Ann Inst Stat Math (2014) 66:165–191
DOI 10.1007/s10463-013-0410-4

# Robust and efficient variable selection for semiparametric partially linear varying coefficient model based on modal regression

**Weihua Zhao · Riquan Zhang · Jicai Liu · Yazhao Lv**

Received: 5 July 2012 / Revised: 20 March 2013 / Published online: 29 May 2013


**Abstract** Semiparametric partially linear varying coefficient models (SPLVCM) are frequently used in statistical modeling. With high-dimensional covariates both in parametric and nonparametric part for SPLVCM, sparse modeling is often considered in practice. In this paper, we propose a new estimation and variable selection procedure based on modal regression, where the nonparametric functions are approximated by $B$-spline basis. The outstanding merit of the proposed variable selection procedure is that it can achieve both robustness and efficiency by introducing an additional tuning parameter (i.e., bandwidth $h$). Its oracle property is also established for both the parametric and nonparametric part. Moreover, we give the data-driven bandwidth selection method and propose an EM-type algorithm for the proposed method. Monte Carlo simulation study and real data example are conducted to examine the finite sample performance of the proposed method. Both the simulation results and real data analysis confirm that the newly proposed method works very well.

The research was supported in part by National Natural Science Foundation of China (11171112, 11101114, 11201190), National Statistical Science Research Major Program of China (2011LZ051) and the Natural Science Foundation of Zhejiang Province Education Department (Y201121276).

W. Zhao · R. Zhang (✉) · J. Liu · Y. Lv
School of Finance and Statistics, East China Normal University, Shanghai 200241,
People's Republic of China
e-mail: zhangriquan@163.com

W. Zhao
School of Science, NanTong University, NanTong 226007, People's Republic of China

R. Zhang
Department of Mathematics, Shanxi Datong University, Datong 037009,
People's Republic of China

**Keywords** Partially linear varying coefficient model · Modal regression · $B$-spline · Variable selection · Oracle property · Robustness · Efficiency

## 1 Introduction

Semiparametric partially linear varying coefficient model (SPLVCM) is an extension of partially linear model and varying coefficient model (Hastie and Tibshirani 1993; Cai et al. 2000; Fan and Zhang 1999; Fan and Zhang 2000). It allows some coefficient functions to vary with certain covariates, such as time or age variable. If $Y$ is a response variable and $(U, \mathbf{X}, \mathbf{Z})$ is the associated covariates, then SPLVCM takes the form

$$Y = \mathbf{X}^T \boldsymbol{\alpha}(U) + \mathbf{Z}^T \boldsymbol{\beta} + \varepsilon, \tag{1}$$

where $U$ is the so called index variable, without loss of generality, we assume it ranges over the unit interval [0,1]; $\boldsymbol{\alpha}(\cdot) = (\alpha_1(\cdot), \ldots, \alpha_p(\cdot))^T$ is a unknown $p$-dimensional coefficient vector; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ is a $d$-dimensional vector of unknown regression coefficients; $\mathbf{Z} = (Z_1, \ldots, Z_d)^T \in \mathbb{R}^d$ and $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ are two vector predictors; $\varepsilon$ is random error with mean zero.

SPLVCM retains the virtues of both parametric and nonparametric modelling. It is a very flexible model and not only the linear interactions as in parametric model are considered but also general interactions between the index variable $U$ and these covariates are explored nonparametrically. Many papers have been focused on SPLVCM. For example, Li et al. (2002) introduced a local least-square method with a kernel weight function for SPLVCM; Zhang et al. (2002) studied SPLVCM based on local polynomial method (Fan and Gijbels 1996); Lu (2008) discussed the SPLVCM in the framework of generalized linear model based on two step estimation procedure; Xia et al. (2004) investigated the efficient estimation problem of parametric part for SPLVCM; Fan and Huang (2005) presented the profile likelihood inferences for SPLVCM based on profile least-square technique. As an extension of Fan and Huang (2005), a profile likelihood estimation procedure was developed in Lam and Fan (2008) under the generalized linear model framework with a diverging number of covariates.

However, all the above mentioned papers were built on either least square or likelihood based methods, which are expected to be very sensitive to outliers and their efficiency may be significantly reduced for many commonly used non-normal errors. Due to the well-known advantages of quantile regression, researchers set foot on SPLVCM in the framework of quantile regression method. For example, Wang et al. (2009) considered quantile regression SPLVCM by $B$-spline and developed rank score test; Cai and Xiao (2012) presented the model based on local polynomial smoothing. Although the QR based method is a robust modeling tool, it has limitations in terms of efficiency and uniqueness of estimation. Specially, since the check loss function for QR is not strictly convex, its estimation may not necessarily be unique in general. Moreover, the quantile method may lose some efficiency when there are no outliers or the error distribution is normal.

Recently, Yao et al. (2012) investigated a new estimation method based on local modal regression in a nonparametric model. A distinguishing characteristic of their proposed method is that it introduces an additional tuning parameter that is automatically selected using the observed data in order to achieve both robustness and efficiency of the resulting estimate. Their estimation method not only is robust when the data sets include outliers or heavy-tail error distribution but also as asymptotically efficient as least square based method when there are no outliers and the error distribution follows normal distribution. In other words, their proposed estimator is almost as efficient as an omniscient estimator. This fact motivates us to extend the modal regression method to SPLVCM by borrowing the idea of Yao et al. (2012).

In practice, there are often many covariates in both in parametric part and nonparametric part of the model (1). With high-dimensional covariates, sparse modeling is often considered superior, owing to enhanced model predictability and interpretability. Various powerful penalization methods have been developed for variable selection in parametric models, such as the Lasso (Tibshirani 1996), the SCAD (Fan and Li 2001), the elastic net (Zou and Hastie 2005), the adaptive lasso (Zou 2006), the Dantzig selector (Candes and Tao 2007), one step sparse estimation (Zou and Li 2008), more recently the MCP (Zhang 2010), etc. Similar to linear models, variable selection for semiparametric regressions is equally important and even more complex because model (1) involves both nonparametric and parametric parts.

There are only a few papers on variable selection in semiparametric regression models. Li and Liang (2008) considered the problem of variable selection for SPLVCM, where the parametric components are identified via the smoothed clipped absolute deviation (SCAD) procedure and the varying coefficients are selected via the generalized likelihood ratio test. Xie and Huang (2009) discussed SCAD-penalized regression in partially linear models, which is a special case of SPLVCM. Zhao and Xue (2009) investigated a selection procedure via SCAD which can select parametric components and nonparametric components simultaneously based on $B$-spline for SPLVCM. In addition, Leng (2009) proposed a simple approach of model selection for varying coefficient models and Lin and Yuan (2012) studied the variable selection of the generalized partially linear varying coefficient model based on basic function approximation. More importantly, Kai et al. (2011) introduced a robust variable selection method for SPLVCM based on composite quantile regression and local ploynomial method, but they only considered variable selection for parametric part of model (1). The main goal of this paper is to develop an effective and robust estimation and variable selection procedure based on modal regression to select significant parametric and nonparametric components in model (1), where the nonparametric components are approximated by $B$-spline. The proposed procedure possesses the oracle property in the sense of Fan and Li (2001) and the computation time is very fast. An important contribution of this paper is to develop a newly robust and efficient variable selection for SPLVCM.

The outline of this paper is as follows. In Sect. 2, following the idea of modal regression method, we describe a new estimation method for SPLVCM, where the varying coefficient functions are approximated by $B$-spline. Then, an efficient and robust variable selection procedure via SACD penalty is developed, which can select both the significant parametric components and nonparametric components in Sect. 3. Meanwhile, we also establish its oracle property for both parametric and nonparametric part. In Sect. 4, we give the details of bandwidth selection both in theory and in practise and propose an EM-type algorithm for the variable selection procedure. Moreover, we develop the CV method to select the optimal knots of $B$-spline approximation and optimal adaptive penalty parameter. In Sect. 5, we conduct simulation study and real data example to examine the finite-sample performance of the proposed procedures. Finally, some concluding remarks are given in Sect. 6. All the regularity conditions and the technical proofs are contained in the Appendix.

## 2 Robust estimation method

### 2.1 Modal regression

As a measure of center, the mean, the median and the mode are three important numerical characteristics of error distribution. Among them, median and mode have the common advantage of robustness, which can be resistent to outliers. Moreover, since the modal regression focuses on the relationship for majority of the data and summaries the "most likely" conditional values, it can provide more meaningful point prediction and larger coverage probability for prediction than others when the error density is skewed if the same length of short intervals, centered around each estimate, are used.

For the linear regression model $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$, Yao and Li (2011) proposed to estimate the regression parameter $\boldsymbol{\beta}$ by maximizing

$$Q(\boldsymbol{\beta}) \equiv \frac{1}{n} \sum_{i=1}^{n} \phi_h \left( y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right), \tag{2}$$

where $\phi_h(t) = h^{-1}\phi(t/h)$, $\phi(t)$ is a kernel density function and $h$ is a bandwidth.

To see why (2) can be used to estimate the modal regression, taking $\boldsymbol{\beta} = \beta_0$ as the intercept term only in linear regression, then (2) is simplified to

$$Q(\beta_0) \equiv \frac{1}{n} \sum_{i=1}^{n} \phi_h(y_i - \beta_0). \tag{3}$$

As a function of $\beta_0$, $Q_h(\beta_0)$ is the kernel estimate of the density function of $y$. Therefore, the maximizer of (3) is the mode of the kernel density function based on $y_1, \ldots, y_n$. When $n \to \infty$ and $h \to 0$, the mode of kernel density function will converge to the mode of the distribution of $y$.

For the univariate nonparametric regression model $y_i = m(x_i) + \varepsilon_i$, Yao et al. (2012) proposed to estimate the nonparametric function $m(x)$ using local polynomial by maximizing

$$Q(\theta) \equiv \frac{1}{n} \sum_{i=1}^{n} K_{\bar{h}}(x_i - x) \phi_h \left( y_i - \sum_{j=0}^{p} \theta_j (x_i - x)^j \right), \tag{4}$$

where $K_{\bar{h}}(\cdot) = K(\cdot/\bar{h})/\bar{h}$ is a rescaled kernel function of $K(\cdot)$ with bandwidth $\bar{h}$ for estimating nonparametric functions and $h$ is another bandwidth setting for $\phi(\cdot)$, and $\theta_j = m^{(j)}(x)/j!$.

Comparing with other estimation methods, modal regression treats $-\phi_h(\cdot)$ as a loss function instead of quadratic loss function for least square and check loss function for quantile regression. It provides the "most likely" conditional values rather than the conditional average or quantile. However, despite the usefulness of modal regression, it has received little attention in the literatures. Lee (1989) used the uniform kernel and fixed $h$ in (3) to estimate the modal regression. Scott (1992) proposed it, but little methodology is given on how to actually implement it. Recently, Yao and Li (2011) and Yao et al. (2012) systematically studied the modal regression for linear model and univariate nonparametric regression model. The main goal of this paper is to extend the modal regression to semiparametric models and discuss the variable selection for SPLVCM to obtain robust and efficient sparse estimator.

### 2.2 Estimation method for SPLVCM

Suppose that $\{\mathbf{X}_i, \mathbf{Z}_i, U_i, Y_i\}_{i=1}^n$ is an independent and identically distributed sample from the model (1). Since $\alpha_j(U)(j = 1, \ldots, p)$ in (1) are some unknown nonparametric functions, following the method of Yao et al. (2012), we can use local linear polynomial to approximate $\alpha_j(U)$ for $U$ in a neighborhood of $u$, i.e.,

$$\alpha_j(U) \approx \alpha_j(u) + \alpha_j'(u)(U - u) \triangleq a_j + b_j(U - u), \quad j = 1, \ldots, p.$$

Then we can obtain $\hat{\boldsymbol{\alpha}}(u)$ and $\hat{\boldsymbol{\beta}}$ by maximizing of local modal function

$$\frac{1}{n}\sum_{i=1}^{n} \phi_h\left(Y_i - \mathbf{X}_i^T(\mathbf{a} + \mathbf{b}(U_i - u)) - \mathbf{Z}_i^T\boldsymbol{\beta}\right) K_{\bar{h}}(U_i - u) \tag{5}$$

with respect to $\mathbf{a}$, $\mathbf{b}$ and $\boldsymbol{\beta}$, where $\mathbf{a} = (a_1, \ldots, a_p)^T$ and $\mathbf{b} = (b_1, \ldots, b_p)^T$.

However, there are two criticisms of local polynomial estimation for semiparametric models. Firstly, noting that the parameter $\beta$ is a global parameter, in order to obtain its optimal $\sqrt{n}$-consistent estimation, we need two-step estimation and under smoothing technique in the first-step estimation; Secondly, the computation task for local polynomial estimation is very heavy especially for high dimensional SPLVCM.

To avoid these drawbacks of local polynomial estimation, we propose to use basis function approximations for nonparametric functions. More specially, let $B(u) = (B_1(u), \ldots, B_q(u))^T$ be $B$-spline basis functions with the order of $\hbar$, where $q = K + \hbar + 1$, and $K$ is the number of interior knots. Then $\alpha_j(u)$ can be approximated by

$$\alpha_j(u) \approx B(u)^T \boldsymbol{\gamma}_j, \quad j = 1, \ldots, p.$$

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

$$Q(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \phi_h\left(Y_i - \mathbf{W}_i^T\boldsymbol{\gamma} - \mathbf{Z}_i^T\boldsymbol{\beta}\right), \tag{6}$$

with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, where $\mathbf{W}_i = I_p \otimes B(U_i) \cdot \mathbf{X}_i$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$. According to Yao et al. (2012), the choice of $\phi(\cdot)$ is not very crucial. For ease of computation, we use the standard normal density for $\phi(t)$ throughout this paper. The bandwidth $h$ in $\phi_h(\cdot)$ plays the role of the bandwidth, which determines the degree of robustness of the estimator.

## 3 Variable selection for SPLVCM

In this section, we develop a robust and efficient variable selection procedure for SPLVCM and prove its oracle property via SCAD penalty.

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

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

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

We define the penalized estimation for SPLVCM based on modal regression as

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}) = Q(\boldsymbol{\gamma}, \boldsymbol{\beta}) - n \sum_{j=1}^{p} p_{\lambda_{1j}} \left( \|B(\cdot)^T \boldsymbol{\gamma}_j\| \right) - n \sum_{k=1}^{d} p_{\lambda_{2k}}(|\beta_k|), \tag{7}$$

where $\lambda_{1j} (j = 1, \ldots, p)$ and $\lambda_{2k} (k = 1, \ldots, d)$ are penalized parameters for the $j$th varying coefficient function and the $k$th parameter component, respectively.

Note that the regularization parameters for the penalty functions and in (7) are not necessarily the same for $\boldsymbol{\gamma}_j, j = 1, \ldots, p$ and $\boldsymbol{\beta}_k, k = 1, \ldots, d$, which can provide with flexibility and adaptivity. By this adaptive strategy, the tuning parameter for zero coefficient could be larger than that for nonzero coefficient, which can simultaneously unbiasedly estimate large coefficients and shrink the small coefficients toward zero. By maximizing the above objective function with proper penalty parameters, we can get sparse estimators and hence conduct variable selection.

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ be the solution by maximizing (7). Therefore, the estimator of $\alpha_j(u)$ can be obtained by $\hat{\alpha}_j(u) = B(u)^T \hat{\boldsymbol{\gamma}}_j, j = 1, \ldots, p$. We call $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}_j(u)$ as the penalized estimator of $\boldsymbol{\beta}$ and $\alpha_j(u)$ based on spline and robust modal regression (SMR) for SPLVCM. Next, we discuss the asymptotic properties of the resulting penalized estimators. Denote $\boldsymbol{\alpha}_0(\cdot)$ and $\boldsymbol{\beta}_0$ to be the true values of $\boldsymbol{\alpha}(\cdot)$ and $\boldsymbol{\beta}$, respectively. Without loss of generality, we assume that $\alpha_{j0}(\cdot) = 0, j = s_1 + 1, \ldots, p$, and $\alpha_{j0}(\cdot), j = 1, \ldots, s_1$ are all nonzero components of $\boldsymbol{\alpha}_0(\cdot)$. Furthermore, we assume that $\beta_{k0} = 0, k = s_2 + 1, \ldots, d$, and $\beta_{k0}, k = 1, \ldots, s_2$ are all nonzero components of $\boldsymbol{\beta}_0$. Let

$$F(\mathbf{x}, \mathbf{z}, u, h) = \mathrm{E}\left\{\phi_h''(\varepsilon)|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}, U = u\right\}$$

and

$$G(\mathbf{x}, \mathbf{z}, u, h) = \mathrm{E}\left\{\phi_h'(\varepsilon)^2|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}, U = u\right\}.$$

Denote

$$a_n = \max_{j,k}\left\{|p'_{\lambda_{1j}}(\|\boldsymbol{\gamma}_{j0}\|_H)|, |p'_{\lambda_{2k}}(|\beta_{k0}|)| : \boldsymbol{\gamma}_{j0} \neq 0, \beta_{k0} \neq 0\right\}$$

and

$$b_n = \max_{j,k}\left\{|p''_{\lambda_{1j}}(\|\boldsymbol{\gamma}_{j0}\|_H)|, |p''_{\lambda_{2k}}(|\beta_{k0}|)| : \boldsymbol{\gamma}_{j0} \neq 0, \beta_{k0} \neq 0\right\},$$

where $\|\boldsymbol{\gamma}_{j0}\|_H = \sqrt{\boldsymbol{\gamma}_{j0}^T H \boldsymbol{\gamma}_{j0}}$, $H = \int_0^1 B(u)B^T(u)du$, $\boldsymbol{\gamma}_{j0}$ is the best approximation coefficient of $\alpha_j(u)$ in the $B$-spline space. Then, we have the following Theorem 1 which gives the consistency of the proposed penalized estimators.

**Theorem 1** *Suppose that the regularity conditions (C1)–(C8) in the Appendix hold and the numbers of knots $K = O(n^{1/(2r+1)})$. If $b_n \to 0$, then we have*

(i) $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p\left(n^{\frac{-r}{2r+1}} + a_n\right)$,
(ii) $\|\hat{\alpha}_j(\cdot) - \alpha_{j0}\| = O_p\left(n^{\frac{-r}{2r+1}} + a_n\right), \quad j = 1, \ldots, p,$
*where $r$ is defined in the condition (C2) in the Appendix.*

Let $\lambda_{\max} = \max_{j,k}\{\lambda_{1j}, \lambda_{2k}\}$ and $\lambda_{\min} = \min_{j,k}\{\lambda_{1j}, \lambda_{2k}\}$. Under some conditions, we can show that the consistent estimators in Theorem 1 possess the sparse property, which is stated as follows.

**Theorem 2** *Suppose that the regularity conditions (C1)–(C8) in the Appendix hold and the numbers of knots $K = O(n^{1/(2r+1)})$. If $\lambda_{\max} \to 0$ and $n^{\frac{r}{2r+1}}\lambda_{\min} \to \infty$ as $n \to \infty$, then with probability tending to 1, $\hat{\boldsymbol{\beta}}$ and $\hat{\alpha}_j(\cdot)$ satisfy*

(i) $\hat{\beta}_k = 0, \quad k = s_2 + 1, \ldots, d,$
(ii) $\hat{\alpha}_j(\cdot) = 0, \quad j = s_1 + 1, \ldots, p.$

*Remark 1* For SCAD penalty function, we know that if $\lambda_{\max} \to 0$ as $n \to \infty$, then $a_n = 0$. Therefore, from Theorems 1 and 2, it is clear that, by choosing proper tuning parameters, our proposed variable selection method is consistent and the estimators of nonparametric components achieve the optimal convergence rate as if the subset of true zero coefficients is already known (Stone 1982).

Next, we show that the estimators for nonzero coefficients in the parametric components have the same asymptotic distribution as that based on the oracle model. To demonstrate this, we need more notations to present the asymptotic property of the resulting estimators. Let $\boldsymbol{\gamma}_a = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_{s_1}^T)^T$, $\boldsymbol{\beta}_a = (\beta_1, \ldots, \beta_{s_2})^T$, and $\boldsymbol{\gamma}_{a0}$ and $\boldsymbol{\beta}_{a0}$ be the true values of $\boldsymbol{\gamma}_a$ and $\boldsymbol{\beta}_a$, respectively. Corresponding covariates are denoted by $\mathbf{W}_a$ and $\mathbf{Z}_a$. In addition, denote

$$\Phi = \mathrm{E}\left(\phi_h''(\varepsilon)\mathbf{W}_a\mathbf{W}_a^T\right) = \mathrm{E}\left(F(\mathbf{X}, \mathbf{Z}, U, h)\mathbf{W}_a\mathbf{W}_a^T\right)$$

and

$$\Psi = \mathrm{E}\left(\phi_h''(\varepsilon)\mathbf{W}_a\mathbf{Z}_a^T\right) = \mathrm{E}\left(F(\mathbf{X}, \mathbf{Z}, U, h)\mathbf{W}_a\mathbf{Z}_a^T\right).$$

Then we have the following theorem.

**Theorem 3** *Under the conditions of Theorem 2, we have*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{a0}\right) \xrightarrow{\mathrm{d}} N(0, \Sigma^{-1}\Delta\Sigma^{-1}), \tag{8}$$

*where* $\Delta = \mathrm{E}(G(\mathbf{X}, \mathbf{Z}, U, h)\check{\mathbf{Z}}_a\check{\mathbf{Z}}_a^T)$, $\Sigma = \mathrm{E}(F(\mathbf{X}, \mathbf{Z}, U, h)\check{\mathbf{Z}}_a\check{\mathbf{Z}}_a^T)$, $\check{\mathbf{Z}}_a = \mathbf{Z}_a - \Psi^T\Phi^{-1}\mathbf{W}_a$.

Let $\tilde{\boldsymbol{\alpha}}_j(u) = B^T(u)\boldsymbol{\gamma}_{j0}$ for $j = 1, \ldots, s_1$, denote $\tilde{\boldsymbol{\alpha}}_a(u) = (\tilde{\boldsymbol{\alpha}}_1(u), \ldots, \tilde{\boldsymbol{\alpha}}_{s_1}(u))^T$ and $\hat{\boldsymbol{\alpha}}_a(u) = (\hat{\boldsymbol{\alpha}}_1(u), \ldots, \hat{\boldsymbol{\alpha}}_{s_1}(u))^T$, then the following result holds.

**Theorem 4** *Under the conditions of Theorem 2, for any vector* $\mathbf{c}_n$ *with dimension* $q \times s_1$ *and components not all 0, then we have*

$$\left\{\mathbf{c}_n^T \mathrm{var}(\hat{\boldsymbol{\alpha}}_a(u))\mathbf{c}_n\right\}^{-1/2} \mathbf{c}_n^T\left(\hat{\boldsymbol{\alpha}}_a(u) - \tilde{\boldsymbol{\alpha}}_a(u)\right) \xrightarrow{\mathrm{d}} N(0, 1). \tag{9}$$

The proofs of Theorems 1–4 are given in the Appendix.

## 4 Bandwidth selection and estimation algorithm

In this section, we first discuss the selection of bandwidth both in theoretical and in practice. Then, we develop estimation procedure for SPLVCM based on MEM algorithm (Li et al. 2007) and LQA algorithm (Fan and Li 2001). Note that the bandwidth selection discussing in this section is not the same as the bandwidth selection in local polynomial fitting for SPLVCM (Li and Palta 2009).

### 4.1 Optimal bandwidth

In this subsection, we give the optimal bandwidth in theoretical. For simplicity, we assume that the error variable independent of $\mathbf{X}$, $\mathbf{Z}$ and $U$, based on (8) and the asymptotic variance of least-square $B$-spline estimator (LSB) given in Zhao and Xue

(2009), we can show that the ratio of the asymptotic variance of the SMR estimator to that of the LPB estimator is given by

$$r(h) \triangleq \frac{G(h)F^{-2}(h)}{\sigma^2}, \tag{10}$$

where $\sigma^2 = \mathrm{E}(\varepsilon^2)$, $F(h) = \mathrm{E}\{\phi_h''(\varepsilon)\}$ and $G(h) = \mathrm{E}\{\phi_h'(\varepsilon)\}^2$. The ratio $r(h)$ depends on $h$ only, and it plays an important role in efficiency and robustness of estimators. Therefore, the ideal choice of $h$ is

$$h_{\text{opt}} = \text{argmin}_h r(h) = \text{argmin}_h G(h)F^{-2}(h). \tag{11}$$

From (11), we can see that $h_{\text{opt}}$ does not depend on $n$ and only depends on the conditional error distribution of $\varepsilon$.

*Remark 2* Based on the expression of the ratio $r(h)$, for all $h > 0$, we can prove that $\inf_h r(h) = 1$ if the error follows normal distribution, and $\inf_h r(h) \le 1$ regardless of the error distribution. Hence, SMR is better than or at least as well as LSB. In particular, if the error distribution has heavy tails or has large variance, the performance of SMR is much better than LSB.

### 4.2 Bandwidth selection in practice

In practice, we do not know the error distribution, hence we cannot obtain $F(h)$ and $G(h)$. An feasible method is to estimate $F(h)$ and $G(h)$ by

$$\hat{F}(h) = \frac{1}{n}\sum_{i=1}^{n} \phi_h''(\hat{\varepsilon}_i) \quad \text{and} \quad \hat{G}(h) = \frac{1}{n}\sum_{i=1}^{n} \left\{\phi_h'(\hat{\varepsilon}_i)\right\}^2,$$

respectively.

Then $r(h)$ can be estimated by $\hat{r}(h) = \hat{G}(h)\hat{F}(h)^{-2}/\hat{\sigma}^2$, where $\hat{\varepsilon}_i = Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\alpha}}(U_i) - \mathbf{Z}_i^T\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}(\cdot)$, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ are estimated based on the pilot estimates. Then, using the grid search method, we can easily find $h_{\text{opt}}$ to minimize $\hat{r}(h)$. According to the advise of Yao et al. (2012), the possible grids points for $h$ can be $h = 0.5\hat{\sigma} \times 1.02^j$, $j = 0, 1, \ldots, k$, for some fixed $k$ (such as $k = 50$ or 100).

### 4.3 Algorithm

In this subsection, we extend the modal expectation-maximization (MEM) algorithm (Li et al. 2007) and local quadratic algorithm (LQA, Fan and Li 2001) to maximize (7) for SPLVCM. Here, we assume $\phi(\cdot)$ is the density function of a standard normal distribution.

Because $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is irregular at the origin. Directly maximizing (7) may be difficult. Following Fan and Li (2001), we first locally approximate the penalty function $p_\lambda(\cdot)$ by a quadratic function at every step of iteration. More specifically, in a neighborhood of a given nonzero $\omega_0$, an approximation of the penalized function at value $\omega_0$ can be given by

$$p_\lambda(|\omega|) \approx p_\lambda(|\omega_0|) + \frac{1}{2}\left\{p'_\lambda(|\omega_0|)/|\omega_0|\right\}\left(\omega^2 - \omega_0^2\right), \quad \text{for } \omega \approx \omega_0.$$

Hence, if initial estimators $\beta_k^{(0)}$ and $\boldsymbol{\gamma}_j^{(0)}$ are very close to 0, then set $\hat{\beta}_k = 0$ and $\hat{\boldsymbol{\gamma}}_j = 0$; otherwise, for the given initial value $\beta_k^{(0)}$ with $|\beta_k^{(0)}| > 0$, $k = 1, \ldots, d$, and $\boldsymbol{\gamma}_j^{(0)}$ with $\|\boldsymbol{\gamma}_j^{(0)}\|_H > 0$, $j = 1, \ldots, p$, we can obtain

$$p_{\lambda_{2k}}(|\beta_k|) \approx p_{\lambda_{2k}}\left(|\beta_k^{(0)}|\right) + \frac{1}{2}\frac{p'_{\lambda_{2k}}\left(|\beta_k^{(0)}|\right)}{|\beta_k^{(0)}|}\left(|\beta_k|^2 - |\beta_k^{(0)}|^2\right) \quad \text{and}$$

$$p_{\lambda_{1j}}(\|\boldsymbol{\gamma}_j\|_H) \approx p_{\lambda_{1j}}\left(\|\boldsymbol{\gamma}_j^{(0)}\|_H\right) + \frac{1}{2}\frac{p'_{\lambda_{1j}}\left(\|\boldsymbol{\gamma}_j^{(0)}\|_H\right)}{\|\boldsymbol{\gamma}_j^{(0)}\|_H}\left(\|\boldsymbol{\gamma}_j\|_H^2 - \|\boldsymbol{\gamma}_j^{(0)}\|_H^2\right).$$

Denote $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$, $\mathbf{Z}_i^* = (\mathbf{Z}_i^T, \mathbf{W}_i^T)^T$ and set $m = 0$. Let

$$\Sigma_\lambda(\boldsymbol{\theta}^{(m)}) = \operatorname{diag}\left\{\frac{p'_{\lambda_{21}}\left(|\beta_1^{(m)}|\right)}{|\beta_1^{(m)}|}, \ldots, \frac{p'_{\lambda_{2d}}\left(|\beta_d^{(m)}|\right)}{|\beta_d^{(m)}|}, \frac{p'_{\lambda_{11}}\left(\|\boldsymbol{\gamma}_1^{(m)}\|_H\right)}{\|\boldsymbol{\gamma}_1^{(m)}\|_H}H, \ldots, \right.$$
$$\left. \frac{p'_{\lambda_{1p}}\left(\|\boldsymbol{\gamma}_p^{(m)}\|_H\right)}{\|\boldsymbol{\gamma}_p^{(m)}\|_H}H\right\}.$$

With the aid of LQA and MEM algorithm, we can obtain the sparse estimators as follows:

**Step 1 (E-step):** We first update $\pi(i|\boldsymbol{\theta}^{(m)})$ by

$$\pi(i|\boldsymbol{\theta}^{(m)}) = \frac{\phi_h\left(Y_i - \mathbf{Z}_i^{*T}\boldsymbol{\theta}^{(m)}\right)}{\sum_{i=1}^n \phi_h\left(Y_i - \mathbf{Z}_i^{*T}\boldsymbol{\theta}^{(m)}\right)} \propto \phi_h\left(Y_i - \mathbf{Z}_i^{*T}\boldsymbol{\theta}^{(m)}\right), \quad i = 1, \ldots, n,$$

**Step 2 (M-step):** Then, we update $\boldsymbol{\theta}$ obtain $\hat{\boldsymbol{\theta}}^{(m+1)}$

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{(m+1)} &= \operatorname{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^n \left\{\pi(i|\boldsymbol{\theta}^{(m)})\log\phi_h\left(Y_i - \mathbf{Z}_i^{*T}\boldsymbol{\theta}^{(m)}\right)\right\} + \tfrac{n}{2}\boldsymbol{\theta}^T\Sigma_\lambda(\boldsymbol{\theta}^{(m)})\boldsymbol{\theta} \\ &= \left(\mathbf{Z}^{*T}W\mathbf{Z}^* + n\Sigma_\lambda(\boldsymbol{\theta}^{(m)})\right)^{-1}\mathbf{Z}^{*T}W\mathbf{Y},\end{aligned}$$

where $W$ is an $n \times n$ diagonal matrix with diagonal elements $\pi(i|\boldsymbol{\theta}^{(m)})$s.

**Step 3:** Iterate the E-step and M-step until converges, and denote the final estimator of $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}$. Then $\hat{\boldsymbol{\beta}} = (I_{d\times d}, 0_{d\times pq})\hat{\boldsymbol{\theta}}$, and $\hat{\boldsymbol{\gamma}} = (0_{pq\times d}, I_{pq\times pq})\hat{\boldsymbol{\theta}}$.

Similar to the EM algorithm, the MEM algorithm for SPLVCM within each Step also consists of two steps: E-step and M-step. The ascending property of the proposed MEM algorithm can be established along the lines of the study of Li et al. (2007).

Note that the converged value may depend on the starting point as the usual EM algorithms, and there is no guarantee that the MEM algorithm will converge to the global optimal solution. Therefore, it is prudent to run the algorithm from several starting-points and choose the best local optima found.

### 4.4 Selection of tuning parameter

To implement the above estimation procedure, the number of interior knots $K$, and the tuning parameters $a$, $\lambda_{1j}$'s and $\lambda_{2k}$'s in the penalty functions should be chosen appropriately. According to the suggestion of Fan and Li (2001), the choice of $a = 3.7$ performs well in a variety situations. Hence, we also use this value throughout this paper. We note that there are total $(p + d)$-dimension penalty parameters ($\lambda_{1j}$'s and $\lambda_{2k}$'s) need to be selected. To reduce the computation task, we can use following strategy to set

$$\lambda_{1j} = \frac{\lambda}{\|\hat{\boldsymbol{\gamma}}_j^{(0)}\|_H} \quad \text{and} \quad \lambda_{2k} = \frac{\lambda}{|\hat{\beta}_k^{(0)}|}, \tag{12}$$

where $\hat{\boldsymbol{\gamma}}_j^{(0)}$ and $\hat{\beta}_j^{(0)}$ are the initial estimators of $\boldsymbol{\gamma}_j$ and $\beta_k$, respectively, using unpenalized estimator. Then we can use the following two-dimensional cross-validation score maximization problem

$$\mathrm{CV}(K, \lambda) = \sum_{i=1}^{n} \phi_h\left(Y_i - \mathbf{W}_i^T \hat{\boldsymbol{\gamma}}^{(-i)} - \mathbf{Z}_i^T \hat{\boldsymbol{\beta}}^{(-i)}\right), \tag{13}$$

where $\hat{\boldsymbol{\beta}}^{(-i)}$ and $\hat{\boldsymbol{\gamma}}^{(-i)}$ are the solution of (7) after deleting the $i$th subject. Then, the optimal $K_{\text{opt}}$ and $\lambda_{\text{opt}}$ are obtained by

$$(K_{\text{opt}}, \lambda_{\text{opt}}) = \max_{K,\lambda} \mathrm{CV}(K, \lambda).$$

Note that the above strategy of selecting tuning parameters, in some sense, is the same rationale behind the adaptive Lasso (Zou 2006), and from our simulation experience, we found that this method performs well.

## 5 Numerical properties

In this section, we conduct simulation study to assess the finite-sample performance of the proposed procedures and illustrate the proposed methodology on a real-world data set in a health study. In all examples, we use the kernel function to be the Gaussian kernel.

### 5.1 Simulation study

In this example, we generate the random samples from the model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\alpha}(U_i) + \mathbf{Z}_i^T \boldsymbol{\beta} + \varepsilon_i,$$

where $\boldsymbol{\alpha}(u) = (\alpha_1(u), \ldots, \alpha_8(u))^T$, $\alpha_1(u) = 2\sin(2\pi u)$, $\alpha_2(u) = 8u(1-u)$ and $\alpha_j(u) = 0, j = 3, \ldots, 10$; $\beta = (2.5, 1.2, 0.5, 0, 0, 0, 0, 0, 0, 0)^T$; The covariate vector $(\mathbf{X}_i^T, \mathbf{Z}_i^T)^T$ is normally distributed with mean 0, variance 1 and correlation $0.8^{|k-j|}, 1 \le k, j \le p + d, p = d = 10$; The index variable $U_i$ is simulated from $U[0, 1]$ and independent of $(\mathbf{X}_i^T, \mathbf{Z}_i^T)^T$. In our simulations, we considered the following five different error distributions: $N(0, 1)$, $t$-distribution with freedom degree 3: $t(3)$, Laplace distribution: $Lp(0, 1)$, Laplace mixture distribution: $0.8Lp(0, 1) + 0.2Lp(0, 5)$ and mixture of normals: $0.9N(0, 1) + 0.1N(0, 10)$ and error $\epsilon_i$ is independent of all covariates. The sample size $n$ is set to be 200, 400 and 600. A total of 400 simulation replications are conducted for each model setup. In all simulations, we use cubic $B$-spline basis to approximate varying coefficient functions and the optimal knots and penalty parameter obtained by CV method in Sect. 4.4.

The performance of the nonparametric estimator $\hat{\alpha}(\cdot)$ will be assessed using the square root of average square errors (RASE)

$$\text{RASE} = \left\{ n_{\text{grid}}^{-1} \sum_{k=1}^{n_{\text{grid}}} \|\hat{\alpha}(u_k) - \alpha(u_k)\|^2 \right\}^{1/2},$$

where $\{u_k, k = 1, \ldots, n_{\text{grid}}\}$ are the grid points at which the functions $\{\hat{\alpha}_j(\cdot)\}$ are evaluated. The generalized mean square error (GMSE) as defined in Li and Liang (2008) is used to evaluate the performance for parametric part

$$\text{GMSE} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathrm{E}(\mathbf{Z}\mathbf{Z}^T)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

The medians of RASE and GMSE are listed in Table 1. To examine the robustness and efficiency of the proposed procedure (SMR), we also compare the simulation results with least square $B$-spline estimator (LSB) (Zhao and Xue 2009). Column "CN" shows the average number of zero coefficients correctly estimated to be zero for varying coefficient functions, and Column "CP" for parametric part. In the column labeled "IN", we present the average number of nonzero coefficients incorrectly estimated to be zero for varying coefficient part, and "IP" for parametric part.

Several observations can be made from the Table 1. Firstly, for the given sample size, penalized SMR estimator performs obviously better than penalized LSB estimator method especially for non-normal error distribution. Secondly, for the given error distribution, the performances of SMR estimator become better and better when the

**Table 1** Simulation results with different error distributions

| $n$ | Method | RASE | GMSE | CN | IN | CP | IP |
|---|---|---|---|---|---|---|---|
| $N(0, 1)$ | | | | | | | |
| 200 | LSB | 0.1861 (0.0579) | 0.0158 (0.0160) | 8.0000 | 0.0025 | 6.2850 | 0 |
| | SMR | 0.2286 (0.1080) | 0.0312 (0.0660) | 7.9925 | 0 | 6.8000 | 0.0025 |
| 400 | LSB | 0.1238 (0.0400) | 0.0070 (0.0069) | 8.0000 | 0 | 6.4150 | 0 |
| | SMR | 0.1257 (0.0423) | 0.0073 (0.0072) | 7.9975 | 0 | 6.8775 | 0 |
| 600 | LSB | 0.1001 (0.0318) | 0.0041 (0.0050) | 8.0000 | 0 | 6.5375 | 0 |
| | SMR | 0.1012 (0.0322) | 0.0043 (0.0054) | 8.0000 | 0 | 6.8900 | 0 |
| $t(3)$ | | | | | | | |
| 200 | LSB | 0.3026 (0.1224) | 0.0451 (0.0632) | 7.9400 | 0.0175 | 6.2350 | 0 |
| | SMR | 0.2879 (0.1234) | 0.0425 (0.0686) | 7.9400 | 0.0100 | 6.7550 | 0 |
| 400 | LSB | 0.2015 (0.0829) | 0.0216 (0.0308) | 7.9375 | 0 | 6.4050 | 0 |
| | SMR | 0.1631 (0.0547) | 0.0128 (0.0147) | 7.9825 | 0 | 6.7950 | 0 |
| 600 | LSB | 0.1770 (0.0684) | 0.0131 (0.0153) | 7.9775 | 0 | 6.5025 | 0 |
| | SMR | 0.1296 (0.0440) | 0.0079 (0.0084) | 7.9900 | 0 | 6.8575 | 0 |
| Laplace $Lp(0, 1)$ | | | | | | | |
| 200 | LSB | 0.2550 (0.0881) | 0.0303 (0.0336) | 7.9975 | 0.0125 | 6.1925 | 0 |
| | SMR | 0.2545 (0.0906) | 0.0292 (0.0363) | 7.9825 | 0.0050 | 6.7175 | 0 |
| 400 | LSB | 0.1741 (0.0552) | 0.0150 (0.0155) | 8.0000 | 0 | 6.2700 | 0 |
| | SMR | 0.1490 (0.0574) | 0.0111 (0.0127) | 7.9850 | 0 | 6.8475 | 0 |
| 600 | LSB | 0.1467 (0.0456) | 0.0085 (0.0084) | 8.0000 | 0 | 6.5300 | 0 |
| | SMR | 0.1198 (0.0428) | 0.0063 (0.0068) | 7.9850 | 0 | 6.8500 | 0 |
| Laplace mixture $0.8Lp(0, 1) + 0.2Lp(0, 5)$ | | | | | | | |
| 200 | LSB | 0.7083 (0.3371) | 0.2032 (0.2891) | 6.2050 | 0.1950 | 6.2175 | 0.0325 |
| | SMR | 0.3030 (0.1423) | 0.0471 (0.0697) | 7.8475 | 0.0050 | 6.7325 | 0.0075 |
| 400 | LSB | 0.4714 (0.2103) | 0.0865 (0.1394) | 6.3675 | 0.0225 | 6.3925 | 0 |
| | SMR | 0.1802 (0.0607) | 0.0159 (0.0167) | 7.9775 | 0 | 6.8175 | 0 |
| 600 | LSB | 0.3568 (0.1485) | 0.0531 (0.0667) | 6.5850 | 0 | 6.4950 | 0 |
| | SMR | 0.1457 (0.0462) | 0.0088 (0.0090) | 7.9825 | 0 | 6.8075 | 0 |
| Normal mixture $0.9N(0, 1) + 0.1N(0, 10)$ | | | | | | | |
| 200 | LSB | 0.6290 (0.3246) | 0.1900 (0.2867) | 6.3875 | 0.2275 | 6.5325 | 0.0025 |
| | SMR | 0.2156 (0.0764) | 0.0272 (0.0270) | 7.9875 | 0 | 6.8025 | 0 |
| 400 | LSB | 0.4364 (0.1950) | 0.0751 (0.0869) | 6.4925 | 0.0150 | 6.6525 | 0 |
| | SMR | 0.1446 (0.0409) | 0.0083 (0.0098) | 7.9975 | 0 | 6.8325 | 0 |
| 600 | LSB | 0.3532 (0.1502) | 0.0498 (0.0599) | 6.6025 | 0 | 6.6800 | 0 |
| | SMR | 0.1171 (0.0364) | 0.0057 (0.0057) | 7.9975 | 0 | 6.8450 | 0 |

The estimated standard deviations for RASE and GMSE based on 400 replications are given in the parentheses

sample size increases. Thirdly, even for the normal error case, the SMR seems to perform no worse than the LSB. Especially, when sample size $n = 600$, there is almost no efficiency lost of RASE and GMSE compared with LSB or even slightly better in term of variable selection. Moreover, it is very interesting to see that the superiority of SMR become more and more obvious when the error follows mixture distribution and sample size is large. The main reason for this is that the larger of the sample size, the more likely the data contain outliers, and when there are some very large outliers in the data, the modal regression will put more weight to the "most likely" data around the true value, which lead to robust and efficient estimator.

To conclude, the penalized SMR estimator is better than or at least as well as LSB estimator.

## 5.2 Real data analysis

As an illustration, we apply the proposed procedures to analyze the plasma beta-carotene level data set collected by a cross-sectional study (Nierenberg et al. 1989). Research has shown that there is a direct relationship between beta-carotene and cancers such as lung, colon, breast, and prostate cancer (Fairfield and Fletcher 2002). This data set consists of 315 observations. The data can be downloaded from the StatLib database via the link "*lib.stat.cmu.edu/datasets/Plasma_Retinol*". Brief description of the variable is shown in Table 2.

Of interest is the relationships between the plasma beta-carotene level and the following covariates: sex, smoking status, quetelet index (BMI), vitamin use, number of calories, grams of fat, grams of fiber, number of alcoholic drinks, cholesterol and age. We fit the data using SPLVCM with $U$ being "Age". The covariates "smoking status" and "vitamin use" are categorical variables and are thus replaced with dummy variables. We take these two dummy variables and other two discrete variables "sex" and "alcohol" as covariates of parametric part. All of the other covariates are standardized as the covariates of varying coefficient part. The index variable $U$ is rescaled

**Table 2** Plasma beta-carotene level data

| Number | Variable | Description |
|---|---|---|
| 1 | AGE | Age (years) |
| 2 | SEX | Sex (1 = male, 2 = female) |
| 3 | SMOKSTAT | Smoking status (1 = never, 2 = former, 3 = current smoker) |
| 4 | QUETELET | Quetelet (weight/(height$^2$)) |
| 5 | VITUSE | Vitamin use (1 = yes, fairly often, 2 = yes, not often, 3 = no) |
| 6 | CALORIES | Number of calories consumed per day |
| 7 | FAT | Grams of fat consumed per day |
| 8 | FIBER | Grams of fiber consumed per day |
| 9 | ALCOHOL | Number of alcoholic drinks consumed per week |
| 10 | CHOLESTEROL | Cholesterol consumed (mg/day) |
| 11 | BETADIET | Dietary beta-carotene consumed (mcg/day) |
| 12 | RETDIET | Dietary retinol consumed (mcg/day) |
| 13 | BETAPLASMA | Plasma beta-carotene (ng/ml) |
| 14 | RETPLASMA | Plasma retinol (ng/ml) |

into interval [0,1]. We applied the SMR and LSB estimators to fit the SPLVCM. We randomly split the data into two parts, where 2/3 observations used as a training data set to fit the model and select significant variables, and the remaining 1/3 observations as test data set to evaluate the predictive ability of the selected model. The prediction performance is measured by the median absolute prediction error (MAPE), which is the median of $\{|Y_i^{\text{test}} - \hat{Y}_i^{\text{test}}|, i = 1, \ldots, 105\}$.

Beside the SCAD penalty, to see the effect of variable selection result for SMR, we also consider other two penalty functions, i.e. LASSO and MCP. We found that both SMR and LSB estimators select all the variables in varying coefficient part for three different penalty. The estimations of varying coefficient functions for SMR with SCAD penalty are depicted in Fig. 1. The resulting estimations for parametric part and MAPE together with the optimal bandwidth and penalized parameter are given in Table 3 (the estimated standard deviance for parametric component is given in the brackets).

As we can see from Table 3, the performances of the three different penalty are very similar and SMR estimator is sparser than LSB method. Meanwhile, for all three penalties, the MAPE of SMR is smaller than LSB, which indicate that SMR model has better prediction performance than LSB model for the plasma beta-carotene level data. Because, for this data, the response $Y$ is left-skewed, which can be seen in Fig. 1h. In addition, to confirm whether the selected variables in nonparametric part are truly relevant, we found that none of their 95 % pointwise confidence intervals (the dot-dashed lines) can completely well cover 0, which can see from the Fig. 1a–g.

*Remark 3* Based on the result of Theorem 4, we can construct pointwise confidence intervals for each varying coefficient function if we know $\text{var}(\hat{\alpha}_j(u))$ or its estimate $\widehat{\text{var}}(\hat{\alpha}_j(u))$. In practice, because $\text{var}(\hat{\alpha}_j(u))$ is unknown, one can use sandwich formula to obtain $\widehat{\text{var}}(\hat{\alpha}_j(u))$. However, the sandwich formula for $\widehat{\text{var}}(\hat{\alpha}_j(u))$ is very complicated and it includes many approximations, sometimes the results of the confidence intervals are not very well. Here, we obtain the 95 % pointwise confidence intervals for each nonzero varying coefficient function using Bootstrap resampling method. With $B$ independent bootstrap samples, we can obtain the $B$ bootstrap estimators of $\tilde{\alpha}_j(u)$, then the sample standard error $\hat{\sigma}_{j,B}(u)$ of $\tilde{\alpha}_j(u)$ can be computed, and a $1-\alpha$ confidence intervals for $\tilde{\alpha}_j(u)$ based on a normal approximation is

$$\hat{\alpha}_j(u) \pm z_{1-\alpha/2}\hat{\sigma}_{j,B}(u), \quad \text{for } j = 1, \ldots, s_1,$$

where $z_p$ is the $100p$th percentile of the standard normal distribution. If the bias $\tilde{\alpha}_j(u) - \alpha_{j0}(u)$ is asymptotically negligible relative to the variance of $\hat{\alpha}_j(u)$ by choosing a large $K$, then $\hat{\alpha}_j(u) \pm z_{1-\alpha/2}\hat{\sigma}_{j,B}(u)$ is also a $1-\alpha$ asymptotic confidence intervals for $\alpha_{j0}(u)$. More details see Huang et al. (2002) and Wang et al. (2008).

## 6 Concluding remarks

Variable selection for SPLVCM has been an interesting topic. However, most existing methods were built on either least square or likelihood based methods, which are very sensitive to outliers and their efficiency may be significantly reduced for heavy tail

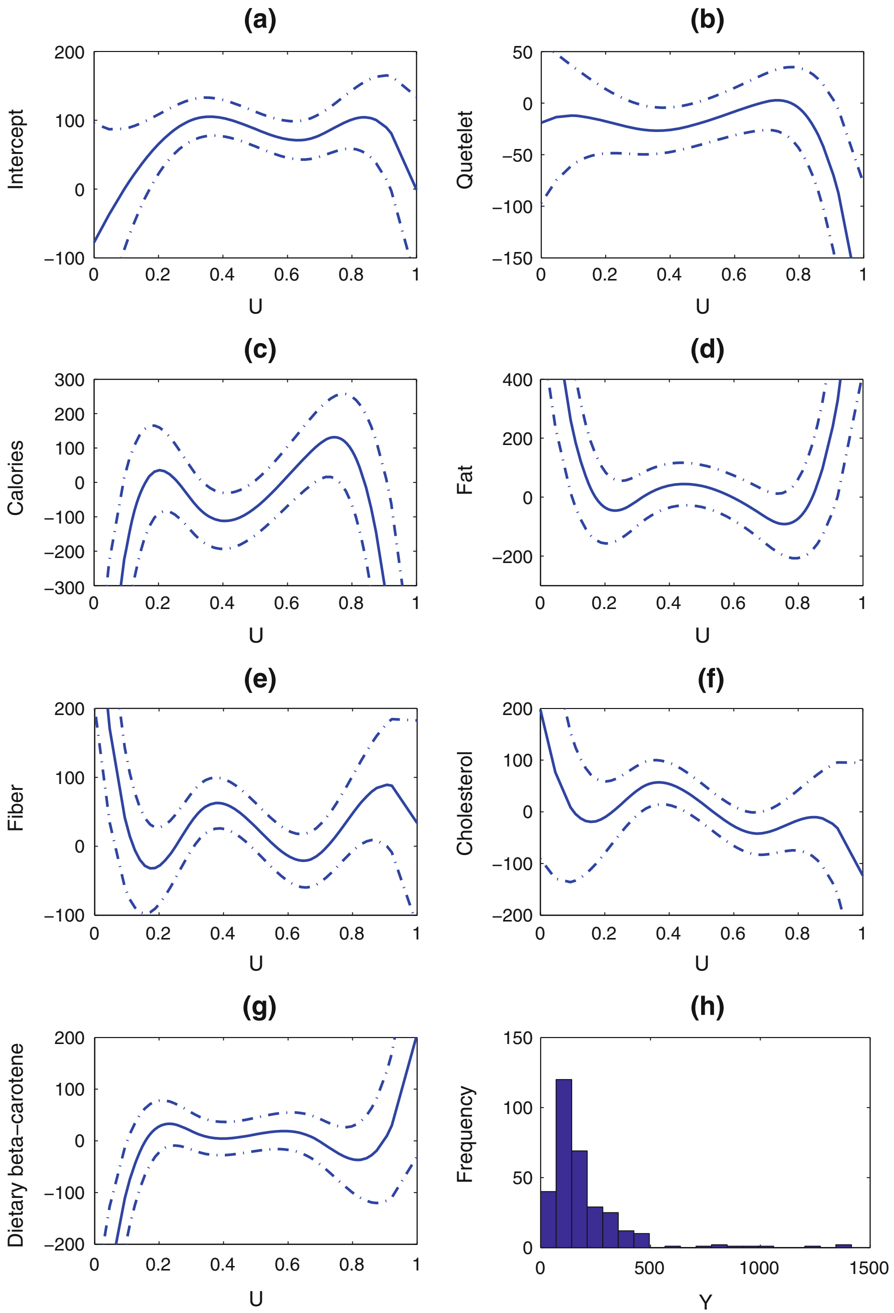


**Fig. 1** Plots of estimated varying coefficient functions with SCAD penalty, the *solid line* is the estimated curve and the *dot-dashed lines* are 95 % pointwise confidence intervals: **a** intercept, **b** quetelet, **c** calories, **d** fat, **e** fiber, **f** cholesterol, **g** dietary beta-carotene. The *histogram* for $Y$ is shown in **h**

**Table 3** Selected parametric components and MAPE with different penalties

| Variable | SCAD | | LASSO | | MCP | |
|---|---|---|---|---|---|---|
| | LSB | SMR | LSB | SMR | LSB | SMR |
| Sex | −40.3798 (19.7231) | 0 (–) | −40.2131 (19.2779) | 0 (–) | −40.3809 (20.0580) | 0 (–) |
| Smoking status (never) | 56.2906 (19.1024) | 46.1699 (10.7457) | 156.0623 (18.3150) | 43.6890 (10.3397) | 56.2903 (19.2835) | 45.8199 (10.6082) |
| Smoking status (former) | 48.5872 (19.3461) | 31.2528 (10.8604) | 48.3586 (18.6002) | 28.4288 (10.4187) | 48.5863 (19.5708) | 30.6077 (10.7295) |
| Vitamin use (fairly often) | 107.9514 (14.2764) | 29.8394 (8.1633) | 107.8889 (14.0674) | 29.7122 (8.0752) | 107.9524 (14.2613) | 30.6979 (8.0708) |
| Vitamin use (not often) | 46.5574 (14.5137) | 34.4075 (8.0915) | 46.4857 (14.2628) | 32.5410 (7.9780) | 46.5579 (14.5752) | 35.3282 (7.9930) |
| Alcohol | 2.5546 (1.3971) | 1.6787 (0.7826) | 2.5494 (1.3819) | 1.6951 (0.7722) | 2.5550 (1.3960) | 1.9908 (0.7921) |
| $\lambda_{opt}$ | 0.0940 | 0.0071 | 0.2381 | 0.0119 | 0.1429 | 0.0095 |
| $h_{opt}$ | – | 95.4157 | – | 95.4327 | – | 95.4120 |
| MAPE | 125.8268 | 111.2356 | 125.7262 | 107.8711 | 125.8276 | 112.1046 |

error distribution. In this paper, we developed an efficient and robust variable selection procedure for SPLVCM based on modal regression method, where the nonparametric functions are approximated by $B$-spline. The proposed procedure can simultaneously estimate and select important variables for both the nonparametric and the parametric part at their best convergence rates. We also established the oracle property for the proposed method. The distinguishing characteristic of newly proposed method is that it introduces an additional tuning parameter $h$ to achieve both robustness and efficiency, which can automatically selected using the observed data. Simulation study and the plasma beta-carotene level data example confirm that the performances of our proposed method outperform than least-square based method.

There is room to improve our method. One limitation is that our proposed variable selection method for SPLVCM is established under the assumption that the varying and constant coefficients can be separated in advance. In fact, we do not know about this prior information when one using SPLVCM to analysis real data, i,e., whether a coefficient is important or not and whether it should be treated as fixed or varying. So, how to simultaneously identify whose predictors are important variables, whose predictors are really varying and whose predictors are only constant effect has been practical interest for researchers, more details see Cheng et al. (2009), Li and Zhang (2011) and Tang et al. (2012). We have embarked some research about it. In addition, one can apply our method to other semiparametric models, such as single-index model, partially linear single-index model and varying coefficient single-index model, to obtain robust and efficient estimation and achieve variable selection. Research in these aspects is ongoing.

**Acknowledgments** We sincerely thank two referees and associate editor for their valuable comments that has led to great improved presentation of our work.

## 7 Appendix

To establish the asymptotic properties of the proposed estimators, the following regularity conditions are needed in this paper. For convenience and simplicity, let $C$ denote a positive constant that may be different at different place throughout this paper.

(C1) The index variable $U$ has a bounded support $\Omega$ and its density function $f_U(\cdot)$ is positive and has a continuous second derivative. Without loss of generality, we assume $\Omega$ be the unit interval [0,1].
(C2) The varying coefficient functions $\alpha_1(u), \ldots, \alpha_p(u)$ are $r$th continuously differentiable on [0,1], where $r > 2$.
(C3) Let $\Sigma_1(u) = \mathrm{E}\{\mathbf{X}\mathbf{X}^T | U = u\}$, $\Sigma_2(u) = \mathrm{E}\{\mathbf{Z}\mathbf{Z}^T | U = u\}$ be continuous with respect to $u$. Furthermore, for given $u$, $\Sigma_1(u)$ and $\Sigma_2(u)$ are positive definite matrix, and their eigenvalues are bounded. In addition, we assume $\max_i \|\mathbf{X}_i\|/\sqrt{n} = o_p(1)$ and $\max_i \|\mathbf{Z}_i\|/\sqrt{n} = o_p(1)$.
(C4) Let $t_1, \ldots, t_K$ be the interior knots of [0,1]. Moreover, let $t_0 = 0$, $t_{K+1} = 1$, $\xi_i = t_i - t_{i-1}$ and $\xi = \max\{\xi_i\}$. Then, there exists a constant $C_0$ such that

$$\frac{\xi}{\min\{\xi_i\}} \leq C_0, \quad \max\{|\xi_{i+1} - \xi_i|\} = o(K^{-1}).$$

(C5) $F(x, z, u, h)$ and $G(x, z, u, h)$ are continuous with respect to $(x, z, u)$.
(C6) $F(x, z, u, h) < 0$ for any $h > 0$.
(C7) $\mathrm{E}(\phi_h'(\varepsilon)|\mathbf{x}, \mathbf{z}, u) = 0$ and $\mathrm{E}(\phi_h''(\varepsilon)^2|\mathbf{x}, \mathbf{z}, u)$, $\mathrm{E}(\phi_h'(\varepsilon)^3|\mathbf{x}, \mathbf{z}, u)$ and $\mathrm{E}(\phi_h'''(\varepsilon)$ $|\mathbf{x}, \mathbf{z}, u)$ are continuous with respect to $x$.
(C8) $\mathrm{liminf}_{n\to\infty}\mathrm{liminf}_{\|\boldsymbol{\gamma}_j\|_H\to 0^+}{\lambda_{1j}}^{-1} p'_{\lambda_{1j}}(\|\boldsymbol{\gamma}_j\|_H) > 0, j = s_1, \ldots, p$, and $\mathrm{liminf}_{n\to\infty}\mathrm{liminf}_{\beta_k\to 0^+}{\lambda_{2k}}^{-1} p'_{\lambda_{2k}}(|\beta_k|) > 0, k = s_2, \ldots, d$.

*Remark 4* The conditions (C1)–(C3) are similar adopted for the SPLVCM, such as in Fan and Huang (2005), Li and Liang (2008) and Zhao and Xue (2009). Condition (C4) implies that $c_0, \ldots, c_{K+1}$ is a $C_0$-quasi-uniform sequence of partitions of [0,1]. (C5)–(C7) are used in modal nonparametric regression in Yao et al. (2012). The condition $\mathrm{E}(\phi_h'(\varepsilon)|\mathbf{x}, \mathbf{z}, u) = 0$ ensures that the proposed estimate is consistent and it is satisfied if the error density is symmetric about zero. However, we do not require the error distribution to be symmetric about zero. If the assumption $\mathrm{E}(\phi_h'(\varepsilon)|\mathbf{x}, \mathbf{z}, u) = 0$ does not hold, the proposed estimate is actually estimating the function $\tilde{m}(\mathbf{x}, \mathbf{z}, u) = \mathrm{argmin}_m \mathrm{E}(\phi_h(Y - m)|\mathbf{x}, \mathbf{z}, u)$. Condition (C8) is the assumption about the penalty function, which is similarly to that used in Fan and Li (2001), Li and Liang (2008) and Zhao and Xue (2009).

## Proof of Theorem 1

*Proof* Let $\delta = n^{-r/(2r+1)} + a_n$ and $\mathbf{v} = (\mathbf{v}_1^T, \mathbf{v}_2^T)^T$ be a vector, where $\mathbf{v}_1$ is $d$-dimension vector and $\mathbf{v}_2$ is $p \times q$-dimension vector, $q = K + \hbar + 1$. Define $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \delta\mathbf{v}_1$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 + \delta\mathbf{v}_2$, where $\boldsymbol{\gamma}_0$ is the best approximation of $\alpha(u)$ in the $B$-spline space. We first show that, for and any given $\varrho > 0$, there exists a large $C$ such that

$$P\left\{\sup_{\|\mathbf{v}\|=C} \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}) < \mathcal{L}(\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0)\right\} \leq 1 - \varrho, \tag{14}$$

where $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is defined in (7). Let $\Xi(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \frac{1}{K}\{\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta}) - \mathcal{L}(\boldsymbol{\gamma}_0, \boldsymbol{\beta}_0)\}$, then by Taylor expansion, we have that

$$\begin{aligned}
\Xi(\boldsymbol{\gamma}, \boldsymbol{\beta}) = &-\frac{\delta}{K}\sum_{i=1}^{n} \phi_h'\left(\varepsilon_i + \mathbf{X}_i^T R(U_i)\right)\left(\mathbf{Z}_i^T \mathbf{v}_1 + \mathbf{W}_i^T \mathbf{v}_2\right) \\
&+ \frac{\delta^2}{K}\sum_{i=1}^{n} \phi_h''\left(\varepsilon_i + \mathbf{X}_i^T R(U_i)\right)\left(\mathbf{Z}_i^T \mathbf{v}_1 + \mathbf{W}_i^T \mathbf{v}_2\right)^2 \\
&+ \frac{\delta^3}{K}\sum_{i=1}^{n} \phi_h'''(\zeta_i)\left(\mathbf{Z}_i^T \mathbf{v}_1 + \mathbf{W}_i^T \mathbf{v}_2\right)^3
\end{aligned}$$

$$+\frac{n}{K}\sum_{j=1}^{p}\left\{p_{\lambda_{1j}}(\|\boldsymbol{\gamma}_j\|_H)-p_{\lambda_{1j}}(\|\boldsymbol{\gamma}_{j0}\|_H)\right\}$$
$$+\frac{n}{K}\sum_{k=1}^{d}\left\{p_{\lambda_{2k}}(|\beta_k|)-p_{\lambda_{2k}}(|\beta_{k0}|)\right\}$$
$$\triangleq I_1+I_2+I_3+I_4+I_5,$$

where $\zeta_i$ is between $\varepsilon_i+\mathbf{X}_i^T R(U_i)$ and $\varepsilon_i+\mathbf{X}_i^T R(U_i)-\delta(\mathbf{Z}_i^T\mathbf{v}_1+\mathbf{W}_i^T\mathbf{v}_2)$,

$$R(u)=(R_1(u),\ldots,R_p(u))^T \quad \text{and} \quad R_j(u)=\alpha_j(u)-B(u)^T\boldsymbol{\gamma}_{j0}, \quad j=1,\ldots,p.$$

By the condition (C1), (C2) and Corollary 6.21 in Schumaker (1981), we have

$$\|R_j(u)\|=O(K^{-r}).$$

Then, by Taylor expansion, we have

$$\sum_{i=1}^{n}\phi_h'\left(\varepsilon_i+\mathbf{X}_i^T R(U_i)\right)\left(\mathbf{Z}_i^T\mathbf{v}_1+\mathbf{W}_i^T\mathbf{v}_2\right)$$
$$=\sum_{i=1}^{n}\left[\phi_h'(\varepsilon_i)+\phi_h''(\varepsilon_i)\mathbf{X}_i^T R(U_i)+\phi_h'''\left(\varepsilon_i^*\right)\left(\mathbf{X}_i^T R(U_i)\right)^2\right]\left(\mathbf{Z}_i^T\mathbf{v}_1+\mathbf{W}_i^T\mathbf{v}_2\right),$$

where $\varepsilon_i^*$ is between $\varepsilon_i$ and $\varepsilon_i+\mathbf{X}_i^T R(U_i)$.

Invoking condition (C4) and (C7), after some direct calculations, we get

$$\sum_{i=1}^{n}\phi_h'\left(\varepsilon_i+\mathbf{X}_i^T R(U_i)\right)\left(\mathbf{Z}_i^T\mathbf{v}_1+\mathbf{W}_i^T\mathbf{v}_2\right)=O_p(nK^{-r}\|\mathbf{v}\|). \tag{15}$$

Hence, we have $I_1=O_p(n\delta K^{-(r+1)}\|\mathbf{v}\|)=O_p(n\delta^2K^{-1}\|\mathbf{v}\|)$.

For $I_2$, we can prove

$$I_2=\mathrm{E}(F(\mathbf{X},\mathbf{Z},U,h))O_p(nK^{-1}\delta^2\|\mathbf{v}\|^2).$$

Therefore, by choosing a sufficiently large $C$, $I_2$ dominates $I_1$ uniformly $\|\mathbf{v}\|=C$. Similarly, we can prove that

$$I_3=O_p(nK^{-1}\delta^3\|\mathbf{v}\|^3).$$

By the condition $a_n\to 0$, hence $\delta\to 0$. It follows that $\delta\|\mathbf{v}\|\to 0$ with $\|\mathbf{v}\|=C$, which lead to $I_3=o_p(J_2)$. Therefore, $I_3$ is also dominated by $I_2$ in $\|\mathbf{v}\|=C$.

Moreover, invoking $p_\lambda(0) = 0$, and by the standard argument of the Taylor expansion, we get that

$$
\begin{aligned}
I_5 \leq& \sum_{k=1}^{s_2} \Big\{ K^{-1} n\delta p'_{\lambda_{2k}}(|\beta_{k0}|)\mathrm{sgn}(\beta_{k0})|v_{1l}| \\
&+ K^{-1} n\delta^2 p'_{\lambda_{2k}}(|\beta_{k0}|)\mathrm{sgn}(\beta_{k0})|v_{1l}|^2(1+o_p(1)) \Big\} \\
\leq& \sqrt{s_2}\left( K^{-1} n\delta a_n\|\mathbf{v}\| + K^{-1} n\delta b_n\|\mathbf{v}\|^2 \right).
\end{aligned}
$$

Then, by the condition $b_n \to 0$, it is easy to show that $I_5$ is dominated by $I_2$ uniformly in $\|\mathbf{v}\| = C$. With the same argument, we can prove that $I_4$ is also dominated by $I_2$ uniformly in $\|\mathbf{v}\| = C$.

By the condition (C6), we know that $F(\mathbf{x}, \mathbf{z}, u, h) < 0$, hence by choosing a sufficiently large $C$, we have $\Xi(\boldsymbol{\gamma}, \boldsymbol{\beta}) < 0$, which implies that with the probability at least $1 - \varrho$, (14) holds. Hence, there exists a local maximizer such that

$$
\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\delta) \quad \text{and} \quad \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(\delta), \tag{16}
$$

which completes the proof of part (i).

Now, we prove part (ii). Note that

$$
\begin{aligned}
\|\hat{\alpha}_j(\cdot) - \alpha_{j0}(\cdot)\|^2 &= \int_0^1 |\hat{\alpha}_j(u) - \alpha_{j0}(u)|^2 \mathrm{d}u \\
&= \int_0^1 \left\{ B^T(u)\hat{\boldsymbol{\gamma}}_k - B^T(u)\boldsymbol{\gamma}_k + R_k(u) \right\}^2 \mathrm{d}u \\
&\leq 2\int_0^1 \left\{ B^T(u)\hat{\boldsymbol{\gamma}}_k - B^T(u)\boldsymbol{\gamma}_k \right\}^2 \mathrm{d}u + 2\int_0^1 R_k(u)^2 \mathrm{d}u \\
&= 2(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k)^T H(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k) + 2\int_0^1 R_k(u)^2 \mathrm{d}u,
\end{aligned}
$$

where $H = \int_0^1 B(u)B^T(u)\mathrm{d}u$. Invoking $\|H\| = O(1)$ and (16), we have

$$
(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k)^T H(\hat{\boldsymbol{\gamma}}_k - \boldsymbol{\gamma}_k) = O_p\left( n^{-\frac{2r}{2r+1}} + a_n^2 \right).
$$

In addition, it is easy to show that

$$
\int_0^1 R_k(u)^2 \mathrm{d}u = O_p\left( n^{-\frac{2r}{2r+1}} \right).
$$

Consequently, $\|\hat{\alpha}_j(\cdot) - \alpha_{j0}\| = O_p\left( n^{\frac{-r}{2r+1}} + a_n \right)$, $j = 1, \ldots, p$, which complete the proof of part (ii). □

**Proof of Theorem 2**

*Proof* By the property of SCAD penalty function, $a_n = 0$ as $\lambda_{\max} \to 0$. Then by Theorem 1, it is sufficient to show that, when $n \to \infty$, for any $\boldsymbol{\gamma}$ that satisfies $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| = O_p(n^{-r/(2r+1)})$, $\beta_k$ that satisfies $\|\beta_k - \beta_{k0}\| = O_p(n^{-r/(2r+1)})$, $k = 1, \ldots, s_2$, and some given small $\nu = Cn^{-r/(2r+1)}$, with probability tending to 1 we have

$$\frac{\partial \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial \beta_k} < 0, \quad \text{for } 0 < \beta_k < \nu,\ k = s_2 + 1, \ldots, d \tag{17}$$

and

$$\frac{\partial \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial \beta_k} > 0, \quad \text{for } -\nu < \beta_k < 0,\ k = s_2 + 1, \ldots, d. \tag{18}$$

Consequently, (17) and (18) imply the maximizer of $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta})$ attains at $\beta_k = 0,\ k = s_2 + 1, \ldots, d$.

By a similar proof of Theorem 1, we can show that

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial \beta_k} &= \frac{\partial Q(\boldsymbol{\gamma}, \boldsymbol{\beta})}{\partial \beta_k} - np'_{\lambda_{2k}}(|\beta_k|)\mathrm{sgn}(\beta_k) \\
&= \sum_{i=1}^{n} Z_{ik}\phi'_h\left(Y_i - \mathbf{W}_i^T\boldsymbol{\gamma} - \mathbf{Z}_i^T\boldsymbol{\beta}\right) - np'_{\lambda_{2k}}(|\beta_k|)\mathrm{sgn}(\beta_k) \\
&= \sum_{i=1}^{n} \Big\{ Z_{ik}\phi'_h\left(\varepsilon_i + \mathbf{X}_i^T R(U_i)\right) - \phi''_h\left(\varepsilon_i + \mathbf{X}_i^T R(U_i)\right) \\
&\quad Z_{ik}\left[\mathbf{W}_i^T(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) + \mathbf{Z}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right] \\
&\quad + \phi'''(\eta_i)Z_{ik}[\mathbf{W}_i^T(\boldsymbol{\gamma} - \boldsymbol{\gamma}_0) + \mathbf{Z}_i^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0)]^2 - np'_{\lambda_{2k}}(|\beta_k|)\mathrm{sgn}(\beta_k)\Big\} \\
&= n\lambda_{2k}\left\{\lambda_{2k}^{-1}p'_{\lambda_{2k}}(|\beta_k|)\mathrm{sgn}(\beta_k) + O_p(\lambda_{2k}^{-1}n^{-\frac{r}{2r+1}})\right\},
\end{aligned}$$

where $\eta_i$ is between $Y_i - \mathbf{W}_i^T\boldsymbol{\gamma} - \mathbf{Z}_i^T\boldsymbol{\beta}$ and $\varepsilon_i + \mathbf{X}_i^T R(U_i)$.

By the condition (C8), $\liminf_{n\to\infty}\liminf_{\beta_k\to 0^+}\lambda_{2k}{}^{-1}p'_{\lambda_{2k}}(|\beta_k|) > 0$, and $\lambda_{2k}n^{\frac{r}{2r+1}} > \lambda_{\min}n^{\frac{r}{2r+1}} \to \infty$, the sign of the derivation is completely determined by that of $\beta_k$, then (17) and (18) hold. This completes the proof of part (i).

For part (ii), apply the similar techniques as in part (i), we have, with probability tending to 1, that $\hat{\alpha}_j(\cdot) = 0$, $j = s_1 + 1, \ldots, p$. Invoking $\sup_u \|B(u)\| = O(1)$, the result is achieved from $\hat{\alpha}_j(u) = B(u)^T\hat{\boldsymbol{\gamma}}_j$. □

**Proof of Theorem 3**

*Proof* From Theorems 1 and 2, we know that, as $n \to \infty$, with probability tending to 1, $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\beta})$ attains the maximal value at $(\hat{\boldsymbol{\beta}}_a^T, 0)^T$ and $(\hat{\boldsymbol{\gamma}}_a^T, 0)^T$. Let $\mathcal{L}_1(\boldsymbol{\gamma}, \boldsymbol{\beta}) =$

$\partial\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\beta})/\partial\boldsymbol{\beta}_a$ and $\mathcal{L}_2(\boldsymbol{\gamma},\boldsymbol{\beta}) = \partial\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\beta})/\partial\boldsymbol{\gamma}_a$, then $(\hat{\boldsymbol{\beta}}_a^T, 0)^T$ and $(\hat{\boldsymbol{\gamma}}_a^T, 0)^T$ must satisfy following two equations

$$\begin{aligned}\frac{1}{n}\mathcal{L}_1&\left(\left(\hat{\boldsymbol{\gamma}}_a^T,0\right)^T,\left(\hat{\boldsymbol{\beta}}_a^T,0\right)^T\right)\\ &=\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_{ia}\phi_h'\left\{Y_i-\mathbf{W}_{ia}^T\hat{\boldsymbol{\gamma}}_a-\mathbf{Z}_{ia}^T\hat{\boldsymbol{\beta}}_a\right\}-p_{\lambda_2}'(|\hat{\boldsymbol{\beta}}_a|)\circ\operatorname{sgn}(\hat{\boldsymbol{\beta}}_a)=0\end{aligned} \tag{19}$$

and

$$\frac{1}{n}\mathcal{L}_2\left(\left(\hat{\boldsymbol{\gamma}}_a^T,0\right)^T,\left(\hat{\boldsymbol{\beta}}_a^T,0\right)^T\right)=\frac{1}{n}\sum_{i=1}^n \mathbf{W}_{ia}\phi_h'\left\{Y_i-\mathbf{W}_{ia}^T\hat{\boldsymbol{\gamma}}_a-\mathbf{Z}_{ia}^T\hat{\boldsymbol{\beta}}_a\right\}-\boldsymbol{\kappa}=0, \tag{20}$$

where "$\circ$" denotes the Hadamard (componentwise) product and the $k$th component of $p_{\lambda_2}'(|\hat{\boldsymbol{\beta}}_a|)$ is $p_{\lambda_{2k}}'(|\hat{\beta}_k|)$, $1\le k\le s_1$; $\boldsymbol{\kappa}$ is a $q\times s_1$-dimensional vector with its $j$th block subvector being $H\frac{\hat{\boldsymbol{\gamma}}_j}{\|\hat{\boldsymbol{\gamma}}_j\|_H}p_{\lambda_1}'(\|\hat{\boldsymbol{\gamma}}_j\|_H)$. Applying the Taylor expansion to $p_{\lambda_{2k}}'(|\hat{\beta}_k|)$, we get that

$$p_{\lambda_{2k}}'(|\hat{\beta}_k|)=p_{\lambda_{2k}}'(|\hat{\beta}_{k0}|)+\{p_{\lambda_{2k}}''(|\hat{\beta}_{k0}|)+o_p(1)\}(\hat{\beta}_k-\beta_{k0}),\quad k=1,\ldots,s_2.$$

By the condition $b_n\to 0$ and note that $p_{\lambda_{2k}}'(|\hat{\beta}_{k0}|)=0$ as $\lambda_{\max}\to 0$, some simple calculations yields

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\mathbf{Z}_{ia}&\left\{\phi_h'(\varepsilon_i)+\phi_h''(\varepsilon_i)\left\{\mathbf{X}_i^TR^*(U_i)-\left[\mathbf{Z}_{ia}^T(\hat{\boldsymbol{\beta}}_a-\boldsymbol{\beta}_{a0})+\mathbf{W}_{ia}^T(\hat{\boldsymbol{\gamma}}_a-\boldsymbol{\gamma}_{a0})\right]\right\}\right.\\ &\left.+\phi_h'''(\zeta_i)\left\{\mathbf{X}_i^TR^*(U_i)-\left[\mathbf{Z}_{ia}^T(\hat{\boldsymbol{\beta}}_a-\boldsymbol{\beta}_{a0})+\mathbf{W}_{ia}^T(\hat{\boldsymbol{\gamma}}_a-\boldsymbol{\gamma}_{a0})\right]\right\}^2\right\}+o_p(\hat{\boldsymbol{\beta}}_a-\boldsymbol{\beta}_{a0})=0,\end{aligned} \tag{21}$$

where $\zeta_i$ is between $\varepsilon_i$ and $Y_i-\mathbf{W}_{ia}^T\hat{\boldsymbol{\gamma}}_a-\mathbf{Z}_{ia}^T\hat{\boldsymbol{\beta}}_a$, $R^*(u)=(R_1(u),\ldots,R_{s_1}(u))^T$. Invoking (20), and using the similar arguments to (21), we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\mathbf{W}_{ia}&\left\{\phi_h'(\varepsilon_i)+\phi_h''(\varepsilon_i)\left\{\mathbf{X}_i^TR^*(U_i)-\left[\mathbf{Z}_{ia}^T(\hat{\boldsymbol{\beta}}_a-\boldsymbol{\beta}_{a0})+\mathbf{W}_{ia}^T(\hat{\boldsymbol{\gamma}}_a-\boldsymbol{\gamma}_{a0})\right]\right\}\right.\\ &\left.+\phi_h'''(\bar{\zeta}_i)\left\{\mathbf{X}_{ia}^TR^*(U_i)-\left[\mathbf{Z}_{ia}^T(\hat{\boldsymbol{\beta}}_a-\boldsymbol{\beta}_{a0})+\mathbf{W}_{ia}^T(\hat{\boldsymbol{\gamma}}_a-\boldsymbol{\gamma}_{a0})\right]\right\}^2\right\}+o_p(\hat{\boldsymbol{\gamma}}_a-\boldsymbol{\gamma}_{a0})=0,\end{aligned} \tag{22}$$

where $\bar{\zeta}_i$ is also between $\varepsilon_i$ and $Y_i-\mathbf{W}_{ia}^T\hat{\boldsymbol{\gamma}}_a-\mathbf{Z}_{ia}^T\hat{\boldsymbol{\beta}}_a$.

Let $\Phi_n = \frac{1}{n}\sum_{i=1}^n \phi''(\varepsilon_i)\mathbf{W}_{ia}\mathbf{W}_{ia}^T$ and $\Psi_n = \frac{1}{n}\sum_{i=1}^n \phi''(\varepsilon_i)\mathbf{W}_{ia}\mathbf{Z}_{ia}^T$, then, by the result of Theorem 2 and regularity conditions (C3) and (C7), after some calculations based on (22), it follows that

$$\hat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_{a0} = (\Phi_n + o_p(1))^{-1}\left\{-\Psi_n(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{a0}) + \Lambda_n\right\}, \tag{23}$$

where $\Lambda_n = \frac{1}{n}\sum_{i=1}^n \mathbf{W}_{ia}\left[\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)\mathbf{X}_{ia}^T R^*(U_i)\right]$. Furthermore, we can prove

$$\Phi_n \xrightarrow{\mathrm{P}} \Phi = \mathrm{E}\left(F(\mathbf{X},\mathbf{Z},U,h)\mathbf{W}_a\mathbf{W}_a^T\right) \quad \text{and} \quad \Psi_n \xrightarrow{\mathrm{P}} \Psi = \mathrm{E}\left(F(\mathbf{X},\mathbf{Z},U,h)\mathbf{W}_a\mathbf{Z}_a^T\right).$$

Therefore, we can write

$$\hat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_{a0} = -(\Phi + o_p(1))^{-1}\Psi(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{a0}) + (\Phi + o_p(1))^{-1}\Lambda_n. \tag{24}$$

Substituting (24) into (21), we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \phi_h''(\varepsilon_i)\mathbf{Z}_{ia}\left[\mathbf{Z}_{ia} - \Psi^T\Phi^{-1}\mathbf{W}_{ia}\right]^T(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{a0}) + o_p(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{a0})\\
&\quad = \frac{1}{n}\sum_{i=1}^n \mathbf{Z}_{ia}\left[\phi_h'(\varepsilon_i) + \phi_h''(\varepsilon_i)\mathbf{X}_{ia}^T R^*(U_i) - \phi_h''(\varepsilon_i)\mathbf{W}_{ia}^T\frac{1}{n}\sum_{j=1}^n \Phi^{-1}\mathbf{W}_{ja}\phi_h'(\varepsilon_j)\right]\\
&\qquad -\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_{ia}\phi_h''(\varepsilon_i)\mathbf{W}_{ia}^T\frac{1}{n}\sum_{j=1}^n \mathbf{X}_{ja}^T R^*(U_j).
\end{aligned} \tag{25}$$

Note that

$$\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^n \phi_h''(\varepsilon_i)\Psi^T\Phi^{-1}\mathbf{W}_{ia}\left[\mathbf{Z}_{ia}^T - \mathbf{W}_{ia}^T\Phi^{-1}\Psi\right]\right) = 0$$

and

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n \phi_h''(\varepsilon_i)\Psi^T\Phi^{-1}\mathbf{W}_{ia}\left[\mathbf{Z}_{ia}^T - \mathbf{W}_{ia}^T\Phi^{-1}\Psi\right]\right) = o_p(1/n).$$

Hence, it is easy to show that

$$\begin{aligned}
&\left\{\frac{1}{n}\sum_{i=1}^n \phi_h''(\varepsilon_i)\check{\mathbf{Z}}_{ia}\check{\mathbf{Z}}_{ia}^T + o_p(1)\right\}\sqrt{n}(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{a0})\\
&\quad = \frac{1}{\sqrt{n}}\sum_{i=1}^n \check{\mathbf{Z}}_{ia}\phi_h'(\varepsilon_i) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \check{\mathbf{Z}}_{ia}\phi_h''(\varepsilon_i)\mathbf{X}_{ia}^T R^*(U_i) \triangleq J_1 + J_2.
\end{aligned} \tag{26}$$

By the definition of $R^*(U_i)$, we can prove $J_2 = o_p(1)$. Moreover, we have

$$\frac{1}{n}\sum_{i=1}^{n}\phi_h''(\varepsilon_i)\check{\mathbf{Z}}_{ia}\check{\mathbf{Z}}_{ia}^T \xrightarrow{\text{P}} \Sigma.$$

It remains to show that

$$J_1 \xrightarrow{\text{d}} N(0, \Delta), \tag{27}$$

where $\Delta = \mathrm{E}(G(\mathbf{X}, \mathbf{Z}, U, h)\check{\mathbf{Z}}_a\check{\mathbf{Z}}_a^T)$.

Then, combine (26) and (27) and use the Slutsky's theorem, it follows that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{a0}) \xrightarrow{\text{d}} N(0, \Sigma^{-1}\Delta\Sigma^{-1}).$$

Next, we prove (27). Note that for any vector $\boldsymbol{\varsigma}$ whose components are not all zero,

$$\boldsymbol{\varsigma}^T J_1 = \sum_{i=1}^{n}\frac{1}{\sqrt{n}}\boldsymbol{\varsigma}^T\check{\mathbf{Z}}_{ia}\phi_h'(\varepsilon_i) = \sum_{i=1}^{n} a_i\xi_i,$$

where $a_i^2 = \frac{1}{n}G(\mathbf{X}_i, \mathbf{Z}_i, U_i, h)\boldsymbol{\varsigma}^T\check{\mathbf{Z}}_{ia}\check{\mathbf{Z}}_{ia}^T\boldsymbol{\varsigma}$ and, conditioning on $\{\mathbf{X}_i, \mathbf{Z}_i, U_i\}$, $\xi_i$ are independent with mean zero and variance one. It follows easily by checking Lindeberg condition that if

$$\frac{\max_i a_i^2}{\sum_{i=1} a_i^2} \xrightarrow{\text{P}} 0, \tag{28}$$

then $\sum_{i=1}^{n} a_i\xi_i/\sqrt{\sum_{i=1}^{n} a_i^2} \xrightarrow{\text{d}} N(0, 1)$. Thus, we can conclude that (27) holds.

Now, we only need to show (28) holds. Noting that $(\boldsymbol{\varsigma}^T\check{\mathbf{Z}}_{ia})^2 \le \|\boldsymbol{\varsigma}\|^2\|\check{\mathbf{Z}}_{ia}\|^2$, hence $a_i^2 \le \frac{1}{n}G(\mathbf{X}_i, \mathbf{Z}_i, U_i, h)\|\boldsymbol{\varsigma}\|^2\|\check{\mathbf{Z}}_{ia}\|^2$. Since

$$\|\check{\mathbf{Z}}_{ia}\| = \|\mathbf{Z}_{ia} - \Psi^T\Phi^{-1}\mathbf{W}_{ia}\| \le \|\mathbf{Z}_{ia}\| + \|\Psi^T\Phi^{-1}\mathbf{W}_{ia}\|,$$

and by the conditions $\max_i \|\mathbf{X}_i\|/\sqrt{n} = o_p(1)$ and $\max_i \|\mathbf{Z}_i\|/\sqrt{n} = o_p(1)$ in (C3), using the property of spline basis (Schumaker 1981) and the definition $\mathbf{W}_{ia} = I_p \otimes B(U_i) \cdot \mathbf{X}_{ia}$ together with the conditions (C5) and (C7), we can prove

$$\max_i \|\check{\mathbf{Z}}_i\|/\sqrt{n} = o_p(1).$$

Applying the Slutsky's theorem, (28) holds obviously, which complete the proof of Theorem 3. □

### Proof of Theorem 4

*Proof* According to the Eq. (24) and the asymptotic normality of $\hat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}_{a0}$ in Theorem 3, for any vector $\mathbf{d}_n$ with dimension $q \times s_1$ and components not all 0, by the conditions (C1)–(C5) and (C7) and use the Slutsky's theorem and the property of multivariate normal distribution, it follows that

$$\left\{\mathbf{d}_n^T \operatorname{var}(\hat{\boldsymbol{\gamma}}_a)\mathbf{d}_n\right\}^{-1/2} \mathbf{d}_n^T(\hat{\boldsymbol{\gamma}}_a - \boldsymbol{\gamma}_{a0}) \xrightarrow{\mathrm{d}} N(0, 1),$$

where

$$\operatorname{var}(\hat{\boldsymbol{\gamma}}_a) = \Phi^{-1}\Psi\frac{\Sigma^{-1}\Delta\Sigma^{-1}}{n}\Psi^T\Phi^{-1}.$$

For any $q \times s_1$-vector $\mathbf{c}_n$ whose components are not all 0, by the definition of $\hat{\boldsymbol{\alpha}}_a$ and $\tilde{\boldsymbol{\alpha}}_a$, choosing $\mathbf{d}_n = \mathbf{W}_a^T\mathbf{c}_n$ yields

$$\left\{\mathbf{c}_n^T \operatorname{var}(\hat{\boldsymbol{\alpha}}_a(u))\mathbf{c}_n\right\}^{-1/2} \mathbf{c}_n^T(\hat{\boldsymbol{\alpha}}_a(u) - \tilde{\boldsymbol{\alpha}}_a(u)) \xrightarrow{\mathrm{d}} N(0, 1).$$

□

## References


Cai, Z., Xiao, Z. (2012). Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. *Journal of Econometrics*, *167*, 413–425.

Cai, Z., Fan, J., Li, R. (2000). Efficient estimation and inference for varying-coefficient models. *Journal of the American Statistical Association*, *95*, 888–902.

Candes, E., Tao, T. (2007). The Dantzig selector: statistical estimation when $p$ is much larger than $n$. *The Annals of Statistics*, *35*, 2313–2351.

Cheng, M., Zhang, W., Chen, L. (2009). Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association*, *104*, 1179–1191.

Fairfield, K., Fletcher, R. (2002). Vitamins for chronic disease prevention in adults: scientific review. *The Journal of the American Medical Association*, *287*, 3116–3126.

Fan, J., Gijbels, I. (1996). *Local polynomial modelling and its application*. New York: Chapman and Hall.

Fan, J., Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, *11*, 1031–1057.

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., Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, *27*, 1491–1518.

Fan, J., Zhang, W. (2000). Simultaneous confidence bands and hypotheses testing in varying-coefficient models. *Scandinavian Journal of Statistics*, *27*, 715–731.

Hastie, T., Tibshirani, R. (1993). Varying-coefficient model. *Journal of the Royal Statistical Society, Series B*, *55*, 757–796.

Huang, J., Wu, C., Zhou, L. (2002). Varying-coefficient models and basis function approximation for the analysis of repeated measurements. *Biometrika*, *89*, 111–128.

Kai, B., Li, R., Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics*, *39*, 305–332.

Lam, C., Fan, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *The Annals of Statistics*, *36*, 2232–2260.

Lee, M. (1989). Mode regression. *Journal of Econometrics*, *42*, 337–349.
Leng, C. (2009). A simple approach for varying-coefficient model selection. *Journal of Statistical Planning and Inference*, *139*, 2138–2146.
Li, J., Palta, M. (2009). Bandwidth selection through cross-validation for semi-parametric varying-coefficient partially linear models. *Journal of Statistical Computation and Simulation*, *79*, 1277–1286.
Li, J., Zhang, W. (2011). A semiparametric threshold model for censored longitudinal data analysis. *Journal of the American Statistical Association*, *106*, 685–696.
Li, J., Ray, S., Lindsay, B. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, *8*, 1687–1723.
Li, Q., Huang, C., Li, D., Fu, T. (2002). Semiparametric smooth coefficient models. *Journal of Business and Economic Statistics*, *3*, 412–422.
Li, R., Liang, H. (2008). Variable selection in semiparametric regression modeling. *The Annals of Statistics*, *36*, 261–286.
Lin, Z., Yuan, Y. (2012). Variable selection for generalized varying coefficient partially linear models with diverging number of parameters. *Acta Mathematicae Applicatae Sinica, English Series*, *28*, 237–246.
Lu, Y. (2008). Generalized partially linear varying-coefficient models. *Journal of Statistical Planning and Inference*, *138*, 901–914.
Nierenberg, D., Stukel, T., Baron, J., Dain, B., Greenberg, E. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, *130*, 511–521.
Schumaker, L. (1981). *Splines function: basic theory*. New York: Wiley.
Scott, D. (1992). *Multivariate density estimation: theory, practice and visualization*. New York: Wiley.
Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, *10*, 1040–1053.
Tang, Y., Wang, H., Zhu, Z., Song, X. (2012). A unified variable selection approach for varying coefficient models. *Statistica Sinica*, *22*, 601–628.
Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.
Wang, H., Zhu, Z., Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *The Annals of Statistics*, *37*, 3841–3866.
Wang, L., Li, H., Huang, J. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, *103*, 1556–1569.
Xia, Y., Zhand, W., Tong, H. (2004). Efficient estimation for semivarying-coefficient models. *Biometrika*, *91*, 661–681.
Xie, H., Huang, J. (2009). SCAD-penalized regression in high-dimensional partially linear models. *The Annals of Statistics*, *37*, 673–696.
Yao, W., Li, L. (2011). A new regression model: modal linear regression. Technical report, Kansas State University, Manhattan. http://www-personal.ksu.edu/~wxyao/
Yao, W., Lindsay, B., Li, R. (2012). Local modal regression. *Journal of Nonparametric Statistics*, *24*, 647–663.
Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*, 894–942.
Zhang, W., Lee, S., Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis*, *82*, 166–188.
Zhao, P., Xue, L. (2009). Variable selection for semiparametric varying coefficient partially linear models. *Statistics and Probability Letters*, *79*, 2148–2157.
Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.
Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, *67*, 301–320.
Zou, H., Li, R. (2008). One-step sparse estimates in nonconcave penalized like-lihood models (with discussion). *The Annals of Statistics*, *36*, 1509–1533.