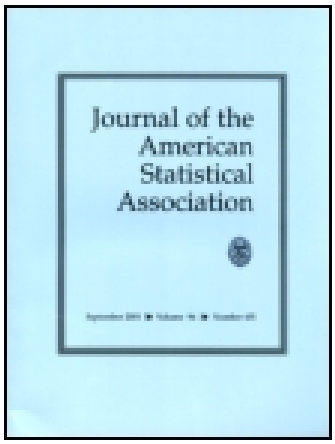

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Shrinkage Estimation of the Varying Coefficient Model

Hansheng Wang[a] & Yingcun Xia[a]

[a] Hansheng Wang is Associate Professor of Statistics, Guanghua School of Management, Peking University, Beijing, People's Republic of China, 100871 . Yingcun Xia is Associate Professor, Department of Statistics and Applied Probability, Risk Management Institute, National University of Singapore, Singapore, 117546 . Hansheng Wang's research is partially supported by a grant from NSFC (10771006) and also a grant from Microsoft Research Asia. Yingcun Xia's research is partially supported by a grant from Institute of Risk Management and also a grant from National Universitiy of Singapore (FRG R-155-000-063-112). The authors thank the editor, associate editor, and referees for their helpful comments and suggestions.
Published online: 01 Jan 2012.

To cite this article: Hansheng Wang & Yingcun Xia (2009) Shrinkage Estimation of the Varying Coefficient Model, Journal of the American Statistical Association, 104:486, 747-757, DOI: 10.1198/jasa.2009.0138

To link to this article: http://dx.doi.org/10.1198/jasa.2009.0138

Shrinkage Estimation of the Varying Coefficient Model

Hansheng WANG and Yingcun XIA

The varying coefficient model is a useful extension of the linear regression model. Nevertheless, how to conduct variable selection for the varying coefficient model in a computationally efficient manner is poorly understood. To solve the problem, we propose here a novel method, which combines the ideas of the local polynomial smoothing and the Least Absolute Shrinkage and Selection Operator (LASSO). The new method can do nonparametric estimation and variable selection simultaneously. With a local constant estimator and the adaptive LASSO penalty, the new method can identify the true model consistently, and that the resulting estimator can be as efficient as the oracle estimator. Numerical studies clearly confirm our theories. Extension to other shrinkage methods (e.g., the SCAD, i.e., the Smoothly Clipped Absolute Deviation.) and other smoothing methods is straightforward.

KEY WORDS: Bayesian information criterion; Kernel smoothing; Least Absolute Shrinkage and Selection Operator; Oracle property; Smoothly Clipped Absolute Deviation; Variable selection; Varying coefficient model.

1. INTRODUCTION

The varying coefficient model is an important generalization of the linear regression model and has gained a lot of popularity during the past decade (Chen and Tsay 1993; Hastie and Tibshirani 1993; Fan and Zhang 1999; Cai, Fan, and Li 2000; Fan and Zhang 2000a b; Huang et al. 2002, 2004; Fan and Huang 2005).

Most existing research about the model focuses on either parameter estimation (Chen and Tsay 1993; Fan and Zhang 1999; Cai et al. 2000; Fan and Zhang 2000a; Huang, Wu, and Zhou 2002; Huang, Wu, and Zhou 2004; Fan and Huang 2005) or hypotheses testing (Fan and Zhang 2000b; Fan, Zhang, and Zhang 2001; Li and Liang 2008). Much less has been done about its variable selection. In a typical linear regression setup, it has been very well understood that ignoring any important predictor can lead to seriously biased results, whereas including spurious covariates can degrade the estimation efficiency substantially. Thus, variable selection is important for any regression problems. In a traditional linear regression setting, many selection criteria [e.g., Akaike information criterion (AIC) and Bayesian information criterion (BIC)] have been extensively used in practice. Nevertheless, those selection methods suffer expensive computational costs (Breiman 1995; Tibshirani 1996; Fan and Li 2001). As computational efficiency is more desirable in many situations, various shrinkage methods have been developed, which include but are not limited to the nonnegative garrotte (Breiman 1995; Yuan and Lin 2007), the Least Absolute Shrinkage and Selection Operator (Tibshirani 1996; Zou 2006; Zhang and Lu 2007; Wang, Li, and Tsai 2007a), the bridge regression (Fu 1998; Knight and Fu 2000), the Smoothly Clipped Absolute Deviation (Fan and Li 2001), and the one-step sparse estimator (Zou and Li 2007).

A number of works have been done to extend the regularized estimation methods to semiparametric models. For example, Fan and Li (2004) extended the SCAD to partially linear models (Härdle, Liang, and Gao 2000), but their focus is on the parametric components. Recently, Li and Liang (2008) carefully studied variable selection for partially linear varying coefficient models, where the parametric components are identified via the SCAD (Fan and Li 2001) but the nonparametric components are selected via a generalized likelihood ratio test, instead of a shrinkage method. Thus, we are motivated to develop a more genuine shrinkage method that is able to select important nonparametric components automatically. Furthermore, given the popularity of the kernel smoothing methods, it is very desirable to have a shrinkage method that can work with kernel smoothing techniques in a very natural way. As we will demonstrate later, our proposal perfectly meets such a requirement. The proposed method can be extended easily to many other semiparametric models, where the local polynomial methods have been found very useful. This also differentiates our proposal from the spline-based methods; see, for example, Zhang and Lin, (2003).

It is remarkable that extending the shrinkage ideas (e.g., LASSO) to semiparametric models (e.g., varying coefficient models) is not trivial (Zhang and Lin 2003; Fan and Li 2004). For a parametric model, it involves only one type of tuning parameters (i.e., the shrinkage parameters). To achieve the oracle property, the appropriate convergence rates for the shrinkage parameters have been well investigated (Fan and Li 2001; Zou 2006; Zhang and Lu 2007; Wang et al. 2007a). Nevertheless, for a semiparametric model, it involves another type of regularization parameters (i.e., the smoothing parameters) (e.g., the bandwidths); see Zhang and Lee (2007). What should be the right convergence speed for the shrinkage parameters under this situation is much less well understood. In a typical regression setting, Wang et al. (2007b) and Wang and Leng (2007) proposed BIC-type criteria for shrinkage parameter selection. They showed theoretically that the shrinkage parameters selected by their BIC criteria can identify the true model consistently. Whether similar criteria can be developed under a semiparametric setup (e.g., the varying coefficient model) is not clear and not investigated to the best of our knowledge.

Hansheng Wang is Associate Professor of Statistics, Guanghua School of Management, Peking University, Beijing, People's Republic of China, 100871 (E-mail: *hansheng@gsm.pku.edu.cn*). Yingcun Xia is Associate Professor, Department of Statistics and Applied Probability, Risk Management Institute, National University of Singapore, Singapore, 117546 (E-mail: *staxyc@nus.edu.sg*). Hansheng Wang's research is partially supported by a grant from NSFC (10771006) and also a grant from Microsoft Research Asia. Yingcun Xia's research is partially supported by a grant from Institute of Risk Management and also a grant from National Universitiy of Singapore (FRG R-155-000-063-112). The authors thank the editor, associate editor, and referees for their helpful comments and suggestions.

Journal of the American Statistical Association
June 2009, Vol. 104, No. 486, Theory and Methods
DOI 10.1198/jasa.2009.0138

As a preliminary but very important first attempt, we consider here the application of LASSO (a typical shrinkage method) to the varying coefficient model (a popular semiparametric model) with local constant kernel estimation. The convexity of L_1 penalty and the simplicity of the local constant estimation enable us to demonstrate the method easily. Nevertheless, the same idea is readily applicable to all combinations of other shrinkage methods and nonparametric smoothing methods. In fact, without much difficulty, we can replace the L_1 penalty by the SCAD (Fan and Li 2001) or implement our general idea via the most recently developed one-step sparse estimator of Zou and Li (2007). We refer to Fan and Li (2001) for an excellent discussion about the advantages of the nonconcave penalty (e.g., the SCAD) and Fan and Gijbels (1996) for the advantages of local polynomial smoothing.

The rest of the article is organized as follows. Section 2 introduces the new method. Its theoretical properties are carefully studied in Section 3. Numerical studies are reported in Section 4. The article is concluded with a brief discussion in Section 5. All technical details are left to the Appendixes.

2. THE METHODOLOGY

2.1 Model and Notations

Let (X_i,Y_i,Z_i) be the observation collected from the ith subject ($1 \le i \le n$), where $Y_i \in \mathbb{R}^1$ is the response of interest, $X_i = (X_{i1}, \ldots, X_{id})^\top \in \mathbb{R}^d$ is the d-dimensional predictor, and $Z_i \in [0,1]$ is the so-called univariate *index* variable. A typical varying coefficient model assumes that

$$Y_i = X_i^\top \beta(Z_i) + e_i, \tag{1}$$

where $e_i \in \mathbb{R}^1$ is the random noise satisfying $E(e_i|X_i,Z_i) = 0$ almost surely. Coefficient vector $\beta(z) = \{\beta_1(z), \ldots, \beta_d(z)\}^\top \in \mathbb{R}^d$ is an unknown but smooth function in z, whose true value is given by $\beta_0(z) = \{\beta_{01}(z), \ldots, \beta_{0d}(z)\}^\top \in \mathbb{R}^d$. Next, we assume without loss of generality that there exists an integer $d_0 \le d$ such that $\infty > E\{\beta_{0j}^2(Z_i)\} > 0$ for any $j \le d_0$ but $E\{\beta_{0j}^2(Z_i)\} = 0$ for any $d_0 < j$. Simply speaking, we assume that the first d_0 predictors are truly relevant but the rest are not.

2.2 The Kernel Least Absolute Shrinkage and Selection Operator Method

For an arbitrary *index* value $z \in [0,1]$, $\beta(z)$ can be estimated by minimizing the following locally weighted least squares function (Fan and Zhang 2000b)

$$Q_z(\beta) = \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 K_h(z - Z_i) \tag{2}$$

with respect to β. Denote the resulting estimator by $\tilde{\beta}(z)$. Based on $\tilde{\beta}(z)$, we further define $\tilde{\beta}_t = \tilde{\beta}(Z_t)$ and $\tilde{B} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_n)^\top \in \mathbb{R}^{n \times d}$. As one can see, $\tilde{B}$ is a natural estimator for $B_0 = \{\beta_0(Z_1), \ldots, \beta_0(Z_n)\}^\top \in \mathbb{R}^{n \times d}$. Furthermore, one can verify that $\tilde{B}$ is also the minimizer of the following global least squares function

$$\begin{aligned} Q(B) &= \sum_{t=1}^{n} Q_{Z_t}(\beta_t) \\ &= \sum_{t=1}^{n} \sum_{i=1}^{n} \{Y_i - X_i^\top \beta_t\}^2 K_h(Z_t - Z_i) \end{aligned} \tag{3}$$

with respect to $B = \{\beta(Z_1), \ldots, \beta(Z_n)\}^\top = (\beta_1, \ldots, \beta_n)^\top \in \mathbb{R}^{n \times d}$. To see this, note that $Q(B)$ is a quadratic function in B. Thus, its minimizer is solely determined by the normal equation $\partial \mathrm{RSS}(B)/\partial \beta_t = 0$ for every $1 \le t \le n$. On the other hand, for $Q(B)$, we find β_t is only involved in $Q_{Z_t}(\beta_t)$; see (2). Consequently, we have $\partial Q(B)/\partial \beta_t = \partial Q_{Z_t}(\beta_t)/\partial \beta_t = 0$, leading to the solution $\tilde{\beta}(Z_t) = \mathrm{argmin}_{\beta_t} Q_{Z_t}(\beta_t)$; see (2). Thus, $\tilde{B}$ is also the minimizer of (3).

Note that, the last $(d - d_0)$ columns of the B_0 matrix should be 0 under our model assumption. Thus, the task of variable selection becomes equivalent to identifying sparse columns in matrix B_0. To claim an irrelevant variable, we need to identify sparse solutions in B_0 in a column-wise manner. Directly applying the LASSO (Tibshirani 1996) or the adaptive LASSO (Zou 2006; Zhang and Lu 2007; Wang et al. 2007a) method to B_0 is not efficient. Following the group LASSO idea of Yuan and Lin (2006), we propose the following penalized estimate

$$\begin{aligned} \hat{B}_\lambda &= \{\hat{\beta}_\lambda(Z_1), \ldots, \hat{\beta}_\lambda(Z_n)\}^\top = (\hat{b}_{\lambda,1}, \ldots, \hat{b}_{\lambda,d}) \\ &= \mathrm{argmin}_{B \in \mathbb{R}^{n \times d}} Q_\lambda(B), \end{aligned} \tag{4}$$

where $\lambda = (\lambda_1, \ldots, \lambda_d)^\top \in \mathbb{R}^d$ is the tuning parameter,

$$\begin{aligned} \hat{\beta}_\lambda(z) &= \{\hat{\beta}_{\lambda,1}(z), \ldots, \hat{\beta}_{\lambda,d}(z)\}^\top \in \mathbb{R}^d, \\ \hat{b}_{\lambda,j} &= \{\hat{\beta}_{\lambda,j}(Z_1), \ldots, \hat{\beta}_{\lambda,j}(Z_n)\}^\top \in \mathbb{R}^n, \\ Q_\lambda(B) &= \sum_{t=1}^{n} \sum_{i=1}^{n} \{Y_i - X_i^\top \beta(Z_t)\}^2 K_h(Z_t - Z_i) \\ &\quad + \sum_{j=1}^{d} \lambda_j \| b_j \|, \end{aligned} \tag{5}$$

$b_j \in \mathbb{R}^{n \times 1}$ is the jth column of B, and $\|\cdot\|$ stands for the usual Euclidean norm.

2.3 Local Quadratic Approximation

In a typical least squares setting, the computational algorithms for the LASSO-type problems have been very well developed. These algorithms include the shooting algorithm (Fu 1998; Yuan and Lin 2006), local quadratic approximation (Fan and Li, 2001), the least angle regression (Efron, Hastie, Johnstone, and Tibshirani 2004), and many others (Zhao and Yu 2004; Park and Hastie 2007). For the purpose of completeness and simplicity, we describe here an easy implementation based on the idea of the local quadratic approximation (Fan and Li 2001). Specifically, our implementation is based on an iterative algorithm with $\tilde{B}$ (i.e., the unpenalized estimator) as the initial estimator. Next, we define

$$\hat{B}_\lambda^{(m)} = \left(\hat{b}_{\lambda,1}^{(m)}, \ldots, \hat{b}_{\lambda,d}^{(m)}\right) = \left\{\hat{\beta}_\lambda^{(m)}(Z_1), \ldots, \hat{\beta}_\lambda^{(m)}(Z_n)\right\}^\top$$

to be the KLASSO estimate obtained in the mth iteration. Then, the loss function in (4) can be locally approximated by (Fan and Li 2001; Hunter and Li 2005)

$$\begin{aligned} &\sum_{t=1}^{n} \sum_{i=1}^{n} \{Y_i - X_i^\top \beta(Z_t)\}^2 K_h(Z_t - Z_i) + \sum_{j=1}^{d} \lambda_j \frac{\| b_j \|^2}{\| \hat{b}_{\lambda,j}^{(m)} \|} \\ &= \sum_{t=1}^{n} \left(\sum_{i=1}^{n} \{Y_i - X_i^\top \beta(Z_t)\}^2 K_h(Z_t - Z_i) + \sum_{j=1}^{d} \lambda_j \frac{\beta_j^2(Z_t)}{\| \hat{b}_{\lambda,j}^{(m)} \|} \right), \end{aligned}$$

whose minimizer is given by $\hat{B}_\lambda^{(m+1)}$ with the tth row given by

$$\hat{\beta}_\lambda^{(m+1)}(Z_t) = \left(\sum_{i=1}^{n} X_i X_i^\top K_h(Z_t - Z_i) + D^{(m)}\right)^{-1} \times \left(\sum_{i=1}^{n} X_i Y_i K_h(Z_t - Z_i)\right), \quad (6)$$

where $D^{(m)}$ is a $d \times d$ diagonal matrix with its jth diagonal component given by $\lambda_j / \| \hat{b}_{\lambda,j}^{(m)} \|, j = 1, \ldots, d$. If the function $\beta(z)$ is interested, then it can be estimated via

$$\hat{\beta}_\lambda^{(m+1)}(z) = \left(\sum_{i=1}^{n} X_i X_i^\top K_h(z - Z_i) + D^{(m)}\right)^{-1} \times \left(\sum_{i=1}^{n} X_i Y_i K_h(z - Z_i)\right). \quad (7)$$

Denote the limit values of $\hat{B}_\lambda^{(m+1)}$ and $\hat{\beta}_\lambda^{(m+1)}(z)$ respectively by $\hat{B}_\lambda$ and $\hat{\beta}_\lambda(z)$. Because the estimate $\hat{\beta}_\lambda(z)$ is obtained based on a combination of the kernel smoothing and the LASSO methods, we refer to it as a Kernel LASSO (KLASSO) estimate.

3. THEORETICAL PROPERTIES

3.1 Technical Conditions

To study the asymptotic properties of the KLASSO method, we define $a_n = \max\{\lambda_j: 1 \le j \le d_0\}$ and $b_n = \min\{\lambda_j: d_0 < j \le d\}$. In other words, a_n and b_n define the maximal and minimal amounts of shrinkages applied to relevant and irrelevant coefficients, respectively (e.g., see Wang et al. (2007a). Moreover, the following standard regularity conditions are needed (Fan and Huang 2005).

(C1) For an $s > 2$, we must have $E|Y_i|^{2s} < \infty$ and $E \| X_{ij} \|^{2s} < \infty$.
(C2) The density function of Z_i, $f(z)$, is continuous and positively bounded away from 0 on [0, 1].
(C3) Matrix $\Omega(z) = E(X_i X_i^\top | Z_i = z)$ is nonsingular and has bounded second order derivatives on [0, 1]. Function $E(\| X_i \|^4 | Z_i = z)$ is also bounded.
(C4) The second order derivative of $f(z)$ and $\sigma^2(z) = E(e_i^2 | Z_i = z)$ are bounded.
(C5) $K(z)$ is a symmetric density function with a compact support.
(C6) The second order derivatives of coefficients $\beta_{0j}(z), j = 1, \ldots d$, are continuous.

Remark 1. Note that (C2) guarantees the maximal distance between two consecutive *index* variables is only of the order $O_p(\log n/n)$ (e.g., see Janson 1987). For an arbitrary *index* value $z \in [0, 1]$, let z^* be its nearest neighbor among the observed *index* values, i.e., $z^* = \text{argmin}_{\tilde{z} \in \{Z_t: 1 \le t \le n\}} |z - \tilde{z}|$. Under the smoothness assumption (C6), we have $\|\beta_0(z) - \beta_0(z^*)\| = O_p(\log n/n)$ also, which is an order substantially smaller than the optimal nonparametric convergence rate (i.e., $n^{-2/5}$). Practically, this means that the observed *index* values are sufficiently dense on the support. Thus, it suffices to approximate the entire coefficient curve $\beta_0(z)$ by $\{\beta(Z_t): 1 \le t \le n\}$. This explains why KLASSO performs so well, even though (7) focus only on $\{\beta(Z_t): 1 \le t \le n\}$, instead of the entire $\{\beta(z): 0 \le z \le 1\}$.

3.2 Basic Theoretical Properties

Based on the regularity conditions listed in the previous subsection, the sparsity and oracle efficiency for the estimators of KLASSO can be established. Specifically, we define $X_{ia} = (X_{i1}, \ldots, X_{id_0})^\top \in \mathbb{R}^{d_0}$ and $X_{ib} = (X_{id_0+1}, \ldots, X_{id})^\top \in \mathbb{R}^{d-d_0}$. Furthermore, we define accordingly $\hat{\beta}_{a,\lambda}(z) = \{\hat{\beta}_{\lambda,1}(z), \ldots, \hat{\beta}_{\lambda,d_0}(z)\}^\top \in \mathbb{R}^{d_0}$ and $\hat{\beta}_{b,\lambda}(z) = \{\hat{\beta}_{\lambda,(d_0+1)}(z), \ldots, \hat{\beta}_{\lambda,d}(z)\}^\top \in \mathbb{R}^{d-d_0}$.

Theorem 1—Estimation Sparsity. Assume (C1)−(C6), $h \propto n^{-1/5}$, $n^{11/10} a_n \to 0$, and $n^{11/10} b_n \to \infty$, then we have $P(\sup_{z \in [0,1]} \| \hat{\beta}_{\lambda,b}(z) \| = 0) \to 1$ for any $d_0 < j \le d$.

By Theorem 1 we know that, as long as Equation (7) is used to compute $\hat{\beta}_\lambda(z)$, sparse solutions can be consistently produced for every irrelevant predictor over the entire *index* support uniformly. To establish the oracle properties, define the oracle estimator (i.e., the unpenalized estimator obtained under the true model) as

$$\hat{\beta}_{ora}(z) = \left\{\frac{1}{n}\sum_{i=1}^{n} X_{ia} X_{ia}^\top K_h(Z_i - z)\right\}^{-1} \times \left\{\frac{1}{n}\sum_{i=1}^{n} X_{ia} Y_i K_h(Z_i - z)\right\}. \quad (8)$$

Then, the following theorem establishes the oracle property.

Theorem 2—Oracle Property. Assume the technical conditions (C1)−(C6) hold. If $h \propto n^{-1/5}$, $n^{11/10} a_n \to 0$, and $n^{11/10} b_n \to \infty$, we then have

$$\sup_{z \in [0,1]} \| \hat{\beta}_{a,\lambda}(z) - \hat{\beta}_{ora}(z) \| = o_p(n^{-2/5}).$$

Note that the optimal pointwise rate of the oracle estimator is $O_p(n^{-2/5})$. Theorem 2 indicates that the difference between the KLASSO estimate and the oracle estimate is negligible uniformly over the entire *index* support. Consequently, we know that $\hat{\beta}_{a,\lambda}(z)$ shares the same asymptotic distribution (and efficiency) as the oracle estimator $\hat{\beta}_{ora}(z)$ (see Fan and Zhang 1999, 2000b).

Remark 2. Based on the oracle properties in Theorem 2, most statistical inference for $\hat{\beta}_{a,\lambda}(z)$ can be made exactly the same as the oracle estimators. Here, we consider the simultaneous confidence band as one example. Suppose $K(t)$ is supported on a compact interval $[-A, A]$ for some constant $A > 0$. Furthermore, assume that $K(A) = 0$ (e.g., the Epanechnikov kernel). Then, for a given confidence level $(1 - \alpha)$, the $(1 - \alpha)$th simultaneous confidence band (without bias correction) is given by $[\hat{\beta}_{\lambda,j}(z) - \Delta(z), \hat{\beta}_{\lambda,j}(z) + \Delta(z)]$, where

$$\Delta(z) = \{d_n + [\log 2 - \log\{-\log(1 - \alpha)\}](-2\log h)^{-1/2} \times \widehat{\text{SD}}\{\hat{\beta}_{\lambda,j}(z)\},$$

$$d_n = (-2\log h)^{1/2} + \frac{1}{(-2\log h)^{1/2}} \log\left\{\frac{1}{4\upsilon_0 \pi}\int \{K'(t)\}^2 dt\right\},$$

$\upsilon_0 = \int K^2(t)dt, K'(t) = \partial K(t)\partial(t), \widehat{\text{SD}}^2\{\hat{\beta}_{\lambda,j}(z)\}$ is the jth diagonal component of

$$\widehat{\operatorname{cov}}\{\hat{\beta}_\lambda(z)\} = \frac{\hat{\sigma}_e^2(z)}{n}\left(n^{-1}\sum_{i=1}^n K_h(z-Z_i)X_iX_i^\top\right)^{-1} \times\left(n^{-1}\sum_{i=1}^n K_h^2(z-Z_i)X_iX_i^\top\right) \times\left(n^{-1}\sum_{i=1}^n K_h(z-Z_i)X_iX_i^\top\right)^{-1},$$

$$\hat{\sigma}_e^2(z) = \left[\frac{1}{n}\sum_{i=1}^n K_h(Z_i-z)\{Y_i - X_i^\top\hat{\beta}_\lambda(z)\}^2\right] \times\left(n^{-1}\sum_{i=1}^n K_h(Z_i-z)\right)^{-1}.$$

We refer to Fan and Zhang (2000b) for more details.

3.3 Tuning Parameter Selection

By the results of Theorem 1 and Theorem 2, we know that as long as the following two technical conditions

$$n^{11/10}a_n \to 0 \quad \text{and} \quad n^{11/10}b_n \to \infty \tag{9}$$

are satisfied, the optimal nonparametric convergence rate can be achieved and the true model can be consistently selected. Nevertheless, in real application, how to simultaneously select a total of d shrinkage parameters (i.e., λ_j, $1 \le j \le d$); satisfying that requirement is challenging. To bypass this difficulty, we follow the idea of (Zou 2006; Zhang and Lu 2007; Wang et al. 2007a; Zou and Li 2007), and simplify the tuning parameters as

$$\lambda_j = \frac{\lambda_0}{n^{-1/2}\parallel\tilde{\beta}_j\parallel}, \tag{10}$$

where $\tilde{\beta}_j$ is the jth column of the unpenalized estimate $\tilde{B}$. Because $\tilde{\beta}_j$ is also a KLASSO estimator with $\lambda_j = 0$, the theoretical results of Theorem 1 and Theorem 2 can be applied, yielding $n^{-1/2}\parallel\tilde{\beta}_j\parallel \to_p (E\{\beta_j^2(Z_i)\})^{1/2} > 0$ for $j \le d_0$ while $n^{-1/2}\parallel\tilde{\beta}_j\parallel = O_p(n^{-2/5})$ for $j > d_0$. One can verify that as long as $\lambda_0 n^{11/10} \to 0$ but $\lambda_0 n^{3/2} \to \infty$, then the two conditions listed in (9) are satisfied. Consequently, the original d-dimensional problem about $\lambda \in \mathbb{R}^d$ becomes a univariate problem about $\lambda_0 \in \mathbb{R}^1$, where λ_0 can be selected according to the following BIC-type criterion

$$\mathrm{BIC}_\lambda = \log(\mathrm{RSS}_\lambda) + df_\lambda \times \frac{\log(nh)}{nh}, \tag{11}$$

where $0 \le df_\lambda \le d$ is simply the number of nonzero coefficients identified by $\hat{B}_\lambda$ (Wang et al. 2007b; Wang and Leng 2007), and RSS_λ is defined as

$$\mathrm{RSS}_\lambda = n^{-2}\sum_{t=1}^n\sum_{i=1}^n\{Y_i - X_i^\top\hat{\beta}_\lambda(Z_t)\}^2 K_h(Z_t - Z_i). \tag{12}$$

It is remarkable that the effective sample size nh is used instead of the original sample size n. Therefore, the BIC_λ in (11) replaces $\log(n)/n$ in the parametric BIC by $\log(nh)/(nh)$. The tuning parameter can be obtained as

$$\hat{\lambda} = \operatorname*{argmin}_\lambda \mathrm{BIC}_\lambda.$$

Define $\mathcal{S} = \{j_1, \ldots, j_{d^*}\}$ as an arbitrary model with a total of $0 \le d^* \le d$ nonzero coefficients (i.e., $X_{ij_1}, \ldots, X_{ij_{d^*}}$). Then, we use $\mathcal{S}_T = \{1, \ldots, d_0\}$ to denote the true model and $\mathcal{S}_\lambda = \{j : \parallel\hat{\beta}_{\lambda,j}\parallel > 0\}$ to represent the model identified by the KLASSO estimate $\hat{B}_\lambda$. Consequently, $\mathcal{S}_{\hat{\lambda}}$ represents the model identified by $\hat{B}_{\hat{\lambda}}$.

Theorem 3—Selection Consistency. Assume conditions (C1)–(C6) hold, the tuning parameter $\hat{\lambda}$ selected by the BIC criterion (11) can indeed identify the true model consistently, i.e., $P(\mathcal{S}_{\hat{\lambda}} = \mathcal{S}_T) \to 1$ as n $\to \infty$.

4. NUMERICAL EXPERIMENTS

4.1 Simulation Examples

To demonstrate the finite sample performance of the proposed KLASSO method, we consider the following three varying coefficient models.

(I) $Y_i = 2\sin(2\pi Z_i)X_{i1} + 4Z_i(1-Z_i)X_{i2} + \sigma_e \times e_i,$

(II) $Y_i = \exp(2Z_i - 1)X_{i1} + 8Z_i(1-Z_i)X_{i2} + 2\cos^2(2\pi Z_i)X_{i3} + \sigma_e \times e_i,$

(III) $Y_i = 4Z_iX_{i1} + 2\sin(2\pi Z_i)X_{i2} + X_{i3} + \sigma_e \times e_i,$

where $X_{i1} = 1$ and $(X_{i2}, \ldots, X_{i7})^\top$ is generated from a multivariate normal distribution with $\operatorname{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$ for any $2 \le j_1, j_2 \le 7$. e_i is simulated from $N(0, 1)$. The *index* variable is simulated from either *Uniform*[0, 1], a symmetric distribution, or *Beta*(4, 1), a highly asymmetric Beta distribution. As one can see, for model (I), only the first two variables are relevant. However, for both models (II) and (III), the first three are relevant. The value of σ_e is given by 3.0 or 1.5. The simulation results are qualitatively similar. To save space, only the results about $\sigma_e = 1.5$ are reported in Table 1. A total of 1,000 simulation replications are conducted for each model setup. For every simulated data, we first fit an unpenalized varying coefficient estimate $\hat{\beta}(Z_i)$, for which the Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$ is used, and the optimal bandwidth is selected via the method of leaving-one-out cross-validation. Then, the same bandwidth is used for KLASSO, where the optimal shrinkage parameter is determined by the BIC criterion (11).

To summarize the variable selection results, we differentiate three different situations. Whenever the estimated model misses at least one relevant predictor, we classify it as an underfitted model. Whenever the estimated model includes at least one irrelevant predictor but does not miss any relevant one, we classify it as an overfitted model. Whenever the resulting model is exactly the same as the true model, we classify it as the correctly fitted model. Then, the percentage of the experiments with correctly (under, over) fitted models are summarized in Table 1. For a better understanding of the results, we also report the average number of correctly and incorrectly identified 0 coefficients. To evaluate the estimation accuracy of KLASSO estimate, we consider the following relative estimation error (REE)

$$\mathrm{REE} = 100 \times \frac{\sum_{i=1}^n\sum_{j=1}^p|\hat{\beta}_{\lambda,j}(Z_i) - \beta_{0j}(Z_i)|}{\sum_{i=1}^n\sum_{j=1}^p|\bar{\beta}_j(Z_i) - \beta_{0j}(Z_i)|},$$

where $\bar{\beta}_j(\cdot)$ is either the unpenalized estimator or the oracle estimator. Thus, the corresponding REE value measures the

Table 1. The simulation results with $\sigma_e = 1.5$ and 1,000 simulation replications

		Number of estimated zeros		Percentage of models			MREE (%)	
$f(z)$	n	Correct	Incorrect	Under fitted	Correctly fitted	Over fitted	Unpenalized estimate	Oracle estimate
					Model I			
U[0,1]	100	4.79	0.09	0.09	0.74	0.16	40.46	121.00
	200	4.97	0.02	0.02	0.95	0.03	36.19	115.45
	400	5.00	0.00	0.00	1.00	0.00	35.08	110.55
B(4,1)	100	4.61	0.21	0.21	0.58	0.21	47.77	127.42
	200	4.94	0.08	0.08	0.86	0.05	46.46	122.12
	400	4.99	0.02	0.02	0.97	0.01	43.98	120.51
					Model II			
U[0,1]	100	3.79	0.01	0.01	0.83	0.16	57.32	109.45
	200	3.99	0.00	0.00	0.99	0.01	52.15	109.46
	400	4.00	0.00	0.00	1.00	0.00	50.44	107.36
B(4,1)	100	3.79	0.01	0.01	0.82	0.18	61.13	111.05
	200	3.96	0.00	0.00	0.96	0.04	58.19	108.07
	400	4.00	0.00	0.00	1.00	0.00	56.30	107.03
					Model III			
U[0,1]	100	3.84	0.02	0.02	0.85	0.13	55.10	116.53
	200	3.99	0.00	0.00	0.99	0.01	50.56	110.59
	400	4.00	0.00	0.00	1.00	0.00	48.20	106.94
B(4,1)	100	3.77	0.02	0.02	0.79	0.19	60.78	118.91
	200	3.96	0.00	0.00	0.96	0.04	56.84	113.43
	400	4.00	0.00	0.00	1.00	0.00	53.91	108.40

NOTE: U[0,1] stands for uniform[0,1] distribution, and B(4,1) stands for beta(4,1) distribution.

estimation accuracy of $\hat{\beta}_\lambda(Z_i)$ relative to that of $\bar{\beta}_j(Z_i)$ (e.g., unpenalized or oracle). Then, for each model and parameter setting, the median of REE values (denoted as MREE) are summarized.

As one can see from Tables 1, all MREE ratios of the penalized estimator to the unpenalized estimator are much less than 100% and can be as small as 35.08% (see Table 1, Model I, $U[0, 1]$, $n = 400$). They clearly indicate that the KLASSO estimates are much more accurate than the unpenalized estimates. Furthermore, for every model and noise level, the percentage of the correctly fitted models steadily increases as the sample size increases, and approaches 100% quickly, which confirms that our BIC criterion (11) can indeed identify the true model consistently. As a consequence, we find the MREE ratios of the penalized estimator to the oracle estimator approaches 100% quickly, which corroborates the oracle properties of the KLASSO estimator. Last, we find that the results of *Beta*(4, 1) is very similar to that of *Uniform*[0, 1]. Thus, we conclude that the role played by the *index* distributions (on KLASSO's finite sample performance) is rather limited.

4.2 The Boston Housing Data

To further illustrate the usefulness of KLASSO, we consider here the Boston Housing Data, which has been analyzed by Fan and Huang (2005) and is publicly available in the R package *mlbench*, (*http://cran.r-project.org/*). Following Fan and Huang (2005), we take MEDV [median value of owner-occupied homes in 1,000 United States dollar (USD)] as the response, LSTAT (the percentage of lower status of the population) as the *index* variable, and the following predictors as the *X*-variables: INT (the intercept), CRIM (per capita crime rate by town), RM (average number of rooms per dwelling), PTRATIO (pupil-teacher ratio by town), NOX (nitric oxides concentration parts per 10 million), TAX (full-value property-tax rate per 10,000 USD), and AGE (proportion of owner-occupied units built prior to 1940). By doing so, different regression models can be fitted at different lower status population percentage (see Fan and Huang 2005). Before applying our method, both the response and the *X*-variables (except INT) are transformed so that their marginal distribution is approximately $N(0, 1)$. Moreover, the *index* variable LSTAT is transformed so that its marginal distribution is $U[0, 1]$.

First, a standard leaving-out-one cross-validation method without penalization suggests an optimal bandwidth $h = 0.1739$. The KLASSO method is then applied with this bandwidth. The optimal shrinkage parameter then selected by the BIC criterion (11) is given by $\hat{\lambda}_0 = 0.060$. The resulting KLASSO estimate suggests that INT, CRIM, RM, and PTRATIO are all relevant variables, whereas NOX, TAX, and AGE are not. To confirm whether the selected variables (i.e., INT, CRIM, RM, and PTRATIO) are truly relevant, we provide in Figure 1 their KLASSO estimates (the solid lines) and their 90% simultaneous confidence bands (the dashed lines) (see Fan and Zhang 2000b and Remark 2). Obviously, they all suggest that those four coefficient functions are unlikely to be constant zero, because none of their simultaneous confidence bands can completely well cover 0. To confirm whether the eliminated variables (i.e., NOX, TAX, and AGE) are truly irrelevant, we provide in Figure 2 their unpenalized estimates and their 90% simultaneous confidence bands (the dashed lines). We find that 0 is almost always well covered by the simultaneous confidence bands, over the entire range of the *index* variable. Thus, Figure 2 further confirms that those variables eliminated by KLASSO are unlikely to be relevant. Those findings corroborate the KLASSO selection results very well.

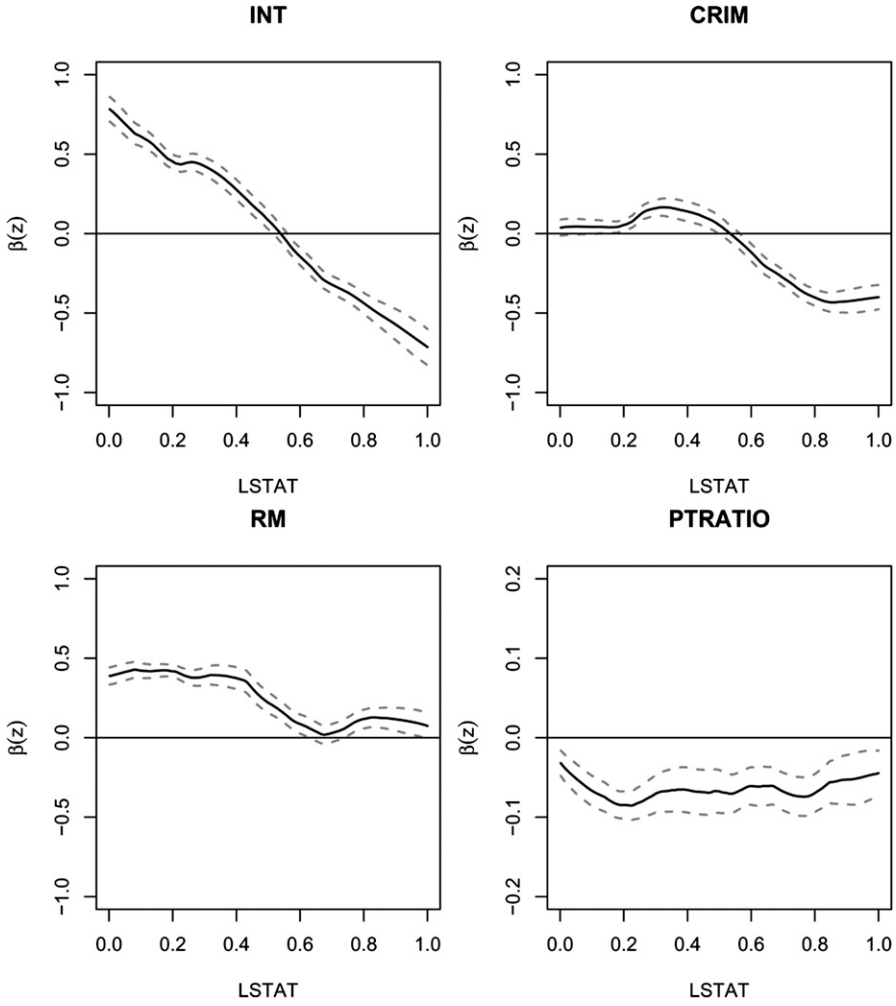

Figure 1. The KLASSO estimates of the relevant coefficients.

5. CONCLUDING REMARKS

We propose in this article a KLASSO method for a shrinkage estimation of the varying coefficient model. The proposed method is able to do variable selection and nonparametric estimation simultaneously. Our preliminary experience seems rather encouraging. To conclude the article, we would like to discuss some interesting topics for future study. First, our proposal is based on LASSO method due to its simplicity. Similar ideas can be extended to other useful shrinkage methods, such as nonnegative garrotte, bridge regression, and SCAD. Second, shrinkage estimation of a semiparametric model inevitably involves two regularization parameters (i.e., smoothing parameter and shrinkage parameter). In our current work, we follow the method of Fan and Li (2004) and select those two parameters separately. Such a simplification makes

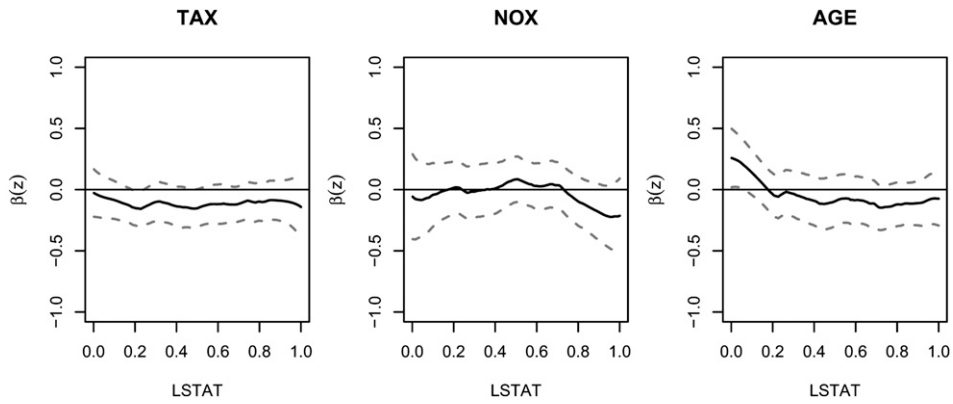

Figure 2. The unpenalized estimates of the irrelevant coefficients.

our method computationally feasible, though possibly not optimal. How to jointly tune both regularization parameters within certain computation complexity constraint is another interesting topic open for discussion.

APPENDIX A: TWO USEFUL LEMMAS

The following two lemmas study the asymptotic properties of $\hat{B}_\lambda$, i.e., a discretized version of $\hat{\beta}_\lambda(z)$, and are important for the subsequent proof of the theorems.

Lemma A.1. If (C1)–(C6), $h \propto n^{-1/5}$, and $n^{11/10}a_n \to 0$, then we must have

$$n^{-1}\sum_{t=1}^{n}\left|\left|\hat{\beta}_\lambda(Z_t) - \beta_0(Z_t)\right|\right|^2 = O_p(n^{-4/5}).$$

Proof. For an arbitrary matrix $A = (a_{ij})$, we define its L_2 norm as $\| A \|^2 = \sum a_{ij}^2$. We use $u = (u_{tj}) \in \mathbb{R}^{n \times d}$ to denote an arbitrary $n \times d$ matrix with rows ${u_1}^\top, \ldots, {u_n}^\top$ and columns $v_1, \cdots, v_d$. Under the theorem condition $h \propto n^{-1/5}$, we know that $(nh)^{-1} = n^{-4/5}$. Let $B_0 = \{\beta_0(Z_1), \ldots, \beta_0(Z_n)\}^\top \in \mathbb{R}^{n\times d}$. Thus, by Fan and Li (2001), it suffices to show that for any small probability $\epsilon > 0$, we can always find a constant $C > 0$, such that

$$\liminf_n P\left(\inf_{n^{-1}\|u\|^2=C} Q_\lambda(B_0 + (nh)^{-1/2}u) > Q_\lambda(B_0)\right) = 1 - \epsilon. \tag{A.1}$$

By definition of $Q_\lambda(B)$, we have

$$\begin{aligned} &hn^{-1}\left\{Q_\lambda(B_0 + (nh)^{-1/2}u) - Q_\lambda(B_0)\right\} \\ &= hn^{-1}\sum_{t=1}^{n}\sum_{i=1}^{n}(Y_i - X_i^\top\left\{\beta_0(Z_t) + (nh)^{-1/2}u_t\right\})^2 K_h(Z_t - Z_i) \\ &\quad -\frac{h}{n}\sum_{t=1}^{n}\sum_{i=1}^{n}(Y_i - X_i^\top\beta_0(Z_t))^2 K_h(Z_t - Z_i) \\ &\quad +\frac{h}{n}\sum_{j=1}^{d}\lambda_j(\| b_{0j} + (nh)^{-1/2}v_j \| - \| b_{0j} \|) \doteq R_1, \end{aligned} \tag{A.2}$$

where b_{0j} stands for the jth column of B_0. By simple algebraic calculation and the fact that $\|b_{0j}\| = 0$ for any $j > d_0$, we have

$$\begin{aligned} R_1 &= n^{-1}\sum_{t=1}^{n}\left\{u_t^\top\hat{\Sigma}(Z_t)u_t - 2u_t^\top\hat{e}_t\right\} \\ &\quad + hn^{-1}\sum_{j=1}^{d_0}\lambda_j(\| b_{0j} + (nh)^{-1/2}v_j \| - \| b_{0j} \|) \\ &\quad + hn^{-1}\sum_{j=d_0+1}^{d}\lambda_j \| (nh)^{-1/2}v_j \| \\ &\geq n^{-1}\sum_{t=1}^{n}\left\{u_t^\top\hat{\Sigma}(Z_t)u_t - 2u_t^\top\hat{e}_t\right\} \\ &\quad + hn^{-1}\sum_{j=1}^{d_0}\lambda_j(\| b_{0j} + (nh)^{-1/2}v_j \| - \| b_{0j} \|) \doteq R_2, \end{aligned}$$

where $\hat{\Sigma}(Z_t) = n^{-1}\sum_i X_i X_i^\top K_h(Z_t - Z_i)$ and $\hat{e}_t = n^{-1/2}h^{1/2}\sum_{i=1}^{n} X_i\{X_i^\top[\beta(Z_t) - \beta(Z_i)] + e_i\}K_h(Z_t - Z_i)$. Let $\hat{\lambda}_t^{\min}$ be the smallest eigenvalue of $\hat{\Sigma}(Z_t)$, $\hat{\lambda}_{\min} = \min\{\hat{\lambda}_t^{\min}, t = 1, \ldots, n\}$ and $\hat{e} = (\hat{e}_1, \ldots, \hat{e}_n)^\top \in \mathbb{R}^{n\times d}$. We have

$$\begin{aligned} R_2 &\geq n^{-1}\sum_{t=1}^{n}\{\| u_t \|^2 \hat{\lambda}_t^{\min} - 2\| u_t \| \cdot \| \hat{e}_t \|\} \\ &\quad - n^{-3/2}h^{1/2}\sum_{j=1}^{d_0}\lambda_j \| v_j \| \\ &\geq \hat{\lambda}_{\min}\left\{n^{-1}\sum \| u_t \|^2\right\} - n^{-1}\left(\sum_{t=1}^{n} 2\| u_t \| \cdot \| \hat{e}_t \|\right) \\ &\quad - n^{-3/2}h^{1/2}\sum_{j=1}^{d_0}\lambda_j \| v_j \| \\ &\geq \hat{\lambda}_{\min}\{n^{-1} \| u \|^2\} - 2\left(n^{-1} \| u \|^2\right)^{1/2}\left(n^{-1} \| \hat{e} \|^2\right)^{1/2} \\ &\quad - n^{-3/2}h^{1/2}\sum_{j=1}^{d_0}\lambda_j \| v_j \| \doteq R_3. \end{aligned}$$

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

$$\begin{aligned} R_3 &= \hat{\lambda}_{\min} \times C^2 - 2C \times (n^{-1} \| \hat{e} \|^2)^{1/2} \\ &\quad - n^{-3/2}h^{1/2}\sum_{j=1}^{d_0}\lambda_j \| v_j \| \\ &\geq \hat{\lambda}_{\min} \times C^2 - 2C \times (n^{-1} \| \hat{e} \|^2)^{1/2} \\ &\quad - n^{-3/2}h^{1/2}a_n\sum_{j=1}^{d_0} \| v_j \| \\ &\geq \hat{\lambda}_{\min} \times C^2 - 2C \times (n^{-1} \| \hat{e} \|^2)^{1/2} \\ &\quad - n^{-1}h^{1/2}a_n\left(n^{-1}\sum_{j=1}^{d} \| v_j \|^2\right)^{1/2} \\ &= \hat{\lambda}_{\min} \times C^2 - 2C \times (n^{-1} \| \hat{e} \|^2)^{1/2} \\ &\quad - n^{-1}h^{1/2}a_n C. \end{aligned} \tag{A.3}$$

As we will show in Appendix C, we have

$$n^{-1} \| \hat{e} \|^2 = O_p(1) \tag{A.4}$$

and

$$P(\hat{\lambda}_{\min} \to \lambda_0^{\min}) \to 1, \tag{A.5}$$

where $\lambda_0^{\min} = \inf_{z\in[0,1]}\lambda_{\min}(f(z)\Omega(z)), \lambda_{\min}(A)$ stands for the minimal eigenvalue of an arbitrary positive definite matrix A. By assumptions (C2) and Equation (A.18) in Appendix C, we have $\lambda_0^{\min} > 0$. Consequently, the last term in (A.3) is dominated by the first two terms because in the last term $nh^{-1/2}a_n \propto n^{11/10}a_n \to 0$. Last, we note that the first term in (A.3) is a quadratic function in C while the second term is linear in C. By (A.4) and (A.5) we know then as long as C is sufficiently large, the right of (A.3) is guaranteed to be positive with probability arbitrarily close to 1. This proves (A.1). The proof is completed.

Lemma A.2. Assume (C1)–(C6), $h \propto n^{-1/5}$, $n^{11/10}a_n \to 0$, and $n^{11/10}b_n \to \infty$, then we must have $P(\| \hat{b}_{\lambda,j} \| = 0) \to 1$ for any $d_0 < j \leq d$.

Proof. We only need to prove that $P(\| \hat{b}_{\lambda,j} \| = 0) \to 1$ with $j = d$. The proofs for $d_0 < j < d$ are similar. If the claim is not true (i.e., $\| \hat{b}_{\lambda,d} \| \neq 0$), then it must be the solution of the following normal equation

$$0 = \frac{\partial Q_\lambda(B)}{\partial b_d}\bigg|_{B=\hat{B}_\lambda} = \alpha_1 + \alpha_2, \tag{A.6}$$

where $\alpha_1 = \lambda_j b_d/\|b_d\|$, α_2 is a n-dimensional vector with its tth component given by $\alpha_{2t} = -2\sum_{i=1}^n X_{id}\{Y_i - X_i^\top \hat{\beta}_\lambda(Z_t)\} K_h(Z_i - Z_t)$. By standard argument of kernel smoothing given in Appendix C, we have

$$\| \alpha_2 \| = \sqrt{\alpha_{21}^2 + \alpha_{22}^2 + \cdots + \alpha_{2n}^2} = O_p(nh^{-1/2}). \tag{A.7}$$

On the other hand, under the theorem condition, we know that $(nh^{-1/2})\|\alpha_1\| = (nh^{-1/2})\,\lambda_j \geq (nh^{-1/2})b_n \propto n^{11/10}b_n \to \infty$. This implies that $P(\|\alpha_1\| > \|\alpha_2\|) \to 1$. Consequently, we know that, with probability tending to one, the normal Equation (A.6) cannot hold. This implies that $\hat{b}_{\lambda,j}$ must be located at the place where the objective function $Q_\lambda(B)$ is not differentiable. Since the only place where $Q_\lambda(B)$ is not differentiable for b_d is the origin, we know immediately that $P(\hat{b}_{\lambda,j} = 0) \to 1$. This completes the proof.

APPENDIX B: PROOFS OF THEOREMS

B.1 Proof of Theorem 1

By Lemma A.2 and Hunter and Li (2005), we know that $\| \hat{b}_{\lambda,j}^{(m)} \| \to \| \hat{b}_{\lambda,j} \|$ for every $1 \leq j \leq d$. Then, as $m \to \infty$, we must have $\| \hat{b}_{\lambda,j}^{(m)} \|$ converge to a positive number for every $j \leq d_0$, while $\| \hat{b}_{\lambda,j}^{(m)} \|$ converges to 0 for every $j > d_0$. Consequently, we define $D_{aa}^{(m)}$ as the upper $d_0 \times d_0$ diagonal submatrix of $n^{-1}D^{(m)}$ and $D_{bb}^{(m)}$ as the lower $(d - d_0) \times (d - d_0)$ diagonal submatrix of $n^{-1}D^{(m)}$. By the definition of $D^{(m)}$, every diagonal component of $D_{aa}^{(m)}$ must converge to some finite number, whereas that of $D_{bb}^{(m)}$ must diverge to infinity (as $m \to \infty$). We next study how this will affect the dynamics of $\hat{\beta}_\lambda^{(m)}(z)$ as $m \to \infty$.

For convenience, we follow (7) and rewrite $\hat{\beta}_\lambda(z) = \{\mathcal{M}(z) + n^{-1}D^{(m)}\}^{-1}\mathcal{N}(z)$, where $\mathcal{M}(z)$ is a 2×2 block matrix given by $\{\mathcal{M}_{aa}(z), \mathcal{M}_{ab}(z); \mathcal{M}_{ba}(z), \mathcal{M}_{bb}(z)$ and $\mathcal{N}(z) = \left\{\mathcal{N}_a^\top(z), \mathcal{N}_b^\top(z)\right\}^\top \in \mathbb{R}^d$, with $\mathcal{M}_{aa}(z) = n^{-1}\sum_{i=1}^n X_{ia} X_{ia}^\top K_h(Z_i - z)$, $\mathcal{M}_{bb}(z) = n^{-1}\sum_{i=1}^n X_{ib}X_{ib}^\top K_h(Z_i - z)$, $\mathcal{M}_{ab}(z) = n^{-1}\sum_{i=1}^n X_{ia}X_{ib}^\top K_h(Z_i - z) = \mathcal{M}_{ba}^{(m)\top}$, $\mathcal{N}_a(z) = n^{-1}\sum X_{ia}Y_i K_h(z - Z_i)$, and $\mathcal{N}_b(z) = n^{-1}\sum X_{ib}Y_i K_h(z - Z_i)$. We then write $\{\mathcal{M}(z) + D^{(m)}\}^{-1}$ as a 2×2 matrix, given by $\left\{\Omega_{aa}^{(m)}(z), \Omega_{ab}^{(m)}(z); \Omega_{ba}^{(m)}(z), \Omega_{bb}^{(m)}(z)\right\}$, where

$$\Omega_{aa}^{(m)}(z) = \left(\mathcal{M}_{aa}(z) + D_{aa}^{(m)} - \mathcal{M}_{ab}(z)\left\{\mathcal{M}_{bb} + D_{bb}^{(m)}\right\}^{-1}\mathcal{M}_{ba}\right)^{-1},$$

$$\Omega_{ab}^{(m)}(z) = -\left\{\mathcal{M}_{aa}(z) + D_{aa}^{(m)}\right\}^{-1}\mathcal{M}_{ab}(z) \times \left(\mathcal{M}_{aa}(z) + D_{aa}^{(m)} - \mathcal{M}_{ab}(z)\left\{\mathcal{M}_{bb} + D_{bb}^{(m)}\right\}^{-1}\mathcal{M}_{ba}\right)^{-1},$$

$$\Omega_{ba}^{(m)}(z) = -\left\{\mathcal{M}_{bb}(z) + D_{bb}^{(m)}\right\}^{-1}\mathcal{M}_{ba}(z) \times \left(\mathcal{M}_{bb}(z) + D_{bb}^{(m)} - \mathcal{M}_{ba}(z)\left\{\mathcal{M}_{aa} + D_{aa}^{(m)}\right\}^{-1}\mathcal{M}_{ab}\right)^{-1},$$

$$\Omega_{bb}^{(m)}(z) = \left(\mathcal{M}_{bb}(z) + D_{bb}^{(m)} - \mathcal{M}_{ba}(z)\left\{\mathcal{M}_{aa} + D_{aa}^{(m)}\right\}^{-1}\mathcal{M}_{ab}\right)^{-1}.$$

Recall that $D_{aa}^{(m)}$ is a diagonal matrix with every diagonal component converging to a positive number (as $m \to \infty$), whereas $D_{aa}^{(m)}$ is a diagonal matrix with every diagonal component diverging to infinity. Thus, we have every component of $\Omega_{ba}^{(m)}(z)$ and $\Omega_{bb}^{(m)}(z)$ converging to 0 uniformly on [0, 1] as m $\to \infty$. Note that

$$\beta_{\lambda,b}^{(m+1)} = \Omega_{ba}^{(m)}(z)\mathcal{N}_a(z) + \Omega_{bb}^{(m)}(z)\mathcal{N}_b(z),$$

and $\mathcal{N}_a(z)$ and $\mathcal{N}_b(z)$ are uniformly bounded. This proves that $\hat{\beta}_{\lambda,j}^{(m)}(z) \to 0$ as $m \to \infty$ uniformly on $z \in [0, 1]$ and $d_0 < j \leq d$. Thus, we must have $\sup \| \hat{\beta}_{\lambda,j}(z) \| = 0$ for every $d_0 < j \leq d$. This completes the entire proof.

B.2 Proof of Theorem 2

By Lemma A.2, we know immediately that $\hat{\beta}_{b,\lambda} = 0$ with probability tending to one. Consequently, we know that $\hat{\beta}_{a,\lambda}(z)$ must be the solution of the following normal equation

$$\frac{1}{n}\sum_{i=1}^n X_{ia}(Y_i - X_{ia}^\top \hat{\beta}_{a,\lambda})K_h(Z_i - z) + n^{-1}\sum_{j=1}^{d_0}\lambda_j \frac{\hat{b}_{\lambda,j}}{\| \hat{b}_{\lambda,j} \|} = 0,$$

which implies that $\hat{\beta}_{a,\lambda}$ must be of the form

$$\hat{\beta}_{a,\lambda}(z) = \left\{\hat{\Sigma}(z)\right\}^{-1} \times \left\{\frac{1}{n}\sum_{i=1}^n X_{ia}Y_i K_h(Z_i - z) + \frac{1}{n}\sum_{j=1}^{d_0}\lambda_j \frac{\hat{b}_{j,\lambda}}{\| \hat{b}_{j,\lambda} \|}\right\},$$

where $\hat{\Sigma}(z) = n^{-1}\sum_{i=1}^n K_h(z - Z_i)X_i X_i^\top$. Comparing with the oracle estimator (8), we know that

$$\begin{aligned}
&\max_{z\in[0,1]} \| \hat{\beta}_{a,\lambda}(z) - \hat{\beta}_{ora}(z) \| \\
&\quad = \max_{z\in[0,1]} \left\| \hat{\Sigma}(z)^{-1}\left(\frac{1}{n}\sum_{j=1}^{d_0}\lambda_j \frac{\hat{b}_{j,\lambda}}{\| \hat{b}_{j,\lambda} \|}\right)\right\| \\
&\leq \hat{\lambda}_{\max}\left\| \left(\frac{1}{n}\sum_{j=1}^{d_0}\lambda_j \frac{\hat{b}_{j,\lambda}}{\| \hat{b}_{j,\lambda} \|}\right)\right\| \leq \frac{\hat{\lambda}_{\max}}{n}\sum_{j=1}^{d_0}\lambda_j \leq \frac{\hat{\lambda}_{\max} d_0 a_n}{n} \\
&\quad = o_p(n^{-21/10}),
\end{aligned}$$

where $\hat{\lambda}_{\max} = \sup_z \lambda_{\max}\{f(z)\Omega(z)\}$ with $\lambda_{\max}(A)$ stands for the maximal eigenvalue of an arbitrary positive definite matrix A. Note that $n^{-21/10} = o(n^{-2/5})$, thus the theorem conclusion is correct. This completes the proof.

B.3 Proof of Theorem 3

For an arbitrary model S, we say it is underfitted if it misses at least one variable with nonzero coefficient (i.e., $\mathcal{S} \supset \mathcal{S}_T$); it is overfitted if $\mathcal{S}$ covers all relevant variables but also includes at least one redundant predictor (i.e., $\mathcal{S} \supset \mathcal{S}_T$ but $\mathcal{S} \neq \mathcal{S}_T$). Then, according to whether the model S_λ is underfitted, correctly fitted, or overfitted, we can create three mutually exclusive sets $\mathbb{R}_+ = \{\lambda \in \mathbb{R}^{\mathrm{d}}: \mathcal{S}_\lambda \supset \mathcal{S}_T, \mathcal{S}_\lambda \neq \mathcal{S}_T\}$, $\mathbb{R}_0 = \{\lambda \in \mathbb{R}^{\mathrm{d}}: \mathcal{S}_\lambda = \mathcal{S}_T\}$, $\mathbb{R}_- = \left\{\lambda \in \mathbb{R}^d : \mathcal{S}_\lambda \supset \mathcal{S}_T\right\}$. Furthermore, following the idea of Wang and Leng (2007), we define a reference tuning parameter sequence λ_n according to (10) with $\lambda_0 = n^{-3/2}\log(n)$. It follows immediately that such a tuning parameter sequence satisfies the technical conditions as specified in (9). Consequently, we know that $P(\mathcal{S}_{\lambda n} = \mathcal{S}_T) \to 1$. Then, the theorem can be proved by comparing BIC_λ and $\mathrm{BIC}_{\lambda \mathrm{n}}$. We consider two cases separately.

B.3.1. Case 1—Underfitted model. For arbitrary $\lambda \in \mathbb{R}_-$ (i.e., $\mathcal{S}_\lambda / \not\supset \mathcal{S}_T$, an underfitted model), we then must have $\sum_{i=1}^n \{Y_i - X_i^{\top\top}\tilde{\beta}(Z_t)\}X_i K_h(Z_t - Z_i) = 0$. Thus,

$$\begin{aligned} n^{-2}\mathrm{RSS}_\lambda &= n^{-2}\sum_{t=1}^n\sum_{i=1}^n \{Y_i - X_i^\top\tilde{\beta}(Z_t)\}^2 K_h(Z_t - Z_i) \\ &\quad + n^{-1}\sum_{t=1}^n \{\tilde{\beta}(Z_t) - \hat{\beta}_\lambda(Z_t)\}^\top \hat{\Sigma}(Z_t)\{\tilde{\beta}(Z_t) \\ &\quad - \hat{\beta}_\lambda(Z_t)\} \doteq RSS_F + R_{2\lambda}. \end{aligned} \tag{A.8}$$

By (A.19) in Appendix C, we have $\mathrm{RSS}_F \to_p \sigma^2 = E(e_i^2)$. Because $\lambda \in \mathbb{R}_-$, without loss of generality, we assume in its estimator $\hat{\beta}_\lambda(z) = \{\hat{\beta}_{1,\lambda}(z), \ldots, \hat{\beta}_{d,\lambda}(z)\}^\top$, the first term is selected as irrelevant, i.e., $\hat{\beta}_{1,\lambda}(z) \doteq 0$. By (A.5) we have

$$R_{2\lambda} \geq \hat{\lambda}_{\min}\left\{\frac{1}{n}\sum_{t=1}^n \|\tilde{\beta}(Z_t) - \hat{\beta}_\lambda(Z_t)\|^2\right\},$$

$$\begin{aligned} \sum_{t=1}^n \|\tilde{\beta}_j(Z_t) - \hat{\beta}_{j,\lambda}(Z_t)\|^2 &\geq \hat{\lambda}_{\min}\frac{1}{n}\sum_{t=1}^n \|\tilde{\beta}_1(Z_t) - \hat{\beta}_{1,\lambda}(Z_t)\|^2 \\ &\geq \hat{\lambda}_{\min}\frac{1}{n}\sum_{t=1}^n \|\tilde{\beta}_1(Z_t)\|^2 \to \\ &\quad \lambda_0^{\min} E\{\beta_1^2(Z_t)\}, \end{aligned}$$

in probability. Equation (A.19) in Appendix C was used for the last equation earlier. Thus, we have in probability

$$n^{-2}\mathrm{RSS}_\lambda \geq \sigma^2 + \lambda_0^{\min} E\{\beta_1^2(Z_t)\}. \tag{A.9}$$

It is remarkable that all the inequalities involved previously hold uniformly over all $\lambda \in \mathbb{R}_-$. We can also do similar decomposition for the reference tuning parameter λ_n, i.e., $n^{-2}\mathrm{RSS}_{\lambda n} = \mathrm{RSS}_F + R_{2\lambda n}$, where

$$\begin{aligned} R_{2\lambda_n} &= \frac{1}{n}\sum_{t=1}^n \big(\tilde{\beta}(Z_t) - \hat{\beta}_{\lambda_n}(Z_t)\big)^\top \hat{\Sigma}(Z_t)\big(\tilde{\beta}(Z_t) \\ &\quad - \hat{\beta}_{\lambda_n}(Z_t)\big) \leq \frac{\hat{\lambda}_{\max}}{n}\sum_{t=1}^n \|\tilde{\beta}(Z_t) - \hat{\beta}_{\lambda_n}(Z_t)\|^2 \leq \hat{\lambda}_{\max} \\ &\quad \times \left(\frac{1}{n}\sum_{t=1}^n \| \tilde{\beta}(Z_t) - \beta_0(Z_t) \|^2 + \frac{1}{n}\| \tilde{\beta}_{\lambda_n}(Z_t) - \beta_0(Z_t) \|^2\right). \end{aligned} \tag{A.10}$$

By Lemma A.1, both $\tilde{\beta}(Z_t)$ and $\hat{\beta}_{\lambda_n}(Z_t)$ are consistent estimates for $\beta_0(Z_t)$. Hence, the right of (A.10) is a $o_p(1)$. Consequently, we know that in probability $\mathrm{RSS}_{\lambda n} \to \sigma^2$. This together with (A.9) and the definition of BIC_λ suggest that

$$P(\inf_{\lambda \in \mathbb{R}^-} \mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}) \to 1. \tag{A.11}$$

B.3.2. Case 2—Overfitted model. Consider an arbitrary $\lambda \in \mathbb{R}_+$ (i.e., $\mathcal{S}_\lambda \supset \mathcal{S}_T$ but $\mathcal{S}_\lambda \neq \mathcal{S}_T$). Define $\tilde{\sigma}^2 = n^{-1}R_1$, where R_1 is defined in (A.8). Recall $\hat{B}_\lambda$ automatically determines a model S_λ. Under such a model S_λ, we can define another unpenalized estimate $\tilde{B}_\lambda$ as

$$\tilde{B}_\lambda = \mathrm{argmin}_{\{\|\beta_j\|=0, \forall j \in \mathcal{S}_\lambda\}}\sum_{t=1}^n\sum_{i=1}^n \{Y_i - X_i^\top\beta(Z_t)\}^2 K_h(Z_t - Z_i).$$

In other words, $\tilde{B}_{\mathcal{S}_\lambda} = (\tilde{\beta}_{\mathcal{S}_\lambda}(Z_1), \ldots, \tilde{\beta}_{\mathcal{S}_\lambda}(Z_n))^\top$ is the unpenalized estimator under the model determined by $\tilde{B}_\lambda$. By definition, we know immediately $\mathrm{RSS}_\lambda \geq \mathrm{RSS}_{\mathcal{S}\lambda}$, where

$$\mathrm{RSS}_\lambda = n^{-2}\sum_{t=1}^n\sum_{i=1}^n \{Y_i - X_i^\top\tilde{\beta}_{\mathcal{S}_\lambda}(Z_t)\}^2 K_h(Z_i - Z_t).$$

It follows that

$$\begin{aligned} \log \mathrm{RSS}_\lambda - \log \mathrm{RSS}_F &\geq \log \mathrm{RSS}_{\mathcal{S}_\lambda} - \log \mathrm{RSS}_F = \log\left(\frac{n^{-2}\mathrm{RSS}_{\mathcal{S}_\lambda}}{\tilde{\sigma}^2}\right) \\ &= \log\left(\frac{\tilde{\sigma}^2}{\tilde{\sigma}^2} + n^{-1}\tilde{\sigma}^{-2}\sum_{t=1}^n \{\tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}(Z_t)\}^\top \hat{\Sigma}(Z_t)\{\tilde{\beta}(Z_t) \right. \\ &\quad \left. - \tilde{\beta}_{\mathcal{S}_\lambda}(Z_t)\}\right) \\ &\geq -n^{-1}\tilde{\sigma}^{-2}\sum_{t=1}^n \{\tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}(Z_t)\}^\top \hat{\Sigma}(Z_t)\{\tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}(Z_t)\} \\ &\geq -\frac{\hat{\lambda}_{\min}}{\tilde{\sigma}^2}\left(n^{-1}\sum_{t=1}^n \|\tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}(Z_t)\|^2\right) = -|O_p(n^{-4/5})|, \end{aligned} \tag{A.12}$$

where the last equality is because

$$\frac{\| \tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}(Z_t) \|^2}{n} \leq \frac{\| \tilde{\beta}_{\mathcal{S}_\lambda}(Z_t) - \beta_0(Z_t) \|^2}{n} + \frac{\| \tilde{\beta}(Z_t) - \beta_0(Z_t) \|^2}{n} = O_p(n^{-4/5})$$

for any $\mathcal{S} \supset \mathcal{S}_T$. Similarly, we can prove that

$$\log \mathrm{RSS}_{\lambda_n} - \log \mathrm{RSS}_F = O_p(n^{-4/5}). \tag{A.13}$$

Combining the results of (A.12) and (A.13) we know that $\inf_{\lambda \in \mathbb{R}_+} \log \mathrm{RSS}_\lambda - \log \mathrm{RSS}_{\lambda_n} \geq -|O_p(n^{-4/5})|$. Consequently, it follows that

$$\begin{aligned} \inf_{\lambda \in \mathbb{R}_+} \mathrm{BIC}_\lambda - \mathrm{BIC}_{\lambda_n} &= \left(\inf_{\lambda \in \mathbb{R}_+} \log \mathrm{RSS}_{\lambda_n} - \log \mathrm{RSS}_F\right) + \\ &\quad (df_\lambda - df_{\lambda_n}) \times \frac{\log(nh)}{n^{4/5}} \\ &\geq -|O_p(n^{-4/5})| + (df_\lambda - df_{\lambda_n}) \times \frac{\log(nh)}{n^{4/5}} \\ &\geq -|O_p(n^{-4/5})| + \frac{\log(nh)}{n^{4/5}}, \end{aligned} \tag{A.14}$$

where the last equality is due to the following three facts. First, because the reference sequence λ_n satisfies (9), by Lemma A.2 we know that $P(df_{\lambda_n} = d_0) \to 1$. Second, because $\lambda \in \mathbb{R}_+$ and $\mathcal{S}_\lambda$ is an overfitted model, we must have $P(df_\lambda \geq d_0 + 1) \to 1$. Last, under the assumption $h \propto n^{-1/5}$, we have $\log(nh) \propto \log n \to \infty$. Consequently, with probability tending to one, we have $df_\lambda - df_{\lambda_n} \geq 1$. It is clear that, with probability tending to one, the right of (A.14) is guaranteed to be positive. Consequently,

$$P(\inf_{\lambda \in \mathbb{R}_+} \mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}) \to 1. \tag{A.15}$$

Combining the results from (A.11) and (A.15), we have

$$P(\inf_{\lambda \in \mathbb{R}_- \cup \mathbb{R}_+} \mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}) \to 1. \tag{A.16}$$

The Equation (A.16) implies that, with probability tending to one, the tuning parameters failing to identify the true model cannot be selected by our BIC criterion, because it is at least not as favorable as our reference sequence λ_n. Consequently,

we know that $P(\hat{\lambda}=T) \to 1$ (Wang and Leng 2007). This completes the proof.

APPENDIX C: SOME TECHNICAL DETAILS

Lemma A.3. Suppose (ξ_i, Z_i), $i = 1, \ldots, n$ are iid random vectors, where $\xi_i s$ are scalar random variables. Assume further that $E|\xi_i|^s < \infty$ and $\sup_z \int |y|^s f(z, v)dv < \infty$, where f denotes the joint density of (ξ_1, Z_1). Let K be a bounded positive function with bounded support, satisfying the Lipschitz condition. Then

$$\sup_{z\in[0,1]} \left| n^{-1} \sum_{i=1}^{n} [K_h(Z_i - z)\xi_i - E\{K_h(Z_i - z)\xi_i\}] \right| = O_p\left\{ \left(\frac{\log(1/h)}{nh} \right)^{1/2} \right\}$$

provided that $n^{2\delta-1}h \to \infty$ for some $\delta < 1 - s^{-1}$.

The proof of the Lemma can be found in Mack and Silverman (1982) or Fan and Zhang (2000a). If $g(z) = E(\xi_i|Z_i = z)$ has bounded derivative, then $E\{K_h(Z_i - z)\xi_i\} = f(z)g(z) + O(h)$. It follows that

$$\sup_{x\in[0,1]} |n^{-1} \sum_{i=1}^{n} [K_h(Z_i - z)\xi_i - f(z)g(z)]| = O_p\left\{ h + \left(\frac{\log(1/h)}{nh} \right)^{1/2} \right\}. \tag{A.17}$$

By assumption (C3), we have

$$\hat{\Sigma}(z) - f(z)\Omega(z) = O_p\left\{ h + \left(\frac{\log(1/h)}{nh} \right)^{1/2} \right\} \tag{A.18}$$

uniformly for all $z \in [0, 1]$. For the varying coefficient model, the estimator is

$$\tilde{\beta}(z) = \left\{ \hat{\Sigma}(z) \right\}^{-1} \sum_{i=1}^{n} K_h(z - Z_i)X_i Y_i = \beta(z) + O_p\left\{ h + \left(\frac{\log(1/h)}{nh} \right)^{1/2} \right\}, \tag{A.19}$$

uniformly for $z \in [0, 1]$.

Proof of (A.4). Let $e_{n,1,t} = n^{-1/2}h^{1/2} \sum_{i=1}^{n} X_i X_i^\top [\beta(Z_t) - \beta(Z_i)]K_h(Z_t - Z_i)$ and $e_{n,2,t} = n^{-1/2}h^{1/2} \sum_{i=1}^{n} X_i e_i K_h(Z_t - Z_i)$. Write

$$E \parallel e_{n,1,t} \parallel^2 = 2n^{-1}h \sum_{i>j} EQ_{i,j,t} + n^{-1}h \sum_{i=1}^{n} ER_{i,t},$$

where $Q_{i,j,t} = [\beta(Z_t) - \beta(Z_i)]^\top X_i X_i^\top X_j X_j^\top [\beta(Z_t) - \beta(Z_j)]K_h(Z_t - Z_i)K_h(Z_t - Z_j)$ and $R_{i,t} = \|X_i X_i^\top [\beta(Z_t) - \beta(Z_i)]\|^2 K_h^2(Z_t - Z_i)$. By (C6), we have

$$\begin{aligned} Q_{i,j,t} &= \{\beta'(Z_t)\}^\top X_i X_i^\top X_j X_j^\top \beta'(Z_t)(Z_i - Z_t)(Z_j - Z_t) \\ &\quad K_h(Z_t - Z_i)K_h(Z_t - Z_j) \\ &\quad + C \parallel X_i \parallel^2 \parallel X_j \parallel^2 (Z_i - Z_t)^2 (Z_j - Z_t)^2 K_h(Z_t - Z_i) \\ &\quad K_h(Z_t - Z_j) \\ &\doteq Q_{i,j,t,1} + Q_{i,j,t,2}, \end{aligned}$$

where C is a constant. If $i = t$ or $j = t$, then $Q_{i,j,t,1} = 0$ and $Q_{i,j,t,2} = 0$. For $i \neq t$ and $j \neq t$, we have

$$\begin{aligned} EQ_{i,j,t,1} &= E\{E(Q_{i,j,t,1}|Z_i, Z_j, Z_t)\} \\ &= \int \{\beta'(Z_t)\}^\top \Omega(Z_i)\Omega(Z_j)\beta'(Z_t)(Z_i - Z_t) \\ &\quad (Z_j - Z_t)K_h(Z_t - Z_i) \\ &\quad \times K_h(Z_t - Z_j)f(Z_t)f(Z_i)f(Z_j)dZ_t dZ_i dZ_j \\ &= h^2 \int \{\beta'(Z_t)\}^\top [\Omega(Z_t + hu)\Omega(Z_t + hv)f(Z_t + hu) \\ &\quad f(Z_t + hv)]\beta'(Z_t)\, uvK(u)K(v)dZ_t du dv. \end{aligned}$$

By (C3), we have Taylor expansion

$$\begin{aligned} &\Omega(Z_t + hu)\Omega(Z_t + hv)f(Z_t + hu)f(Z_t + hv) \\ &= \tilde{\Omega}(Z_t) + \tilde{\Omega}'_1(Z_t)hu + \tilde{\Omega}'_2(Z_t)hv + Ch^2(u^2 + v^2), \end{aligned}$$

where $\tilde{\Omega}(Z_t), \tilde{\Omega}'_1(Z_t)$ and $\tilde{\Omega}'_2(Z_t)$ are bounded continuous functions and C is a constant independent of t. Note that $K(v)$ is a symmetric function, thus

$$\int \left\{ \tilde{\Omega}(Z_t) + \tilde{\Omega}'_1(Z_t)hu + \tilde{\Omega}'_2(Z_t)hv \right\} uvK(u)K(v)du dv = 0.$$

It follows that $EQ_{i,j,t,1} = O(h^4)$ uniformly for t. Similarly, by (C3) we have $EQ_{i,j,t,2} = O(h^4)$ uniformly for t. Thus, $EQ_{i,j,t} = O(h^4)$ uniformly for t. In the same spirit, we have $ER_{i,t} = O(h)$ uniformly for t. It follows that

$$\begin{aligned} E\left\| e_{n,1,t} \right\|^2 &= 2n^{-1}hn(n-1)O(h^4) + n^{-1}hnO(h) = O(nh^5) \\ &= O(1) \end{aligned} \tag{A.20}$$

uniformly for t. Because (X_i, Z_i, e_i), $i = 1, \ldots, n$ are independent and identically distributed and $E(e_i|X_i, Z_i) = 0$ a.s., we have

$$\begin{aligned} E \parallel e_{n,2,t} \parallel^2 &= n^{-1}h \sum_{i=1}^{n} E\{\parallel X_i e_i \parallel^2 K_h^2(Z_t - Z_i)\} \\ &= n^{-1}(n-1)h^{-1}E\left\{ \parallel X_1 e_1 \parallel^2 K^2\left(\frac{Z_2 - Z_1}{h} \right) \right\} \\ &\quad + n^{-1}h^{-1}K^2(0)E\{\parallel X_1 e_1 \parallel^2\}. \end{aligned}$$

By (C2), we have f is bounded. Let $\varrho(z) = \{\parallel X_1 e_1 \parallel^2 Z_1 = z\}$. We have

$$\begin{aligned} &E\left\{ \parallel X_1 e_1 \parallel^2 K^2\left(\frac{Z_2 - Z_1}{h} \right) \right\} = E\left[E\left\{ \parallel X_1 e_1 \parallel^2 K^2 \right. \right. \\ &\left. \left. \times \left(\frac{Z_2 - Z_1}{h} \right) \right| Z_1, Z_2 \right\} \Bigg] \\ &= \int \varrho^2(Z_1)f(Z_2)f(Z_1)K^2\left(\frac{Z_2 - Z_1}{h} \right) dZ_2 dZ_1 \\ &= h \int \varrho^2(Z_1)K^2(v)f(Z_1 + hv)f(Z_1)dv dZ_1 \leq C\mu_2 hE \parallel X_1 e_1 \parallel^2, \end{aligned}$$

where $\mu_2 = \int K^2(v)dv$. Note that $(nh)^{-1} \to 0$. It follows that

$$E \parallel e_{n,2,t} \parallel^2 \leq C\mu_2 E\{\parallel X_1 e_1 \parallel^2\} + n^{-1}h^{-1}K^2(0)E\{\parallel X_1 e_1 \parallel^2\} < \infty. \tag{A.21}$$

By (A.20) and (A.21), we have $En^{-1} \sum_{t=1}^{n} \parallel e_{n,1,t} + e_{n,2,t} \parallel^2 \leq 2n^{-1}\{\sum_{t=1}^{n} E \parallel e_{n,1,t} \parallel^2 + \sum_{t=1}^{n} E \parallel e_{n,2,t} \parallel^2\} = O(1)$.

Therefore (A.4) follows from the previous equation and because $\hat{e}_t = e_{n,1,t} + e_{n,2,t}$.

Proof of (A.5). It follows from (A.18) and (C2).

Proof of (A.7). Note that, we can write

$$\begin{aligned}\alpha_{2t} &= \sum_{i=1}^{n} X_{id} e_i K_h(Z_i - Z_t) + \sum_{i=1}^{n} X_{id} X_i^\top \{\beta_0(Z_i) - \beta_0(Z_t)\} \\ &\quad K_h(Z_i - Z_t) \\ &\quad + \sum_{i=1}^{n} X_{id} X_i^\top \{\beta_0(Z_t) - \hat{\beta}_\lambda(Z_t)\} K_h(Z_i - Z_t) \doteq \alpha_{2t1} + \\ &\quad \alpha_{2t2} + \alpha_{2t3}.\end{aligned}$$

From the proof of (A.4), we have $\{\sum_{t=1}^{n} \alpha_{2t1}^2\}^{1/2} = O_p(nh^{-1/2})$ and $\{\sum_{t=1}^{n} \alpha_{2t2}^2\}^{1/2} = O_p(nh^{-1/2})$. Moreover, by (A.17), we have $n^{-1} \sum_{i=1}^{n} X_{id} X_i^\top K_h(Z_i - Z_t) = O_p(1)$, uniformly for all t. By Lemma A.1, we have

$$\begin{aligned}\left(\sum_{t=1}^{n} \alpha_{2t3}^2\right)^{1/2} &\le \left(\sum_{t=1}^{n} \left\|\beta_0(Z_t) - \hat{\beta}_\lambda(Z_t)\right\|^2 \times \right. \\ &\quad \left. \left\| \sum_{i=1}^{n} X_{id} X_i^\top K_h(Z_i - Z_t) \right\|^2 \right)^{1/2} \\ &= \left[nO_p\left\{(nh)^{-1}\right\} O_p(n^2) \right]^{1/2} = O_p(nh^{-1/2}).\end{aligned}$$

Therefore, (A.7) follows. This completes the proof.

[Received July 2008. Revised October 2008.]

REFERENCES

Breiman, L. (1995), "Better Subset Selection Using Nonnegative Garrote," *Technometrics*, 37, 373–384.

Cai, Z., Fan, J., and Li, R. (2000), "Efficient Estimation and Inferences for Varying-Coefficient Models," *Journal of the American Statistical Association*, 95, 888–902.

Chen, R., and Tsay, R. S. (1993), "Functional-Coefficient Autoregressive Models," *Journal of the American Statistical Association*, 88, 298–308.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–489.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modeling and Its Applications*, New York: Chapman and Hall.

Fan, J., and Huang, T. (2005), "Profile Likelihood Inferences on Semiparametric Varying-Coefficient Partially Linear Models," *Bernoulli*, 11, 1031–1057.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan and Li, (2004), "New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis," *Journal of the American Statistical Association*, 99, 710–723.

Fan, J., Zhang, C., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, 29, 153–193.

Fan, J., and Zhang, J. T. (2000a), "Two-Step Estimation of Functional Linear Model with Application to Longitudinal Data," *Journal of the Royal Statistical Society:* Ser. B, 62, 303–322.

Fan, J., and Zhang, J. T. (2000b), "Simultaneous Confidence Bands and Hypotheses Testing in Varying-Coefficient Models," *Scandinavian Journal of Statistics*, 27, 715–731.

Fan, J., and Zhang, W. (1999), "Statistical Estimation in Varying Coefficient Models," *The Annals of Statistics*, 27, 1491–1518.

Fu, W. J. (1998), "Penalized Regression: The Bridge versus the LASSO," *Journal of Computational and Graphical Statistics*, 7, 397–416.

Härdle, W., Liang, H., and Gao, J. (2000), *Partial Linear Models*, Heidelberg: Springer Physica-Verlag.

Hastie, T. J., and Tibshirani, R. J. (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society:* Ser. B, 55, 757–796.

Huang, J. Z., Wu, C. O., and Zhou, L. (2002), "Varying-Coefficient Models and Basis Function Approximations for the Analysis of Repeated Measurements," *Biometrika*, 89, 111–128.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004), "Polynomial Spline Estimation and Inference for Varying Coefficient Models with Longitudinal Data," *Statistica Sinica*, 14, 763–788.

Hunter, D. R., and Li, R. (2005), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, 33, 1617–1642.

Janson, S. (1987), "Maximal Spacing in Several Dimensions," *Annals of Probability*, 15, 274–280.

Knight, K., and Fu, W. (2000), "Asymptotics for LASSO-type Estimators," *The Annals of Statistics*, 28, 1356–1378.

Li, R., and Liang, H. (2008), "Variable Selection in Semiparametric Regression Model," *The Annals of Statistics*, 36, 261–286.

Mack, Y. P., and Silverman, B. W. (1982), "Weak and Strong Uniform Consistency of Kernel Regression Estimates," *Z. Wahrsch. verw Gebiete*, 61, 405–415.

Park, M. Y., and Hastie, T. (2007), "An $L1$ Regularization-path Algorithm for Generalized Linear Models," *Journal of the Royal Statistical Society:* Ser. B, 69, 659–667.

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society:* Ser. B, 58, 267–288.

Wang, H., and Leng, C. (2007), "Unified LASSO Estimation via Least Squares Approximation," *Journal of the American Statistical Association*, 101, 1418–1429.

Wang, H., Li, G., and Tsai, C. L. (2007a), "Regression Coefficient and Autoregressive Order Shrinkage and Selection via LASSO," *Journal of the Royal Statistical Society:* Ser. B, 69, 63–78.

Wang, H., Li, R., and Tsai, C. L. (2007b), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," *Biometrika*, 94, 553–558.

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society:* Ser. B, 68, 49–67.

Yuan, M., and Lin, Y. (2007), "On the Nonnegative Garrote Estimator," *Journal of the Royal Statistical Society:* Ser. B, 69, 143–161.

Zhang, H. H., and Lin, Y. (2003), "Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models," *The Annals of Statistics*, 34, 2272–2297.

Zhang, W., and Lee, S. Y. (2007), "Variable Bandwidth Selection in Varying Coefficient Models," *Journal of Multivariate Analysis*, 19, 116–134.

Zhang, H. H., and Lu, W. (2007), "Adaptive LASSO for Cox's Proportional Hazard Model," *Biometrika*, 94, 691–703.

Zhao, P., and Yu, B. (2004), "Boosted LASSO," Technical Report, Statistics, UC Berkeley.

Zou, H. (2006), "The Adaptive LASSO and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, 36(4), 1509–1533.