Annals of the Institute of Statistical Mathematics (2021) 73:457–480
https://doi.org/10.1007/s10463-020-00757-0

Model identification and selection for single-index varying-coefficient models

Peng Lai[1] · **Fangjian Wang**[1] · **Tingyu Zhu**[2] · **Qingzhao Zhang**[3]

Received: 9 January 2018 / Revised: 5 March 2020 / Published online: 10 June 2020

Abstract

Single-index varying-coefficient models include many types of popular semiparametric models, i.e., single-index models, partially linear models, varying coefficient models, and so on. In this paper, a two-stage efficient variable selection procedure is proposed to select important nonparametric and parametric components and obtain estimators simultaneously. We also find that the proposed procedure can separate predictors into varying-coefficient and constant-coefficient predictors automatically. Theoretically, it has the selection and estimation consistency properties. Simulation studies and a real data application are conducted to evaluate and illustrate the proposed methods.

Keywords Efficient estimating equation · Group LASSO · Single-index varying-coefficient model · Variable selection

1 Introduction

Consider a single-index varying-coefficient model of the form

$$Y = g^{\top}(X^{\top}\beta)Z + \varepsilon, \tag{1}$$

where $X \in R^p$ and $Z \in R^q$ are vectors of covariates, Y is the response variable, β is a $p \times 1$ vector of unknown parameters with $\|\beta\| = 1$ and its first component being positive for the sake of identifiability ($\|\cdot\|$ denotes the Euclidean metric),

✉ Qingzhao Zhang
qzzhang@xmu.edu.cn

[1] School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China

[2] Department of Statistics, Oregon State University, Corvallis, OR 97331, USA

[3] Department of Statistics, School of Economics, The Wang Yanan Institute for Studies in Economics, MOE Key Lab of Economics and Fujian Key Lab of Statistics, Xiamen University, Xiamen 361005, China

$g(\cdot) = (g_1(\cdot), \ldots, g_q(\cdot))^\top$ is a $q \times 1$ vector of unknown functions and ε is a random error with $E(\varepsilon|X, Z) = 0$ and $Var(\varepsilon|X, Z) = \sigma^2 < \infty$. Model (1) includes many important statistical models such as the linear regression model, varying-coefficient model and single-index model. More details refer to Xue and Wang (2012) and Lai et al. (2016).

In this work, we are interested in estimating parametric coefficients β and functions $g(\cdot)$, where β and $g(\cdot)$ are sparse in the sense that some of their elements are zero, and some $g_k(\cdot)$'s may be nonzero constants. Sparsity plays a crucial role in high dimensional analysis, as it can improve interpretability and the accuracy of prediction. In addition, separation of the varying and constant effects have important implications, for example, in gene-environment interaction studies (Wu et al. 2014, 2015, 2018). Many studies have investigated statistical inference for single-index varying-coefficient models, such as Xue and Wang (2012), Xue and Pang (2013), Huang and Zhang (2013), and so on. However, these methods give nonzero estimates to all coefficients. Various penalization methods that can automatically select relevant parameters and simultaneously estimate them have been developed. Examples include LASSO (Tibshirani 1996), SCAD (Fan and Li 2001), bridge (Huang et al. 2008), and so on. Most recently, variable selection using penalty functions for nonparametric or semiparametric models has been developed, see Ma and Du (2012) and Huang et al. (2012) for example.

The aforementioned works motivate us to develop a penalization-based approach for variable selection and identification in the single-index varying coefficient model. There are two related studies on this topic: Feng and Xue (2015) and Song et al. (2016). However, the nonparametric parts in both studies are approximated by the B-spline basis functions. As Fan and Zhang (2008) notes, varying coefficient models, as they stand, are locally linear models. It is more reasonable to use the kernel smoothing method for estimation. In this paper, we propose a two-stage efficient variable selection procedure based on the local linear smooth technique and efficient estimating equation for the single-index varying coefficient model. It is noted that our approach is different from Lai et al. (2014), which focuses on the partially linear single-index models and performs variable selection only on parameters without identification of the nonparametric function.

The rest of this paper is organized as follows. A two-stage efficient variable selection procedure is developed and its theoretical properties are carefully studied in Sect. 2. Numerical studies are reported in Sect. 3. The article is concluded with a brief discussion in Sect. 4. All technical details are delayed to "Appendix."

2 A two-stage efficient variable selection procedure

In practice, we often face problems with high dimensions. In order to enhance the predictability and interpretability of the model, sparse modeling assumes that many covariates in the studied model are not relevant. This motivates us to apply penalty methods to simultaneously estimate parameters and select the relevant ones.

In the following, our goal is to not only select the significant variables in X and Z but also identify which components of $g(\cdot)$ are constants or varying coefficients, through the following two stages.

- *Stage 1* For a given β, construct the local least squares penalized loss function to select the important components in Z, and meanwhile, identify constant coefficients and varying coefficients in the components of $g(\cdot)$. Here β can be obtained by approaches in Xue and Wang (2012), Xue and Pang (2013), Huang and Zhang (2013), or Lai et al. (2016), which are all $\sqrt{n}$ consistent. In our numerical studies, we adopt Lai et al. (2016)'s estimate.
- *Stage 2* Construct the penalized efficient estimating equations for parametric components based on the estimated $\hat{g}_\lambda(\cdot)$ and $\hat{g}'_\lambda(\cdot)$ from Stage 1. Select the relevant single-index variables of X and obtain their estimated coefficients.

There is no need to iterate for the proposed two-stage approach. The existing studies have shown that the techniques in each stage have satisfactory convergence properties. In the following, we first study the variable selection procedure for nonparametric components.

2.1 Stage 1: variable selection for nonparametric components

For model (1), it is easy to see that for a given β, we can employ the local linear regression technique (Fan and Gijbels 1996) to estimate $g(\cdot)$ and $g'(\cdot)$. The local linear estimators for $g(t)$ and $g'(t)$ are defined as $\hat{g}(t) = \hat{a}$ and $\hat{g}'(t) = \hat{b}$, where $\hat{a}$ and $\hat{b}$ minimize the sum of weighted squares

$$\sum_{i=1}^{n} \left[Y_i - \{a + b(X_i^\top \beta - t)\}^\top Z_i\right]^2 K_h(X_i^\top \beta - t),$$

with respect to a and b, $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function, and h is the bandwidth. Let $t_j = X_j^\top \beta$, $G = (g(t_1), \dots, g(t_n))$ and $G' = (g'(t_1), \dots, g'(t_n))$. Then, $\hat{G}_M = (\hat{G}^\top, h\hat{G}'^\top)^\top$ is a natural estimator for $G_M = (G^\top, hG'^\top)^\top$, which can be obtained by minimizing the following global least squares function

$$Q(G_{\mathrm{ML}}) = \sum_{j=1}^{n}\sum_{i=1}^{n} \left[Y_i - a^\top(t_j)Z_i - (hb(t_j))^\top Z_i\left(\frac{X_i^\top \beta - t_j}{h}\right)\right]^2 K_h(X_i^\top \beta - t_j) \quad (2)$$

with respect to $G_{\mathrm{ML}} = (G_a^\top, hG_b^\top)^\top$, where $G_a = (a(t_1), \dots, a(t_n)) \in R^{q\times n}$ and $G_b = (b(t_1), \dots, b(t_n)) \in R^{q\times n}$.

Denote $c_k = (a_k(t_1), a_k(t_2), \dots, a_k(t_n))^\top$ and $d_k = (b_k(t_1), b_k(t_2), \dots, b_k(t_n))^\top$, which correspond to the kth column of G_a and G_b, respectively. Inspired by Wang and Xia (2009), we propose the following penalized estimator

$$\hat{G}_{M_\lambda} = (\hat{G}_\lambda^\top, h\hat{G}_\lambda'^\top)^\top = \mathop{\arg\min}_{G_{\mathrm{ML}} \in R^{2q\times n}} Q_\lambda(G_{\mathrm{ML}}),$$

where

$$Q_\lambda(G_{\mathrm{ML}}) = Q(G_{\mathrm{ML}}) + \sum_{k=1}^{q} \lambda_{1k}\|c_k\| + \sum_{k=1}^{q} \lambda_{2k}\|d_k\|. \tag{3}$$

The tuning parameters λ_{1k} and $\lambda_{2k}, k = 1, \dots, q$ are discussed later in Sect. 2.3.

To solve the above minimization problem, we describe here an easy implementation based on the idea of the local quadratic approximation (Fan and Li 2001). Our implementation is based on an iterative algorithm with $\hat{G}_M$ from (2) (i.e., the unpenalized estimator) as the initial value. The objective function (3) can be locally approximated by

$$Q(G_{\mathrm{ML}}) + \sum_{j=1}^{n}\sum_{k=1}^{q} \frac{\lambda_{1k} a_k^2(t_j)}{\|\hat{c}_{\lambda k}^{(m)}\|} + \sum_{j=1}^{n}\sum_{k=1}^{q} \frac{\lambda_{2k} b_k^2(t_j)}{\|\hat{d}_{\lambda k}^{(m)}\|},$$

where $\hat{c}_{\lambda k}^{(m)}$ and $\hat{d}_{\lambda k}^{(m)}$ are obtained in the mth iteration. Let $\eta(t) = (a(t)^\top, hb(t)^\top)^\top$ and $\tilde{Z}_i(t) = (Z_i^\top, \frac{t_i - t}{h} Z_i^\top)^\top$. For any t, we can define the $(m+1)$th iteration estimator by

$$\hat{\eta}_\lambda^{(m+1)}(t) = \left[\sum_{i=1}^{n} \tilde{Z}_i(t)\tilde{Z}_i^\top(t) K_h(t_i - t) + D^{(m)}\right]^{-1} \left[\sum_{i=1}^{n} Y_i \tilde{Z}_i(t) K_h(t_i - t)\right]. \tag{4}$$

where $D^{(m)} = \operatorname{diag}\,(\frac{\lambda_{11}}{\|\hat{c}_{\lambda 1}^{(m)}\|}, \dots, \frac{\lambda_{1q}}{\|\hat{c}_{\lambda q}^{(m)}\|}, \frac{\lambda_{21}}{\|\hat{d}_{\lambda 1}^{(m)}\|}, \dots, \frac{\lambda_{2q}}{\|\hat{d}_{\lambda q}^{(m)}\|})$ is a $2q \times 2q$ diagonal matrix. In practice, we see that the absolute values of some coefficients get smaller and smaller with iterations. The coefficients are regarded as zero if the absolute values are less than a certain threshold (10^{-3} in our numerical studies).

Next we establish the consistency of the nonparametric component selection. Assume that the number of significant variables in Z is $q_0, q_0 \le q$. Furthermore, some components of $g(\cdot)$ may be nonzero constants. Assume that the number of varying coefficient components is d_0. Then the number of nonzero components of $g'(t)$ is d_0, $d_0 \le q_0$. Without loss of generality, assume that the last $(q - q_0)$ rows of G and G' are zero, the first d_0 components are varying coefficients and the following $(q_0 - d_0)$ components are nonzero constants. Denote the subscript set of nonzero components as $\mathbf{A}_g^* = \{1, \dots, q_0\}$, and the subscript set of nonzero varying coefficients as $\mathbf{A}_{g'}^* = \{1, \dots, d_0\}$. Let $\phi_n = \max\{\lambda_{1j}, 1 \le j \le q_0\}$, $\phi_n' = \max\{\lambda_{2j}, 1 \le j \le q_0\}$, $\varphi_n = \min\{\lambda_{1j}, q_0 + 1 \le j \le q\}$, $\varphi_n' = \min\{\lambda_{2j}, q_0 + 1 \le j \le q\}$, and $\psi_n = \min\{\lambda_{2j}, d_0 + 1 \le j \le q_0\}$. The following theorem shows the selection consistency for the nonparametric part. Suppose that β lies in a small neighbor of β_0: $\mathbf{B}_n = \{\beta : \|\beta - \beta_0\| \le Cn^{-1/2}\}$ (at the usual parametric rate), where C is a positive constant. Some standard assumptions are imposed, which are similar to those in Wang and Xia (2009).

- *C1* The density function $f(t)$ of $X^\top\beta$ is bounded away from zero and infinity on $\mathbf{T} = \{t : t = X^\top\beta, X \in \mathbf{A}_X, \beta \in \mathbf{B}_n\}$, where $\mathbf{A}_X$ is the compact support of X. Moreover, $f(t)$ has continuous derivatives up to order two on $\mathbf{T}$.
- *C2* The coefficients $g_j(\cdot), j = 1, \dots, q$, are both twice continuously differentiable.

- *C3* The kernel K is a bounded and symmetric probability density function, satisfying

$$\int_{-\infty}^{\infty} u^2 K(u)\mathrm{d}u \neq 0, \quad \int_{-\infty}^{\infty} |u|^i K(u)\mathrm{d}u < \infty, \quad i = 1, 2, \ldots.$$

- *C4* For any $s > 2$, $E|Y|^{2s} < \infty$, $E|Z_j|^{2s} < \infty, j = 1, \ldots, q$ with $Z = (Z_1, \ldots, Z_q)^\top$.
- C5. The function $D_1(t) = E[ZZ^\top|t]$ is nonsingular and twice continuously differentiable with bounded derivatives. Function $E(\|Z\|^4|t)$ is also bounded.

Theorem 1 *Under conditions C1–C5, if* $h = O_p(n^{-1/5})$, $n^{-11/10}\phi_n \to 0$, $n^{-11/10}\phi'_n \to 0, n^{-11/10}\varphi_n \to \infty, n^{-11/10}\varphi'_n \to \infty$, *and* $n^{-9/10}\psi_n \to \infty$, *we have*

(i) $P\big(\sup_{t\in T} \|\hat{g}_{\lambda j}(t)\| = 0\big) \to 1$ *for any* $q_0 < j \leq q$,

(ii) $P\Big(\sup_{t\in T} \|\hat{g}'_{\lambda j}(t)\| = 0\Big) \to 1$ *for any* $d_0 < j \leq q$.

By Theorem 1, the irrelevant predictors' coefficients are shrunken to zero consistently over the entire index support uniformly. Let $\mathbf{A}^*_{G_M} = \{\mathbf{A}^*_g, q_0 + \mathbf{A}^*_{g'}\}$. Next, we study the asymptotic normality for the estimators of the nonzero components $\eta_{0\mathbf{A}^*_{G_M}} = (g_1(t), \ldots, g_{q_0}(t), hg'_1(t), \ldots, hg'_{d_0}(t))^\top$ of $g(t)$ and $g'(t)$. Define $Z^* = (Z_1, \ldots, Z_{q_0})^\top$ and $Z^{**} = (Z_1, \ldots, Z_{d_0})^\top$.

Theorem 2 *Suppose that the conditions of Theorem 1 hold. Then*

$$\sqrt{nh}\left(\hat{\eta}_{\lambda\mathbf{A}^*_{G_M}}(t) - \eta_{0\mathbf{A}^*_{G_M}}(t) - \frac{h^2}{2}\Sigma_1^{*-1}(t)\begin{pmatrix} E(Z^*Z^{*\top}|t)g''_{\mathbf{A}^*_g}(t)\mu_2 f(t) \\ 0 \end{pmatrix}\right)$$

$$\xrightarrow{\mathcal{D}} N(0, \sigma^2\Sigma_1^{*-1}(t)V_1^*\Sigma_1^{*-1}(t)),$$

where $\mu_2 = \int u^2K(u)\mathrm{d}u, \nu_2 = \int u^2K^2(u)\mathrm{d}u, \kappa_2 = \int K^2(u)\mathrm{d}u$,

$$V_1^* = \begin{pmatrix} f(t)E(Z^*Z^{*\top}|t)\kappa_2 & 0 \\ 0 & f(t)E(Z^{**}Z^{**\top}|t)\nu_2 \end{pmatrix},$$

$$\Sigma_1^*(t) = \begin{pmatrix} f(t)E(Z^*Z^{*\top}|t) & 0 \\ 0 & f(t)E(Z^{**}Z^{**\top}|t)\mu_2 \end{pmatrix}.$$

Remark 1 From Theorem 1, the sparsity of $g'(t)$ indicates that $(\hat{g}_{d_0+1}(t), \ldots, \hat{g}_{q_0}(t))^\top$ is a constant vector. Through the variable selection procedure, we can separate the predictors into the constant coefficient ones and those with varying coefficients. Furthermore, as soon as we identify the constant coefficients, the interested model turns to the single-index varying-coefficient partially linear model, and some methods can be generalized to obtain root-n consistent estimators.

2.2 Stage 2: efficient variable selection for single-index parameters

In this section, we develop a penalized efficient estimating equation method for variable selection and parametric estimation. Note that in the first stage, we select the important nonparametric components, which reduces the dimension of Z from q to q_0. Let $\hat{g}^*_\lambda(t) = (\hat{g}_{\lambda 1}(t), \dots, \hat{g}_{\lambda q_0}(t))^\top$ be the penalized estimator of the nonzero components of $g^*(t)$, $g^*(t) = (g_1(t), \dots, g_{q_0}(t))^\top$, and let $\hat{g}'^*_\lambda(t) = (\hat{g}'_{\lambda 1}(t), \dots, \hat{g}'_{\lambda d_0}(t), 0^\top_{q_0-d_0})^\top$ be the penalized estimators of $g'^*(t)$, where $g'^*(t) = (g'_1(t), \dots, g'_{d_0}(t), 0^\top_{q_0-d_0})^\top$.

Note that $\|\beta\| = 1$ and the first component of β is positive. Let $\beta = (\beta_1, \dots, \beta_p)^\top$ and $\beta^{(1)} = (\beta_2, \dots, \beta_p)^\top$. Then we have $\beta(\beta^{(1)}) = ((1 - \|\beta^{(1)}\|^2)^{1/2}, \beta_2, \dots, \beta_p)^\top$. The true parameter vector $\beta^{(1)}_0$ must satisfy the constraint $\|\beta^{(1)}_0\| < 1$. Thus, β is infinitely differentiable in a neighborhood of $\beta^{(1)}_0$, and the Jacobian matrix is $J_{\beta^{(1)}} = \partial\beta / \partial\beta^{(1)} = (b_1, \dots, b_p)^\top$, where $b_s (1 < s \le p)$ is a $(p-1)$-dimensional unit vector with the sth component 1, and $b_1 = -(1 - \|\beta^{(1)}\|^2)^{-1/2}\beta^{(1)}$. With the result from Stage 1, we construct the penalized efficient estimating equation for $\beta^{(1)}$ as

$$U_{\lambda_n}(\beta^{(1)}) = \sum_{i=1}^{n} \tilde{S}^{eff}_\beta(Y_i, X_i, Z^*_i) - n\boldsymbol{q}_{\lambda_n}(|\beta^{(1)}|)\boldsymbol{sgn}(\beta^{(1)}), \tag{5}$$

where

$$\begin{aligned}\tilde{S}^{eff}_\beta(Y_i, X_i, Z^*_i) = \{Y_i - Z^{*\top}_i \hat{g}^*_\lambda(X_i^\top\beta)\}\Big\{ & \hat{g}'^{*\top}_\lambda(X_i^\top\beta) Z^*_i J^\top_{\beta^{(1)}} X_i \\ & - \hat{E}[\hat{g}'^{*\top}_\lambda(X_i^\top\beta) Z^*_i J^\top_{\beta^{(1)}} X_i Z^{*\top}_i | X_i^\top\beta]\big\{\hat{E}[Z^*_i Z^{*\top}_i | X_i^\top\beta]\big\}^{-1} Z^*_i \Big\}\end{aligned}$$

with $\sum_{i=1}^n \tilde{S}^{eff}_\beta(Y_i, X_i, Z^*_i) = 0$ being the semiparametric efficient estimating equation proposed in Lai et al. (2016). Here $\boldsymbol{q}_{\lambda_n}(|\beta^{(1)}|) = (q_{\lambda_n}(|\beta_2|), \dots, q_{\lambda_n}(|\beta_p|))^\top$, $q_{\lambda_n} = p'_{\lambda_n}$, $\boldsymbol{sgn}(\beta^{(1)}) = (\mathrm{sgn}(\beta_2), \dots, \mathrm{sgn}(\beta_p))^\top$, and $p_{\lambda_n}(\cdot)$ is the SCAD penalty function with λ_n being a tuning parameter, defined as $p'_\lambda(w) = \lambda\Big\{ I(w \le \lambda) + \frac{(a\lambda - w)_+}{(a-1)\lambda} I(w > \lambda)\Big\}$, with $a > 2, w > 0$ and $p_\lambda(0) = 0$. Denote $U_{\lambda_n}(\beta^{(1)}) = (U_{\lambda_{n2}}(\beta^{(1)}), \cdots, U_{\lambda_{np}}(\beta^{(1)}))^\top$.

Since $U_{\lambda_n}(\beta^{(1)})$ is not continuous, there may not be an exact solution. We introduce a zero-crossing penalized estimating equation defined in Johnson et al. (2008) to accommodate discrete estimating functions. Let $\hat{\beta}^{(1)}_{\lambda_n}$ be a zero-crossing to penalized estimating equation, if for $j = 2, \dots, p$,

$$\overline{\lim_{\epsilon \to 0+}} n^{-1} U_{\lambda_n j}(\hat{\beta}^{(1)}_{\lambda_n} + \epsilon e_j) U_{\lambda_n j}(\hat{\beta}^{(1)}_{\lambda_n} - \epsilon e_j) \le 0,$$

where e_j is the jth canonical unit vector. To study properties of sparsity and asymptotic normality, the following conditions are needed.

- *C6* The function $D^*_2(t) = E[g'^{*\top}(t) Z^* J^\top_{\beta^{(1)}_0} X Z^{*\top} | t]$ are twice continuously differentiable with bounded derivatives,

- *C7* The matrix $\Sigma_2 = E\left[g'^{*\top}(X^\top\beta_0)Z^*J^\top_{\beta_0^{(1)}}X - D_2^*(X^\top\beta_0)D_1^{*-1}(X^\top\beta_0)Z^*\right]^{\otimes 2}$ is positive definite, where $D_1^*(t) = E[Z^*Z^{*\top}|t]$.

Without loss of generality, we define $\mathbf{A}_{\beta^{(1)}} = \{2, \dots, p\}$ and the true nonzero components index $\mathbf{A}^*_\beta = \mathbf{A}^*_{\beta^{(1)}} \cup \{1\}$, where $\mathbf{A}^*_{\beta^{(1)}} = \{2, \dots, p_0\}$. Rewrite $U_{\lambda_n}(\beta^{(1)}) = (U^\top_{\lambda_n\mathbf{A}^*_{\beta^{(1)}}}(\beta^{(1)}), U^\top_{\lambda_n[\mathbf{A}_{\beta^{(1)}}-\mathbf{A}^*_{\beta^{(1)}}]}(\beta^{(1)}))^\top$ and $\beta^{(1)} = (\beta^{(1)\top}_{\mathbf{A}^*_{\beta^{(1)}}}, \beta^{(1)\top}_{[\mathbf{A}_{\beta^{(1)}}-\mathbf{A}^*_{\beta^{(1)}}]})^\top$.

Theorem 3 *Suppose that conditions C1–C7 hold, if* $nh^4 \to \infty, nh^6 \to 0$, $\lambda_n \to 0$ *and* $\sqrt{n}\lambda_n \to \infty$, *then the following results hold:*

(i) *There exists a zero-crossing* $\hat{\beta}^{(1)}_{\lambda_n}$ *to* $U_{\lambda_n}(\beta^{(1)})$ *that satisfies* $\hat{\beta}^{(1)}_{\lambda_n} = \beta^{(1)}_0 + O_p(n^{-1/2})$. *Furthermore, there exists a zero-crossing estimator* $\hat{\beta}^{(1)}_{\lambda_n} = (\hat{\beta}^{(1)\top}_{\lambda_n\mathbf{A}^*_{\beta^{(1)}}}, \mathbf{0}^\top)^\top$ *of* $U_{\lambda_n}(\beta^{(1)})$ *satisfies* $U_{\lambda_n\mathbf{A}^*_{\hat{\beta}^{(1)}}}(\hat{\beta}^{(1)}_{\lambda_n}) = 0$.

(ii) *For any root-n consistent zero-crossing estimator of* $U_{\lambda_n}(\beta^{(1)})$, *denoted by* $\hat{\beta}^{(1)}_{\lambda_n} = (\hat{\beta}_{\lambda_n 2}, \dots, \hat{\beta}_{\lambda_n p})^\top$, *as* $n \to \infty$, *with probability tending to 1,* $\hat{\beta}_{\lambda_n k} = 0$, $k = p_0 + 1, \dots, p$. *Moreover, let* $\hat{\beta}^{(1)*}_{\lambda_n} = (\hat{\beta}_{\lambda_n 2}, \dots, \hat{\beta}_{\lambda_n p_0})^\top, \beta^{(1)*}_0 = (\beta_{02}, \dots, \beta_{0p_0})^\top$ *and* $X^* = (X_1, \dots, X_{p_0})^\top$, *then*

$$\left[\Sigma_3^* + \Sigma_4^*\right]\sqrt{n}\left(\hat{\beta}^{(1)*}_{\lambda_n} - \beta^{(1)*}_0 + \left[\Sigma_3^* + \Sigma_4^*\right]^{-1}B_n^*\right)\xrightarrow{\mathcal{D}}N(0, \Sigma_5^*),$$

where

$$\begin{aligned}
\Sigma_3^* &= E\left[g'^{*\top}(X^\top\beta_0)Z^*J^\top_{\beta_0^{(1)*}}X^* - \tilde{D}_2^*(X^\top\beta_0)D_1^{*-1}(X^\top\beta_0)Z^*\right]^{\otimes 2},\\
\tilde{D}_2^*(X^\top\beta_0) &= E[g'^{*\top}(X^\top\beta_0)Z^*J^\top_{\beta_0^{*(1)}}X^*Z^{*\top}|X^\top\beta_0],\\
B_n^* &= (q_{\lambda_n}(|\beta_{02}|)\mathrm{sgn}(\beta_{02}), \dots, q_{\lambda_n}(|\beta_{0p_0}|)\mathrm{sgn}(\beta_{0p_0}))^\top,\\
\Sigma_4^* &= \text{ diag } (q'_{\lambda_n}(|\beta_{02}|)\mathrm{sgn}(\beta_{02}), \dots, q'_{\lambda_n}(|\beta_{0p_0}|)\mathrm{sgn}(\beta_{0p_0})), \quad \Sigma_5^* = \sigma^2\Sigma_3^*.
\end{aligned}$$

Corollary 1 *Under the conditions of Theorem 3, if* $q_{\lambda_n}(|\beta_j|) = q'_{\lambda_n}(|\beta_j|) = 0$ *for* $\beta_j \neq 0$,

$$\sqrt{n}(\hat{\beta}^*_{\lambda_n} - \beta^*_0)\xrightarrow{\mathcal{D}}N(0, \sigma^2 J_{\beta_0^{(1)*}}\Sigma_3^{*-1}J^\top_{\beta_0^{(1)*}}).$$

Remark 2 From Corollary 1, we can select the relevant predictors and estimate the coefficients simultaneously. Particularly, if the true model has $q_0 = d_0$, then the asymptotic variance of $\hat{\beta}^*_{\lambda_n}$ achieves the semiparametric efficiency bound. If $d_0 < q_0$, the true model turns to the single-index varying-coefficient partially linear model. In order to get the efficient estimator, the efficient score vector should be redefined.

2.3 Tuning parameters selection

From Theorems 1–3, as long as $n^{-11/10}\phi_n \to 0$, $n^{-11/10}\phi'_n \to 0$, $n^{-11/10}\varphi_n \to \infty$, $n^{-11/10}\varphi'_n \to \infty$, and $n^{-9/10}\psi_n \to \infty$, the true model can be consistently selected and the optimal convergence rate is achieved. In simulations and real applications, the tuning parameters $\lambda_{1k}, \lambda_{2k}, 1 \le k \le q$ and λ_n need to be selected. Following the idea of Wang and Xia (2009), the BIC type criterions are adopted.

For λ_{1k} and λ_{2k}, $1 \le k \le q$, let

$$\lambda_{1k} = \frac{\lambda_{10}}{n^{-1/2}\|\hat{c}_k\|} \quad \text{and} \quad \lambda_{2k} = \frac{\lambda_{20}}{n^{-1/2}\|\hat{d}_k\|},$$

where $\hat{c}_k^\top$ and $\hat{d}_k^\top$ are the kth row of the unpenalized estimates $\hat{G}$ and $\hat{G}'$, respectively. Thus, the selection of the sequences $\{\lambda_{1k}, 1 \le k \le q\}$ and $\{\lambda_{2k}, 1 \le k \le q\}$ becomes to select λ_{10} and λ_{20}. Similar to Wang and Xia (2009), we choose λ_{10} and λ_{20} by constructing the BIC-type criterion

$$\mathrm{BIC}_\lambda = \log(\mathrm{RSS}_\lambda) + df_{\lambda_1}\frac{\log(nh)}{nh} + df_{\lambda_2}\frac{\log(nh^3)}{nh^3},$$

where

$$\mathrm{RSS}_\lambda = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i - Z_i^\top \hat{g}_\lambda(X_j^\top\hat{\beta}) - Z_i^\top \hat{g}'_\lambda(X_j^\top\hat{\beta})(X_i^\top\hat{\beta} - X_j^\top\hat{\beta})\}^2 K_h(X_i^\top\hat{\beta} - X_j^\top\hat{\beta}),$$

$\hat{\beta}$ is the unpenalized initial estimator, $0 \le df_{\lambda_1} \le q, 0 \le df_{\lambda_2} \le df_{\lambda_1}$ are the numbers of nonzero coefficients identified in $\hat{G}_{M_\lambda}$ with $\lambda = (\lambda_1, \lambda_2)$. The tuning parameters can be obtained by minimizing BIC_λ. Similar to the proof of Theorem 3 in Wang and Xia (2009), we can prove that the proposed BIC-type criterion can identify the true model consistently.

On the other hand, to select the tuning parameter λ_n for $\hat{\beta}_{\lambda_n}$, we construct the BIC-type criterion as follows

$$\mathrm{BIC}_{\lambda_n} = (\hat{\beta}^{(1)}_{\lambda_n} - \hat{\beta}^{(1)})^\top \hat{\Sigma}_2^{-1}(\hat{\beta}^{(1)}_{\lambda_n} - \hat{\beta}^{(1)}) + df_{\lambda_n}\frac{\log n}{n},$$

where $\hat{\beta}^{(1)}$ is the full model initial estimator and $\hat{\Sigma}_2$ is the estimator of Σ_2. Therefore, similar to the proof of Theorem 4 in Wang and Leng (2007), we can conclude that this BIC-type criterion can identify the true model consistently.

3 Numerical examples

3.1 Simulations

To illustrate the finite sample performance of the proposed method, we consider two examples. The Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$ is used. Moreover, since root-n consistency of the proposed estimators does not require undersmoothing of nonparametric functions, the optimal bandwidth can be selected to be the optimal one for the estimation of nonparametric functions. We use fivefold cross-validation to choose bandwidths for estimators. The tuning parameters are determined by two BIC-type criterions proposed in Sect. 2.3.

Let $\{Y_i, X_i, Z_i; i = 1, \dots, n\}$ be i.i.d samples. $X_i = (X_{i1}, \dots, X_{ip})^\top$ are generated from uniform distribution on $[0, 1]^p$ with independent components, $Z_{i1} = 1$ and $(Z_{i2}, \dots, Z_{iq})^\top$ follows a multivariate normal distribution with $cov(Z_{ij_1}, Z_{ij_2}) = 0.5^{|j_1 - j_2|}$ for $2 \leq j_1, j_2 \leq q$. Consider the model

$$Y_i = \sum_{s=1}^{q} g_s(X_i^\top \beta) Z_{i(s+1)} + 0.5 e_i, \tag{6}$$

where $\beta = (\beta_1, \dots, \beta_p)^\top$ with true value β_0 and e_i is the error term. Additionally, the functions $g_s(\cdot)$ are set, respectively, in the following examples. Let $n = 100$ and 200. A total of 500 simulation replications are conducted for each example.

Example 1 Let $p = 8$ and $q = 8$. In model (6), $\beta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top / \sqrt{15.25}$, $g_1(u) = 4u, g_2(u) = \exp(2u - 1), g_3(u) = 2\cos^2(2\pi u)$ and $g_k(u) \equiv 0$ for $k = 4, \cdots, 8$. Moreover, $e_i \sim N(0, 1)$. As one can see, the first three variables (Z_1, Z_2, Z_3) are relevant and have varying effects.

Example 2 The setup is the same as that in Example 1 except for different design of function g. Specifically, $g_1(u) = 2\sin(2\pi u), g_2(u) = \exp(2u - 1), g_3(u) = 4, g_4(u) = 1.5$ and $g_k(u) \equiv 0$ for $k = 5, \cdots, 8$. That is, four variables (Z_1, Z_2, Z_3, Z_4) are relevant where the first two have varying effects and the other two have constant effects.

To demonstrate the performance of the proposed procedure, we consider the following criterions: NS is the average number of variables selected; NST is the average number of selected that are truly nonzero; NV is the average number of varying-coefficient components selected; NVT is the average number of varying-coefficient components selected that are truly nonzero and varying; NC is the average number of nonzero constant components selected while NCT is the average number of nonzero constant components selected which are truly nonzero constant. Furthermore, we differentiate the following three situations. When one or more relevant predictors are not selected, we label it as an under-fitted model. The following cases are also labeled as under-fitted: A varying-coefficient component is selected as a nonzero constant component; a nonzero constant component is estimated as a varying-coefficient component. When the resulting model is exactly the same as the true model, we refer it to the correctly fitted model. When the resulting model includes at least one irrelevant predictor and all relevant predictors are selected, we label it as an over-fitted model. Following Wang and Xia (2009), to evaluate estimation, we use the relative estimation error (REE)

$$\mathrm{REE}_{\beta} = \frac{\sum_{j=1}^{p} |\hat{\beta}_j - \beta_j|}{\sum_{j=1}^{p} |\bar{\beta}_j - \beta_j|}, \quad \mathrm{REE}_{g} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{q} |\hat{g}_j(X_i^\top \hat{\beta}) - g_j(X_i^\top \beta_0)|}{\sum_{i=1}^{n} \sum_{j=1}^{q} |\bar{g}_j(X_i^\top \bar{\beta}) - g_j(X_i^\top \beta_0)|},$$

where $\bar{\beta}_j, \bar{g}_j(\cdot)$ are either unpenalized estimators or oracle estimators.

We summarize the results in Table 1. In the columns labeled 'U-fit,' 'C-fit' and 'O-fit,' we present proportions of trials that result in under-fitted, correctly fitted and over-fitted models, respectively. The values in the parentheses are the corresponding standard deviations. Several observations can be made. First, the proportion of the correctly fitted model increases as the sample size increases, quickly approaching 1. This also confirms that the BIC-type criterions proposed in Sect. 2.3 can indeed identify the true model consistently. Second, the means of $\mathrm{REE}_\beta(\mathrm{REE}_g)$ of our estimators to unpenalized estimators are much smaller than 1, which indicates that the proposed estimators are much more accurate than the unpenalized estimators. Last, the values of $\mathrm{REE}_\beta(\mathrm{REE}_g)$ of ours to the oracle estimators based on true model are close to 1. These phenomena corroborate the oracle properties of the proposed estimators.

3.2 Data analysis

We analyze a body fat data (http://lib.stat.cmu.edu/datasets/bodyfat). The aim is to build a predictive model for the percentage of body fat. We delete three samples with potential outliers, which lead to records on 249 men and thirteen baseline predictors: AGE (age), BMI (body mass index), NK (neck), CT (chest), AN (2 abdomen), HIP (hip), RAN (the ratio of 2 abdomen to hip), TN (thigh), KE (knee), AK (ankle), BS (biceps), FA (forearm) and WI (wrist). All predictors except AGE, BMI and RAN are standardized. These three variables are transformed so that the marginal distributions are $U[0, 1]$. The response of interest is the logarithm of percentage of body fat.

Lai et al. (2016) consider the following homoscedastic model

$$\begin{aligned} Y = g_0(U) + g_1(U)NK + g_2(U)CT + g_3(U)AN + g_4(U)HIP + g_5(U)TN \\ + g_6(U)KE + g_7(U)AK + g_8(U)BS + g_9(U)FA + g_{10}(U)WI + \varepsilon, \end{aligned} \tag{7}$$

where $U = \beta_1 AGE + \beta_2 BMI + \beta_3 RAN$ and ε is the error term. We label the estimation developed in Lai et al. (2016) for model (7) as 'The unpenalized approach.' Two hundred samples are random selected to fit the model, and the rest 49 observations are to evaluate the prediction ability of the underlying model by the mean absolute prediction error (MAPE), which are defined as $\sum_{i=1}^{49} |y_i - \hat{y}_i|/49$, where $\hat{y}_i$ is the predicted value by the proposed approach or the unpenalized approach. To overcome the randomness, we run the process 100 times and summarize the results. Figure 1 shows the boxplots for 100 MAPEs based on the proposed and unpenalized approaches. The median of MAPEs of the proposed approach is equal to 0.2665

Table 1 Summary of the two-stage procedure

n	For X		For Z				Proportion of models			REE_β		REE_g	
	NS	NST	NV	NVT	NC	NCT	U-fit	C-fit	O-fit	unpenalized	oracle	unpenalized	oracle
Example 1													
Oracle	3	3	3	3	0	0	0	1	0				
100	3.900	3.000	2.934	2.884	0.192	0.000	0.116	0.534	0.350	0.3457	1.7824	0.4074	1.0573
	(1.4759)	(0.0000)	(0.4171)	(0.3205)	(0.4854)	(0.0000)				(0.1491)	(0.7877)	(0.0509)	(0.0565)
200	3.308	3.000	3.000	3.000	0.004	0.000	0.000	0.848	0.152	0.3214	1.1833	0.4496	1.0086
	(0.9396)	(0.0000)	(0.0000)	(0.0000)	(0.0632)	(0.0000)				(0.1193)	(0.4117)	(0.0375)	(0.0116)
Example 2													
Oracle	3	3	2	2	2	2	0	1	0				
100	3.200	3.000	2.000	2.000	2.300	2.000	0.000	0.600	0.400	0.3668	1.5578	0.3816	1.0713
	(0.4216)	(0.0000)	(0.0000)	(0.0000)	(0.4830)	(0.0000)				(0.1372)	(0.5441)	(0.0354)	(0.0846)
200	3.080	3.000	2.000	2.000	2.000	2.000	0.000	0.930	0.070	0.3421	1.1346	0.4357	1.0028
	(0.3253)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)				(0.1181)	(0.3066)	(0.0412)	(0.0064)

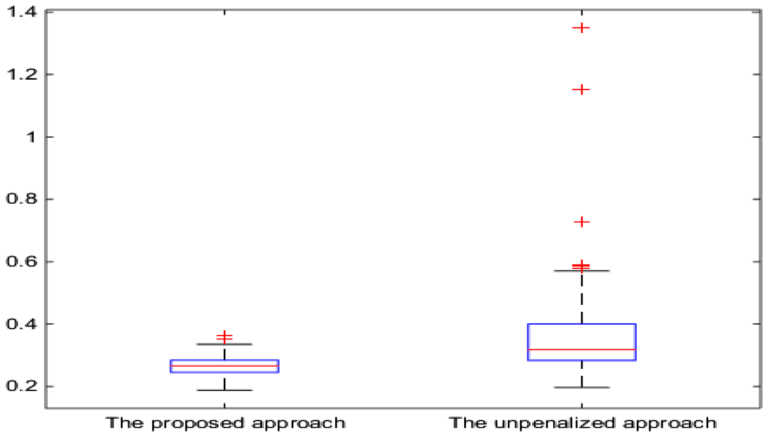

Fig. 1 The boxplots of 100 MAPEs based on the proposed and unpenalized approaches

(0.0238), while that of the unpenalized approach is 0.3176 (0.0947). The values in the parentheses are MAD of APEs. We can conclude that the proposed approach has better prediction performance and stability.

4 Discussion

In this paper, we have proposed a two-stage efficient variable selection procedure for single-index varying-coefficient models. With a proper choice of tuning parameters, we have established that the proposed method can consistently identify the true structure, and the estimators of important parametric and nonparametric components are consistent and asymptotically normal. It is remarkable that the method can also separate predictors into varying-coefficient and constant-coefficient predictors automatically. The numerical studies show that the proposed approach works well. We have focused on variable selection problems with fixed dimension. Extension to a 'large p small n' scenario is of interest in future study. In addition, very little work exists on the inferential aspects (e.g., constructing confidence intervals or statistical testing) of high dimensional single-index varying-coefficient models, which requires further research.

Acknowledgements The authors wish to thank the associate editor and two anonymous referees for their constructive comments that greatly improved the manuscript. This study was supported by the National Natural Science Foundation of China (11771215, 11971404), the Natural Science Foundation of Jiangsu Province (BK20161530), Qing Lan Project of Jiangsu Province (2016), and MOE (Ministry of Education in China) Project of Humanities and Social Sciences (19YJC910010).

Appendix: Technical proofs

Proof of Theorem 1 The proof is similar to the proof of Theorem 1 in Wang and Xia (2009). From Sect. 2, we know that $G_M = (G^\top, hG'^\top)^\top$, and its estimator is $\hat{G}_M$. Firstly, let $\alpha_n = (nh)^{-1/2}$, we prove that there exists a large constant C such that

$$\liminf_n P\left\{ \inf_{\frac{1}{n}\|u\|^2 = C} Q_\lambda(G_M + \alpha_n u) > Q_\lambda(G_M) \right\} = 1 - \epsilon, \tag{8}$$

for any small constant $\epsilon > 0$. (8) implies with probability at least $1-\epsilon$ that there exists a local maximum in $\{G_M + \alpha_n u \,:\, n^{-1}\|u\|^2 \le C\}$, where $u = (u_{ij}) \in R^{2q\times n}$ with columns $u_1, \dots, u_n$ and rows $v_1, \dots, v_q$. Further, we can divide u as $u = (\bar{u}_1^\top, \bar{u}_2^\top)^\top$, where $\bar{u}_k \in R^{q\times n}$ with columns $u_{k1}, \dots, u_{kn}$ and rows $v_{k1}, \dots, v_{kq}$, $k = 1, 2$. Hence, (8) also proves there exists a local maximizer such that

$$n^{-1} \sum_{i=1}^n \|\hat{\eta}_\lambda(t_i) - \eta_0(t_i)\|^2 = O_p(\alpha_n^2).$$

From (4), note that $t_i = X_i^\top \beta, \beta \in \mathbf{B}_n$ and $a_j = g(t_j), b_j = g'(t_j), \eta_j = (a_j^\top, hb_j^\top)^\top$,

$$n^{-1}h\big\{Q_\lambda(G_M + \alpha_n u) - Q_\lambda(G_M)\big\} = L_1 + L_2, \tag{9}$$

where

$$\begin{aligned} L_1 &= \frac{h}{n}\sum_{j=1}^n\sum_{i=1}^n \left[Y_i - Z_i^\top(a_j + \alpha_n u_{1j}) - \frac{t_i - t_j}{h} Z_i^\top(hb_j + \alpha_n u_{2j})\right]^2 K_h(t_i - t_j) \\ &\quad - \frac{h}{n}\sum_{j=1}^n\sum_{i=1}^n \left[Y_i - Z_i^\top a_j - \frac{t_i - t_j}{h} Z_i^\top hb_j\right]^2 K_h(t_i - t_j), \\ L_2 &= \frac{h}{n}\sum_{j=1}^q \lambda_{1j}(\|c_j + \alpha_n v_{1j}\| - \|c_j\|) + \frac{h}{n}\sum_{j=1}^q \lambda_{2j}(\|d_j + \alpha_n v_{2j}\| - \|d_j\|). \end{aligned}$$

For L_1, since $\tilde{Z}_i(t_j) = (Z_i^\top, \frac{t_i - t_j}{h} Z_i^\top)^\top$ and $\alpha_n = (nh)^{-1/2}$, we have

$$\begin{aligned} L_1 &= \frac{h}{n}\sum_{j=1}^n\sum_{i=1}^n [\alpha_n \tilde{Z}_i^\top(t_j)u_j]^2 K_h(t_i - t_i) - \frac{2h}{n}\sum_{j=1}^n\sum_{i=1}^n [Y_i - \tilde{Z}_i^\top(t_j)\eta_j][\alpha_n \tilde{Z}_i^\top(t_j)u_j] K_h(t_i - t_j) \\ &= \frac{1}{n}\sum_{j=1}^n u_j^\top \hat{\Sigma}_1(t_j) u_j - \frac{2}{n}\sum_{j=1}^n u_j^\top \hat{e}_j - \frac{2}{n}\sum_{j=1}^n u_j^\top \tilde{e}_j, \end{aligned} \tag{10}$$

where

$$\hat{\Sigma}_1(t_j) = \frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i(t_j)\tilde{Z}_i^{\top}(t_j)K_h(t_i - t_j), \quad \hat{e}_j = \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\varepsilon_i\tilde{Z}_i(t_j)K_h(t_i - t_j),$$
$$\tilde{e}_j = \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\tilde{Z}_i(t_j)g''^{\top}(t_j)Z_i(\frac{t_i - t_j}{h})^2h^2K_h(t_i - t_j).$$

Let $R_1 = L_1 + L_2$, from the assumptions, we know that $\|c_j\| = 0$ for any $j > q_0$, and $\|d_j\| = 0$ for any $j > q_0$. Furthermore, if some $g_k(t)$ is constant, without loss of generality, $g_{d_0+1}(t), \dots, g_{q_0}(t)$ are constants, then $\|d_j\| = 0$ for any $j > d_0$. Based on this information and $\alpha_n = (nh)^{-1/2}$, we can conclude that

$$\begin{aligned} R_1 \geq & \frac{1}{n}\sum_{j=1}^{n}u_j^{\top}\hat{\Sigma}_1(t_j)u_j - \frac{2}{n}\sum_{i=1}^{2}u_j^{\top}\hat{e}_j - \frac{2}{n}\sum_{i=1}^{2}u_j^{\top}\tilde{e}_j \\ & + \frac{h}{n}\sum_{j=1}^{q_0}\lambda_{1j}(\|c_j + \alpha_n v_{1j}\| - \|c_j\|) + \frac{h}{n}\sum_{j=1}^{q_0}\lambda_{2j}(\|d_j + \alpha_n v_{2j}\| - \|d_j\|). \end{aligned}$$

Let $\hat{\lambda}_j^{\min}$ be the smallest eigenvalue of $\hat{\Sigma}_1(t_j)$, $\hat{\lambda}_{\min} = \min\{\hat{\lambda}_j^{\min}, j = 1, \dots, n\}$. Then, we have

$$\begin{aligned} R_1 \geq & \frac{1}{n}\sum_{j=1}^{n}\left\{\|u_j\|^2\hat{\lambda}_j^{\min} - 2\|u_j\|\|\hat{e}_j\| - 2\|u_j\|\|\tilde{e}_j\|\right\} - \sqrt{\frac{h}{n^3}}\sum_{j=1}^{q_0}\|v_{1j}\|\lambda_{1j} - \sqrt{\frac{h}{n^3}}\sum_{j=1}^{d_0}\|v_{2j}\|\lambda_{2j} \\ \geq & \hat{\lambda}_{\min}\left\{\frac{1}{n}\sum_{j=1}^{n}\|u_j\|^2\right\} - \frac{2}{n}\sum_{j=1}^{n}\|u_j\|\|\hat{e}_j\| - \frac{2}{n}\sum_{j=1}^{n}\|u_j\|\|\tilde{e}_j\| - \sqrt{\frac{h}{n^3}}\sum_{j=1}^{q_0}\left[\|v_{1j}\|\lambda_{1j} + \|v_{2j}\|\lambda_{2j}\right] \\ \geq & \hat{\lambda}_{\min}\left\{\frac{1}{n}\|u\|^2\right\} - 2\left\{\frac{1}{n}\|u\|^2\right\}^{\frac{1}{2}}\left\{\frac{1}{n}\|\hat{e}\|^2\right\}^{\frac{1}{2}} - 2\left\{\frac{1}{n}\|u\|^2\right\}^{\frac{1}{2}}\left\{\frac{1}{n}\|\tilde{e}\|^2\right\}^{\frac{1}{2}} \\ & - \sqrt{\frac{h}{n^3}}\sum_{j=1}^{q_0}\left[\|v_{1j}\|\lambda_{1j} + \|v_{2j}\|\lambda_{2j}\right] := R_2. \end{aligned}$$

By the condition that $\frac{1}{n}\|u\|^2 = C$, we have

$$\begin{aligned} R_2 \geq & \hat{\lambda}_{\min}C^2 - 2C\left\{\frac{1}{n}\|\hat{e}\|^2\right\}^{\frac{1}{2}} - 2C\left\{\frac{1}{n}\|\tilde{e}\|^2\right\}^{\frac{1}{2}} \\ & - \frac{\sqrt{h}}{n}\phi_n\left(\frac{1}{n}\sum_{j=1}^{q}\|v_{1j}\|^2\right)^{\frac{1}{2}} - \frac{\sqrt{h}}{n}\phi_n'\left(\frac{1}{n}\sum_{j=1}^{q}\|v_{2j}\|^2\right)^{\frac{1}{2}} \\ \geq & \hat{\lambda}_{\min}C^2 - 2C\left\{\frac{1}{n}\|\hat{e}\|^2\right\}^{\frac{1}{2}} - 2C\left\{\frac{1}{n}\|\tilde{e}\|^2\right\}^{\frac{1}{2}} - \frac{\sqrt{h}}{n}\phi_nC - \frac{\sqrt{h}}{n}\phi_n'C. \end{aligned} \tag{11}$$

From (10), we know that $\hat{e}_j = \sqrt{\frac{h}{n}}\sum_{i=1}^{n}\varepsilon_i\tilde{Z}_i(t_j)K_h(t_i - t_j)$. Then for a given t_j,

$$E\|\hat{e}_j\|^2 = \frac{h}{n}\sum_{i=1}^{n}E\left[\varepsilon_i^2\tilde{Z}_i^{\top}(t_j)\tilde{Z}_i(t_j)\frac{1}{h^2}K^2\left(\frac{t_i - t_j}{h}\right)\right] = \frac{h}{n}\cdot n\cdot O\left(\frac{1}{h}\right) = O(1). \tag{12}$$

This concludes that $\frac{1}{n}\|\hat{e}\|^2 = O_p(1)$. Similarly,

$$\begin{aligned}E\|\tilde{e}_j\|^2 &= \frac{h}{n}\sum_{i=1}^{n} E\left[\tilde{Z}_i^\top(t_j)\tilde{Z}_i(t_j)g''^\top(t_j)Z_iZ_i^\top g''(t_j)h^4\left(\frac{t_i-t_j}{h}\right)^4\frac{1}{h^2}K^2\left(\frac{t_i-t_j}{h}\right)\right]\\ &\quad+\frac{2h}{n}\sum_{i>k} E\big[\tilde{Z}_i^\top(t_j)\tilde{Z}_k(t_j)g''^\top(t_j)Z_iZ_k^\top g''(t_j)\\ &\quad\times h^4\left(\frac{t_i-t_j}{h}\right)^2\left(\frac{t_k-t_j}{h}\right)^2\frac{1}{h^2}K\left(\frac{t_i-t_j}{h}\right)K\left(\frac{t_k-t_j}{h}\right)\bigg]\\ &= O(h^4)+O(nh^5),\end{aligned} \tag{13}$$

which deduces $E\|\tilde{e}_j\|^2 = O(1)$ since $h \propto n^{-1/5}$. Then $n^{-1}\|\tilde{e}\|^2 = O_p(1)$.

By Lemma A.3 in Wang and Xia (2009) and C1, it follows

$$\hat{\Sigma}_2(t) - f(t)\Sigma_2(t) = O_p\left(h + \left[\frac{\log(1/h)}{nh}\right]^{1/2}\right), \quad \Sigma_2(t) = \begin{pmatrix} E(ZZ^\top|t) & \mathbf{0} \\ \mathbf{0} & E(ZZ^\top|t)\mu_2 \end{pmatrix}. \tag{14}$$

Then we have $P(\hat{\lambda}_{\min} \to \lambda^0_{\min}) \to 1$ and

$$\lambda^0_{\min} = \inf_{t\in\mathbf{T}} \lambda_{\min}(f(t)\Sigma_2(t)) > 0.$$

If $\frac{\sqrt{h}}{n}\phi_n = n^{-\frac{11}{10}}\phi_n \to 0$ and $n^{-\frac{11}{10}}\phi'_n \to 0$, from (11) together with (12)-(14), we can see that the first term of (11) dominates the last two terms for sufficient large C. Then, we prove that

$$\liminf_n P\left\{\inf_{n^{-1}\|u\|^2=C} Q_\lambda(G_M+\alpha_n u) > Q_\lambda(G_M)\right\} = 1-\epsilon.$$

Next, we prove the sparsity. We first prove under some assumptions, we have $P(\|\hat{c}_{\lambda j}\| = 0) \to 1$ for any $q_0 < j \le q$ and $P(\|\hat{d}_{\lambda j}\| = 0) \to 1$ for any $d_0 < j \le q$. For $j = q$ (for $q_0 < j < q$, the proof is similar), if $\{\|\hat{c}_{\lambda q}\| \ne 0, \|\hat{d}_{\lambda q}\| \ne 0\}$, then it must be the solution of $\frac{\partial Q_\lambda(G_{\mathrm{ML}})}{\partial\omega_q}|_{G_{\mathrm{ML}}=\hat{G}_{ML_\lambda}} = 0$, where $\omega_q = (c_q^\top, d_q^\top)^\top$. That is

$$0 = -2\sum_{i=1}^{n} W_i\tilde{Z}_{iq} + \lambda_{1q}\frac{(c_q^\top, 0_{n\times1}^\top)^\top}{\|c_q\|} + \lambda_{2q}\frac{(0_{n\times1}^\top, d_q^\top)^\top}{\|d_q\|} = L_3 + L_4 + L_5, \tag{15}$$

where W_i is a $2n\times 2n$ diagonal matrix with the first n diagonal components $[Y_i - \hat{\eta}_{\lambda k}^\top\tilde{Z}_i(t_k)]K_h(t_i - t_k), k = 1, \ldots, n$ and the last n diagonal components $[Y_i - \hat{\eta}_{\lambda k}^\top\tilde{Z}_i(t_k)]K_h(t_i - t_k), k = 1, \ldots, n$, $\tilde{Z}_{iq} = (Z_{iq}1_{1\times n}, Z_{iq}(t_i - t_1), \ldots, Z_{iq}(t_i - t_n))^\top$. Then for the kth component L_{3k} of L_3, $k = 1, 2, \ldots, n$,

$$\begin{aligned} L_{3k} =& -2\sum_{i=1}^{n}(\eta_k - \hat{\eta}_{\lambda k})^{\top}\tilde{Z}_i(t_k)Z_{iq}K_h(t_i - t_k)\\ &- \sum_{i=1}^{n} g''^{\top}(t_k)Z_iZ_{iq}K_h(t_i - t_k)h^2\Big(\frac{t_i - t_k}{h}\Big)^2 - 2\sum_{i=1}^{n}\varepsilon_i Z_{iq}K_h(t_i - t_q)\\ :&= L_{3k1} + L_{3k2} + L_{3k3}. \end{aligned}$$

Similar to the proofs of (12) and (13), we have $(\sum_{k=1}^{n} L_{3k2}^2)^{1/2} = O_p(nh^2\sqrt{n}) = O_p(nh^{-1/2})$ since $h \propto n^{-1/5}$ and $(\sum_{k=1}^{n} L_{3k3}^2)^{1/2} = O_p(nh^{-1/2})$. Similar to the proof of (A.7) in (Wang and Xia 2009),

$$\begin{aligned} \Big(\sum_{k=1}^{n} L_{3k1}^2\Big)^{1/2} \leq& \Big(\sum_{k=1}^{n} \|\eta_0(t_k) - \hat{\eta}_{\lambda}(t_k)\|^2 \times \|\sum_{i=1}^{n} Z_{iq}Z_i^{\top}K_h(t_i - t_k)\|^2\Big)^{1/2}\\ =& \Big[nO_p\Big(\frac{1}{nh}\Big)O_p(n^2)\Big]^{1/2} = O_p(nh^{-1/2}). \end{aligned}$$

For the last n components of L_3, $k = 1, \dots, n$,

$$\begin{aligned} L_{3(k+n)} =& -2\sum_{i=1}^{n}(\eta_k - \hat{\eta}_{\lambda k})^{\top}\tilde{Z}_i(t_k)Z_{iq}(t_i - t_k)K_h(t_i - t_k)\\ &- \sum_{i=1}^{n} g''^{\top}(t_k)Z_iZ_{iq}(t_i - t_k)K_h(t_i - t_k)h^2\Big(\frac{t_i - t_k}{h}\Big)^2\\ &- 2\sum_{i=1}^{n}\varepsilon_i Z_{iq}(t_i - t_k)K_h(t_i - t_k)\\ :=& L_{3(k+n)1} + L_{3(k+n)2} + L_{3(k+n)3}, \end{aligned}$$

similarly, we have $(\sum_{k=1}^{n} L_{3(k+n)l}^2)^{1/2} < O_p(nh^{-1/2}), l = 1, 2, 3.$ It follows $\|L_3\| = O_p(nh^{-1/2})$.

On the other hand, $\|L_4\| = \lambda_{1q}$, if $n^{-\frac{11}{10}}\varphi_n \to \infty$, then

$$(nh^{-\frac{1}{2}})^{-1}\|L_4\| = (nh^{-\frac{1}{2}})^{-1}\lambda_{1q} \geq (nh^{-\frac{1}{2}})^{-1}\varphi_n \propto n^{-\frac{11}{10}}\varphi_n \to \infty.$$

Similarly, if $n^{-\frac{11}{10}}\varphi'_n \to \infty$, then

$$(nh^{-\frac{1}{2}})^{-1}\|L_5\| = (nh^{-\frac{1}{2}})^{-1}\lambda_{2q} \geq (nh^{-\frac{1}{2}})^{-1}\varphi'_n \propto n^{-\frac{11}{10}}\varphi'_n \to \infty.$$

These imply that

$$P(\|L_4\| + \|L_5\| \geq \|L_3\|) \to 1.$$

Therefore, with probability tending to 1, the estimating equation (15) cannot hold. Then we conclude, $P\big(\|\hat{c}_{\lambda j}\| = 0\big) \to 1$ for any $q_0 < j \leq q$, and $P\big(\|\hat{d}_{\lambda j}\| = 0\big) \to 1$ for any $q_0 < j \leq q$.

Next, we prove that under some conditions, $P\big(\|\hat{d}_{\lambda j}\| = 0\big) \to 1$ for any $d_0 < j \leq q_0$. For $j = q_0$ (for any $d_0 < j < q_0$, the proof is similar), if $\|\hat{d}_{\lambda j}\| \neq 0$, then it must be the solution of $\frac{\partial Q_\lambda(G_{\mathrm{ML}})}{\partial \omega_{q_0}} |_{G_{\mathrm{ML}}=\hat{G}_{ML_\lambda}} = 0$, where $\omega_{q_0} = (c_{q_0}^\top, d_{q_0}^\top)^\top$. Similar to (15), it follows

$$0 = \; -2\sum_{i=1}^{n} W_i \tilde{Z}_{iq_0} + \lambda_{1q_0} \frac{(c_{q_0}^\top, 0_{n\times 1}^\top)^\top}{\|c_{q_0}\|} + \lambda_{2q_0} \frac{(0_{n\times 1}^\top, d_{q_0}^\top)^\top}{\|d_{q_0}\|},$$

For the sparsity of $g'(t)$, we only need to study

$$0 = -2\sum_{i=1}^{n} W_i^{(2)} \tilde{Z}_{iq_0}^{(2)} + \lambda_{2q_0} \frac{d_{q_0}}{\|d_{q_0}\|} \; := L_6 + L_7, \tag{16}$$

where $W_i^{(2)}$ is a $n \times n$ diagonal matrix with elements $[Y_i - \hat{\eta}_{\lambda k}^\top \tilde{Z}_i(t_k)]K_h(t_i - t_k), k = 1, \ldots, n$, and $\tilde{Z}_{iq_0}^{(2)} = (Z_{iq_0}(t_i - t_1), \ldots, Z_{iq_0}(t_i - t_n))^\top$. Then the kth component L_{6k} of L_6 is

$$\begin{aligned}
L_{6k} &= -2\sum_{i=1}^{n} (\eta_k - \hat{\eta}_{\lambda k})^\top \tilde{Z}_i(t_k) Z_{iq_0}(t_i - t_k) K_h(t_i - t_k) \\
&\quad - \sum_{i=1}^{n} g''^\top(t_k) Z_i Z_{iq_0}(t_i - t_k) K_h(t_i - t_k) h^2 \Big(\frac{t_i - t_k}{h}\Big)^2 \\
&\quad - 2\sum_{i=1}^{n} \varepsilon_i Z_{iq_0}(t_i - t_k) K_h(t_i - t_k) \\
&:= L_{6k1} + L_{6k2} + L_{6k3},
\end{aligned}$$

Similar to the proof of $\|L_3\| = O_p(nh^{-1/2})$, we have

$$\begin{aligned}
\Big(\sum_{k=1}^{n} L_{6k1}^2\Big)^{1/2} &= \Big[nO_p\Big(\frac{1}{nh}\Big)O_p(n^2h^4)\Big]^{1/2} = O_p(nh\sqrt{h}), \\
\Big(\sum_{k=1}^{n} L_{6k2}^2\Big)^{1/2} &= O_p(n\sqrt{n}h^4), \quad \Big(\sum_{k=1}^{n} L_{6k3}^2\Big)^{1/2} = O_p(n\sqrt{h}).
\end{aligned}$$

Therefore, $\|L_6\| = O_p(n\sqrt{h})$. From the assumption $n^{-\frac{9}{10}}\psi_n \to \infty$, we have

$$(n\sqrt{h})^{-1}\|L_7\| = (n\sqrt{h})^{-1}\lambda_{2q_0} \geq (n\sqrt{h})^{-1}\psi_n = n^{-\frac{9}{10}}\psi_n \to \infty.$$

This implies that $P(\|L_7\| \geq \|L_6\|) \to 1$, thus (16) does not hold. Then we have $P\big(\|\hat{d}_{\lambda j}\| = 0\big) \to 1$ for any $d_0 < j \leq q_0$.

Therefore, we have

$$\begin{aligned}
&P(\|\hat{c}_{\lambda j}\| = 0) \to 1 \;\; \text{for any } q_0 < j \leq q, \\
&P(\|\hat{d}_{\lambda j}\| = 0) \to 1 \;\; \text{for any } d_0 < j \leq q.
\end{aligned} \tag{17}$$

By (4), (17) and Hunter and Li (2005) we know for $\hat{c}_{\lambda j}^{(m)} = (\hat{g}_{\lambda j}^{(m)}(t_1), \ldots, \hat{g}_{\lambda j}^{(m)}(t_n))^\top$ and $\hat{d}_{\lambda j}^{(m)} = (\hat{g}_{\lambda j}^{\prime(m)}(t_1), \ldots, \hat{g}_{\lambda j}^{\prime(m)}(t_n))^\top$, as $m \to \infty$, $\|\hat{c}_{\lambda j}^{(m)}\| \to \|\hat{c}_{\lambda j}\|$ and $\|\hat{d}_{\lambda j}^{(m)}\| \to \|\hat{d}_{\lambda j}\|$, $1 \le j \le q$. Then, for $1 \le j \le q_0$, $\|\hat{c}_{\lambda j}^{(m)}\|$ converges to a positive number, while $\|\hat{c}_{\lambda j}^{(m)}\|$ converges to 0 for $j > q_0$; for $1 \le j \le d_0$, $\|\hat{d}_{\lambda j}^{(m)}\|$ converges to a positive number, while $\|\hat{d}_{\lambda j}^{(m)}\|$ converges to 0 for $j > d_0$. From (4), it follows

$$\hat{\eta}_{\lambda}^{(m+1)}(t) = \left[\frac{1}{n}\sum_{i=1}^{n} \tilde{Z}_i(t)\tilde{Z}_i^\top(t)K_h(t_i - t) + \frac{1}{n}D^{(m)}\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} Y_i\tilde{Z}_i(t)K_h(t_i - t)\right].$$

Define

$$\begin{aligned}\Omega^{(m)}(t) &= \left[\frac{1}{n}\sum_{i=1}^{n} \tilde{Z}_i(t)\tilde{Z}_i^\top(t)K_h(t_i - t) + \frac{1}{n}D^{(m)}\right] \\ &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} Z_iZ_i^\top K_h(t_i - t) + \frac{D_a^{(m)}}{n} & \frac{1}{n}\sum_{i=1}^{n} Z_iZ_i^\top\left(\frac{t_i-t}{h}\right)K_h(t_i - t) \\ \frac{1}{n}\sum_{i=1}^{n} Z_iZ_i^\top\left(\frac{t_i-t}{h}\right)K_h(t_i - t) & \frac{1}{n}\sum_{i=1}^{n} Z_iZ_i^\top\left(\frac{t_i-t}{h}\right)^2 K_h(t_i - t) + \frac{D_b^{(m)}}{n} \end{pmatrix} \\ &= \begin{pmatrix} \Omega_{aa}^{(m)}(t) & \Omega_{ab}^{(m)}(t) \\ \Omega_{ba}^{(m)}(t) & \Omega_{bb}^{(m)}(t) \end{pmatrix},\end{aligned}$$

where $D^{(m)} = \begin{pmatrix} D_a^{(m)} & 0 \\ 0 & D_b^{(m)} \end{pmatrix}$, $D_a^{(m)}$ and $D_b^{(m)}$ are $(q \times q)$ diagonal matrices. Note that

$$[\Omega^{(m)}(t)]^{-1} = \begin{pmatrix} A_{11.2}^{-1} & -A_{11}^{-1}A_{12}A_{22.1}^{-1} \\ -A_{22}^{-1}A_{21}A_{11.2}^{-1} & A_{22.1}^{-1} \end{pmatrix}, \tag{18}$$

where

$$\begin{aligned} A_{11.2} &= \Omega_{aa}^{(m)}(t) - \Omega_{ab}^{(m)}(t)(\Omega_{bb}^{(m)}(t))^{-1}\Omega_{ba}^{(m)}(t), \quad A_{11} = \Omega_{aa}^{(m)}(t), \quad A_{22} = \Omega_{bb}^{(m)}(t), \\ A_{22.1} &= \Omega_{bb}^{(m)}(t) - \Omega_{ba}^{(m)}(t)(\Omega_{aa}^{(m)}(t))^{-1}\Omega_{ab}^{(m)}(t), \quad A_{12} = \Omega_{ab}^{(m)}(t), \quad A_{21} = \Omega_{ba}^{(m)}(t). \end{aligned}$$

Because $\hat{\eta}_{\lambda}^{(m+1)}(t) = (\left(\hat{a}_{\lambda}^{(m+1)}(t)\right)^\top, h\left(\hat{b}_{\lambda}^{(m+1)}(t)\right)^\top)^\top$, $\hat{a}_{\lambda}^{(m+1)}(t) = (\hat{g}_{\lambda 1}^{(m+1)}(t), \ldots, \hat{g}_{\lambda q}^{(m+1)}(t))^\top$ and $h\hat{b}_{\lambda}^{(m+1)}(t) = (h\hat{g}_{\lambda 1}^{\prime(m+1)}(t), \ldots, h\hat{g}_{\lambda q}^{\prime(m+1)}(t))^\top$, and for a given t, $\Omega_{ab}^{(m)}(t) = \Omega_{ba}^{(m)\top}(t) = hO_p(C_q)$, C_q is a $(q \times q)$ constant matrix, we have

$$\begin{aligned}\hat{a}_{\lambda}^{(m+1)}(t) &= \left[\Omega_{aa}^{(m)}(t) - h^2O_p(C_q)[\Omega_{bb}^{(m)}(t)]^{-1}O_p^\top(C_q)\right]^{-1} N_a(t) \\ &\quad - h\left\{\Omega_{aa}^{(m)}(t)\right\}^{-1} O_p(C_q) \\ &\quad \times \left[\Omega_{bb}^{(m)}(t) - h^2O_p^\top(C_q)[\Omega_{aa}^{(m)}(t)]^{-1}O_p(C_q)\right]^{-1} N_b(t),\end{aligned} \tag{19}$$

where

$$N_a(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i Z_i K_h(t_i - t), \quad N_b(t) = \frac{1}{n}\sum_{i=1}^{n} Y_i Z_i \Big(\frac{t_i - t}{h}\Big) K_h(t_i - t),$$

and

$$\begin{aligned} & h\hat{b}_{\lambda}^{(m+1)}(t) \\ & \quad = -h\Big\{\Omega_{bb}^{(m)}(t)\Big\}^{-1} O_p^{\top}(C_q)\Big[\Omega_{aa}^{(m)}(t) - h^2 O_p(C_q)[\Omega_{bb}^{(m)}(t)]^{-1} O_p^{\top}(C_q)\Big]^{-1} N_a(t) \\ & \qquad + \Big[\Omega_{bb}^{(m)}(t) - h^2 O_p^{\top}(C_q)[\Omega_{aa}^{(m)}(t)]^{-1} O_p(C_q)\Big]^{-1} N_b(t). \end{aligned} \tag{20}$$

Consider $\hat{a}_{\lambda}^{(m+1)}(t) = (\hat{a}_{\lambda a}^{(m+1)\top}(t), \hat{a}_{\lambda b}^{(m+1)\top}(t))^{\top}$, where

$$\hat{a}_{\lambda a}^{(m+1)}(t) = (\hat{g}_{\lambda 1}^{(m+1)}(t), \ldots, \hat{g}_{\lambda q_0}^{(m+1)}(t))^{\top} \text{ and } \hat{a}_{\lambda b}^{(m+1)}(t) = (\hat{g}_{\lambda(q_0+1)}^{(m+1)}(t), \ldots, \hat{g}_{\lambda q}^{(m+1)}(t))^{\top},$$

and $\hat{b}_{\lambda}^{(m+1)} = (\hat{b}_{\lambda a}^{(m+1)\top}(t), \hat{b}_{\lambda b}^{(m+1)\top}(t))^{\top}$, where

$$\hat{b}_{\lambda a}^{(m+1)}(t) = (\hat{g}_{\lambda 1}^{\prime(m+1)}(t), \ldots, \hat{g}_{\lambda d_0}^{\prime(m+1)}(t))^{\top} \text{ and } \hat{b}_{\lambda b}(t) = (\hat{g}_{\lambda(d_0+1)}^{\prime(m+1)}(t), \ldots, \hat{g}_{\lambda q}^{\prime(m+1)}(t))^{\top}.$$

Define $D_a^{(m)} = \begin{pmatrix} D_{aa}^{(m)} & 0 \\ 0 & D_{ab}^{(m)} \end{pmatrix}$, $D_{aa}^{(m)}$ is the $q_0 \times q_0$ diagonal submatrix of $D_a^{(m)}$, and $D_{ab}^{(m)}$ is the lower $(q - q_0) \times (q - q_0)$ diagonal submatrix of $D_a^{(m)}$. Define $D_b^{(m)} = \begin{pmatrix} D_{ba}^{(m)} & 0 \\ 0 & D_{bb}^{(m)} \end{pmatrix}$, $D_{ba}^{(m)}$ is the $d_0 \times d_0$ diagonal submatrix of $D_b^{(m)}$, and $D_{bb}^{(m)}$ is the lower $(q - d_0) \times (q - d_0)$ diagonal submatrix of $D_b^{(m)}$.

Therefore, for any fixed n, with $m \to \infty$, by the definition of $D^{(m)}$ and (17), we have every diagonal component of $D_{aa}^{(m)}$ and $D_{ba}^{(m)}$ converges to some finite number, and the diagonal components of $D_{ab}^{(m)}$ and $D_{bb}^{(m)}$ converge to infinity. Therefore, by (19),

$$\hat{a}_{\lambda}^{(m+1)}(t) = \Big\{[\Omega_{aa}^{(m)}(t)]^{-1} N_a(t) - h[\Omega_{aa}^{(m)}(t)]^{-1} O_p(C_q)[\Omega_{bb}^{(m)}(t)]^{-1} N_b(t)\Big\}(1 + o_p(1)). \tag{21}$$

We find that

$$\begin{aligned} \Omega_{aa}^{(m)}(t) &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} Z_i^{*} Z_i^{*\top} K_h(t_i - t) + \frac{D_{aa}^{(m)}}{n} & \frac{1}{n}\sum_{i=1}^{n} Z_i^{*} Z_i^{**\top} K_h(t_i - t) \\ \frac{1}{n}\sum_{i=1}^{n} Z_i^{**} Z_i^{*\top} K_h(t_i - t) & \frac{1}{n}\sum_{i=1}^{n} Z_i^{**} Z_i^{**\top} K_h(t_i - t) + \frac{D_{ab}^{(m)}}{n} \end{pmatrix} \\ &= \begin{pmatrix} \Omega_{aa1}^{(m)}(t) & \Omega_{aa2}^{(m)}(t) \\ \Omega_{aa3}^{(m)}(t) & \Omega_{aa4}^{(m)}(t) \end{pmatrix}, \end{aligned}$$

where $Z_i = (Z_i^{*\top}, Z_i^{**\top})^{\top}$, the dimensions of Z_i^{*} and Z_i^{**} are q_0 and $q - q_0$, respectively. Similar to (18), define

$$[\Omega_{aa}^{(m)}(t)]^{-1} = \begin{pmatrix} B_{11.2}^{-1} & -B_{11}^{-1}B_{12}B_{22.1}^{-1} \\ -B_{22}^{-1}B_{21}B_{11.2}^{-1} & B_{22.1}^{-1} \end{pmatrix},$$

where

$$\begin{aligned} B_{11.2} &= \Omega_{aa1}^{(m)}(t) - \Omega_{aa2}^{(m)}(t)(\Omega_{aa4}^{(m)}(t))^{-1}\Omega_{aa3}^{(m)}(t), \quad B_{11} = \Omega_{aa1}^{(m)}(t), \quad B_{22} = \Omega_{aa4}^{(m)}(t), \\ B_{22.1} &= \Omega_{aa4}^{(m)}(t) - \Omega_{aa3}^{(m)}(t)(\Omega_{aa1}^{(m)}(t))^{-1}\Omega_{aa2}^{(m)}(t), \quad B_{12} = \Omega_{aa2}^{(m)}(t), \quad B_{21} = \Omega_{aa3}^{(m)}(t). \end{aligned}$$

Since every diagonal component of $D_{aa}^{(m)}$ converges to some finite number, and the diagonal components of $D_{ab}^{(m)}$ converge to infinity, $-B_{22}^{-1}B_{21}B_{11.2}^{-1}$ and $B_{22.1}^{-1}$ converge to 0 uniformly on $t \in \mathbf{T}$ as $m \to \infty$. Furthermore, $N_a(t) = (N_{aa}^\top(t), N_{ab}^\top(t))$, $N_b(t) = (N_{ba}^\top(t), N_{bb}^\top(t))$, $N_{aa}(t), N_{ab}(t), N_{ba}(t)$ and $N_{bb}(t)$ are uniformly bounded, $[\Omega_{bb}^{(m)}(t)]^{-1}$ is uniformly bounded, after expanse (21) it proves that $\hat{a}_{\lambda j}^{(m+1)}(t) \to 0$, as $m \to \infty$, for $q_0 < j \le q$, uniformly on $t \in \mathbf{T}$. Therefore, we conclude that $P\big(\sup_{t \in T} \|\hat{g}_{\lambda j}(t)\| = 0\big) \to 1$ for any $q_0 < j \le q$. Then, (i) in Theorem 1 is proved.

To prove (ii) in Theorem 1, we consider (20), we have

$$\begin{aligned} h\hat{b}_{\lambda}^{(m+1)}(t) = & -h\Big\{\Omega_{bb}^{(m)}(t)\Big\}^{-1} O_p^\top(C_q)\big[\Omega_{aa}^{(m)}(t)\big]^{-1} N_a(t) \\ & + \Big[\Omega_{bb}^{(m)}(t)\Big]^{-1} N_b(t). \end{aligned} \tag{22}$$

Define

$$\begin{aligned} \Omega_{bb}^{(m)}(t) &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n Z_i^* Z_i^{*\top}\left(\frac{t_i-t}{h}\right)^2 K_h(t_i - t) + \frac{D_{ba}^{(m)}}{n} & \frac{1}{n}\sum_{i=1}^n Z_i^* Z_i^{**\top}\left(\frac{t_i-t}{h}\right)^2 K_h(t_i - t) \\ \frac{1}{n}\sum_{i=1}^n Z_i^{**} Z_i^{*\top}\left(\frac{t_i-t}{h}\right)^2 K_h(t_i - t) & \frac{1}{n}\sum_{i=1}^n Z_i^{**} Z_i^{**\top}\left(\frac{t_i-t}{h}\right)^2 K_h(t_i - t) + \frac{D_{bb}^{(m)}}{n} \end{pmatrix} \\ &= \begin{pmatrix} \Omega_{bb1}^{(m)}(t) & \Omega_{bb2}^{(m)}(t) \\ \Omega_{bb3}^{(m)}(t) & \Omega_{bb4}^{(m)}(t) \end{pmatrix}, \end{aligned}$$

where $Z_i = (Z_i^{*\top}, Z_i^{**\top})^\top$, the dimensions of Z_i^* and Z_i^{**} are d_0 and $q - d_0$, respectively. Since every diagonal component of $D_{ba}^{(m)}$ converges to some finite number, and the diagonal components of $D_{bb}^{(m)}$ converge to infinity, similar to the proof of estimation sparsity of $\hat{g}_\lambda(t)$, we have $P\Big(\sup_{t \in T} \|\hat{g}'_{\lambda j}(t)\| = 0\Big) \to 1$ for any $d_0 < j \le q$. Thus, this theorem is proved. □

Proof of Theorem 2 Theorem 1 proves the estimation sparsity of $\hat{g}_\lambda(t)$ and $\hat{g}'_\lambda(t)$, together with (4), we have the estimating equation

$$0 = -2\sum_{i=1}^n [Y_i - \eta_{\lambda \mathbf{A}_{G_M}^*}^\top(t)\tilde{Z}_i^*(t)]\tilde{Z}_i^*(t)K_h(t_i - t) + \begin{pmatrix} D_c & 0 \\ 0 & D_d \end{pmatrix}\hat{\eta}_{\lambda \mathbf{A}_{G_M}^*}(t), \tag{23}$$

where $\tilde{Z}_i^*(t) = (Z_i^{*\top}, Z_i^{**\top}(\frac{t_i-t}{h}))$, $Z_i^* = (Z_{i1}, \dots, Z_{iq_0})^\top$ and $Z_i^{**} = (Z_{i1}, \dots, Z_{id_0})^\top$, $D_c =$ diag $(\frac{\lambda_{11}}{\|\hat{c}_1\|}, \dots, \frac{\lambda_{1q_0}}{\|\hat{c}_{q_0}\|})$ and $D_d =$ diag $(\frac{\lambda_{21}}{\|\hat{d}_1\|}, \dots, \frac{\lambda_{2d_0}}{\|\hat{d}_{d_0}\|})$. By (23),

$$
\begin{aligned}
\hat{\eta}_{\lambda \mathbf{A}^*_{G_M}}(t) &= \left[\frac{1}{n}\sum_{i=1}^{n} \tilde{Z}_i^*(t)\tilde{Z}_i^{*\top}(t)K_h(t_i-t) + \frac{1}{2n}\begin{pmatrix} D_c & 0 \\ 0 & D_d \end{pmatrix}\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\tilde{Z}_i^*(t)K_h(t_i-t)\right] \\
&= \left[\hat{\Sigma}_1^*(t)\right]^{-1}\hat{\Sigma}_1^*(t)\eta_{0\mathbf{A}^*_{G_M}}(t) + \left[\hat{\Sigma}_1^*(t)\right]^{-1}\frac{h^2}{2n}\sum_{i=1}^{n}\left(\frac{t_i-t}{h}\right)^2 g''^{\top}_{\mathbf{A}^*_g}(t)Z_i^*\tilde{Z}_i^*K_h(t_i-t) \\
&\quad + \left[\hat{\Sigma}_1^*(t)\right]^{-1}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\tilde{Z}_i^*K_h(t_i-t),
\end{aligned} \tag{24}
$$

where $\hat{\Sigma}_1^*(t) = \frac{1}{n}\sum_{i=1}^{n}\tilde{Z}_i^*(t)\tilde{Z}_i^{*\top}(t)K_h(t_i-t)$.

To prove the second equation in (24), according to

$$
\|\hat{c}_j\|^2 = \sum_{i=1}^{n}(\hat{g}_j(t_i)-g_j(t_i))^2 + 2\sum_{i=1}^{n} g_j(t_i)(\hat{g}_j(t_i)-g_j(t_i)) + \sum_{i=1}^{n} g_j^2(t_i) = O_p(n),
$$

for $j = 1, \ldots, q_0$, together with $n^{-\frac{11}{10}}\phi_n \to 0$, we have

$$
\frac{1}{n}\frac{\lambda_{1j}}{\|\hat{c}_j\|} = \frac{n^{-11/10}\lambda_{1j}}{n \cdot n^{-11/10} \cdot n^{1/2}} = n^{-11/10}\lambda_{1j} \cdot O_p(n^{-2/5}) = o_p(n^{-2/5}).
$$

Similarly, we can obtain $\frac{1}{2n}\begin{pmatrix} D_c & 0 \\ 0 & D_d \end{pmatrix} = o_p(n^{-2/5})$. Thus the second equation in (24) is proved.

Note that

$$
\hat{\Sigma}_1^*(t) \to \Sigma_1^*(t) = \begin{pmatrix} f(t)E(Z^*Z^{*\top}|t) & 0 \\ 0 & f(t)E(Z^{**}Z^{**\top}|t)\mu_2 \end{pmatrix},
$$

and $\sqrt{\frac{h}{n}}\sum_{i=1}^{n}\varepsilon_i\tilde{Z}_i^*K_h(t_i-t)$ follows the normal distribution $N(0, \sigma^2 V_1^*)$,

$$
V_1^* = \begin{pmatrix} f(t)E(Z^*Z^{*\top}|t)\kappa_2 & 0 \\ 0 & f(t)E(Z^{**}Z^{**\top}|t)\mu_2 \end{pmatrix},
$$

thus, by (24), we conclude

$$
\begin{aligned}
&\sqrt{nh}\left(\hat{\eta}_{\lambda\mathbf{A}^*_{G_M}}(t) - \eta_{0\mathbf{A}^*_{G_M}}(t) - \frac{h^2}{2}[\Sigma_1^*(t)]^{-1}\begin{pmatrix} f(t)g''_{\mathbf{A}^*_g}(t)E(Z^*Z^{*\top}|t)\mu_2 \\ 0 \end{pmatrix}\right) \\
&\xrightarrow{\mathcal{D}} N(0, \sigma^2[\Sigma_1^*(t)]^{-1}V_1^*[\Sigma_1^*(t)]^{-1}).
\end{aligned}
$$

□

Proof of Theorem 3 To prove (i), we consider $\hat{\beta}^{(1)}_{\lambda_n} = \beta^{(1)}_0 + Cn^{-1/2}$ and let $\hat{\beta}^{(1)}_{\lambda_n} = (\hat{\beta}^{(1)*\top}_{\lambda_n}, \mathbf{0}^{\top}_{(p-p_0)\times 1})^{\top}$, where $\hat{\beta}^{(1)*}_{\lambda_n} = \beta^{(1)*}_0 + n^{-1/2}\Sigma_3^{*-1}\frac{1}{\sqrt{n}}U_*(\beta^{(1)}_0)$, $\beta^{(1)*}_0 = (\beta_{02}, \ldots, \beta_{0p_0})^{\top}$ and

$$U_*(\beta_0^{(1)}) = \sum_{i=1}^{n}(Y_i - Z_i^{*\top} g^*(t_i))[g'^{*\top}(t_i) Z_i^* J^{\top}_{\beta_0^{(1)*}} X_i^* - \tilde{D}_2^*(t_i) D_1^{*-1}(t_i) Z_i^*],$$

where $t_i = X_i^\top \beta_0 = X_i^{*\top}\beta_0^*$, $X_i^* = (X_{i1}, \ldots, X_{ip_0})^\top$, $D_1^*(X^\top\beta_0) = E[Z^* Z^{*\top} | X^\top \beta_0]$, $\tilde{D}_2^*(X^\top\beta_0) = E[g'^{*\top}(X^\top\beta_0) Z^* J^\top_{\beta_0^{*(1)}} X^* Z^{*\top} | X^\top\beta_0]$ and $\Sigma_3^* = E\Big[g'^{*\top}(X^\top\beta_0) Z^* J^\top_{\beta_0^{(1)*}} X^* - \tilde{D}_2^*(X^\top\beta_0) D_1^{*-1}(X^\top\beta_0) Z^*\Big]^{\otimes 2}$.

Similar to the proof of Theorem 2 in Lai et al. (2016), we can conclude

$$\frac{1}{\sqrt{n}} U_{\lambda_n}(\hat{\beta}_{\lambda_n}^{(1)}) = \frac{1}{\sqrt{n}} \tilde{U}(\beta_0^{(1)}) - \Sigma_2 \cdot \sqrt{n}(\hat{\beta}_{\lambda_n}^{(1)} - \beta_0^{(1)}) - \sqrt{n}\boldsymbol{q}_{\lambda_n}(|\hat{\beta}_{\lambda_n}^{(1)}|)\mathrm{sgn}(\hat{\beta}_{\lambda_n}^{(1)}) + o_p(1), \tag{25}$$

where

$$\begin{aligned}\tilde{U}(\beta_0^{(1)}) = &\sum_{i=1}^{n}(Y_i - Z_i^{*\top} g^*(X_i^\top\beta_0))[g'^{*\top}(X_i^\top\beta_0) Z_i^* J^\top_{\beta_0^{(1)}} X_i \\ &- D_2^*(X_i^\top\beta_0) D_1^{-1}(X_i^\top\beta_0) Z_i^*].\end{aligned}$$

Since the true parameter vector $\beta_0 = (\beta_0^{*\top}, \mathbf{0}^\top_{(p-p_0)\times 1})^\top$, $\beta_0^* = (\beta_{01}, \ldots, \beta_{0p_0})^\top$, thus

$$J_{\beta_0^{(1)*}} = \begin{pmatrix} -\beta_0^{(1)\top} / \sqrt{1 - \|\beta_0^{(1)}\|^2} \\ \boldsymbol{I}_{p_0-1} \end{pmatrix}, \quad J_{\beta_0^{(1)}} = \begin{pmatrix} J_{\beta_0^{(1)*}} & \mathbf{0}_{1\times(p-p_0)} \\ \mathbf{0}_{(p-p_0)\times 1} & \boldsymbol{I}_{(p-p_0)\times(p-p_0)} \end{pmatrix}.$$

Then $J^\top_{\beta_0^{(1)}} X = \begin{pmatrix} J^\top_{\beta_0^{(1)*}} X^* \\ X^{**} \end{pmatrix}$, where $X^{**} = (X_{p_0+1}, \ldots, X_p)^\top$. It is easy to see that $U_*(\beta_0^{(1)})$ are same with the first $(p_0 - 1)$ components of $U(\beta_0^{(1)})$.

By the definition of $\hat{\beta}_{\lambda_n}^{(1)*}$ and (25), for $j = 2, \ldots, p_0$,

$$\frac{1}{\sqrt{n}} U_{\lambda_n j}([\hat{\beta}_{\lambda_n} \pm \epsilon e_j]^{(1)}) = o_p(1) - \sqrt{n} q_{\lambda_n}(|\hat{\beta}_{\lambda_n j} \pm \epsilon|)\mathrm{sgn}(\hat{\beta}_{\lambda_n j} \pm \epsilon) = o_p(1). \tag{26}$$

The last equation is due to the fact that $p_{\lambda_n}(\cdot)$ is a SCAD penalty and $\sqrt{n} q_{\lambda_n}(|\hat{\beta}_{\lambda_n j}|) \to 0$ for $j = 2, \ldots, p_0$, by the condition $\lambda_n \to 0$, $\sqrt{n}\lambda_n \to \infty$.

In addition, we denote a^{**} as the last $p - p_0$ components of $p - 1$ dimensional vector a for convenience. For example, $\hat{\beta}_{\lambda_n}^{(1)**} = (\hat{\beta}_{\lambda_n(p_0+1)}, \ldots, \hat{\beta}_{\lambda_n p})^\top$ and $\beta_0^{(1)**} = (\beta_{0(p_0+1)}, \ldots, \beta_{0p})^\top$. From the definition, we know that $\hat{\beta}_{\lambda_n}^{(1)**} = \beta_0^{(1)**} = \mathbf{0}_{(p-p_0)\times 1}$. Moreover, we find that $\frac{1}{\sqrt{n}}\tilde{U}(\beta_0^{(1)}) = O_p(1)$. Thus, the last $(p - p_0)$ components of (25) reduce to

$$\frac{1}{\sqrt{n}} U_{\lambda_n}^{**}([\hat{\beta}_{\lambda_n} \pm \epsilon e_j]^{(1)}) = O_p(1) - \sqrt{n}\boldsymbol{q}_{\lambda_n}^{**}(|\epsilon|)\boldsymbol{sgn}^{**}(\pm\epsilon).$$

Since $\sqrt{n}\lambda_n \to \infty$, $\frac{1}{\sqrt{n}} U_{\lambda_n}^{**}([\hat{\beta}_{\lambda_n} \pm \epsilon e_j])$ are dominated by $\sqrt{n}\boldsymbol{q}_{\lambda_n}^{**}(|\epsilon|)\boldsymbol{sgn}^{**}(\pm\epsilon)$ as ϵ tends to zero. Based on the above discussions, we have proved that there exists a $\sqrt{n}$

-consistent zero-crossing to $U_{\lambda_n}(\beta^{(1)})$. Recall that $\mathbf{A}^*_{\beta^{(1)}} = \{2, \dots, p_0\}$. Following the discussion of Theorem 1(c) in Johnson et al. (2008), we conclude that there exists an exact zero-crossing estimator $\hat{\beta}^{(1)}_{\lambda_n} = (\hat{\beta}^{(1)\top}_{\lambda_n \mathbf{A}^*_{\beta^{(1)}}}, \mathbf{0}^\top)^\top$ of $U_{\lambda_n}(\beta^{(1)})$ satisfied $U_{\lambda_n \mathbf{A}^*_{\beta^{(1)}}}(\hat{\beta}^{(1)}_{\lambda_n}) = 0$. Then (i) in Theorem 3 is proved.

Next we prove the estimation sparsity. Similar to the proof of Theorem 1(b) in Johnson et al. (2008), define $B_j = \{\hat{\beta}_{\lambda_n j} \neq 0\}, j = p_0 + 1, \dots, p$. To prove (ii) in Theorem 3, we only need to prove for any $\epsilon > 0$, with $n \to \infty$, $P(B_j) < \epsilon$. Because $\hat{\beta}_{\lambda_n j} = O_p(n^{-1/2})$, there exists some C such that with n is large enough,

$$P(B_j) < \frac{\epsilon}{2} + P\{\hat{\beta}_{\lambda_n j} \neq 0, |\hat{\beta}_{\lambda_n j}| < Cn^{-1/2}\}.$$

On the other hand, from (26), the first two terms on the right side are of order $O_p(1)$, which implies that there exists some C' such that for large n,

$$P\{\hat{\beta}_{\lambda_n j} \neq 0, |\hat{\beta}_{\lambda_n j}| < Cn^{-1/2}, \sqrt{n} q_{\lambda_n}(|\hat{\beta}_{\lambda_n j}|) > C'\} < \frac{\epsilon}{2}.$$

Since $\sqrt{n} \inf_{|\beta_j| \leq Cn^{-1/2}} q_{\lambda_n}(|\beta_j|) = \sqrt{n}\lambda_n \to \infty$, we have

$$\begin{aligned} &P\{\hat{\beta}_{\lambda_n j} \neq 0, |\hat{\beta}_{\lambda_n j}| < Cn^{-1/2}\} \\ &\quad = P\{\hat{\beta}_{\lambda_n j} \neq 0, |\hat{\beta}_{\lambda_n j}| < Cn^{-1/2}, \sqrt{n} q_{\lambda_n}(|\hat{\beta}_{\lambda_n j}|) > C'\} < \frac{\epsilon}{2}. \end{aligned}$$

Therefore, $P(B_j) < \epsilon$, for any $\epsilon > 0$, $j = p_0 + 1, \dots, p$. The consistent selection of Theorem 3 (ii) is proved.

To prove the asymptotic normality of Theorem 3, by (25), after the Taylor expansion of the penalty term, we have

$$o_p(1) = \frac{1}{\sqrt{n}} U_*(\beta^{(1)}_0) - \Sigma^*_3 \cdot \sqrt{n}(\hat{\beta}^{(1)*}_{\lambda_n} - \beta^{(1)*}_0) - \Sigma^*_4 \sqrt{n}(\hat{\beta}^{(1)*}_{\lambda_n} - \beta^{(1)*}_0) - \sqrt{n} B^*_n,$$

where

$$\begin{aligned} B^*_n &= (q_{\lambda_n}(|\beta_{02}|)\mathrm{sgn}(\beta_{02}), \dots, q_{\lambda_n}(|\beta_{0p_0}|)\mathrm{sgn}(\beta_{0p_0}))^\top, \\ \Sigma^*_4 &= \mathrm{diag}(q'_{\lambda_n}(|\beta_{02}|)\mathrm{sgn}(\beta_{02}), \dots, q'_{\lambda_n}(|\beta_{0p_0}|)\mathrm{sgn}(\beta_{0p_0})). \end{aligned}$$

Therefore, we prove that

$$\left[\Sigma^*_3 + \Sigma^*_4\right] \sqrt{n}\left(\hat{\beta}^{(1)*}_{\lambda_n} - \beta^{(1)*}_0 + \left[\Sigma^*_3 + \Sigma^*_4\right]^{-1} B^*_n\right) \xrightarrow{\mathcal{D}} N(0, \Sigma^*_5),$$

where $\Sigma^*_5 = \sigma^2 \Sigma^*_3$. □

References

Fan, J., Gijbels, I. (1996). *Local polynomial modelling and its applications*, Vol. 66. Boca Raton: Chapman & Hall/CRC.

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. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, *1*(1), 179–195.

Feng, S., Xue, L. (2015). Model detection and estimation for single-index varying coefficient model. *Journal of Multivariate Analysis*, *139*(C), 227–244.

Huang, J., Horowitz, J. L., Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, *36*(2), 587–613.

Huang, J., Wei, F., Ma, S. (2012). Semiparametric regression pursuit. *Statistica Sinica*, *22*(4), 1403–1426.

Huang, Z., Zhang, R. (2013). Profile empirical-likelihood inferences for the single-index-coefficient regression model. *Statistics and Computing*, *23*, 455–465.

Hunter, D., Li, R. (2005). Variable selection using mm algorithms. *Annals of Statistics*, *33*, 1617–1642.

Johnson, B., Lin, D., Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, *103*, 672–680.

Lai, P., Wang, Q., Zhou, X. H. (2014). Variable selection and semiparametric efficient estimation for the heteroscedastic partially linear single-index model. *Computational Statistics and Data Analysis*, *70*(2), 241–256.

Lai, P., Zhang, Q., Lian, H., Wang, Q. (2016). Efficient estimation for the heteroscedastic single-index varying coefficient models. *Statistics and Probability Letters*, *110*, 84–93.

Ma, S., Du, P. (2012). Variable selection in partly linear regression model with diverging dimensions for right censored data. *Statistica Sinica*, *22*(3), 1003–1020.

Song, Y., Jian, L., Lin, L. (2016). Robust exponential squared loss-based variable selection for high-dimensional single-index varying-coefficient model. *Journal of Computational and Applied Mathematics*, *308*, 330–345.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)*, *58*, 267–288.

Wang, H., Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, *102*, 1039–1048.

Wang, H., Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, *104*, 747–757.

Wu, C., Cui, Y., Ma, S. (2014). Integrative analysis of gene-environment interactions under a multi-response partially linear varying coefficient model. *Statistics in Medicine*, *33*(28), 4988–4998.

Wu, C., Shi, X., Cui, Y., Ma, S. (2015). A penalized robust semiparametric approach for gene-environment interactions. *Statistics in Medicine*, *34*(30), 4016–4030.

Wu, C., Zhong, P. S., & Cui, Y. (2018). Additive varying-coefficient model for nonlinear gene-environment interactions. *Statistical Applications in Genetics and Molecular Biology*. https://doi.org/10.1515/sagmb-2017-0008.

Xue, L., Pang, Z. (2013). Statistical inference for a single-index varying-coefficient model. *Statistics and Computing*, *23*(5), 589–599.

Xue, L., Wang, Q. (2012). Empirical likelihood for single-index varying-coefficient models. *Bernoulli*, *18*, 836–856.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.