Journal of Multivariate Analysis 112 (2012) 117–129

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

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

Parametric component detection and variable selection in varying-coefficient partially linear models

Dewei Wang, K.B. Kulasekera *

Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975, United States

ARTICLE INFO

Article history:
Received 24 August 2011
Available online 7 June 2012

AMS subject classification:
62G08

Keywords:
Parametric component detection
Variable selection
Adaptive LASSO
Oracle property
Varying-coefficient partially linear model

ABSTRACT

In this paper we are concerned with detecting the true structure of a varying-coefficient partially linear model. The first issue is to identify whether a coefficient is parametric. The second issue is to select significant covariates in both nonparametric and parametric portions. In order to simultaneously address both issues, we propose to combine local linear smoothing and the adaptive LASSO and penalize both the coefficient functions and their derivatives using an adaptive L_1 penalty. We give conditions under which this new adaptive LASSO consistently identifies the significant variables and parametric components along with estimation sparsity. Simulated and real data analysis demonstrate the proposed methodology.

1. Introduction

Semiparametric regression models have recently gained much popularity due to their flexibility of nonparametric modeling and explanatory power of parametric modeling. Let Y be a response associated with covariates $(U, \mathbf{W}, \mathbf{Z})$. Further, denote $E(Y|U = u, \mathbf{W} = \mathbf{w}, \mathbf{Z} = \mathbf{z}) = \mu(u, \mathbf{w}, \mathbf{z})$. The varying-coefficient partially linear model (VCPLM) assumes that

$$\mu(u, \mathbf{w}, \mathbf{z}) = \mathbf{w}^\top \alpha(u) + \mathbf{z}^\top \theta, \tag{1}$$

where $\alpha(\cdot)$ is a vector consisting of unknown coefficient functions and θ is a vector of unspecified regression coefficients. We do not specify the lengths of these vectors at this point except that there are a total of p components, some of which can be zero. In the sequel when we rewrite (1) in a suitable form, we shall define the lengths of each coefficient vector. The term $\mathbf{w}^\top \alpha(u)$ is referred to as the nonparametric component and the term $\mathbf{z}^\top \theta$ is called the parametric component. The structure of model (1) covers many existing parametric, semiparametric or nonparametric models, such as linear models, partially linear models [13], semi-varying coefficient models [23,6] varying coefficient models [14,2] and nonparametric regression models [4]. Different model structures warrant different estimation procedures. Thus, it is important to find out which model structure is the most suitable for the given data. Our proposed methodology provides a general framework of model structure detection for the VCPLM.

Variable selection plays a fundamental role in statistical model detection. It is well-known that missing significant coefficients would result in huge estimation bias; and including spurious coefficients would degrade the estimation efficiency. In parametric variable selection, traditional methods such as AIC, BIC, the best subset selection, etc., suffer from huge computational burden [7]. In order to simultaneously select the significant variables and estimate the unknown regression coefficients, Tibshirani [20] investigated the LASSO method which shrinks estimates with an L_1 penalty. Based

* Corresponding author.
E-mail addresses: dwang@clemson.edu (D. Wang), kk@clemson.edu (K.B. Kulasekera).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.05.006

on a non-concave penalized likelihood, Fan and Li [7] proposed a family of shrinkage methods using the smoothly clipped absolute deviation (SCAD) penalty which achieves the oracle properties. Zou [25] remedied the possible inconsistencies in selection by LASSO using adaptive weights in the L_1 penalty. Alternative shrinkage methods have been discussed by Breiman [1], Fu [11], Yuan and Lin [22] and Zou and Li [26] among others.

Extending these shrinkage ideas to the VCPLM is challenging due to the complexity of the nonparametric construction. Li and Liang [18] presented a variable selection method in generalized varying-coefficient partially linear models where they use the SCAD method to select parametric coefficients and use a generalized likelihood ratio test for identifying nonparametric coefficient functions. Another possible approach is the spline based selection method [24,15]. Recently, Wang and Xia [21] have studied a nonparametric variable selection method for a varying coefficient model based on local-constant kernel smoothing and the group LASSO (referred to as KLASSO), which enjoys the asymptotic estimation sparsity and the same efficiency as the oracle estimator. However, the computational burden of solving the group LASSO is very high [22] and without local linear smoothing the estimation accuracy may be impacted [5].

Besides variable selection, parametric component detection is important to discover the underlying structure of the VCPLM. Since the optimal parametric estimation rate is root n and the optimal nonparametric estimation rate is $n^{2/5}$, treating a parametric component as a nonparametric function would be inefficient. Hypothesis testing has been developed for parametric component detection in varying coefficient models for testing if an $\alpha_j(\cdot)$ is an unknown constant by Fan and Zhang [8] whose test was based on maximum deviation. Fan et al. [10] constructed generalized likelihood ratio tests for the same hypothesis testing problem. In an attempt to simultaneously select variables, detect parametric components, and estimate regression coefficients, we investigate a new shrinkage method combining local linear smoothing and the adaptive LASSO.

In our approach we assume that θ in model (1) is also a vector consisting of unknown functions of u. Then we define a p-vector $\beta(u)$ by suitably rearranging the components of the vector $(\alpha^\top(u), \theta^\top)^\top$ of all the coefficients in (1). Our objective is to determine which components of $\beta(u)$ are nonzero functions or nonzero constants. For any given index u, we apply the adaptive LASSO to estimate components of $\beta(u)$ from a locally weighted least squares. In contrast to [21]'s local-constant approach, we use a local linear approximation to each component of $\beta(\cdot)$ [5,9]. A typical assumption in the inference for varying coefficient models is that the coefficient functions are smooth (second order derivative is bounded). Continuing to make this assumption, we propose to detect a parametric component by detecting if a varying coefficient function has a zero derivative. Hence, we penalize not only the coefficients but also the derivatives with random uniform (in u) weights. Under this setting, we show the correct model sparsity and consistency in identifying all the components. Furthermore, the oracle properties of the nonzero coefficient function estimators are established. Moreover, our simulations show that by solving the adaptive LASSO problem using the popular Least Angle Regression (LARS) procedure [3], the proposed method provides significant computational efficiency over the KLASSO method. The application of the proposed method to Boston Housing Data reveals that a prior analysis using varying coefficients might have missed an important feature of a coefficient function. That is, a coefficient function that shows up as nonzero in [21]'s analysis might in fact be a step function.

The paper is organized as follows. In Section 2, we provide the new estimation procedure and its theoretical properties and its implementation. In Section 3, simulated numerical experiments and a real data analysis are reported followed by a discussion in Section 4. All technical proofs are relegated to the Appendix.

2. Methodology

Suppose $\{(Y_i, U_i, \mathbf{W}_i, \mathbf{Z}_i), i = 1, \ldots, n\}$ constitutes a random sample generated from model (1). We have

$$Y_i = \mathbf{W}_i^\top \alpha(U_i) + \mathbf{Z}_i^\top \theta + \varepsilon_i, \quad i = 1, \ldots, n$$

where the components of $\alpha(\cdot)$ are all smooth functions with bounded second order derivatives, a commonly used technical assumption; ε_i's are the error terms satisfying $E(\varepsilon_i|U_i, \mathbf{W}_i, \mathbf{Z}_i) = 0$ and $\mathrm{Var}(\varepsilon_i|U_i, \mathbf{W}_i, \mathbf{Z}_i) = \sigma^2(U_i)$. Here we assume that there are exactly r component functions of the $\alpha(u)$ and exactly $q - r$, $q \geq r$, components of the θ vector are nonzero. Without loss of generality, let $\beta(u)$ be the p-vector that is a permutation of the components of $(\alpha^\top(u), \theta^\top)^\top$ where its first r components are the nonzero functions of the $\alpha(\cdot)$ vector, the next $q - r$ components are the nonzero components of the θ vector followed by $p - q$ zeros. We denote the corresponding permutation of $(\mathbf{W}_i^\top, \mathbf{Z}_i^\top)^\top$ by X_i. Then, without prior information on the partial linear structure, model (1) can be written as a varying coefficient model,

$$Y_i = X_i^\top \beta(U_i) + \varepsilon_i. \tag{2}$$

Denote the true value of $\beta_j(\cdot)$ as $\beta_j^*(\cdot)$. Then, our assumptions on $\alpha(\cdot)$ and θ give us that for $j > q$, $\beta_j^*(\cdot) = 0$ and for $k > r$, $\beta_k^{*'}(\cdot) = 0$. Let $B_1 = \{1, \ldots, q\}, B_2 = \{1, \ldots, r\}$. Then, detecting the model structure of (1) is equivalent to identifying these two sets.

2.1. Uniform adaptive LASSO

Since model (2) is simply a varying coefficient model, we adopt [9]'s estimation procedure. For any index value $u \in [0, 1]$, we employ the local linear smoothing technique, i.e. $\beta(U_i) \approx \beta(u) + (U_i - u)\beta'(u)$ [5], to estimate $\beta(u)$ and $\beta'(u)$ by

minimizing the locally weighted least squares

$$L_n\left(\beta(u), \beta'(u)\right) = \sum_{i=1}^{n} \left\{Y_i - X_i^\top \beta(u) - X_i^\top \beta'(u)\,(U_i - u)\right\}^2 K_h(U_i - u) \tag{3}$$

with respect to $\beta(u)$ and $\beta'(u)$, where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth. We denote the minimizer of (3) by $(\tilde{\beta}(u), \tilde{\beta}'(u))$ and let $U_u = \text{diag}\,(U_1 - u, \ldots, U_n - u)$, $X = (X_1, \ldots, X_n)^\top$, $D_u = (X, U_u X/h)$, $W_u = \text{diag}\,(K_h(U_1 - u), \ldots, K_h(U_n - u))$, and $Y = (Y_1, \ldots, Y_n)^\top$. Then it can be shown that $(\tilde{\beta}^\top(u), h\tilde{\beta}'^\top(u))^\top = (D_u^\top W_u D_u)^{-1} D_u^\top W_u Y$.

Now, we define the adaptive LASSO estimator $(\hat{\beta}^{(n)}_{\lambda,\gamma}(u), \hat{\beta}'^{(n)}_{\lambda,\gamma}(u))$ for each u in $[0, 1]$ as the minimizer of the convex function

$$Q_n\left(\beta(u), \beta'(u)\right) = L_n\left(\beta(u), \beta'(u)\right) + \lambda_n \sum_{j=1}^{p} \frac{|\beta_j(u)|}{w_j} + \gamma_n \sum_{j=1}^{p} \frac{|\beta_j'(u)|}{v_j} \tag{4}$$

where $\lambda_n \geq 0$, $\gamma_n \geq 0$ are the tuning parameters, w_j's and v_j's are all user determined positive random quantities free of u (depending on n) satisfying the properties: (1) for $j \in B_1$ and $k \in B_2$, as $n \to \infty$, w_j and v_k converge in probability to two positive numbers; (2) for $j \notin B_1$, there exists a sequence α_{1n} such that $\alpha_{1n}/\sqrt{\log n} \to \infty$, $\alpha_{1n} w_j = O_p(1)$ as $n \to \infty$; (3) for $k \notin B_2$, there exists a sequence α_{2n} such that $\alpha_{2n}/\sqrt{\log n} \to \infty$, $\alpha_{2n} v_k = O_p(1)$, as $n \to \infty$. Note that the conditions on the weights w_j's and v_j's accommodate different degrees of smoothness of $\beta_j(\cdot)$'s and their derivatives. In Section 2.3 we provide a suitable set of w_j's and v_j's based on $(\tilde{\beta}(U_t), \tilde{\beta}'(U_t))$, $t = 1, \ldots, n$. Under this adaptive penalty setting, since the adaptive weights of the above L_1 penalty are set to be free of u, we can show that this procedure can consistently identify the uniform sparsity in $\beta^*(\cdot)$ and $\beta^{*'}(\cdot)$ (see Theorem 2 below). We refer to this setting as uniform adaptive LASSO (ULASSO). Moreover, for the purpose of selecting nonzero coefficients only, we can simply set $\gamma_n = 0$ and in identifying only the parametric components we can set $\lambda_n = 0$ and minimize the corresponding Q. For notational convenience we write $\hat{\beta}^{(n)}_{\lambda,\gamma}(u) = \hat{\beta}^{(n)}_{\lambda}(u)$ and $\hat{\beta}'^{(n)}_{\lambda,\gamma}(u) = \hat{\beta}'^{(n)}_{\gamma}(u)$ to emphasize that λ_n is the shrinkage parameter for the coefficients and γ_n is the shrinkage parameter for the derivatives of coefficients. Similar notations are used in Section 2.3 for $\hat{B}_{1\lambda}$ and $\hat{B}_{2\gamma}$.

2.2. Technical assumptions and theoretical properties

In this section, we first present a few technical assumptions followed by the uniform consistency, the uniform sparsity and the pointwise asymptotic normality of the ULASSO estimator.

C1. The density function f of U is positively bounded away from 0 on $[0, 1]$ and has bounded second order derivative.
C2. The second order derivatives of $\{\beta_j^*(\cdot), j = 1, \ldots, p\}$ are bounded.
C3. The kernel function $K(\cdot)$ is a symmetric density function with a compact support.
C4. The $p \times p$ matrix $\Gamma(u) = E(X_i X_i^T | U_i = u)$ is non-singular for each $u \in [0, 1]$; and its elements have bounded second order derivatives. The function $E(\|X_i\|^4 | U_i = u)$ is bounded.
C5. The function $\sigma^2(u) = E(\varepsilon_i^2 | U_i = u)$ has bounded second order derivative.
C6. There is an $s > 2$ such that $E|X_{ij}|^{2s} < \infty$ and $E|Y_i|^{2s} < \infty$.

Remark 1. Following [16], (C1) assures that the distance between two consecutive index variables is at most of order $O_p(\log n/n)$. For an arbitrary $u \in [0, 1]$, let $\tilde{u} = \text{argmin}_{\{U_t : 1 \leq t \leq n\}} |U_t - u|$. Combining with (C2), it is seen that both $|\beta^*(\tilde{u}) - \beta^*(u)|$ and $|\beta^{*'}(\tilde{u}) - \beta^{*'}(u)|$ are of the same order as $|\tilde{u} - u|$, which is $O_p(\log n/n)$. Since the estimation rates of the coefficient function and its derivative are $n^{-2/5}$ and $n^{-1/5}$, respectively, both of which converge to zero substantially slower than $\log n/n$, it suffices to approximate the entire coefficient curve $\beta^*(\cdot)$ and the entire derivative cure $\beta^{*'}(\cdot)$ by $\{\beta^*(U_t) : 1 \leq t \leq n\}$ and $\{\beta^{*'}(U_t) : 1 \leq t \leq n\}$. This allows us to focus only on the index observations instead of the whole index interval $[0, 1]$.

Now we state two asymptotic results regarding the ULASSO estimator. We begin with a uniform consistency result for the estimated coefficient functions and their derivatives.

Theorem 1. *Let $h \propto n^{-1/5}$. Suppose conditions* (C1)–(C6) *hold. When $h\lambda_n/\sqrt{nh \log n} \to 0$, and $\gamma_n/\sqrt{nh \log n} \to 0$ as $n \to \infty$, the ULASSO estimator satisfies*

$$\sup_u \left\| \hat{\beta}^{(n)}_{\lambda}(u) - \beta^*(u) \right\| = O_p(c_n) \quad \text{and} \quad \sup_u \left\| h\hat{\beta}'^{(n)}_{\gamma}(u) - h\beta^{*'}(u) \right\| = O_p(c_n),$$

where $c_n = \sqrt{\log(1/h)/(nh)}$ and $\|\cdot\|$ is the Euclidean norm.

This theorem shows that the ULASSO estimator is uniformly consistent and the next theorem shows the ULASSO estimator has oracle properties [7] for suitably chosen λ_n and γ_n.

Theorem 2 (*Oracle Properties*). *Let $h \propto n^{-1/5}$. Suppose conditions* (C1)–(C6) *hold. When $h\lambda_n/\sqrt{nh} \to 0$, $h\lambda_n\alpha_{1n}/\sqrt{nh\log n} \to \infty$, $\gamma_n/\sqrt{nh} \to 0$, and $\gamma_n\alpha_{2n}/\sqrt{nh\log n} \to \infty$ as $n \to \infty$, we have*

1. $P\left(\sup_u \left|\hat{\beta}^{(n)}_{\lambda,B_1^c}(u)\right| = 0, \sup_u \left|\hat{\beta}^{'(n)}_{\gamma,B_2^c}(u)\right| = 0\right) \to 1$;
2. $\sqrt{nh}\left\{\hat{\beta}^{(n)}_{\lambda,B_1}(u) - \beta^*_{B_1}(u) - \frac{1}{2}h^2\mu_2\beta^{*''}_{B_1}(u)\right\} \to^d N\left(0, \nu_0\sigma^2(u)\left[f(u)\Gamma(u)\right]^{-1}_{11}\right)$, *for any* $u \in (0, 1)$;
3. $\sqrt{nh^3}\left\{\hat{\beta}^{'(n)}_{\gamma,B_2}(u) - \beta^{*'}_{B_2}(u)\right\} \to^d N\left(0, \nu_2\sigma^2(u)\left[\mu_2^2 f(u)\Gamma(u)\right]^{-1}_{22}\right)$, *for any* $u \in (0, 1)$,

where $\nu_i = \int u^i K^2(u)du$, $\mu_i = \int u^i K(u)du$. Here d_B represents the subvector of the elements of a vector d according to set B and $[A]_{11}$ represents the first $q \times q$ submatrix and $[A]_{22}$ represents the first $r \times r$ submatrix of a $p \times p$ matrix A.

2.3. Implementation and tuning parameter selection

In applications it is impossible to perform the ULASSO method for all $u \in [0, 1]$. Following Remark 1, it is sufficient for us to focus only on the index sample $\{U_t, t = 1, \ldots, n\}$. Hence we minimize the convex function

$$\frac{1}{n}\sum_{t=1}^{n} L_n(\beta(U_t), \beta'(U_t)) + \lambda_n \sum_{j=1}^{p} \frac{1}{w_j}\left\{\frac{\sum_{t=1}^{n}|\beta_j(U_t)|}{n}\right\} + \gamma_n \sum_{j=1}^{p}\frac{1}{v_j}\left\{\frac{\sum_{t=1}^{n}|\beta'_j(U_t)|}{n}\right\} \tag{5}$$

with respect to $\beta(U_t)$ and $\beta'(U_t)$ for $t = 1, \ldots, n$ and let $(\hat{\beta}^{(n)}_\lambda(U_t), \hat{\beta}^{'(n)}_\gamma(U_t))$, $t = 1, \ldots, n$ be the minimizer. Then $\hat{\beta}^{(n)}_\lambda(\tilde{u})$ is the proposed shrinkage estimator of $\beta^*(u)$. Now, denote $\hat{B}_{1\lambda} = \{j : \sum_{t=1}^{n}|\hat{\beta}^{(n)}_{\lambda,j}(U_t)| \neq 0\}$ where $\hat{\beta}^{(n)}_{\lambda,j}(U_t)$ is the jth element of $\hat{\beta}^{(n)}_\lambda(U_t)$ and $\hat{B}_{2\gamma} = \{j : \sum_{t=1}^{n}|\hat{\beta}^{'(n)}_{\gamma,j}(U_t)| \neq 0\}$ where $\hat{\beta}^{'(n)}_{\gamma,j}(U_t)$ is the jth element of $\hat{\beta}^{'(n)}_\gamma(U_t)$. The two sets $\hat{B}_{1\lambda}$ and $\hat{B}_{2\gamma}$ are taken as the estimators of B_1 and B_2 respectively. Since the derivative of an insignificant coefficient is zero, we suggest using $\hat{B}_{2\gamma} \cap \hat{B}_{1\lambda}$ for estimating B_2 in finite sample applications to improve accuracy. The following theorem shows that these two sets can consistently identify B_1 and B_2 respectively and that $\hat{\beta}^{(n)}_\lambda$ can consistently estimate β^*.

Theorem 3. *Let $h \propto n^{-1/5}$. Suppose conditions (C1)–(C6) hold. When $h\lambda_n/\sqrt{nh} \to 0$, and $\gamma_n/\sqrt{nh} \to 0$, as $n \to \infty$, we have*

$$\frac{1}{n}\sum_{t=1}^{n}\left\|\hat{\beta}^{(n)}_\lambda(U_t) - \beta^*(U_t)\right\|^2 = O_p(n^{-4/5});$$

$$\frac{1}{n}\sum_{t=1}^{n}\left\|\hat{\beta}^{'(n)}_\gamma(U_t) - \beta^{*'}(U_t)\right\|^2 = O_p(n^{-2/5}).$$

If in addition $h\lambda_n\alpha_{1n}/\sqrt{nh\log n} \to \infty$, and $\gamma_n\alpha_{2n}/\sqrt{nh\log n} \to \infty$ are satisfied as $n \to \infty$, then

$$P(\hat{B}_{1\lambda} = B_1, \hat{B}_{2\gamma} = B_2) \to 1.$$

In the formulation in (5), for each j, our method groups $\beta_j(U_t)$'s, $t = 1, \ldots, n$, by the L_1 norm, whereas KLASSO groups them by L_2 norm $(\sum_{t=1}^{n}\beta_j^2(U_t)/n)^{1/2}$ [21]. A major advantage in our approach is the computational burden for the group LASSO would increase dramatically as the number of predictors (which is np in this case) increases [22]. In addition, the LQA algorithm used in a L_2 problem of this type essentially produces a ridge estimator that is forced to become sparse depending on an arbitrary criteria. The L_1 approach avoids both these issues.

Remark 2. The ULASSO estimator achieves the optimal estimation rate for all the nonparametric coefficient functions. Nevertheless, when $B_1 \neq B_2$, there exist $q - r$ parametric components. For $j \in B_1 \cap B_2^c$, $\beta_j^*(\cdot) = \beta_j^*$, a nonzero constant. Its optimal estimation rate should be $\sqrt{n}$, which cannot be achieved by this estimation procedure. However, the proposed method above can identify the structure of the set $B_1 \cap B_2^c$. Then the inefficiencies can be remedied by choosing an appropriate estimation procedure for a resulting VCPLM [13,23,6].

For suitable w_j's and v_j's and $h \propto n^{-1/5}$, we have to choose tuning parameters, λ_n and γ_n, for the coefficient function part and the derivative part, respectively. In what follows, we set $\alpha_{1n} = n^{2/5}$, $\alpha_{2n} = n^{1/5}$, $w_j = (\sum_{i=1}^{n}\tilde{\beta}_j^2(U_i)/n)^{1/2}$ and $v_j = (\sum_{i=1}^{n}\tilde{\beta}_j^{'2}(U_i)/n)^{1/2}$. It is easy to check that the requirements for w_j's and v_j's are satisfied. For these settings, we find

the suitable rates for these two tuning parameters as $\lambda_n \cdot n^{-3/5} \to 0$, $\gamma_n \cdot n^{-2/5} \to 0$, and $\min(\lambda_n, \gamma_n) \cdot n^{-1/5}/\sqrt{\log n} \to \infty$. We observe that these conditions in fact allow $\lambda_n = \gamma_n$.

In practice it is preferred to select the tuning parameters λ_n and γ_n based on data. To this end, we propose a two dimensional BIC selector [21]. We define

$$\mathrm{BIC}(\lambda, \gamma) = \log\{\mathrm{RSS}(\lambda, \gamma)\} + \mathrm{DF}(\lambda, \gamma) \times \frac{\log(nh)}{nh}, \tag{6}$$

where $\mathrm{RSS}(\lambda, \gamma)$ is defined as

$$\mathrm{RSS}(\lambda, \gamma) = n^{-2} \sum_{t=1}^{n} \sum_{i=1}^{n} \left\{ Y_i - X_i^\top \hat{\beta}_\lambda^{(n)}(U_t) - X_i^\top \hat{\beta}_\gamma'^{(n)}(U_t)(U_i - U_t) \right\}^2 K_h(U_i - U_t),$$

and $0 \le \mathrm{DF}(\lambda, \gamma) \le 2p$ is the sum of the numbers of elements in $\hat{B}_{1\lambda}$ and $\hat{B}_{2\gamma}$ both of which depend on each pair of (λ, γ). Define $(\hat{\lambda}, \hat{\gamma})$ to be the minimizer of (6) over a suitable two dimensional interval. Now, let $\hat{B}_{1\hat{\lambda}}, \hat{B}_{2\hat{\gamma}}$ be the estimators of B_1 and B_2, respectively, identified via the resulting ULASSO estimator denoted by $(\hat{\beta}_{\hat{\lambda}}^{(n)}(U_t), \hat{\beta}_{\hat{\gamma}}'^{(n)}(U_t))$, $t = 1, \ldots, n$. The next theorem assures that the tuning parameters selected by this BIC continue to identify the true model and the true set of parametric components consistently.

Theorem 4 (*Selection Consistency*). *Let $h \propto n^{-1/5}$. Suppose conditions (C1)–(C6) hold. As $n \to \infty$, we have*

$$P(\hat{B}_{1\hat{\lambda}} = B_1, \hat{B}_{2\hat{\gamma}} = B_2) \to 1.$$

One may find the minimizer $(\hat{\lambda}, \hat{\gamma})$ by a suitable minimization procedure such as an optimal gradient search. However, in our simulations we chose $\lambda_n = \gamma_n$ in light of the comments following Remark 2.

3. Numerical experiments

In this section we provide simulation results followed by the results of a real data analysis.

3.1. Simulation results

To evaluate the finite sample performance of the proposed ULASSO method, we conducted a simulation study similar to [21]. We carried out 1000 simulations in each case. We examined 18 different simulation settings created using three sample sizes $n = 200$, 300 and 400, two distributions for the index variable U coupled with three models:

Model 1: $\beta(u) = (4u, 2\sin(2\pi u), 1, 0, 0, 0, 0)^\top$;
Model 2: $\beta(u) = (\exp(2u), 2\cos(2\pi u), 2\sin^2(2\pi u), 0, 0, 0, 0)^\top$;
Model 3: $\beta(u) = (8u(1-u), 0.8, 1, 1.2, 0, 0, 0)^\top$.

The distributions of U_k were chosen to be either *Unif*$[0, 1]$, or *Beta*$(4, 1)$, a highly asymmetric distribution. In each model, $X_k = (x_{k1}, \ldots, x_{k7})^\top$ where $x_{k1} = 1$ and $(x_{k2}, \ldots, x_{k7})^\top$ were generated from a multivariate normal distribution with mean vector 0 and $\mathrm{cov}(x_{ki}, x_{kj}) = 2^{-|i-j|}$ for $2 \le i, j \le 7$. ε_k is simulated from $N(0, \sigma_e^2)$, where $\sigma_e = 1.5$. As one can see, Model 1 is a varying coefficient partially linear model with $B_1 = \{1, 2, 3\}$ and $B_2 = \{1, 2\}$; Model 2 is a varying coefficient model with $B_1 = B_2 = \{1, 2, 3\}$; Model 3 is a partially linear model with $B_1 = \{1, 2, 3, 4\}$ and $B_2 = \{1\}$. All simulations were conducted using the package R.

In each simulation, we first use the leave-one-out cross-validation to select the optimal bandwidth by fitting an unpenalized estimator, $(\tilde{\beta}(U_k), \tilde{\beta}'(U_k))$, with the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$. The same bandwidth is used for the BIC selector (6).

To compare the performance of ULASSO and KLASSO, we perform both methods in each simulation. First, we classify the result of variable selection (i.e. nonzero coefficient selection): 1. underfitted (at least one true nonzero variable is missing); 2. correctly fitted; 3. overfitted (all the significant variables are identified while at least one spurious variable is included). The percentages of experiments in each category by ULASSO and KLASSO are presented in Table 1 under VS. Similarly, we can partition the results for detecting the true coefficient functions and parametric components into these categories. Here a correct selection means that the procedure identified both types correctly, an underfitting indicating missing at least one true function or a coefficient etc. These percentages are given under the heading VS & PCD in Table 1. To evaluate the estimation accuracy, we consider the following relative estimation error (REE),

$$\mathrm{REE} = \frac{\sum_{k=1}^{n} \sum_{j=1}^{p} \left| \hat{\beta}_{\lambda,j}^{(n)}(U_k) - \beta_j^*(U_k) \right|}{\sum_{k=1}^{n} \sum_{j=1}^{p} \left| \bar{\beta}_j^{(n)}(U_k) - \beta_j^*(U_k) \right|},$$

Table 1
VS: variable selection; PCD: parametric components detection; *O*: overfitted; *C*: correctly fitted; *U*: underfitted.

$f(u)$	n	VS by ULASSO			VS by KLASSO			VS & PCD			MREE
		O	*C*	*U*	*O*	*C*	*U*	*O*	*C*	*U*	
Model 1											
Unif[0, 1]	200	0.02	0.97	0.01	0.00	0.99	0.01	0.03	0.94	0.03	1.023
	300	0.01	0.98	0.01	0.00	1.00	0.00	0.03	0.96	0.01	0.989
	400	0.00	1.00	0.00	0.00	1.00	0.00	0.01	0.99	0.00	0.936
Beta[4, 1]	200	0.07	0.91	0.02	0.01	0.95	0.04	0.08	0.89	0.03	0.729
	300	0.04	0.96	0.01	0.01	0.98	0.01	0.04	0.94	0.02	0.699
	400	0.00	0.99	0.00	0.00	0.99	0.00	0.01	0.98	0.01	0.624
Model 2											
Unif[0, 1]	200	0.05	0.94	0.01	0.01	0.98	0.01	0.05	0.91	0.04	0.999
	300	0.04	0.96	0.00	0.01	0.99	0.00	0.04	0.94	0.02	0.978
	400	0.02	0.98	0.00	0.01	0.99	0.00	0.02	0.98	0.00	0.923
Beta[4, 1]	200	0.09	0.89	0.02	0.03	0.93	0.04	0.09	0.88	0.03	0.796
	300	0.07	0.92	0.01	0.01	0.96	0.03	0.07	0.91	0.02	0.691
	400	0.02	0.98	0.00	0.01	0.98	0.01	0.03	0.96	0.01	0.605
Model 3											
Unif[0, 1]	200	0.03	0.97	0.00	0.02	0.98	0.00	0.03	0.96	0.01	1.011
	300	0.01	0.99	0.00	0.01	0.99	0.00	0.01	0.99	0.00	1.002
	400	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.986
Beta[4, 1]	200	0.08	0.92	0.00	0.06	0.93	0.01	0.08	0.88	0.04	0.924
	300	0.05	0.95	0.00	0.04	0.96	0.01	0.06	0.93	0.01	0.857
	400	0.02	0.98	0.00	0.02	0.98	0.00	0.03	0.97	0.00	0.806

where $\bar{\beta}^{(n)}(u)$ is either the unpenalized ULASSO estimator or the KLASSO estimator. The median REE values (denoted as MREE) with respect to the KLASSO estimator are also summarized in Table 1.

As one can see from Table 1, for every model, the percentage of correct selection (when tuning parameters are chosen by the BIC criterion) of the significant variables and parametric components (VS & PCD) is very high and it increases steadily as the sample size increases. If we focus only on the variable selection part, the percentage of correctly fitted models by ULASSO is slightly smaller than that by KLASSO for sample sizes 200 and 300 in few cases. This perhaps results from the slower estimation rate for the derivative possibly affecting the accuracy of the BIC selector. However, the correct model always appeared in ULASSO shrinkage path sets $\hat{B}_{1\lambda}$ and $\hat{B}_{2\lambda}$ for some λ, that λ value was not always picked by the BIC. The MREE with respect to unpenalized estimators (not given here) is much smaller than 1 for all cases. The MREE (with respect to the KLASSO estimator) is slightly smaller than 1 in almost all cases for uniformly distributed index variables. However, the MREE for *Beta*(4, 1) is smaller than 1 and decreasing with n in all examined models. Hence it is reasonable to suggest that the ULASSO with local linear smoothing remedies estimation issues caused by the distribution of the index variable while maintaining very satisfactory selection frequencies for all component types.

For $n = 200$, setting $\lambda_n = \gamma_n$ and using the corresponding BIC for both KLASSO and ULASSO, the computing time for KLASSO is about 9 times that of ULASSO per simulation for the above models. Furthermore, in our simulations we noticed that the optimal λ_n selected via our version of BIC for ULASSO is much smaller than that selected via Wang and Xia's [21] BIC for KLASSO. The selection consistency via KLASSO requires their tuning parameter λ_n to be between the orders $n^{11/10}$ and $n^{7/10}$ and our λ_n and γ_n need to be between orders $n^{4/10}$ and $n^{2/10}\sqrt{\log n}$, perhaps forcing a wider search for KLASSO's BIC than ULASSO's BIC. This can degrade KLASSO's computational efficiency. As our detailed simulations exhibit, taking $\lambda_n = \gamma_n$ in ULASSO has little impact in the selection probabilities and the estimation efficiency. We also noticed that the LQA in KLASSO converges at different speeds depending on the tuning parameter value tested in their BIC.

To show that ULASSO can have much superior performance in some situations compared with KLASSO, we considered the model $\beta(u) = (\beta_1(u), 2, 0, 0, 0, 0, 0)^\top$ where $\beta_1(u)$ is such that $\beta_1(u) = 0$ for $u < 0.9$, $\beta_1(u)$ is a third degree polynomial in [0.9, 1] satisfying $\beta_1(0.9) = \beta_1'(0.9) = \beta_1''(0.9) = 0$ and $\beta_1(1) = 20$. We took $n = 200$ and generated the covariate X and the errors same as above, while U is from *Beta*(4, 1). We fixed the range of λ_n to be $[0, 3n^{11/10}]$ for both BICs. In 200 simulations, the proportion of correctly fitted models via KLASSO was 0.34 while ULASSO's identification rate was 94% with a MREE = 0.680.

3.2. Boston housing data

We now illustrate the ULASSO method by an application to the Boston Housing Data, which has been analyzed by Fan and Huang [6] and Wang and Xia [21] among others. We take MEDV (median value of owner-occupied homes in 1000 United States dollar) as Y, the response, and LSTAT (the percentage of lower status of the population) as U, the index variable. The covariates includes INT (the intercept), CRIM (per capita crime rate by town), NOX (nitric oxides concentration parts per 10 million), RM (average number of rooms per dwelling), AGE (proportion of owner-occupied units built prior to 1940), TAX

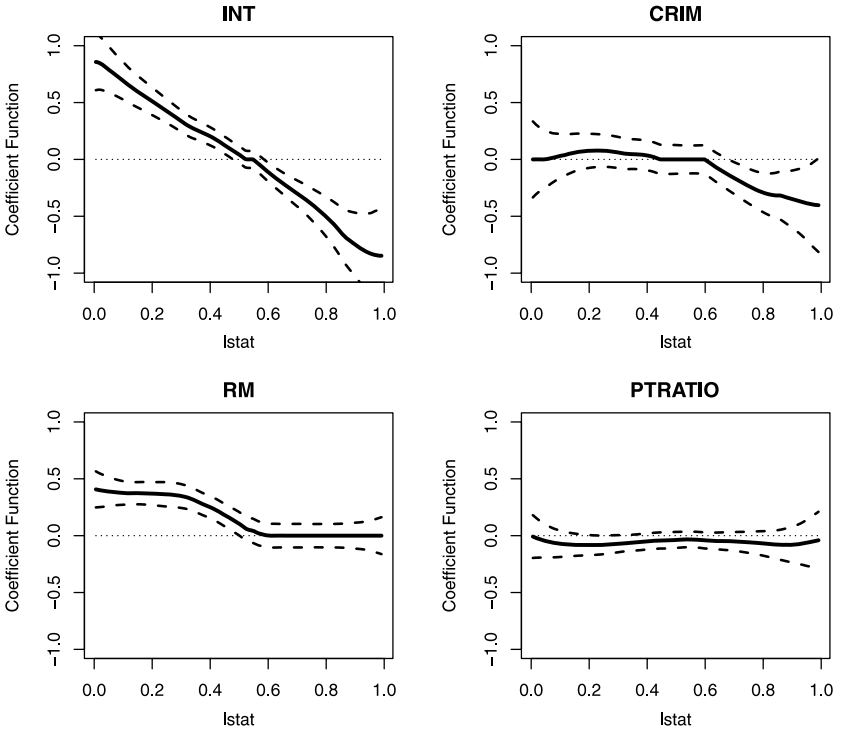

Fig. 1. The ULASSO estimates of the relevant coefficients.

(full-value property-tax rate per 10,000) and PTRATIO (pupil teacher ratio by town). They are denoted by $X_1, X_2, \ldots, X_7$, respectively.

Then a varying coefficient model

$$Y = \sum_{j=1}^{7} X_j^{\top} \beta_j(U) + \varepsilon$$

is fitted by our method. Before applying the proposed shrinkage procedure, we transform the marginal distribution of Y and $X_j, j = 2, \ldots, 7$, to be approximately $N(0, 1)$ via the method of Box–Cox transformation and the marginal distribution of U is transformed to be $U[0, 1]$. The method of leave-one-out cross-validation without penalization suggested an optimal bandwidth $\hat{h} = 0.2809$. By using this bandwidth, the BIC selector picks the optimal tuning parameters as $\hat{\lambda} = 5.074$, $\hat{\gamma} = 5.054$. The resulting ULASSO estimators indicate that X_1 (INT), X_2 (CRIM), X_4 (RM) and X_7 (PTRATIO) are significant. This is the same as the conclusion in [21]. However, the result of parametric component detection further tells that the coefficient functions $\beta_4(\cdot)$ and $\beta_7(\cdot)$ should be constant.

The nonparametric curve estimates (the solid lines) are presented in Fig. 1. To confirm that NOX, AGE and TAX are spurious variables and that RM and PTRATIO are in fact parametric components, following [8] we construct the 90% simultaneous confidence bands (the dashed lines) of the unpenalized estimators of $\beta_3(\cdot)$, $\beta_5(\cdot)$ and $\beta_6(\cdot)$ (Fig. 2) and the 90% simultaneous confidence bands (the dashed lines) of the nonzero ULASSO estimators [21] in Fig. 1. It can be seen that all the simultaneous confidence bands in Fig. 2 almost cover the complete zero line. In addition, in Fig. 1, for PTRATIO, we can draw a constant straight line within its simultaneous confidence bands, suggesting that it is reasonable to consider that $\beta_7(\cdot)$ is a constant. However, for RM, its simultaneous confidence bands indicate that $\beta_4(\cdot)$ is not constant. We believe the following argument explains this discrepancy. Although our theory holds under the assumption that all the true coefficient functions should have bounded second order derivatives everywhere, we estimate these coefficient functions only at the sampled index values. In this real application, the true $\beta_4(\cdot)$ is perhaps not continuous. Yet, the results of parametric component detection tells us that the derivatives of this true coefficient function is zero at all the observed index values suggesting a step function behavior for the coefficient function $\beta_4(\cdot)$.

As observed in Fig. 1, the estimated $\beta_4(u)$ curve by ULASSO is zero when u is greater than some number a. In this application we chose a to be the smallest U_i such that $\hat{\beta}^{(n)}_{\lambda,4}(U_i) = 0$ which equals to 0.6025429. Consequently, we chose to fit the given data to the following model via the profile least-squares estimation method [6]

$$Y = \beta_1(U) + X_2\beta_2^{\top}(U) + X_4^{\top}\beta_4(U) + X_7^{\top}\beta_7 + \varepsilon, \tag{7}$$

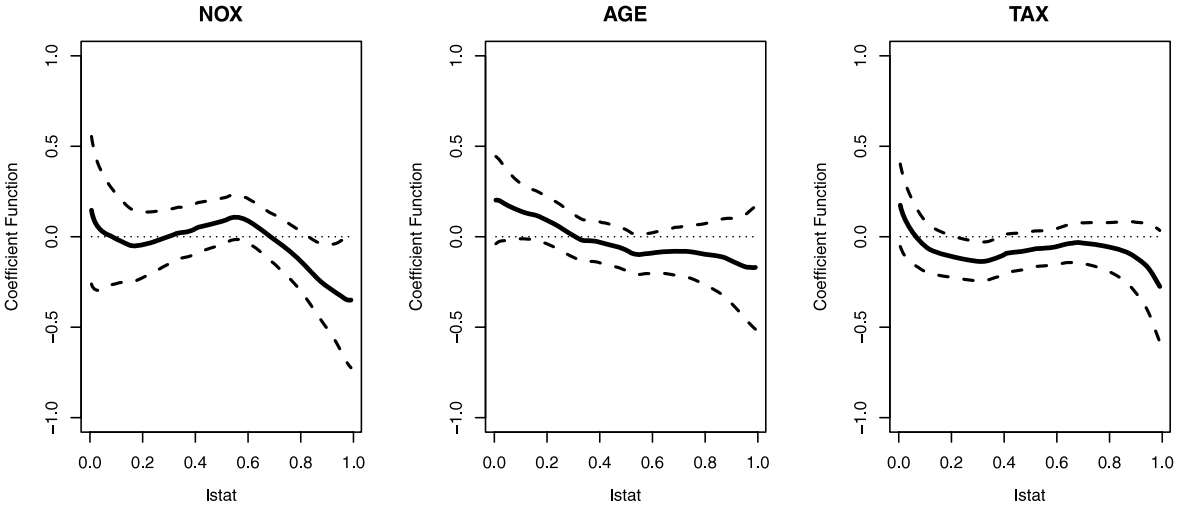

Fig. 2. The unpenalized estimates of the irrelevant coefficients.

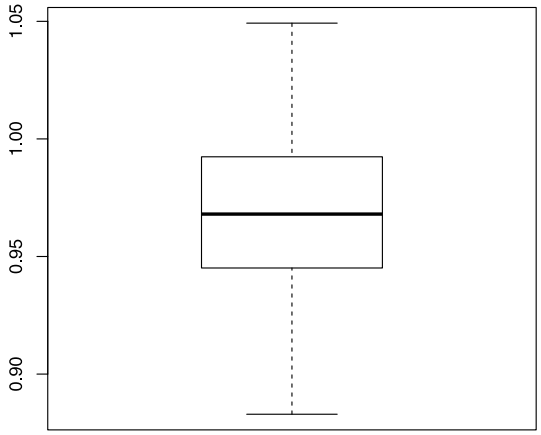

Fig. 3. Box plot of 100 RPE numbers.

where $\beta_4(u) = \beta_{41}$, if $u < a$; $\beta_4(u) = 0$, if $u \geq a$. The resulting estimate of β_{41} is 0.341. To compare with the model suggested by Wang and Xia [21]

$$Y = \beta_1(U) + X_2\beta_2^\top(U) + X_4^\top\beta_4(U) + X_7^\top\beta_7(U) + \varepsilon. \tag{8}$$

We randomly split the data into two equal sized groups and use one group to estimate and the other group to predict. The relative prediction error

$$\text{RPE} = \frac{\sum |Y_i - \hat{Y}_i|}{\sum |Y_i - \bar{Y}_i|}$$

is calculated, where the summation is over the second group and $\bar{Y}_i$ is the predicted value of the ith observation under model (8). We repeated this calculation 100 times. In Fig. 3 a box plot of the resulting 100 RPE numbers is presented. Here 86% of the RPE values are below 1. Thus model (7) appears to fit the Boston Housing data better than model (8) suggesting that for high percentages of lesser status, the room variable may have a diminishing impact on median home values.

4. Conclusion

We have proposed the ULASSO shrinkage method for simultaneously identifying the constant coefficients, selecting variables and estimating the unknown coefficients in the VCPLM. The proposed method has very desirable selection properties and estimation efficiency while being computationally thrifty. It compares well with existing methods which only target the variable selection. Our data analysis suggest that the proposed method is able to reveal hidden features of the coefficient functions such as step behavior. In addition, the ULASSO method has all the desirable asymptotic properties of a shrinkage method.

Appendix

Lemma 1. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed random vectors, where the Y_i's are scalar random variables. Further assume that $E|y|^s < \infty$ and $\sup_x \int |y|^s f(x, y)dy < \infty$, where f denotes the joint density of (X, Y). Let K be a bounded positive function with a bounded support, satisfying a Lipschitz condition. Given that $n^{2\varepsilon-1}h \to \infty$ for some $\varepsilon < 1-s^{-1}$, we have*

$$\sup_x \left| \frac{1}{n} \sum_{i=1}^{n} [K_h(X_i - x)Y_i - E\{K_h(X_i - x)Y_i\}] \right| = O_p\left(c_n\right).$$

Proof. For the proof we refer the reader to [19]. □

In what follows, we let $\nu(u) = (\beta^\top(u), h\beta'^\top(u))^\top$ and rewrite $L_n(\beta(u), \beta'(u))$ as

$$L_n(\nu(u)) = (Y - D_u\nu(u))^\top W_u(Y - D_u\nu(u));$$

and $Q_n(\beta(u), \beta'(u))$ as

$$Q_n(\nu(u)) = L_n(\nu(u)) + \lambda_n \sum_{j=1}^{p} \frac{|\nu_j(u)|}{w_j} + \gamma_n \sum_{j=1}^{p} \frac{|\nu_{p+j}(u)|}{hv_j}.$$

Similarly, denote $\hat{\nu}_\lambda^{(n)}(u) = (\hat{\beta}_\lambda^{(n)\top}(u), h\hat{\beta}_\gamma'^{(n)\top}(u))^\top$ which is the minimizer of $Q_n(\nu(u))$, $\nu^*(u) = (\beta^{*\top}(u), h\beta^{*'\top}(u))^\top$, $\tilde{\nu}(u) = (\tilde{\beta}^\top(u), h\tilde{\beta}'^\top(u))^\top$ and define $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ as the minimum and maximum eigenvalue of a matrix A, respectively. Further, we denote $\lambda^{\min} = \min(1, \mu_2) \cdot \inf_u \lambda_{\min}(\Gamma(u)f(u))$, $\lambda^{\max} = \max(1, \mu_2) \cdot \sup_u \lambda_{\max}(\Gamma(u)f(u))$.

Proof of Theorem 1. We only need to show that

$$\sup_u \|\hat{\nu}_\lambda^{(n)}(u) - \nu^*(u)\| = O_p(c_n).$$

Following [7], consider the ball $B_C = \{\nu(u) : \nu(u) = \nu^*(u) + c_n r, \|r\| \le C\}$, $C > 0$. Note that for each u, $Q_n(\nu(u))$ is strictly convex. It is sufficient to show that, for any given $\delta > 0$, there exists a large constant C (does not depend on u), such that

$$P\left\{ \inf_{\|r\|=C} Q_n(\nu^*(u) + c_n r) > Q_n(\nu^*(u)) \text{ for every } u \right\} \ge 1 - \delta.$$

This implies that, with probability $1 - \delta$, for each u, there exists a minimum in the ball B_C. Hence, for each u, the minimizer $\hat{\nu}_\lambda^{(n)}(u)$ must satisfy that $\|\hat{\nu}_\lambda^{(n)}(u) - \nu^*(u)\| = O_p(c_n)$. Furthermore, since C does not depend on u, we must have $\sup_u \|\hat{\nu}_\lambda^{(n)}(u) - \nu^*(u)\| = O_p(c_n)$.

Now, for any u define

$$R_1 = \frac{h}{\log 1/h} \left\{ Q_n\left(\nu^*(u) + \sqrt{\frac{\log 1/h}{nh}} r\right) - Q_n(\nu^*(u)) \right\}.$$

When $j \notin B_1$ and $k \notin B_2$, $\nu_j^*(u) = 0$ and $\nu_{p+k}^*(u) = 0$. Some simplifications show that

$$\begin{aligned} R_1 \ge\ & \frac{r^\top D_u^\top W_u D_u r}{n} - 2\frac{r^\top}{\sqrt{\log 1/h}} \cdot \sqrt{\frac{h}{n}} D_u^\top W_u(Y - D_u\nu^*(u)) \\ & + \sum_{j\in B_1} \frac{h\lambda_n}{(\log 1/h)w_j} \left(\left| \nu_j^*(u) + \sqrt{\frac{\log 1/h}{nh}} \cdot r_j \right| - |\nu_j^*(u)| \right) \\ & + \sum_{j\in B_2} \frac{\gamma_n}{(\log 1/h)v_j} \left(\left| \nu_{p+j}^*(u) + \sqrt{\frac{\log 1/h}{nh}} \cdot r_{p+j} \right| - |\nu_{p+j}^*(u)| \right) \\ \doteq\ & R_2. \end{aligned}$$

Let $\lambda_n^{\min} = \inf_u \lambda_{\min}(\frac{D_u^\top W_u D_u}{n})$. Then

$$\begin{aligned} R_2 \ge\ & \|r\|^2 \lambda_n^{\min} - 2\|r\| \cdot \frac{1}{\sqrt{\log 1/h}} \sup_u \left\| \sqrt{\frac{h}{n}} X^\top W_u(Y - D_u\nu^*(u)) \right\| \\ & - \frac{h\lambda_n}{\sqrt{nh\log 1/h}} \cdot \frac{\sqrt{2p}}{\min_{j\in B_1} w_j} \|r\| - \frac{\gamma_n}{\sqrt{nh\log 1/h}} \cdot \frac{\sqrt{2p}}{\min_{j\in B_2} v_j} \|r\| \\ \doteq\ & R_3. \end{aligned}$$

When $\|r\| = C$,

$$\begin{aligned} R_3 &= \lambda_n^{\min} \times C^2 - 2C \times \frac{1}{\sqrt{\log 1/h}} \sup_u \left\| \sqrt{\frac{h}{n}} D_u^\top W_u (Y - D_u \nu^*(u)) \right\| \\ &\quad - C \times \left\{ \frac{h\lambda_n}{\sqrt{nh \log 1/h}} \cdot \frac{\sqrt{p}}{\min_{j \in B} w_j} + \frac{\gamma_n}{\sqrt{nh \log 1/h}} \cdot \frac{\sqrt{p}}{\min_{j \in B_2} v_j} \|r\| \right\} \\ &\doteq H_n(C). \end{aligned}$$

Clearly, $H_n(C)$ does not depend on u and,

$$H_n(C) > 0 \Rightarrow \forall u, \qquad \inf_{\|r\|=C} Q_n(\nu^*(u) + c_n r) > Q_n(\nu^*(u)).$$

Then it suffices to show that $\forall \delta > 0$, there exists a large constant C, such that

$$P(H_n(C) > 0) \geq 1 - \delta. \tag{9}$$

Since $\lambda^{\min} > 0$ by (C4) and,

$$\frac{D_u^\top W_u D_u}{n} \to^p \begin{pmatrix} \Gamma(u) f(u) & 0 \\ 0 & \mu_2 \Gamma(u) f(u) \end{pmatrix}$$

uniformly in u, we have

$$\lambda_n^{\min} \to^p \lambda^{\min}. \tag{10}$$

Since

$$\sqrt{\frac{h}{n}} D_u^\top W_u (Y - D_u \nu^*(u)) = \sqrt{\frac{h}{n}} D_u^\top W_u \varepsilon + \sqrt{\frac{h}{n}} D_u^\top W_u \begin{pmatrix} X_1^\top (\beta^*(U_1) - \beta^*(u) - \beta^{*'}(u)(U_1 - u)) \\ \vdots \\ X_n^\top (\beta^*(U_n) - \beta^*(u) - \beta^{*'}(u)(U_n - u)) \end{pmatrix},$$

it is sufficient to show each term in the right side is uniformly $O_p(\sqrt{\log 1/h})$ by applying Lemma 1 to each term. Then, we have

$$\sup_u \left\| \sqrt{\frac{h}{n}} D_u^\top W_u (Y - D_u \nu^*(u)) \right\| = O_p(\sqrt{\log 1/h}). \tag{11}$$

Furthermore, by the definition of w_j and v_j, $\min_{j \in B_1} w_j$ and $\min_{j \in B_2} v_j$ converge in probability to two positive constants. Since $h\lambda_n / \sqrt{nh \log n} \to 0$ and $\gamma_n / \sqrt{nh \log n} \to 0$, it provides us that

$$\frac{h\lambda_n}{\sqrt{nh \log 1/h}} \cdot \frac{\sqrt{p}}{\min_{j \in B_1} w_j} = o_p(1), \quad \text{and} \quad \frac{\gamma_n}{\sqrt{nh \log 1/h}} \cdot \frac{\sqrt{p}}{\min_{j \in B_2} v_j} = o_p(1). \tag{12}$$

Combining (10)–(12), if we choose a sufficiently large C, the second and third term of $H_n(C)$ are dominated by its first term. This proves (9) and completing the proof. □

Proof of Theorem 2. We first prove the pointwise asymptotic normality part. Let $\hat{r}_u^{(n)} = \sqrt{nh}(\hat{\nu}_\lambda^{(n)}(u) - \nu^*(u))$. Then

$$\begin{aligned} \hat{r}_u^{(n)} &= \operatorname*{argmin}_{r_u} h \left\{ Q_n \left(\nu^*(u) + \frac{r_u}{\sqrt{nh}} \right) - Q_n(\nu^*(u)) \right\} \\ &\doteq \operatorname*{argmin}_{r_u} V_n(r_u) \end{aligned}$$

where

$$\begin{aligned} V_n(r_u) &= \frac{r_u^\top D_u^\top W_u D_u r_u}{n} - 2 \frac{\sqrt{h}}{\sqrt{n}} r_u^\top D_u^\top W_u (Y - D_u \nu^*(u)) + \sum_{j=1}^p \frac{h\lambda_n}{\sqrt{nh} w_j} \sqrt{nh} \left(\left| \nu_j^*(u) + \frac{r_{uj}}{\sqrt{nh}} \right| - |\nu_j^*(u)| \right) \\ &\quad + \sum_{j=1}^p \frac{\gamma_n}{\sqrt{nh} v_j} \sqrt{nh} \left(\left| \nu_{p+j}^*(u) + \frac{r_{u,p+j}}{\sqrt{nh}} \right| - |\nu_{p+j}^*(u)| \right). \end{aligned}$$

Denote I_{1n}, I_{3n} and I_{4n} as the first, third and fourth terms of the right side of the last equation respectively, and $I_{2n} = h^{1/2}n^{-1/2}D_u^\top W_u(Y - D_u v^*(u))$. For I_{1n}, we know

$$I_{1n} \to^p r_u^\top \begin{pmatrix} \Gamma(u)f(u) & 0 \\ 0 & \mu_2\Gamma(u)f(u) \end{pmatrix} r_u.$$

Since the true coefficients have bounded second order derivatives, we can write

$$\begin{aligned} I_{2n} &= \frac{\sqrt{h}}{\sqrt{n}} D_u^\top W_u \begin{pmatrix} X_1^\top \dfrac{\beta^{*''}(u)}{2}(U_1-u)^2 \\ \vdots \\ X_n^\top \dfrac{\beta^{*''}(u)}{2}(U_n-u)^2 \end{pmatrix} (1+o_p(1)) + \frac{\sqrt{h}}{\sqrt{n}} D_u^\top W_u \varepsilon \\ &\doteq I_{21n} + I_{22n}. \end{aligned}$$

Applying Lemma 1 we have

$$I_{21n} \to^p \begin{pmatrix} \dfrac{d}{2}\mu_2\Gamma(u)\beta^{*''}(u)f(u) \\ 0 \end{pmatrix},$$

where d is the limit of $\sqrt{nh^5}$. For I_{22n}, the expectation of I_{22n} is zero, and its covariance matrix is

$$\begin{aligned} \mathrm{Cov}(I_{22n}) &\to^p \begin{pmatrix} \nu_0\Gamma(u)\sigma^2(u)f(u) & 0 \\ 0 & \nu_2\Gamma(u)\sigma^2(u)f(u) \end{pmatrix} \\ &\doteq \Sigma(u). \end{aligned}$$

Thus, for each u, $I_{22n} \to^d W \sim N(0, \Sigma(u))$. Now,

$$I_{3n} = \sum_{j=1}^{p} \frac{h\lambda_n}{\sqrt{nh}w_j}\sqrt{nh}\left(\left|\beta_j^*(u) + \frac{r_{uj}}{\sqrt{nh}}\right| - |\beta_j^*(u)|\right).$$

If $j \in B_1$, we know that $w_j = O_p(1)$ and $h\lambda_n/\sqrt{nh} \to 0$. At the same time, $\sqrt{nh}(|\beta_j^*(u) + r_{uj}/\sqrt{nh}| - |\beta_j^*(u)|) \to r_{uj}\,\mathrm{sign}(\beta_j^*(u))$. Hence

$$\frac{h\lambda_n}{\sqrt{nh}w_j}\sqrt{nh}\left(\left|\beta_j^*(u) + \frac{r_{uj}}{\sqrt{nh}}\right| - |\beta_j^*(u)|\right) \to^d 0.$$

If $j \notin B_1$ which means $\beta_j^*(u) = 0$, then $\sqrt{nh}(|\beta_j^*(u) + r_{uj}/\sqrt{nh}| - |\beta_j^*(u)|) = |r_{uj}|$. Since $\alpha_{1n}w_j = O_p(1)$ and $h\lambda_n\alpha_{1n}/\sqrt{nh\log n} \to \infty$, we have

$$\frac{h\lambda_n}{\sqrt{nh}\cdot\alpha_{1n}w_j}\cdot\frac{\alpha_{1n}}{\sqrt{\log n}}\cdot\sqrt{\log n}\cdot\sqrt{nh}\left(\left|\beta_j^*(u) + \frac{r_{uj}}{\sqrt{nh}}\right| - |\beta_j^*(u)|\right) \to^d \infty.$$

A similar result holds for I_{4n}. Let $B = \{1, \ldots, q, p+1, \ldots, p+r\}$. Hence, for every r_u, $V_n(r_u) \to^d V(r_u)$, where

$$V(r_u) = \begin{cases} r_{uB}^\top f(u) \begin{pmatrix} [\Gamma(u)]_{11} & 0 \\ 0 & [\mu_2\Gamma(u)]_{22} \end{pmatrix} r_{uB} - 2r_{uB}^\top \begin{pmatrix} \dfrac{d}{2}\mu_2 f(u)[\Gamma(u)\beta^{*''}(u)]_{B_1} \\ 0 \end{pmatrix} & \\ -2r_{uB}^\top [W]_B & \text{if } r_{uj} = 0 \forall j \notin B \\ 0 & \text{otherwise.} \end{cases}$$

Since $V(\cdot)$ is convex, following the epi-convergence results of [12,17], we have

$$\hat{r}_u^{(n)} \to^d \hat{r}_u = \operatorname*{argmin}_{r_u} V(r_u).$$

Finally, the distribution of $\hat{r}_u$ proves the asymptotic normality part.

To prove the uniform sparsity, it suffices to show that

$$P\left(\sup_u \left|\hat{v}_{\lambda,B^c}^{(n)}(u)\right| = 0\right) \to 1.$$

For any $j \notin B$, we know that the event $\{\sup_u |\hat{\nu}^{(n)}_{\lambda,j}(u)| = 0\}$ is equivalent to the event $\{\forall u, \hat{\nu}^{(n)}_{\lambda,j}(u) = 0\}$. Denoted this by A_{1n}. By the KKT condition, A_{1n} can be implied by the following event, denoted by A_{2n},

$$\sup_u \left| e_j D_u^\top W_u \left(Y - D_u \hat{\nu}^{(n)}_\lambda(u) \right) \right| < \begin{cases} \dfrac{\lambda_n}{w_j}, & j \in \{q+1, \ldots, p\} \\ \dfrac{\gamma_n}{h v_{j-p}}, & j \in \{p+r+1, \ldots, 2p\}, \end{cases}$$

where e_j is a $1 \times 2p$ vector with jth component being 1 while others being 0. Then it suffices to show that $P(A_{2n}) \to 1$. We multiply $h^{1/2} n^{-1/2}$ at both sides of the above inequality. For $j \in \{q+1, \ldots, p\}$, since $\alpha_{1n} w_j = O_p(1)$ and $h\lambda_n \alpha_{1n} / \sqrt{nh \log n} \to \infty$, we can see that $h^{1/2} n^{-1/2} \lambda_n / (w_j \sqrt{\log n}) \to^p \infty$. Similarly, for $j \in \{p+r+1, \ldots, 2p\}$, we can show $(nh)^{-1/2} \gamma_n / (v_{j-p} \sqrt{\log n}) \to^p \infty$. Then, we only need to show that, for $j \notin B$,

$$\sup_u \left| \sqrt{\frac{h}{n}} e_j D_u^\top W_u \left(Y - D_u \hat{\nu}^{(n)}_\lambda(u) \right) \right| = O_p\left(\sqrt{\log n} \right).$$

This follows by noting

$$\begin{aligned} \sup_u \left| \sqrt{\frac{h}{n}} e_j D_u^\top W_u \left(Y - D_u \hat{\nu}^{(n)}_\lambda(u) \right) \right| \le & \sup_u \left| e_j \sqrt{\frac{h}{n}} D_u^T W_u (Y - D_u \nu^*(u)) \right| \\ & + \sup_u \left| e_j \frac{D_u^\top W_u D_u}{n} \sqrt{nh} \left(\hat{\nu}^{(n)}_\lambda(u) - \nu^*(u) \right) \right| \end{aligned}$$

and using (11), along with Theorem 1. □

Proof of Theorem 3. The proof of the first part is similar in spirit to that of [21] and hence omitted. Then the consistent identification of B_1 and B_2 can be implied from part 1 of Theorem 2. □

Proof of Theorem 4. Since $B_2 \subset B_1$, it is equivalent to prove that

$$P\left(\hat{B}_{1\hat{\lambda}} = B_1, \left\{ j : \sum_{t=1}^n |\hat{\beta}^{\prime(n)}_{\hat{\gamma},j}| \neq 0 \right\} = B_2 \right) \to 1.$$

Without a confusion in the notation, we rewrite $\hat{B}_{2\gamma} = \{j : \sum_{t=1}^n |\hat{\beta}^{\prime(n)}_{\gamma,j}| \neq 0\}$. Let $\mathbb{R} = \{(\lambda, \gamma) : \lambda \ge 0, \gamma \ge 0\}$. By Theorem 3, we know that as $n \to \infty$, the probability of the existence of (λ, γ), such that $\hat{B}_{1\lambda} = B_1$ and $\hat{B}_{2\gamma} = B_2$, in $\mathbb{R}$ would go to 1. Then, we can partition $\mathbb{R}$ into three sets $\mathbb{R}_+ = \{(\lambda, \gamma) : B_T \subsetneq B_{\lambda,\gamma}\}$; $\mathbb{R}_0 = \{(\lambda, \gamma) : B_T = B_{\lambda,\gamma}\}$; $\mathbb{R}_- = \{(\lambda, \gamma) : B_T \supsetneq B_{\lambda,\gamma}\}$, corresponding to overfitted (all the correct objectives plus at least one incorrect objective is included); correctly fitted; underfitted (at least one correct objective is missing) respectively. Let $\lambda_n = \gamma_n = n^{1/5} \log n$, by the comments following Remark 2 we know $P(\hat{B}_{1\lambda_n} = B_1, \hat{B}_{2\gamma_n} = B_2) \to 1$. Then the theorem can be proved by comparing $\inf_{\mathbb{R}_-} \mathrm{BIC}(\lambda, \gamma)$, $\inf_{\mathbb{R}_+} \mathrm{BIC}(\lambda, \gamma)$ with $\mathrm{BIC}(\lambda_n, \gamma_n)$. We only present the proof of underfitted case here. The proof of overfitted case is similar to that of [21].

For an arbitrary $(\lambda, \gamma) \in \mathbb{R}_-$, $\mathrm{RSS}(\lambda, \gamma)$ can be written as

$$\mathrm{RSS}(\lambda, \gamma) = \mathrm{RSS}(0, 0) + R(\lambda, \gamma),$$

where $R(\lambda, \gamma) = n^{-1} \sum_{t=1}^n \{\tilde{\nu}(U_t) - \hat{\nu}^{(n)}_\lambda(U_t)\}^\top \hat{\Sigma}(U_t) \{\tilde{\nu}(U_t) - \hat{\nu}^{(n)}_\lambda(U_t)\}$ with $\hat{\Sigma}(u) = D_u^\top W_u D_u / n$. Note that the underfitting may be for B_1 or B_2. If it is in fitting B_1, without loss of generality we assume $\hat{\beta}^{(n)}_{\lambda,1}(U_t) = 0, t = 1, \ldots, n$. We know that $\hat{\lambda}^{\min} = \min\{\lambda_{\min}(\hat{\Sigma}(U_t)), t = 1, \ldots, n\} \to^p \lambda^{\min} > 0$. By Theorem 3, we have

$$\begin{aligned} R(\lambda, \gamma) \ge & \ \hat{\lambda}^{\min} \frac{1}{n} \sum_{t=1}^n \left\| \tilde{\beta}_1(U_t) \right\|^2 \\ \to^p & \ \lambda^{\min} E\{\beta_1^{*2}(U_t)\} > 0. \end{aligned}$$

If the underfitting is in B_2, without loss of generality we assume $\hat{\beta}^{\prime(n)}_{\gamma,1}(U_t) = 0, t = 1, \ldots, n$. Then

$$\begin{aligned} R(\lambda, \gamma) & \ge \hat{\lambda}^{\min} \left\{ \frac{1}{n} \sum_{t=1}^n \left\| h\tilde{\beta}'(U_t) - h\hat{\beta}^{\prime(n)}_\gamma(U_t) \right\|^2 \right\} \\ & \ge \hat{\lambda}^{\min} \frac{1}{n} \sum_{t=1}^n \left\| h\tilde{\beta}'_1(U_t) \right\|^2 = O_p(h^2). \end{aligned}$$

For $\mathrm{RSS}(\lambda_n, \gamma_n)$, we also have $\mathrm{RSS}(\lambda_n, \gamma_n) = \mathrm{RSS}(0,0) + R(\lambda_n, \gamma_n)$. Since $\hat{\lambda}^{\max} = \max\{\lambda_{\max}(\hat{\Sigma}(U_t)), t = 1, \ldots, n\} \to^p \lambda^{\max} > 0$, then

$$\begin{aligned} R(\lambda_n, \gamma_n) &\le \hat{\lambda}^{\max} \left\{ \frac{1}{n} \sum_{t=1}^{n} \left\| \tilde{\nu}(U_t) - \nu^*(U_t) \right\|^2 + \frac{1}{n} \sum_{t=1}^{n} \left\| \hat{\nu}_{\lambda}^{(n)}(U_t) - \nu^*(U_t) \right\|^2 \right\} \\ &= O_p(n^{-4/5}). \end{aligned} \tag{13}$$

The last convergence part in (13) is due to the results of Theorem 3. Hence, combining the fact $\mathrm{RSS}(0,0) \to E[\sigma^2(U)] > 0$ and the definition of $\mathrm{BIC}(\lambda, \gamma)$, we have that

$$\begin{aligned} \inf_{\mathbb{R}_-} \mathrm{BIC}(\lambda, \gamma) - \mathrm{BIC}(\lambda_n, \gamma_n) &\ge \log\{\mathrm{RSS}(0,0) + O_p(n^{-2/5})\} - \log\{\mathrm{RSS}(0,0) + O_p(n^{-4/5})\} - 2p \times O_p\left(\frac{\log(nh)}{nh}\right) \\ &\ge \frac{O_p(n^{-2/5}) - O_p(n^{-4/5})}{O_p(1) + O_p(n^{-2/5})} - 2p \times O_p\left(\frac{\log n}{n^{4/5}}\right) \\ &> 0 \end{aligned} \tag{14}$$

holds with probability converging to 1 as $n \to \infty$. □

References

[1] L. Breiman, Better subset selection using nonnegative garrote, Techonometrics 37 (1995) 373–384.
[2] Z. Cai, J. Fan, R. Li, Efficient estimation and inferences for varying-coefficient models, Journal of the American Statistical Association 95 (2000) 888–902.
[3] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, The Annals of Statistics 32 (2004) 407–489.w.
[4] L.R. Eubank, Nonparametric Regression and Spline Smoothing, second ed., Marcel-Dekker, New York, 1999.
[5] J. Fan, I. Gijbels, Local Polynomial Modeling and Its Applications, Chapman and Hall, New York, 1996.
[6] J. Fan, T. Huang, Profile likelihood inferences on semiparametric varying-coefficient partially linear models, Bernoulli 11 (2005) 1031–1057.
[7] J. Fan, R. Li, Variable selection via non-concave penalized likelihood and its oracle properties, Journal of the American Statistical Association 96 (2001) 1348–1360.
[8] J. Fan, W. Zhang, Simultaneous confidence bands and hypothesis testing in varying-coefficient models, Scandinavian Journal of Statistics 27 (2000) 715–731.
[9] J. Fan, W. Zhang, Statistical estimation in varying coefficient models, The Annals of Statistics 27 (1999) 1491–1518.
[10] J. Fan, C. Zhang, J. Zhang, Generalized likelihood ratio statistics and wilks phenomenon, The Annals of Statistics 29 (2001) 153–193.
[11] W.J. Fu, Penalized regression: the bridge versus the LASSO, Journal of Computational and Graphical Statistics 7 (1998) 397–416.
[12] C. Geyer, On the asymptotics of constrained *M*-estimation, The Annals of Statistics 32 (1994) 928–961.
[13] W. Härdle, Hu. Liang, J.T. Gao, Partially Linear Models, Springer, Heidelberg, 2000, MR 1787637.
[14] T. Hastie, R. Thibshirani, Varying-coefficient models, Journal of the Royal Statistical Society: Series B 55 (1993) 757–796.
[15] J. Huang, J. Horowitz, F. Wei, Variable selection in nonparametric additive models, The Annals of Statistics 38 (2010) 2282–2313.
[16] S. Jason, Maximal spacing in several dimensions, Annals of Probability 15 (1987) 274–280.
[17] K. Knight, W. Fu, Asymptotics for LASSO-type estimators, The Annals of Statistics 28 (2000) 1356–1378.
[18] R. Li, H. Liang, Variable selection in semiparametric regression model, The Annals of Statistics 36 (2008) 261–286.
[19] Y.P. Mack, B.W. Silverman, Weak and strong uniform consistency of kernel regression estimates, Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete 61 (1982) 405–415.
[20] R. Tibshirani, Regression shrinkage and selection via the LASSO, Journal of the American Statistical Association 101 (1996) 1418–1429.
[21] H. Wang, Y. Xia, Shrinkage estimation of the varying coefficient model, Journal of the American Statistical Association 104 (2009) 747–757.
[22] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B 68 (2006) 49–67.
[23] W. Zhang, S.Y. Lee, X. Song, Local polynomial fitting in semi-varying coefficient models, Journal of Multivariate Analysis 82 (2002) 166–188.
[24] H.H. Zhang, Y. Lin, Component selection and smoothing in smoothing spline analysis of variance models, The Annals of Statistics 34 (2003) 2272–2297.
[25] H. Zou, The adaptive LASSO and its oracle properties, Journal of the American Statistical Association 101 (2006) 1418–1429.
[26] H. Zou, R. Li, One-step sparse estimates in non-concave penalized likelihood models, The Annals of Statistics 36 (4) (2008) 1509–1533.