Statistica Sinica **22** (2012), 575-599
doi:http://dx.doi.org/10.5705/ss.2010.105

ADAPTIVE SEMI-VARYING COEFFICIENT MODEL SELECTION

Tao Hu[1,2] and Yingcun Xia[3]

[1]*Guizhou College of Finance and Economics,* [2]*Capital Normal University and* [3]*National University of Singapore*

Abstract: Identification of constant coefficients in a semi-varying coefficient model is an important issue (Zhang et al (2002)). We propose a novel method for this by combining local polynomial smoothing (Fan and Zhang (1999)) with shrinkage estimation (Tibshirani (1996)). Unlike the stepwise procedure (Xia, Zhang, and Tong (2004)), our method can identify the constant coefficients and estimate the model simultaneously. By imposing the adaptive LASSO penalty and starting with the Nadaraya-Watson estimator, the method can identify the constant coefficients and varying coefficients consistently, and estimate the model with oracle efficiency (Fan and Li (2001)).

Key words and phrases: BIC, kernel smoothing, LASSO, model selection, oracle property, SCAD, semi-varying coefficient model.

1. Introduction

Over the past decade, the varying-coefficient model (Hastie and Tibshirani (1993); Chen and Tsay (1993)) has gained in popularity, and has been widely used in such disciplines as finance, economics, medicine, ecology, and biology, due to its good interpretability. Let (X_i, Y_i, Z_i) be the observation collected from the ith subject $(1 \le i \le n)$, where $Y_i \in \mathbb{R}$ 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 varying coefficient model assumes that

$$Y_i = X_i^\top \beta(Z_i) + e_i, \tag{1.1}$$

where $e_i \in \mathbb{R}$ is the random noise satisfying $E(e_i|X_i, Z_i) = 0$ almost surely. For simplicity, we set $X_{i1} \equiv 1$ if there is an intercept in the model. 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 $\beta_0(z) = \{\beta_{01}(z), \ldots, \beta_{0d}(z)\}^\top \in \mathbb{R}^d$. A series of papers (Fan and Zhang (1999); Cai, Fan, and Li (2000); Fan and Zhang (2000a,b); Huang, Wu, and Zhou (2002, 2004); Fan and Huang (2005); Fan and Zhang (2008); Wang, Li, and Huang (2008); Wang and Xia (2009); Wang, Zhu, and Zhou (2009) and references therein) have considered the estimation of the model.

Zhang, Lee, and Song (2002) noticed that in practice some of the coefficients are constant rather than varying, and proposed the so-called semi-varying coefficient model. Statistically, treating constant coefficients as varying degrades estimation efficiency. Without loss of generality, we assume that coefficients for the first $d_0 \leq d$ predictors vary with z but the rest are constant and do not depend on z, i.e. $\beta_0(z) = \{\beta_{a0}^\top(z), \beta_{b0}^\top\}^\top$, $\beta_{a0}(z) \in \mathbb{R}^{d_0}$ and $\beta_{b0} \in \mathbb{R}^{d-d_0}$. Rarely can the analysts know a priori which coefficients are constant and which are varying with the index variable. Therefore, it is of great interest to develop fast and efficient methods to differentiate constant coefficients from varying ones.

The identification of coefficients can be done in a hypothesis testing framework which treats estimation and model selection separately. See for example Fan and Li (2001), Huang, Wu, and Zhou (2002), Fan and Huang (2005), and Wang, Zhu, and Zhou (2009). Alternatively, determining constant coefficients can be done in a variable selection framework. Variable selection is an important topic in modern statistical inference and has been extensively studied. In the linear regression models, many selection criteria (e.g. AIC and BIC) have been used in practice. For the varying coefficient models, Xia, Zhang, and Tong (2004) developed a cross-validation procedure for selecting constant and varying coefficients. Suppose $\mathcal{I}_l$ is any subset of $\{1, 2, \ldots, d\}$. They considered model

$$Y_i = \sum_{k \notin \mathcal{I}_l} \beta_{ak}(Z_i) X_{ik} + \sum_{k \in \mathcal{I}_l} \beta_{bk} X_{ik} + e_i.$$

For every i, they first estimated $\hat{\beta}_{bk}$ and $\hat{\beta}_{ak}(u)$ based on observations $\{(Y_j, X_j, Z_j), j \neq i\}$, say $\hat{\beta}_{bk}^{-i}$ and $\hat{\beta}_{ak}^{-i}(u)$, respectively. The cross-validation sum of squares is calculated,

$$\mathrm{CV}(\mathcal{I}_l) = n^{-1} \sum_{i=1}^{n} \Big\{ Y_i - \sum_{k \in \mathcal{I}_l} \hat{\beta}_{bk}^{-i} X_{ik} - \sum_{k \notin \mathcal{I}_l} \hat{\beta}_{ak}^{-i}(Z_i) X_{ik} \Big\}^2.$$

If $\mathrm{CV}(\mathcal{I}_{l_0}) = \min_l \mathrm{CV}(\mathcal{I}_l)$, the model with constant coefficients for variables in $\mathcal{I}_{l_0}$ is the preferred model. However, as pointed out by the authors, practical implementation of the cross-validation model selection procedure can be time-consuming when d is large, since there are $\sum_{k=0}^{d} \binom{d}{k}$ candidate models.

As computational efficiency is more desirable in many situations, various shrinkage methods have been developed; these include, but are not limited to, the nonnegative garrotte (Breiman (1995); Yuan and Lin (2007)), the LASSO (Tibshirani (1996); Zou (2006); Zhang and Lu (2007); Wang, Li, and Tsai (2007)), bridge regression (Fu (1998); Knight and Fu (2000)), the SCAD (Fan and Li (2001)), and the one-step sparse estimator (Zou and Li (2008)).

Recently, Fan and Li (2004) extended the SCAD to the partially linear model (Härdle, Liang, and Gao (2000)) with their focus on the parametric components. Wang, Chen, and Li (2007) and Wang, Li, and Huang (2008) used the SCAD method to remove variables from the varying-coefficient models. Li and Liang (2008) used SCAD to select parametric components via the generalized likelihood ratio test (Fan, Zhang, and Zhang (2001)) under the assumption that the varying and invariant coefficients are known a priori. Wang and Xia (2009) proposed a method combining the ideas of the local polynomial smoothing and the shrinkage estimation to conduct variable selection for the varying coefficient models. Leng (2009) proposed another model selection approach for varying-coefficient model.

In this paper, we develop a shrinkage method that is able to identify the constant coefficients and estimate the model simultaneously. Furthermore, given the popularity of kernel smoothing methods, it is desirable to have a shrinkage method that can work with kernel smoothing techniques in a natural way. We show that the shrinkage parameters selected by the BIC criterion can identify the constant coefficients and varying coefficients consistently, and that the resulting estimator can be as efficient as the oracle estimator (Fan and Li (2001)). The proposed method can be easily extended to many other semiparametric models where local polynomial methods are used. See for example Zhang and Lin (2003), Wang, Li, and Huang (2008), and Zou and Li (2008).

The model selection procedure in this paper is different from Wang and Xia (2009), Wang, Li, and Huang (2008), and Wei, Huang, and Li (2010) in several aspects. First, the aforementioned papers are concerned about the selection of variables. For the generalized linear or varying coefficient models specifically, their goal is to identify those covariates with zero coefficients. The purpose here is to identify the constant coefficients in a generalized semi-varying coefficient model. Since the varying coefficient model is an extension of the (constant) linear regression model, it is important to identify which coefficients remain constant and which need to be varying; see Zhang et al (2002) and Xia, Zhang, and Tong (2004). Second, and technically, penalizing a varying coefficient to zero is much easier than to a constant; see Subsection 2.1. We propose to combine local polynomial smoothing and adaptive L_1 penalization for our purpose, which as far as we know has not been investigated in the literature. Moreover, our discussion is in a framework for the generalized varying-coefficient model that includes varying-coefficient models, logistic varying-coefficient models, and Poisson varying-coefficient models as special cases (Cai, Fan, and Li (2000)), for which the methods of Wang and Xia (2009), Wang, Li, and Huang (2008), and Wei, Huang, and Li (2010) cannot be used directly. Since the objective function we optimize is a penalized likelihood function, the proof is very different from that for penalized least squares estimation (Wang and Xia (2009); Wei, Huang, and Li (2010)). Third, we further suggest a two-stage adaptive LASSO approach

to identify constant coefficients in the generalized varying coefficient models in sparse and high-dimensional settings where the number of variables can be larger than the sample size; see Section 5 for details.

This paper is organized as follows. In Section 2, we introduce our model selection method, present a computational algorithm and a method for selecting the tuning parameter. We state our theoretical results in Section 3. In Section 4, we report some simulation results and a data analysis. Possible extensions of the proposed method are discussed in Section 5. The article concludes with a brief discussion in Section 6. All technical details are left to the Appendix.

2. Methodology

2.1. Penalized least squares method

Without loss of generality, assume that $\beta_0(z) = \{\beta_{a0}^\top(z), \beta_{b0}^\top\}^\top$, $\beta_{a0}(z) \in \mathbb{R}^{d_0}$ and $\beta_{b0} \in \mathbb{R}^{d-d_0}$, and that $Z_1 \le \cdots \le Z_n$. Let $B_0 = \{\beta_0(Z_1), \ldots, \beta_0(Z_n)\}^\top$. For any $1 \le j \le d$, let $b_j = \{\beta_j(Z_2) - \beta_j(Z_1), \ldots, \beta_j(Z_n) - \beta_j(Z_{n-1})\}^\top \in \mathbb{R}^{n-1}$ and $\|b_j\| = \{\sum_{k=2}^n \{\beta_j(Z_k) - \beta_j(Z_{k-1})\}^2\}^{1/2}$. It is easy to see that if coefficient $\beta_k(z)$ is constant, then $\|b_k\| = 0$, otherwise $\|b_k\| > 0$. Then, following the grouped LASSO idea of Yuan and Lin (2006), we propose a penalized loss function

$$Q_\lambda(B) = \sum_{t=1}^n \sum_{i=1}^n \Big\{Y_i - X_i^\top \beta(Z_t)\Big\}^2 K_h(Z_t - Z_i) + \sum_{j=1}^d \lambda_j \|b_j\|, \tag{2.1}$$

where $\lambda = (\lambda_1, \ldots, \lambda_d)^\top \in \mathbb{R}^d$ is the tuning parameter, $\hat\beta_\lambda(z) = \{\hat\beta_{\lambda,1}(z), \ldots, \hat\beta_{\lambda,d}(z)\}^\top \in \mathbb{R}^d$, $B = \{\beta(Z_1), \ldots, \beta(Z_n)\}^\top$, $K(s)$ is a symmetric density function, $h > 0$ is a bandwidth and $K_h(s) = h^{-1}K(s/h)$. Our estimator of the model is

$$\Big\{\hat\beta_\lambda(Z_1), \ldots, \hat\beta_\lambda(Z_n)\Big\}^\top = \arg\min_{B \in \mathbb{R}^{n\times d}} Q_\lambda(B). \tag{2.2}$$

2.2. Local quadratic approximation

In a typical least squares regression, computational algorithms for the LASSO-type problems have been well developed. These include the shooting algorithm (Fu (1998); Yuan and Lin (2006)), local quadratic approximation (Fan and Li (2001)), least angle regression (Efron et al. (2004)), and many others (Zhao and Yu (2004); Park and Hastie (2007); Zou and Li (2008)). We describe here an easy implementation based on local quadratic approximation (Fan and Li (2001)).

Specifically, our implementation is based on an iterative algorithm with the unpenalized Nadaraya-Watson estimator as the initial estimator,

$$\hat B_\lambda^{(m)} = \Big\{\hat\beta_\lambda^{(m)}(Z_1), \ldots, \hat\beta_\lambda^{(m)}(Z_n)\Big\}$$

with $m = 0$. Let $\hat{b}_{\lambda,j}^{(m)} = \{\hat{\beta}_{\lambda,j}(Z_2) - \hat{\beta}_{\lambda,j}(Z_1), \ldots, \hat{\beta}_{\lambda,j}(Z_n) - \hat{\beta}_{\lambda,j}(Z_{n-1})\}^\top$. Then, the loss function in (2.1) can be locally approximated by (Fan and Li (2001); Hunter and Li (2005))

$$\sum_{t=1}^{n}\sum_{i=1}^{n}\left\{Y_i - X_i^\top \beta(Z_t)\right\}^2 K_h(Z_t - Z_i) + \sum_{j=1}^{d}\lambda_j \frac{\|b_j\|^2}{\|\hat{b}_{\lambda,j}^{(m)}\|}. \tag{2.3}$$

We update the estimator by the solution of $\beta(Z_t)$ that minimizes (2.3), denoted as $\hat{B}_\lambda^{(m+1)}$. It is easy to see that the minimizer has the closed form

$$\mathrm{vec}(\hat{B}_\lambda^{(m+1)}) = \left\{\mathcal{M} + D^{(m)}\right\}^{-1}\mathcal{N},$$

where $\mathrm{vec}(A)$ denotes the vectorization of matrix A,

$$\mathcal{N} = \begin{pmatrix} \sum_{i=1}^n Y_i X_{i1} K_h(Z_1 - Z_i), \ldots, \sum_{i=1}^n Y_i X_{i1} K_h(Z_n - Z_i) \\ \sum_{i=1}^n Y_i X_{i2} K_h(Z_1 - Z_i), \ldots, \sum_{i=1}^n Y_i X_{i2} K_h(Z_n - Z_i) \\ \vdots, \qquad\qquad \ldots, \qquad\qquad \vdots \\ \sum_{i=1}^n Y_i X_{id} K_h(Z_1 - Z_i), \ldots, \sum_{i=1}^n Y_i X_{id} K_h(Z_n - Z_i) \end{pmatrix}^\top \in \mathbb{R}^{nd},$$

where $\mathcal{M} = (m_{\tau,\iota})_{1\le\tau,\iota\le d}$, $m_{\tau,\iota}$ is a $n \times n$ diagonal matrix with its lth diagonal component given by $\sum_{i=1}^n X_{i\tau} X_{i\iota} K_h(Z_l - Z_i)$, and $D^{(m)}$ is blocked diagonal matrix whose jth diagonal block is given by

$$\frac{\lambda_j}{\|\hat{b}_{\lambda,j}\|}\begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & 1 & \cdots & 0 \\ \vdots & & & & \vdots \\ 0 & \cdots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{pmatrix}.$$

As $m \to \infty$, denote the limiting values of $\hat{B}_\lambda^{(m+1)}$ and $\hat{\beta}_\lambda^{(m+1)}(z)$, respectively, by $\hat{B}_\lambda$ and $\hat{\beta}_\lambda(z)$; these are our final estimators.

When n is very large, the above calculation might be time-consuming as $\mathcal{M}$ is a $nd \times nd$ matrix. One way to simplify the calculation is to use sparser grids, as one anonymous referee suggested. Suppose $Z_{[i]}, i = 1, 2, \ldots, n$ are the order statistics of Z_i. Consider griding points $Z_{[k]}, Z_{[2k]}, \ldots, Z_{[mk]}$ such that $mk \le n < (m+1)k$, and let

$$\tilde{b}_j = (\beta_j(Z_{[2k]}) - \beta_j(Z_{[k]}), \ldots, \beta_j(Z_{[mk]}) - \beta_j(Z_{[(m-1)k]}))^\top.$$

The estimation loss function in (2.1) can be replaced by

$$\tilde{Q}_\lambda(B) = \sum_{t=1}^{m}\sum_{i=1}^{n}\left\{Y_i - X_i^\top \beta(Z_{[tk]})\right\}^2 K_h(Z_{[tk]} - Z_i) + \sum_{j=1}^{d}\lambda_j \|\tilde{b}_j\|.$$

Since bandwidth is roughly the range within which the function can be treated as constant, with n samples, the number of griding points can be $O\{n/(nh)\} = O(1/h)$ asymptotically, which is $O(n^{1/5})$ if the optimal bandwidth is used. In practice, one can fix k as a small integer, for example 5. The computational burden in minimizing $\tilde{Q}_\lambda(B)$ is significantly lighter than that of (2.1). Theoretical justification of this simplification needs to be investigated, but is beyond the scope of this paper.

We can further extend the theory and methodology to the generalized varying-coefficient models (GVCM). Consider the penalized local log-likelihood estimation using the grouped LASSO penalty described in Section 2.1 with the likelihood function belonging to the exponential family (McCullagh and Nelder (1989)). A generalized varying-coefficient model (Cai, Fan, and Li (2000)) has the form

$$\eta(z, x) = g\{m(z, x)\} = x^\top \beta(z) \tag{2.4}$$

for some given link function $g(\cdot)$, where $m(z, x)$ is the mean regression function of the response variable Y given the covariates $Z = z$ and $X = x$, where $X = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$ is the d-dimensional predictor, and $Z \in [0, 1]$. The local likelihood version of the grouped LASSO loss function (2.1) is then

$$Q_\lambda^E(B) = \sum_{t=1}^{n}\sum_{i=1}^{n} -\mathcal{L}\Big\{g^{-1}\{X_i^\top \beta(Z_t)\}, Y_i\Big\}K_h(Z_t - Z_i) + \sum_{j=1}^{d}\lambda_j\|b_j\|. \tag{2.5}$$

For logistic varying-coefficient models, (2.5) becomes

$$Q_\lambda^E(B) = \sum_{t=1}^{n}\sum_{i=1}^{n}\Big\{ - Y_i\big\{X_i^\top \beta(Z_t)\big\} + \log\big\{1 + \exp\{X_i^\top \beta(Z_t)\}\big\}\Big\}K_h(Z_t - Z_i) + \sum_{j=1}^{d}\lambda_j\|b_j\|.$$

In Poisson log-linear varying-coefficient models, (2.5) can be written as

$$Q_\lambda^E(B) = \sum_{t=1}^{n}\sum_{i=1}^{n}\Big\{ - Y_i\big\{X_i^\top \beta(Z_t)\big\} + \exp\big\{X_i^\top \beta(Z_t)\big\}\Big\}K_h(Z_t - Z_i) + \sum_{j=1}^{d}\lambda_j\|b_j\|.$$

Again, denote the solution of $\beta(z)$ to (2.5) by

$$\Big\{\hat{\beta}_\lambda^E(Z_1), \ldots, \hat{\beta}_\lambda^E(Z_n)\Big\}^\top = \arg\min_{B\in\mathbb{R}^{n\times d}} Q_\lambda^E(B). \tag{2.6}$$

2.3. Tuning parameter

Let $a_n = \max\{\lambda_j : 1 \le j \le d_0\}$ and $b_n = \min\{\lambda_j : d_0 < j \le d\}$. In words, a_n and b_n are the maximal and minimal amounts of shrinkages applied to varying and constant coefficients, respectively. By Proposition 1 and Theorem 1, we know that as long as

$$nh^{-1/2}a_n \to 0 \qquad \text{and} \qquad nh^{-1/2}b_n \to \infty \tag{2.7}$$

are satisfied, the optimal convergence rate can be achieved, the true model can be consistently identified. Note that there are d shrinkage parameters (i.e., λ_j, $1 \le j \le d$). Selecting them to satisfy that requirement is challenging. To bypass this difficulty, we follow the idea of Zou (2006), Zhang and Lu (2007), Wang, Li, and Tsai (2007), Zou and Li (2008), to simplify the tuning parameters as

$$\lambda_j = \frac{\lambda_0}{n^{-1/2}\|\tilde{b}_j\|}, \tag{2.8}$$

where $\tilde{b}_j = \{\tilde{\beta}_j(Z_2) - \tilde{\beta}_j(Z_1), \ldots, \tilde{\beta}_j(Z_n) - \tilde{\beta}_j(Z_{n-1})\}^\top$, and $\tilde{\beta}_j$ is the jth column of the unpenalized estimate $\tilde{B}$. Because $\tilde{\beta}_j$ is an estimator with $\lambda_j = 0$, the results of Proposition 1 and Theorem 1 can be applied. If $Z_i, i = 1, \ldots, n$, are quasi-uniform (Eggermont and LaRiccia (2009)), one can verify that as long as $\lambda_0 nh^{-1/2} \to 0$ but $\lambda_0 n^{3/2} \to \infty$, the two conditions listed in (2.7) are satisfied. Consequently, the original d-dimensional problem about $\lambda \in \mathbb{R}^d$ becomes a univariate problem about $\lambda_0 \in \mathbb{R}$.

We select λ_0 as follows. For each $\lambda \ge 0$, the procedure described in Section 2.2 gives an estimator $\hat{B}_\lambda$. Denote by df_λ the number of varying coefficients identified by B_λ with $0 \le df_\lambda \le d$. Then $d - df_\lambda$ is the number of non-varying coefficients. The corresponding RSS_λ is

$$\mathrm{RSS}_\lambda = n^{-2} \sum_{t=1}^{n} \sum_{i=1}^{n} \left\{ Y_i - X_i^\top \hat{\beta}_\lambda(Z_t) \right\}^2 K_h(Z_t - Z_i). \tag{2.9}$$

Then λ_0 is selected according to the following BIC-type criterion

$$\mathrm{BIC}_\lambda = \log(\mathrm{RSS}_\lambda) + df_\lambda \times \frac{\log(nh)}{nh} + (d - df_\lambda) \times \frac{\log(n)}{n}. \tag{2.10}$$

Note that there are two penalty terms in (2.10). In the first, the effective sample size nh is used instead of the original sample size n. The second penalty is for the global constant coefficients, thus the effective sample size is n. When n is large and h tends to 0, the penalty is dominated by the first term. However, for medium sample size, our calculation suggests that adding the second penalty

term is helpful in identifying the model. The tuning parameter can be obtained as

$$\hat{\lambda} = \arg\min_{\lambda} \mathrm{BIC}_{\lambda}.$$

Take $\mathcal{S} = \{j_1, \ldots, j_{d^*}\}$ as an arbitrary model with a total of $0 \le d^* \le d$ varying coefficients (i.e., $X_{ij_1} \ldots, X_{ij_d^*}$). Then, $\mathcal{S}_T = \{1, \ldots, d_0\}$ denotes the true model, and $\mathcal{S}_\lambda = \{j : \|\hat{b}_{\lambda,j}\| > 0\}$ represents the model identified by the proposed estimate $\hat{B}_\lambda$. Consequently, $\mathcal{S}_{\hat{\lambda}}$ represents the model identified by $\hat{B}_{\hat{\lambda}}$.

3. Theoretical Properties

Let $X_{ia} = (X_{i1}, \ldots, X_{id_0})^\top \in \mathbb{R}^{d_0}$, $X_{ib} = (X_{i(d_0+1)}, \ldots, X_{id})^\top \in \mathbb{R}^{d-d_0}$ and, 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 $\hat{\beta}_{b,\lambda} = \{\hat{\beta}_{\lambda,d_0+1}, \ldots, \hat{\beta}_{\lambda,d}\}^\top \in \mathbb{R}^{d-d_0}$.

Proposition 1 (Estimation sparsity)**.** *Assume* (C1)−(C6) *in the appendix hold,* $nh^{-1/2}a_n \to 0$, *and* $nh^{-1/2}b_n \to \infty$. *Then there is a constant vector* $\tilde{\beta}_b$ *such that* $P(\hat{\beta}_{b,\lambda}(z) \equiv \tilde{\beta}_b) \to 1$, *so the identification is consistent. For the generalized varying coefficient model, if* (E1)−(E4) *hold as well, then there is a constant vector* $\tilde{\beta}_b^E$ *such that* $P(\hat{\beta}^E_{b,\hat{\lambda}}(z) \equiv \tilde{\beta}_b^E) \to 1$, *so identification is consistent.*

If the constant coefficients are ideally specified, (2.2) becomes a standard semi-varying-coefficient model that can be estimated by some existing methods; see Fan and Huang (2005) or Xia, Zhang, and Tong (2004). Since this specification is not always available in practice, we call the estimator under the ideal specification the oracle estimator. Specifically, for any z,

$$\hat{\beta}_{ora}(z) = \Big\{\frac{1}{n}\sum_{i=1}^{n} X_{ia}X_{ia}^\top K_h(Z_i - z)\Big\}^{-1}\Big\{\frac{1}{n}\sum_{i=1}^{n} X_{ia}\{Y_i - X_{ib}^\top \beta_{b0}\}K_h(Z_i - z)\Big\}. \quad (3.1)$$

For the generalized varying-coefficient model, the method of Cai, Fan, and Li (2000) can be used to get the oracle estimate of $\beta_{a0}(\cdot)$,

$$\hat{\beta}^E_{ora}(z) = \mathrm{argmin}_{\beta(z)} \sum_{i=1}^{n} \mathcal{L}\Big\{g^{-1}\{X_{ia}^\top \beta_a(z) + X_{ib}^\top \beta_{b0}\}, Y_i\Big\}K_h(z - Z_i). \quad (3.2)$$

The following theorem establishes the oracle property.

Theorem 1 (Oracle property)**.** *Assume* (C1)−(C6) *in the appendix hold. If* $nh^4 = o(1)$, $nh^{-1/2}a_n \to 0$, *and* $nh^{-1/2}b_n \to \infty$, *we have*

$$\sup_z \|\hat{\beta}_{a,\lambda}(z) - \hat{\beta}_{ora}(z)\| = o_p(n^{-2/5}).$$

For the generalized varying coefficient model, if further (E1)−(E4) *hold, we have*

$$\sup_z \|\hat{\beta}^E_{a,\lambda}(z) - \hat{\beta}^E_{ora}(z)\| = o_p(n^{-2/5}).$$

Based on the oracle properties in Theorem 1, most statistical inferences for $\hat{\beta}_{a,\lambda}$ can be made exactly the same as with the oracle estimators. For example, we can construct a simultaneous confidence band, inspired by Fan and Zhang (2000b). Finally, we establish a theorem indicating that the tuning parameters selected by the BIC criterion in (2.10) can identify the true model consistently.

Theorem 2 (Selection consistency)**.** *Assume conditions* (C1)−(C6) *in the appendix hold. Then the tuning parameter* $\hat{\lambda}$ *selected by the BIC criterion* (2.10) *can indeed identify the true model consistently,* $P(\mathcal{S}_{\hat{\lambda}} = \mathcal{S}_T) \to 1$ *as* $n \to \infty$.

4. Numerical Experiments

In this section, we apply the approach to both simulated and actual data to check the performance of the proposed methods in finite samples. In these calculations, we first fit an unpenalized varying coefficient model and estimate $\hat{\beta}(z)$, where the kernel function $K(t) = \exp(-t^2/2)/\sqrt{2\pi}$ is used and the optimal bandwidth is selected via the leaving-one-out cross-validation. The same bandwidth is then used for the proposed procedure, and $\hat{\beta}(z)$ is used as the initial estimator. The optimal tuning parameter $\hat{\lambda}$ is determined by the BIC criterion (2.10).

4.1. Simulation examples

We generated random samples with $p = 7$ from three models:

$$\begin{aligned}
\text{(I)}\quad & Y_i = 2\sin(2\pi Z_i)X_{i1} + 4Z_i(1-Z_i)X_{i2} + 0X_{i3} \\
& \quad +0.5X_{i4} + 0.5X_{i5} + X_{i6} + 0.1X_{i7} + \sigma_e \times e_i, \qquad (4.1)\\
\text{(II)}\quad & Y_i = 3\sin(2\pi Z_i)X_{i1} + 8Z_i(1-Z_i)X_{i2} + \cos^2(2\pi Z_i)X_{i3} \\
& \quad +X_{i4} + 0.5X_{i5} + X_{i6} - 0.5X_{i7} + \sigma_e \times e_i, \qquad (4.2)\\
\text{(III)}\quad & Y_i = 3Z_iX_{i1} + 2\sin(2\pi Z_i)X_{i2} + 15Z_i(1-Z_i)X_{i3} \\
& \quad +X_{i4} - X_{i5} + X_{i6} + 0X_{i7} + \sigma_e \times e_i, \qquad (4.3)
\end{aligned}$$

where $X_{i1} = 1$ and $(X_{i2}, \ldots, X_{i7})^\top$ were generated from a multivariate normal distribution with $\text{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$ for any $2 \le j_1, j_2 \le 7$; the e_is were $N(0,1)$. The *index* variable was $Uniform[0,1]$. The value of σ_e was 0.5. The simulation results are reported in Table 1, with 200 simulation replications conducted for each model setup.

To evaluate estimation accuracy, we considered the relative estimation error (REE, Wang and Xia (2009))

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

Table 1. The simulation results based on 200 simulation replications.

		Correct identification	MREE(%)	
n	Model	frequency	Unpenalized estimate	Oracle estimate
100	Model I	0.49	53.5%	116.8%
200	Model I	0.91	49.7%	102.2%
400	Model I	0.99	47.9%	100.4%
100	Model II	0.52	59.4%	116.3%
200	Model II	0.92	58.7%	110.4%
400	Model II	0.99	57.6%	108.2%
100	Model III	0.72	64.7%	126.8%
200	Model III	0.85	62.5%	113.4%
400	Model III	1.00	56.7%	102.8%

where $\bar{\beta}_j(\cdot)$ is either the unpenalized estimator or the oracle estimator. The corresponding REE value measures the estimation accuracy of $\hat{\beta}_\lambda(Z_i)$ relative to that of $\bar{\beta}_j(Z_i)$ (e.g., unpenalized or oracle). For each model and parameter setting, the medians of REE values (MREE) are summarized in Table 1. The frequency of the experiments with correct model identifications, the identified model has the same varying coefficient terms and invariant coefficient terms as the true model, are also summarized in Table 1.

As one can see from Table 1, all MREE ratios of the penalized estimator to the unpenalized estimator are much less than 100%, clearly indicating that the proposed estimates are more efficient than the unpenalized estimates. Furthermore, for every model and noise level, the frequency of the experiments with correct model identifications steadily increases as the sample size increases, and approaches 100% quickly, which suggests that our BIC criterion (2.10) can indeed identify the true model consistently. Moreover, we find the MREE ratios of the penalized estimator to the oracle estimator approach 100% quickly, which corroborates the oracle properties of the proposed estimator.

4.2. The Boston housing data

To illustrate the usefulness of the proposed procedure, we consider the Boston housing data that concerns the median value of owner-occupied homes for 506 census tracts of Boston from the 1970 census. Following Fan and Huang (2005), we take MEDV (median value of owner-occupied homes in 1,000USD) as the response, LSTAT (the percentage of lower status of the population) as the index variable, and as predictors: 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

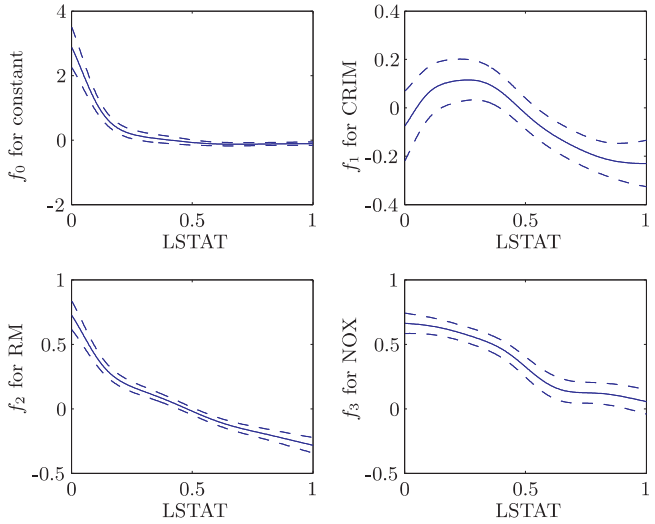

Figure 1. The fitted coefficients for Boston housing data. The solid lines are the estimated coefficients, and the dashed lines are their 95% point-wise confidence bands.

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 for INT) are transformed to have zero mean and unit variance. The index variable LSTAT is transformed so that its marginal distribution is $U[0, 1]$.

First, a standard leave-one-out cross-validation method without penalization suggested a bandwidth $h = 0.1739$. The optimal tuning parameter was then selected by the BIC criterion (2.10). The resulting estimate suggests that INT, CRIM, RM, and NOX have truly varying coefficients depending on the index variable LSTAT, whereas PTRATIO, TAX, and AGE have non-varying coefficients as $-0.1499, 0.0521$, and 0.0124, respectively. The coefficients of the relevant varying coefficients are shown in Figure 1 together with their 95% confidence bands (the dashed lines, see Zhang and Fan (2000b)). Our identification lends support to the model used in Fan and Huang (2005).

5. Some Extensions

In this section, we consider the identification of constant coefficients of the generalized varying coefficients models in the sparse and high-dimensional setting where the number of covariates is larger than the sample size. In this case, the proposed method in Section 2 is not directly applicable for the reasons discussed in Fan and Lv (2008). We propose a two-stage adaptive grouped LASSO approach below. The approach consists of a screening stage to reduce the model

dimension and a selection stage to identify constant coefficients. In the screening stage, we apply the grouped LASSO and basis function expansion to simultaneously select the important variables and estimate the nonzero varying coefficient functions by treating the constant coefficients as varying coefficients. Suppose that the coefficient function $\beta_j(z)$ can be approximated by a linear combination of basis functions, $\beta_j(z) = \sum_{l=1}^{N} \gamma_k B_{jl}(z), 1 \le j \le d$, where $B_{jl}(z)$, $l = 1, 2, \ldots, N$, are basis functions and N is the number of basis functions and allowed to increase with the sample size n. For demonstration, we take that the number of basis functions, N, is the same for all coefficients. When the number of variables d or $d \times N$ is larger than sample size n, the likelihood method might be not applicable. In such case, regularized methods are needed. Applying the grouped LASSO (Yuan and Lin (2006)), consider

$$\operatorname{argmin}_\gamma - \sum_{i=1}^{n} \mathcal{L}\Big\{ g^{-1}\{\sum_{j=1}^{d}\sum_{l=1}^{N} X_{ij}\gamma_k B_{jl}(z), Y_i\Big\} + \sum_{j=1}^{d} \lambda_j \|\gamma_j\|. \tag{5.1}$$

where λ_j is the penalty parameter, $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jN})$ is a N-dimensional coefficient vector corresponding to the jth variable, and $\gamma = (\gamma_1^\top, \ldots, \gamma_d^\top)^\top$. To reduce computational burden, the group coordinate descent algorithm (Fu (1998); Friedman, Hastie, and Tibshirani (2007); Meier, van de Geer, and Bühlmann (2008)) is used to compute the adaptive grouped LASSO estimation defined in (5.1). The BIC type criterion is used to select the spline knots and shrinkage parameters. Under appropriate conditions, following Wei, Huang, and Li (2010), we can show that the adaptive grouped LASSO estimation has the oracle selection property. In the selection stage, the constant coefficients are selected by the method described in Section 2. We report on a simulation study to show the performance of two-stage adaptive grouped LASSO approach. The datasets were generated from a varying coefficients model (VCM) and a logistic varying-coefficient model (Logit-VCM). The response for the former was generated from $N(u_i, 0.1)$ and the latter a Bernoulli random variable with $P(Y_i = 1) = \exp(u_i)/(1 + \exp(u_i))$ for all i, with

$$\begin{aligned} u_i = {} & 3\sin(2\pi Z_i)X_{i1} + 8Z_i(1 - Z_i)X_{i2} + \cos^2(2\pi Z_i)X_{i3} - 0.5X_{i4} - 0.5X_{i6} \\ & + 0.5X_{i7} + 0.5X_{i8}, \end{aligned}$$

where Z_i and $X_i = (X_{i1}, \ldots, X_{id})^\top$ are the same as those in Section 4.1. Since we focus on high-dimensional predictors, $d = 50$ and $d = 150$ were considered respectively. For both scenarios, 200 datasets each with sample size $n = 100$ were generated and fitted. Table 2 shows the average number of identified varying coefficients (avgNV) and constant coefficients (avgNC), median mean absolute

Table 2. Model selection results of two-stage adaptive LASSO estimators based on 200 replications with $n = 100$ and different number of covariates d.

Model	d	avgNV	avgNC	median MAPE	average computational time (in second)
VCM	$d = 50$	3.13	4.03	0.014	19.26
VCM	$d = 150$	2.83	4.24	0.031	69.54
Logit-VCM	$d = 50$	2.91	3.97	0.026	129.53
Logit-VCM	$d = 150$	2.57	4.43	0.039	166.53

prediction error (median MAPE), and the average computational time to perform the two-stage procedures, in seconds. From Table 2, we can see the two-stage approach can help us quickly perform model section procedure with acceptable computational time; we used a computer with Intel CPU750. Limited as it is, however, this short study provides, we hope, a good picture of the performance of our two-stage adaptive LASSO approach.

6. Conclusion

We propose in this article a method which is able to identify constant coefficients and make nonparametric estimation simultaneously. Under some mild regular conditions, the sparsity and oracle efficiency for the proposed estimators can be established. A BIC-type criterion was suggested to choose the regularization parameters. An algorithm was developed based on local quadratic approximation to the criterion function. Numerical experiments indicated that the proposed procedure was very effective in identifying constant coefficients in a semi-varying coefficients model and in estimating the regression coefficient functions. Although our proposal is based on a LASSO method due to its simplicity, similar ideas can be extended to other shrinkage methods, such as the nonnegative garrotte, bridge regression, and SCAD. Further research includes extending the proposed procedure to longitude data and to the case with a diverging number of parameters (Kim, Choi, and Oh (2008), Lam and Fan (2008)).

More numerical studies need to be done to evaluate the the finite sample performance of the two-stage adaptive LASSO approach, and to compare the two-stage adaptive LASSO approach with other methods. In addition, more work is needed to establish the the oracle property of the two-stage adaptive LASSO approach in sparse and high-dimensional settings when the number of variables is larger than the sample size.

Acknowledgements

We are very grateful to an associate editor and two anonymous referees for their constructive comments. The research was partially supported by the Education Department of Nature Science Research of Guizhou Province (Grant No. 2010028), the Nomarch Foundation of Guizhou Province (Grant No. 2010025), China, and a grant from the Risk Management Institute, National University of Singapore.

Appendix: Assumptions and Proofs

To study the asymptotic properties of the proposed method, standard regularity conditions are needed (Fan and Huang (2005)).

(C1) For an $s > 2$, $E|Y_i|^{2s} < \infty$ and $E\|X_i\|^{2s} < \infty$.

(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 non-singular 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 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 $O_p(\log n/n)$; see for example Janson (1987). For an arbitrary *index* value $z \in [0, 1]$, let z^* be its nearest neighbor among the observed *index* values, $z^* = \arg\min_{\tilde{z} \in \{Z_t : 1 \le t \le n\}} |z - \tilde{z}|$. Under (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 $\{\beta(Z_t) : 1 \le t \le n\}$.

To develop the asymptotic results for the method in the exponential family, we impose regularity conditions (Cai, Fan, and Li (2000)). Let $q_j(s, y) = (\partial^j / \partial s^j)\mathcal{L}\{g^{-1}(s), y\}$.

(E1) The function $q_2(s, y) < 0$ for $s \in \mathbb{R}$ and y in the range of the response variable.

(E2) The density function of Z, $f(z)$,, $\Gamma(z)$, $\mathrm{Var}\{Y | Z = z, X = x\}$, the first order derivative of $\mathrm{Var}\{Y | Z = z, X = x\}$ and the third order derivatives of $g(\cdot)$ are continuous at z. Furthermore, $f(z) > 0$ and $\Gamma(u) > 0$.

(E3) $E\{|X|^3|Z=z\}$ is continuous at z.

(E4) $E(Y^4|Z=z, X=x)$ is bounded in a neighborhood of z.

Lemma A.1. *Suppose $(\xi_i, Z_i), i=1,\ldots,n$ are i.i.d random vectors, where ξ_is are scalar random variables. Suppose $E|\xi_i|^s<\infty$ 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*

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

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

The proof of the Lemma can be found in Mack and Silverman (1982), or Fan and Zhang (2000a).

Lemma A.2. *If* (C1)−(C6) *hold, and $nh^{-1/2}a_n\to 0$, then we must have*

$$n^{-1}\sum_{t=1}^{n}\|\hat\beta(Z_t)-\beta(Z_t)\|^2 = O_p\{(nh)^{-1/2}\}.$$

Proof. For an arbitrary matrix $A=(a_{ij})$, $\|A\|^2=\sum a_{ij}^2$. 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 $v_1,\ldots,v_d$. Let $B_0=\{\beta_0(Z_1),\ldots,\beta_0(Z_n)\}\in\mathbb{R}^{n\times d}$. 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

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

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

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

where $b_{0j} = \{\beta_{0j}(Z_2) - \beta_{0j}(Z_1), \ldots, \beta_{0j}(Z_n) - \beta_{0j}(Z_{n-1})\}$. By simple algebraic calculation and the fact that $\|b_{0j}\| = 0$ for any $j > d_0$, we have

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

where $\hat{\Sigma}(Z_t) = n^{-1}\sum_i X_i X_i^\top K_h(Z_t - Z_i)$, $\hat{e}_t = (n^{-1}h)^{1/2}\sum_{i=1}^{n} X_i\{X_i^\top[\beta(Z_t) - \beta(Z_i)] + e_i\}K_h(Z_t - Z_i)$. 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

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

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

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

After some algebraic calculations, we have $n^{-1}\|\hat{e}\|^2 = O_p(1)$. By Lemma A.1 and (C3), we have $P(\lambda_{\min} \to \lambda_0^{\min}) \to 1$, 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 (C2), (C3), and Lemma A.1, 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, $nh^{-1/2}a_n \to 0$. Lastly, note that the first term in (A.3) is quadratic in C while the second term is linear in C. As long as C is sufficiently large, the right hand side of (A.3) is guaranteed to be positive with probability arbitrarily close to 1. This proves (A.1). The proof is complete.

Lemma A.3. *If* (E1)−(E4), (C5)−(C6) *hold and* $nh^{-1/2}a_n \to 0$, *then*

$$n^{-1}\sum_{t=1}^{n}\|\hat{\beta}^E(Z_t)-\beta(Z_t)\|^2 = O_p\{(nh)^{-1}\}.$$

Proof. As in Lemma A.2, it suffices to show that for any small probability $\epsilon > 0$, we can always find a constant $C > 0$, such that

$$\liminf_n P\Big\{\inf_{n^{-1}\|u\|=C} Q_\lambda^E(B_0+(nh)^{-1/2}u) > Q_\lambda^E(B_0)\Big\} = 1-\epsilon. \tag{A.4}$$

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

$$\begin{aligned}
&hn^{-1}\Big\{Q_\lambda^E(B_0+(nh)^{-1/2}u) - Q_\lambda^E(B_0)\Big\}\\
&\quad= hn^{-1}\sum_{t=1}^{n}\sum_{i=1}^{n}\mathcal{L}\Big\{g^{-1}\big\{X_i^\top\{\beta_0(Z_t)+(nh)^{-1/2}u_t\}\big\},Y_i\Big\}K_h(Z_t-Z_i)\\
&\qquad-\frac{h}{n}\sum_{t=1}^{n}\sum_{i=1}^{n}\mathcal{L}\Big\{g^{-1}\big\{X_i^\top\beta_0(Z_t)\big\},Y_i\Big\}K_h(Z_t-Z_i)\\
&\qquad+\frac{h}{n}\sum_{j=1}^{d}\lambda_j\Big(\|b_{0j}+(nh)^{-1}v_j\|-\|b_{0j}\|\Big)\\
&\quad:= E_1+E_2,
\end{aligned}$$

where $b_{0j} = \{\beta_{0j}(Z_2)-\beta_{0j}(Z_1),\ldots,\beta_{0j}(Z_n)-\beta_{0j}(Z_{n-1})\}$. Using Taylor expansion of $\mathcal{L}\{g^{-1}(\cdot),y\}$, we have

$$E_1 = n^{-1}\sum_{t=1}^{n}\Big\{u_t^\top W_t + 2^{-1}u_t^\top\Delta_n(Z_t)u_t\{1+o_p(1)\}\Big\},$$

where

$$W_t = (n^{-1}h)^{1/2}\sum_{i=1}^{n}q_1\{X_i^\top\beta_0(Z_t),Y_i\}X_iK_h(Z_t-Z_i),$$

$$\Delta_n(Z_t) = n^{-1}\sum_{i=1}^{n}q_2\{X_i^\top\beta_0(Z_t),Y_i\}X_iX_i^\top K_h(Z_t-Z_i),$$

and η_i is between $X_i^\top\{\beta_0(Z_t)+(nh)^{-1/2}u_t\}$ and $X_i^\top\beta_0(Z_t)$. Let $\hat\lambda_t^{\min}$ be the smallest eigenvalue of $\Delta_n(Z_t)$, $\hat\lambda_{\min}=\min\{\hat\lambda_t^{\min}, t=1,\dots,n\}$, and $W=(W_1,\dots,W_n)^\top \in \mathbb{R}^{n\times d}$. We have

$$\begin{aligned} E_1 &\geq n^{-1}\sum_{t=1}^{n}\Big\{\|u_t\|^2\hat\lambda_t^{\min}-\|u_t\|\|W_t\|\{1+o_p(1)\}\Big\} \\ &\geq \hat\lambda_{\min}\Big\{n^{-1}\sum_t\|u_t\|^2\Big\}-n^{-1}\Big(\sum_{t=1}^{n}\|u_t\|\cdot\|W_t\|\{1+o_p(1)\}\Big) \\ &\geq \hat\lambda_{\min}\{n^{-1}\|u\|^2\}-(n^{-1}\|u\|^2)^{1/2}(n^{-1}\|W\|^2)^{1/2}\{1+o_p(1)\}. \end{aligned}$$

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

$$E_1 \geq \hat\lambda_{\min}\times C^2 - C\times(n^{-1}\|W\|^2)^{1/2}\{1+o_p(1)\}.$$

After some algebraic calculations, we have $n^{-1}\|W\|^2=O_p(1)$. The rest of the proof follows that of Lemma A.2.

Proof of Proposition 1. 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, $\|\hat b_{\lambda,d}\|=0$, since $b_{\lambda,d}=\{\beta_d(Z_2)-\beta_d(Z_1),\dots,\beta_d(Z_n)-\beta_d(Z_{n-1})\}^\top$, then $\nu=(\beta_d(Z_1),\dots,\beta_d(Z_n))^\top$ is the solution to the normal equation

$$0=\frac{\partial Q_\lambda(B)}{\partial\nu}=\alpha_1+\alpha_2, \tag{A.5}$$

where α_1 is a n-dimensional vector with its tth component given by α_{1t}. If $t=1$, then $\alpha_{1t}=\lambda_j\{\beta_d(Z_1)-\beta_d(Z_2)\}/\|b_d\|$; if $1<t<n$, then $\alpha_{1t}=\lambda_j\{2\beta_d(Z_t)-\beta_d(Z_{t-1})-\beta_d(Z_{t+1})\}/\|b_d\|$; if $t=n$, then $\alpha_{1t}=\lambda_j(\beta_d(Z_n)-\beta_d(Z_{n-1}))/\|b_d\|$. α_2 is a n-dimensional vector with its tth component given by $\alpha_{2t}=-2\sum_{i=1}^n X_{id}\{Y_i-X_i^\top\hat\beta_\lambda(Z_t)\}K_h(Z_t-Z_i)$. By standard arguments of kernel smoothing, and applying Lemma A.1 and Lemma A.2, we have $\|\alpha_2\|=O_p(nh^{-1/2})$. On the other hand, we know that $P(\|\alpha_1\|>\|\alpha_2\|)\to 1$. Consequently, we know that, with probability tending to one, (A.5) cannot hold. This implies that $\hat b_{\lambda,j}$ must be located 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 that $P(\|\hat b_{\lambda,j}\|=0)\to 1$. This completes the proof.

For the generalized model, we only need to prove that $P(\|\hat b^E_{\lambda,j}\|=0)\to 1$ with $j=d$. The proofs for $d_0<j<d$ are similar. If the claim is not true, $\|\hat b^E_{\lambda,d}\|=0$, since $b^E_{\lambda,d}=\{\beta^E_d(Z_2)-\beta^E_d(Z_1),\dots,\beta^E_d(Z_n)-\beta^E_d(Z_{n-1})\}^\top$, then $\nu=(\beta^E_d(Z_1),\dots,\beta^E_d(Z_n))^\top$ satisfies

$$0=\frac{\partial Q^E_\lambda(B)}{\partial\nu}=\alpha_1+\alpha_2, \tag{A.6}$$

where α_1 is a n-dimensional vector with its tth component given by α_{1t}. If $t = 1$, then $\alpha_{1t} = \lambda_j\{\beta_d^E(Z_1) - \beta_d^E(Z_2)\}/\|b_d^E\|$; if $1 < t < n$, then $\alpha_{1t} = \lambda_j\{2\beta_d^E(Z_t) - \beta_d^E(Z_{t-1}) - \beta_d^E(Z_{t+1})\}/\|b_d^E\|$; if $t = n$, then $\alpha_{1t} = \lambda_j\{\beta_d^E(Z_n) - \beta_d^E(Z_{n-1})\}/\|b_d^E\|$. α_2 is a n-dimensional vector with its tth component given by

$$\alpha_{2t} = -2\sum_{i=1}^{n} X_{id} q_1\{g^{-1}\{X_i^\top \hat{\beta}_\lambda(Z_t)\}, Y_i\} K_h(Z_t - Z_i).$$

By standard arguments of kernel smoothing, and applying Lemma A.1 and Lemma A.2, we have $\|\alpha_2\| = O_p(nh^{-1/2})$. The rest of the proof follows that of the simple varying coefficient model and is omitted.

Proof of Theorem 1. By Lemma A.2, we know that $\hat{b}_{\lambda,j} = 0$ $(d_0 < j \le d)$ with probability tending to one. Let $\mathcal{L} \in \mathbb{R}^{d_0}$, with jth component given by $\mathcal{L}_j$. If $t = 1$, then $\mathcal{L}_j = \lambda_j\{\beta_j(Z_1) - \beta_j(Z_2)\}/\|b_{\lambda,j}\|$; if $1 < t < n$, then $\mathcal{L}_j = \lambda_j\{2\beta_j(Z_t) - \beta_j(Z_{t+1}) - \beta_j(Z_{t-1})\}/\|b_{\lambda,j}\|$; if $t = n$, then $\mathcal{L}_j = \lambda_j\{2\beta_j(Z_n) - \beta_j(Z_{n-1})\}/\|b_{\lambda,j}\|$. Consequently, we know that $\hat{\beta}_{a,\lambda}(Z_t)$ must solve

$$\frac{1}{n}\sum_{i=1}^{n} X_{ia}\big\{Y_i - X_{ia}^\top \hat{\beta}_{a,\lambda} - X_{ib}^\top \hat{\beta}_{b,\lambda}\big\} K_h(Z_i - Z_t) + n^{-1}\mathcal{L} = 0,$$

which implies that $\hat{\beta}_{a,\lambda}(Z_t)$ is of the form

$$\hat{\beta}_{a,\lambda}(Z_t) = \Big\{\Sigma_1(Z_t)\Big\}^{-1}\Big\{n^{-1}\sum_{i=1}^{n} X_{ia}\{Y_i - X_{ib}^\top \hat{\beta}_{b,\lambda}\} K_h(Z_i - Z_t) + n^{-1}\mathcal{L}\Big\},$$

where $\Sigma_1(Z_t) = n^{-1}\sum_{i=1}^{n} X_{ia}X_{ia}^\top K_h(Z_i - Z_t)$. Comparing with the oracle estimator (3.1), we know that

$$\begin{aligned}
&\|\hat{\beta}_{a,\lambda}(Z_t) - \hat{\beta}_{ora}(Z_t)\| \\
&\quad = \Big\|\{\Sigma_1(Z_t)\}^{-1}\big\{n^{-1}\mathcal{L} + \Sigma_2(Z_t)\{(\hat{\beta}_b - \beta_{b0}) + (\beta_{b0} - \hat{\beta}_{b,\lambda})\}\big\}\Big\| \\
&\quad \le \lambda_{1,\min}^{-1}\|n^{-1}\mathcal{L}\| + \lambda_{1,\min}^{-1}\lambda_{2,\max}\|(\hat{\beta}_b - \beta_{b0})\| + \lambda_{1,\min}^{-1}\lambda_{2,\max}\|(\beta_{b0} - \hat{\beta}_{b,\lambda})\| \\
&\quad := J_1 + J_2 + J_3,
\end{aligned}$$

where $\Sigma_2(Z_t) = \{n^{-1}\sum_{i=1}^{n} X_{ib}X_{ib}^\top K(Z_i - Z_t)\}$, $\lambda_{1,\min} = \min\{\lambda_{\min}(\Sigma_1(Z_t)), t = 1, \ldots, n\}$, and $\lambda_{2,\max} = \max\{\lambda_{\max}(\Sigma_2(Z_t)), t = 1, \ldots, n\}$. For J_1, applying Lemma A.1, we have $J_1 \le C\{n\lambda_{1,\min}\}^{-1} d_0 a_n = o_p(n^{-2}h^{1/2}) = o_p(n^{-2/5})$. By Theorem 4.1 and Lemma A.1, we have $J_2 = O_p(n^{-1/2})$. By Lemma A.2, we have $J_3 = o_p(n^{-2/5})$. Therefore, the theorem follows.

Now consider the generalized varying coefficient model. By Lemma A.3, we know that $\hat{b}_{\lambda,j} = 0$ $(d_0 < j \le d)$ with probability tending to one. Let $\mathbb{L} \in \mathbb{R}^{d_0}$,

with jth component given by $\mathbb{L}_j$. If $t=1$, then $\mathbb{L}_j=\lambda_j\{\beta_j^E(Z_1)-\beta_j^E(Z_2)\}/\|b_{\lambda,j}^E\|$; if $1<t<n$, then $\mathbb{L}_j=\lambda_j\{2\beta_j^E(Z_t)-\beta_j^E(Z_{t+1})-\beta_j^E(Z_{t-1})\}/\|b_{\lambda,j}^E\|$; if $t=n$, then $\mathbb{L}_j=\lambda_j\{2\beta_j^E(Z_n)-\beta_j^E(Z_{n-1})\}/\|b_{\lambda,j}^E\|$. Consequently, using a Taylor expansion, we know that $\hat{\beta}_{a,\lambda}(Z_t)$ must solve

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\mathcal{L}\Big\{g^{-1}\{X_i^\top\beta(Z_t)\},Y_i\Big\}\\
&\quad+\Big\{n^{-1}\sum_{i=1}^{n}q_1\{X_i\beta(Z_t),Y_i\}X_{ia}(\hat{\beta}_a(Z_t)-\beta_a(Z_t))K_h(Z_t-Z_i)\Big\}\\
&\quad+\frac{1}{2}(\hat{\beta}_a(Z_t)-\beta_a(Z_t))^\top\\
&\quad\times\Big\{n^{-1}\sum_{i=1}^{n}q_2\{X_i\beta(Z_t),Y_i\}X_{ia}X_{ia}^\top K_h(Z_t-Z_i)\Big\}(\hat{\beta}_a(Z_t)-\beta_a(Z_t))\\
&\quad+\frac{1}{6}\sum_{i=1}^{n}q_3\Big\{X_i\bar{\beta}(Z_t),Y_i\Big\}\Big\{X_{ia}^\top(\hat{\beta}_a(Z_t)-\beta_a(Z_t))\Big\}^3K_h(Z_t-Z_i)+n^{-1}\mathbb{L}=0,
\end{aligned}
$$

where $\bar{\beta}(Z_t)$ is between $\hat{\beta}(Z_t)$ and $\beta(Z_t)$.

Using the arguments used for Theorem 2 in Carroll et al. (1997), it can be shown that

$$
\begin{aligned}
\hat{\beta}_a(Z_t)-\beta(Z_t)=&\Big\{n^{-1}\sum_{i=1}^{n}q_2\{X_i\beta(Z_t),Y_i\}X_{ia}X_{ia}^\top K_h(Z_t-Z_i)\Big\}^{-1}\\
&\times\Big\{n^{-1}\sum_{i=1}^{n}q_1\{X_i\beta(Z_t),Y_i\}X_{ia}K_h(Z_t-Z_i)+n^{-1}\mathbb{L}\Big\}\\
&+o_p\big\{(nh)^{-1/2}\big\}.
\end{aligned}
$$

From (A.10) in the Appendix of Cai, Fan, and Li (2000), we obtain

$$
\begin{aligned}
\hat{\beta}_{ora}(Z_t)-\beta(Z_t)=&\Big\{n^{-1}\sum_{i=1}^{n}q_2\{X_i\beta(Z_t),Y_i\}X_{ia}X_{ia}^\top K_h(Z_t-Z_i)\Big\}^{-1}\\
&\times\Big\{n^{-1}\sum_{i=1}^{n}q_1\{X_i\beta(Z_t),Y_i\}X_{ia}K_h(Z_t-Z_i)\Big\}+o_p\big\{(nh)^{-1/2}\big\}.
\end{aligned}
$$

Comparing with the oracle estimator (3.2), we know that

$$
\|\hat{\beta}_{a,\lambda}(Z_t)-\hat{\beta}_{ora}(Z_t)\|=\|\Sigma_3(Z_t)^{-1}\big\{n^{-1}\mathbb{L}\big\}\|\le\lambda_{3,\min}^{-1}\|n^{-1}\mathbb{L}\|:=J_4,
$$

where $\Sigma_3(Z_t)=n^{-1}\sum_{i=1}^{n}q_2\{X_i\beta(Z_t),Y_i\}X_{ia}X_{ia}^\top K_h(Z_t-Z_i)$, $\lambda_{3,\min}=\min\{\lambda_{\min}(\Sigma_3(Z_t)),t=1,\ldots,n\}$. For J_4, applying Lemma A.1, we have $J_4\le C\{n\lambda_{3,\min}\}^{-1}d_0a_n=o_p(n^{-2}h^{1/2})=o_p(n^{-2/5})$. This completes the proof.

Proof of Theorem 2. For an arbitrary model $\mathcal{S}$, we say it is underfitted if it misses at least one variable with non-zero coefficient, i.e., $\mathcal{S} \not\supset \mathcal{S}_T$; it is overfitted if $\mathcal{S}$ covers all relevant variables with varying coefficient, but also includes at least one predictor with constant coefficient, i.e., $\mathcal{S} \supset \mathcal{S}_T$; but $\mathcal{S} \neq \mathcal{S}_T$; Then, according to whether the model $\mathcal{S}_\lambda$ is underfitted, correctly fitted, or overfitted, we can create three mutually exclusive sets $\mathbb{R}_+ = \{\lambda \in \mathbb{R}^d : \mathcal{S}_\lambda \supset \mathcal{S}_T, \mathcal{S}_\lambda \neq \mathcal{S}_T\}$, $\mathbb{R}_0 = \{\lambda \in \mathbb{R}^d : \mathcal{S}_\lambda = \mathcal{S}_T\}$, $\mathbb{R}_- = \{\lambda \in \mathbb{R}^d : \mathcal{S}_\lambda \not\supset \mathcal{S}_T\}$. Following Wang and Leng (2007), we define a reference tuning parameter sequence λ_n according to (2.8) with $\lambda_0 = n^{-3/2}h\log(n)$. It follows that such a tuning parameter sequence satisfies the technical conditions as specified in (2.7). Consequently, we know that $P(\mathcal{S}_{\lambda_n} = \mathcal{S}_T) \to 1$. Then, the theorem can be proved by comparing BIC_{λ_n} and BIC_λ. We consider two cases separately.

Case 1. (Underfitted model) Recall that $\hat{B}_\lambda$ automatically determines a model $\mathcal{S}_\lambda$. Under such a model $\mathcal{S}_\lambda$, we can define another unpenalized estimate $\tilde{B}_{\mathcal{S}_\lambda}$ as

$$\tilde{B}_{\mathcal{S}_\lambda} = \arg \min_{\{\|b_j\|=0, \forall j \notin \mathcal{S}_\lambda\}} \sum_{t=1}^{n}\sum_{i=1}^{n} \Big\{Y_i - X_i^\top \beta(Z_t)\Big\}^2 K_h(Z_t - Z_i).$$

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 have $\mathrm{RSS}_\lambda \geq \mathrm{RSS}_{\mathcal{S}_\lambda}$, where

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

Due to the fact that $B_{\mathcal{S}} \neq B_0$ for any $\mathcal{S} \not\supset \mathcal{S}_T$, we know that

$$\begin{aligned}
\mathrm{RSS}_\lambda - \mathrm{RSS}_{\lambda_n} &> n^{-1}\sum_t \Big\{\hat{\beta}_{\mathcal{S}_\lambda}(Z_t) - \hat{\beta}_\lambda(Z_t)\Big\}^\top \hat{\Sigma}(Z_t)\Big\{\hat{\beta}_{\mathcal{S}_\lambda}(Z_t) - \hat{\beta}_\lambda(Z_t)\Big\} \\
&\geq \hat{\lambda}_{\min}\Big\{n^{-1}\|\hat{\beta}_{\mathcal{S}_\lambda}(Z_t) - \hat{\beta}_\lambda(Z_t)\|^2\Big\} \\
&= \hat{\lambda}_{\min}\Big\{\|\hat{B}_{\mathcal{S}_\lambda} - \hat{B}_\lambda\|\Big\} \to \lambda_0^{\min}\|B_{\mathcal{S}_\lambda} - B_0\| > 0,
\end{aligned}$$

in probability, where $\hat{\Sigma}(Z_t) = n^{-1}\sum_i X_i X_i^\top K_h(Z_t - Z_i)$. This, together with the definition of BIC_λ, suggest that

$$P(\inf_{\lambda \in \mathbb{R}_-} \mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}) \to 1. \tag{A.7}$$

Case 2. (Overfitted model) Consider an arbitrary $\lambda \in \mathbb{R}_+$ (i.e. $\mathcal{S}_\lambda \supset \mathcal{S}_T$ but $\mathcal{S}_\lambda \neq \mathcal{S}_T$). For an unpenalized estimator $\tilde{\beta}(Z_t)$, we must have $\sum_{i=1}^n\{Y_i -$

$X_i^\top \tilde{\beta}(Z_t)\}X_i K_h(Z_t - Z_i) = 0$. Thus,

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

It follows that

$$\begin{aligned} &\log(\mathrm{RSS}_\lambda) - \log(\mathrm{RSS}_F) = \log\Big(\frac{\mathrm{RSS}_\lambda}{\mathrm{RSS}_F}\Big) \\ &= \log\bigg(\frac{\mathrm{RSS}_F}{\mathrm{RSS}_F} + n^{-1}\mathrm{RSS}_F \sum_{t=1}^n \Big\{\tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}\Big\}^\top \hat{\Sigma}(Z_t)\Big\{\tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}\Big\}\bigg) \\ &\geq -n^{-1}\mathrm{RSS}_F \sum_{t=1}^n \Big\{\tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}\Big\}^\top \hat{\Sigma}(Z_t)\Big\{\tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}\Big\} \\ &\geq \frac{\hat{\lambda}_{\min}}{\mathrm{RSS}_F}\Big(n^{-1}\sum_{t=1}^n \|\tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}\|^2\Big) = -|O_p\{(nh)^{-1}\}|, \end{aligned} \tag{A.9}$$

where the last equality is due to the fact that

$$\frac{\|\tilde{\beta}(Z_t) - \tilde{\beta}_{\mathcal{S}_\lambda}(Z_t)\|^2}{n} \leq \frac{\|\tilde{\beta}_{\mathcal{S}_\lambda}(Z_t) - \beta_0(Z_t)\|^2}{n} + \frac{\|\tilde{\beta}(Z_t) - \beta_0(Z_t)\|^2}{n} = O_p\{(nh)^{-1}\}$$

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

$$\log(\mathrm{RSS}_\lambda) - \log(\mathrm{RSS}_F) = O_p\{(nh)^{-1}\}. \tag{A.10}$$

Combining the results of (A.9) and (A.10) we know that $\inf_{\lambda\in\mathbb{R}_+} \log \mathrm{RSS}_\lambda - \log \mathrm{RSS}_{\lambda_n} \geq -|O_p\{(nh)^{-1}\}|$. Consequently, it follows that

$$\begin{aligned} \inf_{\lambda\in\mathbb{R}_+} \mathrm{BIC}_\lambda - \mathrm{BIC}_{\lambda_n} &= \big(\inf_{\lambda\in\mathbb{R}_+} \log \mathrm{RSS}_\lambda - \log \mathrm{RSS}_{\lambda_n}\big) \\ &\quad + (df_\lambda - df_{\lambda_n}) \times \Big\{\frac{\log(nh)}{nh} - \frac{\log(n)}{n}\Big\} \\ &\geq -|O_p\{(nh)^{-1}\}| + (df_\lambda - df_{\lambda_n}) \times \frac{\log(nh)}{nh}\big\{1 + o(1)\big\} \\ &\geq -|O_p\{(nh)^{-1}\}| + \frac{\log(nh)}{nh}\big\{1 + o(1)\big\}, \end{aligned} \tag{A.11}$$

where the last equality is due to the following. First, because the reference sequence λ_n satisfies (2.7), by Proposition 1 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 $P(df_{\lambda_n} \geq d_0+1) \to 1$. Third, note that $\log(nh) \propto \log n \to \infty$. 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 side of (A.11) is guaranteed to be positive. Consequently,

$$P\big(\inf_{\lambda \in \mathbb{R}_+} \mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}\big) \to 1. \tag{A.12}$$

Combining the results from (A.7) and (A.12), we have

$$P\big(\inf_{\lambda \in \mathbb{R}_- \cup \mathbb{R}_+} \mathrm{BIC}_\lambda > \mathrm{BIC}_{\lambda_n}\big) \to 1. \tag{A.13}$$

Then (A.13) 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 as unfavorable as our reference sequence λ_n. Consequently, we know that $P(\mathcal{S}_{\hat{\lambda}} = \mathcal{S}_T) \to 1$. This completes the proof.

References

Breiman, L. (1995). Subset selection using nonnegative garrote. *Techonometrics* **37**, 373-384.

Cai, Z., Fan, J. and Li, R. (2000). Efficient estimation and inferences for varying- coefficient models. *J. Amer. Statist. Assoc.* **95**, 888-902.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**, 477-489.

Chen, R. and Tsay, R. S. (1993). Functional-coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88**, 298-308.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-489.

Eggermont, P. P. B. and LaRiccia, V. N. (2009). *Maximum Penalized Likelihood Estimation. Volume II : Regression.* Springer-Verlag, New York.

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. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99**, 710-723.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27**, 1491-1518.

Fan, J. and Zhang, J. T. (2000a). Two-step estimation of functional linear model with application to longitudinal data. *J. Roy. Statist. Soc. Ser. B* **62**, 303-322.

Fan, J. and Zhang, W. (2000b). Simultaneous confidence bands and hypotheses testing in varying-coefficient models. *Scand. J. Statist.* **27**, 715-731.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface* **1**, 179-195

Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153-193.

Friedman, J., Hastie, T. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1**, 302-332.

Fu, W. J. (1998). Penalized regression: the bridge versus the LASSO. *J. Comput. Graph. Statist.* **7**, 397-416.

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

Hastie, T. J. and Tibshirani, R. J. (1993). Varying-Coefficient models. *J. Roy. Statist. Soc. 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. *Statist. Sinica* **14**, 763-788.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617-1642.

Janson, S. (1987). Maximal spacing in several dimensions. *Ann. Probab.* **15**, 274-280.

Lam, C. and Fan, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *Ann. Statist.* **36**, 2232-2260.

Leng, C. (2009). A simple approach for varying-coefficient model selection. *J. Statist. Plann. Inference* **139**, 2138-2146.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.* **36**, 261-286.

Kim, Y., Choi, H. and Oh, H. S. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Amer. Statist. Assoc.* **103**, 1665-1673.

Knight, K. and Fu, W. (2000). Asymptotics for LASSO-type estimators. *Ann. Statist.* **28**, 1356-1378.

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. verw. Gebiete* **61**, 405-415.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Meier, L., van de Geer, S. and Bühlmann, P. (2008). The Group Lasso for logistic regression. *J. Roy. Statist. Soc. Ser. B*, **70**, 53-71.

Park, M. Y. and Hastie, T. (2007). An L_1 regularization-path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B* **69**, 659-667.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**, 1486-1494.

Wang, H. and Leng, C. (2007). Unified LASSO estimation via least squares approximation. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Wang, L., Li, H. and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103**, 1556-1569.

Wang, H., Li, G. and Tsai, C. L. (2007). Regression coefficient and autoregressive order shrinkage and selection via LASSO. *J. Roy. Statist. Soc. Ser. B* **69**, 63-78.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Amer. Statist. Assoc.* **104**, 747-757.

Wang, H. J., Zhu, Z. and Zhou, J. (2009). Quantile regression in partially linear varying coefficient models. *Ann. Statist.* **37**, 3841-3866.

Wei, X, Huang, J. and Li, H. (2010). Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica*, (to appear).

Xia, Y., Zhang, Y. and Tong, H. (2004). Efficient estimation for semivarying coefficient models. *Biometrika* **91**, 661-681.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Yuan, M. and Lin, Y. (2007), On the nonnegative garrote estimator. *J. Roy. Statist. Soc. Ser. B* **69**, 143-161.

Zhang, H. H. and Lin, Y. (2003). Component selection and smoothing in smoothing spline analysis of variance models. *Ann. Statist.* **34**, 2272-2297.

Zhang, W. Y., Lee, S. and Song, X. (2002). Local polynomial fitting in semivarying coefficient models. *J. Multivariate Anal.* **82**, 166-188.

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

Zhang, H. H. and Lu, W. (2007). Adaptive LASSO for Cox's proportional hazard model. *Biometrika* **94**, 691-703.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509-1566.

Department of mathematics, Capital Normal University, Beijing, China.

E-mail: hutaomath@yahoo.com.cn

Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore.

E-mail: staxyc@nus.edu.sg

(Received May 2010; accepted May 2011)