(\|\gamma_j\|_H) > 0$, where $j = d_1+1, ..., q$, $s = d_2, ..., p$.

These assumptions, while look a bit lengthy, in fact, are quite mild and similar assumptions can be found in Zhao and Xue (2009), Zhao, Zhang, Liu and Lv (2013) and Feng and Xue (2013). The condition $E(\phi'_h(\varepsilon)|u,t,z) = 0$ ensures that the proposed estimate is consistent and it is satisfied if the error density is symmetric about zero. More detail can be found in Yao, Lindsay and Li (2012).

Proof of Theorem 1. Let $\delta = K^{-r} + a_n$ and $v = (v_1^T, v_2^T, v_3^T)^T$. Define $\beta^{(1)} = \beta_0^{(1)} + \delta v_1$, $\eta = \eta_0 + \delta v_2$ and $\gamma = \gamma_0 + \delta v_3$. Let us first show that, for any given $\xi > 0$, there exists a large C such that

$$P\Big\{\inf_{\|v\|=C} \widetilde{\mathfrak{L}}(\alpha_0 + \delta v) > \widetilde{\mathfrak{L}}(\alpha_0)\Big\} \geq 1 - \xi. \tag{A.1}$$

This implies that, with probability at least $1-\xi$, there exists a local minimizer in the ball $\{\alpha_0 + \delta v : \|v\| \leq C\}$. Using the Taylor expansion, it follows that

$$\begin{aligned}
D_n(v) &\equiv nK^{-1}\{\widetilde{\mathfrak{L}}(\alpha_0 + \delta v) - \widetilde{\mathfrak{L}}(\alpha_0)\} \\
&= K^{-1}\sum_{i=1}^{n}\Big\{-\phi_h\Big(Y_i - W_i^T\gamma - B_2(X_i^T\beta(\beta^{(1)}))^T\eta\Big) + \phi_h\Big(Y_i - W_i^T\gamma_0 - B_2(X_i^T\beta(\beta_0^{(1)}))^T\eta_0\Big)\Big\} \\
&\quad + nK^{-1}\sum_{j=1}^{q}\{p_{\lambda_{1j}}(\|\gamma_j\|_H) - p_{\lambda_{1j}}(\|\gamma_{0j}\|_H)\} + nK^{-1}\sum_{s=1}^{p-1}\{p_{\lambda_{2s}}(|\beta_s^{(1)}|) - p_{\lambda_{2s}}(|\beta_{0s}^{(1)}|)\} \\
&\geq K^{-1}\sum_{i=1}^{n}\Big\{-\phi_h\Big(Y_i - W_i^T\gamma - B_2(X_i^T\beta(\beta^{(1)}))^T\eta\Big) + \phi_h\Big(Y_i - W_i^T\gamma_0 - B_2(X_i^T\beta(\beta_0^{(1)}))^T\eta_0\Big)\Big\} \\
&\quad + nK^{-1}\sum_{j=1}^{d_1}\{p_{\lambda_{1j}}(\|\gamma_j\|_H) - p_{\lambda_{1j}}(\|\gamma_{0j}\|_H)\} + nK^{-1}\sum_{s=1}^{d_2-1}\{p_{\lambda_{2s}}(|\beta_s^{(1)}|) - p_{\lambda_{2s}}(|\beta_{0s}^{(1)}|)\} \\
&= \frac{\delta}{K}\sum_{i=1}^{n}\phi'_h\Big(\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))\Big)Q_i - \frac{\delta^2}{2K}\sum_{i=1}^{n}\phi''_h\Big(\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))\Big)Q_i^2 \\
&\quad + \frac{\delta^3}{6K}\sum_{i=1}^{n}\phi'''_h(\varsigma_i)Q_i^3 + \frac{n}{K}\sum_{s=1}^{d_2-1}\{p_{\lambda_{2s}}(|\beta_s^{(1)}|) - p_{\lambda_{2s}}(|\beta_{0s}^{(1)}|)\} + \frac{n}{K}\sum_{j=1}^{d_1}\{p_{\lambda_{1j}}(\|\gamma_j\|_H) - p_{\lambda_{1j}}(\|\gamma_{0j}\|_H)\} \\
&\equiv: I_1 + I_2 + I_3 + I_4 + I_5,
\end{aligned} \tag{A.2}$$

where ς_i is between $\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))$ and $\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)})) - \delta Q_i$, $R_2(t) = g(t) - B_2(t)^T\eta$, $R_1(u) = (R_{11}(u), ..., R_{1q}(u))^T$ with $R_{1j}(u) = \theta_j(u) - B_1(u)^T\gamma_{0j}$, $j = 1, ..., q$ and $Q_i = W_i^T v_3 + B_2(X_i^T\beta(\beta_0^{(1)}))^T v_2 + B_2'(X_i^T\beta(\beta_0^{(1)}))^T\eta_0 v_1^T J_{\beta^{(1)}}^T X_i + \delta B_2'(X_i^T\beta(\beta_0^{(1)}))^T\eta_0 v_2 v_1^T J_{\beta^{(1)}}^T X_i$. Let us first consider I_1, using Taylor expansion, we obtain that

$$\begin{aligned}
\sum_{i=1}^{n}\phi'_h\Big(\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))\Big)Q_i = \sum_{i=1}^{n}\Big\{&\phi'_h(\varepsilon_i) + \phi''_h(\varepsilon_i)\big(Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))\big) \\
&+ \phi'''_h(\epsilon_i)\big(Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))\big)^2\Big\}Q_i,
\end{aligned}$$

where ϵ_i is between ε_i and $\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))$. By the condition C1, C2, C7 and Corollary 6.21 in Schumaker (1981), we have $\|R_{1j}(u)\| = O(K^{-r})$ and $\|R_2(t)\| = O(K^{-r})$, and $|g'(X_i^T\beta(\beta^{(1)})) - B_2'(X_i^T\beta(\beta^{(1)}))\eta_0| \le CK^{-r+1}$, then we can prove

$$\begin{aligned}\sum_{i=1}^n \phi_h''(\varepsilon_i)R_1(U_i)^T Z_i Q_i = &\sum_{i=1}^n \phi_h''(\varepsilon_i)R_1(U_i)^T Z_i\big\{B_2(X_i^T\beta(\beta_0^{(1)}))^T v_2 + \delta B_2'(X_i^T\beta(\beta_0^{(1)}))^T\eta_0 v_2 v_1^T J_{\beta^{(1)}}^T X_i \\ &+ \big[B_2'(X_i^T\beta(\beta_0^{(1)}))^T\eta_0 - g'(X_i^T\beta(\beta_0^{(1)}))\big]v_1^T J_{\beta^{(1)}}^T X_i + g'(X_i^T\beta(\beta_0^{(1)}))v_1^T J_{\beta^{(1)}}^T X_i \\ &+ W_i^T v_3\big\} \\ = &\, O_p(nK^{-r}\|v\|), \end{aligned} \tag{A.3}$$

using an argument similar to the above, we have $\sum_{i=1}^n \phi_h''(\varepsilon_i)R_2(X_i^T\beta(\beta_0^{(1)}))Q_i = O_p(nK^{-r}\|v\|)$, then invoking condition C6 and C7, and some simple calculations, we obtain

$$\sum_{i=1}^n \phi_h'\Big(\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))\Big)Q_i = O_p(nK^{-r}\|v\|).$$

Thus, we have $I_1 = O_p(n\delta^2 K^{-1}\|v\|)$. Similarly, we can prove that $I_2 = F(U,T,Z,h)O_p(n\delta^2K^{-1}\|v\|^2)$ and $I_3 = O_p(n\delta^3K^{-1}\|v\|^3)$. Hence, we can choose C large enough such that I_2 dominates both I_1 and I_3 uniformly in $\|v\| = C$ by noting that $F(U,T,Z,h) < 0$. Furthermore, invoking $p_\lambda(0) = 0$, and by the standard argument of the Taylor expansion, we have

$$\begin{aligned} I_4 &\le nK^{-1}\sum_{s=1}^{d_2-1}\big\{\delta p_{\lambda 2s}'(\|\beta_{0s}^{(1)}\|)\mathrm{sgn}(\beta_{0s}^{(1)})\|v_{1s}\| + \delta^2 p_{\lambda 2s}''(\|\beta_{0s}^{(1)}\|)\|v_{1s}\|^2(1+o_p(1))\big\} \\ &\le n\sqrt{d_2-1}K^{-1}\delta a_n\|v\| + nK^{-1}\delta b_n\|v\|^2. \end{aligned}$$

Then, it is easy to show that I_4 is dominated by I_2 uniformly in $\|v\| = C$. Using an argument similar to I_4, we can prove that I_5 is also dominated by I_2 uniformly in $\|v\| = C$. Therefore, by choosing a sufficiently large C, A.1 holds. Namely, there exists a local minimizers $\hat\beta$, $\hat\gamma$ and $\hat\eta$ such that

$$\|\hat\beta - \beta_0\| = O_p(\delta), \quad \|\hat\gamma - \gamma_0\| = O_p(\delta) \ \text{ and } \ \|\hat\eta - \eta_0\| = O_p(\delta), \tag{A.4}$$

which completes the proof of part (a).

Next, we prove part (b).

$$\begin{aligned}\|\hat\theta_j(u) - \theta_{0j}(u)\|^2 &= \int_{\mathscr{U}}\{B_1(u)^T\hat\gamma_j - B_1(u)^T\gamma_{0j} + R_{1j}(u)\}^2\mathrm{d}u \\ &\le 2\int_{\mathscr{U}}\{B_1(u)^T\hat\gamma_j - B_1(u)^T\gamma_{0j}\}^2\mathrm{d}u + 2\int_{\mathscr{U}} R_{1j}(u)^2\mathrm{d}u \\ &= 2(\hat\gamma_j - \gamma_{0j})^T H(\hat\gamma_j - \gamma_{0j}) + 2\int_{\mathscr{U}} R_{1j}(u)^2\mathrm{d}u, \end{aligned}$$

where $H = \int_{\mathscr{U}} B_1(u)B_1(u)^T\mathrm{d}u$. Then, invoking $\|H\| = O(1)$ and $\|\hat\gamma - \gamma_0\| = O_p(\delta)$, after a simple calculation yields

$$(\hat\gamma_j - \gamma_{0j})^T H(\hat\gamma_j - \gamma_{0j}) = O_p(K^{-2r} + a_n^2). \tag{A.5}$$

Furthermore, it is easy to obtain that

$$\int_{\mathscr{U}} R_{1j}(u)^2 \mathrm{d}u = O_p(K^{-2r}). \tag{A.6}$$

Therefore, combing A.5 and A.6, the proof of part (b) is completed.

Proof of Theorem 2. We only show part (a) as an illustration and part (b) is similar. From $\lambda_{max} \to 0$, it is easy to show that $a_n = 0$ for large n. Then by Theorem 1, it is sufficient to show that, for any $\beta_s^{(1)}$ which satisfies $\|\beta_s^{(1)} - \beta_{0s}^{(1)}\| = O_p(K^{-1})$ for $s = 1, ..., d_2 - 1$, and some given small $\zeta = CK^{-1}$, as $n \to \infty$, with probability approaching one, we have

$$\begin{cases} \dfrac{\partial \widetilde{\mathfrak{L}}(\alpha)}{\partial \beta_s^{(1)}} > 0, \quad 0 < \beta_s^{(1)} < \zeta, \ s = 1, ..., d_2 - 1, \\ \dfrac{\partial \widetilde{\mathfrak{L}}(\alpha)}{\partial \beta_s^{(1)}} < 0, \quad -\zeta < \beta_s^{(1)} < 0, \ s = 1, ..., d_2 - 1. \end{cases}$$

Note that

$$\begin{aligned} \frac{n\partial \widetilde{\mathfrak{L}}(\alpha)}{\partial \beta_s^{(1)}} &= \sum_{i=1}^n \phi_h'\Big(Y_i - W_i^T\gamma - B_2(X_i^T\beta(\beta^{(1)}))^T\eta\Big) B_2'(X_i^T\beta(\beta^{(1)}))^T\eta\Gamma_{\beta_s^{(1)}}^T X_i + np'_{\lambda_{2s}}(|\beta_s^{(1)}|)\mathrm{sgn}(\beta_s^{(1)}) \\ &= \sum_{i=1}^n \phi_h'\Big(\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)})) + \Omega_i\Big) B_2'(X_i^T\beta(\beta^{(1)}))^T\eta\Gamma_{\beta_s^{(1)}}^T X_i + np'_{\lambda_{2s}}(|\beta_s^{(1)}|)\mathrm{sgn}(\beta_s^{(1)}) \\ &= \sum_{i=1}^n \Bigg\{ \phi_h'\Big(\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))\Big) + \phi_h''\Big(\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))\Big)\Omega_i \\ &\quad + \phi_h'''(\omega_i)\Omega_i^2 \Bigg\} \Big([g'(X_i^T\beta(\beta^{(1)})) - B_2'(X_i^T\beta(\beta^{(1)}))^T\eta_0] - g'(X_i^T\beta(\beta^{(1)})) + B_2'(X_i^T\beta(\beta_0^{(1)}))^T(\eta_0 - \eta) \\ &\quad + [B_2'(X_i^T\beta(\beta_0^{(1)})) - B_2'(X_i^T\beta(\beta^{(1)}))]^T\eta\Big)\pi_{\beta_s^{(1)}}^T X_i + np'_{\lambda_{2s}}(|\beta_s^{(1)}|)\mathrm{sgn}(\beta_s^{(1)}), \end{aligned} \tag{A.7}$$

where $\pi_{\beta_s^{(1)}} = (-\beta_s^{(1)}/\sqrt{1 - \|\beta^{(1)}\|^2}, 0, ..., 0, 1, 0, ..., 0)^T$ is a $p \times 1$ vector with the $(s+1)$th component 1, ω_i is between $\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)}))$ and $\varepsilon_i + Z_i^T R_1(U_i) + R_2(X_i^T\beta(\beta_0^{(1)})) + \Omega_i$, and

$$\Omega_i = W_i^T(\gamma_0 - \gamma) + B_2(X_i^T\beta(\beta_0^{(1)}))(\eta_0 - \eta) + [B_2(X_i^T\beta(\beta_0^{(1)})) - B_2(X_i^T\beta(\beta^{(1)}))]^T\eta.$$

Using an argument similar to A.3, and invoking A.4, it is easy to show that

$$\frac{n\partial \widetilde{\mathfrak{L}}(\alpha)}{\partial \beta_s^{(1)}} = n\lambda_{2s}\big\{\lambda_{2s}^{-1}p'_{\lambda_{2s}}(|\beta_s^{(1)}|)\mathrm{sgn}(\beta_s^{(1)}) + O_p(\lambda_{2s}^{-1}K^{-r})\big\},$$

since by the condition C9, $\lim_{n\to\infty} \liminf_{\beta_s^{(1)} \to 0} \lambda_{2s}^{-1} p'_{\lambda_{2s}}(|\beta_s^{(1)}|) > 0$ and $\lambda_{2s}K^r \geq \lambda_{\min}K^r \to \infty$, thus, the sign of A.7 is completely determined by that of $\beta_s^{(1)}$. This completes the proof of part (a).

Proof of Theorem 3. By Theorems 1 and 2, we know that, as $n \to \infty$, with probability tending

to 1, $\widetilde{\mathfrak{L}}(\alpha)$ attains the minimal vale at $(\hat{\gamma}_{\mathcal{A}}^T, 0)^T$, $(\hat{\beta}_{\mathcal{A}}^{(1)T}, 0)^T$ and $\hat{\eta}$. Define $\hat{\alpha}_{\mathcal{A}} = \big((\hat{\gamma}_{\mathcal{A}}^T, 0)^T, \hat{\eta}, (\hat{\beta}_{\mathcal{A}}^{(1)T}, 0)^T\big)^T$,

$$\begin{cases} \dfrac{\partial \widetilde{\mathcal{L}}(\hat{\alpha}_{\mathcal{A}})}{\partial \beta_{\mathcal{A}}^{(1)}} = -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \phi_h'\Big(Y_i - W_{i\mathcal{A}}^T\hat{\gamma}_{\mathcal{A}} - \hat{g}(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)}))\Big)\hat{g}'(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)}))J_{\hat{\beta}_{\mathcal{A}}^{(1)}}^T X_{i\mathcal{A}} \\ \qquad\qquad + p_{\lambda_2}'(|\beta_{\mathcal{A}}^{(1)}|) * \mathrm{sgn}(\beta_{\mathcal{A}}^{(1)}) = 0, \\ \dfrac{\partial \widetilde{\mathcal{L}}(\hat{\alpha}_{\mathcal{A}})}{\partial \gamma_{\mathcal{A}}} = -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \phi_h'\Big(Y_i - W_{i\mathcal{A}}^T\hat{\gamma}_{\mathcal{A}} - \hat{g}(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)}))\Big)W_{i\mathcal{A}} + \Delta = 0, \end{cases}$$

where " $*$ " denotes the Hadamard product and the sth component of $p_{\lambda_2}'(|\beta_{\mathcal{A}}^{(1)}|)$ is $p_{\lambda_{2s}}'(|\beta_{s\mathcal{A}}^{(1)}|)$ for $s = 1, ..., d_2 - 1$, and

$$\Delta = \left(p_{\lambda_{11}}'(\|\hat{\gamma}_1\|_H)\frac{\hat{\gamma}_1^T H}{\|\hat{\gamma}_1\|_H}, ..., p_{\lambda_{1d_1}}'(\|\hat{\gamma}_{d_1}\|_H)\frac{\hat{\gamma}_{d_1}^T H}{\|\hat{\gamma}_{d_1}\|_H}\right)^T.$$

Using the Taylor expansion to $p_{\lambda_{2s}}'(|\hat{\beta}_s^{(1)}|)$, we get that

$$p_{\lambda_{2s}}'(|\hat{\beta}_s^{(1)}|) = p_{\lambda_{2s}}'(|\hat{\beta}_{0s}^{(1)}|) + \{p_{\lambda_{2s}}''(|\hat{\beta}_{0s}^{(1)}|) + o_p(1)\}(\hat{\beta}_s^{(1)} - \beta_{0s}^{(1)}).$$

By condition C8, we have $p_{\lambda_{2s}}''(|\hat{\beta}_{0s}^{(1)}|) = o_p(1)$, and note that $p_{\lambda_{2s}}'(|\hat{\beta}_{0s}^{(1)}|) = 0$ as $\lambda_{\max} \to 0$. Then, from Theorems 1$\sim$2, we have $p_{\lambda_{2s}}'(|\hat{\beta}_s^{(1)}|)\mathrm{sgn}(\hat{\beta}_s^{(1)}) = o_p(\hat{\beta}_s^{(1)} - \beta_{0s}^{(1)})$. Similarly, we can prove that $p_{\lambda_{1j}}'(\|\hat{\gamma}_j\|_H)(H\hat{\gamma}_j/\|\hat{\gamma}_j\|_H) = o_p(\hat{\gamma}_j - \gamma_{0j})$. Thus, invoking Lemma A.1 in Feng and Xue (2013), after a simple calculation yields

$$\begin{aligned} 0 = &-\frac{1}{n}\sum_{i=1}^{n}\phi_h'\Big(\varepsilon_i + Z_{i\mathcal{A}}^T R_{1\mathcal{A}}(U_i) + g(X_{i\mathcal{A}}^T\beta(\beta_{0\mathcal{A}}^{(1)})) - \hat{g}(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)})) - W_{i\mathcal{A}}^T(\hat{\gamma}_{\mathcal{A}} - \gamma_{0\mathcal{A}})\Big) \\ &g'(X_{i\mathcal{A}}^T\beta(\beta_{0\mathcal{A}}^{(1)}))J_{\beta_{0\mathcal{A}}^{(1)}}^T X_{i\mathcal{A}} + o_p(\hat{\beta}_{\mathcal{A}}^{(1)} - \beta_{0\mathcal{A}}^{(1)}) + o_p(n^{-1/2}), \qquad (A.8) \\ 0 = &-\frac{1}{n}\sum_{i=1}^{n}\phi_h'\Big(\varepsilon_i + Z_{i\mathcal{A}}^T R_{1\mathcal{A}}(U_i) + g(X_{i\mathcal{A}}^T\beta(\beta_{0\mathcal{A}}^{(1)})) - \hat{g}(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)})) - W_{i\mathcal{A}}^T(\hat{\gamma}_{\mathcal{A}} - \gamma_{0\mathcal{A}})\Big) \\ &W_{i\mathcal{A}} + o_p(\hat{\gamma}_{\mathcal{A}} - \gamma_{0\mathcal{A}}). \qquad (A.9) \end{aligned}$$

To make our mathematical formula short, let $\Pi_i = Z_{i\mathcal{A}}^T R_{1\mathcal{A}}(U_i) + g(X_{i\mathcal{A}}^T\beta(\beta_{0\mathcal{A}}^{(1)})) - \hat{g}(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)})) - W_{i\mathcal{A}}^T(\hat{\gamma}_{\mathcal{A}} - \gamma_{0\mathcal{A}})$ and $G_{i\mathcal{A}} = g'(X_{i\mathcal{A}}^T\beta(\beta_{0\mathcal{A}}^{(1)}))J_{\beta_{0\mathcal{A}}^{(1)}}^T X_{i\mathcal{A}}$. Applying the Taylor expansion to A.8 and A.9, we obtain that

$$0 = -\frac{1}{n}\sum_{i=1}^{n}\Big\{\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)\Pi_i + \frac{1}{2}\phi_h'''(\varepsilon_i^*)\Pi_i^2\Big\}G_{i\mathcal{A}} + o_p(\hat{\beta}_{\mathcal{A}}^{(1)} - \beta_{0\mathcal{A}}^{(1)}) + o_p(n^{-1/2}), \qquad (A.10)$$

$$0 = -\frac{1}{n}\sum_{i=1}^{n}\Big\{\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)\Pi_i + \frac{1}{2}\phi_h'''(\varepsilon_i^{**})\Pi_i^2\Big\}W_{i\mathcal{A}} + o_p(\hat{\gamma}_{\mathcal{A}} - \gamma_{0\mathcal{A}}), \qquad (A.11)$$

where both ε_i^* and ε_i^{**} lie between ε and $\varepsilon + \Pi_i$.

Define $\Lambda_n = n^{-1}\sum_{i=1}^n \phi_h''(\varepsilon_i)W_{i\mathcal{A}}W_{i\mathcal{A}}^T$, $\Xi_n = n^{-1}\sum_{i=1}^n \phi_h''(\varepsilon_i)W_{i\mathcal{A}}[g(X_{i\mathcal{A}}^T\beta(\beta_{0\mathcal{A}}^{(1)})) - \hat{g}(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)}))]$, $\Phi_n = n^{-1}\sum_{i=1}^n W_{i\mathcal{A}}[\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)Z_{i\mathcal{A}}^T R_{1\mathcal{A}}(U_i)]$ and $\Psi_n = n^{-1}\sum_{i=1}^n \phi_h''(\varepsilon_i)W_{i\mathcal{A}}G_{i\mathcal{A}}^T$. Then, for A.11, by

conditions C5∼C6, and Theorem 1∼2, we have

$$\hat{\gamma}_{\mathcal{A}} - \gamma_{0\mathcal{A}} = (\Lambda_n + o_p(1))^{-1}(\Phi_n + \Xi_n). \tag{A.12}$$

Note that

$$\begin{aligned} g(X_{i\mathcal{A}}^T\beta(\beta_{0\mathcal{A}}^{(1)})) - \hat{g}(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)})) &= g(X_{i\mathcal{A}}^T\beta(\beta_{0\mathcal{A}}^{(1)})) - g(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)})) + g(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)})) - \hat{g}(X_{i\mathcal{A}}^T\beta(\hat{\beta}_{\mathcal{A}}^{(1)})) \\ &= G_{i\mathcal{A}}^T(\beta_{0\mathcal{A}}^{(1)} - \hat{\beta}_{\mathcal{A}}^{(1)}) + o_p(\beta_{0\mathcal{A}}^{(1)} - \hat{\beta}_{\mathcal{A}}^{(1)}) + O_p(K^{-r}). \end{aligned} \tag{A.13}$$

Substituting A.12 into A.10, and a simple calculation yields

$$\begin{aligned} \Big\{n^{-1}\sum_{i=1}^n \phi_h''(\varepsilon_i)G_{i\mathcal{A}}\widetilde{G}_{i\mathcal{A}}^T + o_p(1)\Big\}\sqrt{n}(\hat{\beta}_{\mathcal{A}}^{(1)} - \beta_{0\mathcal{A}}^{(1)}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\Big\{G_{i\mathcal{A}}\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)G_{i\mathcal{A}}Z_{i\mathcal{A}}^T R_{1\mathcal{A}}(U_i) \\ &\quad + \phi_h''(\varepsilon_i)G_{i\mathcal{A}}W_{i\mathcal{A}}^T(\Lambda_n + o_p(1))^{-1}\Phi_n\Big\}. \end{aligned} \tag{A.14}$$

Note that $n^{-1}\sum_{i=1}^n \Psi_n^T\Lambda_n^{-1}W_{i\mathcal{A}}\widetilde{G}_{i\mathcal{A}} = 0$ with $\widetilde{G}_{i\mathcal{A}} = G_{i\mathcal{A}}^T - W_{i\mathcal{A}}^T\Lambda_n^{-1}\Psi_n$, and $n^{-1}\sum_{i=1}^n \Psi_n^T\Lambda_n^{-1}W_{i\mathcal{A}}(\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)Z_{i\mathcal{A}}^T R_{1\mathcal{A}}(U_i) - \phi_h''(\varepsilon_i)W_{i\mathcal{A}}^T\Lambda_n^{-1}\Phi_n) = 0$. Therefore, A.14 can be rewritten as follows

$$\begin{aligned} \Big\{n^{-1}\sum_{i=1}^n \phi_h''(\varepsilon_i)\widetilde{G}_{i\mathcal{A}}\widetilde{G}_{i\mathcal{A}}^T + o_p(1)\Big\}\sqrt{n}(\hat{\beta}_{\mathcal{A}}^{(1)} - \beta_{0\mathcal{A}}^{(1)}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\Big\{\widetilde{G}_{i\mathcal{A}}\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)\widetilde{G}_{i\mathcal{A}}Z_{i\mathcal{A}}^T R_{1\mathcal{A}}(U_i) \\ &\quad + \phi_h''(\varepsilon_i)\widetilde{G}_{i\mathcal{A}}W_{i\mathcal{A}}^T\Lambda_n^{-1}\Phi_n\Big\} + o_p(1). \\ &\equiv: L_1 + L_2 + L_3 + o_p(1). \end{aligned} \tag{A.15}$$

It is easy to show that $n^{-1}\sum_{i=1}^n \phi_h''(\varepsilon_i)\widetilde{G}_{i\mathcal{A}}W_{i\mathcal{A}}^T = 0$ and by the definition of $R_{1\mathcal{A}}(U_i)$, we can prove that $\frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_h''(\varepsilon_i)\widetilde{G}_{i\mathcal{A}}Z_{i\mathcal{A}}^T R_{1\mathcal{A}}(U_i) = o_p(1)$. Now let us deal with the first term L_1, By directly calculating its expectation and variance, we have $\mathrm{E}(L_1) = 0$ and $\mathrm{cov}(L_1) = \mathrm{E}(G(U,T,Z,h)\widetilde{G}_{i\mathcal{A}}\widetilde{G}_{i\mathcal{A}}^T)$, this follows easily by checking Linderbergs condition. In addition, by the law of large numbers, we have

$$n^{-1}\sum_{i=1}^n \phi_h''(\varepsilon_i)\widetilde{G}_{i\mathcal{A}}\widetilde{G}_{i\mathcal{A}}^T \xrightarrow{p} \Sigma_1, \tag{A.16}$$

and from the definition of $J_{\beta_{0\mathcal{A}}^{(1)}}$ of Eq.(2.4), it follows $(\hat{\beta}_{\mathcal{A}} - \beta_{0\mathcal{A}}) = J_{\beta_{0\mathcal{A}}^{(1)}}(\hat{\beta}_{\mathcal{A}}^{(1)} - \beta_{0\mathcal{A}}^{(1)}) + O_p(n^{-1})$. Thus, we have

$$\sqrt{n}(\hat{\beta}_{\mathcal{A}} - \beta_{0\mathcal{A}}) = J_{\beta_{0\mathcal{A}}^{(1)}}\Sigma_1^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \widetilde{G}_{i\mathcal{A}}\phi_h'(\varepsilon_i) + o_p(1).$$

Therefore, we have

$$\sqrt{n}(\hat{\beta}_{\mathcal{A}} - \beta_{0\mathcal{A}}) \xrightarrow{D} N(0, J_{\beta_{0\mathcal{A}}^{(1)}}\Sigma_1^{-1}\Sigma\Sigma_1^{-1}J_{\beta_{0\mathcal{A}}^{(1)}}^T), \tag{A.17}$$

by using the Slutsky theorem. Thus, the proof of Theorem 3 is completed.

References

[1] Eddy, W. F. (1980). Optimum kernel estimators of the mode. *The Annals of Statistics*, 870-882.

[2] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348-1360.

[3] Feng, S., & Xue, L. (2013). Variable selection for partially varying coefficient single-index model. *Journal of Applied Statistics*, 2637-2652.

[4] Frank, L. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109-135.

[5] Huang, Z. (2011). Empirical likelihood-based inference in varying-coefficient single-index models. *Journal of the Korean Statistical Society*, 40(2), 205-215.

[6] Huang, Z., Lin, B., Feng, F., & Pang, Z. (2013). Efficient penalized estimating method in the partially varying-coefficient single-index model. *Journal of Multivariate Analysis*, 114, 189-200.

[7] Huang, Z., & Zhang R. (2010). Empirical likelihood for the varying-coefficient single-index model. *The Canadian Journal of Statistics*, 38, 434-452.

[8] Li, J., Ray, S., & Lindsay, B. G. (2007). A Nonparametric Statistical Approach to Clustering via Mode Identification. *Journal of Machine Learning Research*, 8(8), 1687-1723.

[9] Liu, J., Zhang, R., Zhao, W., & Lv, Y. (2013). A robust and efficient estimation method for single index models. *Journal of Multivariate Analysis*, 122, 226-238.

[10] Li, R., & Liang, H. (2008). Variable selection in semiparametric regression modeling. *Annals of Statistics*, 36(1), 261-286.

[11] Romano, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *Annals of Statistics*, 16, 629-647.

[12] Schumaker, L. L. (1981). *Spline functions: basic theory*. New York: Wiley.

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

[14] Wang, J. L., Xue, L., Zhu, L., & Chong, Y. S. (2010). Estimation for a partial-linear single-index model. *The Annals of statistics*, 38(1), 246-274.

[15] Wang, Q., & Wu, R. (2013). Shrinkage estimation of partially linear single-index models. *Statistics and Probability Letters*. 83(10), 2324-2331.

[16] Wang, Q., & Xue, L. (2011). Statistical inference in partially-varying-coefficient single-index model. *Journal of Multivariate Analysis*, 102(1), 1-19.

[17] Wong, H., Ip, W. C.,& Zhang, R. (2008). Varying-coefficient single-index model. *Computational statistics and data analysis*, 52(3), 1458-1476.

[18] Wu, T. Z., Yu, K., & Yu, Y. (2010). Single-index quantile regression. *Journal of Multivariate Analysis*, 101(7), 1607-1621.

[19] Yao, W., Lindsay, B. G., & Li, R. (2012). Local modal regression. *Journal of nonparametric statistics*, 24(3), 647-663.

[20] Yu, Y., & Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460), 1042-1054.

[21] Zhang, R., Zhao, W., & Liu, J. (2013). Robust estimation and variable selection for semiparametric partially linear varying coefficient model based on modal regression. *Journal of Nonparametric Statistics*, 25(2), 523-544.

[22] Zhao, P., & Xue, L. (2009). Variable selection for semiparametric varying coefficient partially linear models. *Statistics and Probability Letters*, 79(20), 2148-2157.

[23] Zhao, W., Zhang, R., Liu, J., & Lv, Y. (2014). Robust and efficient variable selection for semiparametric partially linear varying coefficient model based on modal regression. *Annals of the Institute of Statistical Mathematics*, 66(1), 165-191.

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

Table 1 :Simulation results

Error distribution	Method	n	GMSE	RASE	No. of zeros			
					C_β	IC_β	$C_{\theta(\cdot)}$	$IC_{\theta(\cdot)}$
N(0, 1)	MR	100	0.054	0.314	5.543	0.006	4.401	0.013
	LSBS		0.039	0.231	5.623	0.003	4.663	0.009
	MR oracle		0.035	0.265	6	0	5	0
	MR	200	0.038	0.200	5.642	0	4.600	0
	LSBS		0.030	0.106	5.761	0	4.762	0.001
	MR oracle		0.010	0.020	6	0	5	0
	MR	400	0.017	0.093	5.818	0	4.850	0
	LSBS		0.017	0.083	5.877	0	4.894	0
	MR oracle		0.016	0.088	6	0	5	0
t(3)	MR	100	0.091	0.512	5.529	0	4.693	0.005
	LSBS		0.106	0.576	5.431	0	4.587	0.012
	MR oracle		0.062	0.487	6	0	5	0
	MR	200	0.074	0.328	5.634	0	4.801	0
	LSBS		0.081	0.351	5.578	0	4.799	0
	MR oracle		0.049	0.227	6	0	5	0
	MR	400	0.042	0.125	5.856	0	4.938	0
	LSBS		0.053	0.150	5.765	0	4.812	0
	MR oracle		0.044	0.112	6	0	5	0
Laplace(0, 2)	MR	100	0.049	0.287	5.675	0	4.432	0.007
	LSBS		0.068	0.308	5.613	0	4.321	0.011
	MR oracle		0.037	0.197	6	0	5	0
	MR	200	0.034	0.190	5.747	0	4.606	0
	LSBS		0.045	0.203	5.695	0	4.532	0
	MR oracle		0.029	0.138	6	0	5	0
	MR	400	0.021	0.094	5.875	0	4.794	0
	LSBS		0.032	0.130	5.806	0	4.685	0
	MR oracle		0.019	0.085	6	0	5	0
0.95N(0, 1)+0.05N(0, 10^2)	SMR	100	0.053	0.242	5.655	0	4.612	0
	LSBS		0.182	0.496	5.428	0.019	4.218	0.065
	MR oracle		0.045	0.180	6	0	5	0
	MR	200	0.035	0.157	5.805	0	4.781	0
	LSBS		0.106	0.303	5.503	0	4.332	0.008
	MR oracle		0.032	0.101	6	0	5	0
	MR	400	0.011	0.072	5.922	0	4.868	0
	LSBS		0.089	0.182	5.697	0	4.593	0
	MR oracle		0.010	0.061	6	0	5	0